

THE UNIVERSITY OF CHICAGO

DISMANTLING THE SCHOOL-TO-PRISON PIPELINE: THREE ESSAYS ON THE
POLICY IMPLICATIONS OF SYSTEMS THINKING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
DAVID MCMILLON

CHICAGO, ILLINOIS

JUNE 2021

Copyright © 2021 by David McMillon
All Rights Reserved

To Joel McMillon, for giving me faith in the possibilities. That common men are capable of such radical love proves that the world can be saved, even if it never is.

Information is not knowledge.

-ALBERT EINSTEIN

Table of Contents

LIST OF FIGURES	viii
LIST OF TABLES	x
ACKNOWLEDGMENTS	xi
ABSTRACT	xiv
1 OVERVIEW	1
1.1 Background	1
1.2 Methodological Philosophy	3
2 A DYNAMICAL SYSTEMS ANALYSIS OF CRIMINAL BEHAVIOR AMONG NATIONAL LONGITUDINAL SURVEY OF YOUTH RESPONDENTS	5
ABSTRACT	6
2.1 Introduction	6
2.1.1 Background	6
2.1.2 Systems Approach	7
2.1.3 Racial disparities	8
2.1.4 State dependence	9
2.1.5 Overview	10
2.2 The Model	10
2.2.1 Modeling overview	10
2.2.2 States in the Model	11
2.2.3 Transitions in the Model	12
2.2.4 Equilibrium	13
2.2.5 Effect of Independent Parameter Changes on X^*	14
2.2.6 $\alpha_{xa} : X \rightarrow A$ is a complex, but important transition	16
2.3 Data: NLSY97	16
2.4 Methods	19
2.4.1 Assignment to model state in each round	19
2.5 Results for State and Parameter Estimation	22
2.5.1 Descriptive Results	22
2.5.2 Statistical estimation of transition rates by race and sex	25
2.5.3 Results of the Sensitivity Simulation	30
2.6 Results on Racial Differences via Simulations	36
2.7 Discussion	40
2.7.1 Summary	40
2.7.2 Real World Policy	41
2.7.3 Simplifying Assumptions	42
2.8 Appendix	43
2.8.1 Simple 3D Model: Model Structure	43
2.8.2 Simple 3D Model: Long Run Equilibrium Analytic Expression	46

2.8.3	Simplified 3D Model: Long Run Equilibrium Dependence on Parameters	46
2.8.4	Computing the equilibrium of the 5D System (1)	48
2.8.5	Empirical Compartment Sizes	51
2.8.6	Transition Probabilities	53
2.8.7	Systems Effects	55
3	A THEORETICAL ANALYSIS OF BEHAVIORAL INFRACTIONS AND SCHOOL DISCIPLINARY POLICY	59
ABSTRACT		60
3.1	Introduction	60
3.1.1	The Contributions	61
3.2	Literature Review	64
3.2.1	Approaches to School Discipline, Deterrence, and Student Affect	64
3.2.2	Dynamics of Deviant Behavior	67
3.2.3	Peer Effects and Social Interactions in Misbehavior	68
3.3	Model	70
3.3.1	Behavioral Decisions	71
3.3.2	The Dynamics of Affect	74
3.3.3	The School Official	78
3.4	Simplification: No Deterrence or Background effect, Symmetric Punishment and Rewards	80
3.5	Full Set of Conditions for Multiple Equilibria and Tipping Points	84
3.5.1	The Role of Personal Affect and Attitude Transference	85
3.5.2	The Role of Deterrence and Background Characteristics	86
3.5.3	The Role of the Impact of Disciplinary Response on Affect	88
3.6	Changes in Disciplinary Policy can Induce Phase Transitions	89
3.6.1	Changes in Punitive Discipline can Induce Phase Transitions	90
3.6.2	Increasing Rewards for Pro-Social Behavior Can Induce Phase Transitions	91
3.7	Optimal Punishing and the Deterrence-Affect Tradeoff in the Case of a Unique Equilibrium	92
3.8	What happens when Student Affect is Time-Sensitive?	94
3.9	Discussion, Limitations, and Conclusion	96
4	THE SELF-SUSTAINING EFFECTS OF ONE-TIME INTERVENTIONS: SYSTEMS THINKING TO REDUCE SYSTEMIC DISADVANTAGE	99
ABSTRACT		100
4.1	Introduction	100
4.2	The EASE Model	105
4.3	Theoretical Results	108
4.3.1	Persistence vs Fadeout through Peer Effects: Early Childhood Education	110
4.3.2	Persistence vs Fadeout Through Phase Transitions: Fixed vs Growth Mindset	112

4.3.3	Persistence vs Fadeout Through Dynamic Complementarity: Double-Dose Algebra	114
4.3.4	Persistence vs Fadeout through Feedback: Self-Affirmation	116
4.4	Empirical Analysis	119
4.4.1	Purpose	119
4.4.2	Data	120
4.4.3	Identification and Statistical Estimation	122
4.4.4	Parameter Estimates	125
4.4.5	Empirical Simulation	126
4.5	Discussion	129
4.5.1	Summary	129
4.5.2	Limitations and Conclusion	130
4.6	Appendix	133
4.6.1	Simplified Model	133
4.6.2	Derivation of Dynamical System for Simplified Model	136
4.6.3	Appendix: Phase Portrait Analysis of Simplified Model	140
4.6.4	Descriptive Trends for Empirical Analysis	145
5	CLOSING	148
	REFERENCES	152

List of Figures

2.1	Flow Diagram for the 5D Model	11
2.2	Transition Estimates I	27
2.3	Transition Estimates II	28
2.4	Derivatives of Equilibrium Proportion in X	31
2.5	Elasticities of Equilibrium Proportion in X	32
2.6	Large Parameter Changes Towards Crime: Black Men	34
2.7	Large Parameter Changes Away from Crime: Black Men	35
2.8	Large Parameter Changes: Black Men	36
2.9	Large Parameter Changes: Black Men	37
2.10	Elasticity of Black-White Male Disparity in X	39
2.11	Black-White Male Disparity in X When Equalizing Each Transition	40
2.12	Flow Diagram for the 3D Model	44
2.13	1997 and 2003 Distributions for X	51
2.14	1997 and 2003 Distributions for C_1	51
2.15	1997 and 2003 Distributions for A	52
2.16	1997 and 2003 Distributions for R	52
2.17	1997 and 2003 Distributions for C_2	53
2.18	$X \rightarrow A$ by Age and Subgroup	53
2.19	$C_1 \rightarrow A$ by Age and Subgroup	54
2.20	Transition Probabilities, Via Regression	54
2.21	System Dynamics	55
2.22	Equilibria	55
2.23	Partial Derivatives of X^*	56
2.24	Give Black Men White Men's Probabilities	56
2.25	Give Black Men White Men's Probabilities	57
2.26	Give White Men Black Men's Probabilities	57
2.27	Give White Men Black Men's Probabilities	58
3.1	Weak Peer Effects	81
3.2	Strong sensitivity to peer behavior, weak sensitivity to peer affect:	82
3.3	Equal sensitivity to peer affect and peer behavior	83
3.4	Strong sensitivity to peer affect, weak sensitivity to peer behavior	84
3.5	Personal Affect and Attitude Transference	86
3.6	Deterrence and Background Effects	87
3.7	Impact α_{-1} of Disciplinary Response on Affect	89
3.8	Changes in Punishment Can Induce Phase Transitions	91
3.9	Changes in Rewards for Pro-Social Behavior Can Induce Phase Transitions	92
3.10	Optimal Punishment Depends on Measurable Tradeoffs	93
4.1	Potential Consequences of Systems Thinking	109
4.2	Persistence and Fadeout Under Strong Peer Effects	111
4.3	Growth Mindset: Persistence Particularly for the Disadvantaged through Phase Transitions	114
4.4	Double Dose Algebra: Persistence through Dynamic Complementarity . .	116
4.5	Self-Affirmation: Fadeout Due to Weak Feedback	116
4.6	Self-Affirmation: Persistence Due to Strong Feedback	118

4.7	Model 1 Validation Exercise	127
4.8	Model 1: Simulating a Self-Affirmation Intervention	128
4.9	Model 2: Simulating a Self-Affirmation Intervention	129
4.10	No Peer Effects: One Stable Equilibrium	142
4.11	Strong Peer Effects: Three Stable Equilibria	143
4.12	Mild Peer Effects: Five Stable Equilibria	144
4.13	Self-Concept Predicts Effort	145
4.14	Effort and Self-Concept Predict Achievement	146
4.15	Current Achievement History Predicts Subsequent Self-Concept	146
4.16	Current Self-Concept History Predicts Subsequent Self-Concept	147

List of Tables

2.1	Proportional Sizes of 5D Model Compartments by Year and Demographic Group in NLSY-97 Analytic Sample	22
2.2	Average Wave-to-Wave Transitions for full NLSY-97 Analytic Sample . . .	25
2.3	Estimated Coefficients from Random Effects Regression	25
4.1	Baseline Structural Equation Model	125
4.2	Full Model With Random Intercepts and Time Dummies	126

ACKNOWLEDGMENTS

I will first acknowledge that GOD is the head of my life, always has been, and always will be. I acknowledge the privilege of having a dissertation committee that I trust. Stephen Raudenbush, Steven Durlauf, Kerwin Charles, Michael Dinerstein, and Susan Levine, thank you for demonstrating both the intellectual prowess and the moral fiber of the scholar I hope to become. I acknowledge the privilege of having some of the most unselfish parents and grandparents in the world. The whole of their investments in me-financial, social, and spiritual capital-has become more than the sum of its parts, and will be paid forward through the generations. I am thankful to my brothers for keeping me well-rounded. Each has served a different purpose in my life, some raising the ceiling and others raising the floor. I love each of you. And Melanie-I love you always and forever.

I am one of the wealthiest men on Earth, because I have never lived a day of my life without love. While my childhood is filled with loving memories from the Saginaw community (extended family, church family including St. Paul and New Mount Calvary Baptist church; a host of quality teachers and staff from Herig and S.A.S.A; John Pugh, Roland Wallace, and the ACT-SO program; Mrs. Currie and the Gamma Kappa Kudos youth group; lifelong friends like Christen Robinson and Newton Davis; and much more), my young adulthood has been filled with love all the same. Always there were close mentors and friends, at the University of Michigan (including but not limited to Carl Simon, Jeffrey Morenoff, Trachette Jackson, Ben Hansen, Mark Conger, and the late John Holland; Darrick Ervin II and the Ervin family, Christiana Turner, Scott Mcpherson, Khory Davenport, Jeremy Cook, Alexis Cook, Louis Ly, and Adrian Ochoa, and so many others), at Princeton (including but not limited to Amir Ali Ahmadi, William Massey, and Nana Kobby Aboagye and Raymond Perkins), and at the University of Chicago (including but not limited to Cynthia Cook-Conley, Milvia Rodriguez, and many other incredible staff members that imbue the Harris School community with love, support, and hope; a litany of close Master's and PhD friends ranging from the class of 2017-2023;

Damon Jones, Anjali Adukia, Scott Ashworth, Bocar Ba, and colleagues and professors affiliated with the IES and the Committee on Education such as Derek Neal, James Heckman, and Margaret Beale-Spencer). No matter where I go, I can see love shining in others. I can still see a faint but eternal love in the hearts of those most would call monsters. Even when it isn't operationalized, love revitalizes my spirit. It gives me focus. Makes me strong. Thank you to everyone who has shared that with me...you, who are too numerous to name. Because of you, I have lived the meaning behind my name.

Too few have the chance to experience love on that scale. And the consequences for our world have been breath-taking.

In 2020 we watched the world suffocate to death with no remorse and no long-term solutions. Starting with the earth itself, from issues with carbon emissions. No remorse, and no long-term solution. Then a pandemic struck based on a virus that attacks the lungs. No remorse, too few long-term investments in public health. In the midst of it, as we watched George Floyd suffocate to death beneath an officer with no remorse, we saw no long-term solution. And as protesters lined the streets of Minneapolis, tear gas filled their lungs, administered with no remorse, and no long-term solution.

As someone with asthma, with friends (such as Marcus George) who suffocated to death from Covid-19 and a grandfather, Joel McMillon, who suffocated to death from lung cancer, I have to say that there is nothing more painful than seeing a doctor offer a short-term solution that is too little, too late. But each of those loved ones would agree with me now when I say that in the grand scheme of things we cannot breathe without each other. We cannot breathe without remorse for those who suffocate, and we cannot breathe without long-term solutions. We cannot breathe without radical collective investment in clean energy to clear the lungs of the earth. We cannot breathe without radical collective investment in public health systems that protect us from airborne

pandemics. We cannot breathe allowing the US military industrial complex to take the breath away from innocent Brown people abroad. We cannot breathe allowing the noose of the prison industrial complex to choke Black families and Black mobility. We cannot breathe allowing the legacy of colonialism to block the airways of self-determination and its economic consequences within the U.S. territories of American Samoa, Guam, the Northern Mariana Islands, the Virgin Islands, and Puerto Rico. We cannot breathe with the noxious fumes of anti-Semitism nor the searing smoke of anti-Palestinian racism. We cannot breathe with the toxic stench of anti-Asian hatred. We cannot breathe drowning in a violent ocean of sexism, human trafficking, gaslighting, silencing, rape, and abuse that impacts women every day. We cannot breathe focusing only on symptoms and short-term solutions, and we cannot breathe without addressing the root causes of our societal asphyxia.

I believe that we can breathe again if we ensure that everyone experiences the same kind of love that has protected me from the time I was a small boy. Every word in this dissertation ultimately stems from that core belief. In that regard, the reader should know that I am indeed biased. I do have an agenda. And every equation, however rigorous or flawed, is a means to this ultimate end.

ABSTRACT

Where, when, and how should we intervene on the School to Prison Pipeline (STPP)? Most scholars agree that there are many factors that influence the outcomes of youth that lead to incarceration, including, but not limited to: identity formation, peer effects, recidivism, and underachievement. Much of the literature has statically investigated these factors in isolation.

However, despite the literature's emphasis on static models of isolated mechanisms, most scholars believe that the STPP persists due to an ecological system of interrelated factors that evolve dynamically (over time). What happens to human developmental trajectories when all of these mechanisms are at work simultaneously? And how might that inform where, when, and how we should intervene? This dissertation proposes a research paradigm that can eventually help answer these questions, presenting the first set of dynamic mathematical models that approach the STPP from a holistic, systems perspective. This perspective allows us to foresee surprising policy consequences that would not be obvious when analyzing these factors in isolated static models. It illuminates conditions under which small policy changes can lead to transformative, self-sustaining effects on school climate; large policy changes can have essentially no effect; and the same policy can have opposite effects under different initial conditions.

This dissertation combines three interrelated studies, each focusing on a different subsystem within the STPP. The first is a macro-dynamic perspective of crime and arrest among youth, capturing the part of the STPP that exists outside of the school. Motivated by Study 1's finding that biggest systemic contribution to the STPP is happening prior to the first arrest, the next two studies focus on parts of the STPP that exist inside the school. Study 2 focuses on how best to intervene on behavioral infractions through disciplinary policy. Finally, because achievement and positive academic identities are so preventive of antisocial behavior in schools, Study 3 focuses on underachievement and

academic identity formation. This dissertation operationalizes an innovative research paradigm to dismantle the STPP, under the premise that reducing systemic disadvantage requires systems thinking.

CHAPTER 1

OVERVIEW

1.1 Background

It is difficult to overstate the salience of mass incarceration in the United States as a high-priority concern among researchers, activists, and politicians ([1]; [27]). The United States holds 5 percent of the world’s population and, according to the American Psychological Association, holds a dramatically disproportionate 25 percent of the world’s prisoners. This has led to a burgeoning of research in several literatures on how best to reduce incarceration and crime, much of which has identified inextricable links between school experiences (including academic achievement, behavioral infractions, suspensions, and drop-out), arrest, and incarceration. These efforts have combined with tremendous zeal from activists and culminated in the use of the phrase “School-to-Prison Pipeline” to describe the system of educational, criminological, and public safety factors that harm behavioral and academic outcomes and funnel youth from schools to prisons.

While the STPP is widely regarded as a high-priority policy problem, there is considerably less agreement on how best to intervene on this issue. Although some scholars argue that punitive criminal justice interventions act as deterrents from criminal behavior ([25]; [11]), other scholars argue that harsh punishments can reinforce criminal behavior by decreasing access to employment and through increased social interactions with other criminals [68]. This debate depends on whether criminal behavior involves a state-dependent process whereby experiences in the criminal justice system reinforce future criminality [67]. This debate cannot be reconciled without a dynamic, systems perspective. In the education literature, there exists a similar debate on whether schools should adopt “Zero-Tolerance” policies that implement harsh punishments even for minor offenses that focus on deterrence, or less punitive “Restorative” practices that focus on improving relationships and positive reinforcement [84]; [45]. The mechanisms under-

girding these alternative approaches to school discipline are very different, and depend in part on the importance of a dynamically evolving set of psychological traits that can be harmed through harsh punishment. They have also largely been studied in isolated, static models, despite the fact that students experience them jointly and dynamically—again requiring a dynamic, systems perspective.

Finally, a policy solution to the STPP that fails to address academic underachievement is fundamentally incomplete. Academic underachievement is tightly coupled with behavioral infractions and future criminality and incarceration [56]. There are many widely accepted mechanisms believed to drive the prevalence of underachievement, including but not limited to the notion that a student’s self-concept with respect to a subject is a strong driver of the effort he/she puts into that subject [37] which in turn impacts his/her academic achievement [96], [51], which impacts self-concept [42]. Together, these constitute a feedback loop, which has important implications in a dynamic context that considers developmental trajectories, and (once again) not just static analyses of outcomes induced by a single policy or mechanism in isolation.

Because these important debates and considerations depend crucially on several jointly experienced mechanisms that are inherently dynamic, adjudicating conditions under which alternative policy approaches would be more or less effective requires that we adopt a dynamic perspective that considers the evolution of behaviors over time within an ecological system of salient interrelated mechanisms. In this dissertation I propose a research paradigm that operationalizes this dynamic “systems perspective” of the STPP using three studies that holistically approach the STPP by focusing on 1) dimensions of the STPP that exist outside the school, 2) disciplinary policy for behavioral infractions inside the school, and 3) prevention of antisocial behavior through academic achievement.

Importantly, this “systems perspective” is not one that solely exists in academia. This dissertation is the culmination of work that began as an effort to operationalize common

colloquial claims I’ve heard my whole life as an African-American male from an inner-city community, such as “the system is set up for us to fail” and “you can’t fix one thing without fixing another.” I heard these claims later while working with advocates in the Children’s Defense Fund, following an intense debate between the advocates on which particular parts of the STPP are most in need of intervention. In 2014 I published my first attempt at a systems perspective of the STPP, (Study 1 is its follow-up). That paper showed how a unified dynamic theoretical model of a system of mechanisms can foresee important policy consequences to the STPP [63] that are not accessible under dominant research paradigms in education. However, this approach does not replace them-it enhances them. Qualitative approaches yield, through a large collection of case studies, a set of working “priors” on human behavior (e.g., criminals induce others into crime). Empirical work allows us to estimate the causal effect of a policy on an outcome (e.g. graduation decreases criminal onset by X%). And a dynamic, systems approach tells us, conditional on reasonable priors/assumptions, how the empirical estimates of those mechanisms can jointly lead to potentially dangerous or exciting policy consequences (e.g., if criminals induce others into crime, increasing prevention efforts through education beyond a threshold could dramatically and disproportionately decrease crime by fundamentally altering the dynamics of the whole system). What I am proposing is a missing complement to what already exists.

1.2 Methodological Philosophy

This dissertation employs the methodological philosophy of applied theory. In all three studies, I construct theoretical mathematical models that are intended to capture certain portions of the STPP and explicitly account for certain salient mechanisms in the literature. Importantly, each of these models is constructed under a certain well-defined set of assumptions. Conditional on these assumptions, the research questions can be answered through uncontestable mathematical proofs. Because those proofs depend on

the assumptions, I systematically test the robustness of the conclusions to deviations in the assumptions. This approach complements empirical work, shedding light on which relationships need to be estimated more carefully; providing a careful, rigorous research paradigm to adjudicate empirically driven academic debates; and allowing us to foresee potentially surprising, unintended consequences of the fact that human beings experience each causally identified mechanism in the literature jointly and dynamically. As we will see, these consequences can have powerful policy implications.

CHAPTER 2

**A DYNAMICAL SYSTEMS ANALYSIS OF CRIMINAL
BEHAVIOR AMONG NATIONAL LONGITUDINAL
SURVEY OF YOUTH RESPONDENTS**

This project is joint work with Carl Simon and Jeffrey Morenoff.

ABSTRACT

Building on previous systems work on the spread and sustenance of crime, we construct and analyze a dynamical systems model of criminal involvement, arrest, desistance, and rehabilitation to be estimated empirically for interviewees in the National Longitudinal Survey of Youth. Using assumptions familiar to stable population theorists, we examine how marginal increases in flows between states interact to decrease or increase the long-run level of crime, and whether this varies by subgroup. We study how observed racial disparities along certain pathways interact to generate macro-level disparities in criminal involvement as measured by arrest and self-report. Finally, we discuss the implications of the model for a broader policy debate on crime control and for competing explanations of the Black-White gap in criminal involvement.

2.1 Introduction

2.1.1 Background

Crime, arrest, and incarceration rates in the United States have been long been key concerns among researchers, activists, and politicians alike [1] [27]. According to a 2014 report, one third of American men between ages 26 and 34 had been arrested before age 26 [85]. The corresponding rate for African-American males within that age range was 42.7 percent. And despite evidence of the preventative impact of education on crime, arrest and incarceration [40] [57], burgeoning arrest rates have been followed by proposals for large budget cuts to the US public education. High profile coverage of racial disparities throughout the criminal justice system and police brutality that has disproportionately harmed Black Americans, including Sandra Bland, Breonna Taylor, and George Floyd, have increased the political salience of these policy concerns in recent years.

Across the ideological spectrum, various policy interventions have been proposed to address these concerns and decrease crime in the United States, including harsh mandatory minimum sentencing vs restorative justice, increasing the police force vs “defunding”

or demilitarizing the police, investing in rehabilitative or preventive social programs, and differential enforcement for first-time vs repeat offenders. While many scholars disagree on policy solutions, most agree that America’s crime and criminal justice problem is a systemic failure with several possible, mutually interacting components. Too often interventions focus on a single component of what is truly an integrated system, a system sometimes referred to as the “School-to-Prison Pipeline.” For example, a policy intervention that increases the incarceration rate of first-time offenders can actually increase crime in the long-run if other interventions have failed to lower recidivism or to increase rehabilitation rates. Long mandatory sentences for nonviolent crimes can lead to overcrowded prisons, necessitating the release of more violent criminals, and to increased recidivism because of the difficulties in finding employment with a prison record. Instead, solving systemic problems requires systems thinking.

2.1.2 Systems Approach

The core of a systems approach requires the construction and analysis of models that explicitly illustrate interdependencies in a complex system’s components, which are often studied in isolation, in order to simulate potential unintended policy consequences, identify “high leverage” points for policy, and systematically adjudicate academic debates.

Systems modeling in criminology began with the seminal work of Alfred Blumstein, who served on the President’s Commission on Law Enforcement and Administration of Justice in 1967. Blumstein helped demonstrate how such models can project the operating costs of various components of the criminal justice system (such as police, courts, and corrections), and how interventions on the various components or demographic shifts in the population influence overall crime rates [71]. More recently, Perc et al [73] show, among other conclusions, that under certain conditions, increasing punishment may not decrease, and could actually increase crime levels, using a game-theoretic approach with interacting agents (criminals, non-criminals, and punishing inspectors).

This paper builds on prior work (references omitted) in which we presented a theo-

retical model for the spread of crime using parameters that capture the stocks and flows through key states in the life course, including initiation into crime, incarceration, recidivism, desistance, and rehabilitation. In one paper (reference omitted), we derived theoretical conditions that lead to high vs low-crime equilibria and found that increasing the incarceration rate of first-time offenders can actually increase the level crime in the system if recidivism is large compared with successful post-incarceration rehabilitation. Another key finding from these models is that of all pathways into crime, long-run crime levels are most sensitive to reductions in the rate of initiation through social influence. A follow-up paper allowed for more heterogeneous agents, especially by age, a step that required use of agent-based models (reference omitted). These two papers include an extensive literature review of mathematical modeling approaches to studying the spread of crime.

In this paper, we take the critical next step of refining and calibrating our theoretical models in prior work. To illustrate the societal dynamics of crime, we use data from the National Longitudinal Survey of Youth 1997 cohort (NLSY97) to estimate a dynamic model separately for three races (Black, White, Hispanic) and two sexes. We present a refined theoretical model of crime dynamics that focuses on arrest rather than incarceration as the initial societal response to crime, with separate states for returning to or abandoning criminal activity after arrest.

2.1.3 Racial disparities

Though our models are simple, our approach sheds light on the dynamics of racial disparities in long-run crime levels. Explanations of racial disparities in crime and criminal justice have long been debated in academic and political circles [26], [27]. Black Americans are more likely to be arrested for the same offense compared to comparable Whites [65]. Black Americans who have been arrested receive harsher sentencing compared to comparable Whites [76], and are more likely to be victims of police brutality compared with comparable Whites [54] [80]. Even outside of the criminal justice system, Black

Americans disproportionately suffer from the social and economic disadvantages that lead to crime, in part due to past injustices [77].

Mallet et.al [58] suggests that there are at least four dominant theories to explain the overrepresentation of minorities in the criminal justice system: larger offense rates among minorities, racial bias in decision-making among criminal justice officials, a disproportionate accumulation of individual risk factors such as neglect and mental health problems, and contextual differences (such as disparities in school quality or communities with fewer employment opportunities). Blumstein [10] attributed racial disparities in U.S. prisons to differences in arrest rates rather than racial bias in the criminal justice system. Re-havi and Starr [76] showed that on average, Blacks receive almost ten percent longer sentences than comparable Whites for the same crimes, and that at least half of this gap can be explained by differences in initial charging choices of prosecutors. Lochner et.al [57] found that an additional year of schooling is associated with 3 to 4 percent decrease in likelihood of incarceration for Whites and an 8 to 9 percent decrease for Blacks, and an 11 percent reduction in the likelihood of arrest, with the effect being larger for Blacks than Whites. Regardless of what explains racial disparities in crime and criminal justice involvement, researchers agree that the disparities exist along multiple pathways. This paper proposes a series of simple dynamic systems models that can be used to study how racial disparities in criminal involvement can magnify as these pathways interact.

2.1.4 State dependence

Finally, our systems approach can be used to shed light on the implications of certain criminological debates. The tight correlation between past and future offending has received considerable attention in criminological literature. While some argue this correlation is due to time-invariant differences in individual propensity for crime, others argue that the correlation may also be due to a time-varying, state-dependent process whereby past behavior alters the trajectory of future behavior [67]. In the latter case, studying racial disparities in crime and criminal justice requires that we not only explain criminal

behavior with static models, but that we also investigate how disparities in flows between states interact to magnify the disparities we observe. Perhaps even more importantly, devising a crime control strategy in a world of state-dependent criminality requires that we have an understanding of how independent changes in one pathway of intervention could have reinforcing or counter-productive effects along other pathways.

2.1.5 Overview

The remainder of this paper proceeds as follows. We present an analysis of the dynamic system model which we have tailored to the data set of interest. Next, we describe that data set and the calibration processes we use to estimate the parameters of the model. Following an analysis of interesting patterns we find in the data for various subgroups, we use the model and the estimated transition parameters to investigate how marginal increases in flows between states interact to decrease or increase the long-run level of crime and disparities in criminal involvement.

2.2 The Model

2.2.1 Modeling overview

In this section we present a mathematical model of the flow of populations in and out of states of criminal activity and arrest. We conduct a theoretical analysis that reveals how the parameters of the model interactively lead to long-run equilibrium levels of crime. This illuminates how our conclusions depend on the assumptions we make about the parameters. In the section that follows, we take the important next step of estimating the parameter values with real data. Combined, the theoretical and empirical analysis will reveal both how common targets of intervention interactively generate long-run crime levels depending on the current parameter values, and estimates for those parameter values.

This five-dimensional model arose from a simpler three-dimensional model (See Ap-

pendix) that misses a critical piece of the dynamic – the distinction between those with and without an arrest history. It is important to keep track of those who returned to crime after arrest and those who avoided criminal behavior after their arrest. Figure 2.1 below summarizes these states and flows.

2.2.2 States in the Model

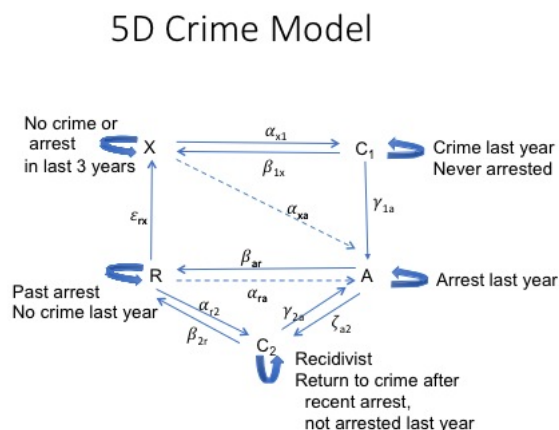


Figure 2.1: Flow Diagram for the 5D Model

Write the five states in the Model as X, C_1, A, R, C_2 :

- X Those who either
 - 1) Have never been arrested, and have not offended in the last year, or
 - 2) Have been arrested before, but have not offended in the last three years (rehabilitation).
- C_1 Those who have never been arrested before, but have offended in the last year.
- A Those who have been arrested in the last year, regardless of self-reported offending.
- R Those who have been arrested or have offended in the last three years, but not in the last year.

C_2 Those who

- 1) Have been arrested before but not in the last year and
- 2) Have offended in the last year.

2.2.3 Transitions in the Model

- Transitions from X

- No crime in last year: $\sigma_{xx} : X \rightarrow X$.
- Crime in last year, no arrest: $\alpha_{x1} : X \rightarrow C_1$ (onset)
- Crime and arrest in last year: $\alpha_{xa} : X \rightarrow A$ (onset)

- Transitions from C_1

- No crime in last year and no arrest in last year: $\beta_{1x} : C_1 \rightarrow X$. (desistance)
- Crime in last year, but no arrest: $\sigma_{11} : C_1 \rightarrow C_1$.
- Arrest (and crime) in last year: $\gamma_{1a} : C_1 \rightarrow A$.

- Transitions from A

- No crime in last year: $\beta_{ar} : A \rightarrow R$.
- Crime and arrest in last year: $\sigma_{aa} : A \rightarrow A$.
- Crime, but no arrest in last year: $\zeta_{a2} : A \rightarrow C_2$.

- Transitions from C_2

- Crime and arrest in last year: $\gamma_{2a} : C_2 \rightarrow A$.
- Crime, but no arrest last year: $\sigma_{22} : C_2 \rightarrow C_2$.
- No crime last year: $\beta_{2r} : C_2 \rightarrow R$.

- Transitions from R

- Crime last year, no arrest: $\alpha_{r2} : R \rightarrow C_2$ (recidivism)

- Crime and arrest last year: $\alpha_{ra} : R \rightarrow A$.
- No crime last year, but crime in last three years: $\sigma_{rr} : R \rightarrow R$.
- No crime in last three years: $\epsilon_{rx} : R \rightarrow X$. (rehabilitation)

The transition parameters within each of the above five bullets must sum to 1.

The dynamic system can be described with the following set of difference equations:

$$\begin{aligned}
X(t+1) &= \sigma_{xx}X(t) + \beta_{1x}C_1(t) + \epsilon_{rx}R(t) \\
C_1(t+1) &= \sigma_{11}C_1(t) + \alpha_{x1}X(t) \\
A(t+1) &= \sigma_{aa}A(t) + \gamma_{1a}C_1(t) + \alpha_{xa}X(t) + \alpha_{ra}R(t) \\
C_2(t+1) &= \sigma_{22}C_2(t) + \zeta_{a2}A(t) + \alpha_{r2}R(t) \\
R(t+1) &= \sigma_{rr}R(t) + \beta_{2r}C_2(t) + \beta_{ar}A(t)
\end{aligned}$$

Because the transition probabilities from each of the five states add to 1, we can replace the σ s in this system and, after simplifying, rewrite it as:

$$\begin{aligned}
X(t+1) - X(t) &= \beta_{1x}C_1 + \epsilon_{rx}R - \alpha_{x1}X - \alpha_{xa}X \\
C_1(t+1) - C_1(t) &= \alpha_{x1}X - \beta_{1x}C_1 - \gamma_{1a}C_1 \\
A(t+1) - A(t) &= \gamma_{1a}C_1 + \alpha_{ra}R + \gamma_{2a}C_2 + \alpha_{xa}X - \beta_{ar}A - \zeta_{a2}A \quad (2.1) \\
R(t+1) - R(t) &= \beta_{ar}A + \beta_{2r}C_2 - \alpha_{r2}R - \alpha_{ra}R - \epsilon_{rx}R \\
C_2(t+1) - C_2(t) &= \alpha_{r2}R + \zeta_{a2}A - \beta_{2r}C_2 - \gamma_{2a}C_2
\end{aligned}$$

Just as for our 3D system, the right hand sides of system (2.1) sum to 0. Population has constant size $N = X(t) + C_1(t) + A(t) + R(t) + C_2(t)$.

2.2.4 Equilibrium

We can therefore eliminate one variable and one equation from (2.1) and work with a linear system of four equations in four unknowns. Furthermore, as we did for the 3D

model, we can divide each of these equations through by N so that the five variables become population fractions instead of total numbers. Finally, as we did in the 3D system, we set the right hand sides of system (2.1) equal to zero to compute the long run equilibrium to which the dynamics flow from any starting ($t = 0$) values of the variables. As we elaborate in the Appendix, the long run equilibrium is:

$$X^* = \frac{\epsilon_{rx}(\beta_{1x} + \gamma_{1a})(\gamma_{2a}\beta_{ar} + \beta_{2x}(\beta_{ar} + \zeta_{a2}))}{D} \quad (2.2)$$

where the denominator is:

$$\begin{aligned} D = & \{ \alpha_{x1} [\epsilon_{rx}\gamma_{1a}\gamma_{2a} + \epsilon_{rx}\gamma_{2a}\beta_{ar} + \gamma_{1a}\gamma_{2a}\beta_{ar} + \gamma_{1a}\gamma_{2a}\alpha_{ra} + \epsilon_{rx}\gamma_{1a}\zeta_{a2} + \gamma_{1a}\alpha_{ra}\zeta_{a2} + \alpha_{x2}\gamma_{1a}(\gamma_{2a} + \beta_{ar} + \zeta_{a2}) \\ & + \beta_{2x}(\epsilon_{rx}(\gamma_{1a} + \beta_{ar} + \zeta_{a2}) + \gamma_{1a}(\beta_{ar} + \alpha_{ra} + \zeta_{a2}))] \\ & + (\beta_{1x} + \gamma_{1a}) [\epsilon_{rx}(\gamma_{2a}(\beta_{ar} + \alpha_{xa}) + \alpha_{xa}\zeta_{a2}) + \alpha_{xa}(\gamma_{2a}(\beta_{ar} + \alpha_{ra}) + \alpha_{ra}\zeta_{a2} + \alpha_{x2}(\gamma_{2a} + \beta_{ar} + \zeta_{a2})) \\ & + \beta_{2x}(\epsilon_{rx}(\beta_{ar} + \alpha_{xa} + \zeta_{a2}) + \alpha_{xa}(\beta_{ar} + \alpha_{ra} + \zeta_{a2}))] \}. \end{aligned}$$

2.2.5 Effect of Independent Parameter Changes on X^*

We investigate how the long run equilibrium level of crime depends on the underlying parameters. Roughly speaking, if the effect of a particular intervention can be interpreted as an independent change in one of the parameters in our model, we investigate how the long-run level of criminal activity will be affected by that change. Note that while the parameters may have interdependencies in reality, we are considering cases in which each parameter is allowed to change exogenously/independently of the others. We again focus on X^* , the fraction of those in the population who have not had any criminal activity or been arrested at least for the past three years. We loosely call $1 - X^* = C_1^* + A^* + C_2^* + R^*$ the “crime rate.”

None of the α s appears in the numerator of X^* . By the quotient rule, $\frac{\partial X^*}{\partial \alpha}$ is < 0 for all four α s. This leads to the obvious conclusion that an increase in the rate of entry

or reentry into crime leads to a decrease in X^* and an increase in the crime rate.

It takes a little more work to show that

$$\frac{\partial X^*}{\partial \beta_{1x}} > 0, \quad \frac{\partial X^*}{\partial \beta_{2r}} > 0, \quad \frac{\partial X^*}{\partial \beta_{ar}} > 0, \quad \frac{\partial X^*}{\partial \epsilon_{rx}} > 0,$$

as it did in the 3D model. See Appendix for details. We conclude that increasing the rate of movement into R or back to X decreases the crime rate.

As before, the effects of an increase in the arrest rates are more subtle. For re-arrests, the derivative is:

$$\frac{\partial X^*}{\partial \gamma_{2a}} = \frac{\epsilon_{rx}(\beta_{1x} + \gamma_{1a})(\beta_{ar} - \beta_{2r})[\alpha_{x1}\gamma_{1a} + (\beta_{1x} + \gamma_{1a})\alpha_{xa}][(\epsilon_{rx} + \alpha_{ra})\zeta_{a2} + \alpha_{r2}(\beta_{ar} + \zeta_{a2})]}{D^2}$$

The sign of this derivative depends on β_{2r} and β_{ar} . Specifically, the derivative is nonnegative if $\beta_{ar} \geq \beta_{2r}$ and negative otherwise. Simply put, arresting criminals with an arrest history reduces crime if they are more likely to desist after arrest than before arrest.

The derivative for first arrests is:

$$\begin{aligned} \frac{\partial X^*}{\partial \gamma_{1a}} = & \frac{\alpha_{x1}\epsilon_{rx}(\gamma_{2a}\beta_{ar} + \beta_{2r}(\beta_{ar} + \zeta_{a2}))}{D^2} \times \left[\epsilon_{rx}[\gamma_{2a}\beta_{ar} + \beta_{2r}(\beta_{ar} + \zeta_{a2})] \right. \\ & \left. - \beta_{1x}[\gamma_{2a}(\epsilon_{rx} + \beta_{ar} + \alpha_{ra}) + \zeta_{a2}(\epsilon_{rx} + \alpha_{ra}) + \alpha_{r2}(\gamma_{2a} + \beta_{ar} + \zeta_{a2}) + \beta_{2r}(\epsilon_{rx} + \beta_{ar} + \alpha_{ra} + \zeta_{a2})] \right]. \end{aligned}$$

Its sign depends on the relative size of ϵ_{rx} , the rehabilitation rate of those with an arrest history, to β_{1x} , the desistance rate of those never arrested. In short, if people are more likely to desist before they have any arrest history than after they obtain an arrest history, then increasing the arrest rate of the never-arrested can increase the overall crime rate.

Finally, our analysis indicates that the first and second derivatives of X^* are always opposite in sign, which implies a decreasing returns effect of each parameter on the equilibrium level X^* . For example, as β_{1x} increases, its positive impact on X^* decreases; and as α_{x1} increases, its negative impact on X^* decreases. Those calculations are available upon request.

2.2.6 $\alpha_{xa} : X \rightarrow A$ is a complex, but important transition

One of our new transition parameters, $\alpha_{xa} : X \rightarrow A$, plays an important role in Section 2.5.3 in the discussion about the effects of parameter changes. This is a complex parameter in that it includes two changes within one year: first criminal activity ($X \rightarrow C_1$) and arrest ($C_1 \rightarrow A$) in the same time step. From a perpetrator’s point of view, it is about moving into crime, but not avoiding detection and arrest in early criminal activity. From a law enforcement perspective, it is about arresting previously law-abiding people with a clean record. From the perspective of many social activists, this captures the first step of the “School to Prison Pipeline;” a step that is targeted by a large number of prevention-oriented investments in education, including authoritative parenting and teaching strategies, disciplinary policy, and early childhood education, among others. Finally, although we do not explore it in this paper, it is possible in principle for law-abiding citizens who are wrongfully arrested for crimes to experience this transition as well.

2.3 Data: NLSY97

We empirically estimate parameter values from our mathematical models using data from the National Longitudinal Survey of Youth 1997 cohort (NLSY97), a representative sample of people born between 1980 and 1984 and living in the United States at the time of the initial survey in 1997. We use data from annual interviews conducted from 1997 to 2003 (Rounds 1-7), when questions about self-reported criminal activity were asked. Our analytic sample consists of 6,459 who responded to all seven survey waves and who identify as Black, White, or Hispanic.¹

To operationalize the states from our theoretical model, we used data from self-administered survey questions about participation in criminal and delinquent activity as well as arrests. We consider someone to have committed an offense in the past year if they reported being involved in any of the following activities within the last year:

¹There were 8,984 interviewees in 1997; 7,754 of these were still being interviewed in 2003. We dropped 1,295 of these individuals who did not respond to all seven survey waves or who did not identify as Black, White, or Hispanic, leaving our final sample size of 6,459.

purposely damaging or destroying property not belonging to the respondent, stealing something worth less than \$50, stealing something worth \$50 or more (including a car), other property crimes (including fencing stolen property, possessing or receiving stolen property, or selling something for more than it was worth), or attacking or assaulting someone. To measure arrests, we used data on whether the respondent had been arrested by the police or taken into custody for an illegal or delinquent offense (not including arrests for minor traffic violations) within the last year. Using these self-reports of criminal activity and arrests, we constructed binary variables for each of the five states of our theoretical model as follows:

- X (not active in criminal behavior) = 1 if the respondent (1) has never been arrested and not offended within the last year or (2) has been arrested but has not offended or been arrested within the last three years;
- C_1 (criminally active without arrest) = 1 if the respondent (1) has never been arrested but (2) has offended within the last year;
- A (recently arrested) = 1 if the respondent has been arrested within the last year, regardless of offense activity;
- C_2 (criminally active after an arrest) = 1 if the respondent (1) has been arrested but not within the last year and (2) has offended in the last year;
- R (criminally inactive after an arrest, i.e., in rehabilitation) = 1 if the respondent (1) reported an arrest in a previous survey wave, (2) has offended or been arrested within the last three years, but (3) has neither offended nor been arrested in the last year.

