Looking Inside the Black Box: The Importance of Causal Mechanism and Treatment Effect Heterogeneity in Experimentally Evaluated Criminal Justice Interventions*

Chongmin Na**

Abstract: This paper discusses limitations of the "black-box" experimental archetype by highlighting the narrowness of outcome-focused approaches. For a more complete understanding of the nuanced implications of policies and programs, this study calls for an investigation of causal mechanism and treatment effect heterogeneity in experimentally evaluated interventions. This study draws on two distinct but closely related empirical studies, one undertaken by Na and Paternoster (2012) and the other by Na, Loughran, and Paternoster (2015), that go beyond the estimation of a population average treatment effect by adopting more recent methodological advancements that are still underappreciated and underutilized in evaluation research.

Keywords: Causal Mechanism, Treatment Effect Heterogeneity

INTRODUCTION

When assessing the effect of policy/program interventions, social scientists in multiple disciplines have given more credibility to the findings from randomized experiments because random assignment—if properly designed and implemented—rules out many rival explanations to the observed effect by creating an ideal counterfactual situation (Rubin, 1974). In noble pursuit of the unbiased estimate of treatment effect, however, researchers have paid relatively less attention to other substantial and

---

* This paper is based on two peer-reviewed journal articles the author published with his colleagues (Na & Paternoster, 2012; Na, Loughran, & Paternoster, 2015). An earlier draft of this paper was presented at the 2016 "Smart Governance" conference hosted by Lehman College-CUNY and the Graduate School of Public Administration at Seoul National University.

** Chongmin Na is an assistant professor in the Department of Criminal Justice at the John Jay College of Criminal Justice, City University of New York. E-mail: cna@jjay.cuny.edu.

Manuscript received February 10, 2016; out for review March 1, 2016; review completed March 25, 2016; accepted March 25, 2016.

The Korean Journal of Policy Studies, Vol. 31, No. 1 (2016), pp. 89-112.

methodological issues that might also be of great interest to both policy makers and researchers: identification of causal mechanism and treatment effect heterogeneity (Heckman & Smith, 1995; Imai, Tringley, & Yamanoto, 2013).

In addition to making a rigorous comparison of the average outcomes between treatment and control groups, disentangling causal mechanisms that link treatment to subsequent changes in the outcome and assessing potentially heterogeneous treatment effects within segments of a larger population are beneficial to policy design/implementation as well as theoretical development. By presenting a more complete picture of how each treatment functions in multiple domains over a relatively long period of time, such evaluation studies can provide scientific evidence regarding not just whether an intervention works but also provide answers to additional questions such as why, how, and for whom it works.

Largely due to the evidence-based movement that has emerged as a new paradigm for evaluation studies in the field of criminology and criminal justice (Sherman et al., 1998)—which requires high standards for policy evaluation so as to ensure that governments do not waste money on ineffective approaches in the face of populist pressure—there is a high demand for a way to test of policies with "well-controlled experiments before spending vast sums in the name of crime control/prevention" (Sherman, 2009, p. 7). In their report to the United States Congress, which was based on a systematic review of more than 500 scientific evaluations of crime prevention practices, Sherman et al. (1998) elegantly defined what works in crime prevention/control policies:

> Clear conclusions about what works and what doesn't require a high level of confidence in the research results. These are programs that we are reasonably certain of preventing crime or reducing risk factors for crime in the kinds of social contexts in which they have been evaluated, and for which the findings should be generalizable to similar settings in other places and times. Programs coded as "working" by this definition must have at least two level 3 evaluations with statistical significance tests showing effectiveness and the preponderance of all available evidence supporting the same conclusion.[1]

Obviously, Sherman et al.'s answer is admirably predicated on advocating for the use of social experiments and randomization to achieve high internal validity and reliability

1. Here, "level 3 evaluation" refers to the Maryland Report. Specifically, an evaluation is at least a level 3 if it includes "a comparison between two or more comparable units of analysis, one with and one without the program." A level 3 evaluation is more rigorous than levels 1 (correlation between program and outcome) or 2 (correlation with proper temporal sequencing) but not as rigorous as levels 4 (additional controls) or 5 (random assignment of the intervention).

of program effectiveness estimates. In this vein, Sherman (2009) and Weisburd (2010), among others, have strongly advocated for the use of randomization in evaluation studies.

Others, however, are more cautious in their enthusiasm for the ubiquity of experiments and critical of the narrowness of the "what works" framework in criminological research. For example, Sampson (2010) argues that experimental results can have certain inferential deficiencies, for instance, an absence of a theoretical basis for causal explanation. Although not directly contradicting Sherman et al.'s points, Heckman and Smith (1995) essentially argue against the narrowness of the "black-box" experimental archetype, which focuses on determining if a program has an overall positive mean impact at the expense of considering important and policy-relevant quantities. In a similar vein, the evidence-based paradigm has been criticized as being "extraordinarily conservative" in nature (Clear, 2010, p. 6) because it uses several experimental evaluations suggesting strong and positive effect sizes to justify a strong level of confidence in a program. Although the "what works" literature should be able to tell policy makers what deserves their support and the consumers of this evidence are willing to see unequivocal answers instead of modest or nonsignificant results, "most experimental criminologists fail to find the big effects that could make criminology in general, and experimental criminology in particular, more central to making policy" (Sherman, 2007, p. 300). Focusing merely on the mean effect aggravates this problem, further "shackling us," preventing researchers "from taking bold action" (Clear, 2010, p. 14).

Using two distinct but closely related articles that have been published in the field of criminology and criminal justice, Na and Paternoster's "Can Self-Control Change Substantially over Time?" (2012) and Na, Loughran, and Paternoster's "On the Importance of Treatment Effect Heterogeneity in Experimentally-Evaluated Criminal Justice Interventions (2015), I explore these issues. In these studies, the researchers go beyond the estimation of an average treatment effect in the general population, drawing on more recent methodological advancements that are still underappreciated and underutilized in the evaluation research, including the hierarchical linear model (HLM) (Raudenbush & Bryk, 2002), the structural equation model (SEM) (Bollen, 1989), the group-based trajectory model (GBTM) (Nagin, 2005), and the growth mixture model (GMM) (Muthén, 2001). I aim to demonstrate how randomized experiments can still be utilized as an important policy tool by exploring not just unbiased average treatment effects in the population but also the much broader and more detailed implications of interventions such as underlying causal mechanisms and potentially differential effects across some distinct segments of the population. All the examples draw on data on an experimental intervention from the Johns Hopkins Prevention Intervention Research Center (JHU PIRC).

