CONTENTS

Text

VOLUME 12 • ISSUE 3 • AUGUST 2013

EVALUATION OF THE G.R.E.A.T. PROGRAM

EDITORIAL INTRODUCTION

Two Decades of G.R.E.A.T.: Considering the History and Evaluation of One of the Longest-Running Gang Prevention Programs 367
Andrew V. Papachristos

EXECUTIVE SUMMARY

Overview of: "Short- and Long-Term Outcome Results from a Multisite Evaluation of the G.R.E.A.T. Program" 373
Finn-Aage Esbensen, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Dena C. Carson

RESEARCH ARTICLE

Short- and Long-Term Outcome Results from a Multisite Evaluation of the G.R.E.A.T. Program 375
Finn-Aage Esbensen, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Dena C. Carson

POLICY ESSAYS

GREAT Results: Implications for PBIS in Schools 413
James C. Howell

Do Not Shoot the Messenger: The Utility of Gang Risk Research in Program Targeting and Content 421
Cheryl L. Maxson

Gangs, Criminal Offending, and an Inconvenient Truth: Considerations for Gang Prevention and Intervention in the Lives of Youth 427
David C. Pyrooz

VICTIM-CENTERED PROSECUTORIAL POLICIES

EDITORIAL INTRODUCTION

Empowering Women across Systems: The Impact of Intimate Partner Violence Intervention 437
Jill Theresa Messing

EXECUTIVE SUMMARY

Overview of: "Evidence-Based and Victim-Centered Prosecutorial Policies: Examination of Deterrent and Therapeutic Jurisprudence Effects on Domestic Violence" 441
Mary A. Finn

RESEARCH ARTICLE

Evidence-Based and Victim-Centered Prosecutorial Policies: Examination of Deterrent and Therapeutic Jurisprudence Effects on Domestic Violence 443
Mary A. Finn

POLICY ESSAYS

Victim Engagement in the Prosecution of Domestic Violence Cases 473
Richard R. Peterson

The Importance of Prosecution Policies in Domestic Violence Cases 481
Krista R. Flannigan

Evidence Based Prosecution: Is it Worth the Cost? 491
Eve S. Buzawa and Aaron D. Buzawa

FORECASTING CRIMINAL BEHAVIOR

EDITORIAL INTRODUCTION

Machine Learning Approaches as a Tool for Effective Offender Risk Prediction 507
William Rhodes

EXECUTIVE SUMMARY

Overview of: "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment" 511
Richard A. Berk and Justin Bleich

RESEARCH ARTICLE

Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment 513
Richard A. Berk and Justin Bleich

POLICY ESSAYS

Linking Prediction and Prevention 545
Greg Ridgeway

The Emergence of Machine Learning Techniques in Criminology: Implications of Complexity in our Data and in Research Questions 551
Tim Brennan and William L. Oliver

Is There Any Logic to Using Logit: Finding the Right Tool for the Increasingly Important Job of Risk Prediction 563
Shawn D. Bushway

Guide to Preparing Manuscripts

Editorial Policy—*Criminology & Public Policy* (CPP) is a peer-reviewed journal devoted to the study of criminal justice policy and practice. The central objective of the journal is to strengthen the role of research findings in the formulation of crime and justice policy by publishing empirically based, policy-focused articles. Authors are encouraged to submit papers that contribute to a more informed dialogue about policies and their empirical bases. Papers suitable for CPP not only present their findings, but also explore the policy-relevant implications of those findings. Specifically, appropriate papers for CPP do one or more of the following:

- Strengthen the role of research in the development of criminal justice policy and practice
- Empirically assess criminal justice policy or practice, and provide evidence-based support for new, modified, or alternative policies and practices
- Provide more informed dialogue about criminal justice policies and practices and the empirical evidence related to these policies and practices
- Advance the relationship between criminological research and criminal justice policy and practice

The policy focus of the journal requires articles with a slightly different emphasis than is found in most peer-reviewed academic journals. Most academic journals look for papers that have comprehensive literature reviews, provide detailed descriptions of methodology, and draw implications for future research. In contrast, CPP seeks papers that offer literature reviews more targeted to the problem at hand, provide efficient data descriptions, and include a more lengthy discussion of the implications for policy and practice. The preferred paper describes the policy or practice at issue, the significance of the problem being investigated, and the associated policy implications. This introduction is followed by a description and critique of pertinent previous research specific to the question at hand. The methodology is described briefly, referring the reader to other sources if available. The presentation of the results includes only those tables and graphs necessary to make central points (additional descriptive statistics and equations are provided in appendices). The paper concludes with a full discussion of how the study either provides or fails to provide empirical support for current, modified, or new policies or practices. The journal is interdisciplinary, devoted to the study of crime, deviant behavior, and related phenomena, as found in the social and behavioral sciences and in the fields of law, criminal justice, and history. The major emphases are theory; research; historical issues; policy evaluation; and current controversies concerning crime, law, and justice.

Manuscript Submissions—**Manuscripts are to be submitted electronically to cpp@fsu.edu. The manuscript should be submitted in one Word (.doc) file with tables and figures in the same document as the manuscript text.** Additional documents, including cover letters or memos to the editor, may also be e-mailed as supplemental files. Although we strongly encourage on-line submission, those who prefer not to submit on-line may send a CD to Julie Mestre Brancale, Managing Editor, Florida State University, Center for Criminology & Public Policy Research, 325 John Knox Road Building L-102, Tallahassee, FL 32303.

An executive summary of approximately 150 words and a brief biographical paragraph describing each author's current affiliation, research interests, and recent publications should accompany the manuscript.

Papers accepted for publication should comply with American Psychological Association guidelines concerning nonsexist language. We accept three formats for digital artwork submission: Encapsulated PostScript (EPS), Portable Document Format (PDF), and Tagged Image Format (TIFF). We suggest that line art be saved as EPS files. Alternately, these may be saved as PDF files at 600 dots per inch (dpi) or better at final size. Tone art, or photographic images, should be saved as TIFF files with a resolution of 300 dpi at final size. For combination figures, or artwork that contains both photographs and labeling, we recommend saving figures as EPS files, or as PDF files with a resolution of 600 dpi or better at final size. More detailed information on the submission of electronic artwork can be found at:

http://authorservices.wiley.com/bauthor/illustration.asp

The American Society of Criminology regards submission of a manuscript to this journal as a commitment to publish herein; simultaneous submission to another journal is unacceptable. Every effort will be made to notify authors of editorial decisions within 3 months of manuscript and processing fee receipt. Note that CPP publication decisions arrive by postal mail, so please indicate whether you prefer e-mail correspondence to postal mail upon submission of your manuscript.

Please consult and conform to the CPP Style Sheet, which is available at cpp.fsu.edu, prior to submission.

Criminology & Public Policy (Print ISSN #1538-6473/On-line ISSN #1745-9133) is published quarterly on behalf of the American Society of Criminology, 1314 Kinnear Road, Suite 212, Columbus, OH 43212 by Wiley Subscription Services, Inc., a Wiley Company, 111 River St., Hoboken, NJ 07030-5774.

Information for subscribers. *Criminology & Public Policy* is published in four issues per year. Institutional print and on-line subscription prices for 2013 are: US$328 (U.S.), US$403 (rest of world), €258 (Europe), £206 (U.K.). Prices are exclusive of tax. Asia-Pacific GST, Canadian GST and European VAT will be applied at the appropriate rates. For more information on current tax rates, please go to www.wileyonlinelibrary.com/tax-vat. The price includes online access to the current and all online back files to January 1st 2009, where available. For other pricing options, including access information and terms and conditions, please visit www.wileyonlinelibrary.com/access. A subscription to *Criminology & Public Policy* also includes 4 issues of *Criminology.*

Delivery Terms and Legal Title. Where the subscription price includes print issues and delivery is to the recipient's address, delivery terms are Delivered at Place (DAP); the recipient is responsible for paying any import duty or taxes. Title to all issues transfers FOB our shipping point, freight prepaid. We will endeavour to fulfil claims for missing or damaged copies within six months of publication, within our reasonable discretion and subject to availability.

Journal customer service. For ordering information, claims, and any inquiry concerning your journal subscription, please visit www.wileycustomerhelp.com or contact your nearest office. *Americas* e-mail: cs-journals@wiley.com; telephone: +1 781 388 8598 or +1 800 835 6770 (toll free inside the U.S. and Canada). *Europe, Middle East, and Africa* e-mail: cs-journals@wiley.com; telephone: +44 (0) 1865 778315. *Asia Pacific* e-mail: cs-journals@wiley.com; telephone: +65 6511 8000. *Japan* For Japanese speaking support, email: cs-japan@wiley.com; telephone: +65 6511 8010 or telephone (toll free): 005 316 50 480. ***Visit our Online Customer Get-Help*** available in 6 languages at www.wileycustomerhelp.com.

Membership. American Society of Criminology membership is available to all interested people. Individuals interested in joining should complete and return the application forms on the American Society of Criminology Web site (www.asc41.com). Membership in the American Society of Criminology includes a subscription to *Criminology* and *Criminology & Public Policy*.

Back issues. Information about the availability and prices of back issues can be obtained from the American Society of Criminology by e-mail (asc@asc41.com) or telephone (614-292-9207).

Mailing and on-line access. Journal Customer Services, John Wiley & Sons, Inc., 350 Main St., Malden, MA 02148-5020. Postmaster: Send all address changes to *Criminology & Public Policy.* View this journal online at wileyonlinelibrary.com. For submission instructions, subscription and all other information, visit http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1745-9133.

Access to this journal is available free online within institutions in the developing world through the AGORA initiative with the FAO, the HINARI initiative with the WHO and the OARE initiative with UNEP. For information, visit www.aginternetwork.org, www.healthinternetwork.org, and www.oarescience.org. *Criminology & Public Policy* is published by Wiley Periodicals, Inc., Commerce Place, 350 Main Street, Malden, MA 02148; Tel: (781) 388-8200; Fax: (781) 388-8210.
ISSN 1538-6473 (Print)
ISSN 1745-9133 (Online)

Wiley contacts. Ang Wen Ting, Production Editor (e-mail: CAPP@wiley.com); Kristin McCarthy, Advertising (e-mail: kmccarthy@wiley.com).

Wiley's Corporate Citizenship initiative seeks to address the environmental, social, economic, and ethical challenges faced in our business and which are important to our diverse stakeholder groups. Since launching the initiative, we have focused on sharing our content with those in need, enhancing community philanthropy, reducing our carbon impact, creating global guidelines and best practices for paper use, establishing a vendor code of ethics, and engaging our colleagues and other stakeholders in our efforts. Follow our progress at www.wiley.com/go/citizenship.

Printed in the USA by The Sheridan Group.

THANKS TO OUR REVIEWERS

Criminology & Public Policy (CPP) relies on the expertise and judgment of our blind peer reviewers in identifying manuscripts for publication. I would like to take this opportunity to extend my sincere appreciation to my colleagues for their contributions that have assisted in producing my sixth and final volume of CPP. Each of the reviewers listed below returned at least one review from July 2012 to July 2013.

—Tom Blomberg

Lynn Addington, American University
Edwin Bakker, Centre for Terrorism and Counterterrorism
Bill Bales, Florida State University
Rodger Bates, Clayton State University
Kevin Beaver, Florida State University
Allen Beck, Bureau of Justice Statistics
Richard Berk, University of Pennsylvania
Alfred Blumstein, Carnegie Mellon University
Jeff Bouffard, Sam Houston State University
Anthony Braga, Harvard University
Chester Britt, Northeastern University
James Brunet, North Carolina State University
Anette Brunovskis, Fafo, Institute of Applied International Studies
Shawn Bushway, SUNY, Albany
Kristin Butcher, Wellesley College
Beth Caldwell, Thomas Jefferson School of Law
Scott Camp, Federal Bureau of Prisons
Lauren Cattaneo, George Mason University
Alan Cavaiola, Monmouth University
Jennifer Chacon, University of California, Irvine
Meda Chesney-Lind, University of Hawaii
Ted Chiricos, Florida State University
Philip Cook, Duke University
Terry Coonan, Florida State University
Caroline Cooper, American University
Nicholas Corsaro, University of Cincinnati
Dean Dabney, Georgia State University
Scott Decker, Arizona State University
James Derzon, Battelle Memorial Institute
Bill Doerner, Florida State University
John Donohue, Stanford University
Finn-Aage Esbensen, University of Missouri, St. Louis
Walter Ewing, Immigration Policy Center
Abby Fagan, Florida State University
Susan Faircloth, North Carolina State University
Amy Farrell, Northeastern University
Malcolm Feeley, University of California, Berkeley
James C. Fell, PIRE
Thomas Feucht, National Institute of Justice
Daniel Flannery, Case Western Reserve University
Brian Flay, Oregon State University
Gerry Gaes, Florida State University
Chris Gibson, University of Florida
Paul Gill, University College London
Denise Gottfredson, University of Maryland
Elzbieta Gozdziak, Georgetown University
Kelly Hannah-Moffat, University of Toronto
Karl Hanson, Corrections and Criminal Justice Research, Public Safety Canada
Scott Henggeler, Medical University of South Carolina
Karen Hennigan, University of Southern California
Laura Hickman, Portland State University
Matthew Hickman, Seattle University
Mark Hoekstra, Texas A&M University
Alexander Holsinger, University of Missouri, Kansas City
Beth Huebner, University of Missouri, St. Louis
Haeil Jung, Indiana University
David Kennedy, John Jay College of Criminal Justice
Gary Kleck, Florida State University
Christoper Koper, George Mason University
Roger Koppl, Fairleigh Dickinson University
Tomislav Kovandzic, University of Texas, Dallas
Lauren Krivo, Rutgers University
Marv Krohn, University of Florida
Gary LaFree, University of Maryland
Linda Ledray, Sane-Sart
Mathew Lee, University of Akron
Mike Leiber, University of South Florida
Andrea Leverentz, University of Massachusetts, Boston

Michael Light, Pennsylvania State University
David Lilley, University of Michigan
Jennifer Louden, University of Texas, El Paso
David Lovell, Napa County Executive Office
Cynthia Lum, George Mason University
Peter Manning, Northeastern University
Liz Marciniak, University of Pittsburgh
Doug Marlowe, National Association of Drug Court Professionals
Ramiro Martinez, Northeastern University
Cheryl Maxson, University of California, Irvine
Dan Mears, Florida State University
Jill Messing, Arizona State University
Merry Morash, Michigan State University
Sophia Moskalenko, Bryn Mawr College
Laura Napolitano, Chapin Hall
Thomas Nochajski, SUNY, Buffalo
Michael Ostermann, Rutgers University
Donna Pavetti, Center on Budget and Policy Practices
Justin Pickett, SUNY, Albany
Alex Piquero, University of Texas, Dallas
David Pyrooz, Sam Houston State University
Bill Rhodes, Abt Associates
Angela Robertson, Mississippi State University
Richard Rosenfeld, University of Missouri, St. Louis
Jon Shane, John Jay College of Criminal Justice
Sonja Siennick, Florida State University
LeeAnn Slocum, University of Missouri, St. Louis
Jon Sorensen, East Carolina University
Cassia Spohn, Arizona State University
Hank Steadman, Policy Research Associates, Inc.
Eric Stewart, Florida State University
Kevin Strom, RTI International
Brian Stults, Florida State University
Naomi Sugie, Princeton University
Jody Sundt, Portland State University
Alex Tabarrok, George Mason University
Faye Taxman, George Mason University
Judy Temple, University of Minnesota
Nikolaj Tollenaar, Ministry of Security and Justice
Susan Turner, University of California, Irvine
Patricia Van Voorhis, University of Cincinnati
Tim Wadsworth, University of Colorado
Xia Wang, Arizona State University
Ron Weitzer, George Washington University
David Wexler, University of Arizona
Michael White, Arizona State University
David Wilson, George Mason University
Kevin Wright, Arizona State University

EDITORIAL INTRODUCTION

EVALUATION OF THE G.R.E.A.T. PROGRAM

Two Decades of G.R.E.A.T.

Considering the History and Evaluation of One of the Longest-Running Gang Prevention Programs

Andrew V. Papachristos
Yale University

The Gang Resistance Education and Training (G.R.E.A.T.) program is more than a delinquency prevention initiative: It is a cultural entity. During the past 20 years, thousands of students in hundreds of schools have gone through the curriculum. Members of the very first cohort are now old enough that their own children may be sitting through the G.R.E.A.T. program. More than that, however, people know about G.R.E.A.T. regardless of whether they have experienced it or know exactly what the program does. For instance, when I asked a nonacademic friend if he "had ever heard of the G.R.E.A.T. program," he replied quickly: "Isn't that like D.A.R.E. but with gangs?" Although a highly unscientific sampling strategy, my friend's response illustrates the pervasiveness of G.R.E.A.T. in the public consciousness. People know about the program, have seen its logo, or have heard its name mentioned. I am, literally, writing this introduction with a G.R.E.A.T. pen I received at a conference, and a matching coffee mug is buried in a desk drawer. If you still doubt my argument, then simply type "great" into a Google search box: The *second* result is the webpage for the G.R.E.A.T. program (the first result is the dictionary definition of the word). As far as branding goes, G.R.E.A.T. is a huge success.

But the successful branding of G.R.E.A.T. does not necessarily mean it achieved its stated programmatic goals of decreasing gang delinquency and violence. Luckily for policy makers and academics, G.R.E.A.T. is one of only a handful of instances in the gang-prevention world where rigorous and systematic evaluation has been engaged at virtually every turning point of the program's life. Esbensen, Osgood, Peterson, Taylor, and Carson's (2013, this issue) article represents the latest and greatest evaluation of G.R.E.A.T. providing, in my opinion, the "best yet" evidence of the efficacy and ongoings of the program. It builds

Direct correspondence to Andrew V. Papachristos, Yale University, 493 College St., Room 201, New Haven, CT 06511–8907 (e-mail: andrew.papachristos@yale.edu).

DOI:10.1111/1745-9133.12061

on an already impressive body of research and provides new and significant insights into the program's key objectives and domains. Although the current study answers some old questions, it raises some new questions as well.