Individuals in our sample are classified into one of these mutually exclusive states at each survey wave but over time they may move in and out of different states. Our primary outcomes for the statistical analysis presented below are wave-to-wave transitions between different states in our model, as illustrated by the arrows in Figure 1. We analyze the following transitions:

- $X \rightarrow X$: Someone who was not criminally active at a given wave and the next wave;
- $X \rightarrow C_1$: Someone who was not criminally active at a given wave but reports involvement in criminal offending at the next wave;
- $X \rightarrow A$: Someone who was not criminally active at a given wave but reports an arrest at the next wave;
- $C_1 \rightarrow C_1$: Someone who was criminally active (without a recent arrest) at a given wave and the next wave;
- $C_1 \rightarrow X$: Someone who was criminally active (without a recent arrest) at a given wave but is no longer criminally active at the next wave;
- $C_1 \rightarrow A$: Someone who was criminally active (without a recent arrest) at a given wave and reports an arrest at the next wave;
- $A \rightarrow A$: Someone who reports an arrest at a given wave and the next wave;
- $A \rightarrow C_2$: Someone who reports an arrest at a given wave and reports involvement in criminal offending at the next wave but no subsequent arrests;
- $A \rightarrow R$: Someone who reports an arrest at a given wave and has no subsequent arrests or reports of criminal behavior at the next wave;
- $C_2 \rightarrow C_2$: Someone who is criminally active (without a recent arrest) at a given wave and the next wave;
- $C_2 \rightarrow A$: Someone who is criminally active (without a recent arrest) at a given wave and reports a new arrest at the next wave;
- $C_2 \rightarrow R$: Someone who is criminally active (without a recent arrest) at a given wave but is not criminally active and has no further arrests at the next wave;

- $R \rightarrow R$: Someone who has a recent history of arrest but has not been criminally active or arrested in the last year at a given wave and the next wave;
- $R \rightarrow A$: Someone who has a history of arrest but has not been criminally active or arrested in the last year at a given wave but reports an arrest at the next wave;
- $R \rightarrow C_2$: Someone who has a history of arrest but has not been criminally active or arrested in the last year at a given wave but reports involvement in criminal offending (without an arrest) at the next wave;
- $R \rightarrow X$: Someone who has a history of arrest but has not been criminally active or arrested in the last year at a given wave and who, by the next wave, has gone three years without any reports of criminal activity or arrest.

We analyze differences in these transitions by gender ², age, and race/ethnicity, using data on gender (men/women) and race/ethnicity (non-Hispanic Black, non-Hispanic White, and Hispanic) collected at Round 1 of NLSY97, and a time-varying measure of age. Our analytic sample is 51% men and 49% women, 26% non-Hispanic Black, 51.9% non-Hispanic White, and 21.2% Hispanic. The mean age at baseline (Round 1) was 14.9.

2.4 Methods

2.4.1 Assignment to model state in each round

Our analysis proceeds in three stages. First, we present descriptive results showing trends over time and across subgroups in the proportional sizes of the five compartments of our theoretical model. We also present a transition matrix showing flows between each of the compartments in our theoretical model from one survey wave to the next.

Second, we analyze differences across demographic groups in each of the 16 possible transitions in our theoretical model by running logistic regression models with random

²We use the terms "gender" and "sex" interchangeably throughout the paper to refer to differences between men and women. The survey questions used the two interchangeably, although we recognize the two are not at all synonymous in reality.

effects, using the following specification:

$$\log \left[\frac{\Pr(y_{i(t \rightarrow t+1)} = 1)}{1 - \Pr(y_{i(t \rightarrow t+1)} = 1)} \right] = \beta_0 + \beta_1 (\text{Age} - 16)_{it} + \beta_2 (\text{Age} - 16)_{it}^2 + \beta_3 \text{Woman}_i \\ + \beta_4 \text{Black}_i + \beta_5 \text{Hispanic}_i + \beta_6 (\text{Woman}_i * \text{Black}_i) + \beta_7 (\text{Woman}_i * \text{Hispanic}_i) + u_i$$

where $y_{i(t \rightarrow t+1)}$ is a binary outcome indicating the transition between states in our theoretical model from survey wave t to wave $t + 1$. Since our study period consists of seven waves of NLSY data, we observe six wave-to-wave transitions for each person. We control for age as a quadratic function, but the main coefficients of interest are the dummy variables for gender and race/ethnicity and the interactions between them. The model also includes a random intercept, u_i , which is assumed to be marginally normal. The purpose of this intercept is to proxy for person-specific, time-invariant traits, such as cultural background, childhood personality traits, and (possibly) a “natural” propensity to commit crimes. Although this approach does not fully account for such “population heterogeneity” concerns, it addresses them. Using these regression models, we calculate the average predicted probabilities for subgroups defined by the interaction between gender and race/ethnicity, and we graph these predicted probabilities along with their 95% confidence intervals.

Third, we carry out simulations to analyze how each of the transition probabilities affect the long-run equilibrium of the system. For each of the six subpopulations, and then for the total population, we substitute the estimated values of the transition parameters from Figures 2 and 3 for that subpopulation into dynamical system (2). Then, for each subpopulation, we use the Round 1 survey values from Table 1 to set the initial values $X(0), C_1(0), A(0), C_2(0), R(0)$ of each of the five states. We simulated system (2) to calculate the corresponding values at $t = 1$ from the values at $t = 0$, and so on for $t = 2, 3, 4, \dots$. By Markov Theory, this process converges such that there exists a unique “long run” equilibrium proportion X^* of people in state X . This convergence occurs from any choice of initial conditions. Expression (4) provides a formula for this equilibrium for

any choice of the transition parameters; we show that this is indeed the value to which the system dynamics converges.

We test how sensitive the equilibrium X^* is to one-percentage point changes in each of the transition probabilities through simulations, which we compare to the partial derivatives computed in Section 2.5. Our simulations also show how sensitive X^* is to larger percentage point changes and percentage changes in each parameter within demographic subgroups defined by race/ethnicity and gender.

Finally, we use the simulations to analyze how racial differences in transition probabilities contribute to racial differences in the equilibrium level of crime, X^* .

Our final task in this vein was to analyze which parameter(s) played the largest role in Black/White differences. We first computed and examined the absolute and relative differences in the transition parameters for Black males and White males; for the relative differences, we used the quotient: (Black rate - White rate)/White rate. Finally, we considered the possible systems-level effects behind some of these racial differences. Could the *complex system* of interrelated transitions lead long-run crime differences to be more sensitive to certain parameters than others, regardless of the size of the racial differences in those transition parameters? To shed light on this systems effect of racial differences in any given transition probability on the overall racial differences in levels of criminal activity, we ran our dynamic simulation (2.1) with the transition parameters for White males with one exception – we changed one parameter to the corresponding value for Black males. We repeated this for each parameter. Then we reversed the process, using all but one parameter from the Black male list and replacing that parameter by the corresponding parameter from the White male list. We carried out this replacement for each transition parameter. We separately studied the equilibrium proportion of law-abiding citizens X^* and the long-run proportion of criminally active and arrested individuals $(C_1 + A + C_2)$.

2.5 Results for State and Parameter Estimation

2.5.1 Descriptive Results

Table 2.1: Proportional Sizes of 5D Model Compartments by Year and Demographic Group in NLSY-97 Analytic Sample

	X	C_1	A	R	C_2	Total
Full Sample						
Round 1 (1997)	0.855	0.079	0.041	0.013	0.012	1.000
Round 2 (1998)	0.765	0.136	0.052	0.021	0.025	1.000
Round 3 (1999)	0.760	0.113	0.054	0.025	0.048	1.000
Round 4 (2000)	0.764	0.090	0.056	0.029	0.061	1.000
Round 5 (2001)	0.771	0.070	0.059	0.031	0.069	1.000
Round 6 (2002)	0.779	0.054	0.057	0.029	0.081	1.000
Round 7 (2003)	0.798	0.041	0.046	0.028	0.086	1.000
White Men						
Round 1 (1997)	0.827	0.094	0.044	0.017	0.017	1.000
Round 2 (1998)	0.733	0.147	0.064	0.025	0.031	1.000
Round 3 (1999)	0.722	0.121	0.074	0.030	0.054	1.000
Round 4 (2000)	0.708	0.107	0.078	0.037	0.071	1.000
Round 5 (2001)	0.721	0.073	0.073	0.035	0.098	1.000
Round 6 (2002)	0.717	0.059	0.084	0.044	0.096	1.000
Round 7 (2003)	0.741	0.050	0.059	0.036	0.114	1.000
Black Men						
Round 1 (1997)	0.796	0.092	0.074	0.025	0.014	1.000
Round 2 (1998)	0.670	0.161	0.094	0.030	0.045	1.000
Round 3 (1999)	0.654	0.118	0.110	0.032	0.087	1.000
Round 4 (2000)	0.653	0.084	0.112	0.045	0.106	1.000
Round 5 (2001)	0.647	0.064	0.127	0.048	0.115	1.000
Round 6 (2002)	0.638	0.056	0.110	0.041	0.155	1.000
Round 7 (2003)	0.656	0.038	0.103	0.054	0.149	1.000
Hispanic Men						
Round 1 (1997)	0.786	0.106	0.073	0.018	0.018	1.000
Round 2 (1998)	0.698	0.142	0.091	0.035	0.034	1.000
Round 3 (1999)	0.682	0.125	0.080	0.039	0.075	1.000
Round 4 (2000)	0.676	0.100	0.084	0.043	0.097	1.000
Round 5 (2001)	0.701	0.074	0.089	0.048	0.088	1.000
Round 6 (2002)	0.699	0.066	0.074	0.043	0.118	1.000
Round 7 (2003)	0.732	0.036	0.083	0.035	0.114	1.000
White Women						
Round 1 (1997)	0.902	0.053	0.027	0.007	0.011	1.000
Round 2 (1998)	0.828	0.103	0.035	0.014	0.020	1.000
Round 3 (1999)	0.828	0.089	0.031	0.012	0.041	1.000
Round 4 (2000)	0.849	0.059	0.031	0.020	0.042	1.000
Round 5 (2001)	0.847	0.057	0.032	0.018	0.046	1.000
Round 6 (2002)	0.864	0.038	0.030	0.014	0.054	1.000
Round 7 (2003)	0.882	0.028	0.022	0.013	0.054	1.000
Black Women						
Round 1 (1997)	0.878	0.074	0.029	0.009	0.010	1.000
Round 2 (1998)	0.823	0.109	0.028	0.009	0.031	1.000
Round 3 (1999)	0.811	0.108	0.025	0.016	0.040	1.000
Round 4 (2000)	0.848	0.064	0.034	0.012	0.042	1.000
Round 5 (2001)	0.853	0.053	0.033	0.014	0.047	1.000
Round 6 (2002)	0.876	0.035	0.024	0.011	0.054	1.000
Round 7 (2003)	0.882	0.032	0.021	0.016	0.050	1.000
Hispanic Women						
Round 1 (1997)	0.889	0.065	0.026	0.010	0.010	1.000
Round 2 (1998)	0.822	0.117	0.024	0.011	0.026	1.000
Round 3 (1999)	0.868	0.058	0.025	0.011	0.039	1.000
Round 4 (2000)	0.861	0.066	0.019	0.017	0.037	1.000
Round 5 (2001)	0.881	0.045	0.023	0.013	0.039	1.000
Round 6 (2002)	0.888	0.040	0.022	0.014	0.037	1.000
Round 7 (2003)	0.900	0.032	0.026	0.009	0.033	1.000

Table 2.1 presents the proportion of the sample in each of the five states at each survey wave, by gender and race/ethnicity. At any given survey wave, most sample members are in state X (not yet criminally active or arrested), but this share drops over time (from 83% at Round 1 to 74% at Round 7). A greater proportion of women than men are in state X at each survey wave, and this gender gap widens over time, as more men than women leave state X because they begin offending or are arrested. The flow of people out of state X is especially high among Black men, 79% of whom are in X at Round 1 compared to only 66% at Round 7.

The share of people in C_1 (criminally active without an arrest) doubles between Rounds 1 (7%) and 2 (14%) but declines over subsequent survey waves, reaching a low of 4% at Round 7. More men than women are in C_1 at any given wave, but the gender gap narrows over time, and there are relatively small differences by race/ethnicity.

The overall share of people in state A (recently arrested) remains fairly stable over time, but there is considerable variation by gender and race/ethnicity. Especially noteworthy is the large share of Black and Hispanic men among those who have been recently arrested compared to White men. At Round 1, roughly 7% of Black and Hispanic men are in state A compared to 4% of white men, and these gaps widen over time, such that by Round 7, 10% of Black men and 8% of Hispanic men in state A compared to 5% of White men.

There are fewer people in state C_2 (those with a history of arrest who are currently criminally active) than any other state at any given survey wave. More men than women are in state C_2 at all time periods, and the gender gap increases over time.

The share of people in state R (previously arrested but without a record of offending or arrest in the past year) starts off very low in early survey waves but increases over time. A gender gap also emerges at later survey waves, with more men than women represented in R , and this is especially noticeable among Blacks by Round 7, 15% of Black men are in R compared to only 5% of Black women.

Table 2.2 shows the population “flows” between states from one wave to the next.

First, we consider flows from state X . Most people (90%) who are in state X at a given wave remain in X at the next wave. The biggest flow out of X is into state C_1 (7%), indicating an onset of offending. A smaller group of people (3%) who start in X are arrested at the next wave (state A).

Second, we examine transitions from state C_1 , which are people who have offended within the past year but not yet been arrested during the study period. Most people (58.2%) in state C_1 at a given wave return to X , desisting from criminal activity, while roughly a third remain in C_1 . A smaller share (10%) report an arrest at the next wave, moving to A .

Third, we consider transitions from state A , people who have been arrested within the past year at a given wave. About half (51%) of the people in state A at a given wave move to R at the subsequent wave, which means that by that time they had not offended or been arrested in the past year. A little more than a quarter (29%) of those in state A at a given wave remain there at the following wave because they have been rearrested, while 22% move to state C_2 , meaning that they remain criminally active at the next wave but do not report another arrest.

Fourth, we examine transitions from state C_2 , who have been arrested during the study period (but not within the last year) and have offended within the last year. The most common transition for those in state C_2 is to state R (46%), which indicates that someone has desisted from offending and has not been arrested in the last year. A little more than one-third (35%) of those in C_2 at a given wave remain there in the subsequent wave, while 20% move to state A , meaning that they are rearrested.

Finally, we consider transitions among people in state R at a given wave, which means that they have been arrested during the study period but have neither offended nor been arrested during the past year. Over three-quarters of people in state R at a given wave remain criminally inactive and arrest-free by the subsequent wave, either because they remain in state R (48%) or transition back to X (29%), which means that they have gone at least three years without offending or being arrested. Other people in state R

transition to C_2 (13%), indicating a resumption of offending, or A (11%) meaning that they are rearrested by the subsequent wave.

Table 2.2: Average Wave-to-Wave Transitions for full NLSY-97 Analytic Sample

	Wave $t + 1$					
Wave t	X	C1	A	C2	R	Total
X	0.902	0.070	0.026	NA	NA	1.00
C1	0.582	0.323	0.096	NA	NA	1.00
A	NA	NA	0.286	0.221	0.508	1.00
C2	NA	NA	0.2	0.346	0.455	1.00
R	0.288	NA	0.108	0.127	0.477	1.00

2.5.2 Statistical estimation of transition rates by race and sex

Table 2.3: Estimated Coefficients from Random Effects Regression

	$X \rightarrow C_1$	$X \rightarrow A$	$C_1 \rightarrow X$	$C_1 \rightarrow A$	$A \rightarrow R$	$C_2 \rightarrow A$	$R \rightarrow C_2$	$R \rightarrow A$	$R \rightarrow X$
Age	-0.29** (0.02)	-3.60** (0.02)	0.11 (0.03)	-0.11** (0.04)	0.23*** (0.03)	-0.05 (0.04)	-0.27*** (0.05)	-0.05 (0.05)	0.24*** (0.04)
Age Squared	-0.02 (0.00)	-0.02** (0.02)	0.02 (0.01)	-0.04** (0.01)	0.00 (0.01)	0.01 (0.02)	0.05** (0.02)	0.01 (0.02)	-0.05*** (0.02)
Female	-0.57 (0.09)	-0.99 (0.13)	0.21 (0.13)	-0.36* (0.16)	0.51** (0.17)	-0.51 + (0.28)	0.06 (0.26)	-0.70** (0.29)	0.12 (0.15)
Black	0.11 (0.11)	0.56*** (0.12)	0.01 (0.14)	0.13 (0.16)	0.08 (0.15)	0.41 + (0.22)	-0.02 (0.25)	0.24 (0.23)	-0.11 (0.15)
Hispanic	0.01 (0.12)	0.01 (0.14)	-0.12 (0.15)	-0.23 (0.19)	-0.23 (0.18)	-0.17 (0.26)	0.30 (0.28)	0.26 (0.27)	-0.18 (0.17)
Female*Black	-0.52 (0.15)	-0.52 (0.21)	0.45 (0.21)	-0.5 (0.22)	0.13 (0.3)	-1.37** (0.63)	-0.44 (0.45)	-0.64 (0.51)	0.28 (0.25)
Female*Hispanic	-0.06 (0.17)	-0.53 (0.25)	0.27 (0.25)	0.05 (0.24)	0.59 + (0.36)	0.29 (0.53)	-0.95 (0.53)	-0.87 (0.62)	0.49 + (0.38)
Constant	-3.10*** (0.07)	-3.70*** (0.14)	0.48*** (0.10)	-1.97*** (0.052)	0.09 (0.10)	-1.40*** (0.17)	-0.247*** (0.24)	-2.39*** (0.23)	-0.91*** (0.10)

In the next stage of our analysis, we use a regression analysis to consider differences across subgroups defined by gender and race/ethnicity in each of the transitions just described. The results are tabulated in 2.3, although we omit results for $A \rightarrow C_2$ and $C_2 \rightarrow R$ due to spatial constraints. We present the results graphically, by plotting predicted probabilities of each transition from the random effects logistic regression models for each subgroup. Subgroup differences in the predicted probabilities for each of the 16 transitions are shown in Figures 2.2 and 2.3.

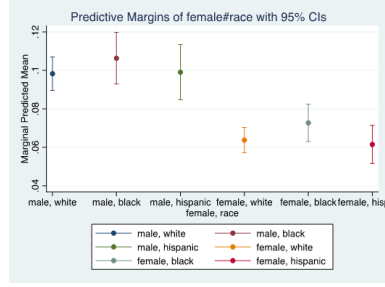
The graphs in the top panel of Figure 2.2 show the predicted transition probabilities for people who were in state X at a given wave, meaning that they had not yet reported any criminal behavior or arrests. Men were less likely than women to remain in X at the subsequent wave and more likely to transition to states $C1$ or A . Blacks were less likely than whites to remain in state X from one wave to the next and more likely to move

from X to A , meaning that they had a higher risk of arrest. There were also significant interactions between gender and race/ethnicity in the $X \rightarrow A$ transition such that the gap between men and women was larger among Blacks and Hispanics compared to Whites. The probability of transitioning from X to A was significantly higher for Black men than any other group and was lowest among Hispanic women.

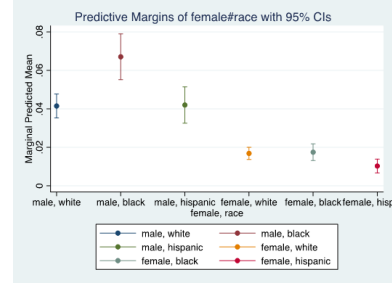
The graphs in the middle panel of Figure 2.2 show transition probabilities for people who were in state C_1 at a given wave, meaning that they committed an offense within the last year. There were no significant differences by gender or race in the probability of remaining in C_1 from one wave to the next. There was a significant interaction between gender and race in predicting the probability of moving from C_1 to X , a transition away from criminal behavior, such that Black women were especially likely to experience this transition. There was also a marginally significant ($p < .10$) interaction between being race and gender in predicting the transition from C_1 to A , the risk of getting arrested among those who were already criminally active, such that Black men were significantly more likely than Black women to make this transition, but the gender gap was smaller among Whites and Hispanics.

Next, we consider the transitions of people in state A at a given wave, meaning they were recently arrested. The bottom panel of Figure 2.2 shows that among those in state A , men were more likely than women to remain there at the subsequent wave, meaning that they were at higher risk of being rearrested. Hispanic men were also more likely than White men to remain in A . There were no significant differences by gender or race/ethnicity in the probability of moving from A to C_2 . The probability of transitioning from A to R , desisting from criminal activity among those who were arrested, was higher for women compared to men, and it was especially high among Hispanic women (the interaction term between female and Hispanic was marginally significant, $p < .10$).

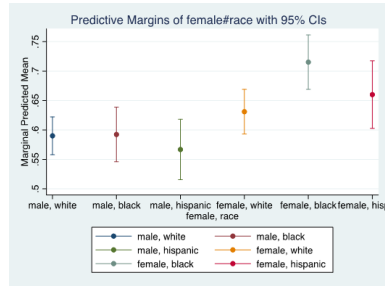
The graphs in the top panel of Figure 2.3 show subgroup differences in the transition probabilities for people in state C_2 at a given wave, meaning that they had been arrested at some point during the study period and had offended within the last year. Blacks were



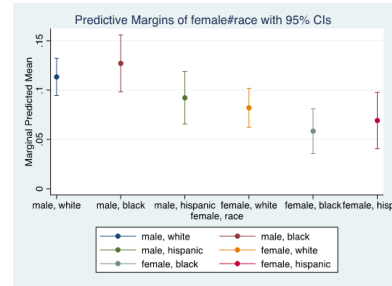
$X \rightarrow C_1$ by Subgroup



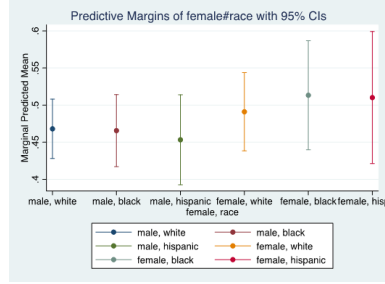
$X \rightarrow A$ by Age and Subgroup



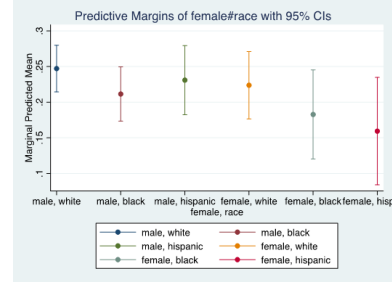
$C_1 \rightarrow X$ by Age and Subgroup



$C_1 \rightarrow A$ by Age and Subgroup

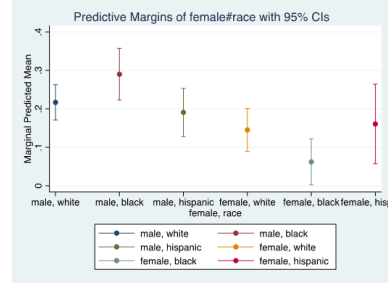
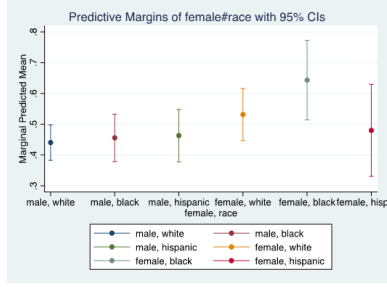


$A \rightarrow R$ by Age and Subgroup



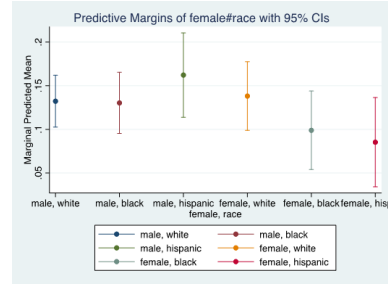
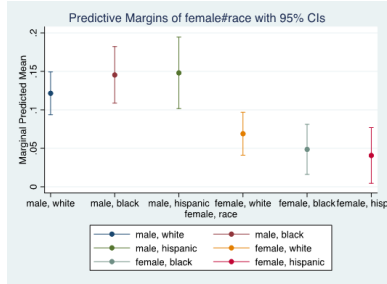
$A \rightarrow C_2$ by Age and Subgroup

Figure 2.2: Transition Estimates I



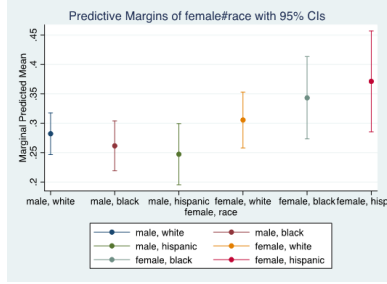
$C_2 \rightarrow R$ by Subgroup

$C_2 \rightarrow A$ by Age and Subgroup



$R \rightarrow A$ by Age and Subgroup

$R \rightarrow C_2$ by Age and Subgroup



$R \rightarrow X$ by Age and Subgroup

Figure 2.3: Transition Estimates II

less likely than Whites to remain in C_2 from one wave to the next.³ Black men in C_2 were more likely to move to state A , meaning that they were rearrested by the subsequent wave, compared to either White men ($p < .07$) or Black women ($p < .03$). Women in C_2 were marginally more likely ($p < .08$) than men to transition from C_2 to R at the subsequent wave, meaning that by the subsequent wave they had desisted from offending and were not rearrested. There were no significant differences across racial/ethnic groups in the transition from C_2 to R .

Finally, the graphs in the middle and bottom panel of Figure 2.3 show subgroup differences in the transition probabilities for people in state R at a given wave, meaning that they had previously been arrested but had neither offended nor been arrested during the past year. There were no significant differences by gender or race/ethnicity in the likelihood of remaining in R at the subsequent wave or transitioning to C_2 (which would indicate a resumption of offending). Men were more likely than women to transition from R to A , meaning that they were rearrested, but there were no differences across racial/ethnic groups in this transition. In the transition from R to X , which indicates that the person had gone three years without being arrested or involved in offending, there was a marginally significant ($p < .08$) positive interaction between being Hispanic and female, such that women were more likely than men to make this transition, but the gender gap was only significant only among Hispanics.

We now summarize the main findings from the analysis presented in Figures 2.2- 2.3 with regard to differences by gender and race/ethnicity. The likelihood of remaining in state X , without onset of criminal behavior or arrest, was significantly lower for men (compared to women) and Blacks (compared to Whites). Men were more likely to experience all transitions to A , meaning that they had a higher risk of arrest or rearrest. There were also race differences in some transitions leading to A . Most notably, Black men were at a significantly higher risk of transitioning from $X \rightarrow A$. In contrast, there

³Although the gap between Blacks and Whites in the predicted probabilities for the transition from C_2 to C_2 was larger among men (0.257 for Blacks compared to 0.343 for Whites) than women (0.305 for Blacks compared to 0.330 for Whites), there was no significant interaction between being Black and female.

were no racial/ethnic differences in the $X \rightarrow C_1$ transition, which marks the onset of criminal offending without an arrest. There were also some instances in which women of color were more likely to make transitions away from crime and arrest. Black women were more likely (compared to either White women or Black men) to transition from $C_1 \rightarrow X$, and Hispanic women were more likely (compared to Hispanic men) to transition from $A \rightarrow R$ and $R \rightarrow X$. Hispanic women were also less likely than all other groups to be arrested for the first time by transitioning from $X \rightarrow C_1$.

2.5.3 *Results of the Sensitivity Simulation*

Small changes

We also computed and verified the expressions in Section 2.2.4 for the partial derivatives of X^* , which measure the sensitivity of the value of equilibrium X^* to changes in each parameter. For Black and White males, we increased the value of each parameter by 1 percentage point and computed the simulated change in X^* . We found in all cases that this marginal impact matched the partial derivatives calculations in Section 2.2.4. These results are available upon request.

If a partial derivative of X^* for some transition parameter is close to zero, then the value of X^* will not change much if slightly inaccurate values of that parameter are reported. On the other hand, a marginal independent increase in such a transition parameter will have little impact. A partial derivative with a large magnitude indicates both a parameter that requires a more careful estimate from the data, and a potentially sensitive intervention target that could have a large impact on the long-run level of crime.

The partial derivatives are illustrated in Figure 2.4. However, these partial derivatives represent the impact of a one percentage point change, which can result in a very different “percent change” depending on the current value of a parameter. Similarly, a resulting percentage point change the equilibrium proportion of people in X from 0.55 is very different from a percentage point change in that proportion from 0.9. To get a better view of the sensitivity of the equilibrium X^* to deviations in the parameters in a way

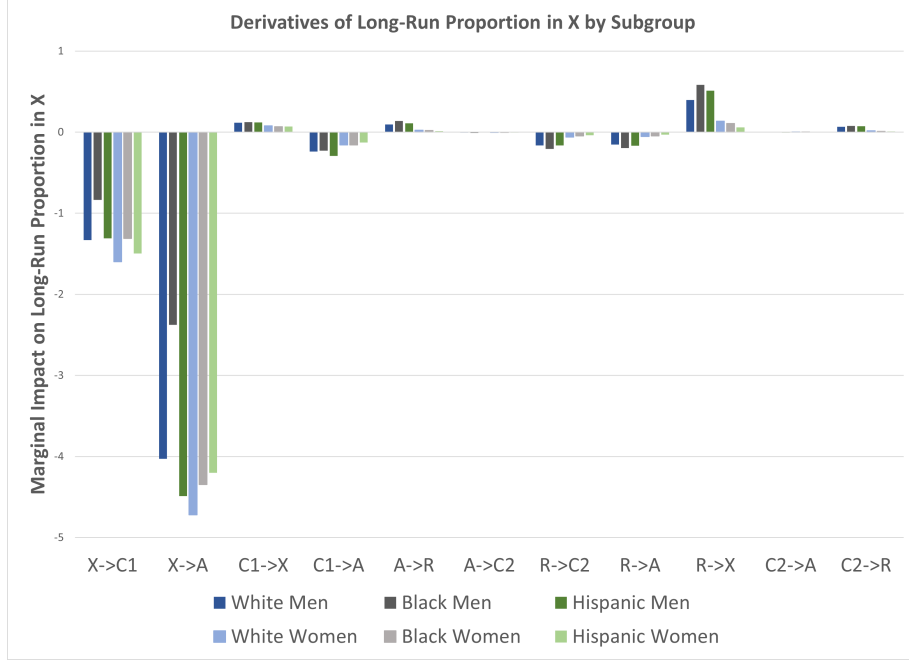


Figure 2.4: Derivatives of Equilibrium Proportion in X

that accounts for these concerns, we present elasticities in Figure 2.5.

As expected, increasing transition probabilities from X into crime or arrest α_{x1} , α_{rc} , α_{xa} , α_{ra} decreased the long-run proportion X^* of law-abiding citizens; increasing desistance rates β_{cx} , β_{ar} , from criminal behavior, and increasing rehabilitation ϵ_{rx} increased X^* .

Our mathematical analysis in Section 2.2.5 indicated that increases in the arrest parameters γ_{1a} ($C_1 \rightarrow A$) and γ_{2a} ($C_2 \rightarrow A$) have ambiguous effects on X^* . Essentially, if an arrest is more likely to lead to eventual desistance and rehabilitation, higher arrest rates lead to lower crime levels ($\partial X^*/\partial \gamma > 0$). Otherwise, higher arrest rates can increase the equilibrium level of crime. This ambiguous effect could work out differently for different subgroups, depending on the values of the other parameters. Using the NLSY97 data we estimated this effect for all subgroups. As indicated in Figure 2.4, for all six subgroups in our sample, $\partial X^*/\partial \gamma_{1a}$ is negative; a marginal increase in arrest rates increases long-run crime for all criminally active subgroups with no arrest history. On the other hand, $\partial X^*/\partial \gamma_{2a}$ is essentially negligible; increasing arrest rates of recidivists has basically no effect on long-run crime levels. Decreasing the virgin arrest rate for the

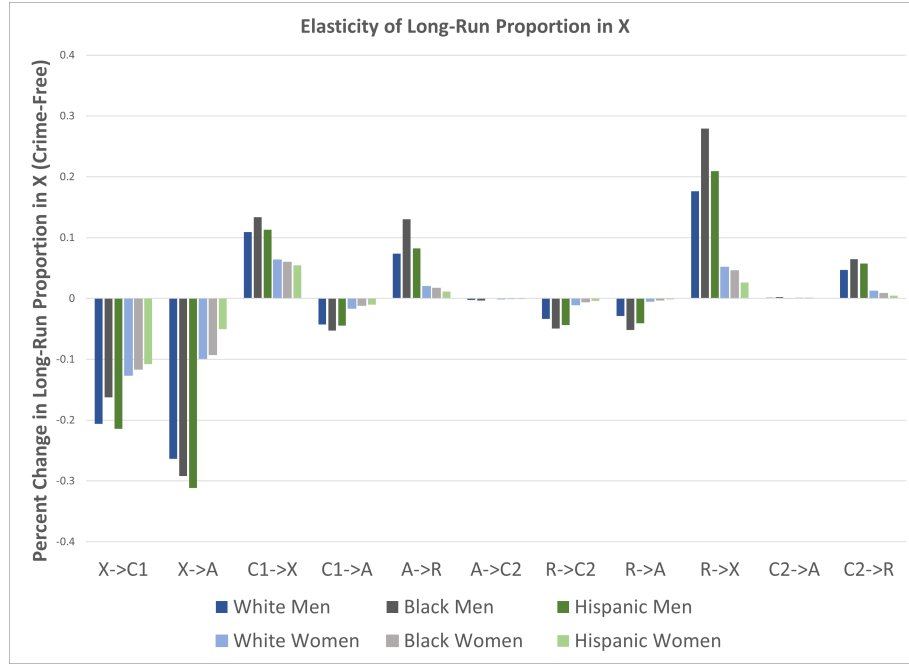


Figure 2.5: Elasticities of Equilibrium Proportion in X

criminally active would have a larger impact on male crime than female crime levels, and Black males most of all.

Overall, the long-run proportion of law-abiding citizens X^* is generally most sensitive to changes in the initial transitions into criminal activity ($X \rightarrow A$ and $X \rightarrow C_1$). However, $R \rightarrow X$ appears to be more important than $X \rightarrow C_1$ particularly for Black men, and nearly as important for White and Hispanic men. For example, a 10 percent increase in rehabilitation would increase the equilibrium proportion of law-abiders by nearly 2.8 percent for Black men, 1.8 percent for White men, and 2.1 percent for Hispanic men. An analogous 10 percent decrease in the $X \rightarrow C_1$ would increase the equilibrium proportion of law abiders by 1.6 percent for Black men, 2.1 percent for White men, and 2.1 percent for Hispanic men. For men, pre-arrest desistance $C_1 \rightarrow X$ appears to be at least important than post-arrest desistance ($A \rightarrow R$, $C_2 \rightarrow R$). Male crime levels appear to be more sensitive to post-arrest transitions than female crime rates, and Black males most of all. Finally, it's worth noting that preventing flow into crime among recent arrestees (e.g. changing $R \rightarrow C_2$ or $R \rightarrow A$) doesn't have nearly the same systemic effect as preventing those in X from flowing into crime or arrest in the first place (e.g., $X \rightarrow A$

and $X \rightarrow C_1$).

For women, focusing percent changes on “pre-arrest” parameters is most effective with no exceptions. For example, decreasing $X \rightarrow C_1$ is most effective for women: decreasing $X \rightarrow C_1$ by 10 percent increases the equilibrium proportion of law-abiders by just over 1 percent for all women. A 10 percent increase in the desistance rate for the criminally active with no arrest history $C_1 \rightarrow X$ increases X^* by 0.6, 0.63, and 0.54 percent for Black, White, and Hispanic women, respectively. The next most important transition for women appears to be $R \rightarrow X$, with the remaining transition making very little difference.

In the next section we reveal how larger independent changes in each parameter impact the long-run level of crime. In doing so we also investigate how the local sensitivities depend on the current values, were the current values estimated at different levels. Elasticities are readily obtainable from the analysis.

Large changes

We have examined the effects of small changes in the transition parameters. But many proposed interventions involve large changes in some parameter, e.g., closing prisons or decriminalizing certain drugs. In this Section we investigate the sensitivity of X^* to large changes in each parameter. The simulation also allows us to investigate how the sensitivity of X^* to deviations in each parameter depends each parameter’s current level. To investigate the possible effects of larger changes for each subpopulation, we vary each transition parameter one at a time over a wide range that still ensures all probabilities are between zero and one.

We present the simulation for Black males in Figures 2.6, 2.7, 2.8, and 2.9. In Figures 2.6 and 2.7, the solid black line indicates the equilibrium X^* , which is the long-run proportion of people in state X (law-abiding, rehabilitated/never-arrested citizens). Increasing curves indicate that the proportion of these law-abiders will increase (so that crime decreases) as the parameter increases, and decreasing curves indicate that the proportion of law-abiders will decrease (so that crime increases) as the parameter increases.

For each parameter curve in these figures, the estimated value of the underlying parameter lies at its intersection with the solid black line. The slopes of the curves at these points are consistent with the partial derivative calculations in the previous Section. They represent the marginal impact of a percentage point increase in a parameter on the equilibrium proportion of people in X . Notice that the marginal impact of a change in a transition probability depends on the current level of that probability in the data, but to very different extents for the various parameters.

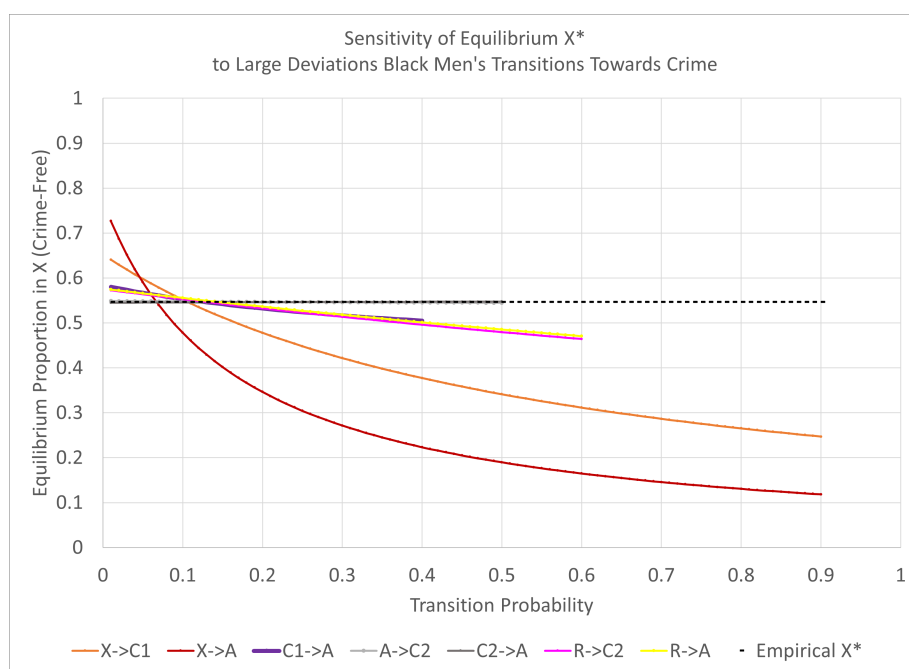


Figure 2.6: Large Parameter Changes Towards Crime: Black Men

As one might expect, the biggest impacts on the level of crime-free X^* arise from percentage-point changes in the rates for leaving X^* via new criminal activity – both with ($\alpha_{xa} : X \rightarrow A$) and without ($\alpha_{x1} : X \rightarrow C_1$) early arrest. For example, changing α_{xa} can drop the equilibrium percentage of crime-free Black males 43 percentage points from its current equilibrium at 55%, or raise it by 18 percentage points. Similar changes for Black females can decrease X^* from 82% to 16%. Interestingly, increases in α_{xa} have larger negative effects than increases in α_{x1} . Note again that increases in $\gamma_{1a} : C_1 \rightarrow A$ increase crime to some degree, holding the other transitions constant. In our sample, increasing the intensity of virgin arrests can lead to increases in the equilibrium crime

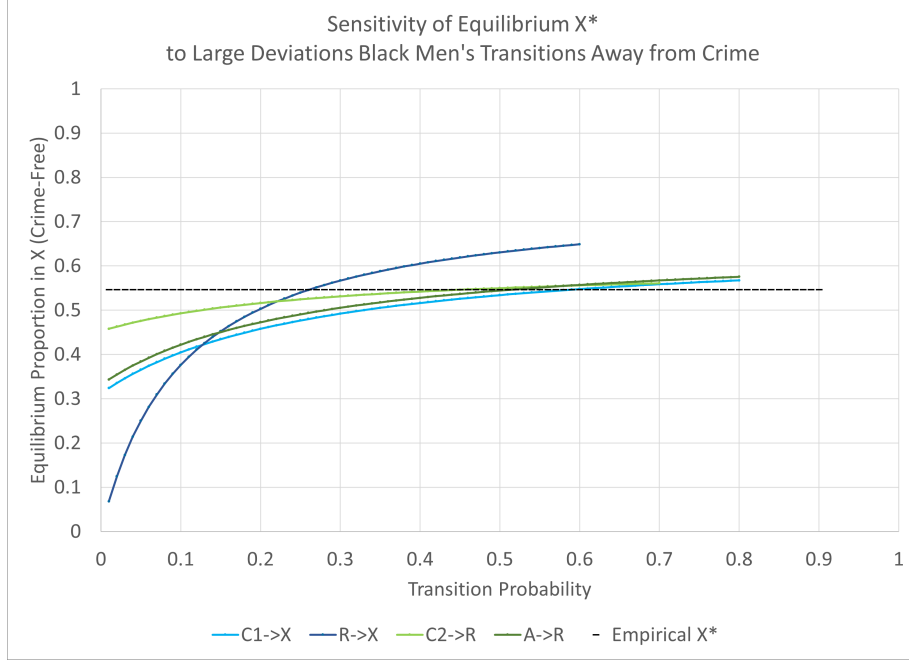


Figure 2.7: Large Parameter Changes Away from Crime: Black Men

rate (decreases in X^*) for all subgroups. However, increases in $\gamma_{2a} : C_2 \rightarrow A$ has no impact on the equilibrium crime rate, as indicated by its nearly horizontal curve. So increasing the arrest rate for those with an arrest history appears to have no effect in contrast to the effect of arresting those with no arrest history. This turns out to be true for all subgroups.