RANDOMIZED EXPERIMENT AS A GOLD STANDARD IN THE EVALUATION RESEARCH

When assessing the impact of public policies/programs, we generally focus on comparing the outcomes under the intervention to what that outcome would have been without the intervention. Because subjects might self-select themselves into the treatment or control group based on their own anticipated response to the treatment, most evaluation research has focused primarily on the unbiased and consistent estimation of an average treatment effect in the population by rigorously accounting for the initial differences between comparison groups. Since the randomization process generates two equivalent groups that are similar in terms of most relevant characteristics (measured or not), evaluators can reasonably assume that the observed between-group difference in the outcome after treatment results primarily from the variation in the treatment condition. This potential outcome model of causality, originally demonstrated by Neyman (1923) and later developed into the Rubin's causal model (Rubin, 1974; 1977; 1978), is still considered to be the most powerful methodological way to identify unbiased causal effects in the evaluation research. Nonetheless, researchers have also recognized the limitations of such a "black-box" experimental archetype due to its inability to provide a more complete understanding of the impact of interventions (e.g., Angrist, 2004; Cook, 2002; Heckman, 1992; Heckman & Smith, 1995).

Looking Inside the Black Box

Causal Mechanisms

One important criticism of a mere comparison of the subsequent outcomes between treatment and control groups in randomized experiments is that it fails to identify the causal mechanisms underlying the intervention effects. To simply assume that any observed treatment effect arises through changes in some hypothesized mediating factors is less useful for theoretical sophistication and successful policy design/implementation (Heckman & Smith, 1995) than testing it explicitly by incorporating those mediators into the analytic model. Such a process-based approach focusing on the question of how and why better informs program developers and policy makers about the elements and conditions of the program that are related to the observed successful outcome, especially when there are multiple components in a universally implemented (vs. targeted) intervention. In addition, understanding the actual process in which any treatment affects an outcome can either confirm or dispute key theoretical propositions on which the intervention is grounded. Despite such theoretical and policy implications,

the importance of disentangling explicitly why and how the intervention works has been less appreciated in the evaluation research.

Treatment Effect Heterogeneity

When researchers attempt to identify the effect of any intervention, there is another source of bias besides selection bias that potentially threatens the validity of a clean causal inference: treatment effects may not be homogeneous but vary systematically across individuals or distinct subgroups within population. While experimental research provides an unbiased estimate of an average treatment effect through the power of randomization, it says little about the differential effects of the treatment on specific individuals or distinct subgroups that depart from the average. Indeed, if treatment effects are heterogeneous in the population, the estimators of treatment effects will vary regardless of confounding bias. While the utility of heterogeneous treatment effects has received particular attention in the medical sciences and biostatistics literatures (Kravitz, Duan, & Braslow, 2004), it has important policy relevance for social scientists as well. Simply evaluating an intervention by whether or not it on average produces an impact greater than zero in essence reduces it to an oversimplified and inefficient binary condition, that is, assesses it only in terms of whether it "works" or it does not. Heterogeneity of treatment effects acknowledges that the given treatment can have a discernible positive or beneficial effect on some respondents, a null or only weak effect on other respondents, and even a negative or harmful effect on still other respondents. Explicitly considering the broader impact of a program intervention beyond just its average improves the intervention's function as a policy tool, as it provides detailed information on what works for whom.

THEORETICAL FRAMEWORK: GOTTFREDSON AND HIRSCHI'S GENERAL THEORY OF CRIME

The current study is concerned specifically with investigating causal mechanism and treatment effect heterogeneity in experimentally evaluated interventions. The argument is twofold in nature: first, evaluation research should more explicitly acknowledge and address the limitations of outcome-focused approaches and of the common- effect assumption when assessing the treatment effect. Second, evaluation research should incorporate other innovative methodological approaches into the study of program effectiveness in order to go beyond a simple comparison of the average outcomes between study groups.

In the field of criminology and criminal justice, Gottfredson and Hirschi's (1990) general theory of crime has been very successful in generating empirical research over the past two decades largely because of its predictive ability (e.g., "trait-like" low self-control is the primary cause of criminal and analogous behaviors), its stability postulate (e.g., once the level of self-control is established in the early childhood as a result of effective socialization by primary caregivers, it remains relatively stable over time, not being influenced by subsequent social experiences and circumstances), and its generalizability (e.g., self-control has very general effect on not just crime but so many other behavioral outcomes, across all times and places). In particular, considering that children exhibiting antisocial propensities become increasingly resistant to change over the course of their lives, many scholars and practitioners have argued for the cost effectiveness of prevention/intervention programs targeting high-risk children and their families. It should be noted that if the theory's stability postulate holds, any effort aimed at the improving self-control after the formative period of early childhood would be less cost effective or even wholly ineffective.

Despite the substantive implications for theory and policy, empirical scrutiny of the nondeterminate nature and role of self-control has been rare and limited in scope. Although most studies support the central proposition of Gottfredson and Hirschi's theory—low self-control is the one of the strongest and significant correlates of crime and deviance (see Pratt & Cullen, 2000), little is known about how self-control develops over a longer period of time. Indeed, research is surprisingly limited with respect to explicitly isolating the causal mechanisms by which interventions targeting self-control reduce violence and crime.

In addition, Gottfredson and Hirschi claim that their theory is a "general" theory of crime, and they remain strongly opposed to offender taxonomies. They make a prediction about a more uniform developmental commonality among individuals in the population. Thus, they never predict the existence of subgroups within a more general population that manifest distinct developmental patterns with inherently different etiological implications. In particular, they never predict a decreasing pattern of self-control or a reshuffling of its trajectory across individuals over time, because "socialization continues to occur throughout life" and "differences between people in the likelihood that they will commit criminal acts persist over time" (1990, p. 107). Thus, self-control should continue to increase over time for everyone and the level of self-control in one individual relative to another should remain stable over time. Accordingly, most theory testing and the evaluation of programs that were designed to improve self-control have involved the comparison of the average outcomes between study groups under the "common-effect" assumption that the program affects every individual in the same way. However, what if there are indeed some segments of the population that manifest

etiologically distinct patterns in the development of self-control? What if these individuals respond to the same interventions in different ways? Indeed, limited but growing research suggests that there are some distinct clusters of individuals that follow non-normative trajectories (e.g., decreasing self-control over time) and shift in the level of self-control in an individual relative to others is an empirical regularity rather than an exception across different study samples (Burt, Sweeten, & Simons, 2014; Hay & Forrest, 2006; Na, Loughran, & Paternoster, 2015).