Esbensen et al.'s (2013) study situates itself firmly within the prior findings and debates surrounding G.R.E.A.T. The first evaluation of G.R.E.A.T. presented somewhat mixed evidence. On the one hand, the earliest cross-sectional study found some evidence of G.R.E.A.T.'s success with a 1-year post treatment reduction in drug use and minor forms of delinquency, as well as in more negative attitudes toward gangs and increased involvement with prosocial peers (Esbensen and Osgood, 1999). On the other hand, the results from the longitudinal evaluation study were less supportive of program efficacy: Whereas Esbensen, Osgood, Peterson, Taylor, and Freng (2001) found some support for small lagged program effects, there was scant evidence that G.R.E.A.T. achieved its main goal of mitigating gang joining or behaviors. These somewhat contradictory findings raised questions as to whether G.R.E.A.T. could be considered a true success. Although the program demonstrated clear effects on several key behavioral outcomes and attitudes, the evidence for changes in gang behaviors and attitudes—G.R.E.A.T.'s central focus—were less than obvious. This led some to posit that G.R.E.A.T. showed more promise as a general delinquency prevention program than as a gang delinquency and crime-prevention program (Klein and Maxson, 2006).

Politically, a negative evaluation does not ensure programmatic death, just as a positive program evaluation does not ensure subsequent funding and implementation.[1] G.R.E.A.T. did not stop in 1999 or 2001 when the evaluation results were released—indeed, the program was in full swing in 2002 when I received my G.R.E.A.T. swag. Instead, G.R.E.A.T. did what researchers and evaluators hope policy makers and program officers do: They take to heart the results, change things up, and try to make a better program. In her policy essay, Maxson (2013, this issue) retells the experience Esbensen had breaking the news to a room full of cops and public officials of the less than glowing evaluation results. "Instead of reaching for their guns," Maxson writes, "the program managers deserve a lot of credit for asking 'What can we do to make it better?'" And that is what they did: Program managers worked *with* the researchers to improve and essentially retool G.R.E.A.T.

As yet another significant contribution, Esbensen et al. (2013) chronicle the massive overhaul of G.R.E.A.T. in light of the evaluation results. Although the central objectives, goals, and themes of G.R.E.A.T. remained unchanged, Esbensen et al. detail the modifications made to the curriculum and implementation strategies based on other highly touted school-based interventions (especially, Life Skills Training [LST] and Seattle Social Development Model [SSDM]). Such changes, I argue, have created an updated version of G.R.E.A.T. that is distinct from its predecessor—it is, in some key ways, a *new* program. Esbensen et al. provide both readers and researchers with sufficient details to make sense of the

1. Elsewhere, I have written about the science and politics of violence-prevention programs (Papachristos, 2011).

program's history and its current state of operation, which is yet another rare contribution to the evaluation literature.

Esbensen et al.'s (2013) study situates itself squarely within these programmatic changes and evaluation history. In so doing, the study provides continuity in research as well as offers new insights into the revised G.R.E.A.T. As such, Esbensen et al.'s study represents perhaps the foremost evaluation of G.R.E.A.T. for at least three reasons. First, the current study is easily one of the most statistically rigorous in this line of research. Multiple models and modeling considerations are used to examine both 1-year and 4-year post-treatment effects disaggregated by city. Second, a range of attitude, behavioral, and programmatic outcomes is examined to delve into as many program dimensions as possible. Third, the results themselves are thoroughly documented and submitted to a variety of robustness checks. Taken together, these three aspects of the study highlight the rigor of the evaluation as well as the intellectual honesty of the research team. Although it is a point of debate whether G.R.E.A.T. represents a true model of success—and hence the policy essays in this issue—it is undeniable that the researchers serve as an exemplar of how to do evaluation research.

Like its predecessors, Esbensen et al.'s (2013) evaluation of G.R.E.A.T. offers some mixed results on the overall efficacy of the program. Overall, much evidence exists of post-treatment increases in positive attitudes toward the police and prosocial peers, as well as decreases in favorable attitudes toward gangs. In contrast to prior evaluations, Esbensen et al.'s study finds statistically significant decreases in gang membership among 6th and 7th graders at both the 1-year and 4-year post-treatment time periods. In other words, the current evaluation finds important evidence that G.R.E.A.T. reduces levels of gang participation among those in the treatment group. Thus, it seems that the revised G.R.E.A.T. program is much better at decreasing gang joining—one of its foundational goals—than its predecessor. Surprisingly, however, there is no evidence of reduced criminal or delinquent behavior among the treatment group—a somewhat puzzling finding given the consistent research on the facilitative effect of gang membership on crime and delinquency (Thornberry, Krohn, Lizotte, Smith, and Tobin, 2003).

The strength and importance of Esbensen et al.'s (2013) evaluation can be clearly seen in the three policy essays by Pyrooz (2013, this issue), Maxson (2013, this issue), and Howell (2013, this issue). Several themes emerge from these essays. First, Pyrooz and Maxson both raise the question of "how big is enough" for the observed effects size on gang membership. The same issue is raised by Esbensen et al. Given the relatively low dosage of the intervention (a 13-week program), the observed diminished levels of gang membership at 1-year (39%) and the persistence of the effect at the 4-year post-treatment period (24%) seem rather impressive. Furthermore, as Pyrooz argues, virtually no prior evaluation evidence exists to determine whether the observed effects are large. Given the disproportionate amount of crime committed by gang members, both Maxson and Pyrooz point out that the potential returns to such a reduction in gang membership might be larger still. Unfortunately, the null effects on crime and delinquency in Esbensen et al.'s study do not allow such an assessment.

Second, Pyrooz (2013) and Maxson (2013) agree that regardless of the magnitude of the observed effects, G.R.E.A.T. might have been even more effective with a better targeting of the treatment population. Pyrooz suggests that some of the observed effects may be diminished because G.R.E.A.T. treats gang membership as a simple dichotomy—one either is or is not a self-identified gang member. Although important for research and evaluation purposes, Pyrooz argues that such a binary demarcation may dilute the program effects for those who do not easily fit this distinction. Future efforts, Pyrooz argues, should start from a more dynamic view of gang membership such as how "embedded" one is in gang networks (e.g., Pyrooz, Sweeten, and Piquero, 2013).

In much the same spirit, Maxson (2013) wonders whether the effects of G.R.E.A.T. would have been greater if the program recipients were better targeted—in particular, if efforts were directed more toward those "most in need." Maxson suggests that the nonsignificant pretreatment gang risk indicator might imply that the program is effective for youth "who are unlikely to join gangs in any case, as well as those at higher risk." If the program focused more squarely on those at elevated risks of joining a gang, Maxson contends, then the effects might have been even more pronounced.

Howell (2013) offers the most positive assessment of Esbensen et al.'s (2013) study, considering it to be a true model of program effectiveness and one of the most promising strategies for schools interested in gang prevention. Rather than focus on the nuances of the statistical evaluation, Howell concentrates on the necessary things schools should consider when implementing evidence-based programs like G.R.E.A.T. In other words, Howell attempts to translate the usual academic caveats into manageable steps for school practitioners—a noteworthy endeavor given both the importance of G.R.E.A.T. and the stated mission of *Criminology & Public Policy*.

Any evaluation research program would be lucky to have both the name recognition of G.R.E.A.T. as well as such a strong voice in (re)shaping a program's implementation strategies based on empirical findings. But it would not be fair to close this introduction without acknowledging the extent to which G.R.E.A.T. also has changed the much larger criminological discourse on modern street gangs. The treasure trove of G.R.E.A.T. data has been used not only to evaluate program efficacy but also to push the boundaries of gang research. As all of the policy essays note, the data from G.R.E.A.T. have been used to produce hundreds of research reports, peer-reviewed papers, book chapters, dissertations, theses, and fact sheets (Howell, 2013; Maxson, 2013; Pyrooz, 2013). It is one of only a few existing datasets that permit the longitudinal study of gang delinquency. To name just a few topics, G.R.E.A.T. data have been used to study definitional issues around gangs and gang membership, gender and racial differences in gang behaviors, patterns of violent victimization and offending, attitudes toward the police, and drug use and delinquency. To state it more plainly, the field of gang research simply would not be where it is today without G.R.E.A.T.

References

Esbensen, Finn-Aage and D. Wayne Osgood. 1999. Gang resistance education and training (GREAT): Results from the national evaluation. *Journal of Research in Crime and Delinquency*, 36: 194–225.

Esbensen, Finn-Aage, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Dena C. Carson. 2013. Short- and long-term outcome results from a multisite evaluation of the G.R.E.A.T. program. *Criminology & Public Policy*, 12: 375–411.

Esbensen, Finn-Aage, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Adrienne Freng. 2001. How great is G.R.E.A.T.? Results from a quasi-experimental design. *Criminology & Public Policy*, 1: 87–118.

Howell, James C. 2013. GREAT results: Implications for PBIS in schools. *Criminology & Public Policy*, 12: 413–420.

Klein, Malcolm W. and Cheryl L. Maxson. 2006. *Street Gang Patterns and Policies*. New York: Oxford University Press.

Maxson, Cheryl L. 2013. Do not shoot the messenger: The utility of gang risk research in program targeting and content. *Criminology & Public Policy*, 12: 421–426.

Papachristos, Andrew V. 2011. Too big to fail: The science and politics of violence prevention. *Criminology & Public Policy*, 10: 1053–1061.

Pyrooz, David C. 2013. Gangs, criminal offending, and an inconvenient truth: Considerations for gang prevention and intervention in the lives of youth. *Criminology & Public Policy*, 12: 427–436.

Pyrooz, David C., Gary Sweeten, and Alex R. Piquero. 2013. Continuity and change in gang membership and gang embeddedness. *Journal of Research in Crime and Delinquency*, 60: 259–277.

Thornberry, Terence P., Marvin D. Krohn, Alan J. Lizotte, Carolyn A. Smith, and Kimberly Tobin. 2003. *Gangs and Delinquency in Development Perspective*. New York: Oxford University Press.

Andrew V. Papachristos is an associate professor of Sociology, Public Health, and Law at Yale University. His research examines neighborhood social organization, street gangs, interpersonal violence, illegal gun markets, and social networks.

EXECUTIVE SUMMARY

EVALUATION OF THE G.R.E.A.T. PROGRAM

Overview of: "Short- and Long-Term Outcome Results from a Multisite Evaluation of the G.R.E.A.T. Program"

Finn-Aage Esbensen
University of Missouri—St. Louis

D. Wayne Osgood
Pennsylvania State University

Dana Peterson
University at Albany

Terrance J. Taylor and Dena C. Carson
University of Missouri—St. Louis

Research Summary

This article presents results from a randomized control trial of the Gang Resistance Education and Training (G.R.E.A.T.) program; 3,820 students enrolled in 195 classrooms in 31 schools in seven cities were surveyed six times over 5 years (pretests and posttests in Year 1 and four annual follow-up surveys). The results indicate that during the 4 years posttreatment, students who received the program had lower odds of gang membership compared with the control group. The treatment group also reported more prosocial attitudes on several program-specific outcomes. In addition to examining the effectiveness for the full sample, we also report analyses that examine program effects by (a) site and (b) initial levels of risk for gang membership.

Policy Implications

Effective youth violence-prevention programs continue to be few in number; effective youth gang-prevention programs are even rarer. Various rating systems exist (e.g., University of Colorado's Blueprint Model, Helping America's Youth, OJJDP Model Program Guide, and NIJ's Crime Solutions), but even application of the least rigorous standards fails to identify many promising or effective programs. Based on results

DOI:10.1111/1745-9133.12045

reported in this article, the Gang Resistance Education and Training (G.R.E.A.T.) program holds promise as a universal gang-prevention program.

Keywords

gang prevention, *G.R.E.A.T.*, *RCT*, *evaluation*

RESEARCH ARTICLE

Short- and Long-Term Outcome Results from a Multisite Evaluation of the G.R.E.A.T. Program

Finn-Aage Esbensen
University of Missouri—St. Louis

D. Wayne Osgood
Pennsylvania State University

Dana Peterson
University at Albany

Terrance J. Taylor and Dena C. Carson
University of Missouri—St. Louis

Youth gangs continue to garner substantial attention from the media, public, and academic researchers as a result, in large part, of the violence attributed to gang members. Several prevention, intervention, and suppression programs have been introduced to address problems associated with youth gangs, but to date, relatively few have been deemed as promising, let alone as effective (e.g., Esbensen, Freng, Taylor, Peterson, and Osgood, 2002; Howell, 2012; Klein and Maxson, 2006; Maxson, Egley, Miller, and Klein, 2013; Reed and Decker, 2002).

Given the disruptive influence that gangs pose on school safety and academic performance (as well as on communities), gangs and associated violence are targets of prevention

This research was made possible, in part, by the support and participation of seven school districts, including the School District of Philadelphia. This project was supported by Award No. 2006-JV-FX-0011 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Department of Justice or of the seven participating school districts. We would like to express our appreciation to the students who made this project possible by completing the student questionnaires. And this project would have been impossible without our team of colleagues and research assistants; special thanks to Adrienne Freng, Kristy Matsuda, J. Michael Vecchio, and Stephanie A. Wiley for their invaluable assistance. Direct correspondence to Finn-Aage Esbensen, Department of Criminology and Criminal Justice, University of Missouri—St. Louis, One University Blvd., St. Louis, MO 63121-4499 (e-mail: esbensen@umsl.edu).

DOI:10.1111/1745-9133.12048

and intervention efforts. Several programs have been developed and promoted as "effective," and school administrators often are confronted with slick promotional materials advocating the "wonderfulness" of a wide array of programs claiming they will either reduce problem behaviors, increase social skills, promote positive youth behavior, or all of the above. Whenever possible, these school administrators should be encouraged to choose programs with a history of evaluation findings supporting program effectiveness. Although many programs exist, few have been subjected to rigorous program evaluations. Of particular importance is the lack of programs subjected to randomized control trials (RCTs). The current study presents one example of short- and long-term findings from a recent RCT assessing the effectiveness of a gang-prevention program——Gang Resistance Education and Training (G.R.E.A.T.). The findings from this study can aid recent efforts to provide empirically based information to school administrators and community leaders seeking to implement evidence-supported programs.

Despite the relative absence of the most rigorous evaluation designs (i.e., RCTs) assessing gang-prevention programs, an increasing number of agencies/organizations has developed criteria for classifying programs into various categories ranging from "not effective" to "effective" or "model" programs based on the findings of empirical evaluations. For example, the Blueprints Series (Mihalic, Fagan, Irwin, Ballard, and Elliott, 2002; Mihalic and Irwin, 2003) identifies model violence-prevention programs that have withstood rigorous scientific evaluations, and the Maryland Report (Sherman et al., 1997) assessed the effectiveness of a broad range of projects. In 2005, the Helping America's Youth (HAY) Community Guide (Howell, 2009) rated programs identified by nonfederal agencies on three levels: Level 1 (exemplary or model programs based on evaluation designs of the "highest quality"), Level 2 (effective programs based on quasi-experimental research), and Level 3 (promising programs). Similarly, the Office of Juvenile Justice and Delinquency Prevention provides a listing of exemplary, effective, or promising programs (OJJDP, 2010), and in 2010, the National Institute of Justice introduced its "Crime Solutions" website, which identifies effective and promising programs (crimesolutions.gov).

Of particular relevance to the current study, the G.R.E.A.T. program is currently rated as "promising" by OJJDP and by Crime Solutions, and it is designated as Level 2 (effective) in the Helping America's Youth rating scale (findyouthinfo.gov). Additionally, a recent systematic review found that the G.R.E.A.T. program was one of only a handful of gang-awareness programs meeting strict guidelines for determining program effectiveness (Gravel, Bouchard, Descormiers, Wong, and Morselli, 2013). These designations were initially based on findings from two multisite evaluations of the "original" program curriculum: one cross-sectional study conducted in 1995 (Esbensen and Osgood, 1999) and one longitudinal study conducted between 1995 and 1999 (Esbensen, Osgood, Taylor, Peterson, and Freng, 2001), but the current classifications are based on short-term findings from an evaluation of the revised G.R.E.A.T. program (Esbensen, Peterson, Taylor, and Osgood, 2012).

The G.R.E.A.T. program has been in existence since 1991 and has received some acclaim since its inception. Originally developed as a nine-lesson curriculum based on Drug Abuse Resistance Education (DARE), the program underwent a substantial curriculum revision after the findings of the two aforementioned studies. Once these revisions were made, there was considerable interest in determining whether the program would be found to be more effective at meeting program goals than was the case in the evaluations of the original G.R.E.A.T. program. In a recent publication, we reported on the 1-year posttreatment effects of the revised G.R.E.A.T. program (Esbensen et al., 2012). This article provides an overview of those results but focuses on the long-term program effects (up to 4 years posttreatment) while reporting additional analyses that examine (a) site-specific program outcomes and (b) the extent to which preexisting risk factors impact program effectiveness. Our findings contribute to the sparse body of knowledge about effective gang-prevention strategies.

We begin with a description of the G.R.E.A.T. program. Next, we turn to a recap of findings from previous evaluations, with a particular emphasis on critiques levied at both the program and the evaluation findings, and how the current program and evaluation overcome many of the limitations highlighted previously. We then describe the methodology employed and the results of the current evaluation of the revised G.R.E.A.T. program. We conclude with a discussion of how the current results fit with those of previous evaluations and what this means for gang-prevention programming.

G.R.E.A.T. Program[1]

The G.R.E.A.T. program is a school-based gang- and violence-prevention program with three primary goals:

(1) To teach youth to avoid gang membership
(2) To prevent violence and criminal activity
(3) To assist youth in developing positive relationships with law enforcement.

Developed as a universal prevention program targeting youth in early adolescence (i.e., 6th or 7th graders), the G.R.E.A.T. program was classified as a gang-awareness program in a recent review of gang programs (Gravel et al., 2013). The original G.R.E.A.T. program, developed by Phoenix-area police departments in 1991, was a cognitive-based program that taught students about crime and its effect on victims, cultural diversity, conflict resolution skills, meeting basic needs (without a gang), responsibility, and goal setting.[2,3] Uniformed

1. This section describing the G.R.E.A.T. program is partially excerpted from Esbensen et al. (2012).
2. The core program component of G.R.E.A.T. is its middle-school curriculum, and often this is what is referred to with the term "G.R.E.A.T. program." Other optional components of G.R.E.A.T. are an elementary-school curriculum, a summer program, and G.R.E.A.T. Families.
3. For a detailed account of the political context surrounding the development of the original G.R.E.A.T. program, consult Winfree, Peterson Lynskey, and Maupin (1999).

law enforcement officers taught the curriculum in schools, and teachers were requested to complement the program content during regular classes.