While large percentage point changes in $X \rightarrow A$ can cause X^* to deviate more than large changes in rehabilitation $R \rightarrow X$, at sufficiently small levels, X^* is more locally sensitive to $R \rightarrow X$ than to any other transition including $X \rightarrow A$. It happens to be the case that the $R \rightarrow X$ estimate is within a range that allows percentage-point deviations to have smaller effects than the onset pathways ($X \rightarrow A$ and $X \rightarrow C_1$). Nonetheless, even at the estimated level it appears to be the third most-important transition for long-run Black male crime levels from the perspective of percentage-point changes.

Finally, Figures 2.8 and 2.7 indicate the “percent” changes in the crime-free equilibrium X^* in response to “percent” changes in the transition probabilities. This takes into account how percentage-point changes above very small percentages are very dif-

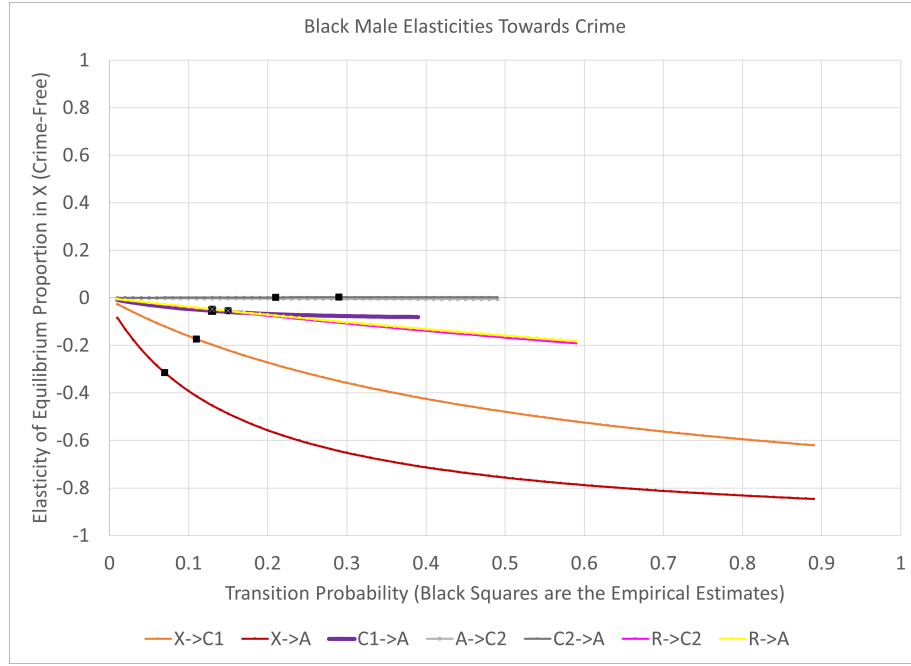


Figure 2.8: Large Parameter Changes: Black Men

ferent from percentage-point changes above very large percentages. In these plots, the black dots indicate the empirically estimated levels. We can see that the equilibrium is most sensitive to percent changes in $X \rightarrow A$ and $R \rightarrow X$ for Black men, and that the equilibrium X^* response $X \rightarrow A$ would become even more elastic for larger values of $XtoA$, while the response to $R \rightarrow X$ would become even more elastic for smaller values of $R \rightarrow X$. These plots tell an otherwise similar story as the previous two Figures, although they illustrate more relative importance for the other transitions besides $X \rightarrow A$.

2.6 Results on Racial Differences via Simulations

In Section 2.5.3 we examined the impacts of independent changes in the transition parameters of our model on the equilibrium level of crime. The objective was to discover which paths might most efficiently lead to decreases in the level of crime. In this Section we instead use our model to shed light on which transitions contribute most to disparities in the equilibrium proportion of law-abiding citizens X^* for Black and White males.

We use the transition probabilities estimated for Black and White men in Figures 2.2

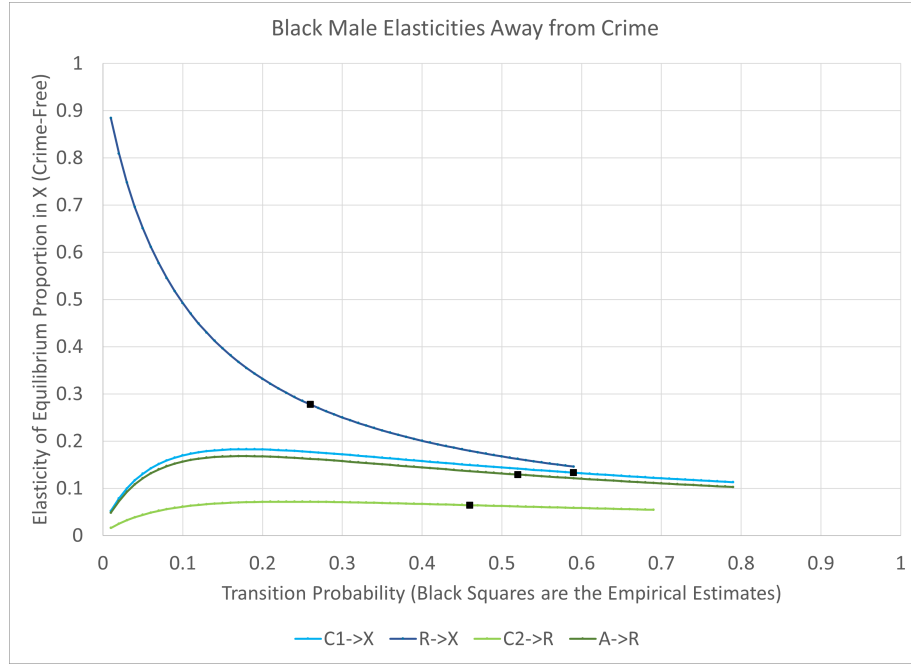


Figure 2.9: Large Parameter Changes: Black Men

and 2.3, calculating the Black-White male percent differences for each parameter. We then simulate the percent decrease in the Black-White male disparity in X^* for every one-percent decrease in the racial gap in each transition probability. Finally, we simulate the percent decrease in the Black-White male disparity in X^* that occurs as a result of giving Black males the transition probabilities of White males, one at a time. Altogether, these exercises reveal 1) the largest and smallest disparities in the transition pathways; 2) the sensitivity of the disparity in the equilibrium X^* to comparable changes in each transition probability; and 3) which gaps, if fully closed would make the most productive, smallest, or counterproductive difference in the disparity in the equilibrium X^* .

Percent Differences in Transition Probabilities

We first calculated the percent differences in the transition probabilities for White and Black males using the formula

$$\frac{\text{White rate} - \text{Black rate}}{\text{Black rate}}.$$

The largest percent difference was in the arrest rate for new offenders: 38% for $\alpha_{xa} : X \rightarrow A$. This was partially due to the small size of the actual probabilities: about 4.1% for Black males and 6.7% for White males. The next largest percent difference was 25.2% for re-arrest after returning to crime ($\gamma_{2a} : C_2 \rightarrow A$); 17% for those arrested in the last year who continued committing crimes without arrest ($\zeta_{a2} : A \rightarrow C_2$); 16% for the arrest of criminally inactive people with a recent arrest history ($\gamma_{ra} : R \rightarrow A$); 10/8% for first arrest among the criminally active with no arrest history ($\gamma_{1a} : C_1 \rightarrow A$); and 7.8% for rehabilitation ($\epsilon_{rx} : R \rightarrow X$).

It is worth noting that Black males are better off than White males along certain pathways. For example, the Black male desistance rates $\rho_{ar} : A \rightarrow R$ and $\beta_{2r} : C_2 \rightarrow R$ are both about 3.3% larger than that of White males. Black males are about 1.5% less likely than White males to make the $\alpha_{r2} : R \rightarrow C_2$ transition into crime, and a negligible 0.4% more likely to desist from crime $\beta_{1x} : C_1 \rightarrow X$ among those with no arrest history. Therefore, closing the racial gap in those transition rates would only exacerbate the racial disparity in the equilibrium X^* .

Systems Effects

To understand the systemic effect of racial differences in any given transition probability on the overall racial differences in levels of criminal activity, we ran our dynamic simulation (2.1) with the transition parameters for Black males with one exception – we changed one parameter to the corresponding value for White males. We repeated this for each parameter. The results are pictured in Figures ?? through 2.24.

First, Figure 2.10 displays how a comparable percent reduction in the transition probabilities leads to a percent reduction in the racial disparity in the equilibrium X^* . Some of the racial gaps in transitions are much larger than others, so it is worth understanding how a comparably realistic attempt to close each gap could lead to different reductions in the overall X^* disparity. Using these elasticities we see that reducing the $C_1 \rightarrow A$ gap by 1% would reduce the X^* racial difference by about 0.36%, while reducing the $C_2 \rightarrow A$

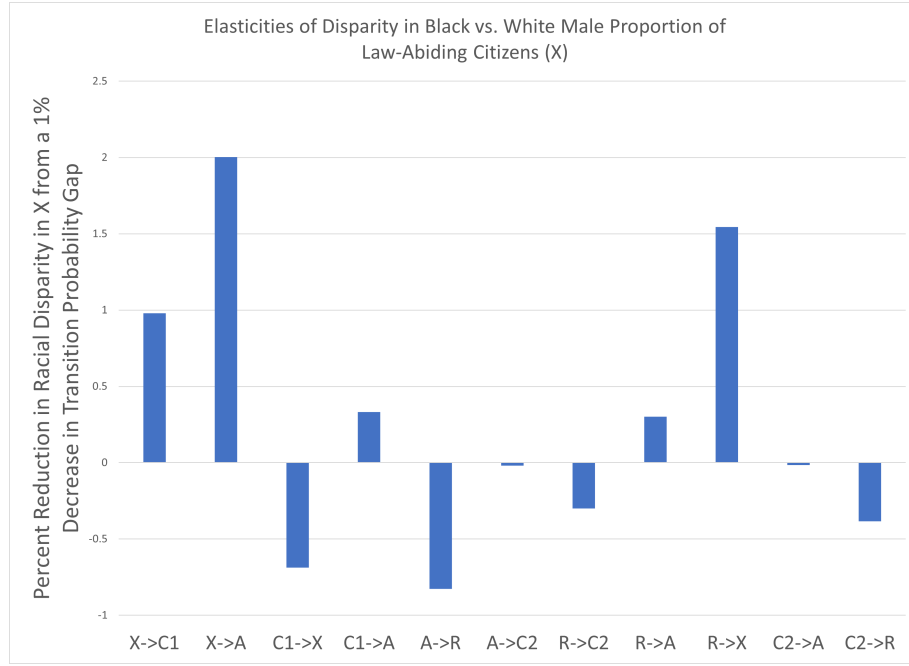


Figure 2.10: Elasticity of Black-White Male Disparity in X

gap would have essentially no effect. A reduction in the $R \rightarrow A$ gap has a comparable effect as reducing the $C_1 \rightarrow A$ gap. Reducing the $X \rightarrow A$ gap by 1% reduces the X^* racial difference by 2%. This is a genuine systemic effect, above and beyond the fact that the $X \rightarrow A$ racial gap is itself the largest gap. Reducing the $R \rightarrow X$ and $X \rightarrow C_1$ gaps by 1% would reduce the X^* racial difference by 1.5% and 0.98%, respectively.

Finally, Figure 2.11 depicts the overall reduction in the racial percent difference in X^* when the racial gap in each transition is fully closed (and not just closed by 1%). The actual sizes of the gaps in the transitions matters more in this figure. Indeed, both because $X \rightarrow A$ has a strong genuine systemic effect and because it constitutes the largest percent gap, the impact of closing its gap dwarfs the impact of closing each of the other gaps. Taken together, the figures imply that the $X \rightarrow A$ transition is most important for the racial gap in X^* ; and that small reductions in the $R \rightarrow X$ and $X \rightarrow C_1$ gaps would also have significant reductions in the racial gap in X^* . There is, however, more room for improvement, in a relative sense, for reductions in the $X \rightarrow A$ gap.

As explained in the previous subsection, closing the racial gap in $A \rightarrow R$, $C_2 \rightarrow R$, $R \rightarrow C_2$, and $C_1 \rightarrow X$ only exacerbate the racial disparity in the equilibrium X^* (they

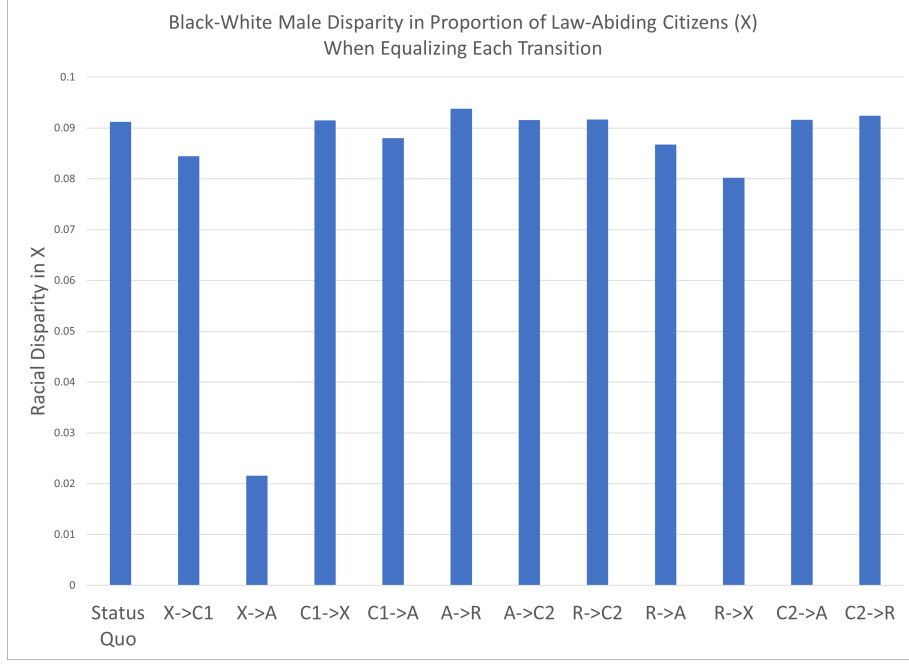


Figure 2.11: Black-White Male Disparity in X When Equalizing Each Transition

“decrease the reduction” in the disparity). This is because Black males are better off than White males along these transition pathways.

2.7 Discussion

2.7.1 Summary

This paper presents our first effort in combining an analytic model with a rich data set to gain insights on the spread of crime. Our model has led us to focus on the eleven transition parameters in Figure 2.1. We have noted many strong sex-based differences in yearly transitions into and out of self-reported crime and arrest. We have also seen fewer strong racial differences, most of which were specific to males. The strongest racial differences appeared in yearly transitions into arrest. There was one exception: the rate at which criminally active Black women with no arrest history desisted from crime in the next year ($\alpha_{1x} : C_1 \rightarrow X$) dwarfed that of their White counterparts, and represented the largest racial difference for women.

The $\alpha_{2a} : C_2 \rightarrow A$ transition parameter (rate of re-arrest) differed the most between

White males and Black males. However, increasing the $\alpha_{1a} : C_1 \rightarrow A$ transition parameter (rate of first arrest after a period of criminal activity) had a larger impact, increasing long-run crime for all subgroups. In contrast, increasing $C_2 \rightarrow A$ has a negligible impact on long-run crime. *Equalizing* the $C_1 \rightarrow A$ transition parameter across races had a larger impact on reducing racial differences in long run crime than the $\alpha_{2a} : C_2 \rightarrow A$ transition parameter.

The $X \rightarrow C_1$ “onset” probability was the second most important transition in terms of the local sensitivity of the long-run crime level for all subgroups, in line with the mantra “prevention is the best cure.”

The $\alpha_{xa} : X \rightarrow A$ transition parameter (rate of first arrest in the first year of criminal activity, as discussed in Section 2.2.6) had the largest *marginal* effect in reducing both crime levels and racial differences in crime level. However, we cannot state whether this is primarily due to actual racial differences in initial crime participation (which could reflect differences in neighborhood context, socioeconomics and educational opportunity), to racial differences in the policing of citizens with no arrest history, or both.

2.7.2 Real World Policy

While the analysis presented here provides insights into possible consequences of certain types of interventions, there is a lot to do before this systems approach can offer real world policy advice. For example, a real policy intervention might be focused on reducing the actual level of crime, and not simply the number of those engaged in criminal activity. We have also said nothing about the effect of crime on its victims – the real cost of crime.

Furthermore, most real world interventions can affect a number of different model parameters in our model. For example, increasing the level of police activity can affect all of the arrest parameters, $\alpha_{xa}, \alpha_{1a}, \alpha_{2a}, \alpha_{ra}$, but it can also affect the desistance rates β_{1x} and β_{2x} . Mandatory prison sentences could raise or lower arrest rates. Our focus has been on the systemic effects of marginal, independent changes in particular target pathways of intervention – not the interventions themselves.

The most obvious weakness of our model in terms of policy recommendations lies in its basic simplicity – in the simplifying assumptions we made to keep this our first model analytically tractable. We list some of these assumptions below, noting that we plan to relax these in future work.

2.7.3 Simplifying Assumptions

Heterogeneity of Agents. We made the standard first-model assumption that the population under study is relatively homogeneous. Our analysis attempts to explore the possible consequences of a state-dependent criminal lifecycle, in contrast to Nagin’s notion of population heterogeneity (that person-varying, time-invariant traits lead to the correlation of offending over time). Under population heterogeneity we would expect agents in the various states to be systematically different across a wide range of unobservable person-varying time-invariant traits; changing one transition probability can change other transition probabilities by changing the distribution of “types” of people in the various states. We did address this to an extent by estimating our probabilities using person-specific time-invariant random intercepts.

Linear transitions In our first two papers (references omitted), we assumed that crime spread like a contagious disease through interactions between those engaged in criminal activity and those who are law-abiders. In this paper, we assumed that each agent’s entry into criminal activity is unrelated to other agents (linear transitions). It is likely that both of these dynamics play a role. We intend to quantify the role of contagion vs linear flow in commencement of criminal activity. Eventually, we will add a network or community structure, including gang behavior, in our study of the spread of crime.

Heterogeneity in Crime. A major simplifying assumption was that a wide range of crimes are categorized as a generic crime in this model. The NLSY97 data set includes different kinds of crime, and we plan to incorporate those differences. We hope to understand how the results of the previous analysis might vary by crime type.

Heterogeneity in Transition Parameters. To keep our model a tractable Markov

system, we assumed that the transition parameters are independent of state sizes and fixed over time; and we focused on long-run equilibrium to compare sexes and races. Yet, we know that crime has been decreasing over the last decade and that the seven-year window of the NLSY97 data may not constitute a long-run equilibrium. We hope to account for these concerns in future work.

Criminal Justice System. Our first two papers (references omitted) focused on imprisonment as the vehicle of punishment as opposed to arrest. We did not include information on imprisonment in our model because imprisonment is a relatively rare event in the NLSY97 data set. More developed models will need to account for both. Furthermore, as we expand to focus on criminal justice policy, we will include compartments for police, prosecutors, and judges, along with transition parameters for probation, parole, and sentence length.

2.8 Appendix

2.8.1 *Simple 3D Model: Model Structure*

We begin with a simplified three-dimensional modification of this paper’s five-dimensional model (see Figure 2.12). The three states are X, C, A :

X: Those who have neither committed crime nor been arrested in the last year.

C: Those who have committed a crime in the last year but were not arrested.

A: Those who have been arrested in the last year, with or without crime commission.

For a person in state X (someone who has not committed a crime or been arrested in the past year) in a given time period, σ_{xx} represents the probability of remaining in X (remaining free of crime and arrest) during the next time period, α_{xc} represents the probability of moving into C (becoming criminally active without arrest), and α_{xa} represents the probability of moving into A (moving to criminal activity and being arrested for that activity in the next year).

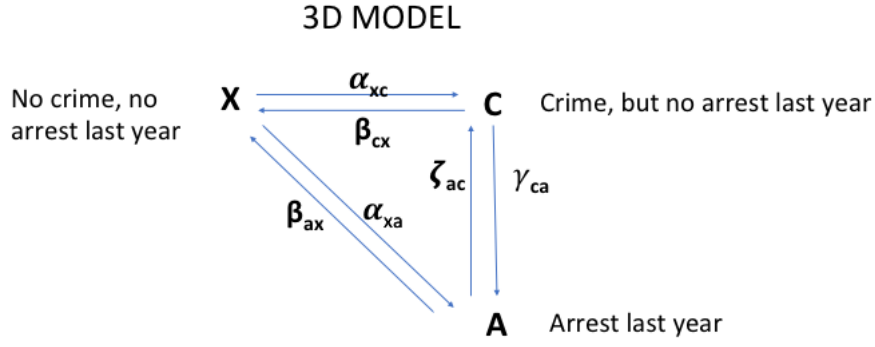


Figure 2.12: Flow Diagram for the 3D Model

For a person in state C (someone who has committed a crime in the past year but not been arrested) in a given time period, σ_{cc} represents the probability of remaining in C (continuing to commit crime) during the next time period, γ_{ca} is the probability of moving to state A (being arrested), and β_{cx} is the probability of moving to state X (not having committed a crime or been arrested in the past year).

For a person in state A (someone who has been arrested in the past year) in a given time period, σ_{aa} represents the probability of remaining in A (being arrested again) during the next time period, ζ_{ac} is the probability of moving to state C (committing crimes without being arrested), and β_{ax} is the probability of moving to state X (not having committed a crime or been arrested in the past year).

We use the Greek letter α for flows from crime-free to crime (with or without arrest), β for flows back to a crime-free status, γ for arrest, and ζ for return to crime after arrest. In each of the above sentences, the conditional probabilities sum to 1, as illustrated in the following equation.

$$\sigma_{xx} = 1 - \alpha_{xc} - \alpha_{xa}, \quad \sigma_{cc} = 1 - \gamma_{ca} - \beta_{cx}, \quad \sigma_{aa} = 1 - \zeta_{ac} - \beta_{ax}. \quad (2.3)$$

For simplicity, we did not include the σ s in Figure 2.12. The following linear (Markov)

system of equations captures this dynamic:

$$X(t+1) = \sigma_{xx}X(t) + \beta_{cx}C(t) + \beta_{ax}A(t) \quad (2.4a)$$

$$C(t+1) = \sigma_{cc}C(t) + \alpha_{xc}X(t) + \zeta_{ac}A(t) \quad (2.4b)$$

$$A(t+1) = \sigma_{aa}A(t) + \gamma_{ca}C(t) + \alpha_{xa}X(t) \quad (2.4c)$$

Using (2.3), rewrite system (2.4) as:

$$X(t+1) - X(t) = \beta_{cx}C(t) + \beta_{ax}A(t) - \alpha_{xc}X(t) - \alpha_{xa}X(t) \quad (2.5a)$$

$$C(t+1) - C(t) = \alpha_{xc}X(t) + \zeta_{ac}A(t) - \beta_{cx}C(t) - \gamma_{ca}C(t) \quad (2.5b)$$

$$A(t+1) - A(t) = \gamma_{ca}C(t) + \alpha_{xa}X(t) - \zeta_{ac}A(t) - \beta_{ax}A(t) \quad (2.5c)$$

Since the right hand sides of system (2.5) sum to 0,

$$X(t+1) + C(t+1) + A(t+1) = X(t) + C(t) + A(t), \quad (2.6)$$

for all times t . The population has constant size N . If we divide each term in linear system (2.4) or linear system (2.5) by N , we replace X, C, A by $X/N, C/N, A/N$ and the three variables in these systems become population *fractions* instead of total numbers. We will make these changes, but without changing the notation, so that we regard the three variables as population fractions.

Since the equations sum to zero, there is a redundant equation. Since the variables sum to 1, if we know two of the variables, we know the third. This means that we can write system (2.5) as system of two equations in two unknowns; for example, as:

$$C(t+1) - C(t) = -(\alpha_{xc} + \beta_{cx} + \gamma_{ca})C + (\zeta_{ac} - \alpha_{xc})A + \alpha_{xc} \quad (2.7a)$$

$$A(t+1) - A(t) = (\gamma_{ca} - \alpha_{xa})C - (\alpha_{xa} + \zeta_{ac} + \beta_{ax})A + \alpha_{xa} \quad (2.7b)$$

2.8.2 Simple 3D Model: Long Run Equilibrium Analytic Expression

Because of (2.3), system (2.4) is a Markov system of equations, which implies that every solution of (2.4) converges over time to the same long-run equilibrium X^* in population frequency space, from any non-zero starting point. To find X^* , we set the right-hand sides of system (2.7) equal to zero, which we rewrite as:

$$\alpha_{xc} = (\alpha_{xc} + \beta_{cx} + \gamma_{ca})C - (\zeta_{ac} - \alpha_{xc})A \quad (2.8a)$$

$$\alpha_{xa} = -(\gamma_{ca} - \alpha_{xa})C + (\alpha_{xa} + \zeta_{ac} + \beta_{ax})A \quad (2.8b)$$

and then solve for C and A , using standard linear algebra techniques. We then use $X = 1 - C - A$ to solve for the equilibrium X^* :

$$X^* = \frac{\beta_{cx}\zeta_{ac} + \beta_{cx}\beta_{ax} + \gamma_{ca}\beta_{ax}}{\alpha_{xc}(\zeta_{ac} + \beta_{ax} + \gamma_{ca}) + \alpha_{xa}(\zeta_{ac} + \beta_{cx} + \gamma_{ca}) + \beta_{cx}\zeta_{ac} + \beta_{cx}\beta_{ax} + \beta_{ax}\gamma_{ca}} \quad (2.9)$$

To understand (2.9) better, write it as:

$$X^* = \frac{1}{1 + \frac{\alpha_{xc}(\zeta_{ac} + \beta_{ax} + \gamma_{ca}) + \alpha_{xa}(\zeta_{ac} + \beta_{cx} + \gamma_{ca})}{\beta_{cx}\beta_{ax} + \zeta_{ac}\beta_{cx} + \gamma_{ca}\beta_{ax}}}. \quad (2.10)$$

The fraction in the denominator of (2.10) is an **output-input ratio**. The α_{xc} and α_{xa} in the numerator of this fraction parameterize the two direct paths out of X . The terms in the denominator parameterize the four paths into X : β_{cx} directly from C to X , β_{ax} directly from A to X , $\zeta_{ac}\beta_{cx}$ from A to X via C , and $\gamma_{ca}\beta_{ax}$ from C to X via A .

2.8.3 Simplified 3D Model: Long Run Equilibrium Dependence on Parameters

Understanding how the equilibrium X^* depends on the various parameters of the model is crucial to the systems approach. How does increasing the arrest rate γ or decreasing the rate of recidivism ζ affect the relative number of non-law-abiding citizens. Some parameters are easy to work with. For example, the rates α_{xc} and α_{xa} appear only in

the denominator of expression (2.9). Increasing either α decreases X^* ; the faster the flow into criminal activity, the higher the crime rate.

Understanding the effects associated with other parameters requires taking partial derivatives of (2.9). For example, taking the partial derivative of X^* with respect to β_{cx} (after applying the quotient rule of calculus with some algebraic simplification) shows how increasing probability of desistance (β_{cx}) increases the fraction of people in X^* , which means lowering the crime rate):

$$\frac{\partial X^*}{\partial \beta_{cx}} = \frac{[(\zeta_{ac} + \beta_{ax})\alpha_{xc} + \alpha_{xa}\zeta_{ac}](\zeta_{ac} + \beta_{ax} + \gamma_{ca})}{(\alpha_{xc}(\zeta_{ac} + \beta_{ax} + \gamma_{ca}) + \alpha_{xa}(\zeta_{ac} + \beta_{cx} + \gamma_{ca}) + \beta_{cx}\zeta_{ac} + \beta_{cx}\beta_{ax} + \beta_{ax}\gamma_{ca})^2} > 0. \quad (2.11)$$

A parallel computation shows that $\frac{\partial X^*}{\partial \beta_{ax}} > 0$.

The effect on X^* of increasing the arrest rate (γ_{ca}) is more nuanced, because the effect depends on the relative size of β_{ax} and β_{cx} : Once again, the quotient rule and some algebraic simplification yields:

$$\frac{\partial X^*}{\partial \gamma_{ca}} = \frac{(\beta_{ax} - \beta_{cx})(\alpha_{xc}\beta_{ax} + \alpha_{xc}\zeta_{ac} + \alpha_{xa}\zeta_{ac})}{(\alpha_{xc}(\zeta_{ac} + \beta_{ax} + \gamma_{ca}) + \alpha_{xa}(\zeta_{ac} + \beta_{cx} + \gamma_{ca}) + \beta_{cx}\zeta_{ac} + \beta_{cx}\beta_{ax} + \beta_{ax}\gamma_{ca})^2}. \quad (2.12)$$

If being arrested has a positive effect on desistance, $\beta_{ax} > \beta_{cx}$, then increasing the arrest rate decreases crime. However, if the probability moving from being arrested makes becoming crime-free less likely, $\beta_{ax} < \beta_{cx}$, then increasing the arrest rate can increase the crime level.

For each of the first partial derivatives we have computed, the second partial derivative has the opposite sign. Therefore, any independent increase in a transition probability has decreasing returns on X^* . For example, any positive effect on the crime level of reducing the transition rate α_{xc} from no-crime to crime has a smaller effect the larger the value of α_{xc} .

Finally, we compute that

$$\frac{\partial X^*}{\partial \zeta_{ac}} = \frac{(\beta_{cx} - \beta_{ax})(\alpha_{xc}\gamma_{ca} + \beta_{cx}\alpha_{xa} + \gamma_{ca}\alpha_{xa})}{(\alpha_{xc}(\zeta_{ac} + \beta_{ax} + \gamma_{ca}) + \alpha_{xa}(\zeta_{ac} + \beta_{cx} + \gamma_{ca}) + \beta_{cx}\zeta_{ac} + \beta_{cx}\beta_{ax} + \beta_{ax}\gamma_{ca})^2}. \quad (2.13)$$

Expression (2.13) suggests that if desistance is more likely for those not recently arrested ($\beta_{cx} > \beta_{ax}$), then decreasing ζ_{ac} might not decrease the crime rate.

We next expand our model to consider recidivism more directly. Such an expansion is important because recidivism plays an important role in the spread of crime and needs to be considered in any model that tries to capture real world crime dynamics.

2.8.4 Computing the equilibrium of the 5D System (1)

We compute the long run equilibrium of system (1) below:

$$X^* = \frac{\epsilon_{rx}(\beta_{1x} + \gamma_{1a})(\gamma_{2a}\beta_{ar} + \beta_{2x}(\beta_{ar} + \zeta_{a2}))}{D} \quad (2.14a)$$

$$C_1^* = \frac{\alpha_{x1}\epsilon_{rx}(\gamma_{2a}\beta_{ar} + \beta_{2x}(\beta_{ar} + \zeta_{a2}))}{D} \quad (2.14b)$$

$$A^* = \frac{(\alpha_{x1}\gamma_{1a} + (\beta_{1x} + \gamma_{1a})\alpha_{xa})(\beta_{2x}(\epsilon_{rx} + \alpha_{ra}) + \gamma_{2a}(\alpha_{x2} + \epsilon_{rx} + \alpha_{ra}))}{D} \quad (2.14c)$$

$$R^* = \frac{(\alpha_{x1}\gamma_{1a} + (\beta_{1x} + \gamma_{1a})\alpha_{xa})(\gamma_{2a}\beta_{ar} + \beta_{2x}(\beta_{ar} + \zeta_{a2}))}{D} \quad (2.14d)$$

$$C_2^* = \frac{(\alpha_{x1}\gamma_{1a} + (\beta_{1x} + \gamma_{1a})\alpha_{xa})((\epsilon_{rx} + \alpha_{ra})\zeta_{a2} + \alpha_{x2}(\beta_{ar} + \zeta_{a2}))}{D} \quad (2.14e)$$

where the common denominator is:

$$\begin{aligned} D = & \{ \alpha_{x1}[\epsilon_{rx}\gamma_{1a}\gamma_{2a} + \epsilon_{rx}\gamma_{2a}\beta_{ar} + \gamma_{1a}\gamma_{2a}\beta_{ar} + \gamma_{1a}\gamma_{2a}\alpha_{ra} + \epsilon_{rx}\gamma_{1a}\zeta_{a2} + \gamma_{1a}\alpha_{ra}\zeta_{a2} \\ & + \alpha_{x2}\gamma_{1a}(\gamma_{2a} + \beta_{ar} + \zeta_{a2}) \\ & + \beta_{2x}(\epsilon_{rx}(\gamma_{1a} + \beta_{ar} + \zeta_{a2}) + \gamma_{1a}(\beta_{ar} + \alpha_{ra} + \zeta_{a2}))] \\ & + (\beta_{1x} + \gamma_{1a})[\epsilon_{rx}(\gamma_{2a}(\beta_{ar} + \alpha_{xa}) + \alpha_{xa}\zeta_{a2}) + \alpha_{xa}(\gamma_{2a}(\beta_{ar} + \alpha_{ra}) + \alpha_{ra}\zeta_{a2} + \end{aligned}$$

$$\alpha_{x2}(\gamma_{2a} + \beta_{ar} + \zeta_{a2})) \\ + \beta_{2x}(\epsilon_{rx}(\beta_{ar} + \alpha_{xa} + \zeta_{a2}) + \alpha_{xa}(\beta_{ar} + \alpha_{ra} + \zeta_{a2})))\}$$

We claim:

$$\frac{\partial X^*}{\partial \beta_{1x}} > 0, \quad \frac{\partial X^*}{\partial \beta_{2r}} > 0, \quad \frac{\partial X^*}{\partial \beta_{ar}} > 0, \quad \frac{\partial X^*}{\partial \epsilon_{rx}} > 0,$$

We will prove the first of these; the proofs of the rest are similar. Write $D = D_1 + D_2$, where D_1 is the expression in the first two lines of D and D_2 is the rest. Write M for the numerator of X^* . Then,

$$\frac{\partial D}{\partial \beta_{1x}} = \frac{D_2}{\beta_{1x} + \gamma_{1a}} \quad \text{and} \quad \frac{\partial M}{\partial \beta_{1x}} = \frac{M}{\beta_{1x} + \gamma_{1a}}.$$

Then,

$$\frac{\partial X^*}{\partial \beta_{1x}} = \frac{1}{D^2} \left[\frac{M}{\beta_{1x} + \gamma_{1a}} (D_1 + D_2) - M \frac{D_2}{\beta_{1x} + \gamma_{1a}} \right] = M \frac{D_1}{\beta_{1x} + \gamma_{1a}} > 0.$$

As before, the effects of an increase in the arrest rates are quite a bit more subtle.

For re-arrests, the derivative is:

$$\frac{\partial X^*}{\partial \gamma_{2a}} = \frac{\epsilon_{rx}(\beta_{1x} + \gamma_{1a})(\beta_{ar} - \beta_{2r})[\alpha_{x1}\gamma_{1a} + (\beta_{1x} + \gamma_{1a})\alpha_{xa}][(\epsilon_{rx} + \alpha_{ra})\zeta_{a2} + \alpha_{r2}(\beta_{ar} + \zeta_{a2})]}{D^2}$$

The sign of this derivative of X^* with respect to γ_{2a} depends on β_{2r} and β_{ar} . Specifically, the derivative is nonnegative if $\beta_{ar} \geq \beta_{2r}$ and negative otherwise. Simply put, arresting criminals with an arrest history reduces crime if they are more likely to desist after arrest than before arrest.