Data

The data used in Na and Paternoster's and Na, Loughran, and Paternoster's studies come from a second generation of the JHU PIRC's field trials, which featured both classroom-centered and family-school partnership interventions directed at improving school achievement and reducing conduct problems among low-income, high-risk youth in Baltimore, MD. The intervention design involved 678 first-graders and their families recruited from 27 classrooms in 9 Baltimore City public elementary schools that were followed up to the 12th grade. Of these 678 children, 53.2% were male, 86.8% were African American, and 63.4% were on free or reduced-cost lunch. At the entrance into first grade in 1993, the age of the children ranged from 5.3 to 7.7 years with a mean age of 6.2 years (SD = .34). A randomized block design was employed, with schools serving as the blocking factor. Three first grade classrooms in each of 9 elementary schools were randomly assigned to one of the two intervention conditions or to a control condition. The studies in question focus exclusively on the family-school partnership intervention, which was designed to improve self-control through enhancement of the social bond between caregivers and children. Accordingly, the analyses examined only 448 individuals assigned to either the family-school partnership intervention or control condition after excluding those who participated in the classroom-based intervention. The analyses in these studies were conducted using the data from grade 6-12 that focuses solely on the long-term effect of an early intervention program administered during the first grade.

Measurement

Independent Variable: Family-School Partnership Intervention

The family-school partnership intervention was designed to provide parents with effective teaching and child behavior management strategies via a series of workshops led by the child's first grade teacher and school psychologist or social worker. The

caregivers in the treatment group were instructed in interventions to help with discipline, strategies such as monitoring conduct, recognizing bad conduct, and properly dealing with bad behavior—all of which are crucial in developing child's self-control according to Gottfredson and Hirschi. In addition, caregivers also learned how to strengthen their bond with their children and become more involved in their children's lives, the goal being to enhance parent-teacher communication and thereby improve the academic achievement and behavioral outcomes of their children.

Dependent Variable: Self-Control

Considering that the preventive intervention was designed to improve the self-control of the students and so lead to changes in physical, mental, and behavioral outcomes, the immediate outcome of primary interest in this study is self-control. There were five domains of self-control assessed in the Teacher Report of Classroom Behavior Checklist (TRCBC), including accepting authority (the inability to accept authority manifests as conduct problems and oppositional defiant behavior), social participation (failure to participate takes the form of shy or withdrawn behavior), self-regulation (lack of self-regulation manifests as impulsivity), motor control (poor motor control is reflected in hyperactivity), concentration (inability to concentrate manifests as inattention), and peer likeability (failure to be liked by peers results in rejection). Given that a common set of items/indicators is necessary in studies of developmental pattern over time, the TRCBC items for grades 6 to 12 have remained constant over the course of the study. Five subscales of self-control created by JHU PIRC have strong face validity because they capture the behavioral manifestations of some combination of the defining elements of self-control in Gottfredson and Hirschi's (1990) original theory. The coefficient alphas for these measures in grades 6-12 ranged from .65 to .79 for impulsivity, from .76 to .88 for hyperactivity, from .90 to .93 for inattention, from .87 to 93 for oppositional-defiant behavior, and from .83 to .86 for socially withdrawn behavior. For the analyses undertaken by Na and Paternoster and by Na, Loughran, and Paternoster, composites of the multiple items were created to model the growth of the latent constructs by taking the mean of the scale's items (see Na & Paternoster, 2012, for more details).

Mediating Variable: Social Bond

The extent of the social bond between parents and children was measured with the Structured Interview of Parent Management Skills and Practices (SIPMSP). SIMPSP includes questions about parent disciplinary practices and practices associated with the

development of antisocial behavior. The relevant parental disciplinary practice constructs are parental monitoring, discipline, reinforcement, rejection, and problem solving. In collaboration with the Oregon Social Learning Center, JHU PIRC modified the SIMPSP to include items that assess parent-teacher communication and involvement and support for the child's academic achievement. Using extant theories and research (Hirschi, 2004; Tittle, Ward, & Grasmick, 2004; Hay & Forrest, 2006), researchers at JHU PIRC created five subscales that represent the key elements of the caregiver-child attachment component of the social bond that function as a source of self-control: monitoring, punishment, attachment, involvement, and support (see Na & Paternoster, 2012, for more details). A summated scale of the social bond was created for each grade from grade 6 to grade 11. The coefficient alphas for the subscales ranged from .25 to .67 for monitoring, from .75 to .80 for punishment, from .59 to .85 for attachment, from .33 to .59 for involvement, and from .50 to .72 for support.

Table 1. Descriptive Statistics for Key Variables

Variable	n	Min.	Max.	Mean	S.D.
treatment (control = 0, treatment = 1)	448	.00	1.00	.5112	.50043
male (female = 0, male = 1)	448	.00	1.00	.5201	.50015
black (white = 0, black = 1)	448	.00	1.00	.8571	.35032
self-control (grade 6)	339	.53	4.97	3.4261	.93981
self-control (grade 7)	340	.83	5.00	3.5077	.87617
self-control (grade 8)	348	.82	5.00	3.4819	.89670
self-control (grade 9)	329	.63	4.98	3.4962	.88585
self-control (grade 10)	308	1.01	5.00	3.6049	.79090
self-control (grade 11)	260	1.45	5.00	3.6806	.77691
self-control (grade 12)	289	.50	5.00	3.8047	.76703
social control (grade 6)	342	2.76	4.88	4.1286	.41046
social control (grade 7)	354	2.77	4.90	4.0436	.41154
social control (grade 8)	356	2.22	4.85	3.9893	.43856
social control (grade 9)	349	2.21	4.96	3.9225	.47536
social control (grade 10)	320	2.17	4.70	3.8280	.51412
social control (grade 11)	319	2.29	4.83	3.6098	.48178

Analytic Strategy

The investigation of patterns among and sources of change in self-control trajectories over time and the roles played by key mediating factors through an ongoing process of dynamic interaction requires longitudinal panel data capturing the within-individual changes of key variables over multiple time points. While there has been a growth of interest in using the appropriate statistical methods to describe and explain individual trajectories of interest, these analyses require making decisions about appropriate statistical models to be employed. Some of the modern approaches that are gaining popularity for modeling longitudinal panel data include, as already noted, HLM, GBTM, and GMM. Each of these different approaches has strengths and weaknesses depending on the particular research topics and contexts, but they all attempt to describe and explain population variation in developmental trajectories by relying on inherently different assumptions about the distribution of trajectory variation in the population than more traditional modeling strategies. While a decision as to how to model this variation should be made based on a priori justification that is primarily guided by theoretical rationales rather than data-driven approaches (Nagin and Piquero, 2010, p. 109; Sampson and Laub, 2005, p. 911), here I explore more than one final model and discuss explicitly how results from different modeling alternatives can be combined in order to better understand the nature and cause of the variation in the outcome profile. Through the progression from conventional HLM to more complex GMM approaches—primarily driven or supported by strong theoretical and empirical justifications—this study illustrates what additional insights can be gained for theory and policy as well as how the selection of alternative models can be carried out in a confirmatory rather than purely exploratory manner.