The revised G.R.E.A.T. program contains much of the substance of the original program but, importantly, was also informed by the work of educators and prevention specialists and the growing body of risk factor research (see Esbensen et al., 2002; Esbensen, Peterson, et al., 2011, for a detailed account of the program review that informed the curriculum revision). As a result, the revised G.R.E.A.T. program was expanded from 9 to 13 lessons. It is still taught primarily by uniformed law enforcement officers (largely police officers and sheriff's deputies, but federal agents from the U.S. Marshals and the Bureau of Alcohol, Tobacco, and Firearms as well as District Attorneys also have been trained and certified to teach G.R.E.A.T.), and it incorporates classroom management training of officers and a focus on students' skill development through cooperative learning strategies—important pedagogical tools for educational settings (Gottfredson, 2001).[4]

Two school-based programs, the Seattle Social Development Model (SSDM) and Life Skills Training (LST), guided the revision of the G.R.E.A.T. program. LST is classified as a model program by the rigorous Blueprint standards, whereas the SSDM has received acclaim from a variety of sources. The SSDM is a comprehensive model that seeks to reduce delinquency and violence by building a positive learning environment incorporating several different classroom management components, such as cooperative learning, proactive classroom management, and interactive teaching (Catalano, Arthur, Hawkins, Berglund, and Olson, 1998). The LST program is a 3-year intervention in which two annual booster sessions supplement the initial program (Dusenbury and Botvin, 1992). LST consists of three components:

(1) Self-management skills
(2) Social skills
(3) Information and skills directly related to the problem of drug abuse.

The revised G.R.E.A.T. program adopted some of the strategies from LST (in fact, some of the LST curriculum writers participated in the rewriting of the G.R.E.A.T. program), including an emphasis on the development of skills rather than on the assimilation of knowledge, and it incorporated problem-solving exercises and cooperative learning strategies.

During the revision of the G.R.E.A.T. program, incorporation of findings from research identifying risk factors for gang affiliation and violent offending was a primary enhancement to the program. While recognizing the importance of risk factors in all five domains (i.e., community, school, peer, family, and individual), the curriculum writers acknowledged that

4. Information about the G.R.E.A.T. program and an overview of the G.R.E.A.T. lessons included in the middle-school curriculum can be found at great-online.org/.

a school-based program could best address risk factors in the school, peer, and individual domains. As such, the revised curriculum addresses the following risk factor areas: school commitment, school performance, association with conventional or delinquent peers, susceptibility to peer influence, involvement in conventional activities, empathy, self-control (impulsivity, risk-seeking, self-centeredness, and anger control), perceived guilt, neutralization techniques (for lying, stealing, and hitting), and moral disengagement (e.g., Battin, Hill, Abbott, Catalano, and Hawkins, 1998; Esbensen and Deschenes, 1998; Esbensen, Huizinga, and Weiher, 1993; Esbensen, Peterson, Taylor, and Freng, 2010; Hill, Howell, Hawkins, Battin-Pearson,1999; Howell and Egley, 2005; Klein and Maxson, 2006; Maxson and Whitlock, 2002; Maxson, Whitlock, and Klein, 1998; Pyrooz, Fox, and Decker, 2010; Thornberry, 1998; Thornberry, Krohn, Lizotte, Smith, and Tobin, 2003).

Research also has demonstrated the deleterious cumulative effects of risk exposure; the greater the number of risk factors or the greater the number of risk domains experienced, the greater the odds of youth gang and violence involvement, with these increases in risk associated with exponential increases in odds of becoming gang involved (Esbensen et al., 2010; Thornberry et al., 2003). This collective body of risk factor research suggests that prevention programs should attempt to address risk factors in multiple domains and to do so earlier, rather than later, in adolescence, both before the factors accumulate and before the typical age of onset for gang involvement—i.e., prior to the age of approximately 14 years of age (Esbensen and Huizinga, 1993; Hill et al., 1999; Thornberry et al., 2003). To this end, the revised G.R.E.A.T. curriculum addresses multiple risk factors across multiple domains and is taught in 6th or 7th grade, when students average 11–13 years of age.

Comparing Previous Evaluations with the Current Evaluation

Two previous multisite evaluations of the original G.R.E.A.T. program were conducted (Esbensen and Osgood, 1999; Esbensen, Osgood, et al., 2001). These evaluations found different degrees of "success" of the G.R.E.A.T. program at meeting its stated goals. A brief background on these studies provides context for the current study's findings.

The first was a cross-sectional study of nearly 6,000 8th graders attending public schools in 11 U.S. cities conducted in 1995 (Esbensen and Osgood, 1999). The study found many results supportive of the original G.R.E.A.T. program's effectiveness at reaching its goals. A variety of modeling strategies was employed, with three increasingly restrictive samples examined. Under the most restrictive analyses, G.R.E.A.T. students were found to be significantly "better" than non-G.R.E.A.T. students on 14 of 33 outcome measures examined. Program participants were consistently found to have lower levels of drug use and minor delinquent offending than nonparticipants. Examining attitudinal measures with consistent findings across modeling strategies, G.R.E.A.T. students had more negative attitudes about gangs, fewer delinquent friends, more friends involved in prosocial activities, greater commitment to peers promoting prosocial behaviors, less likelihood of acting impulsively, higher self-esteem, more commitment to success at school, and higher levels of attachment to both

mothers and fathers than their non-G.R.E.A.T. counterparts. Additionally, program effects on five outcome measures—peer delinquency, friends' involvement in prosocial activities, commitment to peers who promote prosocial activities, self-esteem, and commitment to success at school—were found to be stronger for males (relative to females), and the effects for two outcomes—commitment to and involvement with prosocial peers—were stronger for Black and Hispanic youth (relative to White youth).

The second evaluation was a prospective longitudinal study of more than 2,000 youth attending public schools in six U.S. school districts. The students were followed from 7th grade (6th in one site) until 11th grade (10th in one site). In 15 of the 22 schools that participated, the classrooms were randomly assigned to treatment and control conditions; in the remaining schools, because of constraints such as G.R.E.A.T. officers' schedules, the classrooms were assigned to condition based on matching procedures (e.g., one teacher's morning class was assigned to the treatment condition, whereas the same teacher's afternoon class was assigned to the control condition). The results of the longitudinal analyses were less supportive of the program than the cross-sectional results. Specifically, 5 of the 32 outcome measures were found to be consistent with beneficial program effects in preprogram versus postprogram (all 4 years combined) contrasts; G.R.E.A.T. students were found to have lower rates of victimization, more negative views of gangs, more favorable attitudes toward the police, more involvement with prosocial peers, and reduced levels of risk seeking. The results examining trends over time were less pronounced, with only three outcomes reaching statistical significance (victimization and involvement with and commitment to prosocial peers) and evidence that effects were delayed (rather than immediate). It is important to note, however, that 25 of the 32 outcome measures examined were in a direction consistent with positive program effects. Also, in contrast to the previous cross-sectional analyses, no significant differences in program effects were found across subgroups by sex or race/ethnicity.

Many of the accolades the G.R.E.A.T. program has received were based, in some part, on the relatively positive findings of the cross-sectional study and on the finding of small lagged effects on some program outcomes in the longitudinal evaluation. That is not to say that these studies were definitive "proof" that the original G.R.E.A.T. program was an undeniable "success." In fact, the results from the longitudinal evaluation were viewed as evidence of a lack of program effect and contributed to the comprehensive program review and revision. Some commentators were critical of the G.R.E.A.T. program and raised concerns about the previous evaluations. Klein and Maxson (2006), for example, noted that the most promising results were found employing the least rigorous methodological design: the cross-sectional study. The more rigorous longitudinal design found less support for the program, as demonstrated by the relative lack of significant differences between treatment and control groups after program exposure and only modest program effects when differences were found. They also highlight the lack of a significant program effect on gang membership, which is the key program outcome.

Klein and Maxson (2006) identified three factors that could account for the failure of the program to reduce the odds of gang membership. First, the original G.R.E.A.T. program was based on a "failed" program model: DARE. Second, the original G.R.E.A.T. program was not "gang specific"; rather, it was based on more general social skills targeted at delinquency prevention. Third, the program was aimed at a population with relatively low rates and probabilities of gang membership. Specifically, Klein and Maxson argued that this universal program focusing on all 7th-grade classrooms would be unlikely to reach the target group because few 7th graders attending schools are involved with gangs.

Ludwig (2005) presented additional concerns about the effectiveness of the original G.R.E.A.T. program. In addition to reinforcing the point that evaluations of the G.R.E.A.T. program found no effect on key dependent variables of gang involvement, drug use, or delinquency, Ludwig also noted that sample attrition throughout the study reduces the confidence that we should have about program effectiveness found in the longitudinal study.

Interest was renewed in the question of program effectiveness after the revised curriculum was fully implemented in 2003. In July 2006, the National Institute of Justice selected the University of Missouri—St. Louis to conduct a process and outcome evaluation of the revised G.R.E.A.T. program. The current program and evaluation address many limitations of the previous program and evaluation designs and build on the results of those previous studies. First, as described, the G.R.E.A.T. program underwent major changes after a substantial curriculum review based in large part on the findings of the previous evaluations. Of particular importance was an emphasis on linking specific program lessons with risk factors found to be important in gang joining and delinquency. In short, whereas the revised program still deals with general social skills and the prevention of delinquency, greater attention is now paid to the risk factors found to be associated with gangs. Second, the criticism that the original G.R.E.A.T. program was modeled after the DARE program was addressed during the curriculum review, with the revised G.R.E.A.T. program now modeled after two highly acclaimed school-based prevention programs (LST and SSDM). Third, Klein and Maxson's (2006) critique of the universal targeted population raised the issue of efforts attempting to reduce statistically rare events. As many gang researchers have noted, gang membership is a rare event, even in the most at-risk neighborhoods or subpopulations. At the same time, the past 20+ years of gang research have demonstrated that gangs and gang-involved youth are found in communities not only across the United States but across the world (e.g., Covey, 2010; Esbensen and Maxson, 2012; Hagedorn, 2008). Although one can question the utility of trying to prevent a statistically rare event, it does not seem reasonable to abandon general prevention efforts, especially given researchers' and practitioners' inability to identify unique risk factors for gang membership and recent studies indicating a great deal of overlap in risk factors for gang membership and violence (Esbensen et al., 2010; Peterson and Morgan, 2013).[5] Finally, with respect to

5. See Klein and Maxson's (2006) review of the gang risk factor literature.

methodological issues raised by Ludwig and others, extensive efforts were made to increase both the active consent rates and the survey completion rates in the current evaluation. The results of these efforts are reported in the Methods section.

Although a previous study reporting short-term program effects of the revised G.R.E.A.T. program was published in 2012 (Esbensen et al., 2012), the current study focuses on long-term effects across 4 years posttreatment.[6] This long-term emphasis is important not only to determine whether short-term effects are sustained over time but also because it captures youth at the ages of highest risk of gang joining (Klein and Maxson, 2006) and because delayed effects were detected in the previous longitudinal evaluation (Esbensen, Osgood, et al., 2001). Additionally, supplemental analyses reported in the current study (a) investigate the extent to which the overall results are replicated at each of the seven individual research sites and (b) control for preexisting risk factors. These important questions address the universality of program effects and introduce a more rigorous assessment than was possible in the previous study. As such, the current study goes well beyond the 1-year program effects reported in the 2012 study.

Methods

Site and School Selection

Seven cities (Albuquerque, NM; Chicago, IL; a Dallas-Fort Worth [DFW], TX, area district; Greeley, CO; Nashville, TN; Philadelphia, PA; and Portland, OR) were selected to provide a diverse sample of schools and students. Sample selection was guided by three main criteria:

(1) Geographic and demographic diversity
(2) A substantial number of officers delivering the program to some, but not all, students
(3) Information provided by the National Gang Center about cities' level of gang activity.

The goal was to develop a sample that was geographically and demographically diverse across cities with varying degrees of gang activity. The student and school sample is representative of the students and schools in each of the seven cities' school districts. The final sample consists of 3,820 students (for whom active consent was obtained) nested within 195 classrooms (102 received G.R.E.A.T. and 93 did not receive the program) in 31 schools.

Active Parental Consent

Active parental consent was required for student participation (see Esbensen, Melde, Taylor, and Peterson, 2008, for a detailed description of the active consent process), and as stated previously, significant effort was made to improve these rates over what was achieved in the previous evaluation. Teachers were recruited to assist with the process, and the

6. In several sections of this article, we report long-term effects alongside the previously reported short-term effects for comparison purposes.

combined effort of teachers and evaluators produced a commendable active consent rate of 78%. Of the 4,905 students represented on the classroom rosters at the time of the consent process, 89.1% of youths ($n = 4{,}372$) returned a completed consent form, with 77.9% of parents/guardians ($n = 3{,}820$) allowing their child's participation and 11.3% ($n = 552$) declining.[7]

Research Design and Random Assignment of Classrooms

The outcome evaluation employs an experimental longitudinal panel design (a randomized control trial with long-term follow-up) in which classrooms in each of the participating schools were randomly assigned to the treatment (i.e., G.R.E.A.T.) or control condition.[8] Once it was determined in which grade level (6th grade in 26 schools and 7th grade in 5 schools) and in which core subject area (commonly social studies but also in English and science classes) the program would be taught, we enumerated all the grade-level classrooms (ranging from 3 to 12). In situations with an odd number of classes, we made the a priori decision to oversample treatment classes (in partial recognition of the fact that many of the principals were reluctant to "deprive" any of their students of the program). The list of classes was then numbered from one through highest, and a table of random numbers was consulted to select the classrooms in which G.R.E.A.T. would be taught. Unselected classrooms comprised the control group.

All students in the treatment and control classrooms were eligible to participate in the evaluation, and those for whom active parental consent was obtained ($N = 3{,}820$) were then asked to participate in the evaluation by completing a confidential group-administered pretest questionnaire. After completion of the G.R.E.A.T. program in each school, students in both the experimental and control groups were then requested to complete posttests and four annual follow-up surveys. The retention rates across the six waves of data included in the outcome analyses reported in this article were 98.3%, 94.6%, 87.3%, 82.8%, 74.2%, and 71.9%, respectively, for Wave 1 (pretest) through Wave 6 (4 years posttreatment).[9] These response rates reflect the diligent efforts of the research assistants working on this project. It is particularly challenging to track students through multiple schools and school districts, especially in a highly mobile sample: Although initially enrolled in 31 middle schools at pretest, students were surveyed in more than 200 different schools in Waves 5 and 6 when the students were in high school. We tracked students in each of the seven cities, identifying the schools (or cities) to which students had transferred. In several

7. This might be compared with an active consent rate of 57% of students being allowed to participate in the previous longitudinal evaluation of the original G.R.E.A.T. program (Esbensen, Osgood, et al., 2001).

8. This is an improvement over the previous longitudinal evaluation design, in which random assignment was possible in only 15 of 22 participating schools (Esbensen, Osgood, et al., 2001).

9. This compares with completion rates of 87%, 80%, 86%, 76%, 69%, and 67% in the previous longitudinal evaluation.

instances (especially for students who had moved outside of the district), this required soliciting information from school administrative assistants, teachers, or other students because, somewhat surprisingly, this information often was not available from the central district office or from computerized records. These efforts at locating students, combined with multiple visits to individual schools (in some instances more than 10 trips to survey chronically truant students), contributed to the fact that we could survey virtually all the students still enrolled in schools in the original districts. We obtained permission from principals at each of the new schools to survey the transfer students—clearly, a time- and labor-intensive effort, but one well worth achieving these high response rates.

Student Sample Characteristics

Based on responses provided at Wave 1, the sample is split evenly between males and females; most (55%) youth reside with both biological parents, and the majority (88%) was born in the United States (see Table 1). The sample is racially/ethnically diverse, with Hispanic youth (37%), White youth (27%), and African American youth (18%) accounting for 82% of the sample. Approximately two thirds of the youth (61%) were aged 11 years or younger at the pretest, representing the fact that 26 of the 31 schools delivered the G.R.E.A.T. program in 6th grade. Three of the six Chicago schools and two of four schools in Albuquerque taught G.R.E.A.T. in 7th grade; thus, students in these sites were somewhat older than students in the other sites.

Measurement

Program Goals

To assess program effectiveness, it was essential that measures of the three program goals be included in the student surveys. Additionally, the G.R.E.A.T. lessons targeted several secondary outcomes that sought to reduce known risk factors for delinquency and gang membership. We developed a student questionnaire that captured the essence of this skills building program, including many of the risk factors associated with gang membership as well as lesson-specific social skills (e.g., dealing with peer pressure and being able to say no). To reiterate, the G.R.E.A.T. program has three primary goals, as follows:

(1) To help youth avoid gang membership
(2) To reduce violence and criminal activity
(3) To help youth develop a positive relationship with law enforcement.