The derivative for first arrests is:

$$\frac{\partial X^*}{\partial \gamma_{1a}} = \frac{\alpha_{x1}\epsilon_{rx}(\gamma_{2a}\beta_{ar} + \beta_{2r}(\beta_{ar} + \zeta_{a2}))}{D^2} \times \left[\epsilon_{rx}[\gamma_{2a}\beta_{ar} + \beta_{2r}(\beta_{ar} + \zeta_{a2})] \right]$$

$$-\beta_{1x}[\gamma_{2a}(\epsilon_{rx}+\beta_{ar}+\alpha_{ra})+\zeta_{a2}(\epsilon_{rx}+\alpha_{ra})+\alpha_{r2}(\gamma_{2a}+\beta_{ar}+\zeta_{a2})+\beta_{2r}(\epsilon_{rx}+\beta_{ar}+\alpha_{ra}+\zeta_{a2})]\Big].$$

2.8.5 Empirical Compartment Sizes

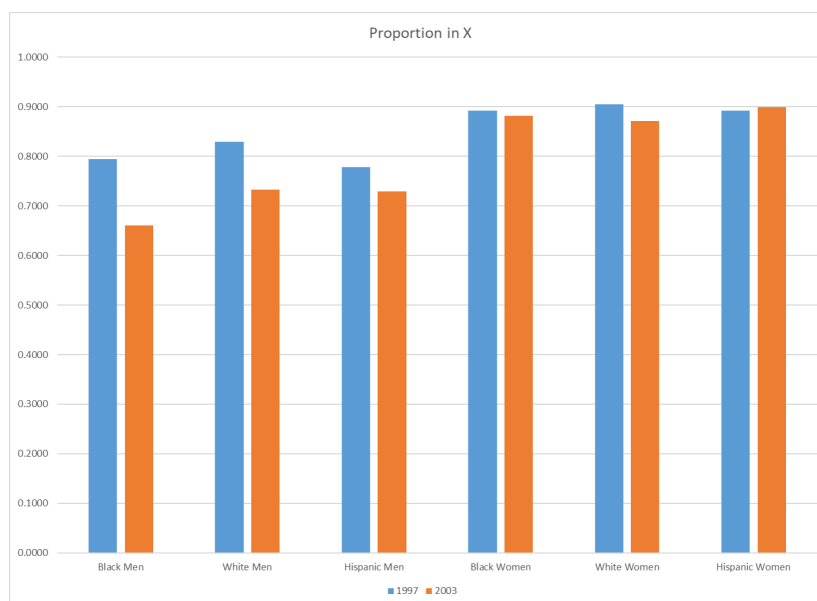


Figure 2.13: 1997 and 2003 Distributions for X

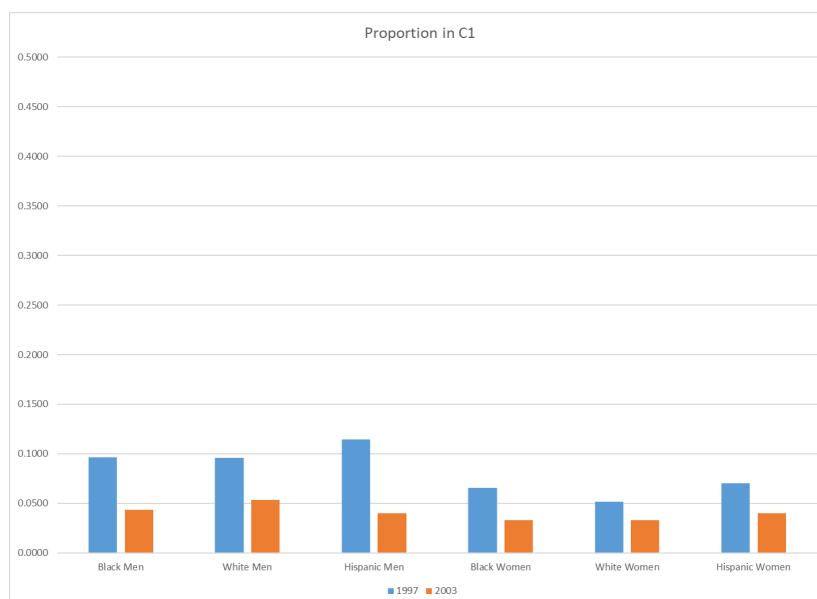


Figure 2.14: 1997 and 2003 Distributions for C1

,

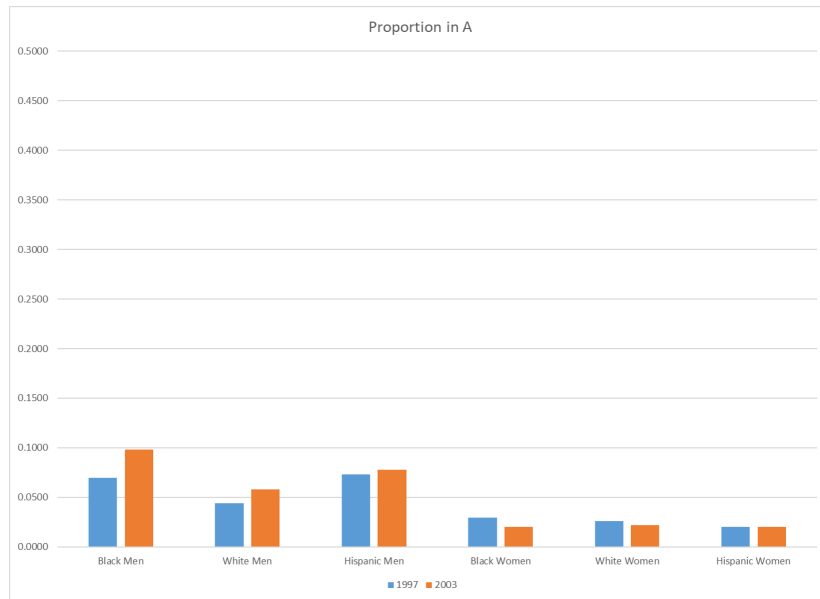


Figure 2.15: 1997 and 2003 Distributions for A

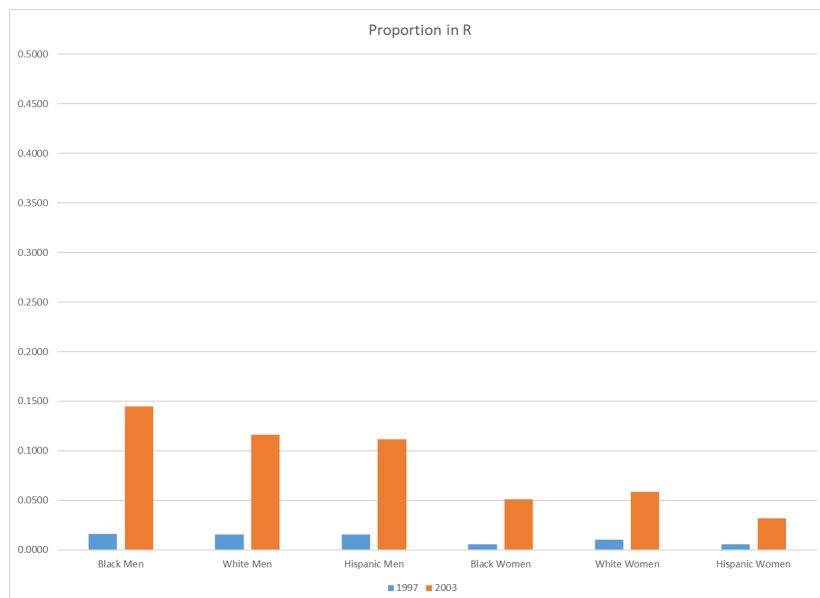


Figure 2.16: 1997 and 2003 Distributions for R

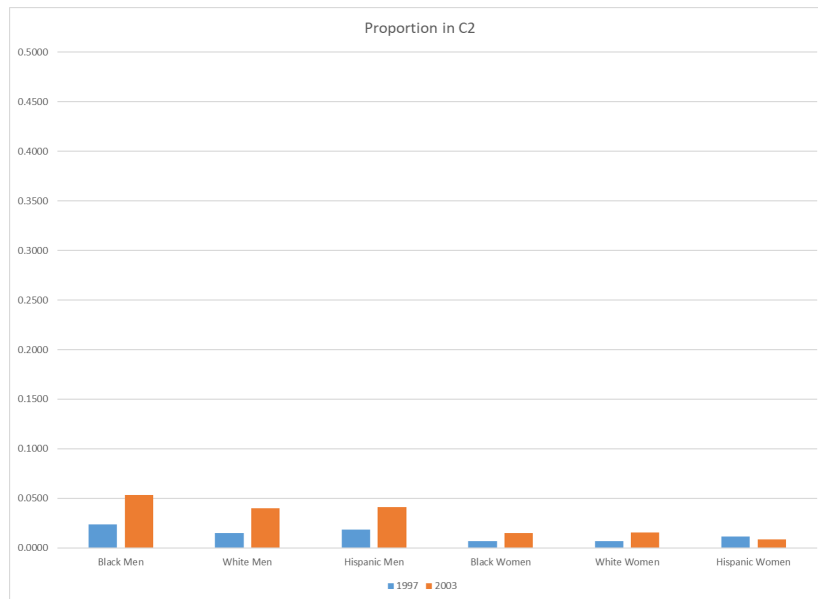


Figure 2.17: 1997 and 2003 Distributions for C2

2.8.6 Transition Probabilities

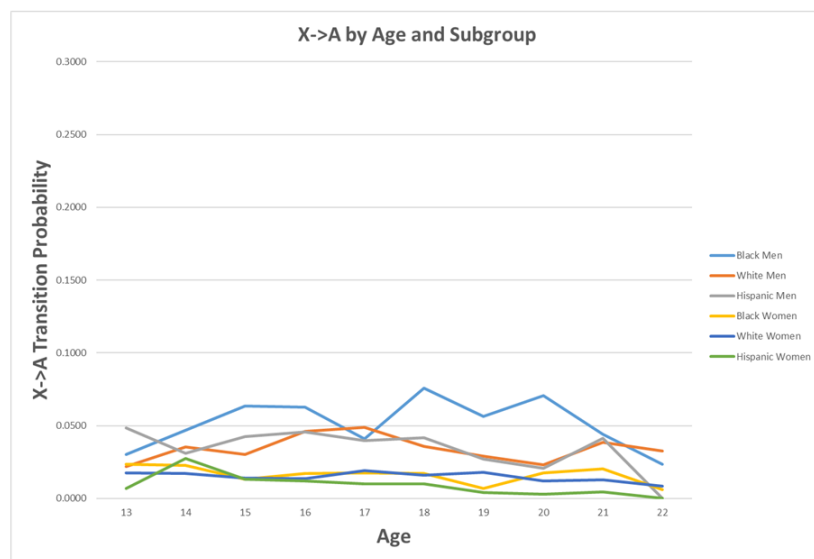


Figure 2.18: $X \rightarrow A$ by Age and Subgroup

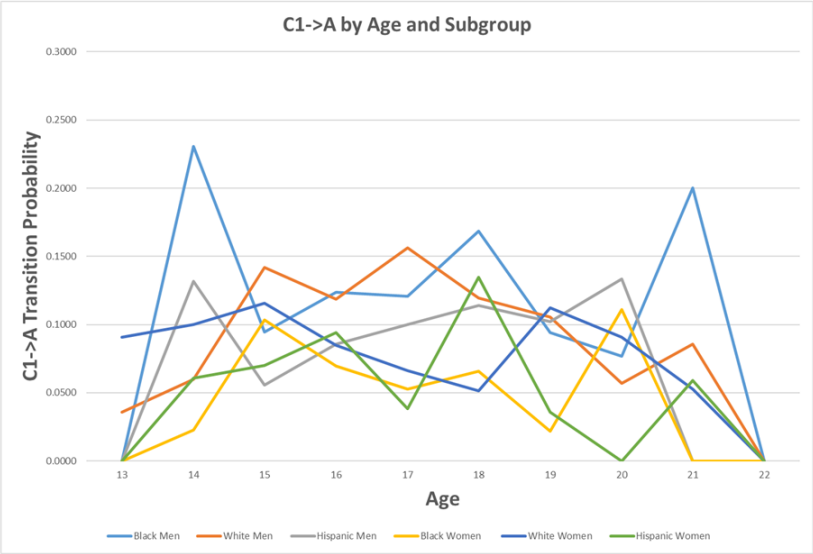


Figure 2.19: $C_1 \rightarrow A$ by Age and Subgroup



Figure 2.20: Transition Probabilities, Via Regression

Black Men	Initial	6 Steps	10 Steps	50 Steps	Equilibrium (300 steps)
X	0.7943	0.6799693	0.656729174	0.642341313	0.642340351
C1	0.0967	0.064828677	0.061796278	0.059919934	0.059919809
A	0.0694	0.088617147	0.092567589	0.095013734	0.095013897
R	0.0161	0.124774134	0.141555694	0.151943913	0.151944608
C2	0.0235	0.041810741	0.047351265	0.050781106	0.050781335
White Men	Initial	6 Steps	10 Steps	50 Steps	Equilibrium (300 steps)
X	0.8295	0.761490678	0.745119657	0.734131933	0.734130733
C1	0.096	0.068066833	0.066068803	0.064729525	0.064729378
A	0.0439	0.056158321	0.058620894	0.060274565	0.060274746
R	0.0156	0.080903362	0.092245283	0.0998554	0.099856231
C2	0.015	0.033380807	0.037945363	0.041008578	0.041008912

Figure 2.21: System Dynamics

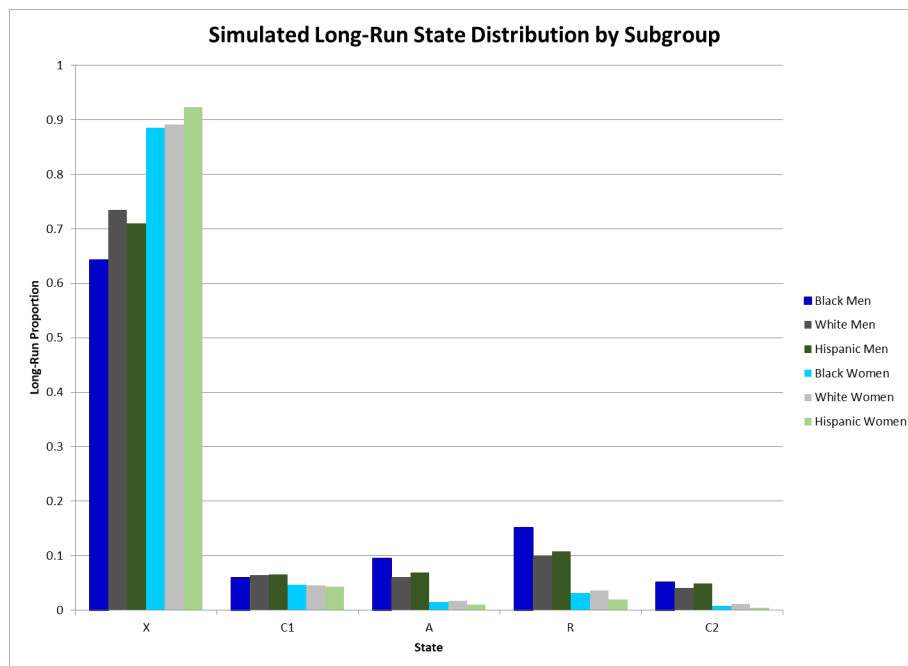


Figure 2.22: Equilibria

2.8.7 Systems Effects

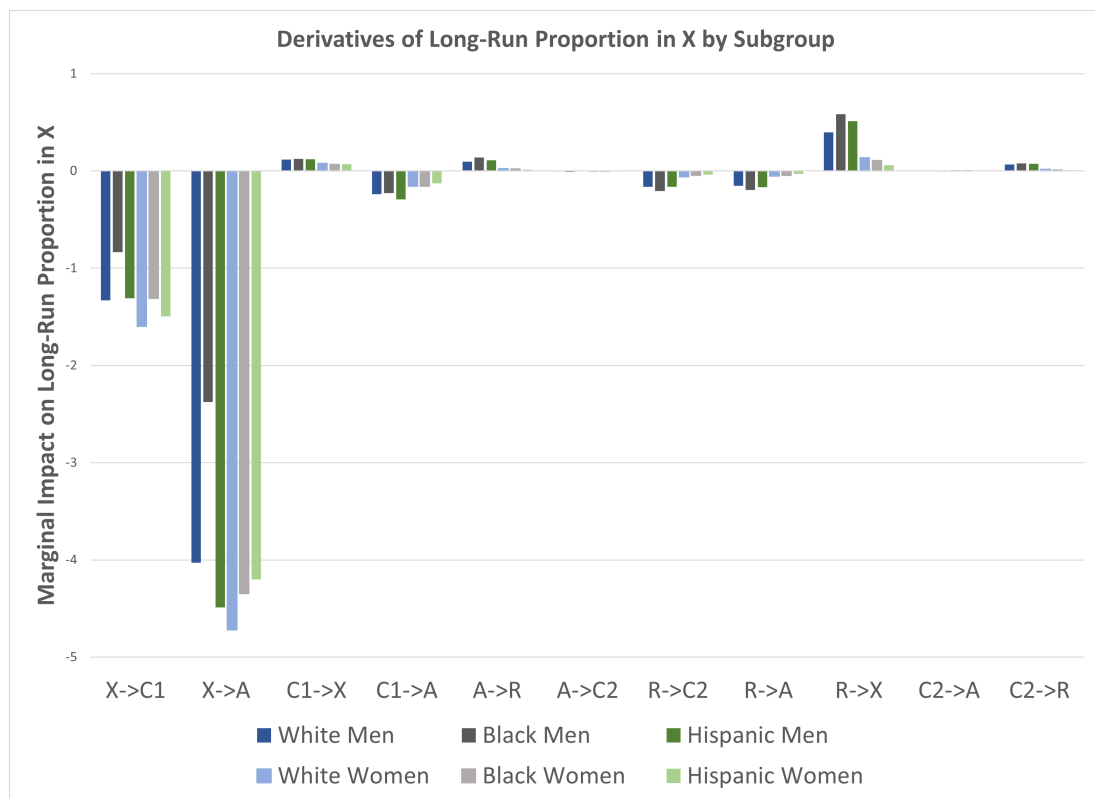


Figure 2.23: Partial Derivatives of X^*

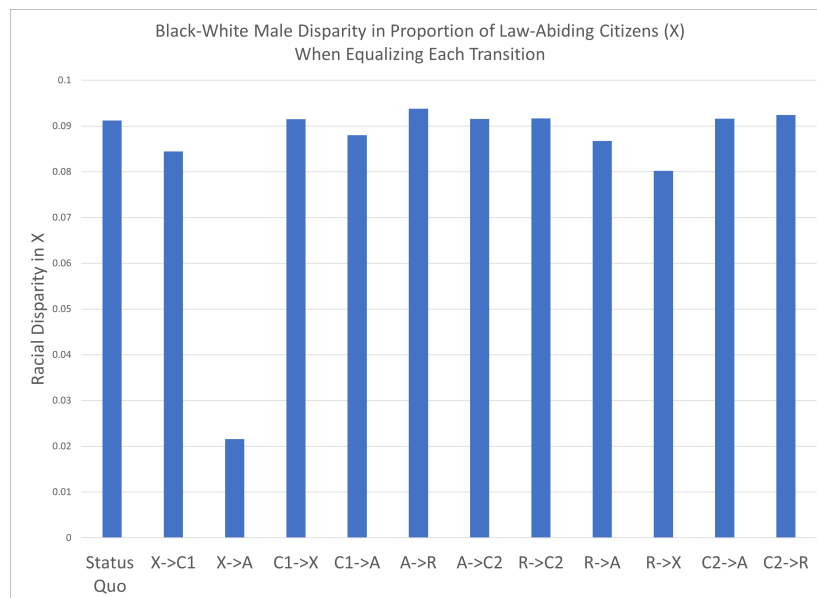


Figure 2.24: Give Black Men White Men's Probabilities

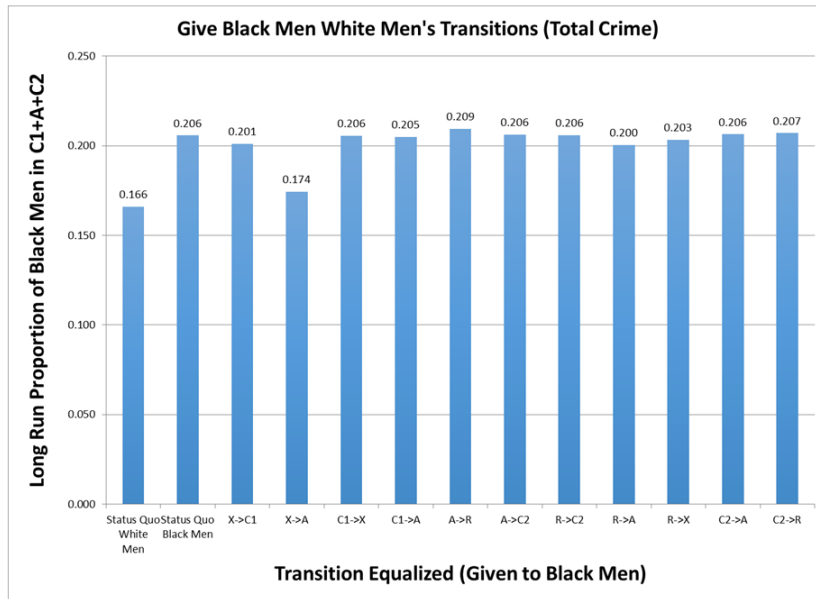


Figure 2.25: Give Black Men White Men's Probabilities

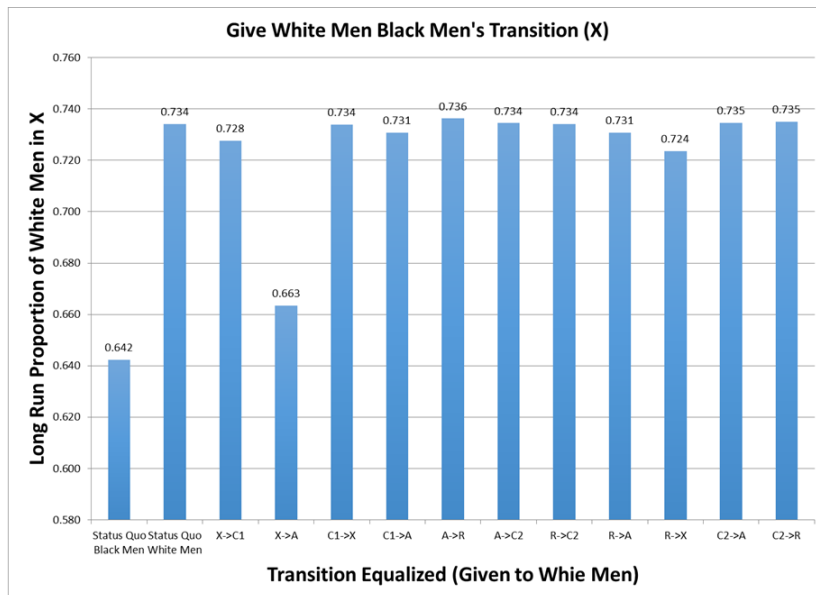


Figure 2.26: Give White Men Black Men's Probabilities

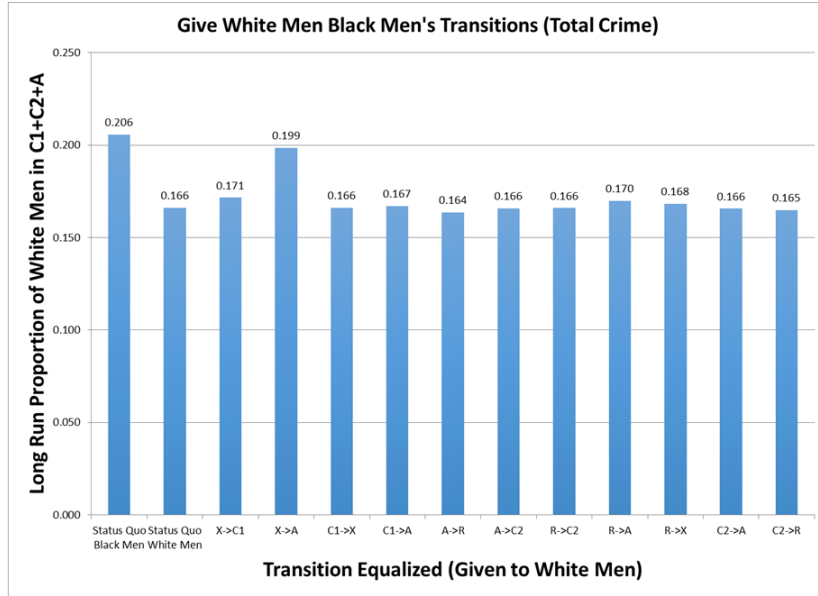


Figure 2.27: Give White Men Black Men's Probabilities

When the outcome of interest is equilibrium X^* , the next most consequential transitions to equalize were $R \rightarrow X$ (rehabilitation), $X \rightarrow C_1$, $R \rightarrow A$, and $C_1 \rightarrow A$ (first arrest), in that order, for reductions of 11.5, 8.9, 5.2, and 4.2 percentage points, respectively. When the outcome of interest is the racial gap in total crime and arrest ($C_1 + A + C_2$), the next most consequential transitions to equalize were the arrest (in the following year) of previously criminally inactive men with an arrest history, $X \rightarrow C_1$, $R \rightarrow A$, $R \rightarrow X$, and $C_1 \rightarrow A$ in that order, for reductions of 16.2, 13.9, 5.5, and 2.5 percentage points, respectively. In both cases, reducing the racial gap in the arrest of criminally active men decreased the racial gap in the law-abiding X^* population when they didn't have an arrest history, but had essentially no effect (and if anything, worsens the racial gap) if they did have an arrest history.

CHAPTER 3

A THEORETICAL ANALYSIS OF BEHAVIORAL INFRACTIONS AND SCHOOL DISCIPLINARY POLICY

In the end, tipping points are a reaffirmation of the potential for change and the power of intelligent action. Look at the world around you. It may seem like an immovable, implacable place. It is not. With the slightest push, in just the right place, it can be tipped.

-Malcolm Gladwell

ABSTRACT

Reducing behavioral infractions through school disciplinary policy has been a major research and policy priority in the U.S. within the last two decades. Much of the best available research seeks to identify the impact of a policy change on behavior, holding constant other conditions. However, nothing is held constant in practice, and generating policy without considering how a broader dynamic system of mechanisms is impacted can lead to unintended consequences. Applied theory can help us explore these considerations and foresee such consequences. I present a theoretical model that is based on salient mechanisms in relevant empirical literatures. With this model I explore how the consequences of disciplinary policy for the dynamics of behavioral infractions depend on certain interrelationships between these mechanisms (punitive vs positive discipline, peer effects, deterrence, attitude transference, attitude towards authority and self concept, etc.) that are usually studied separately and statically (not over time) in the literature. As we will see, in a dynamic setting, these mechanisms can lead to tradeoffs, tipping points, and phase transitions that could help explain the policy resistance and mixed results of school disciplinary practices. They can also illuminate conditions under which appropriate policy changes can lead to disproportionately large changes in the school disciplinary climate. An in-depth bifurcation analysis reveals that these counter-intuitive dynamics are more likely to be present in predominately African-American schools.

3.1 Introduction

Activists and researchers have long coined the phrase “School-to-Prison Pipeline,” to refer to the phenomenon, intentional or not, through which students’ cumulative negative school experiences eventually lead to arrest and mass incarceration [79]. Although there is much disagreement on the specific policies, practices, and mechanisms behind the trend, students’ misbehavior and how schools respond to this misbehavior are often cited as key components of this “pipeline.” Like crime and criminal justice, behavioral infractions and disciplinary practices in schools are characterized by strong disparities across the color line [29], and there are long-standing debates regarding the efficacy of harsh discipline ([33], [30], [43]). While punitive approaches (such as Zero Tolerance)

are based on economic notions of deterrence and Beckerian penalties, other approaches (such as Restorative Justice) are based on psychological research on affective entities, putting less emphasis on deterrence through punishment and more emphasis on building relationships, empathy, and positive feedback for good behavior [7]. It is often argued that a student's sense of self, school connectedness, attitude towards school and authority, sense of disciplinary fairness, and relationship with school officials can be impacted by disciplinary practices([38], [48], [84]) and can strongly influence behavior ([84], [92], [46], [94], [55]). This feedback process between behavior, disciplinary response, and affective entities is inherently dynamic, requiring an analysis of the underlying dynamics of disciplinary policy and misbehavior in schools that should also respect other important mechanisms, such as peer influences.

The purpose of applied theory is rigorously explore the systemic consequences of various mechanisms believed to drive the statistical and causal relationships established by applied research. This is missing in the literature surrounding the school disciplinary policy. As we will see, these mechanisms can lead to tradeoffs, tipping points, and phase transitions that could help explain the policy resistance and mixed results of school disciplinary practices. They can also illuminate conditions under which small policy changes can lead to sweeping, persistent, transformative change in the larger school disciplinary climate. As we will see, these happen to be conditions that may be more likely in predominately African-American schools.

3.1.1 The Contributions

The ultimate goal of this research is to shed light on how the optimal design of disciplinary policy in schools can shape the dynamics of deviant behavior in schools. Static models measuring the causal effect of single interventions are insufficient to capture the complex dynamic behavior that can arise from the feedback processes in school discipline predicted from psychological theory. This approach in this paper allows us to study how

assumptions and mechanisms studied in various fields can interactively lead to various (sometimes unintended) consequences.

There are several substantive contributions of this work. First, I invoke a dynamic discrete choice approach that jointly analyzes several empirically validated mechanisms studied in the literature: the impacts of harsh disciplinary practices and positive rewards on student affective entities; the impact of these affective entities on student behavior; the impact of peers; deterrence; background factors; and random heterogeneity. This approach is designed to consider how developmental trajectories evolve dynamically through a complex of interacting factors which have, up to this point, largely been studied separately in the literature; but which nonetheless produce important, synergistic effects. I consider the never-studied tradeoff between a harsh punishment's deterrent impact on misbehavior, and its potentially reinforcing impact on misbehavior through affective entities. I show conditions under which the same disciplinary policy can lead to different outcomes (multiple equilibria) depending on the initial conditions of the school. I offer a rigorous, theoretical justification for the importance of positive feedback/rewards for positive behavior as both a deterrent and an affect-improving mechanism, which helps explain the success of positive behavioral supports. I also show conditions under which tipping is more likely for a group that is more likely to be harshly punished for the same offense, for a more homogeneous group, and for a group that, due to background conditions, is predisposed to what is considered deviant behavior. Finally, a certain kind of tipping behavior becomes more likely when a student's affect becomes time-sensitive, suggesting a need for empirical estimation of affective trajectories. When a system exhibits tipping points, small policy changes can make a significant difference. Hence, this research also provides a theory of school turnaround with respect to student behavior.

The methodological contribution of this work is to extend the discrete choice approach of Brock and Durlauf [12] with intra-group interactions to a system with inter-group in-

teractions (a school official and interacting students), and to mathematically derive emergent macro-level dynamic systems from agent-based micro-foundations. This approach is readily adapted for a myriad of other problems in social policy governed by feedback loops-particularly in cases in which an individual's preferences are impacted by his/her behavior in the previous time step due to affect evolution (e.g., feedback between effort, math performance, and one's beliefs about math).

The remainder of the paper proceeds as follows. I begin with a review of the relevant literature on school disciplinary policy and various mechanisms believed to be important for students' behavioral decisions, including deterrence, affective considerations, and peers. I also discuss the methodological grounding of this work in the literature on developmental trajectories and discrete choice with social interactions. Second, I present the building blocks of the theoretical model, deriving dynamical systems from the micro-foundational, discrete choice approach of Brock and Durlauf (2001) [12]. These systems will provide a simple view of the inter-agent dynamics between students and school officials that respects social interactions, affective entities, behavior, and school disciplinary policy. To analyze the model, I begin with a simple case (in which there is no deterrent or effect of background characteristics), and show conditions under which trajectories exhibit "tipping," or sensitivity to initial conditions. Next, I present the full model, analyzing the contribution of each mechanism to the existence of tipping. I illustrate two (of many) examples in which changes in disciplinary policy can induce phase transitions-dramatic qualitative changes in the dynamics of deviant behavior. In the following section I demonstrate how the optimal level of punishment for deviant behavior, from the perspective of a school official with perfect information, depends on tradeoffs between the deterrent effect of punishment, the impact of punishment on student affect, and the functional forms of these effects. Finally, I alter the model to show how these results change when students' affective bundles are time-sensitive, with implications for the dynamic stability of racial disparities.

3.2 Literature Review

3.2.1 *Approaches to School Discipline, Deterrence, and Student Affect*

The literature on school disciplinary policy and misbehavior in schools is extremely extensive, spanning multiple fields. George Bear [7] proposes a concise set of best practices from the literature, as well as an explanation of common unintended consequences. He argues that both “authoritarian” approaches that expressly emphasize deterrence through punishments rather than positive rewards for pro-social behavior, and “permissive” approaches that tend to particularly de-emphasize punishment for deviant behavior, have been criticized in the literature. [7] Instead, much of the evidence supports the use of an “authoritative” approach to school discipline that essentially adopts the strengths of both approaches while hedging against the weaknesses. In particular, an authoritative approach still emphasizes deterrence through firm punishment for deviant behavior, but focuses on ensuring punishment is consistent, fair, and not damaging to the student-teacher relationship. Such an approach also puts more emphasis on positively reinforcing pro-social behavior through rewards. Finally, authoritative approaches focus on purposefully improving students’ self-discipline outside of extrinsic rewards and punishments. Self-discipline, Bear argues, requires intrinsic motivation that should be taught explicitly in curriculum, through classroom and extracurricular activities.

The belief that punishment deters the misbehavior of a rational agent dates back to Becker’s work in the economics of crime [8], and this has motivated a punitive approach to school discipline. There is also evidence that exclusionary discipline can improve student achievement among the peers of misbehaving students [45], which is consistent with the academic success of “No-Excuses” Charters [15]. Finally, there is evidence that more punitive disciplinary policies help reduce the Black-White achievement gap [45].

However, punishment is widely believed to have limitations, particularly in that it can reinforce misbehavior through negative effects on the student-teacher relationship, school connectedness, sense of disciplinary fairness, and attitude towards school and authority. As stated by Walker in 1999: “As antisocial behavior increases, school officials may administer even more aversive consequences, increasing a student’s antipathy toward requirements of the student role. Those perceived to be at risk of antisocial conduct, particularly boys and impoverished minority students, are more likely to be excluded and controlled than have their problems addressed in a therapeutic way.” [61]. Similarly, as stated by Margaret Beale Spencer and co-authors: “...Thus, African-American students experiencing negative feedback about the self in school might shut off from this critically important experience. This decision may result in the diverting of psychic energy and attention to their peer group and reorganizing themselves or more closely attaching to a group which provides positive, or at least less threatening feedback. The decision-which could lead to problem behaviors, school drop-out, and delinquency-may be of help in the short term in preserving the self...” [92]. Spencer’s seminal PVEST (Phenomenological Variant of Ecological Systems Theory) framework captures “dynamic processes that occur throughout the lifespan as individuals encounter new risks and protective factors, engage new stressors, establish more expansive coping strategies, and redefine how they view themselves, which affects how others view them.” [87] PVEST suggests that students’ repeated use of reactive coping mechanisms from encounters with school officials and students determines the emergent school identity (their underlying attitudes about the self in the school context), leading to adaptive or maladaptive behaviors to which teachers and students respond with positive or negative feedback.

Recognition of the limitations of punitive discipline, along with the importance of teaching self-discipline, improving empathy, student-teacher relationships, school connectedness, attitudes, and sense of disciplinary fairness has led to a surge in the adoption of

Positive Behavior Supports and Restorative Justice approaches [91]. Positive Behavior Supports seek to decrease punitive responses to deviant behavior, focus on positive reinforcement, increase time for academic engagement, and match the level of intervention resources to the level of behavioral challenges that students present. They generally involve a three-tiered system, with the first tier focused on improving positive relationships on a school-wide level, and the last two tiers focused on a smaller group of chronically deviant students who require help from behavioral specialists. Positive Behavior Supports have demonstrated strong improvements in both student discipline and academic achievement based on experimental evidence [66]. Restorative Justice approaches were first used in the criminal justice system [60], and their use in schools can vary greatly in format. However, the common themes of Restorative Justice include building and restoring relationships between offenders, victims, families, and school officials; victim-centered responses to misbehavior to replace the focus on rule-violation; and empathy. The use of “restorative circles” frequently involves increases in empathy between an offender and a number of his/her peers who may have been affected by the offense. Overall there is some correlational evidence that these approaches improve achievement and pro-social behavior in schools, but the internal validity of most of those studies is questionable [34]. Experiments to improve the evidence is ongoing.

These approaches have been criticized as difficult and expensive to implement, and lacking strong deterrent effects due to softer punitive discipline that could lead to negative peer effects. On the other hand, racial disparities in disciplinary response are believed to be pervasive at schools under more punitive regimes than schools focusing on Positive Behavior Supports or Restorative Justice [29], [34], as well as worsened school connectedness, student-teacher relationships, attitude towards authority and sense of disciplinary fairness (as previously stated).

Clearly, there could be several important tradeoffs to consider in the mechanisms

behind disciplinary policy and behavioral infractions in schools. It is important to systematically simulate how students' behavioral outcomes depend on this complex system of factors. This paper does just that, by building a mathematical model that includes the impacts of disciplinary responses on a bundle of affective entities (school connectedness, student-school official relationships, sense of disciplinary fairness, attitude towards school and authority, and self-concept) referred to as “affect” hereafter, their deterrent effects, the impact of affect on behavior, peer effects, background effects (differential propensities to commit, or be accused of, a behavioral infraction on the basis of a myriad of person-specific variables, potentially including socioeconomic variables and parental upbringing), and random heterogeneity.

3.2.2 Dynamics of Deviant Behavior

The dynamic nature of deviant behavior has received a lot of attention in the criminology literature in the form of growth mixture models. This is in part due to the recognition that a static view of crime outcomes in adulthood fails to account for developmental trajectories that precede them. While a static view considers how best to intervene on criminal behavior, a dynamic perspective also considers “when” to intervene. For example, Nagin and coauthors found that the onset of aggressive behavior early in life significantly predicts offending during adolescence, which suggests that earlier interventions may have serious consequences for long-run crime. This is related to other literature by Nagin and coauthors suggesting that past criminal behavior predicts future behavior not only through time-invariant personal characteristics, but also through the experiences related to the past behavior itself [67]. Some of these models have been applied to the school context [72], showing distinct developmental trajectories in deviance and/or achievement for students endowed with different background factors, different exposure to deviant peers, and different teacher-student relationship qualities [88]. However, to date, no mathematical model of student behavior has been proposed that captures how the

theorized underlying dynamics behind these mechanisms can lead to distinct emergent developmental trajectories. In this paper, beginning with microfoundations specifying the decision rules of students and a school official, I will derive a dynamical system of differential equations that is flexible enough to explain a wide range of behavioral trajectories in terms of key parameters of interest in the literature. In this framework, the developmental trajectories are solutions to the theorized system.

3.2.3 Peer Effects and Social Interactions in Misbehavior

Social interactions and peer influences are widely cited as drivers of deviant behavior in both economics and psychology [36]. Broadly speaking, adolescent students are theorized and empirically shown to be sensitive to the behavioral decisions of their peers, although to different extents depending on the type of misbehavior. In adolescence, cognitive systems associated with higher order critical thinking are far less developed than limbic systems, which focus on immediate and social rewards as opposed to long-term utility maximization [52]. Behavioral economics suggests that adolescents have problems with delayed gratification due to hyperbolic discounting, and care a great deal about the immediate benefits of fitting in with peer groups [35].

Distinct from the sensitivity of behavior to peer behavior is the sensitivity of behavior to peer attitudes. Megens and coauthors ([64]) investigate each of these, concluding that previous literature has underestimated the predictive power of peer attitudes specifically, for misbehavior. Theoretically, a student can be induced to violate school rules based on how his/her peers “feel:” for instance, through the recognition that his/her peers have negative attitudes towards a teacher, even when the peers are not openly misbehaving. Finally, a student may be induced to violate school rules due to his/her own affect, which can be impacted by the affect of peers. This process, called ‘attitude transference,’ has been studied in the criminology literature. In short, there are many different kinds of

peer-based mechanisms that can impact misbehavior. These distinctions will be made explicit in the model presented by this paper.

To mathematically operationalize these kinds of interactions in the dynamics of behavioral decisions, I will employ what are called discrete choice models with social interactions. The literature on these models has been heavily impacted by the seminal work of Brock and Durlauf [12]. Agent-based decisions are modeled using interdependent utility functions, in that outside of a private deterministic utility and random heterogeneity, an agent's likelihood to make a choice depends on the likelihood of other agents making that same choice. These models predict the existence of multiple equilibria (that is, the same group of people could end up with very different average outcomes depending on the initial choice profile of the group) for the same set of parameter values when agents are sufficiently sensitive to the choices of their peers. Consider, for example, the dynamics of a standing ovation. If initially a critical proportion of an audience stands following a performance, most of the others may then choose to stand. However, below that threshold most people may choose to remain seated. What is driving the existence of multiple equilibria depending on initial conditions in this case is the fact that the individuals in the audience, to some extent, care what other individuals are doing. More to the point here, the fact that students care how their peers behave could lead to classrooms or schools in which the long run proportion of misbehaving students depends nonlinearly on the initial proportion of misbehaving students. In all of these cases, weak social influence makes multiple equilibria less likely (e.g., if social influence hardly matters then the long run outcome should not depend much on the initial activity of others), while strong social influence makes such cases more likely.

This discrete choice with social interactions literature has led to some important papers on the implications of social interactions for multiple equilibria in deviant behavior specifically. David Card used the Add Health dataset to estimate a model of deviant

behavior that accounts for the risky behavior of friends [28]. He finds significant peer effects for sexual initiation, smoking, marijuana use, and truancy. Among best friend pairs he finds that multiple equilibria cases are rare, as reciprocated best-friendships are rare. However, Soetevent and Kooreman [86] estimate a discrete choice model focused on deviant high school teen behavior, finding that peer effects are strong enough to be consistent with multiple equilibria for deviant behaviors most closely related to school, such as truancy. In an unpublished manuscript, Bisin and coauthors [9] find that peer effects related to deviant behaviors in school are strong enough to be consistent with multiple equilibria.

Peer effects and social interactions can have important policy implications in the context of deviant behavior, because they can lead to “tipping points-” levers in a system which, when targeted, can produce dramatic effects. This is true even when friendships are not reciprocated. For example, a disciplinary response that changes the behavior of a particularly influential student can have ripple effects, affecting influenced students’ behavior, who may then influence other students’ behavior, and so on. In this paper I will investigate what these peer influence considerations suggest about optimal disciplinary policy, within a system that also considers deterrence, student affect, background characteristics and random heterogeneity.

3.3 Model

In this section I derive the full dynamic model that is of central interest. The end result will be a system of differential equations describing how the average choice of students coevolves with the average affect (representing a bundle of affective attributes as described in the Introduction). The school official, having full information, then chooses a disciplinary policy that maximizes long-run pro-social behavior given that system (hence solving a costless optimal control problem).

3.3.1 Behavioral Decisions

Consider a dynamic model with N students, in which each student, indexed by i , decides, in each time period t , on deviant ($y_{it} = -1$) or pro-social ($y_{it} = 1$) behavior. Following Brock and Durlauf, the preferences of each student are defined by a utility function V that can be linearly separated into a deterministic private utility u , a social utility function S that depends on the choices y_{-it} and affect A_{-it} of peers at time t , and a random utility function ϵ , assumed to be independent across individuals and decisions, and observable only to the individual in question. Suppose the private utility for the choice of student i at time t depends on the affect of the student $A_{it} \in [A_*, A^*]$ (where $A_*, A^* \in \mathbb{R}$), the disciplinary response $T(y_{it}) \in [-1, 1]$, and common, time-invariant background characteristics c . In each period t , each student i solves the decision problem

$$\max_{y_{it} \in \{1, -1\}} u(y_{it}, A_{it}, T(y_{it}), c) + S(y_{it}, y_{-it}, A_{-it}) + \epsilon_i(y_{it}t)$$

Next, let m_{ijt}^e and A_{ijt}^e be the expectations of student $j \neq i$'s choice and affect from the perspective of student i at time t , respectively. Assume the social utility function takes the form

$$S(y_{it}, y_{-it}, A_{-it}) = \frac{Jy_{it}\bar{m}_{it}^e + By_{it}\bar{A}_{it}^e}{2}$$

where $\bar{m}_{it}^e = (N-1)^{-1} \sum_{j \neq i} m_{ijt}^e$ is student i 's belief of the average choice level of peers at time t and similarly, $\bar{A}_{it}^e$ is student i 's belief of the average affect level of peers at time t . Notice that

$$[S(1, 1, A_{-it}) - S(-1, 1, A_{-it})] - [S(1, -1, A_{-it}) - S(-1, -1, A_{-it})] = 2J.$$

Similarly, let $a_1, a_2 \in \mathbb{R}$ be affect levels such that $a_1 > a_2$. Then

$$[S(1, y_{-it}, a_1) - S(-1, y_{-it}, a_1)] - [S(1, y_{-it}, a_2) - S(-1, y_{-it}, a_2)] = 2B(a_1 - a_2).$$

Hence if $J > 0$ and $B > 0$, the gains from pro-social behavior when others have the same behavior, are larger than the gains from pro-social behavior when others are deviant. Similarly the gains from pro-social behavior when others have more positive affect are larger than the gains when others have a less positive affect. I assume $J > 0$, so that behaviors are complementary, and $B > 0$, implying peer affect is complementary with behavior, as consistent with the literature.

Suppose the school official rewards pro-social behavior and punishes deviant behavior so that $T(1) = \mu$ and $T(-1) = -\delta$, where $\mu, \delta \in [0, 1]$. For mathematical convenience I implicitly assume that

$$u(1, A_{it}, T(1), c) = bA_{it} + 2a_1\mu + c$$

and that

$$u(-1, A_{it}, T(-1), c) = -(bA_{it} + 2a_{-1}\delta + c).$$

Then the private utility difference is

$$u(1, A_{it}, T(1), c) - u(-1, A_{it}, T(-1), c) = 2(bA_{it} + a_1\mu + a_{-1}\delta + c).$$

The 2 is just a normalization constant. In this case, parameter b is the effect of an individual's affect on choice (which is theorized and empirically estimated to be nonnegative: [62], [92]). The sum $a_1\mu + a_{-1}\delta$ captures net deterrence of the disciplinary policy (also believed to be nonnegative), since a_1 represents the impact of rewards on pro-social behavior and a_{-1} represents the impact of punishment on pro-social behavior. Finally, c

is the (negative, positive, or zero-valued) impact of common background characteristics on any student's choice.