Population Average Treatment Effect and Causal Mechanism

Analytic Model: HLM

When assessing the long-term effects of a preventive intervention in a longitudinal randomized evaluation in which subjects are randomized into treatment conditions and measured repeatedly over time, longitudinal panel data allow for the modeling of intervention effects on developmental trajectories (e.g., growth rate) of an outcome rather than its between-group differences at a specific time point. As discussed in the previous section, Gottfredson and Hirschi's (1990) theory is claimed to be a 'general' theory of crime that makes a prediction about a more uniform developmental commonality in the population. In this vein, the 2012 study by Na and Paternoster study

employs a HLM approach that assumes that all subjects in the population follow a similar pattern in the growth of self-control, which develops according to a common functional form, although the growth parameters may vary in their magnitude across individuals. The level 1 model of HLM can be specified as follows:

$$selfcontrol_{ti} = \eta_{0i} + \eta_{1i} Grade_{ti} + e_{ti}$$

And the level-2 model can be represented as follows:

$$\eta_{0i} = \alpha_0 + \zeta_{0i}$$
$$\eta_{1i} = \alpha_1 + \zeta_{1i}$$

The treatment status as a time-invariant covariate can be incorporated into the base model (at level 2) to explain the variation in the growth parameters or, more specifically, to compare the overall patterns of self-control development between treatment and control group members. In addition, the individual's social bond as a time-varying covariate can be added to the base model (at level 1) to make it possible to determine whether any observed between-group difference in the developmental patterns is accounted for—at least in part—by the changing level of social bond.

After determining the causal impact of the intervention on the pattern of change in self-control and assessing the role of social bond as a potential mediator of such relationship, Na and Paternoster attempt to directly examine the longitudinal relationship between self-control and social bond over a relatively long period of time. After first building two latent constructs, one for self-control and another for social bond, they employ a longitudinal SEM with a panel design to explore if there is a time-lagged bidirectional relationship between self and social control over time. In particular, they directly compare the unidirectional model (drawing on the self-selection postulates of the Gottfredson and Hirschi's theory) to the bidirectional model (drawing on the mixed theories ofself-selection and social-causation frameworks; see, e.g., Wikström, 2004) in terms of various fit indices and the significance and magnitude of parameter estimates in order to assess which model fits the data better.

Results

Although more complicated functional forms of a model better capture meaningful patterns of variation, a simpler functional form can still provide an easy to understand, good approximation of the general pattern of growth trajectories of interest. Considering the primary goal of Na and Paternoster's study was to investigate different *rates* of

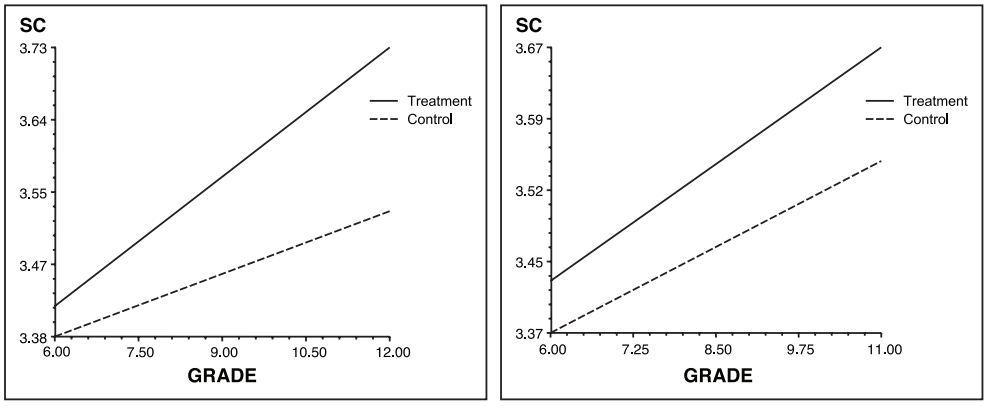
change between two study groups with a common functional form, a simplified model with only a linear growth parameter was adequate.

Consistent with the continued socialization postulate of Gottfredson and Hirschi's theory, the fixed-effects results in table 2 suggest that children have a self-control score of 3.16 points on average at grade 6 and that the level of self-control increases on average by about .04 points with each increasing grade. However, the random effects results indicate that linear growth rates significantly vary across individuals ($p < .001$), which does not conform to the relative stability postulate of the theory. In addition, the significant negative correlation between intercept and slope parameters (-.856, $p < .001$) suggests that individuals with relatively lower levels of self-control tend to gain it at a faster rate than their counterparts, which opens up the possibility of reshuffling of individual trajectories over time.

Table 2. Fixed and Random Effects of Growth Parameters in HLM

Fixed Effect	Coefficient	S.E.	*t*-ratio	*d.f.*	*p*-value
For π_0 β_{00}	3.159790	0.083842	37.687	398	<0.001
For π_1 β_{10}	0.039591	0.007992	4.954	398	<0.001
Random Effect	S.D	Var.	χ^2	*d.f.*	*p*-value
r_0	1.19737	1.43368	828.65794	379	<0.001
r_1	0.09387	0.00881	637.21052	379	<0.001
level-1, *e*	0.54829	0.30062			

Figure 1. HLM Results with Time-Invariant Covariate (Left) and Both Time-Invariant and Time-Varying Covariates (Right)