Gang membership is measured by a single-item question that is part of a larger set of questions about youth gangs. Specifically, students answered the question, "Are you now in a gang?" This self-nomination approach has been found to be a valid and robust measure of gang affiliation (e.g., Esbensen, Winfree, He, and Taylor, 2001; Thornberry

TABLE 1

Sample Characteristics at Wave 1

	Full Sample $N = 3,820$ %	ABQ $n = 591$ %	CHI $n = 500$ %	DFW area $n = 614$ %	GRE $n = 582$ %	NSH $n = 590$ %	PHL $n = 457$ %	POR $n = 486$ %
Sex								
Male	50	50	50	54	52	55	43	42
Female	50	50	50	46	48	46	57	58
Race								
White	27	16	7	20	34	45	12	51
African American	18	4	29	21	2	23	44	7
Hispanic	37	49	56	46	50	17	20	13
American Indian	4	10	1	2	5	1	4	4
Asian	4	2	1	6	1	6	4	9
Multiracial	8	14	2	5	4	4	12	13
Other	4	5	2	1	5	5	5	3
Age								
11 or younger	61	35	18	74	77	80	61	79
12	29	43	44	25	22	19	35	20
13 or older	10	23	38	2	2	<1	4	1
Mean Age	11.48	11.87	12.22	11.27	11.23	11.19	11.42	11.21
Living Arrangement								
Both bioparents	55	52	57	60	58	60	38	58
Single parent	20	20	19	15	14	18	24	15
1 Bio/1 stepparent	13	15	12	14	15	12	18	13
1 Bio/1 other adult	7	7	7	7	7	7	11	8
Other relatives	3	6	3	3	4	2	8	5
Other arrangement	2	1	1	1	3	2	2	1
Immigration Status								
Born outside U.S.	12	10	9	13	11	15	11	15
Born in U.S.	88	90	91	87	89	85	89	85

et al., 2003). To measure *delinquency and violent offending*, students completed a 14-item self-reported delinquency inventory, including response categories that allowed for assessment of both ever and annual prevalence as well as frequency of offending during the past 6 months (past 3 months at Wave 2, the posttest). We treated this self-report inventory as a composite measure of general delinquency (examining both a variety and a frequency score) but also created a separate measure of violent offending consisting of three items (attacked someone with a weapon, used a weapon or force to get money or things from people, and been involved in gang fights). To measure the third specific program goal (improving relations with law enforcement), students answered six questions tapping general *attitudes toward the police* as well as two additional questions measuring students' attitudes about police officers as teachers.

Additional Program Objectives

In addition to these three program goals, the 13 G.R.E.A.T. lessons address risk factors for gang joining and life skills thought necessary to prevent involvement in gangs and delinquency (see, e.g., Hill et al., 1999; Klein and Maxson, 2006; Maxson and Whitlock, 2002; Maxson et al., 1998; Thornberry et al., 2003). These mediating variables are treated as implied program objectives and are included in our outcome analyses. We therefore examined the extent to which students exposed to G.R.E.A.T. had improved or enhanced skills that would enable them to resist the lures of gang membership and resist peer pressure to engage in illegal activities. The G.R.E.A.T. lessons encourage students to make healthy choices such as being involved in more prosocial activities and associating more with prosocial peers and less with delinquent peers. The lessons also teach students to improve their communication skills by being active listeners and being better able to interpret verbal and nonverbal communication, targeting these skills to improve students' empathy for others.

In all, 33 outcomes are assessed in these analyses, comprising five behavioral outcomes (gang affiliation, general delinquency, and violent offending—the latter two measured as both frequency and variety indices) and 28 attitudinal measures, including the two measures of attitudes to the police; guilt associated with norm violation; attitudes about gangs; refusal skills; collective efficacy; neutralizations for lying, stealing, and hitting; resistance to peer pressure; associations with delinquent and prosocial peers; prosocial involvement; commitment to negative and to positive peers; school commitment; guilt; empathy; self-centeredness; anger; impulsivity; risk seeking; conflict resolution; calming others; active listening; problem solving; self-efficacy; awareness of services; and altruism. (For a full listing of scales and scale characteristics, see the Appendix.)

Analysis Strategy

The *posttest-through-4-year posttreatment* analysis strategy is an elaboration of that used by Esbensen et al. (2012) for the first two posttreatment waves of outcome measures. These analyses, using MLwiN software (Rasbash, Steele, Browne, and Goldstein, 2009), include the outcome measures obtained on five occasions after treatment (Waves 2 through 6, Level 1) for a total of 15,693 observations nested within 3,739 individual students (Level 2) in 195 different classrooms (Level 3), in 31 schools (Level 4), in 7 cities (Level 5). The analyses allowed for residual mean differences for students, classrooms, and schools through random intercept terms at each level and for cities through dummy variable fixed effects (because of the small number of cities). By mean centering the treatment versus control explanatory variable within schools, we ensured that differences across schools in mean levels of outcomes did not bias the estimate of program effects inadvertently. The model also included a variance component to allow for the possibility that program impact varied across schools (i.e., a random coefficient for treatment versus control at the school level), which ensured an appropriately conservative significance test of program impact. The

variation in program impact across schools did not reach statistical significance at $p < .05$ for any outcomes. The analyses also controlled for the pretest measure of each outcome. We assessed the pretest comparability of treatment and comparison groups through a version of this model that omits time as a level of analysis.

The model allows for change over time through a quadratic function. We were careful to code this function so that the main effect for treatment would reflect mean differences across the entire posttreatment period. We accomplished this by capturing the function through orthogonal polynomials (coded across Waves 2–6 as linear = –2, –1, 0, 1, 2; quadratic = 2, –1, –2, –1, 2). We then centered these terms within each person to adjust for any individual differences associated with attrition. Analyses included random variance components for the linear and squared terms at the individual, classroom, and school levels, thus allowing for the possibility of systematic differences in trajectories at each of those levels.

Our analytic model is designed so that the coefficient for treatment versus control provides an overall assessment of program impact, and the interactions between that term and the linear and squared terms for time reflect change over time in program impact (with significance assessed by a joint test of those two interaction terms). We applied a linear version of this model to most of the outcomes. The measure of gang membership is dichotomous and thus required a logistic version of the model. The self-report measures of general and violent delinquency were highly skewed integer variables, for which a negative binomial model was most appropriate. For the linear models, our tables show the magnitude of program effects in standard deviation units of difference between treatment and control groups (also known as Cohen's *d*), transformed so that positive values reflect beneficial program effects. For the logistic and negative binomial models, we report the percentage difference between treatment and control in odds (for logistic) or mean rate of offenses (negative binomial).

One objective of this multisite evaluation was to include students from diverse settings to allow us to address the issue of transferability of the program. The seven participating cities were selected to represent large and small cities, racially homogenous and racially heterogeneous populations, and cities across the geographical range of the United States. To examine the generalizability and transferability of the program, we implemented a version of the model that provides separate estimates of program effects and time trends for each city. We accomplished this by replacing all the fixed regression coefficients in the base model (except the pretest outcome measure) by their interactions with dummy variables for every site (leaving no reference site). The variance components remained the same.

A body of literature exists that has suggested that youth with greater preexisting risk might benefit more from some programs than youth at low risk (Andrews et al., 1990; Lipsey, 2009). Indeed, the cross-sectional results reported by Esbensen and Osgood (1999) found some evidence that the G.R.E.A.T. program was more effective for males (relative to females) and African American and Latino youth (relative to White youth)—groups

commonly found to be at higher risk of gang membership. To examine this issue, additional analyses were run to test whether the program impact differed between high- and low-risk youth.[10] To measure risk, we first identified respondents who reported belonging to a gang in Waves 2 through 6. We then conducted a logistic regression analysis with that measure as the outcome and sex, race/ethnicity, and 35 Wave 1 measures (the 33 variables identified previously and 2 measures of school and community disorder) of all of the outcome variables as predictors. The fitted values from that analysis differentiate respondents for their probability of joining a gang by the end of the study. These fitted values were most strongly correlated with Wave 1 gang membership ($r = 0.80$), delinquency ($r = 0.74$), and peer delinquency ($r = 0.57$). We defined high-risk youth as the 25% of the sample with the greatest probability of joining a gang and low risk as the remaining 75% of the sample. We tested for differential program effects on high- versus low-risk youth by adding to the base model the two-way interaction of risk with classroom treatment assignment and the three-way interactions of risk and treatment assignment with linear and quadratic change. Finally, we also assessed the extent to which program effects differed by the subgroups (sex and race/ethnicity) compared in the previous evaluations, conducting sex (or race/ethnicity)-by-treatment interactions and examining group-by-treatment interaction over time. These analyses indicated that only for a few (1 or 2 out of 33) outcomes did program effects differ significantly by sex or race/ethnicity, certainly no more than by chance.

Results

Preliminary analyses examined the comparability of treatment and comparison groups on pretest measures. Across the entire set of 33 outcome measures, the differences tended to be small but slightly favored the treatment group, with the mean Cohen's $d = 0.017$ for the 28 measures to which it applies. The differences reached $p < .05$ for three measures and $p < .10$ for a total of seven, which is somewhat more than expected by chance, but not to a statistically significant degree. For instance, the binomial distribution indicates that $p = .23$ for obtaining three or more "significant" results by chance in 33 tests. Furthermore, the least probability any of these differences was $p = .02$, which is far greater than the Bonferroni standard of .0015 for 33 significance tests. To be cautious, we report results from analyses that control for pretest scores. That control had negligible consequence for the magnitude of estimated program effects, but it did increase their precision.

The results across the posttest-through-4-years posttreatment are consistent with those found for the 1-year posttreatment analyses (see Esbensen et al., 2002); the effect sizes, however, are somewhat smaller (see Table 2). In the 1-year posttreatment analyses, program impact was significant at the .05 level for a total of 11 of 33 outcomes, and an additional

10. In addition, this analysis addresses, in part, Klein and Maxson's (2006) critique that this universal program fails to target the most in-need youth and that effects from universal programs such as this might be diluted because of the large number of low-risk youth in the sample.

TABLE 2

One-Year and Entire Four-Year Postprogram Effect Estimates for Attitudinal and Behavioral Measures Controlling for Between-City Differences, Overall Change over Time, and the Pretest Outcome Measure

	1 Year Postprogram			All 4 Years Postprogram		
	Program Effect	*b*	*SE*	Program Effect	*b*	*SE*
Attitudinal Measures						
Impulsivity	0.015	−0.012	0.024	0.021	−0.017	0.021
Riskseeking	0.041	−0.041	0.030	0.053	−0.051*	0.025
Anger	0.057	−0.056*	0.026	0.049	−0.049*	0.023
Self-centeredness	0.054	−0.046*	0.022	0.038	−0.031	0.025
Attitudes toward the police (ATP)	0.076	0.070*	0.024	0.058	0.055*	0.023
G.R.E.A.T. ATP	0.204	0.190*	0.033	0.144	0.129*	0.029
Prosocial peers	0.051	0.050†	0.030	0.040	0.038	0.024
Peer pressure	0.079	−0.050*	0.020	0.044	−0.031	0.019
Negative peer commitment	0.050	−0.047	0.029	−0.002	0.002	0.030
Positive peer commitment	−0.010	−0.011	0.037	0.007	0.008	0.032
Delinquent peers	0.083	−0.051*	0.021	0.025	−0.017	0.018
Lying neutralizations	0.066	−0.066†	0.034	0.042	−0.041	0.027
Stealing neutralizations	0.018	−0.016	0.030	0.017	−0.015	0.029
Hitting neutralizations	0.105	−0.122*	0.032	0.079	−0.095*	0.030
School commitment	0.020	0.015	0.021	0.031	0.023	0.017
Guilt	0.028	0.016	0.016	0.007	0.004	0.018
Conflict resolution	−0.018	−0.008	0.013	−0.009	−0.004	0.011
Calming others	−0.004	−0.002	0.014	0.010	0.005	0.012
Refusal skills	0.090	0.043*	0.013	0.049	0.022*	0.010
Prosocial involvement index	0.047	0.056†	0.030	0.020	0.039	0.032
Empathy	−0.008	−0.005	0.022	0.012	0.008	0.018
Active listening	0.028	0.019	0.020	0.044	0.028	0.017
Problem solving	0.027	0.025	0.024	−0.019	−0.017	0.022
Self-efficacy	−0.004	−0.003	0.024	0.007	0.004	0.021
Awareness of services	0.015	0.012	0.021	0.016	0.012	0.018
Collective efficacy	0.125	0.075*	0.021	0.096	0.055*	0.015
Attitudes about gangs	0.114	0.102*	0.031	0.094	0.079*	0.024
Altruism	0.051	0.031	0.019	0.058	0.033*	0.017
Behavioral[a]						
Delinquency (frequency)[b]	7.0%	−0.073	0.072	5.0%	−0.053	0.059
Delinquency (variety)[b]	7.0%	−0.072	0.048	5.0%	−0.052	0.039
Violent offending (frequency)[b]	10.0%	−0.107	0.179	11.0%	−0.106	0.122
Violent offending (variety)[b]	−1.0%	0.007	0.108	7.0%	−0.070	0.083
Gang[c]	39.2%	−0.498*	0.162	24.0%	−0.271*	0.135

Note. *SE* = standard error.

[a]Program effect as percent reduction.

[b]Negative binomial model.

[c]Logistic regression model.

†$p < .10$. *$p < .05$.

three were marginally significant at the .10 level (prosocial peers, prosocial involvement, and lying neutralizations). Combining the data for the entire 4 years (waves 2–6) post-treatment, we find 10 significant differences, including 8 of the same outcomes that were significant at 1 year posttreatment. The following list identifies the differences for post-test-through-4-years posttreatment; those identified with an asterisk also were noted in the 1-year posttreatment analyses. Three outcomes were significant at 1 year posttreatment but not for posttest-through-4-years posttreatment (self-centeredness, peer pressure, and delinquent associations).

Lower rates of gang membership (24% reduction in odds)*
More positive attitudes to police (ES = 0.058)*
More positive attitudes about police in classrooms (ES = 0.144)*
Less positive attitudes about gangs (ES = 0.094)*
More use of refusal skills (ES = 0.049)*
Higher collective efficacy (ES = 0.096)*
Less use of hitting neutralizations (ES = 0.079)*
Less anger (ES = 0.049)*
Higher levels of altruism (ES = 0.058)
Less risk seeking (ES = 0.053)

With respect to the three specific program goals, the odds of belonging to a gang during the posttest-through-4-years postprogram were 24% lower for the G.R.E.A.T. students, and they continued to have more positive attitudes toward the police in general and to officers in the classroom, compared with non-G.R.E.A.T students. Estimates of program impact did not reach statistical significance, however, for delinquency (general or violent offending). Importantly, the treatment group continued to express less favorable attitudes about gangs, and several risk factors associated with gang membership also were found to be less pronounced among the G.R.E.A.T. students. Students who had participated in the program were more risk aversive, expressed better anger control, and employed fewer neutralizations regarding the use of violence in response to different scenarios. Additionally, as described, several measures were developed and included in the analyses to assess skills taught in the G.R.E.A.T. lessons. For example, the curriculum teaches (through students' role-playing) strategies for students to use to avoid undesired activities in which their friends encourage them to participate. Students in the treatment group were more apt to report use of these refusal techniques. The G.R.E.A.T. students also reported higher levels of altruism and collective efficacy; that is, they indicated that they value doing things for others (e.g., "It feels good to do something without expecting anything in return") and that they can make a difference in their communities (e.g., "It is my responsibility to do something about problems in our community"). These values are reflected in a component of the G.R.E.A.T. program called the "Making My School a G.R.E.A.T. Place" project. This G.R.E.A.T. project provides students the opportunity to have an impact on their

environment by improving their school, surrounding area, or both. The project is intended to be an ongoing part of the program and to be completed by the end of the 13th lesson.

In contrast to these positive program effects, our long-term (posttest-through-4-years postprogram) analyses failed to discern a difference between the G.R.E.A.T. students and the control group on a range of peer-related factors: prosocial peers, peer pressure, negative peer commitment, positive peer commitment, and delinquent peers. Three of these potential outcomes were marginally significant ($p < .10$) in the 1-year post treatment analyses (prosocial peers, peer pressure, and delinquent peers), suggesting that the peer effect is muted over time. Also, the program did not produce statistically significant differences for several social skills or risk factors emphasized in one or more lessons: conflict resolution, calming others, active listening, problem solving, empathy, self-efficacy, awareness of services, prosocial involvement, neutralizations for lying and stealing, guilt, school commitment, self-centeredness, and impulsivity. The latter two outcomes are subcomponents of the larger self-control measure developed by Grasmick, Tittle, Bursik, and Arneklev (1993). The program impact for two other components of self-control (risk seeking and anger) did reach significance. These aspects of the program that did not differentiate the groups suggest that perhaps attitudes are more easily influenced than is behavior. A large proportion of these remaining nonsignificant factors consists of social skills variables representing program components that teach students factual information or how to modify their behavior (e.g., availability of services, active listening, calming others, and problem solving). That is, students are instructed on where to find assistance when needed and on the importance of listening to others when they speak, how to calm others who are upset, and constructive (and nonviolent) ways to solve problems that develop.

Site-Specific Analyses Posttest-Through-4-Years Posttreatment

One evaluation objective was to address the transportability of the program. That is, can G.R.E.A.T. be effectively taught in a variety of settings? To address this issue, we included seven diverse cities in the study, and in this set of analyses, we explore the extent to which the aggregate-level differences are replicated in the seven different cities. As shown in Table 3, the findings are mixed. At 1 *year posttreatment* (the first columns for each site), the overall findings are largely replicated in three sites (Albuquerque, the DFW area site, and Portland). A few program effects (including lower odds of gang membership) were noted in Philadelphia, but null findings were found in Greeley, Nashville, and Chicago (see Table 3).