Finally, assume $\epsilon_{it}(1)$ and $\epsilon_{it}(-1)$ have the Type I extreme value distribution, so that $\epsilon_{it}(-1) - \epsilon_{it}(1)$ is logistically distributed, as is typical in the discrete choice literature. A student i chooses pro-social behavior ($y_{it} = 1$) if the gains from pro-social behavior are larger than those from deviant behavior. The expected choice level of student i is then calculated using the logistic distribution:

$$\begin{aligned} Prob[y_{it} = 1] &= Prob[V(1) - V(-1) \geq 0] = Prob[\epsilon(-1) - \epsilon(1) \\ &\leq 2(J\bar{m}_{it}^e + B\bar{A}_{it}^e + bA_{it} + a_1\mu + a_{-1}\delta + c)] \\ &= \frac{1}{1 + \exp(-\beta[2(J\bar{m}_{it}^e + B\bar{A}_{it}^e + bA_{it} + a_1\mu + a_{-1}\delta + c)])} \end{aligned}$$

where β characterizes the thinness of the tails of the logistic distribution. The joint choice probability across all agents factors since the random utilities are independent across agents.

In this paper I focus on the solution concept known as symmetric Bayes Nash Equilibrium. Invoking symmetry we impose $E[y_{jt}] = E[y_{it}] = m_t$ and $E[A_{it}] = E[A_{jt}] = A_t$ for all $i \neq j$ in equilibrium. That is, the unconditional average choice level (and average affect level) of all agents is the same for each agent in a given period t . We refer to these simply as m_t and A_t respectively. We also impose rational expectations, so that $\bar{m}_{it}^e = m_{t-1}$ and $\bar{A}_{it}^e = A_{t-1}$. Hence students use the last period's average choice/affect as the basis for formation of expectations for the current period's average choice/affect.

Of course, the unconditional expected choice at time t will depend on the conditional expectations given the current personal affect level A_{it} . In general,

$$E[y_{it+1}|A_{it+1}] = \tanh(\beta[Jm_t + BA_t + bA_{it+1} + a_1\mu + a_{-1}\delta + c]).$$

Therefore, assuming the common density f of A_{it+1} , the unconditional expected choice level is

$$m_{t+1} = \int_{A_*}^{A^*} \tanh(\beta[Jm_t + BA_t + bA_{it+1} + a_1\mu + a_{-1}\delta + c])f(A_{it+1})dA.$$

This equation captures the evolution of the average choice level as a function of peer effects in behavior J , in affect B , of personal affect b , deterrence $a_1\mu + a_{-1}\delta$, and background characteristics c . The deterrent effect has an explicit dependence on punishments for deviant behavior δ and rewards for pro-social behavior μ .

Note that each student is forward-looking in that he/she considers the punishment or reward for his/her behavioral decision. However, in this model students do not consider how their future behavioral trajectories and attitudes will be shaped by the disciplinary response. In this model I take the position that young adolescent students tend not to have such a grasp on introspection or delayed gratification, which seems to be consistent with the literature [90]. Next, we explore how the evolution of affect depends on the disciplinary response (and hence the previous choice).

3.3.2 *The Dynamics of Affect*

Consider the evolution of a students' personal affect. In this model the school official always rewards pro-social behavior (with μ) and punishes deviant behavior (with $-\delta$). Assume α_1 is the impact of rewards on attitude towards school, and α_{-1} is the [negative] impact of punishment on attitude towards school. At any given time, the affect A_{it} of

student i at time t will depend on the student's experienced disciplinary sanction in the previous time period.

There is reason to believe that a student's present affect can also depend directly on his/her past affect (not just indirectly, through previous decisions and sanctions). Recall that I have defined affect as a bundle of entities: attitude towards school, school connectedness, self-concept/identity, attitude towards authority, student relationship with school officials, and beliefs that school discipline is fair. It could be that these entities become more stable over time due to a kind of psychological "stickiness." That is, for example, a student's relationship with a school official may be more malleable during the first few months of interaction than after years of interaction. The school official who implements fair discipline after treating a student unfairly for years may not change the student's sense that discipline is fair as much as the school official who implements fair discipline after treating the student unfairly only once. The literature suggests that a student's self-concept and identity become less malleable with age as well. Therefore, in the full model I propose a cumulative notion of affect that depends explicitly on previous affect levels.

Finally, there is a tradition of literature on attitude transference that suggests these entities can be spread from peers. For example, the relationships between friends and a school official can affect a person's relationship with that school official, and friends' attitudes towards school or authority could potentially impact a student's attitudes.

Suppose that in each period, the updated affect bundle depends on the previous sanctions, the previous affect levels, and the current affect of peers. Then student i 's conditional expected affect at time $t + 1$ can be modeled as

$$E[A_{it+1}|A_1, \dots, A_t]$$

$$= \frac{1}{(t+1)^p} (\alpha_1 \mu \text{Prob}[y_{it} = 1] - \alpha_{-1} \delta \text{Prob}[y_{it} = -1] + K A_{it+1}^e) + d \sum_{s=1}^t \frac{1}{s^p} A_{is}$$

where K is assumed to be the nonnegative impact of the believed average peer affect on the affect of student i at time $t+1$. The parameter d captures the extent to which the updating process is dependent on the affect level accumulated up to the previous time t . Finally, there is an implicit assumption about the way the updating process discounts the past. For example, if $p = 2$, then the total cumulative affect level behaves as a convergent series as time goes on, with later updates carrying less weight than past updates. If $p \leq 1$, then the series diverges over time (a standard result), since later updates still carry less weight than past updates, but not by an amount that is sufficiently small to allow the series to converge.

Once again, we can impose rational expectations and symmetry, so that $\bar{m}_{it}^e = m_{t-1}$ and $\bar{A}_{it}^e = A_{t-1}$, where $m_t = E[y_{it}]$ and $A_t = E[A_{it}]$ is the same for all students. Let g be the joint probability density for realized history $A_1, A_2, \dots, A_t$. Note that this density cannot be factored since the sequence $A_1, A_2, \dots, A_t$ is not a sequence of independent events. Then the unconditional expected affect level is a t -dimensional integral:

$$A_{t+1} = \int_{A_*}^{A^*} \dots \int_{A_*}^{A^*} \left[\frac{1}{(t+1)^p} \left(\frac{\alpha_1 \mu \exp(L) - \alpha_{-1} \delta \exp(-L)}{\exp(L) + \exp(-L)} + K A_{it} \right) + d \sum_{s=1}^t \frac{1}{s^p} A_{is} \right] g(A_1, \dots, A_t) dA.$$

where $L = \beta(Jm_t + BA_t + bA_{it} + a_1\mu + a_{-1}\delta + c)$.

The full system is so complex that it requires simulation to fully analyze. Until the

final sections of this paper, I will assume $d = 0$ and $p = 0$ for simplicity, and that personal affect at time $t + 1$ does not depend directly on the history of affect levels up to time t . I will also assume throughout most of the paper that $A_{it} \in \{-A^*, A_*\}$, where $A^* = \mu\alpha_1 + KA_t$ and $A_* = -\delta\alpha_{-1} + KA_t$, so that at any given time, a student's affect can only take on one of two possible values.

Notice that

$$E[y_{it+1}] = 1\text{prob}(y_{it+1} = 1) - (1 - \text{prob}(y_{it+1} = 1)) \implies \frac{E[y_{it+1}] + 1}{2} = \text{prob}(y_{it+1} = 1).$$

Finally, recall that a pro-social choice results in personal affect A^* in the next period, so that $A_{it} = A^*$ implies $y_{it-1} = 1$, and $A_{it} = A_*$ implies $y_{it-1} = -1$. Hence for most of this paper I use the description

$$A_{t+1} = E[A_{it+1}|A_{it} = A^*]\text{prob}(A_{it} = A^*) + E[A_{it+1}|A_{it} = A_*]\text{prob}(A_{it} = A_*)$$

$$= E[A_{it+1}|A_{it} = A^*]\text{prob}(y_{it-1} = 1) + E[A_{it+1}|A_{it} = A_*]\text{prob}(y_{it-1} = -1)$$

$$\begin{aligned} &= \left(\frac{\alpha_1\mu \exp(L^*) - \alpha_{-1}\delta \exp(-L^*)}{\exp(L^*) + \exp(-L^*)} \right) \left(\frac{1 + m_{t-1}}{2} \right) \\ &+ \left(\frac{\alpha_1\mu \exp(L_*) - \alpha_{-1}\delta \exp(-L_*)}{\exp(L_*) + \exp(-L_*)} \right) \left(1 - \frac{1 + m_{t-1}}{2} \right) + KA_t \end{aligned}$$

where $L^* := \beta[Jm_t + BA_t + bA^* + a_1\mu + a_{-1}\delta + c]$ and $L_* := \beta[Jm_t + BA_t + bA_* + a_1\mu + a_{-1}\delta + c]$.

This captures the evolution of student affect. Our description of the evolution of

choice is then

$$m_{t+1} = \tanh(L^*) \left(\frac{(m_t + 1)}{2} \right) + \tanh(L_*) \left(1 - \frac{m_t + 1}{2} \right).$$

Here we have again conditioned on the two possible values of A_{it+1} and used the fact that the probability that $A_{it+1} = A^*$ is the same as the probability that $y_{it} = 1$, since a student receives affect level A^* if and only if he/she behaved in a pro-social way in the previous period.

Next, we convert the system of difference equations describing student choice and affect evolution into a system of differential equations (see Appendix). The definition of L^* now changes to its continuous analogue (A_t and m_t are replaced by A and m respectively, both implicit functions of continuous time t). The system dynamics are as follows:

$$\dot{A} = \phi(L^*) \left(\frac{1+m}{2} \right) + \phi(L_*) \left(1 - \frac{1+m}{2} \right) + (K-1)A \quad (3.2a)$$

$$\dot{m} = \tanh(L^*) \left(\frac{(m+1)}{2} \right) + \tanh(L_*) \left(1 - \left(\frac{m+1}{2} \right) \right) - m \quad (3.2b)$$

$$(\text{where } \phi(L^*) = \frac{\alpha_1 \mu \exp(L^*) - \alpha_{-1} \delta \exp(-L^*)}{\exp(L^*) + \exp(-L^*)} \text{ and } \phi(L_*) = \frac{\alpha_1 \mu \exp(L_*) - \alpha_{-1} \delta \exp(-L_*)}{\exp(L_*) + \exp(-L_*)})$$

From agent-based micro-foundations, we have derived a dynamical system that can capture a wide range of behavioral trajectories depending on various mechanisms studied in the literature.

3.3.3 The School Official

Importantly, the above system depends on the disciplinary policy a school official establishes at time $t = 0$, which is a rule specifying the extent of rewards for pro-social behavior and the extent of punishments for deviant behavior. For simplicity, this paper will presume a “first-best” scenario in which the school official has reliable information

on all of the parameters captured in the model, and chooses a disciplinary policy that minimizes long-run average deviant behavior in the school. Hence, given N students, the objective of the school official will be

$$\max_{\mu, \delta \in [0,1]} \lim_{t \rightarrow \infty} m$$

such that

$$\dot{A} = \phi(L^*) \left(\frac{1+m}{2} \right) + \phi(L_*) \left(1 - \frac{1+m}{2} \right) + (K-1)A \quad (3.3a)$$

$$\dot{m} = \tanh(L^*) \left(\frac{(m+1)}{2} \right) + \tanh(L_*) \left(1 - \left(\frac{m+1}{2} \right) \right) - m \quad (3.3b)$$

Hence the school official with full information chooses a disciplinary policy that solves an optimal control problem. We will be taking on the perspective of an ideal, fully informed school official in the analyses that follow.

Notice that increasing μ always increases m (decreasing deviant behavior) in this model. Rewards enter the dynamics through two pathways. First, increasing μ has a direct positive impact on student affect through α_1 , the effect of rewards on affect. The literature indicates that $\alpha_1 > 0$. Improving students' affect levels then induces an improvement in average student choice level through b , the effect of student affect on behavioral decisions, and through B , the effect of average peer affect on behavioral decisions. Second, μ has a direct positive impact on student choice levels through deterrence. Specifically, $a_1\mu + a_{-1}\delta$ captures the net deterrent effect of a disciplinary policy, and net deterrence is monotonically increasing in rewards μ .

This has to be taken seriously as a factor in favor of disciplinary policies that stress positive reinforcement for pro-social behavior, such as positive behavioral supports. Although punitive measures can involve tradeoffs in the production of pro-social behavior,

rewards for pro-social behavior positively impact student behavior through both improvements in affect and deterrence. It is therefore optimal, in this model with no cost, for the school official to maximally reward pro-social behavior. In equilibrium, the school official chooses $\mu = 1$.

But to what extent should the school official punish deviant behavior? This is a key policy question, which this model is designed to address while taking into account a myriad of mechanisms studied in the literature: the impact of background characteristics c , peer effects of behavior J and affect B , attitude transference K , individual heterogeneity β , sensitivity of behavioral decisions to personal affect b , the impact of rewards and punishment on affect α_1 and α_{-1} , and the net deterrent effect $a_1\mu + a_{-1}\delta$. These mechanisms can lead to some arguably surprising implications for the consequences of both harsher and softer disciplinary practices.

3.4 Simplification: No Deterrence or Background effect, Symmetric Punishment and Rewards

To understand whether the school official should increase or decrease punishment for deviant behavior, first consider how complex the dynamics of deviant behavior can be even in the absence of a deterrent or background effect. Suppose $a_1 = a_{-1} = 0$ so that the deterrent effect is zero, and $c = 0$ so that there is no impact of background on decisions. Suppose further that $\mu = \delta = 1$ so that rewards and punishments are maximized, and that $\alpha_1 = \alpha_{-1} = 1$ so that the impact of rewards and punishment is symmetric. Suppose further that the attitude transference parameter $K = 0$.

The behavior of this simplified system is as follows. If the importance of peer affect and behavior is sufficiently small compared with random heterogeneity (that is, if $|\beta B + \beta J| < 1$) then there is one unique equilibrium, at $(0, 0)$. The equilibrium is a glob-

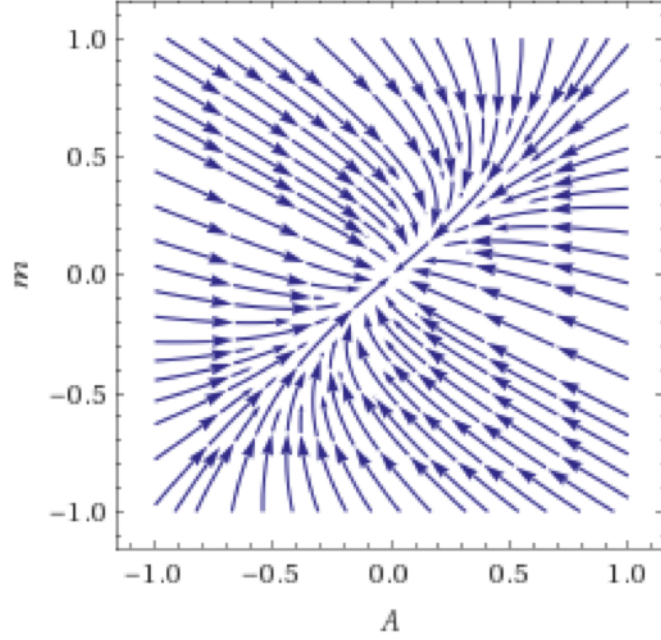


Figure 3.1: Weak Peer Effects

ally stable node, meaning all trajectories lead to it over time (see Figure 3.1). Individual heterogeneity ϵ is very important relative to the total effects of peer affect and behavior, in driving deviant behavior. If $|\beta B + \beta J| > 1$, the total effect of peer affect and behavior is relatively important compared with individual heterogeneity ϵ . The system bifurcates into three equilibria: a stable high-deviance equilibrium $(-\tilde{A}, -\tilde{A})$, a stable low-deviance equilibrium $(\tilde{A}, \tilde{A})$, and a saddle point equilibrium at the origin $(0,0)$. The level of the high/low deviance equilibrium depends on $|\beta B + \beta J|$. As $\beta B + \beta J \rightarrow \infty$, $\tilde{A} \rightarrow 1$.

Importantly, the ultimate destination of any behavioral trajectory is not completely determined solely by the initial average level of deviance in a school, nor by the initial average affect in a school. Consider three sub-cases. First, if $J > B$, (as in Figure 3.2), the system is primarily sensitive to the initial average behavioral decision of students. However, beginning from slightly deviant levels, a school with a sufficiently positive average affect can be “tipped” into the low-deviance equilibrium. Similarly, beginning from average behavior that is slightly pro-social, a school with a sufficiently negative average affect can be “tipped” into the low-deviance equilibrium. If $J \gg B$, “tipping” is deter-

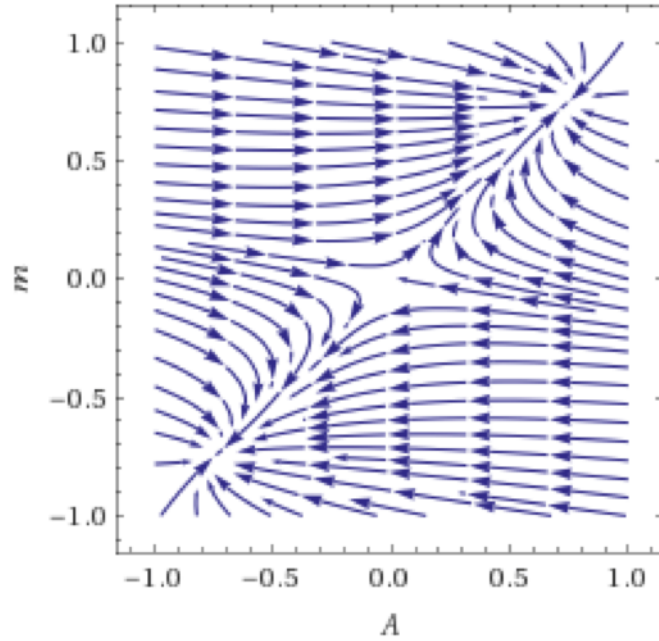


Figure 3.2: Strong sensitivity to peer behavior, weak sensitivity to peer affect:

mined almost entirely by initial choice level rather than initial affect.

Second, if B is close to J , the system is sensitive to both initial average behavioral decisions and initial average affect. When $B = J$ (as in Figure 3.3), the system is still slightly more sensitive to initial behavioral decisions, and for some value of $B > J$ the system is exactly equally sensitive to initial behavioral decisions and attitudes toward school. Hence beginning with a negative affect can be salvagable if the school begins with an equally prosocial level of behavior, and beginning with a deviant level of behavior can be salvagable if the school begins with an equally positive affect.

Finally, if B much larger than J (as in Figure 3.4), the system is primarily sensitive to initial average attitude towards school (whether one begins to the left or right of the origin) than initial behavior (whether one begins above or below the origin). Symmetric to the $J > B$ case, however, beginning with a slightly negative attitude towards school is salvageable if the school begins at a sufficiently prosocial level; and beginning with a slightly positive attitude towards school can still result in a trajectory toward the high-

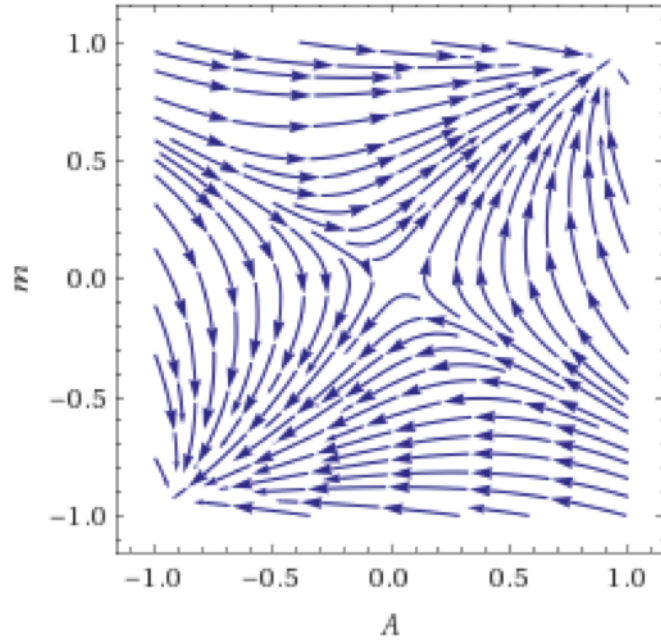


Figure 3.3: Equal sensitivity to peer affect and peer behavior

deviance equilibrium if the school begins at a sufficiently deviant level.

Altogether, tipping behavior is more likely in schools and classrooms with students who are sensitive to peer behavior, who are sensitive to peer affect, and who are more homogeneous.

Recall that we have assumed a certain symmetry in the net impact of rewards and punishment on affect: that is, $\mu\alpha_1 = \delta\alpha_{-1} = 1$. It turns out that this symmetry is causing the unique equilibrium to lie at $(0,0)$. Asymmetry in these parameters shifts the level of the unique equilibrium. For example, if $1 = \mu\alpha_1 > \delta\alpha_{-1}$, then the unique equilibrium will shift in the positive direction along both axes (See Appendix).

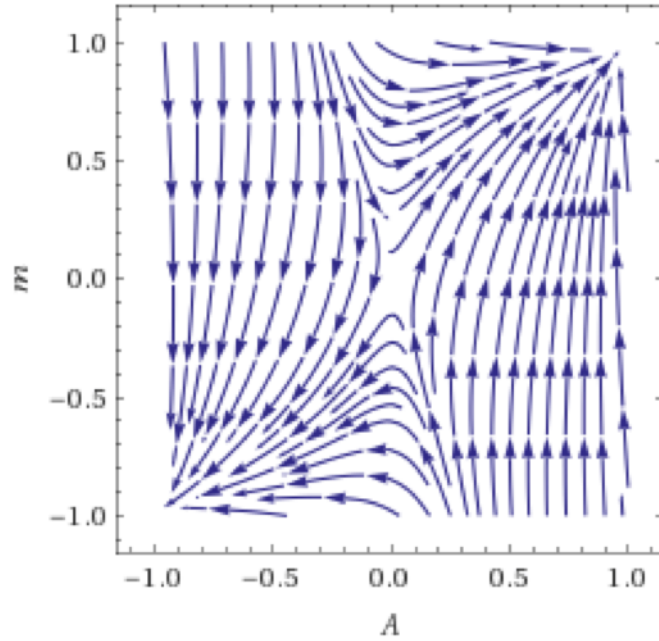


Figure 3.4: Strong sensitivity to peer affect, weak sensitivity to peer behavior

3.5 Full Set of Conditions for Multiple Equilibria and Tipping Points

We have already seen that tipping behavior is more likely when $|\beta(J+B)|$ is large—that is, when students’ decisions are sufficiently sensitive to peer behavior and affect, as opposed to random heterogeneity. Hence tipping is more likely for more homogeneous groups, groups for whom peer affect is extremely important, and groups with strong social complementarities. It turns out that groups with strong attitude transference K in which decisions are sensitive to personal affect (large b) are also more likely to exhibit tipping behavior. A weak deterrent effect makes tipping more likely, and background effects may or may not make tipping more likely.

In this section we consider the full model described at the end of section 3, exploring the roles other mechanisms play in generating multiple equilibria and tipping behavior.

3.5.1 *The Role of Personal Affect and Attitude Transference*

We have seen how the relative sizes of J and B , which represent the sensitivity of behavior to peer behavior and peer affect, respectively, help determine the location of tipping points. In this model, attitude transference K from peers to a student only affects the student's decision if that decision is sensitive to the student's personal affect (captured by b). Hence a larger K increases the likelihood of multiple equilibria if $b > 0$, and a larger b exacerbates the multiple equilibria effects of K .

In the absence of attitude transference K , a large or small b does not affect the multiplicity of equilibria. Importantly, in the current model b only represents the impact of a student's current affect on his/her choice. Recall that we have allowed $A_{it} \in \{A_*, A^*\}$ so that it only takes one of two values. Let $A_* = -1$ and $A^* = 1$. Essentially, the model behaves as a Markov chain with states $A_{it} = 1$ and $A_{it} = -1$, and in the long run, students are expected to spend as much time in one state as the other. Hence, regardless of the size of b , the number of equilibria in the system as determined by the other parameters would be preserved. However, a large b would suggest that a student who began with a negative affect would remain in that state for a long period of time before stochastically switching into the $A_{it} = 1$ state. Essentially, b has the effect of increasing the severity of bottlenecks in the system-severe time delays to reaching the equilibrium.

See the plots in Figures 3.5. The first column (top and bottom left) consists of $b = 0.1$ vs $b = 2$ when there is a single equilibrium ($K = 0, B = 0.2, J = 0.2$). The second column consists of $b = 0.1$ vs $b = 2$ when there are multiple eq ($K = 0, B = 1, J = 1$). Together these plots show that b doesn't affect the number of equilibria when $K = 0$, but instead only affects bottlenecks. Finally, the last column consists of $b = 0.1$ vs $b = 0.8$ when $B = 0.4, J = 0.4, K = 0.1$. These plots suggest that a larger b (impact of personal affect on behavioral decisions) can induce multiple equilibria through attitude transference K .

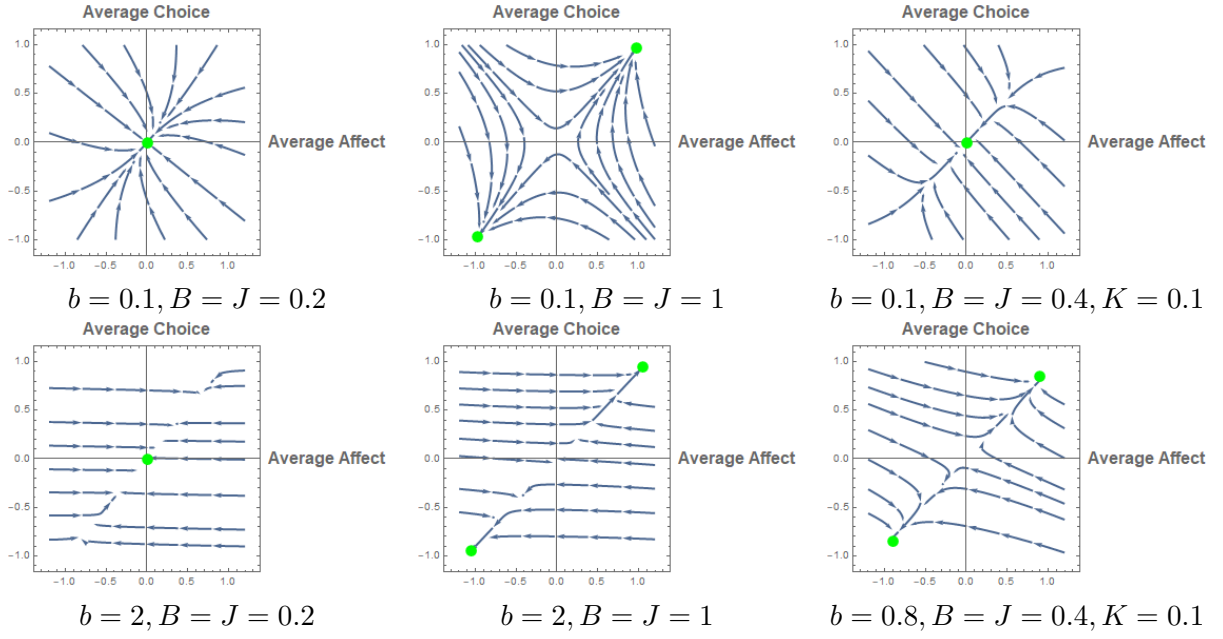


Figure 3.5: Personal Affect and Attitude Transference

3.5.2 The Role of Deterrence and Background Characteristics

Notice the net deterrent effect $a_1\mu + a_{-1}\delta$ and the effect c of common background characteristics on behavioral decisions enter the tanh function as a constant sum. Together, these have the same effect as the private utility parameter h from Brock and Durlauf [12]. There is some critical threshold H (which depends on $\beta, J, B, K, b, \alpha_1, \alpha_{-1}, \mu$, and δ) such that $|a_1\mu + a_{-1}\delta + c| \leq H$ implies multiple equilibria and tipping behavior, and $|a_1\mu + a_{-1}\delta + c| > H$ leads to a unique equilibrium.

Note also by definition that $a_{-1} > 0$ is the reductive effect of punishment for deviant behavior on a student's utility for pro-social behavior, and that $a_1 > 0$ is the positive impact of rewards for pro-social behavior on a student's utility for pro-social behavior. Hence the net deterrent effect $a_1\mu + a_{-1}\delta$ is always positive. However the effect of background characteristics c could be positive, negative, or zero. Tipping is less likely for people who are both easily deterred $a_1\mu + a_{-1}\delta \gg 0$ and are individually predisposed to pro-social behavior ($c \gg 0$). It is also less likely for people who are both heavily predisposed to deviant behavior ($c \ll 0$) and very resistant to deterrence ($a_1\mu + a_{-1}\delta \approx 0$).

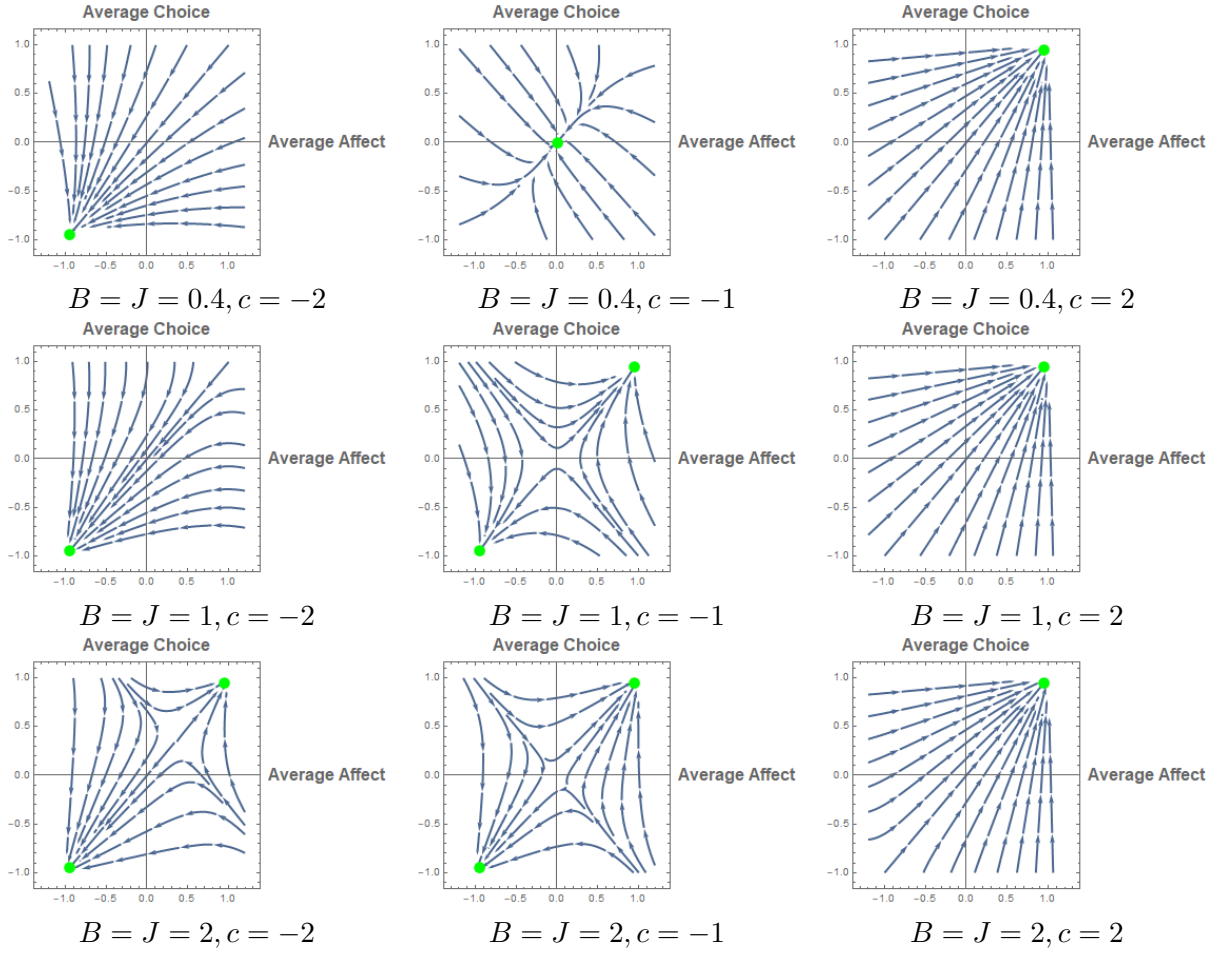


Figure 3.6: Deterrence and Background Effects

Tipping is more likely for people whose background and deterrent effects are more balanced. For example, among people who are mildly sensitive to deterrence, tipping is more likely for people who are mildly pre-disposed to deviant behavior (through c s.t. $|c| \approx a_1\mu + a_{-1}\delta$).

As a graphical proof I present nine cases in Figure 3.6, with $b = K = 0$, $a_1 = a_{-1} = 0.5$, $\mu = \delta = 1$, and $\alpha_1 = \alpha_{-1} = 1$. The first row involves $B = J = 0.4$ with $c = -2$, $c = -1$, and $c = 2$. The second row involves $B = J = 1$, and the final row involves $B = J = 2$. Only when $|\beta(B + J)|$ is sufficiently large is there a chance of multiple equilibria, and the actual size of $|\beta(B + J)|$ determines how large $|a_1\mu + a_{-1}\delta + c|$ can be while the system still retains tipping behavior/multiple equilibria. Notice that B and J need not be so large for multiple equilibria when background effects balance the deterrent

effect.

3.5.3 *The Role of the Impact of Disciplinary Response on Affect*

Each of the previous cases still involved symmetric maximum punishment and rewards $\mu\alpha_1 = \delta\alpha_{-1} = 1$, in that rewards for pro-social behavior induce a positive affect, and punishments for deviant behavior induce a negative affect. In this section I show that the extent to which the disciplinary response impacts student affect is also an important determinant of tipping behavior/multiple equilibria. In this subsection, suppose for simplicity that $|a_1\mu + a_{-1}\delta + c| = 0$ (specifically, let a_1, a_{-1} , and c be zero). Note this supposition is without loss of generality, in that we have already explored how that quantity affects the multiplicity of equilibria. Finally, let $\delta = \mu = 1$

Consider Figure 3.7. In these simulations, $K = 0.3$ and $B = J = 0.3$. Each of the three rows indicate three distinct cases: $\alpha_{-1} = 0.1$, $\alpha_{-1} = 0.7$, and $\alpha_{-1} = 1$. That is, the rows represent varying degrees of the effective impact of punishment on affect. Each of the columns also indicate three distinct cases: $b = 0.1$, $b = 0.5$, and $b = 1$. That is, the rows represent varying degrees of the sensitivity of personal affect to behavioral decisions. These figures illustrate that the impact α_{-1} of punishment on personal affect and the sensitivity b of personal affect to behavioral decisions both increase the likelihood of multiple equilibria. In this case, the role of b in generating multiple equilibria is operating through attitude transference K , which allows for the percolation of affect and therefore behavior throughout peer groups.

Altogether, this suggests that tipping is more likely for people whose affect is sensitive to disciplinary policy and whose behavioral decisions are sufficiently sensitive to affect. For example, students who are more likely to be unfairly punished, see such discipline as unfair, and whose attitudes and resulting behaviors are sensitive to that injustice are

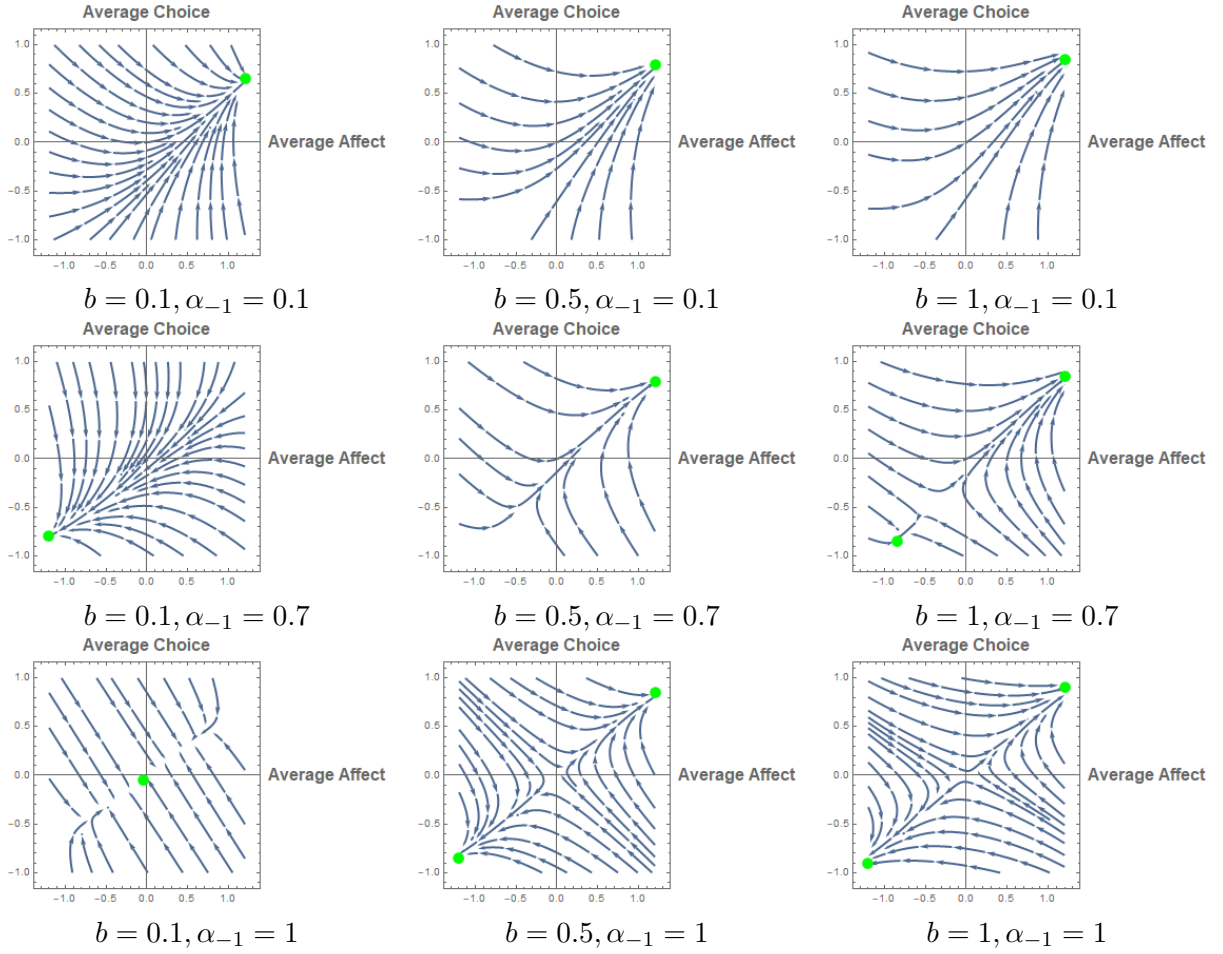


Figure 3.7: Impact α_{-1} of Disciplinary Response on Affect

more likely to exhibit tipping behavior. Also, holding rewards for pro-social behavior constant, tipping is more likely under harsh punishment (again, particularly when behavioral decisions are sensitive to student affect).

3.6 Changes in Disciplinary Policy can Induce Phase Transitions

The purpose of this paper is to propose a theoretical model and demonstrate its usefulness as a means to show how key mechanisms studied in the literature can interactively generate unintended consequences and explain some of the mixed results observed in the literature on disciplinary policy and deviant behavior in schools. In this section speci-

cally, I explore how changes in disciplinary policy can induce phase transitions—significant qualitative transformations of school systems. However, due to the complexity of the model, I will only highlight a few examples of phase transitions, noting that this analysis is not exhaustive.

3.6.1 *Changes in Punitive Discipline can Induce Phase Transitions*

First suppose the school official (rationally) maximally rewards pro-social behavior and partially punishes deviant behavior, so that $\mu = 1$ and $\delta \in [0, 1]$. Assume the parameters are such that there are multiple equilibria (but J is small). Then if b , K , B , and α_{-1} are sufficiently large compared with a_{-1} , decreasing punishment induces a phase transition. Specifically, decreasing punishment causes the system to bifurcate from three distinct equilibria to a single, positive equilibrium. Intuitively, decreasing punishment is improving behavior because the deterrent effect of harsh punishment is not worth its tendency to worsen deviant behavior through worsened affect. Furthermore, a sufficient decrease in punishment for deviant behavior breaks the feedback cycle studied in the psychology literature, in which harsh punishment for deviant behavior worsens student affect which worsens student behavior. The result is a positive school culture in the sense that on average, students engage in pro-social behavior with positive affect, regardless of initial conditions.

This is shown graphically in Figure 3.8. In this example, $\alpha_1 = \alpha_{-1} = 1$, $a_{-1} = 0.5$, $c = -1$, $b = 2$, $K = 0.1$, $B = 1.2$, and $J = 0.1$. In the first (left) plot $\delta = 1$, and in the second (right) plot $\delta = 0.1$. Notice the system has three equilibria for $\delta = 1$ but one, positive equilibrium for $\delta = 0.1$. This demonstrates a phase transition.