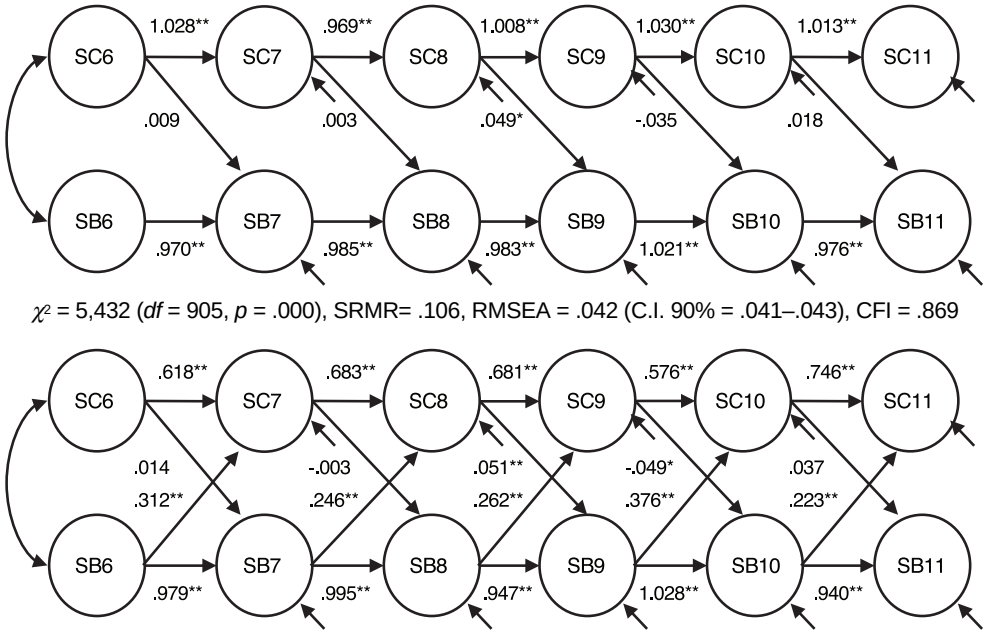
β_{11} = .027, p = .087 (two-tailed)

β_{11} = .013, p = .524 (two-tailed)

Figure 1 compares the average growth rates of self-control of the treatment and control groups. While no difference is observed at the initial level (-.17, p = .453), there is a meaningful difference in the average growth rates between these groups. The cross-level interaction effect shows that the level of self-control increases with time at a higher rate for the members of the treatment group than those in the control group by .027 points each year (p =.088). Interestingly, after incorporating a changing level of social bond as a time-varying covariate at level 1, the observed different rates of change between the two groups in self-control becomes negligible in both magnitude and significance. This suggests that the changing level of self-control is related to the changing level of and social bond even after the formative period of early childhood, which opens up the possibility that self-control is not determined by early childhood history and suggests a dynamic relationship between the two control mechanisms.

HLM results suggest that there is a substantial variability in the growth rate of self-control across individuals and that such variability is accounted for by the changing level of the social bond triggered by the treatment condition. To better isolate the

Figure 2. SEM Results with Longitudinal Latent Variables: Unidirectional (Upper Panel) and Bidirectional (Lower Panel)



χ^2 = 5,432 (df = 905, p = .000), SRMR= .106, RMSEA = .042 (C.I. 90% = .041–.043), CFI = .869

χ^2 = 4,821 (df = 910), SRMR= .076, RMSEA = .039 (C.I. 90% = .038–.040), CFI = .887

Note: SC = self-control; SB = social bond
** p < .01; * p < .05 (two-tailed)

causal mechanism underlying stability and change of self- and social control over time, Na and Paternoster estimated a longitudinal SEM. Figure 2 shows that there is a distinct pattern that supports a social causation model over a self-selection model, social causation being reflected in the path from social bond to self-control and self-selection in the path from social control to social bond. The deletion of five directional paths from the social bond to time-lagged self-control leads to a significant deterioration in model fit relative to the changes in degrees of freedom. This is not surprising considering the magnitude and significance of the path parameters that were included in the bidirectional model but omitted in the unidirectional model, all of which are consistently strong and significant across different model specifications. Interestingly, no meaningful pattern is observed in the directional paths from self-control to the time-lagged social bond measure. In sum, contrary to Gottfredson and Hirschi's assertion, social causation processes continue to occur during adolescence whereas the magnitude and significance of the self-selection process is negligible during the same period.

Treatment Effect Heterogeneity

Analytic Models: GBTM and GMM

HLM assumes that individuals come from a homogeneous population and that a single growth trajectory can adequately approximate the full variation in the entire population. Accordingly, it is better suited for detecting the population average treatment effect, which assumes that a treatment has the same effect for all the individuals under the same treatment condition. While, again, the choice of modeling strategies should be made based on an *a priori* justification primarily guided by theory rather than data-driven approaches because of the obvious dangers of "data snooping" in the model selection (Berk, Brown, & Zhao, 2010), when there is a lack of consensus about the correct model due to the accumulation of empirical anomalies that do not conform to what the theory predicts, it is worthwhile to explore more than one final model and discuss explicitly how results from different modeling alternatives can be used together to better understand the nature and cause of the variation in the outcome profile.

GBTM is an application of finite mixture modeling in which individual trajectories are summarized by a finite number of trajectory groups denoted by the index *k*. When the patterns of heterogeneity do not seem to follow a uniform functional form, GBTM is a useful tool for examining if there are clusters of individuals who do not follow the theoretically normative pattern of development. The level-1 model of GBTM can be

represented as follows:

$$selfcontrol_{ti|ci=k} = \eta_{0k} + \eta_{1k} Grade_{ti} + e_{ti}$$

And the level-2 model can be represented as

$$\eta_{0k} = \alpha_{0k}$$
$$\eta_{1k} = \alpha_{1k}$$

As a nonparametric version of growth mixture modeling, GBTM does not rely on any distributional assumption for the individual variation. Instead, GBTM attempts to approximate an unspecified, potentially nonnormal distribution of unobserved heterogeneity in the population with discrete distributions of distinct clusters of individuals with their own unique growth trajectories. Contrary to HLM, however, the growth parameters (such as intercepts and slopes) for each of these trajectories are no longer allowed to vary across individuals because group trajectories are assumed to capture the full variation across individuals in the population.

As in HLM, treatment status can be included in the basic model as a time-invariant covariate to explain variation in the group membership. While the group-specific treatment effects on the development of self-control cannot be estimated directly because variation around the expected trajectory within each group is assumed to be zero, GBTM analysis is an essential preliminary step before moving on to a more complex model such as GMM when the validity of population homogeneity or common effect assumptions is being questioned. In combination with randomized longitudinal data, GBTM is a useful way to summarize the patterns of development in the counterfactual situation and explore how treatment alters these conventional growth patterns because it allows researchers to directly compare the best-fitting number and shape of trajectories between the two otherwise equivalent comparison groups (see Na, Loughran, & Paternoster, 2015, for more details).