It is important to consider whether these differences across sites in program impact reflect genuine differences in effectiveness or result from a combination of smaller sample sizes and chance variation inevitable among estimates of limited precision. Interaction tests give clear evidence that differences in impact across sites are statistically reliable for only G.R.E.A.T. attitudes toward police and negative peer commitment. For both, $p = .0011$, which surpasses the Bonferroni corrected value of $p < .0015$ (for $p < .05$, 33 tests). For the entire set of 33 outcomes, a total of 4 tests reached the nominal level of $p < .05$ and 6

TABLE 3

One-Year and Entire Four-Year Site-Specific Program Effect Estimates for Attitudinal and Behavioral Measures Controlling for Between-City Differences, Overall Change over Time, and the Pretest Outcome Measure—Significant Effect Sizes Only

	Albuquerque		Chicago		DFW Area		Greeley		Nashville		Philadelphia		Portland	
	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year
Impulsivity														
Risk seeking					0.13†								0.19	0.17
Anger					0.11†	0.10†								
Self-centeredness					0.20	0.14†								0.18†
ATP	0.15				0.17	0.16							0.17	0.14
G.R.E.A.T. ATP	0.39	0.29			0.23	0.16			0.17				0.43	0.36
Prosocial peers		0.13				0.16								0.13
Peer pressure	0.18	0.12†			0.16					–0.14				
Negative peer commit	0.19	0.12†								–0.25			0.14†	
Positive peer commit						0.19								
Peer delinquency	0.32	0.18			0.12†									
Lying neutralizations										–0.13		0.18		
Stealing neutralizations										–0.17				
Hitting neutralizations	0.16	0.13			0.13†								0.17	0.16
School commitment		0.10*								–0.13				
Guilt						0.17								
Conflict resolution														
Calming others						0.16								
Refusal skills	0.15	0.11*			0.15	0.11							0.17	0.11†

Continued

TABLE 3

Continued

	Albuquerque		Chicago		DFW Area		Greeley		Nashville		Philadelphia		Portland	
	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year	1 Year	1–4 Year
Prosocial involvement								−0.12						
Empathy						0.13†								
Active listening					0.17	0.22								
Problem solving													−0.17	−0.22
Self-efficacy														−0.15
Awareness of services														
Collective efficacy	0.19*	0.17			0.24	0.17					0.18†		0.19†	
Attitudes about gangs	0.20				0.28	0.25					0.19	0.17	0.20	
Altruism	0.14*				0.18	0.19							0.14†	
Delinquency (freq)						31%†								
Delinquency (variety)						23%			−43%					
Violent (freq)														
Violent (variety)														
Gang	71%	58%									65%	48%	61%†	

Note. Negative estimates, such as those found in Nashville and Greeley, indicate a negative program effect.
†$p < .10$. *$p < .05$.

reached the nominal level of $p < .10$, which is somewhat more than chance but not notably so. Also recall that we did not find significant school-level variance in program impact for any outcomes. Whether the differences among sites reflect chance fluctuations or genuine differences in effectiveness, the results of Table 3 make clear that any given implementation of the program might or might not achieve results consistent with the overall average.

The results for site-specific program impact *across all 4 years posttreatment* (the second columns for each site in Table 3) are similar to those found at 1 year posttreatment. Once again, the results in Albuquerque, Portland, and the Texas site resemble the aggregate results. Philadelphia experienced a few positive outcomes, whereas Chicago and Greeley once again had null findings. For posttest-through-4-years posttreatment, however, the G.R.E.A.T. students in Nashville reported five negative program effects (more susceptibility to peer pressure, more commitment to negative peers, less school commitment, and greater neutralizations for lying and stealing). Overall, the site-specific results are robust with the posttest-through-4-years posttreatment results similar overall to those found for 1 year posttreatment with the caveat that the 1-year posttreatment effect sizes, as is the case with the full-sample results, are somewhat larger.

Preexisting Risk Analyses Posttest-through-4-Years Posttreatment

To test for the possibility that the G.R.E.A.T. program might be more suitable for high-risk youth, we used Wave 1 data to identify students at risk for gang membership. Specifically, we used sex, race/ethnicity, and 35 attitudinal and behavioral measures (the 33 outcome measures plus school and community disorder) from Wave 1 as predictors of being a gang member in any subsequent wave (i.e., Waves 2 through 6). Then, we saved the predicted probabilities as the risk measure. Although there is no set standard for classifying risk, we dichotomized the risk measure and identified the top 25% as at risk (a method used, e.g., by Farrington and Loeber, 2000; Hill et al., 1999). To minimize missing data, we substituted scale means for any missing Wave 1 predictors when computing the risk score. None of the treatment-by-risk interactions is significant, but to test for the possibility that effects might change over time, we examined also risk-by-treatment-by-time interactions. Several significant three-way interactions emerged, and the pattern is consistent. The three-way interactions suggest that most of the beneficial impact is associated with the high-risk students in the early waves and that the treatment/control difference for high-risk youth fades over time. Some evidence shows that the treatment is increasingly beneficial for low-risk youths over time, but that pattern is far from consistent.

Table 4 provides a summary of the analyses of differential impact in relation to risk. The variables are coded so the main effects retain their original meaning.[11] Four of the 33

11. The overall impact effects reported are similar to, but not exactly the same as, those reported for the aggregate-level analyses in Table 2 because this model adds risk level as a predictor and all its interactions with treatment condition and time (both linear and squared).

TABLE 4

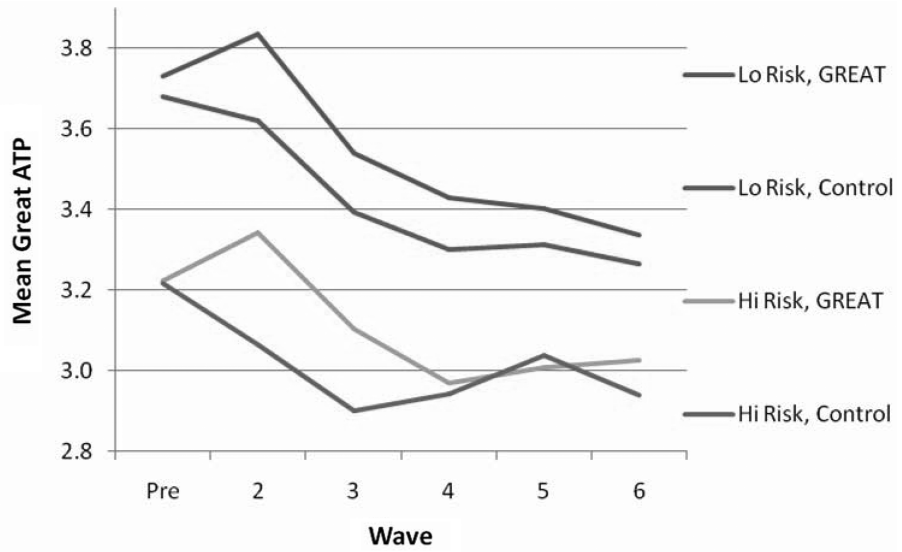
Interaction Effects of Risk by Impact and Risk by Impact by Time

	Risk × Impact			Risk × Impact × Time
	Program Effect Difference (*d*)	*b* Difference	*SE*	Wald Chi Square (2 *df*)
Attitudinal				
Impulsivity	0.036	–0.029	0.044	5.95†
Risk seeking	0.042	–0.041	0.051	9.36*
Anger	0.042	–0.042	0.052	4.35
Self-centeredness	0.062	–0.051	0.043	4.48
(ATP)	–0.053	–0.050	0.048	4.52
G.R.E.A.T. ATP	0.017	0.015	0.047	4.87†
Prosocial peers	0.077	0.073	0.048	0.22
Peer pressure	0.095	–0.066†	0.036	15.96*
Negative peer commitment	–0.032	0.032	0.054	6.04*
Positive peer commitment	0.055	0.064	0.058	3.95
Delinquent peers	0.098	–0.067†	0.036	17.64*
Lying neutralizations	0.099	–0.096†	0.050	15.51*
Stealing neutralizations	0.013	–0.012	0.047	9.62*
Hitting neutralizations	–0.035	0.041	0.059	4.43
School commitment	0.015	0.011	0.039	7.36*
Guilt	0.066	0.041	0.032	10.27*
Conflict resolution	0.013	0.006	0.022	8.39*
Calming others	–0.004	–0.002	0.024	4.23
Refusal skills	0.051	0.024	0.024	3.22
Prosocial involvement index	–0.030	–0.059	0.073	0.76
Empathy	–0.053	–0.036	0.037	4.36
Active listening	0.003	0.002	0.033	1.33
Problem solving	0.004	0.004	0.048	5.03†
Self-efficacy	0.019	0.013	0.037	0.61
Awareness of services	–0.056	–0.043	0.042	2.08
Collective efficacy	0.056	0.032	0.032	0.13
Attitudes about gangs	0.088	0.074†	0.043	2.66
Altruism	–0.022	–0.013	0.031	4.06
Behavioral				
Delinquency (frequency)	11.1%	–0.105	0.127	9.58*
Delinquency (variety)	10.5%	–0.100	0.079	6.21*
Violent offending (frequency)	40.8%	–0.342	0.302	2.54
Violent offending (variety)	26.7%	–0.236	0.181	5.46†
Gang	–16.8%	0.184	0.255	7.84*

Notes. df = degrees of freedom; *SE* = standard error.
†$p < .10$. *$p < .05$.

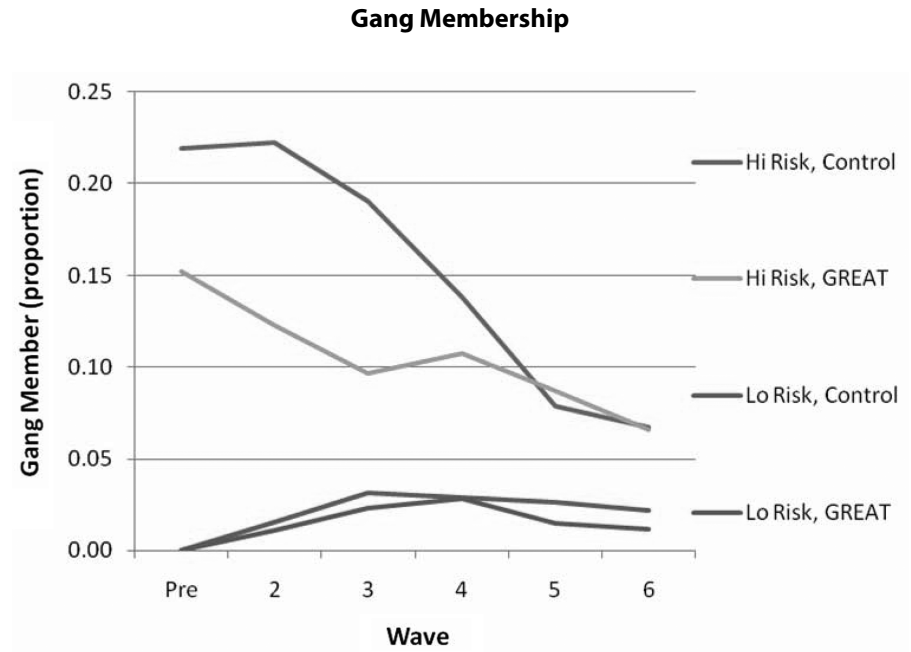
FIGURE 1

G.R.E.A.T. Attitudes toward the Police (ATP)



risk-by-treatment interactions reached the .10 level of significance, but none reached the .05 level, a pattern that could easily arise by chance. Twelve of the three-way interactions (risk by treatment by time) were significant at the .05 level and four more reached the .10 level. Furthermore, the significance levels for three outcomes surpassed the Bonferroni correction criterion of $p < .0015$ and a total of seven reached $p < .01$, giving strong evidence of genuine rather than chance effects for the dataset as a whole. Figures 1 and 2 provide examples of the three-way interactions for the four combinations of high versus low risk and G.R.E.A.T. versus control. Figure 1 shows that for G.R.E.A.T. attitudes toward police, the treatment and control groups are comparable at the pretest for both high-risk and low-risk youth. In Waves 2 and 3, the treatment group shifts toward more favorable attitudes than the control group, and the resulting difference is more pronounced among high-risk youth. Across Waves 4 through 6, the treatment versus control difference largely disappears for the high-risk youth, whereas a moderate difference remains for the low-risk youth. For gang membership, Figure 2 shows that among the high-risk youth, a somewhat larger proportion of control rather than treatment youth were gang members, and that the G.R.E.A.T. program led to greater reductions in membership for the treatment group than controls through Waves 2 and 3. By Wave 6, however, this treatment effect was no longer apparent. The rates of gang membership were much lower in the low-risk group, of course, but we observe suggestions of a beneficial program effect gradually emerging so that at Wave 6, the rate of gang membership was only half as high in the treatment group as the control group.

FIGURE 2



Discussion

Schools are a desirable location to offer universal programs with an emphasis on preventing an array of adolescent problem behaviors including bullying, drug use, dating violence, gang affiliation, and others (Jimerson, Nickerson, Mayer, and Furlong, 2012). Although school-based violence-prevention/intervention programs are widespread, knowledge of their effectiveness is often lacking (Alford and Derzon, 2012; Gottfredson, 2001). Given teacher and administrator concern about the "loss of instructional time" to nonacademic activities, school administrators increasingly rely on "evidence-based practices" when making decisions about which, if any, programs to allow into their schools. The G.R.E.A.T. program is one primary prevention program that, based on our evaluation, holds promise.

In addition to increased placement of prevention programming in schools, the past 20 years have observed an increase in the presence of police officers on school campuses, as both School Resource Officers (e.g., Finn and McDevitt, 2005; Na and Gottfredson, 2013; Petteruti, 2011) and prevention program providers (e.g., DARE and G.R.E.A.T.). The research reported in this article addresses the efficacy of a program that uses law enforcement officers to deliver a gang-prevention and violence-reduction program.

The G.R.E.A.T. program is one choice that school administrators have when selecting from a vast list of prevention programs. G.R.E.A.T. is currently rated as "promising" by OJJDP and by Crime Solutions, and it is designated as "Level 2" (effective) in the Helping

America's Youth rating scale (findyouthinfo.gov). These designations, although initially based on findings from two previously published evaluations of the original G.R.E.A.T. program, have incorporated and are now based on the short-term results reported from the current evaluation of the revised G.R.E.A.T. program (Esbensen et al., 2012). To recap those previous studies, a cross-sectional study conducted in 1995 found that G.R.E.A.T. students were substantially "better" than non-G.R.E.A.T. students on a variety of attitudinal and behavioral outcomes (Esbensen and Osgood, 1999). A more rigorous longitudinal evaluation conducted between 1995 and 1999 found less support for the program (in terms of the number of significant differences, effect sizes, and the presence of delayed—rather than immediate—effects) (Esbensen, Osgood, et al., 2001). Still, because most of the results were in the direction of positive programmatic effects, G.R.E.A.T. was deemed by raters as a program holding "promise." This was particularly true given the relatively short program dosage (i.e., nine 1-hour lessons delivered over a span of 9 weeks).

Previous critiques of the original program and earlier evaluations raised several concerns. First, some commentators labeled G.R.E.A.T. a "failed program" based on a lack of significant effects on delinquency or gang membership (Ludwig, 2005; Klein and Maxson, 2006). Additionally, when positive programmatic effects were found between G.R.E.A.T. and non-G.R.E.A.T. students, effect sizes were modest (Klein and Maxson, 2006). Third, two well-known gang researchers suggested that the lack of significant program effects were not surprising, given the program's emphasis on factors related to general delinquency (as opposed to gang-specific issues), its modeling after the failed DARE program, and the fact that it targeted a population at low risk of gang involvement (Klein and Maxson, 2006). Finally, the previous longitudinal evaluation was criticized for the extent of sample attrition occurring during the examination period (Ludwig, 2005).

The G.R.E.A.T. program underwent a substantial overhaul after a curriculum review (see Esbensen et al., 2002; Esbensen, Peterson, et al., 2011, for a detailed account of the program review). Many changes were sparked by findings from these early evaluations. The program was expanded from 9 to 13 lessons, and substantial effort was made to link specific program lessons to evidence-based risk factors for gang joining and delinquency found in prior research. Practitioners and researchers versed in gangs and school-based prevention were brought together to offer suggestions for program modifications. Then, professional curriculum writers were employed to develop the specific program lessons. This effort led to the "revised" G.R.E.A.T. program that is the focus of the current study.

After the revised program was fully implemented in 2003, interest in assessing the effectiveness of the G.R.E.A.T. program was renewed. Our most recent work (Esbensen et al., 2012), based on a longitudinal evaluation design that included full random assignment and improved active consent and retention rates, reported that a relatively low dosage (13 lessons) primary prevention program can have measurable effects on a diverse sample of students 1 year posttreatment. This article extends that research by reporting the results of treatment effects up to 4 years posttreatment. We also address two additional questions:

(1) Were the aggregate results replicated in each of the seven study sites?
(2) Did the results vary based on youths' preexisting levels of risk?

G.R.E.A.T. Goals and Objectives

The posttest-to-4-year postprogram analyses examined the direct effects of G.R.E.A.T. on the three main program goals (preventing gang involvement, reducing delinquency and violence, and improving views of law enforcement) as well as on several risk factors associated with gang affiliation that were targeted in the curriculum. The results identify positive program effects on a number (10 of 33) of these program objectives. Compared with students in the control classrooms, students in G.R.E.A.T. classrooms expressed more positive attitudes to the police and lower odds of gang membership. They reported also more use of refusal skills, lower support for neutralizations regarding violence, less favorable attitudes about gangs, lower levels of risk seeking and anger, higher levels of altruism, and a higher degree of collective efficacy. It is important to highlight that eight of the ten differences found across 4 years posttreatment also were evident among the 11 differences 1-year post program delivery, indicating a sustained, long-term program effect on those outcomes.

The effect sizes are small (and program rating schemes weight this important aspect of program impact), which remains a criticism lodged by reviewers and by rating schemes. The Blueprints program, for example, declined to classify the revised program as "promising" largely because of the small effect sizes (S. Mihalic, personal communication, April 8, 2013). In our view, that assessment fails to take into account the limited scope and cost of the G.R.E.A.T. program. It is important to note that we are independent evaluators not program developers, and we have no stake in this program's success, financially or otherwise. From our first introduction to the G.R.E.A.T. program in the early 1990s, our shared sentiment was and remains skeptical that there would be a measurable effect of a 9- or 13-lesson program with the average lesson being less than 40 minutes, which is diluted even more by absenteeism and scheduling issues. Finding such beneficial program effects across multiple studies has surprised us, and their consistency forces us to take them seriously. We ask not only ourselves but also the critics what effect size is reasonable to expect given the low dosage and the general audience targeted by this program, and how large must the effects be to justify the use of a program requiring such limited investment?

The revised program and the most recent evaluation design overcome many of the limitations critics noted for the original program and evaluations of it. The program itself is now more "evidence-based," focused on key risk factors found to be important for gang joining. Additionally, more pedagogically sound strategies (such as active learning as opposed to didactic lecture) comprise a bulk of the program lessons. These two factors provide reason for optimism that the revised program should be more effective at preventing gang membership than its original configuration. These program revisions might be responsible for the divergence in findings related to gang membership between the current evaluation

and its previous counterparts. Specifically, the increased focus of the revised curriculum on theoretically and empirically based risk factors for gang joining, coupled with a more effective "skills-based" programmatic structure, might be the primary reasons why the current study finds G.R.E.A.T. participants report reduced odds of gang joining, relative to their non-G.R.E.A.T. counterparts. Conversely, programmatic effects might be more easily uncovered based on the higher rates of study participation relative to the previous longitudinal study.