Notice, however, that J was small in this example. It turns out that if J is sufficiently large, the deviant equilibrium survives the policy change. Intuitively, although reducing

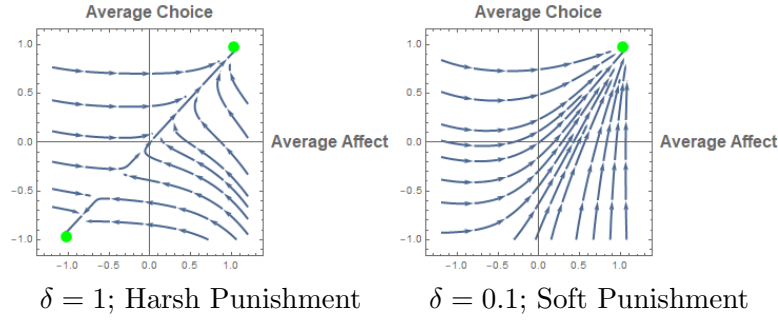


Figure 3.8: Changes in Punishment Can Induce Phase Transitions

punishment improves behavior in this case, the deterrent effect of punishment is too weak to protect from negative peer effects. Hence, if the school begins with slightly deviant behavior on average, it can still “tip” into the high-deviance equilibrium. This adds nuance to the notion that punitive discipline can attenuate the relationship between personal and peer delinquency discussed by Zimmerman [30].

3.6.2 *Increasing Rewards for Pro-Social Behavior Can Induce Phase Transitions*

Throughout much of this paper I have assumed the school official maximally rewards pro-social behavior, which is optimal because rewards always reduce deviance in this model (they improve behavior through improvements in affect, and also improve the net deterrent effect). However, rewards for pro-social behavior are not necessarily included in school disciplinary policies. The focus often surrounds the extent of punishment for deviant behavior. Suppose a school using harsh punishment for deviant behavior with small rewards ($\mu = 0.3, \delta = 1$) switches to a disciplinary policy with moderate rewards ($\mu = 0.6, \delta = 1$) and finally to a policy with larger rewards ($\mu = 0.9, \delta = 1$). Under the same punishment regime, it turns out that each change in rewards can induce a phase transition. Specifically, a school could have only a stable negative/“deviant” equilibrium under small rewards, bifurcate to a system with multiple equilibria under moderate rewards, and finally, bifurcate to a system with only a stable positive/“pro-social” equilibrium under large rewards. This has extremely important policy implications, especially

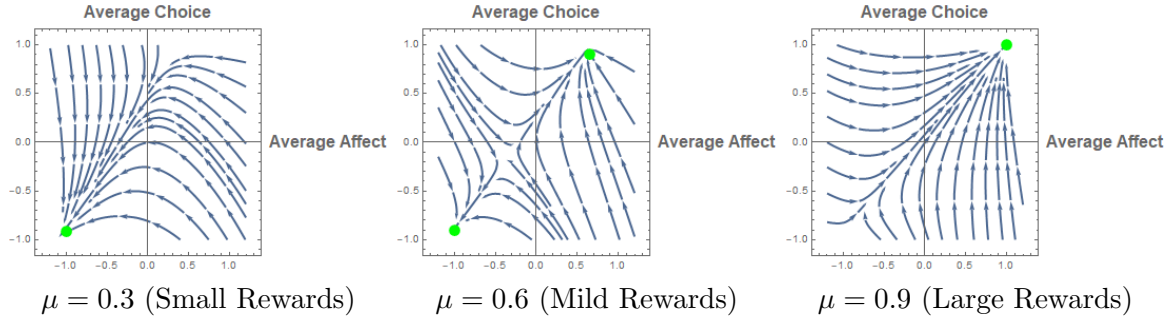


Figure 3.9: Changes in Rewards for Pro-Social Behavior Can Induce Phase Transitions

given the evidence that teachers tend to severely under-use praise for pro-social behavior in student-teacher interactions [32].

See Figure 3.9 for illustration. In these plots, $c = -1$, $K = 0.1$, $b = 1$, $J = 0.2$, $B = 1.2$, $\alpha_1 = \alpha_{-1} = a_1 = a_{-1} = 1$, and $\delta = 1$. When $\mu = 0.3$, all trajectories flow towards the negative equilibrium; when $\mu = 0.6$, trajectories flow to distinct equilibria depending on initial conditions; and when $\mu = 0.9$, all trajectories flow towards the positive equilibrium. For all of the attention on punishments for behavioral infractions, this illustrates the importance of rewards for pro-social behavior. Not only can rewards act as both a deterrent and an affect-improving mechanism, but as a result, they can fundamentally and qualitatively change the nature of the dynamics of behavior in schools. The extent of rewards can create or destroy the existence of positive or negative equilibria, as well as the tipping points that separate them.

3.7 Optimal Punishing and the Deterrence-Affect Tradeoff in the Case of a Unique Equilibrium

Multiple equilibria and tipping points have extremely important policy implications that have not yet been taken seriously in the literature on disciplinary policy. However, this model still has important implications even when the school in question exhibits a single equilibrium. In this section I graphically explore the never-before studied tradeoff between deterrence and affect evolution, with respect to punitive disciplinary practices. In

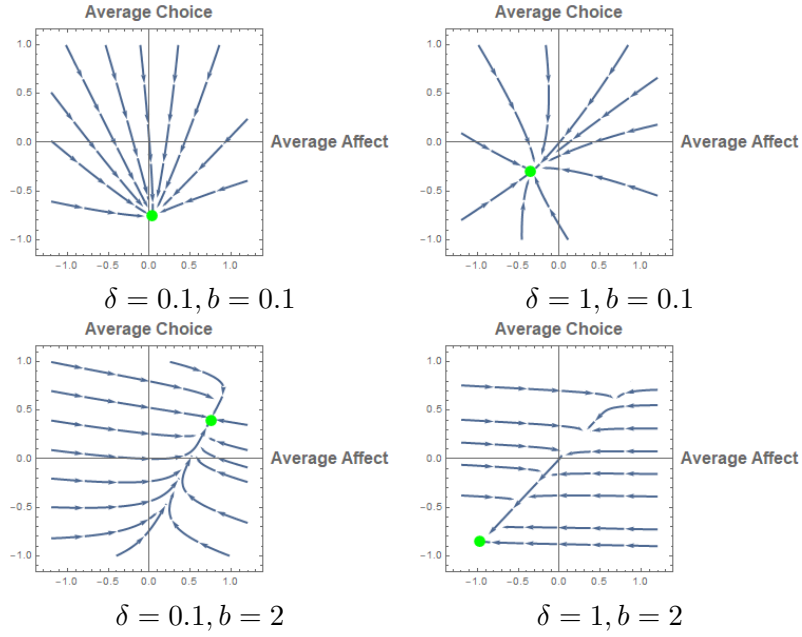


Figure 3.10: Optimal Punishment Depends on Measurable Tradeoffs

the Appendix is an explicit derivation of the relationship between the deterrent effect, the sensitivity of behavior to affect, and the impact of punishment on affect, that characterizes the conditions under which one should increase or decrease punishment.

Suppose the system leads to a unique long-run equilibrium. Choosing the optimal level of punishment is still not straightforward. The net deterrent effect $a_1\mu + a_{-1}\delta$, which improves student behavior, is increasing in the level of punishment. Hence punishments, which induce an incentive for students to avoid deviant behavior, may be important to a rational actor, as in the economics literature. On the other hand, punishment has a negative impact on student affect according to the psychology literature (through α_{-1}), and student affect, in turn, impacts student choice levels (through b). It turns out that the optimal level of punishment depends crucially on the relative size of the deterrent effect of punishment, the impact of punishment on affect, and the sensitivity of choice to affect. In general, punishment is criminogenic when its deterrent effect is sufficiently weak compared with the sensitivity of affect to punishment and the sensitivity of behavioral choices to affect.

To see this, observe the plots in Figure 3.10, corresponding to $b = 0.1$ and $b = 2$

(low and high sensitivity of behavioral decisions to personal affect), with outcomes under soft and harsh discipline ($\delta = 0.1$ and $\delta = 1$) being compared in each case. The other parameter values in this example are $\alpha_1 = \alpha_{-1} = a_1 = 1$, $c = -1.5$, $a_{-1} = 0.4$, and $J = 0.1$. When $b = 0.1$, it is clear that increasing punishment tends to improve average behavioral choice levels. However, when $b = 2$, increasing punishment worsens average behavioral choice levels. Intuitively, if student behavior has little to do with affect, a harsh punishment that harms affect will do little to worsen behavior through affect, and if the deterrent effect of punishment is sufficiently large, punishment will tend to improve behavior. Obviously, the opposite is true: if affect is a key driver of student behavior, punishment that harms affect will tend to worsen behavior unless the deterrent effect of punishment is so large that it supercedes this dynamic.

The key observation here is that these mechanisms have serious implications even when there is a single equilibrium. The impact of harsher punishment can still have qualitatively different systemic impacts under different parameter conditions. The deterrent effect of punishment, the impact of punishment on affect, and the sensitivity of behavior to affect are all empirical questions. Intuitively, groups that are more likely to be unfairly punished, who feel they are being unfairly punished (as a group or as individuals), and/or who are likely to act on this feeling are relatively unlikely to improve their behavior in the face of punishment unless the deterrent effect is truly significant.

3.8 What happens when Student Affect is Time-Sensitive?

Throughout this paper, the parameter b has captured the impact of a student's personal affect level on his/her choice. Importantly, the affect level has been allowed to take on one of two values at any given time, and the current affect level A_{it} only depends on the past through the previous behavioral decision y_{it-1} . This assumes a certain kind of memorylessness in the evolution of affect, such that over time, students can switch back and forth between behavioral states equally and infinitely often. However, there is reason

to believe that a student's personal affect in the current period should depend directly on the student's previous affect levels. As mentioned in section 3.2, the affective bundle considered in this paper consists of: attitude towards school, attitude towards authority, school connectedness, self-concept/identity, sense that school discipline is fair, and relationship with school officials. Several of these entities could have a certain kind of "stickiness" to them, in that their malleability is believed to decrease over time. Personal relationships could depend heavily on first impressions, identity becomes less malleable throughout adolescence and young adulthood, and beliefs are often updated according to a trajectory of events, and not just the most proximal event.

This intuition would suggest that a student's personal affect at time $t + 1$ could depend on some weighted sum of previous affect levels. Recall the updating process from section 3.2, with $d=1$:

$$E[A_{it+1}|A_1, \dots, A_t] \\ = \frac{1}{(t+1)^p} (\alpha_1 \mu Prob[y_{it} = 1] - \alpha_{-1} \delta Prob[y_{it} = -1] + K A_{it+1}^e) + \sum_{s=1}^t \frac{1}{s^p} A_{is}$$

This functional form, with $p \in (0, \infty)$ treats affect evolution as a power series, and implicit is the assumption that the marginal impact of school discipline decreases over time. Suppose that only this time-sensitive personal affect drives behavior. If the power series diverges ($p \leq 1$), then personal affect is allowed to grow in the negative or positive direction without bound, leading to a permanently, chronically pro-social or deviant student in the long-run.

If the power series converges ($p > 1$), then there are infinitely many equilibria in principle. The change in affect induced by the last punishment or reward still takes on only one of two possible values, so the number of equilibria grows as 2^t , which is the

number of possible trajectories up to time t . Because the series is convergent, however, we may establish a tolerance T beyond which the time-sensitive personal affect of each individual, as well as the likelihood of a pro-social choice for that individual, remains essentially unchanged. In that regard the number of equilibria is 2^T , where T is the number of timesteps until an equilibrium is reached.

The full system requires simulation for a full bifurcation analysis. I summarize the key takeaways here. The introduction of a time-sensitive affect has a few interesting impacts on the system dynamics. Most notably, no peer effect of any kind is required for tipping to occur. What drives the existence of tipping points is the presence of some kind of reinforcement process. In the case of peer effects, percolation over time through a system of peers can become self-reinforcing, generating multiple equilibria separated by a tipping point. With the introduction of a time-sensitive affect, a new reinforcement process emerges: student affect reinforces student behavior, reinforcing a behavioral response that led to that student affect. The likelihood of returning to a deviant behavioral state increases with the number of times one has returned to that state. As a result, deviations in initial behavioral decisions, affect, or disciplinary policies can now have large, persistent effects regardless of the peer parameters.

3.9 Discussion, Limitations, and Conclusion

Any theoretical model should be evaluated according to its intended use—in this case, not as a description of the real world, but as a tool to jointly study how key mechanisms in the literature impact developmental trajectories in school discipline in a dynamic context. As discussed in the literature review, there are a number of considerations in disciplinary policy which do not appear in this model—including specific classroom management techniques (e.g. replacement) and intrinsically motivated self-discipline. The model is not designed to account for these factors, although it can be altered to do so in the future.

To maintain tractability, it is essential that complex models of human behavior are initially constructed “as simple as possible, but no simpler,” in the words of Einstein. Any criticism of the model related to entities that it has not yet taken into account is still, of course, an essential step of the model-building process.

A distinct type of criticism relates to challenges in the underlying assumptions of the model. For example, I made a number of linear functional form assumptions in my construction of the utility functions. Student deterrence is linear in the extent of punishment. Because of this, the optimal punishment for students will lie at an extreme—maximum punishment or zero punishment. If instead deterrence were concave in the extent of punishment—so that the marginal amount of deterrence to be gained from a small increase in punishment decays with punishment—then the optimal punishment for students could lie somewhere between the extremes. However, this would not change the major conclusions of the model on the existence of multiple equilibria, tipping points, and tradeoffs between affect evolution and deterrence. I also assumed the random utility is logistically distributed. This was for mathematical convenience, but as explained by Brock and Durlauf, [12] the functional form of the random utility does not impact the key conclusions of this kind of discrete choice model.

In my derivation of system-wide behavior from microfoundational primitives, I invoked the concept of symmetric Baye’s Nash equilibrium. I also assumed that all students have the same utility functions and affect functions, barring random heterogeneity. Together, these assumptions lead easily to conclusions about the “average” choice level and affect level. If I were to allow for heterogeneity in some of the parameters—in b for instance (the impact of affect on choice)—or relax the symmetry assumption, then the calculation of average choice level and average affect level would be much more complex. However, the key results would remain. Students with a larger b would be more likely to exhibit multiple equilibria and to respond poorly to harsher punishment. I do intend to explore

heterogeneity in the parameters in future work through simulations.

Some of these conclusions have very important implications for African-American students. Some researchers have argued that African-American students are more likely to be (or perceive being) unfairly or harshly punished [81] [3], have affective traits (sense of identity, school connectedness, attitude towards school and authority, and quality of student-teacher relationships) that are sensitive to harsh punishment [70], exhibit behaviors and attitudes that are sensitive to those of their peers [35], and mildly more likely to be interpreted as deviant [69]. These very conditions increase the likelihood of the presence of tipping points and phase transitions discussed in this paper. In other words, small, appropriately targeted policy changes in discipline could make significant, persistent, even transformative differences particularly for African-American schools.

CHAPTER 4

**THE SELF-SUSTAINING EFFECTS OF ONE-TIME
INTERVENTIONS: SYSTEMS THINKING TO REDUCE
SYSTEMIC DISADVANTAGE**

ABSTRACT

Large bodies of literature explore the relationships between the effort, academic self-concept, and achievement of students in isolated, static models. However, students experience these mechanisms dynamically within a broader ecological system of coevolving peers. Driven by the utility of systems thinking for reducing systemic disadvantage, I formulate, estimate, and simulate policy counterfactuals using a structural model that reveals how the dynamic interplay between these entities can lead to phase transitions, tipping points, and other counterintuitive properties of complex systems. I show how these properties can help explain persistence and fadeout in a wide range of interventions in the literature, including early childhood education interventions, self-affirmation, growth mindset, and double-dose algebra. They allow for appropriately timed and targeted policy interventions to have self-sustaining impacts particularly on disadvantaged groups. The results have important implications for reducing educational inequities with persistence.

4.1 Introduction

In the backdrop of rising public consciousness of racial inequities at every level of American infrastructure lies a rich collection of literatures on how achieving equity in education is tied to a complex, ecological system of historical, political, social, economic, and psychological factors [53] [50] [49] [87]. Academics have long understood that the academic, psychological, and behavioral well being of students can impact a tidal wave of outcomes, including public health, intergenerational mobility, neighborhood context, political participation, and the infamous school-to-prison pipeline [39] [47] [23] [83] [41]. In turn, the well being of students is believed to be an effect of these causes [16] [74] [79]. It is impacted by practices implemented by many stakeholders (parents, teachers, principals, peers), and characterized by individual-level experiences including but not limited to self-concept, effort, beliefs, stereotype threat, mindset, and past achievement [38] [14] [92] [89] [95]. Altogether, the academic, psychological, and behavioral well being of students is clearly nested within a complex system of interacting causes and effects; and therefore existing academic racial inequities [75] both reinforce and are reinforced by racial inequities

throughout this system. Systemic racism is believed to be sustained by these kinds of reinforcement processes, resulting racial attitudes and stereotypes, and preexisting racial disadvantages stemming from explicitly racist policies and practices for at least 7/8ths of America's history [78] [31].

The interconnectedness of systemic disadvantages that reproduce academic inequities presents both substantive and methodological challenges that have been addressed with an agentic, ecological systems perspective in the education literature. This perspective holds that students have agency, but make decisions within the confines of nested sociopolitical and economic systems that include families, peers, teachers, schools, neighborhoods, school districts, and governments; and which also include the past and present choices and affective attributes of the student [13] [6]. Mathematically and empirically operationalizing such an important perspective in way that informs optimal policy design and evaluation would be a crucial step in the fight to reduce systemic inequities. To evaluate policy interventions, much of the work in economics of education has championed experimental and quasi-experimental methodologies, which, while essential to the production of knowledge, have generally proceeded under research paradigms that explicitly regard the broader system as a set of confounders while studying specific, isolated mechanisms in static models. We learn the value of the causal effect, but we miss the nuances that produce it. The result is a lack of closure on contrasting results, including interventions that fail to produce persistent effects (fadeout).

To fully operationalize ecological systems theory for public policy, one requires a research paradigm that can explicitly model the ecological system in question, while giving credence to the agency of individuals and showing the mechanisms through which individual choices coevolve with the broader system over time. I propose an applied theory approach that harnesses discrete choice modeling to capture fundamental behavioral mechanisms studied in the literature: that effort influences achievement, which influ-

ences academic self-concept, which influences effort; that the behavioral, affective, and academic experiences of peers impact those of the self; and that self-concept stabilizes over time. I show that a formal model of a system in which students are experiencing each of these mechanisms jointly and dynamically produces emergent properties that can help explain persistence and fadeout in a wide range of interventions. This is demonstrated with four theoretical examples and one empirical simulation.

First, the literature on early childhood education interventions shows evidence of both persistence and fadeout [5]. When persistence occurs, it occurs particularly for disadvantaged groups. I show that there is a range of parameter values that can explain these trends on the basis of peer effects. The model predicts that when peer effects are strong, experiments that randomize at higher levels (e.g., school districts, in which case the treated students tend to interact only with other treated students), could generate persistent effects whereas experiments generating the same initial effect size that randomize at lower levels (such as the individual or classroom level, in which case the treated and control students eventually mix) could generate fadeout. It also predicts that persistence would be experienced particularly for disadvantaged groups. Hence, accounting for just one aspect of the broader ecological system-coevolving peers-can generate mixed results that might seem counter-intuitive under dominant quantitative research paradigms, potentially leading to work that misrepresents the potential of Universal Pre-k to sustainably reduce racial inequity.

Second, the contentious growth mindset literature finds both persistence and fadeout [95] [4] [82], and once again, when persistence occurs, it is experienced particularly by disadvantaged groups. Growth mindset interventions aim to change how students view their own intelligence. The rationale is that rather than just directly increasing the skill level of students, switching the mindset of students who view intelligence as a fixed entity so that they view intelligence as a malleable entity should greatly improve skill acquisi-

tion over time. The model predicts that the feedback between effort, achievement, and self-concept makes students with a fixed mindset sensitive to early academic experiences, and that switching to a growth mindset induces a phase transition that particularly helps those who were predisposed to negative early experiences. It qualitatively replicates the mixed results of growth mindset interventions that are disproportionately persistent for disadvantaged groups.

Third, “double dose algebra” has revealed persistent effects, and a bit more-so for disadvantaged groups [20]. One hypothesis behind these effects is that “skill begets skill.” That is, earlier investments in skill make later investments more productive through dynamic complementarity [21]. The model reproduces this result by simulating a double-dose algebra intervention administered in the 9th grade, showing that dynamic complementarity leads to persistence - again, particularly for disadvantaged groups.

Finally, the literature on “self-affirmation” and “belonging” interventions has revealed both fadeout and, particularly for disadvantaged groups, persistence [18]. Cohen and coauthors argue that recursive processes allow an initial boost in academic self-concept to set off a “chain reaction” of effects that reinforce their cause, yielding persistence. I show how the strength of the dynamic interplay between effort, achievement, and academic self-concept, as well as how the academic self-concept stabilizes over time, can explain the mixed results of self-affirmation interventions. Finally, I use the Longitudinal Survey of American Youth (LSAY) to calibrate the EASE model, revealing plausible estimates for the parameters that characterize the strength of these reinforcement processes in a nationally representative sample of students. Conditional on the model with these estimates, I show that a realistic 0.43 standard deviation increase in students’ 9th grade academic self-concept still retains 30 to 60 percent of its original impact on effort, achievement, and self-concept by the end of the 12th grade.

In all of these cases, I show that persistence from a one-time intervention can be explained by properties of the model called reinforcement processes. These processes require a dynamic, systems perspective, and allow the possibility for policy interventions to spark “chain reactions” and harness tipping points and phase transitions in academic development. Because disadvantaged groups are so likely to experience negative early academic circumstances [22], exposure to low-achieving peers [35], exhibit fixed mindsets [24] and succumb to negative notions of the self through stereotype threat [89], they are particularly vulnerable to vicious cycles in the absence of intervention, and are disproportionately impacted by the positive impact of reinforcement processes in the presence of appropriately timed and placed interventions. This reveals how systems thinking can be used to reduce the achievement gap with persistence, in ways that would not be obvious under dominant research paradigms. The main contribution of this paper is to operationalize these notions mathematically and empirically, filling the gap between ecological systems theory and paradigms that focus on isolated causal effects of policy interventions.

The remainder of this paper proceeds as follows. I begin with a qualitative description of the formal “EASE” model (a model of reinforcement processes in student development; e.g., **E**ffort begets **A**chievement begets **S**elf-concept begets **E**ffort). Using the simplifying assumption that the subsequent period’s self-concept level depends only on the previous period’s achievement level, I derive an analytically solvable version of the model that illustrates how the dynamic interplay between effort, self-concept, achievement, and peer effects can yield emergent properties of complex systems in the Appendix. In Section 3 I relax that simplifying assumption, allowing the accumulation of all past achievement and self-concept levels to impact the next period’s academic self-concept. This allows for the self-concept to stabilize over time, as is consistent with the literature. Using variations of the full model I generate simulations to illustrate reinforcement mechanisms that can lead to persistence or fadeout for several very different types of interventions: early childhood education (through peer effects), growth mindset (through “phase tran-

sitions”), Double-dose algebra (through dynamic complementarity), and self-affirmation (through reinforcement between effort, achievement, and academic self-concept). Next, I present the LSAY data, the empirical analysis, and the simulation of a self-affirmation intervention. The last section concludes.

4.2 The EASE Model

In this section I present a brief qualitative description for readers who are less interested in the mathematical detail. Those who are interested in the derivation can skip to the Appendix. “EASE” is a structural model that shows how fundamental behavioral mechanisms that are normally studied in isolation can jointly and dynamically lead to emergent properties with important policy consequences. The model includes an ecosystem of N students, and mutual reinforcement mechanisms between their effort choices, academic self-concept, achievement, and that of their peers. Importantly, the rules of this model are based on literature in economics, education, and psychology. To ease the analysis, I assume each of the three main constructs (effort, academic self-concept, achievement) is binary, containing a “high” or “low” level (e.g., there is only “high” or “low” effort). Each parameter in the model is constructed based on empirical evidence in the literature. Although there are various additional mechanisms that could be included, I omit mechanisms that would not change the overall qualitative results of the simulations, and mechanisms for which I can find no empirical evidence.

At the beginning of the first period, students make effort decisions based on a pre-existing academic self-concept and the effort and academic self-concept levels of their peers. In that same period, each student’s achievement is updated based on her effort decision, academic self-concept, and the achievement level of peers. At the end of the period, the histories of the student’s achievement level and academic self-concept are both updated in the student’s memory, and used to stochastically update the academic

self-concept to be used in second period. The model continues until the terminal period T . The effort decision, the achievement updating process, and the academic self-concept updating process all include stochastic, idiosyncratic factors which are logistically distributed. The resulting system of difference equations is

$$e_{it}^* = d_0 + d_1 S_{it} + d_2 \overline{E}_t^e + d_3 \overline{S}_t^e + \epsilon_{1it} \quad (4.2)$$

$$A_{it}^* = a_0 + a_1 e_{it} + a_2 e_{it} S_{it} + a_3 S_{it} + a_4 e_{it} A_{i0} + a_5 A_t + \epsilon_{2it} \quad (4.3)$$

$$S_{it+1}^* = b_0 + b_1 \widehat{HS}_{it} + b_2 \widehat{HA}_{it} + b_3 \overline{S}_{t+1}^e + \epsilon_{3it+1} \quad (4.4)$$

In this system, e_{it} , A_{it} , and S_{it} represent the effort, achievement, and academic self-concept levels of student i and time t , respectively. The corresponding latent variables that determine whether effort, achievement, or self-concept levels are “high” or “low” are e_{it}^* , A_{it}^* , and S_{it}^* , respectively. The logistically distributed idiosyncratic variables that influence effort, achievement, and academic self-concept are given by ϵ_{1it} , ϵ_{2it} , ϵ_{3it} , respectively. The cumulative histories of academic self concept and achievement are represented by $\widehat{HS}_{it}$ and $\widehat{HA}_{it}$, where $\widehat{HS}_{it} := \sum_{k=1}^t \frac{2S_{ik}-1}{k}$ and $\widehat{HA}_{it} := \sum_{k=1}^t \frac{2A_{ik}-1}{k}$. These functional forms allow for the self-concept to stabilize over time, as is consistent with the literature [19], and for the history variables to be penalized from low self-concept or achievement levels. The main results are robust to other functional forms that preserve this qualitative property. $\overline{E}_t^e$ and $\overline{S}_t^e$ represent students’ beliefs of the average choice and self-concept levels (which will be correct beliefs in the Baye’s Nash equilibrium), while A_t represents the average achievement level across students.

The parameters d_0 , a_0 , and b_0 capture the baseline levels of effort, achievement, and self-concept. Parameters d_1 , d_2 , and d_3 capture the impacts of self-concept, peer effort, and peer self-concept on effort decisions. The direct impacts of effort and self-concept on achievement are captured by a_1 and a_3 . In addition, a_2 represents the synergistic relationship between effort and self-concept in the generation of achievement. This is

motivated by literature that may suggest that students' effort is more productive when certain affective traits are in a healthy state, which is consistent with a trend in the LSAY data which I discuss in a later section. a_4 is the dynamic complementarity parameter, in the sense that, if attaining the high-achieving state is particularly important at $t = 0$ (when a student takes algebra for example), subsequent investments in effort are more productive for students who successfully attained the high-achieving state at that time, by an amount equivalent to a_4 . The direct effect of peer achievement on student achievement is captured by a_5 . Finally, the impacts of the self-concept history, the achievement history, and the subsequent peer self-concept on the subsequent self-concept are captured by parameters b_1 , b_2 , and b_3 .

In this paper I demonstrate how this general model can be used to explain the persistence and fadeout of several famous interventions. First, I will show that the peer effect parameters (such as a_5) can help explain the mixed results of the persistence of early childhood education interventions that increase A_{i0} , the initial proportion of students with high achievement. Second, I show that the dynamic complementarity parameter a_4 can help explain the persistence of double-dose algebra interventions that increase A_{i0} . Third, if growth mindset interventions increase d_0 (the baseline preference for high effort) and decrease d_1 (the sensitivity of effort to academic self-concept), they can generate persistence particularly for disadvantaged groups by inducing a “phase transition” (changing the sensitivity to initial achievement levels). Finally, I show, using both theoretical and empirically calibrated simulations, that a self-affirmation intervention, which increases S_{i0} (the initial proportion of students with high academic self-concept) can generate persistence that depends on a_1 , a_2 , a_3 , d_1 , b_1 , and b_2 . In other words, the persistence of a self-affirmation intervention depends on the strength of the reinforcement processes between effort, self-concept, and achievement.

The remainder of the paper focuses on various simplifications of this most general

version of the EASE model, showing how different parameters interactively generate properties that might be considered counterintuitive under dominant research paradigms. These properties, as we will see, have very important policy implications for persistence and fadeout.

4.3 Theoretical Results

In the Appendix, I present and analyze a simplified, analytical solvable version of the model constructed formally from the model primitives. This version of the model assumes that the updated self-concept is completely determined by the previous period's achievement level. As a pedagogical tool to build the reader's intuition about the policy implications of the properties of complex systems, I present a key result from the simplified model below.

Figure 4.1 is called a phase portrait analysis. It is used to analyze the dynamic properties of an interrelated system of variables. The horizontal axis indicates the proportion of students (in a school district, for example) who are high-achieving, and the vertical axis indicates the proportion of students who are exerting high effort. The arrows constitute a vector field, showing the direction that developmental trajectories will take over time starting from any “initial condition” (that is, the initial proportion of high-achieving and high-effort students in a district). The green dots represent “stable equilibria” or “sink nodes.” They are equilibria because they are solutions to the underlying system of equations that define the system. They are stable because any small deviation from them in any direction will fade out over time, as all trajectories near a stable equilibrium return to the equilibrium. Notice there are three such equilibria in this example: a low-achieving equilibrium, a moderate equilibrium, and a high-achieving equilibrium.

Finally, the red dots indicate “unstable equilibria.” These act as tipping points, in the

sense that small deviations away from from them are self-reinforcing and persistent over time. Imagine how a ball at the top of a pointy hill can be tipped to the left or to the right. A small change on either side of that threshold will appear to have very different and persistent consequences for the ball's trajectory. Similarly, in our context, an intervention that shifts the high achieving proportion of students in a school district beyond some tipping point would appear to have persistent effects, whereas an intervention that fails to successfully shift the achievement beyond that point may appear to have effects that fade out over time.

In the Appendix I show that even this simplified model can actually have one, three, or five equilibria. The number of equilibria depends on the strength of the peer effect parameters relative to the other parameters. If the peer effect parameter values fall within a certain range, the simplified model would generate the properties displayed in Figure 4.1. For the rest of this section I return to the full model and specify the parameter values used in each simulation.

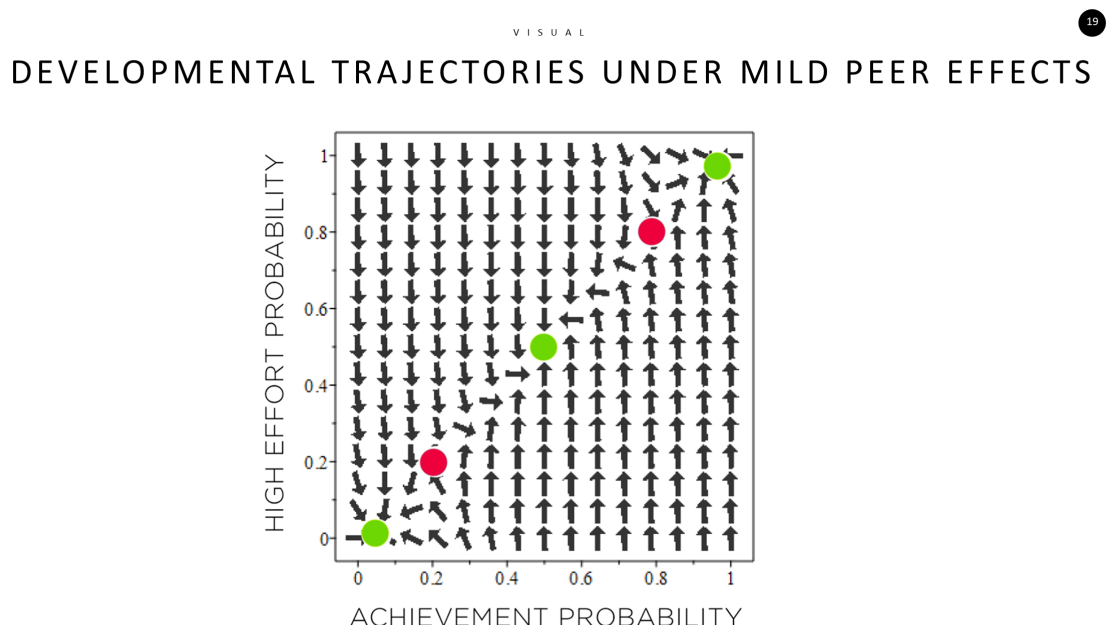


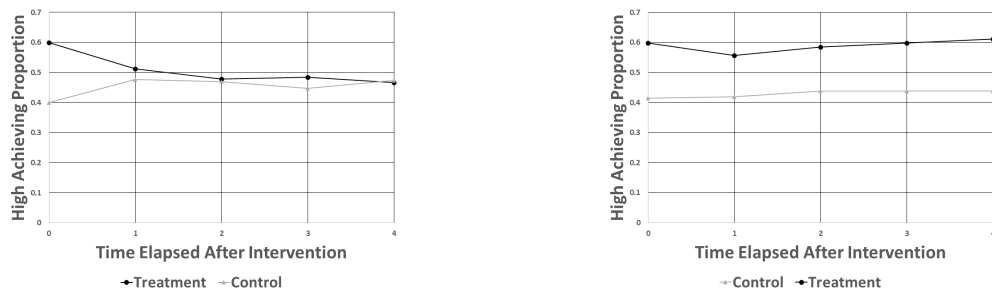
Figure 4.1: Potential Consequences of Systems Thinking

4.3.1 Persistence vs Fadeout through Peer Effects: Early Childhood Education

As Bailey, Duncan and coauthors suggest, one possible explanation for fadeout in early childhood interventions is spillovers from peer effects [5]. For example, early childhood interventions that are randomized at the classroom level generally involve treated students who eventually enroll in classrooms with control students. Negative peer effects could then negatively impact the effort decisions and achievement levels of the treated students, leading to convergence between the achievement levels of the treated and control students. For example, such peer effects could operate through teachers teaching to the median achievement level, so that higher-achieving students who received the intervention are not challenged and lower-achieving students who did not receive the intervention are able to catch up.

Suppose that a policymaker is ultimately interested in RCT's for the purpose of determining what works and implementing it on a large scale. In that case, we may be particularly interested in the intent-to-treat effect of a randomly assigned early childhood education intervention at the district level, to better match the conditions of a Universal Pre-K policy. After all, a Universal Pre-K policy with high take-up would involve classrooms and schools with mostly treated students. When implementing such an intervention at the district level, one would observe considerably less mixing between the treatment and control groups because students are likely to remain in the same school district for much of their schooling. Additionally, as students progress in their schooling, they would experience peer effects particularly from students of the same treatment status. If peer effort, self-concept, and/or achievement increases a student's own effort, self-concept and/or achievement, this suggests both positive peer effects in the treatment group and negative peer effects in the control group throughout students' years of schooling. These peer effects would act as reinforcement processes that help sustain the difference between the treated and control students-hence generating persistent effects.

I illustrate this using the EASE Model with one particular peer effect—the direct impact a_5 of peer achievement on one’s own achievement. In this simulation, $a_1 = 1.3$, $d_1 = 1.15$, $b_1 = 0.83$, and $a_5 = 2.55$, so that the other parameters are of moderate size compared with the peer effect parameter a_5 . This leads to a situation similar to the five-equilibrium case discussed in the previous subsection. If an intervention improves the achievement probability of the treatment group to some threshold amount, but the treatment and control groups are mixed together as in the left plot of Figure 4.2, then the overall achievement level of the full group will be less than that threshold amount, and over time the achievement level of the treatment and control groups will become more similar as the groups both return to the original equilibrium. However, under an intervention that has the same initial effect, but allows the treatment and control groups to remain separate, as in the right plot of Figure 4.2, peer effects can allow the treatment group to “tip” beyond a critical threshold, resulting in self-reinforcing, persistent effects.



Fadeout: Randomizing Within Districts

Persistence: Randomizing Across Districts

Figure 4.2: Persistence and Fadeout Under Strong Peer Effects

I simulate an early childhood intervention that exogenously improves achievement by twenty percentage points. When randomizing across districts (so that the treatment and control groups do not mix), I find large, sustained differences in the achievement levels between the treated and control districts. In contrast, when randomizing within districts so that treated and control students interact, the result is fadeout. This result holds for

a certain range of parameter values. Note that there is fadeout in both cases if the peer effect parameter a_5 is small. If the peer effect parameter for achievement a_5 is sufficiently large, persistence would be observed even when randomizing within districts as long as the treatment results in a shift in the proportion of high-achieving students beyond the tipping point.

Of course, even if a_5 were equal to zero empirically, there are still other reinforcement processes that could help sustain the effects of interventions. I review those in the following three subsections.

4.3.2 Persistence vs Fadeout Through Phase Transitions: Fixed vs Growth Mindset

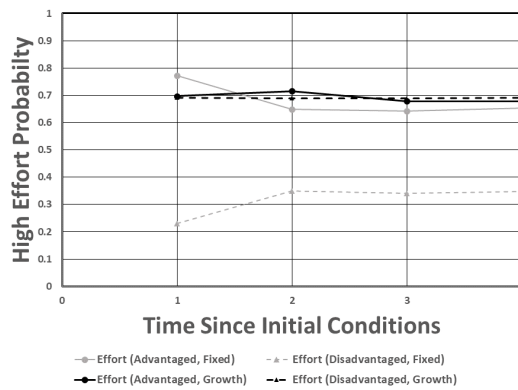
The EASE Model offers a unique perspective into the mixed results of the growth mindset literature, and the fact that when large, persistent effects are found, they are particularly found for disadvantaged groups [95] [4]. While the notions of fixed and growth mindset are very nuanced, one can consider the following simplification to illustrate the intuition. When a student has a fixed mindset, it is particularly true that his/her effort level depends very much on his/her academic self-concept of ability. That is, believing that intelligence is fixed could drive a student who doesn't believe she is good at math to also believe that her efforts to improve her math achievement are relatively fruitless. Similarly, it could drive a student who does believe she is good at math to also believe that her efforts are relatively more fruitful. The result is a dynamic system that is sensitive to initial conditions, so that disadvantaged students who tend to have initially negative academic experiences are caught within a vicious self-fulfilling prophecy, and advantaged students who tend to have initially positive academic experiences are actually caught within a virtuous self-fulfilling prophecy. The latter dynamic implies that there could be an aspect of fixed mindsets that are protective factors for advantaged students with favorable initial conditions—a possible reason for relatively less dramatic impacts of growth

mindset interventions for advantaged students.

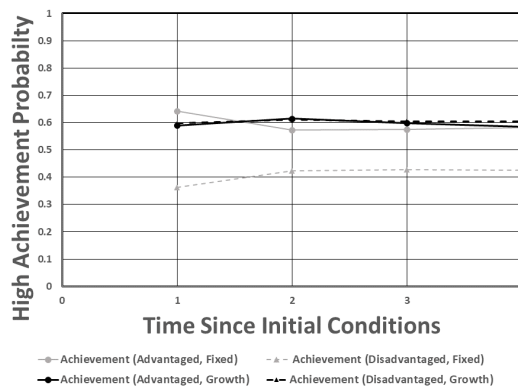
In contrast, growth mindset interventions cause students to change how they make meaning out of their experiences with achievement. Consider two mechanisms. First, and most importantly, they cause students' effort levels to be less sensitive to their academic self-concept of ability (through the parameter d_1). Students now believe that intelligence is malleable through effort, and so believing that they are currently "not good at math" no longer causes them to put in less effort. Similarly, believing that they are currently "good at math" doesn't necessarily increase their effort levels either. Instead, regardless of their current beliefs of their math ability, they are induced to increase effort simply because they believe that effort is what really matters. This activates the parameter d_0 , which characterizes the baseline preference for high effort regardless of academic self-concept. As a result of these mechanisms, students with initially negative experiences are able to persistently close the gap with their counterparts. This would appear to be particularly dramatic and persistent for students with initially negative experiences, because the growth mindset intervention has induced an actual "phase transition" by fundamentally altering the underlying dynamics of students' developmental trajectories-dynamics that previously allowed for the stability of a vicious cycle for initially disadvantaged students.

This is illustrated through a theoretical simulation resulting in Figure 4.3. In this simulation, $a_1 = b_1 = 1.2$. For students with fixed mindsets, $d_1 = 1.2, d_0 = 0$ and for students with a growth mindset, $d_1 = 0, d_0 = 0.8$. Over a four-year period, initially low achieving (disadvantaged) students with fixed mindsets retained low achievement and effort, and initially high achieving (advantaged) students with fixed mindsets retained high achievement and effort. Over that same period, initially low achieving students with growth mindsets reached the achievement levels of their initially high-achieving peers. Therefore, the overall impact of growth mindset is being driving primarily by

initially low-achieving students with fixed mindsets. The impact of growth mindset for initially high-achieving students depends strongly on the size of d_0 , whereas the impact for initially low-achieving students is driven largely by d_1 . Growth mindset interventions that only slightly increase d_0 but strongly decrease d_1 would appear to have modest effects overall, with strong, persistent effects for disadvantaged groups.