In sum, GBTM assumes that individuals come from a heterogeneous population with multiple homogenous subpopulations and that multiple group-specific growth trajectories can adequately approximate an entire population. Accordingly, it is better suited for investigating group-specific treatment effects because it allows researchers to test whether getting a treatment affects the probability of trajectory group membership and if so, how much it affects that probability, assuming that a treatment does not have the same effect for all the individuals under the same treatment condition. Nonetheless, GBTM cannot estimate subgroup-specific treatment effects because it assumes that there is no variation to be explained around each of the group average

trajectories.

GMM was introduced as an alternative modeling strategy to bridge the gap between HLM and GBTM. While GBTM assumes zero within-group variance in the growth parameters, GMM allows for the variation among individuals within each group. The level-1 model of GMM is as follows:

$$selfcontrol_{ti|ci=k} = \eta_{0ki} + \eta_{1ki} Grade_{ti} + e_{ti}$$

And the level-2 model is as follows:

$$\eta_{0ki} = \alpha_{0k} + \zeta_{0ki}$$
$$\eta_{1ki} = \alpha_{1k} + \zeta_{1ki}$$

Unlike with HLM, it is possible with GMM to have unique growth parameters for each of the subgroups with distinct trajectories. And unlike with GBTM, it is possible with GMM for both intercepts and slopes within each group to have random effects. Thus, GMM can be conceived as a more general growth modeling strategy in which HLM can be seen as a GMM with one class and GBTM can be seen as a GMM with zero variances in the growth parameters for each group. The advantage of GMM for the current study is that the variation of growth parameters within each of distinct subpopulations can be predicted by covariates (e.g., treatment status), which allows for the estimation of subgroup-specific average treatment effects. In sum, GMM is useful if one is trying to detect the heterogeneous effects across different classes of distinct trajectories because it allows for the variation of individual trajectories around each of the group average trajectories and because treatment effect parameters can be estimated separately for each subgroup.

Results

Before estimating the group average treatment effects under the GMM framework, Na, Loughran, and Paternoster employed GBTM to better understand the nature of the developmental pattern of self-control and further investigate potentially heterogeneous treatment effects across individuals following distinct trajectories. Using the formal Bayesian information criterion (BIC) for model selection and subjective criteria based on the objective of the analysis (Nagin, 2005), a four group model—in which all the trajectories were specified to follow a linear functional form—was estimated in order to optimally summarize the complex individual trajectories in the population (the detailed results are available upon request). Figure 3 shows the four trajectories estimated by

GBTM for 6th through 12th graders. Each group is labeled based on the distinct pattern of development characterized by the level and direction of growth parameters (low-low, low-high, high-high, and high-low). The estimated percentage of the population in each group is also indicated in the figure. For solely descriptive purposes, GBTM better identified these distinct clusters of individual trajectories that were not clearly depicted in the HLM results. That is, GBTM results show how the initial level and growth rate of individual trajectories vary significantly across individuals as observed in the HLM analysis (table 2). As previously discussed, Gottfredson and Hirschi (1990) do not predict a decreasing pattern of self-control or reshuffling of trajectories over time. However, consistent with the patterns observed in most recent research (Burt et al., 2014; Hay & Forrest, 2006), which also employs GBTM but uses different samples, a non-normative trajectory (high-low) and shifts in an individual's level of self-control relative to others over time were observed. The localized effect of program participation on the developmental process becomes obvious when the joint-group GBTM with treatment status as a time-invariant covariate is estimated. Considering that GBTM does not assume the existence of random effects within each trajectory, subgroup specific treatment effects cannot be examined simultaneously when trajectory models are estimated. Instead, GBTM presents a sketch of long-term, enduring effects

Figure 3. Trajectories of Self-Control Using Joint-Group GBTM, Grades 6-12

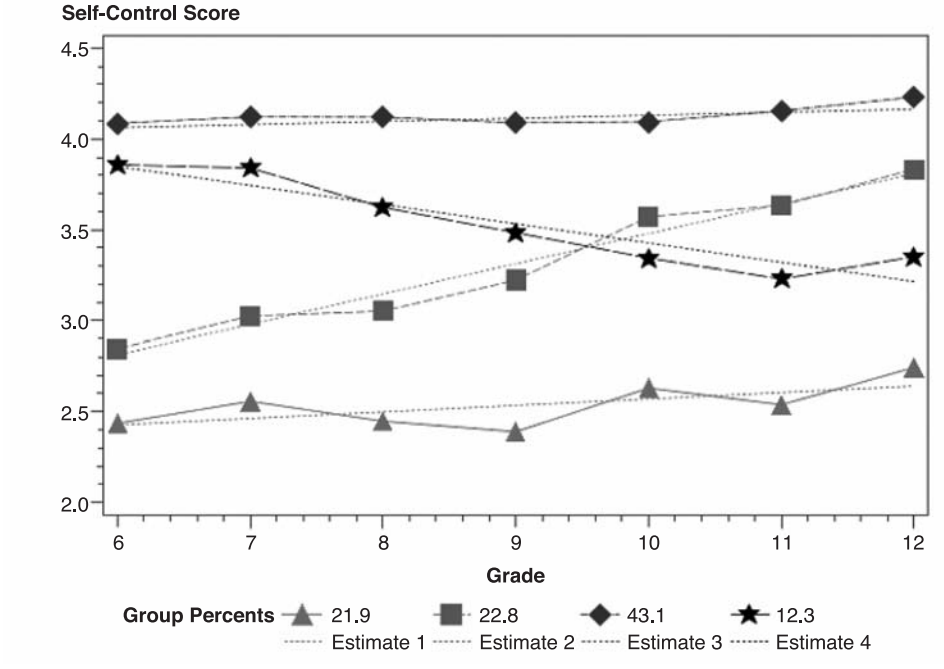


Table 3. The Impact of Treatment on the Probability of Trajectory Group Membership[2]

Group	Parameter	Estimate	Prob > |T|
1	Constant TREATMENT	0.44693 0.59673	0.1687 0.3448
2	Constant TREATMENT	-0.08050 1.73580	0.8585 0.0097
3	Constant TREATMENT	0.75658 1.35966	0.0339 0.0289
4	Constant	(0.00000)	

Note: Group 4 (high-low) is the reference group.