All this being said and despite many significant effects in favor of the G.R.E.A.T. program, our current results also include several effects that failed to reach statistical significance. Next, we focus specifically on some of the findings among the social skills and peer-related measures. We focus on these two areas because of consistent (non) effects. Before discussing them, however, we remind readers that chance might be the source of the weaker results for these outcomes. The lack of significance is definitely not proof of "no effect," and differences in program impact between these outcomes and the others are rarely if ever statistically significant (judging from their standard errors and implied confidence intervals).

Social skills. Our overall lack of findings with regard to several social skills might engender disappointment. In discussing the lack of change in several skills among G.R.E.A.T. students, we speculated previously that effecting attitude change might be easier than stimulating behavioral change. That is, a greater proportion of attitudinal than skills-based behavioral changes existed among the significant differences found between G.R.E.A.T. and control students, and a greater proportion of skills-based factors among the nonsignificant differences. One skill for which we did find a significant difference, however, was G.R.E.A.T. students' greater use of refusal skills. In our classroom observations of lesson delivery, we noted that this component, more than other social skills components, used role plays between students and the officer, with the officer attempting to lure the student into deviant behavior and the student practicing a host of methods to resist involvement. Students relished this exercise, actively paying attention and participating. We suggest it is possible that students' greater interest in and ability to practice this skill might have produced the positive results and that offering students more opportunity to rehearse the other social skills might yield the intended programmatic effect.

Peer effects. Two of the three program effects (resistance to peer pressure and association with delinquent peers) that were found 1 year posttreatment but not for the full 4 years posttreatment are related to the role of the peer group, and one additional peer outcome that reached marginal significance ($p < .10$) at 1-year posttreatment also failed to reach significance across the entire 4 years posttreatment. Two other peer-related variables (commitment to positive and to negative peers) also failed to reach significance at both time periods. These results raise two issues, as follows:

(1) Can an individual-targeted program impact peer factors?
(2) If yes, then can these results be sustained over time?

The answer to the first question is mixed; modest differences were found between treatment and control students on the peer-related outcomes and risk factors at 1 year posttreatment. The answer to the second question seems to be no; for the 4 years posttreatment as a whole, peer-related differences for the full sample were no longer statistically significant. These results, although disappointing, might be expected: Peers play a major role in the lives of adolescents, and a few brief lessons encouraging youth to avoid negative peer influences might not be sufficient to overcome these influences to achieve the intended outcome.[12]

Program Effects by Site

Some questions are raised by the site-specific results regarding the utility of the G.R.E.A.T. program as a general gang-prevention program applicable in a variety of settings. Three diverse cities (Albuquerque, a DFW suburb, and Portland) experienced program results similar to the larger sample. These sites represent cities with a large Hispanic population (Albuquerque), a city that has the largest percentage of White residents in the United States (Portland), and a city that is part of a large megalopolis (the DFW area site). One city has a long history of gangs (Albuquerque), whereas the other two have relatively new gang problems. The cities with null findings are also diverse—one is among the largest cities in the nation (Chicago) with pockets of extreme disadvantage and high rates of violent crime, whereas the other city (Greeley) is the smallest in the sample (less than 100,000 inhabitants) but with a pronounced gang problem that emerged in the past two decades. A few program effects (notably, lower odds of gang involvement and less positive attitudes about gangs) were found in Philadelphia, a city similar to Chicago in many ways, being large and having neighborhoods facing long-standing poverty, violence, and gang activity.

These findings highlight the importance of conducting multisite evaluations not only to assess the transportability of the program or policy but also to allow for the possibility that contextual effects in some sites might not allow for the detection of program effects (Type II error). For example, whereas one of our considerations in selecting the final sites for the evaluation was program saturation (i.e., we excluded sites in which the G.R.E.A.T. program had a long history, thereby introducing the possibility of program contamination in the control group), it was only after agreements had been obtained that we learned that the Nashville Police Department had an extensive involvement in the schools, teaching the G.R.E.A.T. elementary-level component in 3rd or 4th grade, DARE in 5th grade, and then the G.R.E.A.T. middle-school component in 6th grade, as well as a DARE booster session in 9th grade. Thus, the absence of a positive program effect in Nashville might be an artifact of this police saturation in the schools.

Trying to make sense of these site-specific differences led us to consider several potential explanations. First, it is essential to keep in mind that differences of this magnitude are

12. We temper this with the reminder that program effects seem to vary by site, and at least in one site, the program does produce significant and lasting differences on peer-related variables.

little more than would be expected by chance alone. Next, we revisited the results and considered several potential school factors (e.g., school size, school characteristics, and student demographics) but could not isolate factors that shed light on the findings. As part of another project, we revisited all the cities, schools, and neighborhoods in the hope that we could observe neighborhood characteristics that could help explain the disparate results, but again, we gained no satisfactory insights.

We also examined the possibility that the site differences reflect differential program implementation fidelity.[13] Fortunately our research design allowed us to examine this possibility as we went to great lengths to assess officer implementation fidelity by observing 492 unique G.R.E.A.T. classroom deliveries and assigning a fidelity score (ranging from 1 to 5) to each classroom (for more information on the assessment of implementation fidelity, see Esbensen, Matsuda, Taylor, and Peterson, 2011). Analyses failed to identify significant differential program effects associated with program quality; only 1 of the 33 potential outcomes (attitudes toward officers in the classroom) showed a more favorable outcome for students in classrooms in which officers implemented the program with increased fidelity ($p < .05$). One possibility for the overall null finding is that 27 of the 33 officers implemented the program with good-to-excellent fidelity. Only three officers were deemed to have not implemented the program (one each in Albuquerque, Greeley, and Philadelphia) and three (one each in the DFW site, Nashville, and Chicago) to have marginal implementation. A seventh officer was deemed to have implemented the program in three classrooms, but because of classroom management issues, the officer failed to implement the program in two other classrooms. Given the overall program fidelity, there might have been insufficient statistical power to detect implementation effect.

Program Effects by Preexisting Risk

The findings for preexisting risk are complex but straightforward. Although we did not find any risk-by-treatment interaction effects, we did uncover a pattern of three-way interaction of risk by treatment by time. The three-way interactions suggest that most of the beneficial impact is associated with the high-risk students in the early waves and that the treatment/control difference for high-risk youth fades over time. Some evidence indicates that the treatment is increasingly beneficial for low-risk youths over time, but that pattern is far from consistent. What these findings mean for universal versus targeted gang-prevention programming is therefore somewhat ambiguous, although the suggestion might be that

13. Given the literature regarding the importance of implementation fidelity, we investigated the relationship of program impact to the quality of G.R.E.A.T. program delivery. Each officer was observed an average of 15 times by trained research assistants. To address this question, we added to the base model the two-way interaction of the officer rating with classroom treatment assignment and the three-way interactions of officer rating and treatment assignment with linear and quadratic change. We avoid confounding these interactions with overall treatment effects by grand mean centering the officer rating and assigning all control classrooms the mean officer rating.

high-risk students (as demonstrated in prior research) have greater gains than do low-risk students, especially in the short term, but that low-risk students also receive program benefits.

Conclusions

The research team responsible for the current evaluation conducted the original G.R.E.A.T. studies in the 1990s (an 11-city cross-sectional study and a 6-city longitudinal quasi-experimental study). Our familiarity with the original program and the evaluation designs and subsequent results facilitate our ability to place the current results within the larger context of school-based gang-prevention programs. Although we have familiarity with the program as evaluators and did provide recommendations regarding program content and delivery based on findings from our first evaluation, it is important to emphasize that we have not been involved in program development; our sole role has been as program evaluators. We note that findings of positive program effects are unfortunately rare in independent prevention trials (Eisner, 2009). Our previous studies of the original G.R.E.A.T. curriculum found a 1-year posttreatment program effect in the cross-sectional study (see Esbensen and Osgood, 1999), but no effect was observed at that time period in the longitudinal quasi-experimental design (Esbensen, Osgood, et al., 2001). In that latter study, we did find a sleeper or lagged effect (3 and 4 years posttreatment) for five outcomes: more favorable attitudes to police, lower victimization, more negative attitudes about gangs, more prosocial peers, and less risk-seeking behavior. Contrary to that previous longitudinal study, the current longitudinal experimental study of the revised G.R.E.A.T. curriculum did find a positive program effect 1 year posttreatment. Importantly, three of the lagged program effects found across 4 years posttreatment in the previous study were replicated in this study for effects across the 4 years (more favorable attitudes to the police, more negative attitudes about gangs, and less risk seeking). Whereas the original program had no appreciable short- or long-term effect on gang involvement, an evaluation of the revised program found reduced odds of gang membership (39% for the first 12 months and 24% across the entire 48 months postprogram). Given the results of the current evaluation, it is important to restate that the G.R.E.A.T. program underwent a major review and revision subsequent to the previous evaluation results. The original G.R.E.A.T. program was a "canned" nine-lesson program with an emphasis on didactic teaching methods. The current 13-lesson G.R.E.A.T. curriculum emphasizes skills building and the use of cooperative learning strategies—both strategies borrowed from other school-based "model" or acclaimed programs.

The fact that both evaluations (of the original and revised program) found decidedly more favorable attitudes toward the police among the G.R.E.A.T. students suggests that this kind of law-enforcement–based prevention program can have a positive impact on youth–police relations. This is particularly important given recent findings that perceptions of police legitimacy often are muted among gang members (particularly those embedded in criminal networks), a factor associated with their increased involvement in crime (Papachristos,

Meares, and Fagan, 2012). Also, it is important to note that studies of both the original and the revised curriculum produced evidence that the G.R.E.A.T. program is associated with more negative views of gangs. We view these similarities in findings as suggestive of an overall consistency in the program and further speculate that the additional program effects of the revised G.R.E.A.T. program are likely a result of the revised and enhanced curriculum.

The current study is not without limitations. Study participants were enrolled in public schools in seven U.S. cities. Students who attended private schools, other districts, those whose parents declined participation, and those who were absent during survey administration periods were not included. We attempted to survey as many eligible students as possible, making more than 10 trips to schools to try to reach those who were habitually truant or otherwise unavailable. We also attempted to survey students who transferred schools within the original and adjacent districts; those who moved to districts outside of the original metro areas were typically lost. Consequently, we might have lost a disproportionate share of gang members and other "at-risk" youth. Additionally, we have no alternative measures of delinquency or gang membership other than the students' self-reports. Future studies might find it useful to collect measures of school disciplinary reports, police reports, and other indicators.

The G.R.E.A.T. program is no panacea for the gang problems confronting many schools and neighborhoods. However, our findings suggest that G.R.E.A.T. holds promise as a primary gang-prevention program, overall and in several of our seven individual research sites. Although it is important to note that the effect sizes are small (ranging from 0.05 to 0.14 over 48 months posttreatment), it is equally important to emphasize that this is a low-dosage program. The curriculum consists of 13 lessons, generally delivered once a week in less than 40 minutes. Furthermore, the realities of program delivery such as student absenteeism, teacher announcements, fire drills, snow days, officer illness, and shortened day schedules mean that most of the G.R.E.A.T. students do not receive the full recommended dosage. That statistically significant differences were found for 11 outcome measures (and another 3 with marginal significance) 12 months posttreatment and for 10 measures across 4 years posttreatment we find surprising and certainly promising.

References

Alford, Aaron A. and James Derzon. 2012. Meta-analysis and systematic review of the effectiveness of school-based programs to reduce multiple violent and antisocial behavioral outcomes. In (Shane R. Jimerson, Amanda B. Nickerson, Matthew J. Mayer, and Michael J. Furlong, eds.), *Handbook of School Violence and School Safety: International Research and Practice*. New York: Routledge.

Andrews, D.A., Ivan Zinger, Robert D. Hoge, James Bonta, Paul Gendreau, and Francis T. Cullen. 1990. Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, 38: 369–404.

Battin, Sara R., Karl G. Hill, Robert D. Abbott, Richard F. Catalano, and J. David Hawkins. 1998. The contribution of gang membership to delinquency beyond delinquent friends. *Criminology*, 36: 93–115.

Catalano, Richard F., Michael W. Arthur, J. David Hawkins, Lisa Berglund, and Jeffrey J. Olson. 1998. Comprehensive community- and school-based interventions to prevent antisocial behavior. In (Rolf Loeber and David P. Farrington, eds.), *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions*. Thousand Oaks, CA: Sage.

Covey, Herbert C. 2010. *Street Gangs Throughout the World*, 2nd Edition. Springfield, IL: Charles C. Thomas.

Dusenbury, Linda and Gilbert J. Botvin. 1992. Competence enhancement and the development of positive life options. *Journal of Addictive Diseases*, 11: 29–45.

Eisner, Manuel. 2009. No effects in independent prevention trials: Can we reject the cynical view? *Journal of Experimental Criminology*, 5: 163–183.

Esbensen, Finn-Aage and Elizabeth P. Deschenes. 1998. A multisite examination of youth gang membership: Does gender matter? *Criminology*, 36: 799–828.

Esbensen, Finn-Aage, Adrienne Freng, Terrance J. Taylor, Dana Peterson, and D. Wayne Osgood. 2002. Putting research into practice: The national evaluation of the Gang Resistance Education and Training (G.R.E.A.T.) program. In (Winifred L. Reed and Scott H. Decker), *Responding to Gangs: Evaluation and Research*. Washington, DC: U.S. Department of Justice, National Institute of Justice.

Esbensen, Finn-Aage and David Huizinga. 1993. Gangs, drugs, and delinquency in a survey of urban youth. *Criminology*, 31: 565–589.

Esbensen, Finn-Aage, David Huizinga, and Anne W. Weiher. 1993. Gang and non-gang youth: Differences in explanatory factors. *Journal of Contemporary Criminal Justice*, 9: 94–116.

Esbensen, Finn-Aage, Kristy N. Matsuda, Terrance J. Taylor, and Dana Peterson. 2011. Multi-method strategy for assessing program fidelity: The national evaluation of the revised G.R.E.A.T. program. *Evaluation Review*, 35: 14–39.

Esbensen, Finn-Aage and Cheryl L. Maxson. 2012. *Youth Gangs in International Perspective: Results from the Eurogang Program of Research*. New York: Springer.

Esbensen, Finn-Aage, Chris Melde, Terrance J. Taylor, and Dana Peterson. 2008. Active parental consent in school-based research: How much is enough and how do we get it? *Evaluation Review*, 32: 335–362.

Esbensen, Finn-Aage and D. Wayne Osgood. 1999. Gang resistance education and training (GREAT): Results from the national evaluation. *Journal of Research in Crime and Delinquency*, 36: 194–225.

Esbensen, Finn-Aage, D. Wayne Osgood, Terrance J. Taylor, Dana Peterson, and Adrienne Freng. 2001. How great is G.R.E.A.T.? Results from a quasi-experimental design. *Criminology & Public Policy*, 1: 87–118.

Esbensen, Finn-Aage, Dana Peterson, Terrance J. Taylor, and Adrienne Freng. 2010. *Youth Violence: Sex and Race Differences in Offending, Victimization, and Gang Membership*. Philadelphia, PA: Temple University Press.

Esbensen, Finn-Aage, Dana Peterson, Terrance J. Taylor, Adrienne Freng, D. Wayne Osgood, Dena C. Carson, and Kristy N. Matsuda. 2011. Evaluation and evolution of the Gang Resistance Education and Training (G.R.E.A.T.) Program. *Journal of School Violence*, 10: 53–70.

Esbensen, Finn-Aage, Dana Peterson, Terrance J. Taylor, and D. Wayne Osgood. 2012. Results from a multi-site evaluation of the G.R.E.A.T. program. *Justice Quarterly*, 29: 125–151.

Esbensen, Finn-Aage, L. Thomas Winfree, Jr., Ni He, and Terrance J. Taylor. 2001. Young gangs and definitional issues: When is a gang a gang and why does it matter? *Crime & Delinquency*, 47: 105–130.

Farrington, David P. and Rolf Loeber. 2000. Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health*, 10: 100–122.

Finn, Peter and Jack McDevitt. 2005. *National Assessment of School Resource Officer Programs: Final Project Report*. Cambridge, MA: Abt Associates. Retrieved from ncjrs.gov/pdffiles1/nij/grants/209273.pdf.

Gottfredson, Denise C. 2001. *Schools and Delinquency*. New York: Cambridge University Press.

Grasmick, Harold G., Charles R. Tittle, Robert J. Bursik, and Bruce J. Arneklev. 1993. Testing the core empirical implications of Gottfredson and Hirschi's General Theory of Crime. *Journal of Research in Crime and Delinquency*, 30: 5–29.

Gravel, Jason, Martin Bouchard, Karine Descormiers, Jennifer S. Wong, and Carlo Morselli. 2013. Keeping promises: A systematic review and new classification of gang control strategies. *Journal of Criminal Justice*, 41: 228–241.

Hagedorn, John M. 2008. *A World of Gangs: Armed Young Men and Gangsta Culture*. Minneapolis: University of Minnesota Press.

Hill, Karl G., James C. Howell, J. David Hawkins, and Sara R. Battin-Pearson. 1999. Childhood risk factors for adolescent gang membership: Results from the Seattle Social Development Project. *Journal of Research in Crime and Delinquency*, 36: 300–322.

Howell, James C. 2009. *Preventing and Reducing Juvenile Delinquency: A Comprehensive Framework*. Thousand Oaks, CA: Sage.

Howell, James C. 2012. *Gangs in America's Communities*. Thousand Oaks, CA: Sage.

Howell, James C. and Arlen Egley, Jr. 2005. *Gangs in Small Towns and Rural Areas*. NYGC Bulletin. Edited by The Office of Juvenile Justice and Delinquency Prevention. Washington, DC: U.S. Department of Justice.