Growth Mindset: Effort



Growth Mindset: Achievement

Figure 4.3: Growth Mindset: Persistence Particularly for the Disadvantaged through Phase Transitions

4.3.3 Persistence vs Fadeout Through Dynamic Complementarity:

Double-Dose Algebra

One high school intervention that is famous for having pronounced and persistent effects is “Double-Dose Algebra,” in which a group of students who scored just below a cutoff

on an 8th grade math exam were given the chance to enroll in two simultaneous algebra courses in their ninth grade year. Regression discontinuity reveals that students who scored just below the cutoff experienced significantly higher achievement as measured by grades, test scores, high school graduation rates, and college enrollment [20]. Bailey and coauthors [5] discuss this result by pointing to algebra as a “Trifecta Skill-” a skill that is fundamental, malleable, and would not generally be mastered inevitably without the intervention. Interventions targeting such skills are argued to be likely to achieve persistent effects.

One reason fundamental skills can lead to persistent effects is the potential for dynamic complementarity. Dynamic complementarity occurs when past increases in skills make future investments more productive [21], and it is another reinforcement mechanism that can lead to persistent effects of interventions-potentially reducing inequality [44]. In this subsection I present one operationalization of dynamic complementarity that can lead to persistent effects for this intervention. Suppose that ninth grade algebra mastery, being a fundamental skill, is one that makes future investments in effort more productive. For example, perhaps a student gets more out of studying hard for geometry in the tenth grade than that student would have for studying equally hard for geometry had she not mastered ninth grade algebra. This is operationalized by the a_4 parameter. Setting this parameter to a value of 2 with all other parameters set to zero, I use the EASE model to simulate an intervention that exogenously improves the 9th grade achievement level by 20 percentage points in Figure 4.4. I find that while the initial effect dies down, persistent effects remain through the end of high school. This persistence is driven by the fact that the efforts of students who successfully acquired the 9th grade skill were more productive than their counterparts. I repeat this simulation when a_4 is set to 0, reflecting no dynamic complementarity and no mechanism through which the 9th grade success affects the future technology of skill formation. As expected, the difference between the treatment and control groups has completely died down by the end of high school.

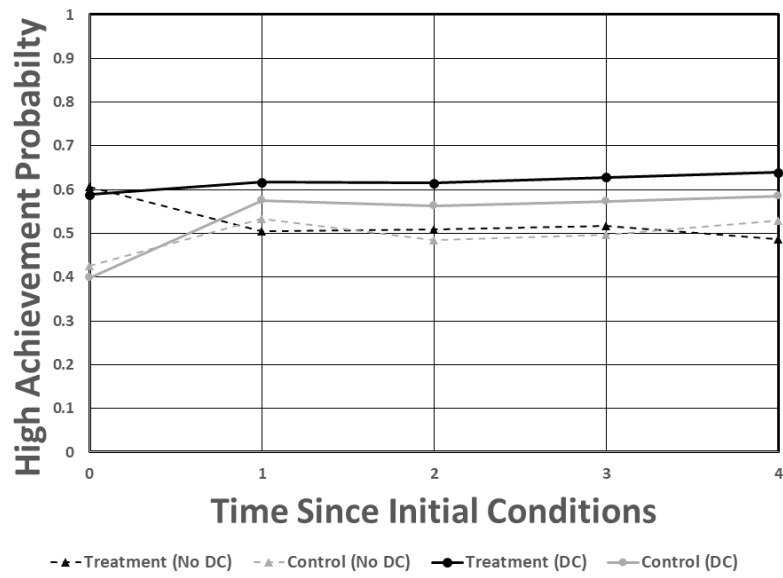


Figure 4.4: Double Dose Algebra: Persistence through Dynamic Complementarity

4.3.4 Persistence vs Fadeout through Feedback: Self-Affirmation

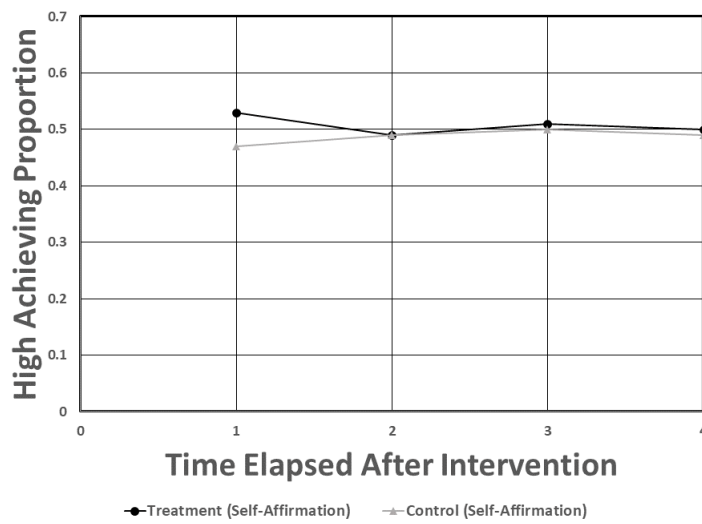
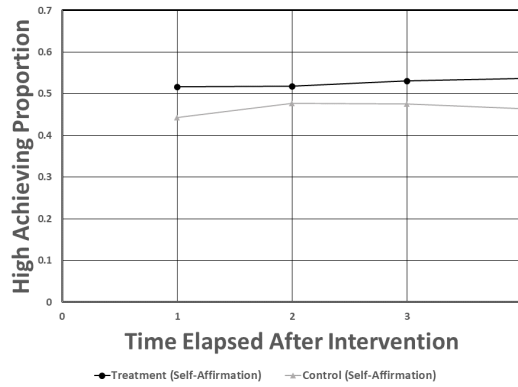


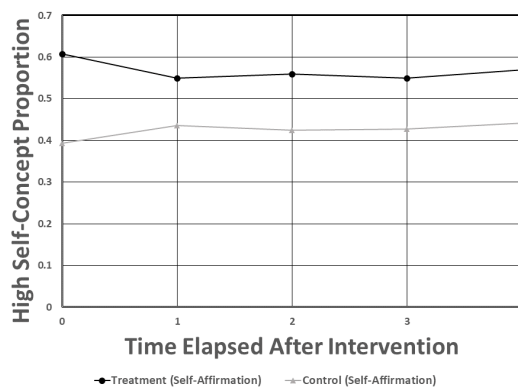
Figure 4.5: Self-Affirmation: Fadeout Due to Weak Feedback

Cohen and coauthors [18] argue that self-affirmation interventions are particularly persistent because of recursive processes through which students respond to achievement in a way that reproduces the achievement. They are able to verify this claim experimentally.

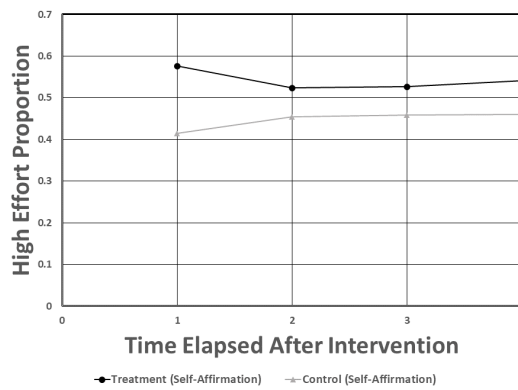
The EASE model is able to formalize this by simulating a randomly assigned self-affirmation intervention that exogeneously increases the academic self-concept. In this theoretical simulation, $a_1 = 1.31$ and $d_1 = 1.59$. When $b_1 = b_2 = 0.1$, so that self-concepts are fairly insensitive to cumulative histories of academic experiences, the model predicts fadeout as in Figure 4.5. This fadeout would exist with respect to achievement, effort, and academic self-concept. When and $b_1 = 1.02$, so that self-concepts are sensitive to cumulative histories of academic experiences, the model predicts persistence, as in Figure 4.6. Persistence operates through the feedback loop between the impact d_1 of academic self-concept on effort, the impact a_1 of effort on achievement, and the impact b_1 of a cumulative history of achievement and academic self-concept on current academic self-concept. When those factors are sufficiently salient, students who are randomly assigned into the treatment group and hence improve their academic self-concept by 20 percentage points initially still retain persistent effects after four years. This spills over into effort and achievement. An intervention that operates on this system of mechanisms should disproportionately impact disadvantaged groups that are more likely to fall into “vicious cycles” absent the intervention, while falling into “virtuous cycles” through the intervention. Indeed, Cohen and coauthors find persistent effects particularly for African-Americans [18].



Self-Affirmation; Persistent Effects on Achievement



Self-Affirmation: Persistent Effects on Academic Self-Concept



Self-Affirmation: Persistent Effects on Effort

Figure 4.6: Self-Affirmation: Persistence Due to Strong Feedback

In this section I have shown that the presence or absence of various types of reinforcement processes can drive the persistence of many different types of one-time interventions. Because of this, it is important to investigate whether plausible estimates of the relevant

parameters that characterize such reinforcement processes are strong enough to generate persistent effects, using real data. In the next section I take the important next step of generating plausible estimates of the parameters, verifying that the model is able to replicate important moments in the data using those parameter values, and running a counterfactual simulation that suggests the extent of persistence we might expect from a one-time self-affirmation intervention.

4.4 Empirical Analysis

4.4.1 *Purpose*

The parameter values of the EASE model have strong implications for the possibility of reinforcement processes that can lead to persistence. For example, we have seen that a self-affirmation intervention retains persistent effects when the parameters that characterize the feedback between effort, achievement, and self-concept are sufficiently strong. Therefore, it is important to consider whether real data suggests that the relevant parameters are strong enough to generate salient reinforcement processes. To test this, I estimate a subset of the parameters by calibrating the model to a national data set. I verify that the model, with the estimated parameters, approximately replicates the trajectories in effort, achievement, and self-concept in the data. After this validation exercise, I simulate a self-affirmation intervention that increases the likelihood of a high academic self-concept for 9th grade students by a realistic 0.43 standard deviations (20 percentage points), and record the impact on self-concept, effort, and achievement by the end of high school. The extent of persistence will suggest whether American students have parameter values that are ripe for interventions that take advantage of reinforcement processes. That, along with obtaining plausible estimates for the parameters, is the purpose of the empirical section.

4.4.2 *Data*

To empirically estimate the EASE model I rely on the nationally representative Longitudinal Survey of American Youth (LSAY) which collected data on students' effort decisions, self-concept, and achievement between 1987 and 1992. The data includes two cohorts. The older cohort was enrolled in the 10th grade in the fall of 1987 while the younger cohort was enrolled in the 7th grade during that same year. The sample contains about 60 different high schools and "feeder" middle schools. The study surveyed about 40 students per school on average. The data includes information about math and science courses, teachers, parents, and socioeconomics. I focus my analysis on mathematics.

I omit information about the older cohort, focusing only on the younger cohort. This is due to the focus of the model on dynamic evolution over time within an individual. I also only include data between 1989 and 1992, representing the years the younger cohort was in high school. The analytic sample consists of 3116 individuals, with 47.8% identifying as women. The sample is coded as 9.6% Hispanic, 73% White, 11.8% Black, 3.8% Asian, and 1.6% Native American. The measures for the outcomes of interest are depicted in the table below. For the empirical analysis I chose the response to "I do my best in school," "I am good at math," and the school-reported math grades to measure student effort, academic self-concept, and achievement, respectively. I code each of these as binary. For example, the response to "I do my best in school," is one of five choices based on a standard five-point scale: "Always," "Usually," "Sometimes," "Rarely," or "Never." I set the effort variable equal to 1 if the response was "Always," or "Usually," and 0 otherwise. Similarly the possible responses to "I am good at math" are "Strongly agree," "Agree," "Somewhat agree," "Disagree," and "Strongly Disagree." I set the academic self-concept variable equal to 1 if the student responded with "Strongly agree" or "Agree" and 0 otherwise. As explained above the achievement variable is equal to 1 if the student received A's and B's that year in math classes and 0 otherwise.

To motivate the statistical models that I estimate for the empirical analysis, I show three descriptive trends from the data that help capture the coevolution between effort, academic self-concept, and achievement. Each of these figures can be found in the Appendix. First, Figure 4.13 shows that in each year, the conditional probability of high effort given high academic self-concept is clearly larger than the conditional probability of high effort given low academic self-concept. This suggests that academic self-concept, as in the literature, predicts effort.

Second, Figure 4.14 shows that 1) effort predicts achievement (since conditional on the same self-concept level, the high effort condition leads to a greater likelihood of high achievement than the low effort condition); 2) academic self-concept predicts achievement (since conditional on the same effort level, the high self-concept condition leads to a greater likelihood of high achievement than the low self-concept condition); and 3) that effort and self-concept may synergistically predict achievement (since the high vs low effort condition appears to make a much greater difference in the likelihood of high achievement when students have a high academic self concept, while the high vs low effort condition barely makes a difference when students have a low academic self-concept). This is consistent with psychology literature suggesting that certain unhealthy affective traits could make studying less productive.

Finally, Figures 4.15 and 4.16 show that, in each year, the likelihood of a high self-concept for the subsequent year is clearly predicted by the self-concept and achievement history variables. In particular, the conditional probability of a high subsequent academic self-concept is substantially larger when a student has a positive achievement or self-concept history, than when a student has a negative achievement or self-concept history.

I proceed with a description of the simplified model to be estimated in the next subsection.

4.4.3 Identification and Statistical Estimation

The purpose of this empirical analysis is not to causally identify the parameters of the model. The goal is to generate plausible estimates for the parameters that are relevant for a self-affirmation intervention, and to run a counterfactual simulation of a self-affirmation intervention to measure the extent of persistence using those estimates. I focus now on a simplified version of the full model, which I refer to as “Model 1:”

$$e_{it}^* = d_1 S_{it} + \epsilon_{1it} \quad (4.5)$$

$$A_{it}^* = a_1 e_{it} + a_2 e_{it} S_{it} + a_3 S_{it} + \epsilon_{2it} \quad (4.6)$$

$$S_{it+1}^* = b_1 \widehat{HS}_{it} + b_2 \widehat{HA}_{it} + \epsilon_{3it+1} \quad (4.7)$$

In this system, e^* , A^* , and S^* are latent effort, achievement, and academic self-concept, respectively; and i indexes an individual while t indexes a year. Importantly, $\widehat{HS}_{it}$ and $\widehat{HA}_{it}$ represent time-weighted histories of self-concept and achievement levels, respectively. These histories appear as sums that decrease in the event of a low academic self-concept or achievement level. That is, $\widehat{HS}_{it} := \sum_{k=1}^t \frac{2S_{ik-1}}{k}$ and $\widehat{HA}_{it} := \sum_{k=1}^t \frac{2A_{ik-1}}{k}$.

For the reader’s convenience I reiterate how the model proceeds. In the first period, each student makes an effort decision based on the impact (d_1) of a pre-existing self-concept level and idiosyncratic factors ϵ_{1it} . In the same period, each student updates her achievement level based on a direct effect (a_1) of effort, a direct effect (a_3) of academic self-concept, a synergistic impact (a_2) of effort and self concept, and idiosyncratic factors ϵ_{2it} . Finally, each student updates the academic self-concept to be used in the subsequent period based on the impact (b_1) of the self-concept history, the impact (b_2) of the achievement history, and idiosyncratic factors (ϵ_{3it}). All idiosyncratic variables are logistically distributed.

Note that there is a potential for simultaneity in the estimates of d_1 , a_1 , a_2 , and a_3 . This specification is one of many possible alternative specifications, each facing important tradeoffs. If I were to rely on the impact of the previous self-concept level on current effort, and the impact of the previous effort and self-concept levels on the current achievement levels, I would remove the simultaneity concern. However, I would lose a period's worth of data, which is quite costly given the number of parameters and the fact that there are only $T = 4$ periods in my analytic sample. Additionally, it is likely that (for example) the true effect of a student's 9th grade effort on her 10th grade achievement level underestimates the true effect of her 10th grade effort on her the 10th grade achievement level. Nonetheless, the qualitative conclusions of the model remain unchanged when I allow for a period's separation between causes and effects. Those results are available upon request.

Although causal identification is not the goal of this section, the assumptions required to causally estimate the parameters are as follows:

- For the first equation: $cov(\epsilon_{1it}, S_{it}) = 0$
- For the second equation: $cov(\epsilon_{2it}, S_{it}) = 0$ and $cov(\epsilon_{1it}, \epsilon_{2it}) = 0$. Together these ensure that $cov(e_{it}, \epsilon_{2it}) = 0$.
- For the last equation: $cov(S_{is}, \epsilon_{3it+1}) = 0$, $cov(\epsilon_{1is}, \epsilon_{3it+1}) = 0$ and $cov(\epsilon_{2is}, \epsilon_{3it+1}) = 0$ for all $s = 1, \dots, t$. Together these ensure that $cov(\hat{H}_{it}, \epsilon_{3it+1}) = 0$.
- Sequential exogeneity: All past error terms are allowed to be correlated with later regressors, but all past regressors must be uncorrelated with future error terms. That is, $cov(\epsilon_{1it}, S_{is}) = cov(\epsilon_{1it}, e_{is}) = cov(\epsilon_{1it}, A_{is}) = 0$ for all $s = 1, \dots, t - 1$ and all interior time periods $t = 2, 3, 4$. (This is a four-period model).

If these assumptions hold I can causally estimate the system using structural equation modeling with maximum likelihood. In a recent paper by Paul Allison and Richard Williams, it is shown that this method can estimate dynamic panel data models with

feedback processes under sequential exogeneity [2]. One could include time varying confounders in the regressions directly (which could lead to over-control bias), or take an approach similar to Wodke’s regression-based adjustment that residualizes time-varying confounders [93]. However, that approach has not been adapted to logistic regression contexts. There is of course no perfect method for dealing with time-varying confounding. I plan to explore methods to address this concern in future work, since the main purpose of this empirical exercise is only to show how the model can be used to reveal how reinforcement processes have consequences for persistence and fadeout.

What would violate these assumptions besides time-varying confounding due to feedback from past errors to future regressors? Perhaps there are person-specific, time-invariant traits such as fixed ability and/or conscientiousness that are correlated with self-concept, achievement, and effort even net of the others. These traits could lead to violations in all of the assumptions. Another class of variables that could lead to violations in those assumptions are person-invariant, time-specific quantities such as national crime trends. Such variables could be correlated with both effort and achievement in a given year just because of the timing of crime trends, and omitting them could induce a correlation between ϵ_{1it} and ϵ_{2it} , for example. As a step towards accounting for these concerns, I estimate a model that uses time trends and a time-invariant, person-specific random intercept for each equation. I refer to this as “Model 2:”

$$e_{it}^* = d_1 S_{it} + c_{1i} + \gamma_t + \nu_{1it} \quad (4.8)$$

$$A_{it}^* = a_1 e_{it} + a_2 e_{it} S_{it} + a_3 S_{it} + c_{2i} + \gamma_t + \nu_{2it} \quad (4.9)$$

$$S_{it+1}^* = b_1 H \hat{S}_{it} + b_2 H \hat{A}_{it} c_{3i} + \gamma_t + \nu_{3it+1} \quad (4.10)$$

where the c_{1i} , c_{2i} , and c_{3i} are person-specific, time invariant random intercepts (which are marginally normal), γ_t captures all time-varying, person-invariant unobservables, and ν_{1it} , ν_{2it} , and ν_{3it} are idiosyncratic, and logistically distributed conditional on the inter-

cepts.

I estimate both Model 1 and Model 2 using generalized structural equation modeling with maximum likelihood. The systems are jointly estimated with logistic regression. I report the results in the next subsection.

4.4.4 *Parameter Estimates*

Table 4.1: Baseline Structural Equation Model

	Effort	Self-Concept (Subsequent)	Achievement
Effort			0.11 (0.09)
Self-Concept	0.87*** (0.04)		***1.22 (0.09)
Effort*Self-Concept			***0.55 (0.1)
Achievement History		0.4*** (0.04)	
Self-Concept History		0.65*** (0.04)	
Constant	0.43*** (0.03)	-0.16*** (0.05)	-1.35*** (0.07)
Notes: Robust standard errors in parenthesis. Asterisk denotes statistical significance (***) implies pvalue smaller than 0.001)			

All empirical results are based on the LSAY data between the years 1989 and 1992 (9th - 12th grade). I refer to the specification with no random intercepts or time dummies as the baseline model, or “Model 1.” I will refer to the full specification that includes random intercepts and time dummies as “Model 2.”

Across both specifications, nearly all parameters are strongly statistically significant. The exception is the impact of effort on achievement, which not statistically significant in any specification. However, there is a salient interaction effect between effort and aca-

Table 4.2: Full Model With Random Intercepts and Time Dummies

	Effort	Self-Concept (Subsequent)	Achievement
Effort			-0.22 (0.12)
Self-Concept	0.81*** (0.06)		***1.21 (0.12)
Effort*Self-Concept			***0.59 (0.14)
Achievement History		0.26* (0.11)	
Self-Concept History		1.15*** (0.1)	
Constant	1.08*** (0.06)	1.15*** (0.3)	-0.69*** (0.1)

Notes: Robust standard errors in parenthesis. Asterisk denotes statistical significance (***) implies pvalue smaller than 0.001)

demic self-concept, suggesting that effort statistically influences achievement particularly through academic self-concept. In these models, effort has little predictive impact on achievement unless a student has a high academic self-concept. The inclusion of the random intercepts and time dummies changes the estimated impact of the history variables on subsequent academic self-concept, increasing the impact of self-concept history and slightly decreasing the impact of achievement history.

4.4.5 Empirical Simulation

These parameter estimates contain information about the salience of reinforcement processes in the data. Before exploring how such reinforcement processes could lead to persistence or fadeout, I first verify the extent to which the mathematical model is able to replicate the trajectories in effort, achievement, and academic self-concept in the data when using the estimated parameters. This exercise is carried out in Figure 4.7 below for Model 1. The mathematical model uses only the 9th grade academic self-concept in the data as its initial condition, and then simulates the resulting 9th grade effort and

achievement levels, as well as the effort, self-concept, and achievement levels through the 12th grade. On average, the euclidean distance between the simulated moments and the moments in the data is about 5 percentage points. That is, the mathematical model replicates the moments in the data within about 0.1 standard deviations on average. This fit improves slightly when I use Model 2 (which includes the random intercepts and time dummies).

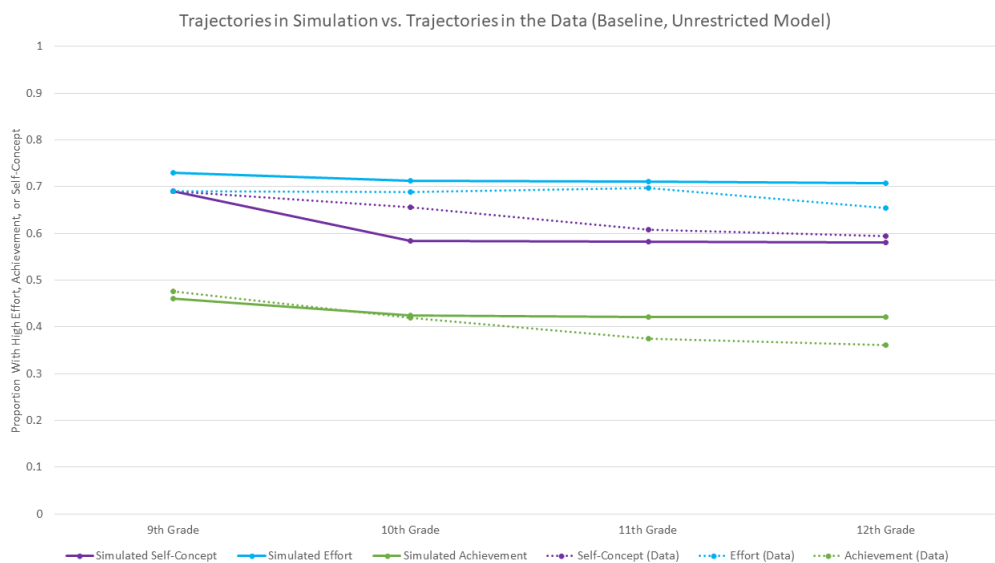


Figure 4.7: Model 1 Validation Exercise

Suppose a large group of American students similar to those in the LSAY data set were exposed to a self-affirmation intervention that successfully improved students' academic self concepts by 20 percentage points (or about 0.43 standard deviations. Note that the literature on self-affirmation interventions contains effect sizes ranging from zero to well above 1 standard deviation). The intervention is randomly assigned to half of the students and the other half, which is otherwise similar to the first half, receives the control. The intervention increases the proportion of 9th students with a high academic self-concept from its 0.69 value in the data to 0.89, while nothing happens to the control group. Will this increase completely fadeout over the 4 years of high school? Will it persist, and to what extent? What will be the initial impact of that increase on achievement and effort, and how will those initial impacts change over time?

The answers to these questions are depicted in Figures 4.8 and 4.9. Each figure shows the trend in the average effect sizes (in standard deviations) on effort, self-concept, and achievement over 20 trial runs. The first figure corresponds with Model 1, which has no random intercepts or time dummies; and the second figure corresponds with Model 2, which includes the full specification with random intercepts and time dummies. Across all specifications, at least one third and at most two thirds of the initial impact of the intervention is maintained by the end of high school. In the Model 1 simulation, just under half of the initial impact of the intervention on self-concept and achievement is maintained by the 12th grade, and about two thirds of the initial impact on effort is maintained by the 12th grade. In the Model 2 simulation, about one third of the initial impact of the intervention on self-concept and achievement is maintained by the 12th grade, and again about two thirds of the initial impact on effort. The inclusion of random intercepts decreases the extent of persistence as expected, but persistence is by no means eliminated.

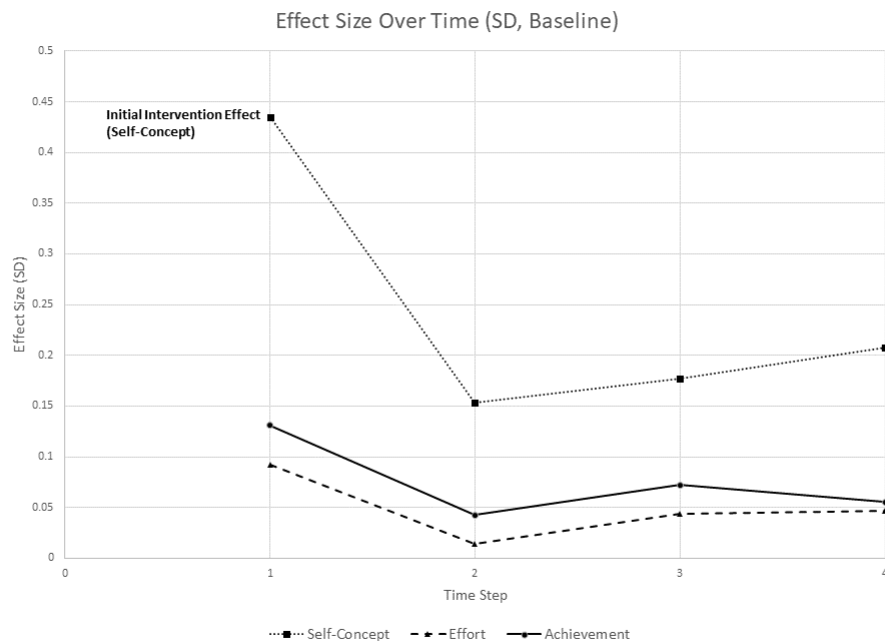


Figure 4.8: Model 1: Simulating a Self-Affirmation Intervention

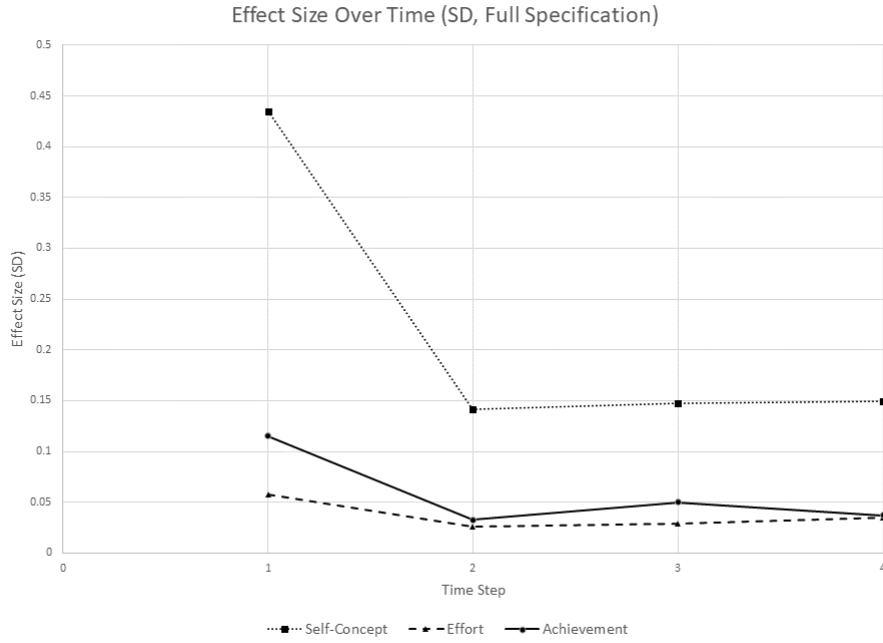


Figure 4.9: Model 2: Simulating a Self-Affirmation Intervention

This shows suggestive evidence that reinforcement processes could be salient in the dynamics of American students’ developmental trajectories, such that even one-shot interventions that take advantage of these reinforcement process stand to obtain modest persistence. Specifically, across all specifications, a simulated self-affirmation intervention still retains 30 to 60% of its initial impact on achievement, effort, and academic self-concept by the end of the 12th grade, and this persistence is driven by the strength of the parameters that characterize reinforcement processes between effort, self-concept, and achievement.

4.5 Discussion

4.5.1 Summary

The research paradigm exemplified in this work serves a special purpose at the intersection between theoretical and empirical literature on student development. As such, it makes both theoretical and empirical contributions-both of which shed light on empirical

literatures. The theoretical contribution is to mathematically operationalize ecological systems theory, synthesizing a generalized explanation for persistence and fadeout for wide range of educational interventions that is based on simple, fundamental behavioral mechanisms. In doing so, I make a methodological contribution that connects micro-foundational discrete choice theory to phase portrait analysis developed from dynamical systems theory, revealing how tipping points can arise from reinforcement processes and could act as important policy levers to generate persistent effects. I show, as examples, that a Universal Pre-K intervention (through peer effects), growth mindset intervention (through “phase transitions”), double-dose algebra (through dynamic complementarity) and self-affirmation interventions (through feedback between academic self-concept, effort, and achievement) can retain persistent effects through reinforcement mechanisms. The empirical contribution, which is not causal, complements the theoretical contributions. I show suggestive evidence that students in a nationally representative study who were in high school from 1989 – 1992 appear to exude salient reinforcement processes, specifically with respect to the dynamic interplay between effort, academic self-concept, and achievement. I show that given the estimated relationships, a substantial but typical 0.43 standard deviation increase in the academic self-concept of students in the ninth grade retains 30 – 60 percent of its original effect by the end of the 12th grade. This is potentially promising news for the potential of a one-time intervention to have modest persistence due to the salience of reinforcement processes in human development. Particularly if a self-affirmation or belonging intervention principally impacts African-Americans, as is the case in the literature, this result suggests specific ways to harness systems thinking to persistently reduce the achievement gap.

4.5.2 *Limitations and Conclusion*

There are several modeling assumptions that can influence the quantitative results, but which have not altered the qualitative conclusions. First, the functional form of the history variable $\hat{H}_{it}$ was important for the exact quantitative extent of persistence, but

qualitatively the results are robust to alternative specifications that still allow the academic self-concept to stabilize over time through past achievement and self-concept levels that matter less and less as time moves forward. The key assumption here that is driving the results is the notion that the academic self-concept crystallizes over time through this accumulation of experiences. Any functional form that respects that assumption should yield qualitatively similar results. Importantly, that assumption comes from psychological research [19]. Another concern is simultaneity from measuring certain variables concurrently. One would expect persistence to decrease if all parameters were identified from variables that were measured one year apart from one another, as this would likely decrease the estimates of the causal effects. This would also cause us to lose a year's worth of data, and the true causal effect of a previous year's effort on the current year's achievement is likely to underestimate the true causal effect of the current year's effort on the current year's achievement level. In any case, I still observe significant persistence in a specification in which all constructs are measured one year apart from one another. Those results are available upon request. Third, the assumption that the parameter values are fixed could be important. If the parameters are allowed to vary, it is possible that this additional heterogeneity would reduce the estimated levels of persistence, because the reinforcement processes may become relatively less salient. Of course, I have also only considered one particular reinforcement process in the empirical simulation. In principle there could be many others-again, for example, through peer effects or dynamic complementarity-both of which, I have shown, enhance persistence.

The parameter estimates are expected to retain some bias in all specifications. There are two main sources of bias. The first is bias from time-invariant confounders such as any innate or crystalized personality trait that is correlated with both the regressors and the dependent variables. Allowing random intercepts both proxy for some of these confounders, but they are not perfect proxies. It would be more ideal to use person-fixed effects with many time periods. Unfortunately, there are only four time periods and

even fewer for the regression of academic self concept, which relies on the previous time period. Fixed effects regressions are biased and inconsistent for specifications with few time periods and large cross-sectional sample sizes.

The second main source of bias is time-varying confounding which violates sequential exogeneity. This can occur when a treatment causes an outcome that itself reinforces the treatment at a later time point. For example, a regression of effort on academic self-concept could be biased because the treatment (academic self-concept) could on one hand cause effort in the same period, but also be caused by a previous effort level through the previous achievement level. Wodtke [93] proposes a regression-based method to control for time-varying confounding under certain assumptions, but that method doesn't work for logistic regression. There is another method that can take steps towards reducing time-varying confounding even for nonlinear regressions, but little has been proven about its properties. Another way to control for time-varying confounding is to simultaneously estimate the entire model and include regressions that control for direct effects of all past observables. However, this approach suffers from a "bad control" problem, in that it requires one to control directly for outcomes caused by treatment in previous periods. There are also other ways to deal with time-varying and time-invariant confounding with dynamic panel models that rely on instruments, but these methods generally require more than the number of time periods I am using in this data set for decent results. I still plan to explore these approaches in the future, but the purpose of the empirical section of this paper was only to generate plausible parameter estimates through a calibration exercise, to be used in a suggestive counterfactual simulation.

There are many mechanisms that govern student achievement, academic self-concept, and effort that are omitted from the theoretical model, including but not limited to teachers and parents-both of which I am working towards as a next step (in fact I focus on dynamics between school officials and students in another working paper focused

on reinforcement processes in school disciplinary policy). Other omitted components include but are not limited to socioeconomic considerations such as neighborhood context. These contextual considerations are only omitted from the model to ensure tractability and clarity of focus. Of course, reinforcement processes are rampant throughout social systems. Math anxious parents reproduce math-anxious students which become math-anxious parents [59]. Racial bias from teachers can increase students’ apprehension about bias, leading to behaviors that reinforce teachers’ racial bias [70]. Early concentrated disadvantage lowers educational attainment and reinforces concentrated disadvantage [16] [17]. As we see the prevalence of reinforcement processes as a key driving force behind what sociologists refer to as “systemic racism,” we also see that the very mechanisms used to sustain it can be targeted to reverse it with persistence given the arguments of this paper. I plan to explore this theme further in future work, revealing how public policy can reduce systemic disadvantage using systems thinking.

4.6 Appendix

4.6.1 *Simplified Model*

Here I present the full derivation of the Simplified Model. Consider a school with N students indexed by i . Over an infinite time horizon indexed discretely by t , each student, in each period, chooses an effort level $e_{it} \in \{-1, 1\}$ (This is only for mathematical convenience. In the empirical exercise, it is redefined as $e_{it} \in \{0, 1\}$). This choice is based on a single measure S_{it} of a preconceived set of affective entities that determines the “academic self-concept” of the student, including how the student identifies with the subject in question (e.g. “I am good at math” or “I am a math person”). The choice also depends on the decisions made by peers via the perceived average choice level of peers, $\bar{E}_t^e$ and on perceived peer attitudes towards the subject as well S_t^e . Finally, the choice contains a component ϵ that is random and Type I extreme value-distributed. This de-

cision then impacts the achievement level at the end of the period, which determines the academic self-concept used in the next period.

To illustrate this I define a utility function $V(e_{it})$ such that:

$$V(1) - V(-1) := d_0 + d_1 S_{it} + d_2 \bar{E}_t^e + d_3 \bar{S}_t^e + \epsilon(1) - \epsilon(-1)$$

Parameters d_1 , d_2 , and d_3 capture the sensitivity of student i 's choice to her own academic self concept, her perception of her peers' average choice, and her perception of her peer's academic self-concept at time t , respectively. As in the discrete choice literature, $\epsilon(-1) - \epsilon(1)$ is logistically distributed and a student chooses effort level $e_{it} = 1$ if and only if $V(1) - V(-1) > 0$. It follows that

$$Pr[e_{it} = 1] = \frac{1}{1 + \exp[-\beta_1(d_0 + d_1 S_{it} + d_2 \bar{E}_t^e + d_3 \bar{S}_t^e)]}$$

where β_1 is a parameter characterizing the thinness of the tails of $\epsilon(-1) - \epsilon(1)$ (hence the relative “unimportance” of random heterogeneity in effort decisions).

The key simplifying assumption in this version of the model is that academic self-concept S_{it} depends directly on the student's achievement in the previous time period, A_{it-1} . Specifically,

$$S_{it} := b_1 A_{it-1}$$

where b_1 is the impact of the previous-time achievement level on self-concept in the current time period, and $A_{it-1} \in \{-1, 1\}$ is the achievement level for student i in time $t - 1$. Hence S_{it} only takes on one of two values in any time period: $S_{it} \in \{S^*, S_*\}$, where $S^* := b_1$ and $S_* := -b_1$. The full version of the model will allow for the inclusion of a stochastic component, dependence on other previous achievement levels and dependence on past academic self-concept levels. We begin with this simplified version of the

academic self-concept updating process.

The achievement level at time t is not realized until the end of period t . At the beginning of the period, the self-concept is updated based on the previous $(t-1)$ achievement level, and among other considerations this updated self-concept is used to make an effort decision at time t . Finally, the self-concept and effort levels at time t impact the achievement level A_{it} for student i at time t . Define $\tilde{e}_{it} := \frac{1+e_{it}}{2}$ and $\tilde{S}_{it} := \frac{S^*+S_{it}}{2}$ to be rescaled versions of e_{it} , A_{it} , and S_{it} that vary from 0 to 1 (or from 0 to S^* in the case of $\tilde{S}_{it}$). Let A_{it}^* be a latent variable such that $A_{it} = 1 \iff y_{it} > 0$ and $A_{it} = -1$ otherwise. In particular, let

$$A_{it}^* := a_0 + a_1 e_{it} + a_2 \tilde{e}_{it} \tilde{S}_{it} + a_3 S_{it} + a_5 A_t + \epsilon_{2it}$$

where τ_{it} is a logistically distributed idiosyncratic component impacting achievement for student i at time t . Parameters a_1, a_2, a_3, a_5 represent the direct impact of effort on achievement, complementarity between effort and self-concept on achievement, and the direct impact of self-concept on achievement, respectively. This illustrates several findings from the literature: that effort positively impacts achievement ($a_1 > 0$), that factors such as decreased anxiety and identifying with a subject makes effort more productive ($a_2 > 0$), and that peer achievement increases a student's achievement ($a_5 > 0$). We will see that some of these factors can lead to developmental trajectories that exhibit properties of complex systems. Note that a_4 has been omitted in this simplification.

Finally, this implies that

$$Pr[A_{it} = 1] = \frac{1}{1 + \exp[-\beta_2(a_0 + a_1 e_{it} + a_2 \tilde{e}_{it} \tilde{S}_{it} + a_3 S_{it} + a_5 A_t)]}$$

where β_2 parameterizes the thinness of the tails of the distribution of ϵ_{2it} .

4.6.2 Derivation of Dynamical System for Simplified Model

In this section I solve for the equilibria of the simplified model. First, by assumption $Pr[S_{it} = S^*] = Pr[A_{it-1} = 1]$. Although there are three important variables in the system, it suffices to solve for $Pr[e_{it} = 1]$ and $Pr[A_{it} = 1]$ (more on this later). I will do this by conditioning and finding Baye's Nash equilibria. Finally I will transform the discrete time system into a continuous time analogue, leading to a system of differential equations. Using that system I conduct a bifurcation analysis exploring the role of each parameter in equilibrium behavior.