Group	Parameter	Estimate	Prob > |T|
1	Constant	(0.00000)	
2	Constant TREATMENT	-0.52740 1.13903	0.1179 0.0058
3	Constant TREATMENT	0.30966 0.76292	0.1486 0.0122
4	Constant TREATMENT	-0.44694 -0.59673	0.1687 0.3448

Note: Group 1 (low-low) is the reference group.

of a preventive intervention program implemented early in life by incorporating program participation as a predictor of the probability of trajectory group membership. Since this specification is a multinomial and the model has identified four distinct trajectory groups, the estimated coefficients measure how the probability of membership in the reference trajectory group varies as a function of whether individuals participate in the treatment program or not. Table 3 shows that unlike in the case of group 4 (high-low), program participation significantly increases the probability of membership in group 2 (low-high: coefficient = 1.74, p <.01) and group 3 (high-high: coefficient = 1.36, p < .05), which seems to be the primary reason why the members of the plot line of the treatment group overall shows a more sharply increasing slope of self-control than that of the members of the control group as observed in the HLM results (figure 1). In addition, unlike in the case of group 1 (low-low), program participation also slightly but significantly increases the probability of membership in group 2 (low-high: coefficient

2. The results with group 2 or group 3 as the reference group are not presented because they were either unnecessary (e.g., no additional significant results were found) or redundant.

= 1.14, $p < .01$), which also adds insight as to why program participation in general increases the average growth rate of self-control faster than for members of the control group.

Since GMM allows for a mixed population with different trajectories as well as a variation within subpopulations, it is gaining popularity as an alternative model to bridge the gap between HLM and GBTM (Kreuter & Muthén, 2008). While HLM estimates a single average growth trajectory for the whole sample and individual variance is captured by random effects, GMM identifies a subset of individuals whose growth trajectories are significantly different from the overall pattern, and individual variance within each group is captured by subgroup-specific random effects. The results are separate sets of growth parameters and variance/covariance estimates for each latent class (unobserved subpopulation) and, more importantly, subgroup-specific treatment effects. When the average growth pattern in the control group within each trajectory is used as a counterfactual, GMM successfully estimates distinct treatment effects within each trajectory by its incorporation of treatment as a time-invariant predictor of variation in growth parameters. Table 4 shows that the level of self-control increases with time at a significantly higher rate for the members of the treatment group than those of the control group by .145 (p =.013) within group 4 (high-low) and by .113 points (p = .005) within group 1 (low-low), respectively. Interestingly, the treatment had negligible and insignificant effects for group 2 (low-high) and group 3 (high-high). These findings demonstrate how population-averaged treatment effects (the result from HLM: .027, p = .87) might underestimate substantively meaningful localized effects among more theoretically and policy-relevant subgroups of individuals such as those with nonnormative growth patterns (high-low) and those with more room for improvement (low-low). In particular, the effect of program participation was strongest among a specific subpopulation—those with a decreasing pattern of self-control. These individuals are not what Sherman (2007) calls the most harmful cases ("power few") given that they initially manifested a relatively high level of self-control. The results demonstrate how difficult it is to identify a small number of individuals that are most likely to be responsive to and benefit from a treatment merely based on a

Table 4. Comparison of the Program Effect Estimates between HLM and GMM

	All (HLM Result)	Group 1 (Low-Low)	Group 2 (Low-High)	Group 3 (High-High)	Group 4 (High-Low)
slope parameter	.027	.113	-.141	-.019	.145
p-value	.080 n=399	.005 n=114	.112 n=18	.300 n=183	.013 n=84

few a priori known background characteristics when assessing the long-term effects of a treatment on developmental patterns over the course of life.

CONCLUSION AND DISCUSSION

In the field of criminology and criminal justice, the developmental/life-course perspective has emerged as a dominant paradigm, and growing attention has been given to defining effective intervention strategies that prevent or deflect trajectories of criminal/delinquent behaviors. While randomized experiment is widely accepted as a gold standard when the primary goal is to identify an unbiased estimate of treatment effect, my study further discusses why evaluation studies should incorporate other innovative methodological approaches in order to go beyond a mere comparison of the average outcomes between the study groups. In particular, the two empirical studies summarized in this paper illustrate how different modeling strategies can be adopted for studying the long-term effects of a preventive intervention using a developmental trajectory as an outcome variable, causal mechanisms linking the treatment to the outcome, and potentially heterogeneous treatment effects within a larger population. These efforts provide valuable insights into each model's relative utility in describing and explaining the nuanced meanings of the observed treatment effect in the study population.

Using HLM and SEM approaches with longitudinal panel data, Na and Paternoster (2012) find that, in contrast to Gottfredson and Hirschi's prediction that any observed differences in self-control among individuals should remain relatively stable after the age of 8 or 10 (Hirschi & Gottfredson, 2001, p. 90), there were meaningful variations across individuals in the developmental pattern of self-control during adolescence within the pooled sample. A subsample, whose caregivers were part of an intervention effort to improve parenting practices, showed substantially greater gains in self-control than the control group. These findings suggest that self-control is not determined during early childhood, is responsive to intentional attempts to increase it, and continues to develop in response to the changing level of social bonding at least until early adulthood. Theoretically, this study has provided evidence that contradicts Gottfredson and Hirschi's stability postulate and claims about self-selection. The improvement in the relationship between caregivers and children as a result of the treatment intervention continued to have an impact on the developmental pattern of self-control. Practically, it provided convincing evidence of the utility of prevention/intervention efforts by tracking and highlighting the long-term implications of such efforts, instead of merely comparing before-and-after mean scores of the outcome variable at specific time

points. Such process-based approaches can help program developers and policy makers understand the elements and conditions of programs that are related to the observed successful outcomes. In particular, this study has provided a strong empirical evidence for the utility of any effort that seeks to enhance self-control, even during adolescence (e.g., in the school setting) by highlighting the fact that individuals can learn how to exert greater self-control in their adolescence and adulthood.