Jimerson, Shane R., Amanda B. Nickerson, Matthew J. Mayer, and Michael J. Furlong (eds.). 2012. *Handbook of School Violence and School Safety: International Research and Practice*, 2nd Edition. New York: Routledge.

Klein, Malcolm W. and Cheryl L. Maxson. 2006. *Street Gang Patterns and Policies*. New York: Oxford University Press.

Lipsey, Mark W. 2009. The primary factors that characterize effective interventions with juvenile offenders: A meta-analysis review. *Victims & Offenders*, 4: 124–147.

Ludwig, Jens. 2005. Better gun enforcement, less crime. *Criminology & Public Policy*, 4: 677–716.

Maxson, Cheryl L., Arlen Egley, Jr., Jody Miller, and Malcolm W. Klein (eds.). 2013. *The Modern Gang Reader*, 4th Edition. New York: Oxford University Press.

Maxson, Cheryl L., and Monica L. Whitlock. 2002. Joining the gang: Gender differences in risk factors of gang membership. In (C. Ronald Huff, ed.), *Gangs in America III*. Thousand Oaks, CA: Sage.

Maxson, Cheryl, Monica L. Whitlock, and Malcolm W. Klein. 1998. Vulnerability to street gang membership: Implications for practice. *Social Service Review*, March: 70–91.

Mihalic, Sharon F., Abigail A. Fagan, Katherine Irwin, Diane Ballard, and Delbert S. Elliott. 2002. *Blueprints for Violence Prevention Replications: Factors for Implementation Success*. Boulder: University of Colorado, Center for the Study and Prevention of Violence.

Mihalic, Sharon F. and Katherine Irwin. 2003. Blueprints for violence prevention: From research to real-world settings—factors influencing the successful replication of model programs. *Youth Violence and Juvenile Justice*, 1: 307–329.

Na, Chongmin and Denise C. Gottfredson. 2013. Police officers in schools: Effects of school crime and the processing of offending behaviors. *Justice Quarterly*, 30: 619–650.

Office of Juvenile Justice and Delinquency Prevention (OJJDP). 2010. *Model Programs Guide*. Retrieved from ojjdp.gov/mpg/.

Papachristos, Andrew V., Tracy L. Meares, and Jeffrey Fagan. 2012. Why do criminals obey the law? The influence of legitimacy and social networks on active gun offenders. *Journal of Criminal Law & Criminology*, 102: 397–440.

Peterson, Dana and Kirstin A. Morgan. 2013. Sex differences and the overlap in risk factors for gang membership and violence. *Journal of Crime and Justice.* DOI:10.1080/0735648X.2013.830393

Petteruti, Amanda. 2011. *Education Under Arrest: The Case Against Police in Schools*. Washington, DC: Justice Policy Institute.

Pyrooz, David C., Andrew M. Fox, and Scott H. Decker. 2010. Racial and ethnic heterogeneity, economic disadvantage, and gangs: A macro-level study of gang membership in urban America. *Justice Quarterly*, 27: 867–892.

Rasbash, Jon, Fiona Steele, William J. Browne, and Harvey Goldstein. 2009. *A User's Guide to MLwiN, Version 2.10*. Bristol, U.K.: Bristol University Press.

Reed, Winifred L., and Scott H. Decker. 2002. *Responding to Gangs: Evaluation and Research*. Washington, DC: U.S. Department of Justice, National Institute of Justice.

Sherman, Lawrence W., Denise C. Gottfredson, Doris MacKenzie, John Eck, Peter Reuter, and Shawn D. Bushway. 1997. *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington, DC: Department of Justice, Office of Justice Programs.

Thornberry, Terence P. 1998. Membership in youth gangs and involvement in serious and violent offending. In (Rolf Loeber and David P. Farrington, eds.), *Serious and Violent Juvenile Offenders: Risk Factors and Successful Interventions*. Thousand Oaks, CA: Sage.

Thornberry, Terence P., Marvin D. Krohn, Alan J. Lizotte, C. A. Smith, and Kimberly Tobin. 2003. *Gangs and Delinquency in Developmental Perspective*. New York: Cambridge University Press.

Winfree, L. Thomas, Jr., Dana Peterson Lynskey, and James R. Maupin. 1999. Developing local police and federal law enforcement partnerships: G.R.E.A.T. as a case study of policy implementation. *Criminal Justice Review*, 24: 145–168.

Appendix

Scale Characteristics of Outcome Measures (Wave 1)

Impulsivity: Four items such as: I often act without stopping to think.
Scale mean = 2.97 (0.81); $\alpha = 0.59$
Response categories: 1) strongly disagree to 5) strongly agree

Risk-Seeking: Four items including: I like to test myself every now and then by doing something a little risky.
Scale mean = 2.60 (0.95); $\alpha = 0.77$
Response categories: 1) strongly disagree to 5) strongly agree

Anger: Four items including: I lose my temper pretty easily.
Scale mean = 3.08 (0.96); $\alpha = 0.74$
Response categories: 1) strongly disagree to 5) strongly agree

Self-Centeredness: Four items such as: If things I do upset people, it's their problem not mine.
Scale mean = 2.50 (0.82); $\alpha = 0.69$
Response categories: 1) strongly disagree to 5) strongly agree

Attitudes Toward Police: Six items such as: Police officers are honest.
Scale mean = 3.81 (0.82); $\alpha = 0.86$
Response categories: 1) strongly disagree to 5) strongly agree

G.R.E.A.T. ATP: Two items such as: Police officers make good teachers.
Mean = 3.58 (0.95)
Response categories: 1) strongly disagree to 5) strongly agree

Prosocial Peers: Four items including: During the last year, how many of your current friends have been generally honest and told the truth?
Scale mean = 3.42 (0.97); $\alpha = 0.83$
Response categories: 1) none of them, 2) few of them, 3) half of them, 4) most of them, and 5) all of them

Peer Pressure: Seven items such as: How likely is it that you would go along with your current friends if they wanted you to bully another student at school?

Scale mean = 1.27 (0.51); $\alpha = 0.82$
Response categories: 1) not at all likely to 5) very likely

Negative Peer Commitment: Three items including: If your group of friends was getting you into trouble at home, how likely is it that you would still hang out with them?
Scale mean = 1.68 (0.85); $\alpha = 0.81$
Response categories: 1) not at all likely to 5) very likely

Positive Peer Commitment: Two items including: If your group of friends told you *not* to do something because it was wrong, how likely is it that you would listen to them?
Scale mean = 4.19 (1.17); $\alpha = 0.80$
Response categories: 1) not at all likely to 5) very likely

Delinquent Peers: Seven items including: During the last year, how many of your current friends have attacked someone with a weapon?
Scale mean = 1.30 (0.54); $\alpha = 0.86$
Response categories: 1) none of them, 2) few of them, 3) half of them, 4) most of them, and 5) all of them

Lying Neutralizations: Three items such as: It's okay to tell a small lie if it doesn't hurt anyone.
Scale mean = 2.60 (0.98); $\alpha = 0.76$
Response categories: 1) strongly disagree to 5) strongly agree

Stealing Neutralizations: Three items such as: It's okay to steal something it that's the only way you could ever get it.
Scale mean = 1.64 (0.80); $\alpha = 0.83$
Response categories: 1) strongly disagree to 5) strongly agree

Hitting Neutralizations: Three items such as: It's okay to beat up someone if they hit you first.
Scale mean = 3.32 (1.11); $\alpha = 0.80$
Response categories: 1) strongly disagree to 5) strongly agree

School Commitment: Seven items such as: I try hard in school.
Scale mean = 3.92 (0.70); $\alpha = 0.77$
Response categories: 1) strongly disagree to 5) strongly agree

Guilt: Seven items including: How guilty or how badly would you feel if you stole something worth *less* than $50?
Scale mean = 2.66 (0.55); $\alpha = 0.93$
Response categories: 1) not very guilty/badly, 2) somewhat guilty/badly, and 3) very guilty/badly

Conflict Resolution: Five items including: During the past year when you've gotten upset with someone, how often have you talked to the person about why you were upset?
Scale mean = 2.17 (0.46); $\alpha = 0.66$
Response categories: 1) never, 2) sometimes, and 3) often

Calming Others: Three items including: When someone was upset, how often have you asked the person why he/she was upset during the past year?
Scale mean = 2.41(0.51); $\alpha = 0.71$
Response categories:1) never, 2) sometimes, and 3) often

Refusal Skills: Four items including: During the past year when you have tried to avoid doing something your friends tried to get you to do, how often have you told the person that I can't do it because my parents will get upset with me.
Scale mean = 2.33 (0.51); $\alpha = 0.70$
Response categories: 1) never, 2) sometimes, and 3) often

Prosocial Involvement Index: Four items including: During the past year, were you involved in school activities, or athletics?
Scale mean: 2.38 (1.14); $\alpha = 0.47$
Response categories: 1) yes and 2) no

Empathy: Five item including: I would feel sorry for a lonely stranger in a group.
Scale mean = 3.63 (0.65); $\alpha = 0.59$
Response categories: 1) strongly disagree to 5) strongly agree

Active Listening: Three items such as: I look at the person talking to me.
Scale mean = 3.66 (0.72); $\alpha = 0.60$
Response categories: 1) strongly disagree to 5) strongly agree

Problem Solving: Two items including: I talk to my friends about my problems.
Scale mean = 3.57 (0.91); $\alpha = 0.45$
Response categories: 1) strongly disagree to 5) strongly agree

Awareness of Services: Four items including: You know where a person can go for help if he/she is victimized.
Scale mean = 3.76 (0.65); $\alpha = 0.72$
Response categories: 1) strongly disagree to 5) strongly agree

Collective Efficacy: Three items including: It is my responsibility to do something about problems in our community.
Scale mean = 3.25 (0.77); $\alpha = 0.62$
Response categories: 1) strongly disagree to 5) strongly agree

Attitudes about Gangs: Two items: Getting involved with gangs will interfere with reaching my goals.

Scale mean = 3.72 (1.12); $\alpha = 0.71$
Response categories: 1) strongly disagree to 5) strongly agree

Altruism: Three items including: It feels good to do something without expecting anything in return.
Scale mean = 3.60 (0.83); $\alpha = 0.66$
Response categories: 1) strongly disagree to 5) strongly agree

---

Finn-Aage Esbensen is the E. Desmond Lee Professor of Youth Crime and Violence and also serves as Chair of the Department of Criminology and Criminal Justice at the University of Missouri—St. Louis. He also serves on the Steering Committee of the Eurogang Program of Research.

D. Wayne Osgood is a professor of criminology and sociology at Pennsylvania State University and lead editor of *Criminology*. His current research interests include relationships between adolescent friendship networks and delinquency, connections between time use and problem behavior, and effects of prevention programs.

Dana Peterson is an associate professor in the School of Criminal Justice, University at Albany (New York) and conducts research primarily in the areas of youth gangs and violence, with particular interest in how these are structured by sex and gender.

Terrance J. Taylor is an associate professor in the Department of Criminology and Criminal Justice at the University of Missouri—St. Louis. His primary research interest involves youth crime and violence. He received his Ph.D. in criminal justice from the University of Nebraska in 2002.

Dena C. Carson is an assistant research professor in the Department of Criminology and Criminal Justice at the University of Missouri—St. Louis. Her general research interests include youth violence, victimization, gangs, and delinquent peer groups. Her recent publications have appeared in *Youth & Society*, *Journal of Criminal Justice*, and *Youth Violence & Juvenile Justice*.

EVALUATION OF THE G.R.E.A.T. PROGRAM

GREAT Results

Implications for PBIS in Schools

James C. Howell
National Gang Center

Esbensen, Osgood, Peterson, Taylor, and Carson (2013, this issue) report four important outcomes from a multisite (seven cities) 4-year follow-up on the effects of the school-based Gang Resistance Education and Training (G.R.E.A.T.) curriculum:

1. Reduced odds of gang joining
2. A positive impact on youth-police relations
3. More negative views of gangs
4. Improvements in several risk factors for delinquency involvement and gang joining

Our comments focus on the policy and program implications of these findings for school systems. Important considerations include the following. Foremost, school administrators should consider the program efficacy of the G.R.E.A.T. curriculum. Is it worth school classroom time? Does the program consistently reach its outcome goals? How does the school benefit from it at the student level and schoolwide? The main aim in this article is to assist school officials in finding answers to these vitally important questions for evidence-based programming.

School administrators are concerned with providing safe learning environments and creating a positive school climate. To help realize these goals, school systems across the United States are implementing the federal Office of Special Education Program's Positive Behavioral Interventions and Supports (PBIS) framework (pbis.org). A three-tiered model for instruction and intervention, PBIS is based on the principle that academic and behavioral supports must be provided at a school-wide level to address the needs of all students in a school effectively (referred to as Tier 1; core, universal instruction, and supports). Because

Direct correspondence to James C. Howell, National Gang Center, P.O. Box 12729, Tallahassee, FL 32317 (e-mail: bhowell@iir.com).

DOI:10.1111/1745-9133.12051

not all students will respond to the same curricula and teaching strategies, some students with identified needs receive supplemental or targeted instruction and intervention in Tier 2 (targeted, supplemental interventions and supports). Last, in Tier 3, intensive, individualized interventions and supports are provided for the relatively few students with the most severe needs, requiring intensive and individualized behavioral treatment and academic support.

The G.R.E.A.T. program has produced several beneficial outcomes that are called for in Tier 1 of the PBIS framework. Thus, it could be incorporated within this framework in any school system to help achieve PBIS delinquency prevention goals and specifically where gang problems exist in the schools themselves and in the surrounding community to prevent gang joining. We first summarize G.R.E.A.T. outcomes, followed by suggested policy and program implications.

G.R.E.A.T. Outcomes

In addition to improvements in several risk factors—such as having less anger, more use of refusal skills, and less risk-seeking among elementary and middle school students—the school-based curriculum generated positive attitudes toward police and less positive attitudes about gangs, and it reduced the odds of gang joining among racially/ethnically diverse groups of youth by 24%. It is noteworthy that these results held up over a 4-year follow up. These findings are reinforced when the evaluation methodology is taken into account. Other strengths of the evaluation such as independent evaluators, randomized control trials, multisite and school selection with the goal to develop a sample that was geographically and demographically diverse, close attention to program implementation with fidelity, examination of treatment interaction effects by preexisting risk factors, and an extraordinary effort that succeeded in maintaining a very high study participation rate (72% of the study youth, 4 years posttreatment). Given these strenuous scientific procedures that meet "The Maryland Scale of Scientific Methods" (Sherman, et al., 1998) and the evidence of a large reduction in gang joining, this G.R.E.A.T. program evaluation meets the rigorous Blueprint standards for proven effectiveness (see Mihalic, Irwin, Elliott, Fagan, and Hansen, 2001). In addition, as shown in the subsequent discussion, the respectable effect sizes for other measured outcomes are remarkable for such a short-term program, producing large cost-benefits because schools are not required to remunerate law enforcement agencies or school resource officers who typically would be on school premises in the absence of G.R.E.A.T.

Esbensen et al. (2013) attribute the favorable outcomes in large part to the research base underlying the revised version of the G.R.E.A.T. curriculum they evaluated. It mainly was improved by incorporating some of the skill-development strategies employed in the evidence-based LifeSkills Training program (Dusenbury and Botvin, 1992).[1] In addition to

1. A classroom-based tobacco-, alcohol-, and drug-abuse-prevention program for upper elementary and junior high school students.

educating students about the dangers of gang involvement, the G.R.E.A.T. lesson content places considerable emphasis on cognitive-behavioral training, social skills development, refusal skills training, and conflict resolution. Although the G.R.E.A.T. lesson content is delivered in varying formats, the primary service is skills-based training, which typically shows relatively modest reductions in delinquency (Lipsey, 2009). The target service dosage for effective social skills training programs for juvenile delinquency reduction is 16 weeks and 24 hours of service (Lipsey and Chapman, 2013). In contrast, the G.R.E.A.T. curriculum consists of just 13 lessons of less than 1 hour each. Thus, it is impressive that Esbensen et al. report effect sizes ranging from .05 to .14 of measured outcomes for G.R.E.A.T. such as lower levels of risk-seeking and anger. Although these reductions are modest, they are not trivial by any means, particularly given the potential for school-wide impacts. Reducing risk factors and antisocial attitudes leads to delinquency reduction and, in turn, gang involvement (Howell and Egley, 2005; Thornberry, Krohn, Lizotte, Smith, and Tobin, 2003).

It also is noteworthy that G.R.E.A.T. produced changes in attitudes toward gangs and police. Many studies show that key psychological variables within the individual risk domain include attitudes accepting of antisocial behavior, delinquency, or drug use, and antiestablishment attitudes and beliefs (antithetical to conventional society, antiauthority) (Tanner-Smith, Wilson, and Lipsey, 2013). These attitudes, along with negative attitudes toward school, are particularly important for subsequent delinquency when formed during the middle school years (ages 11–14). Changes in unconventional beliefs and favorable attitudes toward delinquency almost always precede reductions in delinquency and other problem behaviors (Mulvey et al., 2004). Therefore, the favorable attitudes toward police and unfavorable attitudes toward gangs that G.R.E.A.T. produced are important.

The most salutary accomplishment of the G.R.E.A.T. curriculum, however, is having reduced gang joining with such a short-term intervention. This achievement surely can be attributed to the strength of the skill-based instruction and the high fidelity with which the G.R.E.A.T. curriculum was implemented. Esbensen et al. (2013) note results from an earlier report on fidelity measures that 27 of the 33 officers implemented the program with good-to-excellent fidelity (Esbensen, Matsuda, Taylor, and Peterson, 2011). This finding is based on an unusually thorough assessment of fidelity of the G.R.E.A.T. program in three key areas in which implementation quality could be expected to suffer: officer preparedness and commitment to the program (i.e., program provider training), support and involvement of educators, and program delivery (i.e., officers' actual ability to deliver the program in the schools as designed). The fidelity assessment came from unusually meticulous procedures including (a) observations of G.R.E.A.T. officer trainings to assess the quality of the training that officers receive before being sent into classrooms, (b) surveys and interviews of G.R.E.A.T.-trained officers and supervisors to determine their own perceptions of preparedness and the level of commitment to delivering the program, (c) surveys of school personnel to evaluate officers' abilities as instructors and educators' involvement in

the program, and (d) approximately 500 direct onsite observations of 33 different officers as they delivered the G.R.E.A.T. program in 31 schools in seven cities. Each classroom included in the study was observed multiple times to assess any fidelity loss and any slippage was promptly addressed. Attributable to these quality assurance procedures, and the fundamental quality of service delivery, the current evaluation revealed no significant differences in program outcomes across sites associated with the fidelity with which the G.R.E.A.T. curriculum was implemented.