Using the Law of Total Probability with Baye's Rule and the fact that $Pr[S_{it} = S^*] = Pr[A_{it-1} = 1]$, we arrive at the following equation:

$$\begin{aligned}
& Prob[A_{it} = 1] \\
&= Prob[A_{it} = 1|S_{it} = S^*, e_{it} = 1]Prob[e_{it} = 1|S_{it} = S^*]Prob[A_{it-1} = 1] \\
&+ Prob[A_{it} = 1|S_{it} = S^*, e_{it} = -1](1 - Prob[e_{it} = 1|S_{it} = S^*])Prob[A_{it-1} = 1] \\
&+ Prob[A_{it} = 1|S_{it} = S_*, e_{it} = 1]Prob[e_{it} = 1|S_{it} = S_*](1 - Prob[A_{it-1} = 1]) \\
&+ Prob[A_{it} = 1|S_{it} = S_*, e_{it} = -1](1 - Prob[e_{it} = 1|S_{it} = S_*])(1 - Prob[A_{it-1} = 1])
\end{aligned}$$

Using the formulations in the previous sections, note the following equalities:

- $Prob[A_{it} = 1|S_{it} = S^*, e_{it} = 1] = \frac{1}{1 + \exp[-\beta_2(a_0 + a_1 + a_2 S^* + a_3 S^* + a_5 \bar{A}_t^e)]}$
- $Prob[A_{it} = 1|S_{it} = S^*, e_{it} = -1] = \frac{1}{1 + \exp[-\beta_2(a_0 - a_1 + a_3 S^* + a_5 \bar{A}_t^e)]}$
- $Prob[A_{it} = 1|S_{it} = S_*, e_{it} = 1] = \frac{1}{1 + \exp[-\beta_2(a_0 + a_1 + a_3 S_* + a_5 \bar{A}_t^e)]}$
- $Prob[A_{it} = 1|S_{it} = S_*, e_{it} = -1] = \frac{1}{1 + \exp[-\beta_2(a_0 - a_1 + a_3 S_* + a_5 \bar{A}_t^e)]}$
- $Prob[e_{it} = 1|S_{it} = S^*] = \frac{1}{1 + \exp[-\beta_1(d_0 + d_1 S^* + d_1 \bar{E}_t^e + d_2 \bar{S}_t^e)]}$
- $Prob[e_{it} = 1|S_{it} = S_*] = \frac{1}{1 + \exp[-\beta_1(d_0 + d_1 S_* + d_1 \bar{E}_t^e + d_2 \bar{S}_t^e)]}$

Combining equalities 1-4 allows one to solve the first equation for $Pr[A_{it} = 1]$ in terms of the model primitives and $Pr[A_{it-1} = 1]$. Finally, again using the Law of Total Probability, we obtain

$$\begin{aligned} & Pr[e_{it} = 1] \\ &= Prob[e_{it} = 1|S_{it} = S^*]Pr[A_{it-1} = 1] + Prob[e_{it} = 1|S_{it} = S_*](1 - Pr[A_{it-1} = 1]) \end{aligned}$$

Next I invoke the concept of symmetric Baye's Nash equilibrium. Specifically, I require that in equilibrium, the beliefs of all agents are consistent with the dynamics of the game. Hence, student i 's beliefs about the average choice, self-concept, and achievement level are equal to the actual respective levels from the previous time period: $\bar{S}_t^e = E[S_{it-1}]$, $\bar{E}_t^e = E[e_{it-1}]$, and $\bar{A}_t^e = E[A_{it-1}]$. For convenience, I hereafter defer to the notation $E_t := E[e_{it}]$, $S_t := E[S_{it}]$, and $A_t := E[A_{it}]$.

The result is a coherent system of difference equations. Instead, however, I convert the system to continuous time by doing the following:

- Replace S_t , E_t and A_t with continuous time-dependent variables $S(t)$, $E(t)$, and $A(t)$.
- Similarly, replace $t - 1$ subscripts with $t - \Delta t$, with Δt representing an infinitesimal change in time.
- Replace $Prob[S_{it} = 1]$, $Prob[e_{it} = 1]$, and $Prob[A_{it} = 1]$ with continuous time-dependent variables $P_S(t)$, $P_e(t)$, and $P_A(t)$, noting they are the same for all students i .
- Note by definition that $P_e(t) = \frac{1+E(t)}{2}$
- Note by assumption that $S(t) = b_1 A(t - \Delta t) = b_1(2P_A(t - \Delta t) - 1)$. Therefore $S(t - 1) = b_1 A(t - 2\Delta t) = b_1(2P_A(t - 2\Delta t) - 1)$.

$$\begin{aligned}
& Prob[A_{it} = 1] \\
&= Prob[A_{it} = 1 | S_{it} = S^*, e_{it} = 1] Prob[e_{it} = 1 | S_{it} = S^*] Prob[A_{it-1} = 1] \\
&+ Prob[A_{it} = 1 | S_{it} = S^*, e_{it} = -1] (1 - Prob[e_{it} = 1 | S_{it} = S^*]) Prob[A_{it-1} = 1] \\
&+ Prob[A_{it} = 1 | S_{it} = S_*, e_{it} = 1] Prob[e_{it} = 1 | S_{it} = S_*] (1 - Prob[A_{it-1} = 1]) \\
&+ Prob[A_{it} = 1 | S_{it} = S_*, e_{it} = -1] (1 - Prob[e_{it} = 1 | S_{it} = S_*]) (1 - Prob[A_{it-1} = 1])
\end{aligned}$$

Altogether,

$$\begin{aligned}
P_e(t) &= \frac{P_A(t - 2\Delta t)}{1 + \exp[-\beta_1(d_0 + d_1 S^* + d_2(2P_e(t - \Delta t) - 1) + d_3 b_1(2P_A(t - 2\Delta t) - 1))]} \\
&+ \frac{(1 - P_A(t - 2\Delta t))}{1 + \exp[-\beta_1(d_0 + d_1 S_* + d_2(2P_e(t - \Delta t) - 1) + d_3 b_1(2P_A(t - 2\Delta t) - 1))]}
\end{aligned}$$

Define the function $F(S)$ so that

$$F(S) := \frac{1}{1 + \exp[-\beta_1(d_0 + d_1 S + d_2(2P_e(t - \Delta t) - 1) + d_3 b_1(2P_A(t - 2\Delta t) - 1))]}$$

Now taking the limit we recover a differential equation describing the evolution of the probability of high effort:

$$\begin{aligned}
\dot{P}_e &= \lim_{\Delta t \rightarrow 0} \frac{P_e(t) - P_e(t - \Delta t)}{\Delta t} \\
&= \lim_{\Delta t \rightarrow 0} \frac{F(S^*)P_A(t - 2\Delta t) + F(S_*)(1 - P_A(t - 2\Delta t)) - P_e(t - \Delta t)}{\Delta t} \\
&= F(S^*)P_A + F(S_*)(1 - P_A) - P_e
\end{aligned}$$

Where $\dot{P}_e$ is the derivative of the probability of high effort with respect to time and

P_e and P_A are implicit functions of time. Similarly, define $G(e, S)$ to represent the probability of high achievement conditional on e and S , so that

$$G(e, S) := \frac{1}{1 + \exp[-\beta_2(a_0 + a_1e + a_2S + a_3\tilde{e}\tilde{S} + a_5(2P_A(t - \Delta t) - 1))]}$$

Now taking the limit we recover a differential equation describing the evolution of the probability of high achievement:

$$\begin{aligned}\dot{P}_A &= \lim_{\Delta t \rightarrow 0} \frac{P_A(t) - P_A(t - \Delta t)}{\Delta t} \\ &= (G(1, S^*)F(S^*) + G(-1, S^*)(1 - F(S^*)))P_A \\ &\quad + (G(1, S_*)F(S_*) + G(-1, S_*)(1 - F(S_*)))(1 - P_A) - P_A\end{aligned}$$

Where $\dot{P}_A$ is the derivative of the probability of high achievement with respect to time. Altogether we obtain the dynamical system

$$\begin{aligned}\dot{P}_e &= F(S^*)P_A + F(S_*)(1 - P_A) - P_e \\ \dot{P}_A &= (G(1, S^*)F(S^*) + G(-1, S^*)(1 - F(S^*)) - 1)P_A \\ &\quad + (G(1, S_*)F(S_*) + G(-1, S_*)(1 - F(S_*)))(1 - P_A)\end{aligned}$$

This system of differential equations captures the evolution of achievement and effort over time in the simplified model. This is the system that gives rise to the phase portraits presented in the paper. Note that there is no equation of motion for self-concept. That is because it is redundant given the assumption that the previous achievement level completely determines the subsequent self-concept level. In this simplified system, the probability of being of high self-concept would look the same as the probability of high achievement. This assumption is relaxed in the full model.

4.6.3 Appendix: Phase Portrait Analysis of Simplified Model

In this section I reveal theoretical results of a simple, analytically solvable version of the EASE model. Specifically, I illustrate how assumptions about the underlying peer effect parameters can lead to different long-run outcomes for classrooms or schools. The strength of the peer effect parameters combined with those that characterize the relationships between self-concept, effort, and achievement can cause the average trajectory of a group of students to be policy resistant even to dramatic policies, sensitive to small policies, or exude opposite effects under different initial conditions. These nonlinearities can help explain the persistence and fadeout of a wide range of interventions.

Let us begin with the simplifying assumption that the subsequent self-concept depends only on the achievement level of the previous period, so that $b_1 = 0$ and $\widehat{HA}_{it} = A_{it-1}$, which makes the model analytically solvable at the population level, as shown in the Appendix. That is, the equilibrium proportion of students who are high-achieving, retain a high academic self-concept, and choose high-effort over time can be calculated using clear vector fields starting from any initial condition. I will show how these population-level equilibria depend on the parameters of the model. This turns out to be important for understanding the early childhood intervention simulation.

Conditional on a set of parameter values, a population-level equilibrium is the long-run proportion of students who are high-achieving, exude a high self-concept, and choose high effort when those quantities start with a given initial condition. With the simplifying assumption that the subsequent self-concept depends only on the achievement level of the previous period, so that $b_1 = 0$ and $\widehat{HA}_{it} = A_{it-1}$, the model can generate one, three, or five equilibria at the population level. In other words, for a given set of parameter values, the model can generate 1) a single equilibrium regardless of initial conditions; 2) three equilibria, in which case different initial conditions can lead to different equilibria; and 3) five equilibria, in which case different initial conditions can lead to different equilibria.

If there are no peer effects of any kind, this simplified model always generates a single equilibrium, as shown in Figure 4.10. That is, holding the parameters constant, regardless of how much an intervention alters the initial proportion of students who retained high achievement, self-concept, or effort, the long-run levels of these quantities always flow to a single equilibrium. This would result in fadeout unless the underlying parameters changed. Figure 4.10 shows what is called a “phase portrait analysis,” which is a set of trajectories that each constitute a solution to the underlying dynamical system. The horizontal and vertical axes represent the proportion of students in a school exuding high achievement and high effort, respectively. Starting from any point in the plot, the arrows indicate the direction these proportions would take over time before the system reaches its equilibrium, represented by the green dot. This equilibrium is stable, in the sense that any perturbations away from the equilibrium fade out over time as the system returns to the equilibrium.

The reason the model is only generating a single equilibrium absent peer effects is that it is equivalent to a type of Markov chain. Although high effort tends to lead to high achievement, hence to high self-concept and therefore high effort, the system is still a non-reinforcing stochastic process. With some probability a student who has consistently achieved with high self-concept and effort for many time steps will eventually underachieve, and be very unlikely to attain a high self-concept, hence unlikely to put in high effort and likely underachieve in the next timestep. Even if the feedback between effort, self-concept, and achievement is especially strong, the student essentially oscillates between two three-dimensional states rarely yet equally and infinitely often. Crucial to this dynamic is the fact that past negative or positive experiences no longer matter once the current experience has been taken into account. This leads to situation in which the probability of revisiting a state is not continually accumulating with the number of times one has visited that state. Hence there is no genuine reinforcement process. Importantly,

once we relax this simplifying assumption, the model will no longer be solvable analytically, but it will produce multiple equilibria even without peer effects.

DEVELOPMENTAL TRAJECTORIES UNDER WEAK PEER EFFECTS

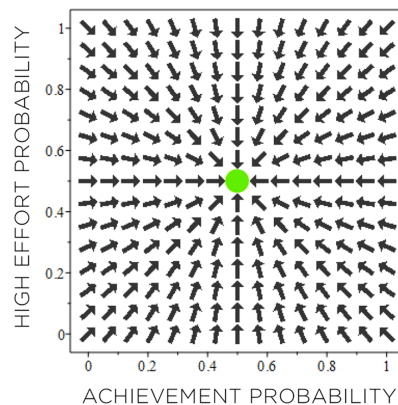


Figure 4.10: No Peer Effects: One Stable Equilibrium

When there are three equilibria in this model, as in Figure 4.11 there is an unstable saddle (the red dot) separating two stable equilibria (the green dots). There are many parameter combinations that could lead to this outcome, but there is always a requirement that one of the peer effect parameters be sufficiently strong. The unstable equilibrium is called a “saddle point” because it is stable along the vertical axis, but unstable along the horizontal axis. Suppose a school were currently in the stable low-achieving state. Holding the parameters constant, if an intervention were to “tip” the proportion of high-achieving students slightly to the right of the unstable equilibrium, trajectories would flow towards the high-achieving stable equilibrium. Such an intervention would appear to have large and persistent effects. That same intervention, were it to fail to change a critical proportion of students to the high-achieving state, would fade out over time. Even in this simplified model, the policy implications of systems thinking for persistence and fadeout become apparent.

DEVELOPMENTAL TRAJECTORIES UNDER Strong PEER EFFECTS

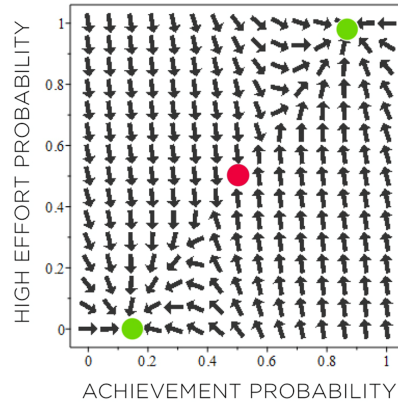


Figure 4.11: Strong Peer Effects: Three Stable Equilibria

Interestingly, there is a five-equilibria case, as shown in Figure 4.12. This case also requires sufficiently strong peer effects. There is another requirement: that the strength of the feedback between effort, achievement, and self-concept is neither too weak nor too strong. If such feedback $(a1, b1, d1)$ is too weak relative to peer effects, the system yields the three-equilibria case. If $a1, b1$, and $d1$ are too strong relative to peer effects, the system yields the single equilibrium case with very strong bottlenecks (meaning the time required to reach the equilibrium becomes very large).

DEVELOPMENTAL TRAJECTORIES UNDER MILD PEER EFFECTS

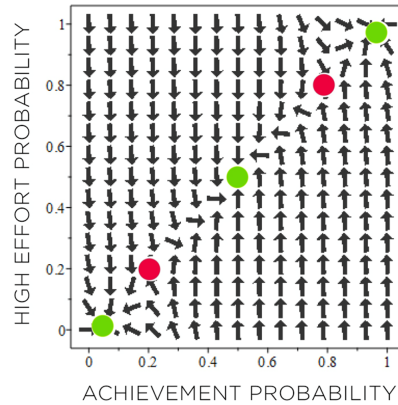


Figure 4.12: Mild Peer Effects: Five Stable Equilibria

For an intermediate range of a_1, b_1 , and d_1 , five equilibria are possible, with three stable equilibria separated by two unstable saddles. If the initial proportion of high-achieving, high-effort students lies above the right-most unstable saddle, peer effects are strong enough to lead the system to a high-achieving, high-effort equilibrium. If the initial proportion of high-achieving, high-effort students lies below the left-most unstable saddles, peer effects are strong enough to lead the system to a low-achieving, low-effort equilibrium. But if the initial proportion of high-achieving, high-effort students lies between the two unstable saddles, the net peer effects are relatively weak. In that case, all trajectories flow towards a stable equilibrium that acts as a “middle ground,” because students oscillate between high-effort/achievement and low effort/achievement, and there is never a sufficient quantity of high or low-effort students to make peer effects strong enough to “escape” the sink. Again, they oscillate because no matter the past, a student who has been high-achieving for some time stochastically becomes low achieving equally and infinitely often. More to the point of this paper, if treatment and control students were to mix together and the treatment were to fail to increase the whole population’s initial achievement level beyond some threshold (“tipping point”), the result would be fadeout. However, if the treatment and control students never mixed, as might be the case in a district-level early childhood education RCT, the same intervention with the

same initial effect size would appear to have persistent effects, because positive peer effects within the treated group sustain the initial effect on the treated, and peer effects in the control group may help sustain the disadvantages within the control group.

The simplifying assumption used in this section is unrealistic, of course. Students' developmental trajectories tend to depend on a full history of experiences. The purpose of this exercise is to show that peer effects alone can cause the model to exude properties of complex systems, using a simplified, analytically solvable version of the model. This fact remains after we return to the full model, allowing trajectories to depend on the history of achievement and self-concept levels, as shown in the body of the paper.

4.6.4 Descriptive Trends for Empirical Analysis

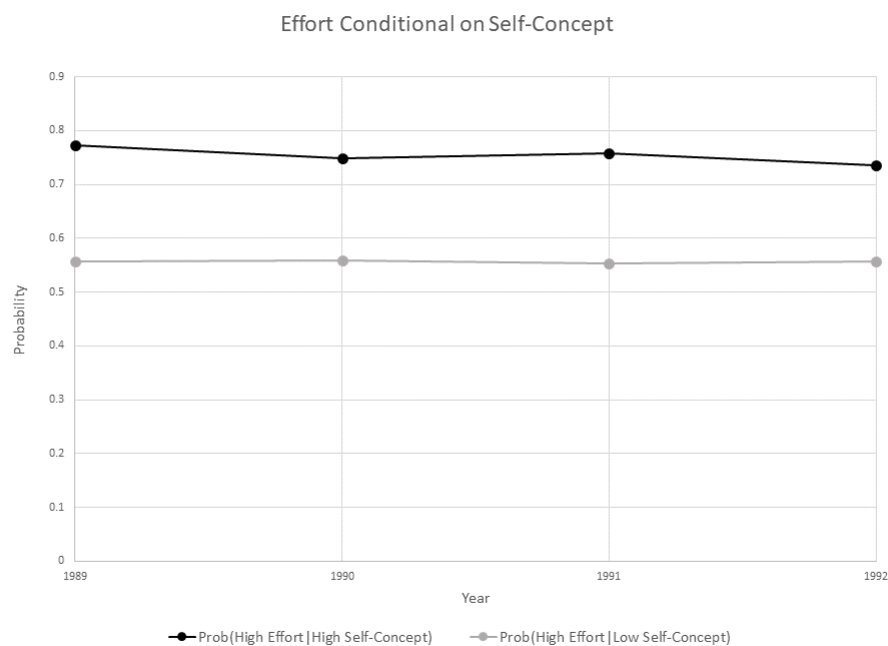


Figure 4.13: Self-Concept Predicts Effort

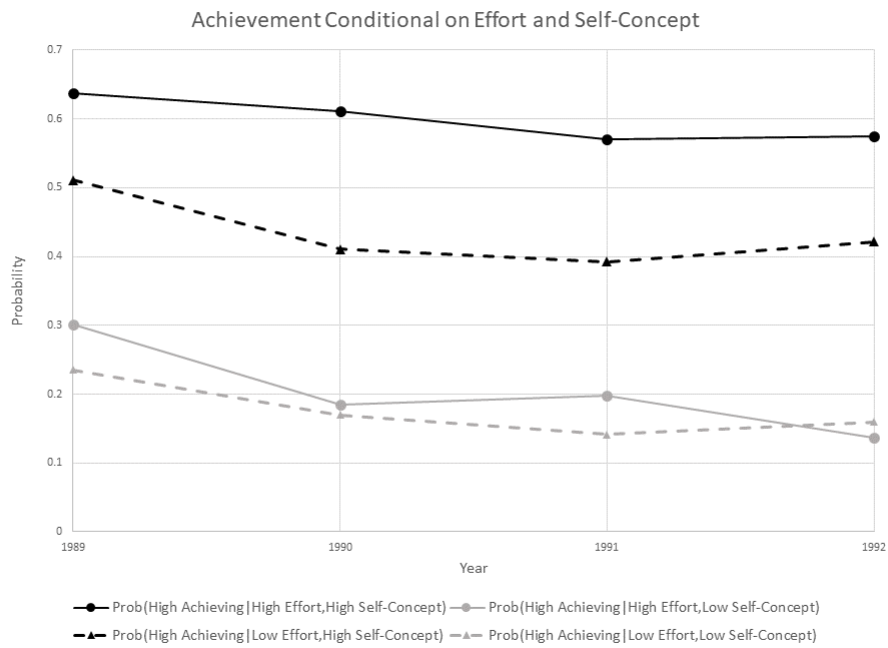


Figure 4.14: Effort and Self-Concept Predict Achievement



Figure 4.15: Current Achievement History Predicts Subsequent Self-Concept



Figure 4.16: Current Self-Concept History Predicts Subsequent Self-Concept

CHAPTER 5

CLOSING

“Everything we see is a shadow cast by that which we do not see.”

-MLK, Jr.

This dissertation shows how systems thinking can be used to fight against systemic problems in public policy. Substantively, this was demonstrated through three inter-related studies on the School-To-Prison Pipeline (STPP). Methodologically, a systems approach to studying the STPP was broadly operationalized through what economists call “applied theory:” the use of theoretical models built from empirically studied mechanisms to simulate policy counterfactuals within complex systems. Used as a complement (not a substitute) to reduced-form work that focuses on causal effects between isolated variables, this approach can be used for ex-ante policy evaluation, allowing research to lead policy and not just learn from policy consequences that harm disadvantaged groups after they happen in the real world. It helps us use the knowledge we have to foresee that which we do not otherwise see.

The first study takes a step towards learning the policy ramifications of state-dependent criminality-whereby the likelihood of offending depends on whether one has had previous experiences with offence and/or arrest. The study theoretically analyzes and empirically estimates a five-state Markov chain that captures the movement of youth in and out of states of criminal activity and arrest using the National Longitudinal Survey of Youth (NLSY97). The parameters of the model indicate common targets of intervention which are usually studied in isolation, such as prevention of criminal onset, arrest of first-time vs repeat-offenders, desistance of first-time vs. repeat offenders, and rehabilitation. The following conclusions are reached under the assumptions of the model. A sensitivity analysis reveals that overall, the equilibrium proportion of law-abiding citizens is most sensitive to independent changes in prevention of criminal onset, but also quite sensitive to changes in rehabilitation rates-especially for Black men. For all subgroups, independently increasing

arrest rates of repeat offenders appears to have no systemic effect, but increasing arrest rates of first-time offenders appears to be criminogenic for all subgroups. Finally, closing the racial gap in the likelihood of initial criminal onset through the first arrest would appear to have the largest impact on the racial disparity in the equilibrium proportion of law-abiding citizens, although this disparity is also quite sensitive to reductions in the racial difference in rehabilitation.

Motivated by the finding that equilibrium crime levels are systemically most sensitive to preventing the likelihood of initial criminal onset in the first place, the second study focuses on behavioral infractions in schools. It uses a complex mathematical model to simultaneously consider the mechanisms prioritized by alternative disciplinary philosophies, broadly referred to as “Zero-Tolerance” and “Restorative.” These mechanisms (such as deterrence, peer effects, how students see themselves with respect to behavior and how that influences behavior, and how punishments and rewards impact how students see themselves) are usually studied in isolation. The approach of this study allows us to foresee potential unintended consequences that wouldn’t be obvious were we to study these mechanisms in isolation. These consequences have important policy implications and could help explain the mixed empirical results on the impact of various disciplinary strategies. I find conditions under which the same school or classroom, under the same disciplinary policy, could end up in very different equilibrium levels of infractions depending on initial conditions. I find that changes in the extent of punishment for infractions or rewards for pro-social behavior can induce multiple “phase transitions,” such as destroying the possibility of a “bad” high-infraction equilibrium and making a high pro-social equilibrium stable. For example, three otherwise identical schools with the same punishment policy for behavioral infractions could end up in very different equilibrium levels of behavioral infractions and even with different sensitivities to initial conditions, depending on the extent to which they reward pro-social behavior. I also characterize how these kinds of nonlinearities depend on the parameters of the model, and conclude that

the required parameter relationships are more likely to hold in predominately African-American schools.

In the final study, I turn to underachievement-a key element in the STPP. I begin with a set of simple mechanisms which are believed to be important for underachievement, and which are normally studied in isolation (that effort, self-concept, and peers matter in the production of achievement; that self-concept matters in the production of effort; and that the self-concept stabilizes over time through repeated academic experiences). I show that a model built from these simple mechanisms generates emergent properties that wouldn't be obvious when studying the elements in isolation, and which have important policy consequences. These emergent properties, such as tipping points, phase transitions, and virtuous and vicious cycles, can help explain the mixed results of a wide range of famous educational interventions with regards to persistence and fadeout. This includes, but is not limited to: early childhood education interventions (fadeout vs persistence could be driven by the level of randomization when peer effects matter), self-affirmation (fadeout vs persistence can be driven by the strength of the feedback process between self-concept, effort, and achievement), growth mindset (fadeout vs persistence can be driven by the extent to which growth mindset lowers the sensitivity of self-concept to one's achievement history), and double-dose algebra interventions (fadeout vs. persistence can be driven by the extent to which the intervention makes later effort levels more productive). In each example, I use theoretical simulations to demonstrate how the model can qualitatively reproduce both fadeout and persistence depending on measurable parameters, and show persistence particularly for disadvantaged groups. Finally, I take a simple version of the model to data to simulate a self-affirmation intervention, estimating the parameters of the model in a calibration exercise. Based on the parameter estimates, I show that a simulated self-affirmation intervention implemented on the students in the Longitudinal Survey of American Youth (LSAY) generates moderate persistence because of the reinforcement processes operating in the system (e.g., academic self-concept reinforces itself

through effort and achievement). Taking advantage of these processes, made apparent through a systems perspective, could then be crucial to making persistent gains against the STPP and reducing educational inequities.

The notion of systems thinking is not new, although it has resurfaced in different methodological and substantive forms in different fields (e.g., “ecological systems theory;” “complex systems;” “management science;” “operations research”). The innovation described here lies in its application and operationalization within issues of systemic disadvantage, in such a way that connects to concrete policy solutions and ex-ante policy evaluation based on empirical studies. More than any specific outcome, the School-To-Prison Pipeline is a systemic problem. Much of the quantitative work on the STPP acknowledges this, yet proceeds to focus on isolated components in the hopes that policy makers will fit the pieces together, resulting in unintended consequences. Another common approach is purely theoretical, generating predictions from models which are too complex to connect with pragmatic policy decisions or clear empirically estimable parameters. In this dissertation, I choose a productive middle ground that is pragmatic and empirical enough for us to see what the systems perspective is contributing to public policy, but theoretical enough to foresee unintended consequences and look beyond the “shadows” in Dr. King’s famous quote.

This approach, which is a complement-not a substitute-for other methodologies, can be extended for the productive empirical study of systemic problems much more broadly. This includes the various forms and iterations of “systemic racism,” which is to date, poorly understood in economics, and poorly operationalized towards the end of quantifiable policy interventions in many other fields. My hope is that this contribution will replicate in others the same optimism I feel towards the power of systems thinking. . . the power to uproot systemic racism by ensuring that the impact of public policy lines up with the alleged righteous intent of policy makers.

REFERENCES

- [1] Michelle Alexander. *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press, 2020.
- [2] Paul D. Allison, Richard Williams, and Enrique Moral-Benito. Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, 2017.
- [3] Kaitlin P Anderson and Gary W Ritter. Do school discipline policies treat students fairly? evidence from arkansas. *Educational Policy*, 34(5):707–734, 2020.
- [4] Štěpán Bahník and Marek A. Vranka. Growth mindset is not associated with scholastic aptitude in a large sample of university applicants.
- [5] Drew Bailey, Greg J. Duncan, Candice L. Odgers, and Winnie Yu. persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, 2017.
- [6] Albert Bandura. Social cognitive theory: An agentic perspective. *Asian journal of social psychology*, 2(1):21–41, 1999.
- [7] G. Bear. Discipline: Effective school practices. *National Association of School Psychologists*, pages 1–5, 2010.
- [8] GS Becker. Crime and punishment: an economic approach. *Journal of Political Economy*, 76:169–217, 1968.
- [9] Alberto Bisin, Andrea Moro, and Giorgio Topa. The empirical content of models with multiple equilibria in economies with social interactions. Technical report, National Bureau of Economic Research, 2011.
- [10] A. Blumstein. On the racial disproportionality of united states’ prison populations. *Journal of Criminal Law and Criminology*, 1982.

- [11] Anthony A Braga, David Weisburd, and Brandon Turchan. Focused deterrence strategies and crime control: An updated systematic review and meta-analysis of the empirical evidence. *Criminology & Public Policy*, 17(1):205–250, 2018.
- [12] William A Brock and Steven N Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001.
- [13] Urie Bronfenbrenner. *Ecological systems theory*. Jessica Kingsley Publishers, 1992.
- [14] Leonardo Bursztyn and Robert Jensen. How does peer pressure affect educational investments? *The quarterly journal of economics*, 130(3):1329–1367, 2015.
- [15] Albert et al. Cheng. “no excuses” charter schools: A meta-analysis of the experimental evidence on student achievement. *Journal of School Choice*, 11(2):209–238, 2017.
- [16] Raj Chetty, Nathaniel Hendren, and Lawrence F Katz. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 2016.
- [17] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.
- [18] Geoffrey L. Cohen, Julio Garcia, Valerie Purdie-Vaughns, Nancy Apfel, and Patricia Brzustoski. recursive processes in self-affirmation: intervening to close the minority achievement gap. *science*, 2009.
- [19] David A Cole, Scott E Maxwell, Joan M Martin, Lachlan G Peeke, Alesha D Seroczynski, Jane M Tram, Kit B Hoffman, Mark D Ruiz, Farrah Jacquez, and Tracy Maschman. The development of multiple domains of child and adolescent self-concept: A cohort sequential longitudinal design. *Child development*, 2001.

- [20] Kalena E Cortes, Joshua S Goodman, and Takako Nomi. Intensive math instruction and educational attainment long-run impacts of double-dose algebra. *Journal of Human Resources*, 2015.
- [21] Flavio Cunha and James Heckman. The technology of skill formation. 2007.
- [22] James Earl Davis. Early schooling and academic achievement of african american males. *Urban education*, 38(5):515–537, 2003.
- [23] Thomas S Dee. Are there civic returns to education? *Journal of public economics*, 88(9-10):1697–1720, 2004.
- [24] Mesmin Destin, Paul Hanselman, Jenny Buontempo, Elizabeth Tipton, and David S Yeager. Do student mindsets differ by socioeconomic status and explain disparities in academic achievement in the united states? *AERA open*, 5(3), 2019.
- [25] Dieter Dölling, Horst Entorf, Dieter Hermann, and Thomas Rupp. Is deterrence effective? results of a meta-analysis of punishment. *European Journal on Criminal Policy and Research*, 15(1):201–224, 2009.
- [26] The Economist. America’s overcrowded prisons: one nation, behind bars, August 17, 2013.
- [27] M.W. Edelman. Losing our children in America’s cradle to prison pipeline. *New York: National Urban League*, 2007.
- [28] Card et al. Peer effects and multiple equilibria in the risky behavior of friends. *Review of Economics and Statistics*, 95(4):1130–1149, 2013.
- [29] Payne et al. Modeling the effects of racial threat on punitive and restorative school discipline practices. *Criminology*, 48(4):1019–1062, 2010.
- [30] Zimmerman et al. Do school disciplinary policies have positive social impacts? examining the attenuating effects of school policies on the relationship between personal and peer delinquency. *Journal of Criminal Justice*, 42(1):54–65, 2014.

- [31] Joe R. Feagin. *Systemic racism: a theory of oppression*. Routledge, 2006.
- [32] Margaret T Floress, Lyndsay N Jenkins, Wendy M Reinke, and Lorena McKown. General education teachers’ natural rates of praise: A preliminary investigation. *Behavioral Disorders*, 43(4):411–422, 2018.
- [33] American Psychological Association Zero Tolerance Task Force et al. Are zero tolerance policies effective in the schools?: an evidentiary review and recommendations. *The American Psychologist*, 63(9):852, 2008.
- [34] Trevor Fronius. Restorative justice in us schools: A research review. *Justice & Prevention Research Center*, pages 1–33, 2016.
- [35] Roland G Fryer Jr and Paul Torelli. An empirical analysis of ‘acting white’. *Journal of Public Economics*, 94(5-6):380–396, 2010.
- [36] Alejandro Gaviria and Steven Raphael. School-based peer effects and juvenile behavior. *Review of Economics and Statistics*, 83(2):257–268, 2001.
- [37] Jasmine Green, Gregory Arief D Liem, Andrew J Martin, Susan Colmar, Herbert W Marsh, and Dennis McInerney. Academic motivation, self-concept, engagement, and performance in high school: Key processes from a longitudinal perspective. *Journal of adolescence*, 35(5):1111–1122, 2012.
- [38] Anne et al. Gregory. Adolescent trust in teachers: Implications for behavior in the high school classroom. *School Psychology Review*, 37(3):337, 2008.
- [39] James J Heckman, John Eric Humphries, and Gregory Veramendi. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126(S1):S197–S246, 2018.
- [40] James J Heckman and Dimitriy V Masterov. The productivity argument for investing in young children. *Applied economic perspectives and policy*, 2007.

- [41] J.J. Heckman and P.A. LaFontaine. The american high school graduation rate: Trends and levels. *The Review of Economics and Statistics*, 92(2):244–262, April 2010.
- [42] Andreas Helmke and Marcel AG van Aken. The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study. *Journal of educational Psychology*, 87(4):624, 1995.
- [43] Paul J. Hirschfield. Schools and crime. *Annual Review of Criminology*, 1:149–169, 2018.
- [44] Rucker C. Johnson and C. Kirabo Jackson. Reducing inequality through dynamic complementarity: Evidence from head start and public school spending. 2019.
- [45] J. Kinsler. School discipline: A source or salve for the racial achievement gap? *International Economic Review*, 54(1):355–383, 2013.
- [46] S. Konopljova. *The role of discipline practices in student connectedness to school, bullying and victimization*. PhD thesis, University of British Columbia, 2017.
- [47] Tomoaki Kotera and Ananth Seshadri. Educational policy and intergenerational mobility. *Review of economic dynamics*, 25:187–207, 2017.
- [48] A. et al. Kupchik. School discipline and security: Fair for all students? *Youth & Society*, 39(4):549–574, 2008.
- [49] Gloria Ladson-Billings. From the achievement gap to the education debt: Understanding achievement in us schools. *Educational researcher*, 2006.
- [50] Gloria Ladson-Billings and William F Tate. Teachers College Press, 2006.
- [51] Kelly Ka Lai Lam and Mingming Zhou. Examining the relationship between grit and academic achievement within k-12 and higher education: A systematic review. *Psychology in the Schools*, 56(10):1654–1686, 2019.

- [52] Adam M Lavecchia, Heidi Liu, and Philip Oreopoulos. Behavioral economics of education: Progress and possibilities. In *Handbook of the Economics of Education*, pages 1–74. Elsevier, 2016.
- [53] Carol D. Lee. The centrality of culture to the scientific study of learning and development: How an ecological framework in education research facilitates civic responsibility. *Educational Researcher*, 2008.
- [54] Carl Lieberman. Variation in racial disparities in police use of force. 2020.
- [55] Jenny F. et al. Liljeberg. Poor school bonding and delinquency over time: Bidirectional effects and sex differences. *Journal of Adolescence*, 34(1):1–9, 2011.
- [56] L. Lochner and E. Moretti. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Association*, 94(1):155–189, March 2004.
- [57] Lance Lochner and Enrico Moretti. The effect of education on crime. *American economic review*, 2004.
- [58] C. Mallett, F. Miyuki, P. Stoddard-Dare, and L. Quinn. Significant race differences in factors related to the detention recidivism of youthful offenders. *Journal of Ethnicity in Criminal Justice*, 10:24–40, January 2012.
- [59] Erin A Maloney, Gerardo Ramirez, Elizabeth A Gunderson, Susan C Levine, and Sian L Beilock. Intergenerational effects of parents’ math anxiety on children’s math achievement and anxiety. *Psychological Science*, 2015.
- [60] Tony F Marshall. *Restorative justice: An overview*. 1998.
- [61] Alan McEvoy and Robert Welker. Antisocial behavior, academic failure, and school climate: A critical review. *Journal of Emotional and Behavioral disorders*, 8(3):130–140, 2000.

- [62] Kevin F McGrath and Penny Van Bergen. Who, when, why and to what end? students at risk of negative student–teacher relationships and their outcomes. *Educational Research Review*, 14:1–17, 2015.
- [63] D.M. McMillon, C.P. Simon, and J.M. Morenoff. Modeling the underlying dynamics of the spread of crime. *PLoS One*, 2014.
- [64] Kim CIM Megens and Frank M Weerman. The social transmission of delinquency: Effects of peer attitudes and behavior revisited. *Journal of Research in Crime and Delinquency*, 49(3):420–443, 2012.
- [65] Ojmarrh Mitchell and Michael S Caudy. Examining racial disparities in drug arrests. *Justice Quarterly*, 2015.
- [66] Howard S. Muscott. Positive behavioral interventions and supports in new hampshire: Effects of large-scale implementation of schoolwide positive behavior support on student discipline and academic achievement. *Journal of positive behavior interventions*, 10(3):190–205, 2008.
- [67] D. Nagin and R. Paternoster. Population heterogeneity and state dependence: State of the evidence and directions for future research. *Journal of Quantitative Criminology*, 16(2):117–144, 2000.
- [68] Daniel S. Nagin, Francis T. Cullen, and Cheryl Lero Jonson. imprisonment and reoffending. *Crime and Justice*, 2009.
- [69] Jason A Okonofua and Jennifer L Eberhardt. Two strikes: Race and the disciplining of young students. *Psychological science*, 26(5):617–624, 2015.
- [70] Jason A Okonofua, Gregory M Walton, and Jennifer L Eberhardt. A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, 11(3):381–398, 2016.

- [71] President’s Commission on Law Enforcement and Administration of Justice. *The Challenge of Crime in a Free Society*. United States Government Printing Office, 1967.
- [72] Debra Pepler, Depeng Jiang, Wendy Craig, and Jennifer Connolly. Developmental trajectories of bullying and associated factors. *Child development*, 79(2):325–338, 2008.
- [73] M. Perc, K. Donnay, and D. Helbing. Understanding recurrent crime as system-immanent collective behavior. *PLoS One*, 8:e76063, 2013.
- [74] Christopher Rauh. Voting, education, and the great gatsby curve. *Journal of Public Economics*, 146:1–14, 2017.
- [75] Sean F Reardon. The widening income achievement gap. *Educational Leadership*, 70(8):10–16, 2013.
- [76] M Marit Rehavi and Sonja B Starr. Racial disparity in federal criminal sentences. *Journal of Political Economy*, 2014.
- [77] Barbara Reskin. The race discrimination system. *Annual Review of Sociology*, 2012.
- [78] Barbara Reskin. The race discrimination system. *Annual Review of Sociology*, 38:17–35, 2012.
- [79] J. Rosenbaum. Educational and criminal justice outcomes 12 years after school suspension. *Youth & Society*, 2018.
- [80] Cody T Ross. A multi-level bayesian analysis of racial bias in police shootings at the county-level in the united states, 2011–2014. *PloS one*, 2015.
- [81] Martin D Ruck and Scot Wortley. Racial and ethnic minority high school students’ perceptions of school disciplinary practices: A look at some canadian findings. *Journal of Youth and Adolescence*, 31(3):185–195, 2002.

- [82] Jérémie Blanchette Sarrasin, Lucian Nenciovici, Lorie-Marlène Brault Foisy, Geneviève Allaire-Duquette, Martin Riopel, and Steve Masson. Effects of teaching the concept of neuroplasticity to induce a growth mindset on motivation, achievement, and brain activity: A meta-analysis. *Trends in neuroscience and education*, 2018.
- [83] Joanne Savage, Christopher J. Ferguson, and Lesli Flores. The effect of academic achievement on aggression and violent behavior: a meta-analysis. *Aggression and violent behavior*, 37:91–101, 2017.
- [84] Russel J Skiba. Zero tolerance, zero evidence: An analysis of school disciplinary practice. policy research report. *ERIC*, 2000.
- [85] James P Smith. The long-term economic impact of criminalization in american childhoods. *Crime & Delinquency*, 65(3):422–444, 2019.
- [86] Adriaan R Soetevent and Peter Kooreman. A discrete-choice model with social interactions: with an application to high school teen behavior. *Journal of Applied Econometrics*, 22(3):599–624, 2007.
- [87] Margaret Beale Spencer, Davido Dupree, and Tracey Hartmann. A phenomenological variant of ecological systems theory (pvest): A self-organization perspective in context. *Development and psychopathology*, 9(4):817–833, 1997.
- [88] Jantine L Spilt, Jan N Hughes, Jiun-Yu Wu, and Oi-Man Kwok. Dynamics of teacher–student relationships: Stability and change across elementary school and the influence on children’s academic success. *Child development*, 83(4):1180–1195, 2012.
- [89] Claude Steele. Stereotype threat and african american student achievement. *The inequality reader: Contemporary and foundational readings in race, class, and gender*, 2003.

- [90] Laurence Steinberg, Sandra Graham, Lia O’Brien, Jennifer Woolard, Elizabeth Cauffman, and Marie Banich. Age differences in future orientation and delay discounting. *Child development*, 80(1):28–44, 2009.
- [91] George Sugai and Robert Horner. The evolution of discipline practices: School-wide positive behavior supports. *Child & Family Behavior Therapy*, 24(1-2):23–50, 2002.
- [92] Dena Phillips Swanson, Margaret Beale Spencer, Vinay Harpalani, Tirzah R Spencer, et al. Identity processes and the positive youth development of african americans: An explanatory framework. *New Directions for Student Leadership*, 2002.
- [93] Geoffrey T. Wodtke. Regression-based adjustment for time-varying confounders. *Sociological Methods & Research*, 2018.
- [94] David S. Yeager. Loss of institutional trust among racial and ethnic minority adolescents: A consequence of procedural injustice and a cause of life-span outcomes. *Child Development*, 88(2):658–676, 2017.
- [95] david s yeager, paul hanselman, gregory m walton, jared s murray, robert crose, chandra muller, elizabeth tipton, barbara schneider, chris s hulleman, cintia p hinojosa, et al. A national experiment reveals where a growth mindset improves achievement. *Nature*, 2019.
- [96] Gema Zamarro, Collin Hitt, and Ildefonso Mendez. When students don’t care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, 13(4):519–552, 2019.