Using GBTM and GMM, Na, Loughran, and Paternoster (2015)suggest that focusing exclusively on the population average treatment effect based on the common-effect assumption might mask some meaningful heterogeneity in the way the individuals respond to and benefit from the same intervention. Given the inherent limitations of social experiments in which each individual's counterfactual is unobservable and needs be to simulated at the aggregate level, researchers have to make a common-effect assumption at a certain level of aggregation. It makes no sense, however, to assume that every subject receiving the same treatment will respond in the same way. For example, many individuals will always demonstrate high levels of self-control and therefore have less room for improvement no matter what treatment they receive. Others will always have low levels of self-control and therefore be less likely to respond to treatment that does have an impact on others. In this more realistic scenario, modeling strategies designed to identify mean effect size and explain the variability of that mean level are far less useful than growth mixture modeling strategies designed to identify distinct clusters of individual trajectories and separate out the localized treatment effect within each cluster. GBTM and GMM allow for the estimation of various effects of a treatment that are dependent on distinct developmental trajectories that will vary as a function of multiple covariates regardless of whether they are time invariant or time varying, measured or not. Thus, this approach has a distinct advantage over traditional interaction models that rely exclusively on one or a few a priori known covariate(s) as a moderator of the treatment-effect relationship. The essence of the approach is to examine heterogeneous treatment effects for meaningful subpopulations including those that are most responsive to and therefore most likely to benefit from a treatment. At the same time, although not addressed in the current study, these alterative modeling strategies have the potential to identify clusters of individuals for whom a given treatment has unintended, detrimental effects. In addition, this study suggests that even if a sample is relatively homogeneous with respect to crucial background characteristics, which therefore would mean that any variation should be more limited than in a more general population, there still exists substantial and meaningful heterogeneity in long-term individual trajectories of interest. If so, the question remains whether and how each individual trajectory might differentially shift in response to the same treatment. The purpose of this demonstration has been to highlight that we might

wrongfully conclude that a given program is not effective when it in fact has a great impact, even if only on the segments of population who need it the most. Future evaluation research should more explicitly assess the causal mechanisms underlying the impact of policies/programs to better understand exactly why and how such interventions do or even do not work. In addition, not only individual studies that are originally designed to evaluate specific policies and programs but also many systematic reviews or meta-analyses that are commonly adopted to assess the current status of research evidence should focus on both population-average and group-average treatment effects for a more complete understanding of program effectiveness.

REFERENCES

Angrist, J. D. 2004. Treatment effect heterogeneity in theory and practice. *Economic Journal, Royal Economic Society*, 114(494): C52-C83.

Berk, R., Brown, L., & Zhao, L. 2010. Statistical inference after model selection. *Journal of Quantitative Criminology*, 26(2): 217-236.

Bollen, K. A. 1989. *Structural equations with latent variables*. New York: Wiley.

Burt, C., Sweeten, G., & Simons, R. 2014. Self-control through emerging adulthood: Instability, multidimensionality, and criminological significance. *Criminology*, 52(3): 450-487.

Clear, T. R. 2010. Policy and evidence: The challenge to the American Society of Criminology. *Criminology*, 48(1): 1-25.

Cook, P. 2012. Calibrating effect size. 12th Annual Jerry Lee Crime Prevention Symposium. Retrieved on April 19, 2016, from http://cebcp.org/wp-content/uploads/2013/05/Cook.pdf.

Gottfredson, M. R., & Hirschi, T. 1990. *A General Theory of Crime*. Stanford, CA: Stanford University Press.

Hay, C., & Forrest, W. 2006. The development of self-control: Examining self-control theory's stability thesis. *Criminology*, 44(4): 739-774.

Heckman, J. J. 1992. Haavelmo and the birth of modern econometrics: A review of the history of econometric ideas by Mary Morgan. *Journal of Economic Literature*, 30(2): 876-886.

Heckman, J. J., & Smith, J. A. 1995. Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2): 85-110.

Hirschi, T. 2004. Self-control and crime. In R. F. Baumeister & K. D. Vohs (eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 537-552). New York: Guilford Press.

Imai, K., Tringley, D., & Yamanoto, T. 2013. Experimental design for identifying causal mechanisms. *Journal of the Royal Statistical Society*, 176(1): 5-51.

Kravitz, R. L., Duan, N., & Braslow, J. 2004. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly*, 82(4): 661-687.

Kreuter, F., & Muthén, B. 2008. Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology*, 24(1): 1-31.

Muthén, B. O. 2001. Latent variable mixture modeling. In G. A. Marcoulides and R. E. Schumacker (eds.), *New developments and techniques in structural equation modeling* (pp. 1-33). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. O., & Brown, H., Masyn, K., Jo, B., Khoo, S., Yang, C., Wang, C., Kellam, S. G., Carlin, J. B., & Liao, J. 2002. General growth mixture modeling for randomized preventive interventions. *Biostatistics*, 3(4): 459-475.

Na, C., Loughran, T., & Paternoster, R. 2015. On the importance of treatment effect heterogeneity in experimentally-evaluated criminal justice interventions. *Journal of Quantitative Criminology*, 31(2): 289-310.

Na, C., & Paternoster, R. 2012. Can self-control change substantially over time? Rethinking the relationship between self- and social control. *Criminology*, 50(2): 427-462.

Nagin, D. S. 2005. Group-based modeling of development over the life course. Cambridge, MA: Harvard University Press.

Nagin, D. S., & Piquero, A. R. 2010. Using the group-based trajectory modeling to study crime over the life course. *Journal of Criminal Justice Education*, 21(2): 105-116.

Neyman, J. 1935. Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society*, 2(2): 107-180.

Pratt, T. C., & Cullen, F. T. 2000. The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. *Criminology*, 38(3): 931-964.

Raudenbush, S. W., & Bryk, A. S. 2002. *Hierarchical linear models* (2nd ed.). Thousand Oaks: Sage Publications.

Rubin, D. B. 1974. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688-701.

Rubin, D. B. 1977. Assignment to treatment groups on the basis of a covariate. *Journal of Educational Statistics*, 2(1): 1-26.

Rubin, D. B. 1978. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1): 34-58.

Sampson, Robert J. 2010. Gold standard myths: Observations on the experimental turn

in quantitative criminology. *Journal of Quantitative Criminology*, 26(4): 489-500.
Sampson, R. J., & Laub, J. H. 2005. When prediction fails: From crime-prone boys to heterogeneity in adulthood. *The Annals of the American Academy of Political and Social Science* 602: 73-81.
Sherman, L. W. 2007. The power few: Experimental criminology and the reduction of harm. *Journal of Experimental Criminology*, 3(4): 299-321.
Sherman, L. W. 2009. Evidence and liberty: The promise of experimental criminology. *Criminology and Criminal Justice*, 9(1): 5-28.
Sherman, L. W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., and Bushway, S. 1998. Preventing crime: What works, what doesn't, what's promising. www.ncjrs.gov/works/index.htm.
Tittle, C. R., Ward, D. A., & Grasmick, H. G. 2004. Capacity for self-control and individuals' interest in exercising self-control. *Journal of Quantitative Criminology*, 20(2): 143-72.
Weisburd, David. 2010. Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: Challenging folklore in evaluation research in crime and justice. *Journal of Experimental Criminology*, 6(2): 209-227.