It is encouraging that high fidelity was achieved across multiple school systems given the enormous difficulty encountered in mounting programs in schools. A national assessment of gang programs in schools (Gottfredson and Gottfredson, 2001) found that common weaknesses include adopting programs without doing careful planning to match school needs, poorly implementing programs with little supervision, and failing to engage youths who are at highest risk of gang involvement. In addition, gangs and other crime problems disproportionately occur in areas of concentrated disadvantage, where many schools have high dropout rates, high rates of teacher turnover, and difficulty attracting and retaining good teachers, and where teachers do not receive the administrative support they need (Gottfredson, 2013). Importing whole-cloth programs into such school settings is difficult, largely for these reasons and because of disruptions to essential academic instruction.

The G.R.E.A.T. skills-based curriculum is more transportable to school environments than other evidence-based programs because the instructors are law enforcement officers who teach the lesson content to entire elementary and middle school classrooms. No other evidence-based program uses law enforcement officers to deliver therapeutic interventions. Thus, from a practical utility viewpoint, the G.R.E.A.T. curriculum has very high transportability potential. In addition to many U.S. school systems over the past decade, the G.R.E.A.T. program has been implemented successfully within schools in El Salvador, Guatemala, Nicaragua, Costa Rica, Panama, Belize, and Honduras.

Policy and Program Implications

PBIS is a framework that guides the selection, integration, and implementation of the best academic and behavioral practices for improving important academic and behavioral outcomes for all students (Sugai and Simonsen, 2012). The three-tiered PBIS support system "allows educators to identify the needs of all students, match the level of support to the severity of the academic and behavior problems, and then assess the students' response to instruction and intervention" (Florida's Positive Behavior Support Project, 2011: 3). In sum, PBIS is a framework for organizing a continuum of evidence-based approaches that have been assessed to be suitable for a school and its characteristics. In Tier 1, the focus is on academic and behavior interventions that are important for all students. At this level, purposes include building a positive school-wide climate; at the individual level, purposes include teaching anger management and social skills to address those risk

factors that negatively impact student behavior and student success. Such a positive youth development approach to prevention and intervention—also called the developmental assets framework—supports a strength-based approach to child development with an emphasis on building resilience (similar to the concept of protective factors) (Edwards, Mumford, and Serra-Roldan, 2007).

Because the G.R.E.A.T. curriculum produced improvements in several risk factors—such as having less anger, more use of refusal skills, and less risk-seeking among elementary and middle school students—this program has merit for adoption within Tier 1 of the PBIS framework as a universal prevention program. As an effective program for reducing gang membership, improvements in student attitudes toward law enforcement, and reductions in several risk factors for delinquency and joining gangs, the G.R.E.A.T. curriculum fits very well within Tier 1 of the PBIS framework. Individually targeted programs (in Tier 2) and intensive programs (in Tier 3) address students' specific intervention needs (pbis.org). For those students potentially in need of treatment, PBIS is a means to address emotional behavior disabilities. Henderson's (2007) Resiliency Wheel provides a useful protocol for screening, assessment, and crisis intervention. The goal is to use a continuum of evidenced-based interventions to achieve goals for well-defined outcomes. However, fidelity of program implementation is of paramount importance (Gottfredson, 2013).

Community-wide benefits also can be achieved from integrating G.R.E.A.T. with other anti-gang programming. Although Esbensen et al. (2013) emphatically state that the G.R.E.A.T. program is no panacea for the gang problems confronting many schools and neighborhoods, preventing gang joining is essential, and G.R.E.A.T. promises to help reduce gang involvement. Youngsters often are exposed to gangs and gang culture in multiple segments of their lives before the end of childhood, including their families, neighborhoods, schools, and the youth subculture. G.R.E.A.T. also is highly compatible with strategies that mitigate other risk factors in the environment and build resiliency to them.

Thus, it is important for communities that currently experience gang activity to address this in a seamless continuum of juvenile delinquency prevention and gang-oriented strategies and services, beginning with troubled families and disruptive children (Howell, 2010; Wyrick, 2006; Wyrick and Howell, 2004). For an added benefit in communities that have substantial gang activity, the Office of Juvenile Justice and Delinquency Prevention (OJJDP) Comprehensive Gang Prevention, Intervention, and Suppression Model helps communities develop a continuum of gang prevention, intervention, and suppression programs and strategies (National Gang Center, 2010). A balanced and integrated approach is most likely to be effective (Cahill and Hayeslip, 2010; Hayeslip and Cahill, 2009; Spergel, 2007; Spergel, Wa, and Sosa, 2006). Before choosing any programs, services, or activities, communities and neighborhoods that have gangs should complete a comprehensive assessment that identifies elevated risk factors for gangs and how gangs affect the local community. An assessment protocol is available to assist communities in conducting such an assessment through the

OJJDP Comprehensive Gang Model, and an implementation guide is available (Office of Juvenile Justice and Delinquency Prevention, 2009a, 2009b).

Summary

Esbensen et al.'s (2013) evaluation of the G.R.E.A.T. curriculum has demonstrated that when implemented with fidelity, it reduces gang membership, improves student attitudes toward law enforcement, and reduces several risk factors for joining gangs. It should not be expected to do more than that as a primary prevention program. Importantly, this evaluation demonstrates the G.R.E.A.T. program's effectiveness across a geographically and demographically diverse sample of classrooms in seven cities in separate states.

These findings suggest that the G.R.E.A.T. program fits well within a continuum of strategies that are part of the PBIS framework, as a school-based primary prevention program. Because the G.R.E.A.T. curriculum showed improvements in students' social skills—such as having less anger, more use of refusal skills, and less risk-seeking among elementary and middle school students—it can make a useful contribution as a universal prevention program. In particular, the G.R.E.A.T. curriculum fits nicely within the PBIS Tier 1 focus at the individual level, teaching anger management and social skills for all students. Any community that has a gang problem should consider providing G.R.E.A.T. in both its elementary and middle schools because this curriculum has an added benefit of preventing gang joining.

References

Cahill, Megan and David Hayeslip. 2010. *Findings from the Evaluation of OJJDP's Gang Reduction Program. Juvenile Justice Bulletin*. Washington, DC: U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention.

Dusenbury, Linda and Gilbert J. Botvin. 1992. Competence enhancement and the development of positive life options. *Journal of Addictive Diseases*, 11: 29–45.

Edwards, Oliver W., Vincent E. Mumford, and Rut Serra-Roldan. 2007. A positive youth development model for students considered at-risk. *School Psychology International*, 28: 29–45.

Esbensen, Finn-Aage, Kristy N. Matsuda, Terrance J. Taylor, and Dana Peterson. 2011. Multi-method strategy for assessing program fidelity: The national evaluation of the revised G.R.E.A.T. program. *Evaluation Review*, 35: 14–39.

Esbensen, Finn-Aage, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Dena C. Carson. 2013. Short- and long-term outcome results from a multisite evaluation of the G.R.E.A.T. program. *Criminology & Public Policy*, 12: 375–411.

Florida's Positive Behavior Support Project. 2011. *Implementing a Multi-Tiered System of Support for Behavior: A Practical Guide*. Tampa, FL: University of South Florida, Florida's Positive Behavior Support Project.

Gottfredson, Gary D. 2013. What can schools do to help prevent gang joining? In (Thomas R. Simon, Nancy M. Ritter, and Reshma R. Mahendra, eds.), *Changing Course:*

Preventing Gang Membership. Washington, DC: U.S. Department of Justice, U.S. Department of Health and Human Services.

Gottfredson, Gary D. and Denise C. Gottfredson. 2001. *Gang Problems and Gang Programs in a National Sample of Schools*. Ellicott City, MD: Gottfredson Associates.

Hayeslip, David and Megan Cahill. 2009. *Community Collaboratives Addressing Youth Gangs: Final Evaluation Findings from the Gang Reduction Program*. Washington, DC: Urban Institute.

Henderson, Nan. 2007. *Resiliency in Action: Practical Ideas for Overcoming Risks and Building Strengths in Youth, Families, and Communities*. San Diego, CA: Resiliency in Action.

Howell, James C. 2010. *Gang Prevention: An Overview of Current Research and Programs. Juvenile Justice Bulletin*. Washington, DC: U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention.

Howell, James C. and Arlen Egley, Jr. 2005. Moving risk factors into developmental theories of gang membership. *Youth Violence and Juvenile Justice*, 3: 334–354.

Lipsey, Mark W. 2009. The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims & Offenders*, 4: 124–147.

Lipsey, Mark W. and Gabrielle L. Chapman. 2013. *Standardized Program Evaluation Protocol (SPEP): A Users Guide*. Nashville, TN: Vanderbilt University, Peabody Research Institute. Retrieved September 21, 2013 from peabody.vanderbilt.edu/research/pri/publications.php.

Mihalic, Sharon, Katherine Irwin, Delbert Elliott, Abigail Fagan, and Diane Hansen. 2001. *Blueprints for Violence Prevention. Juvenile Justice Bulletin*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Mulvey, Edward P., Laurence Steinberg, Jeffrey Fagan, Elizabeth Cauffman, Alex R. Piquero, Laurie Chassin, et al. 2004. Theory and research on desistance from antisocial activity among serious adolescent offenders. *Youth Violence and Juvenile Justice*, 2: 213–236.

National Gang Center. 2010. *Best Practices to Address Community Gang Problems: OJJDP's Comprehensive Gang Model*. Washington, DC: Author. Retrieved September 21, 2013 from nationalgangcenter.gov/Publications.

Office of Juvenile Justice and Delinquency Prevention. 2009a. *OJJDP's Comprehensive Gang Model: A Guide to Assessing Your Community's Youth Gang Problem*. Tallahassee, FL: National Youth Gang Center, Institute for Intergovernmental Research. Retrieved September 21, 2013 from nationalgangcenter.gov/Publications.

Office of Juvenile Justice and Delinquency Prevention. 2009b. *OJJDP's Comprehensive Gang Model: Planning for Implementation*. Tallahassee, FL: National Youth Gang Center, Institute for Intergovernmental Research. Retrieved September 21, 2013 from nationalgangcenter.gov/Publications.

Sherman, Lawrence W., Denise Gottfredson, Doris MacKenzie, John Eck, Peter Reuter, and Shawn D. Bushway. 1998. *Preventing Crime: What Works, What Doesn't, What's Promising*. Research in Brief. Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Spergel, Irving A. 2007. *Reducing Youth Gang Violence: The Little Village Gang Project in Chicago*. Lanham, MD: AltaMira Press.

Spergel, Irving A., Kwai Ming Wa, and Rolando V. Sosa. 2006. The comprehensive, community-wide, gang program model: Success and failure. In (James F. Short, Jr. and Lorine A. Hughes, eds.), *Studying Youth Gangs*. Lanham, MD: AltaMira Press.

Sugai, George and Brandi Simonsen. 2012. *Positive Behavioral Interventions and Supports: History, Defining Features, and Misconceptions*. Storrs, CT: Center for Positive Behavioral Interventions and Supports, University of Connecticut.

Tanner-Smith, Emily E., Sandra J. Wilson, and Mark W. Lipsey. 2013. Risk factors and crime. In (Francis T. Cullen and Pamela Wilcox, eds.), *The Oxford Handbook of Criminological Theory*. New York: Oxford University Press.

Thornberry, Terrence P., Marvin D. Krohn, Alan J. Lizotte, Carolyn A. Smith, and Kimberly Tobin. 2003. *Gangs and Delinquency in Developmental Perspective*. New York: Cambridge University Press.

Wyrick, Phelan A. 2006. Gang prevention: How to make the "front end" of your anti-gang effort work. *United States Attorneys' Bulletin*, 54: 52–60.

Wyrick, Phelan A. and James C. Howell. 2004. Strategic risk-based response to youth gangs. *Juvenile Justice*, 10: 20–29.

James C. Howell is a senior research associate with the National Gang Center in Tallahassee, FL. His gang publication topics include street gang history, gang homicides, drug trafficking, gangs in schools, myths about gangs, risk factors, and what works in preventing gang activity and reducing gang crime. He assists states and localities in addressing gang problems in a balanced approach using the Comprehensive Gang Prevention, Intervention, and Suppression Model.

EVALUATION OF THE G.R.E.A.T. PROGRAM

Do Not Shoot the Messenger

The Utility of Gang Risk Research in Program Targeting and Content

Cheryl L. Maxson
University of California, Irvine

Finn-Aage Esbensen tells the story of walking into a room full of cops to deliver the news that the initial Gang Resistance Education and Training (G.R.E.A.T.) program developed by local police officers and brought to scale by federal law enforcement just was not working (personal communication, September 26, 2013). Instead of reaching for their guns, the program managers deserve a lot of credit for asking "what can we do to make it better?" In concluding comments to their systematic review of gang control strategies, Gravel, Bouchard, Descormiers, Wong, and Morselli (2013: 240) noted:

> It is crucial now more than ever to be able to give an answer to the dreaded question "what works?" But, perhaps more importantly, gang researchers should also be able to answer two questions: "Why is it not working?" And, "how can we make it work?"

G.R.E.A.T. provides a textbook example of how good evaluation can positively impact program content and outcomes. Not only did Esbensen deliver this news, but also he led a team to advise on appropriate program alterations and produced a massive, multi-method evaluation of the revised G.R.E.A.T. program. Along the way, he and his team have mined these data to make contributions to the academic and practice literatures on street gangs, delinquency, school violence, theory, and evaluation methods, among others via a book, more than 50 articles and chapters, and six doctoral dissertations (Finn-Aage Esbensen, personal communication, September 17, 2013). So, along with his colleagues and students, a broad audience has benefited from the fact that the messenger was not shot that

Direct correspondence to Cheryl L. Maxson. Department of Criminology, Law and Society, School of Social Ecology, University of California, Irvine, 2309 Social Ecology II, Irvine, CA 92697-7050 (e-mail: cmaxson@uci. edu).

DOI:10.1111/1745-9133.12052

day. At a time when the field is struggling to produce evidence of effective programs to prevent or reduce gang membership or activity, Esbensen, Osgood, Peterson, Taylor, and Carson's (2013, this issue) article is a welcome addition to the meager batch of studies that report positive results on specifically *gang* outcomes.

Evidence of effective gang prevention is rare. Gravel et al. (2013) combed through 284 "possible" evaluation reports to identify just 11 programs that aimed to reduce gang risk or prevent gang joining, which had been evaluated with a scientifically sound design (see also Wong, Gravel, Bouchard, Morselli, and Descormiers, 2011). Notwithstanding the paucity of independent program evaluations, why have so pitifully few generated positive results? In the past, I have argued that the path to effective gang prevention is not in universal or primary prevention programs but in targeting just those youth at high risk for gang membership with intensive interventions (Hennigan and Maxson, 2012; Klein and Maxson, 2006). Later in this essay, I will describe two current efforts to improve practice along these lines. Both of these ventures rely on using gang risk studies to select youth most likely to benefit from gang programming. In the following discussion, I argue that the findings presented by Esbensen et al. (2013) imply that these efforts toward more specialized client targeting represent promising directions for gang-prevention practitioners.

In this article, Esbensen et al. (2013) have provided us with definitive and statistically sophisticated analyses of four annual posttreatment assessments, extended with supplemental analyses by site and by preprogram risk levels. Esbensen et al. seem to be surprised by the basic finding: This low-dose, relatively inexpensive, relatively easy-to-deliver program seems to produce sustained positive effects on a broad spectrum of youth. Esbensen et al. report that they replicated their findings by sex and race/ethnic background, which suggests that this program is effective in diverse youth populations. Youth in 6th- or 7th-grade classrooms with G.R.E.A.T. are 39% less likely to identify as gang members after 1 year and are 24% less likely to join gangs over the 4-year postprogram period than control students.

That news is truly welcome for practitioners across the country grappling with gang problems. From the 2011 National Youth Gang Survey, Egley and Howell (2013) estimated that 3300 jurisdictions had gang problems. Seventeen percent of the agencies reporting gang activity in this survey delivered the G.R.E.A.T. program in local schools. Consequently, more than 550 law enforcement agencies across all regions of the United States now have the evidence that their investment in G.R.E.A.T. is paying off.

The program also achieved a second goal of producing more positive attitudes toward the police but did not accomplish the third goal of reducing crime and violence. As we learn more and more about the importance of perceptions of legitimacy (Hough and Maffei, 2013; Papachristos, Meares, and Fagan, 2012), fostering more positive attitudes toward law enforcement is a reasonable program goal and a laudatory accomplishment. However, G.R.E.A.T. is primarily perceived as a *gang* program, so that is where my focus will lie in this essay.

Given the well-documented relationship between gang membership and delinquency, it is surprising that the researchers detected no significant program effects on violence or other forms of crime. These patterns raise questions about the mechanisms by which this program might have produced the intended outcomes, and that is certainly grist for the mill of additional analyses of these data and future evaluations of G.R.E.A.T. Gravel et al. (2013) reminded us that it is important for practitioners to know not just whether the program works but also how and why. Esbensen et al. (2013) draw on their close familiarity with the program's evolution, extensive investigation of program implementation, and careful analytic strategies to speculate on their findings. They argue the likelihood that the modifications to G.R.E.A.T. that occurred in response to the first set of evaluations created a better program (i.e., with new content that addresses important gang risk factors as revealed by systematic research) that was delivered more effectively (i.e., as guided by active learning principles). The global analysis finds that program effects on gang membership also were accompanied by a handful of differences in attitudes and behaviors favoring the youth who received G.R.E.A.T. It is surprising that these results apparently were not achieved via the mechanism of altered peer relationships, a consistent and strong predictor of gang membership (Klein and Maxson, 2006). Esbensen et al. address this with a comment about low dosage: "[A] few brief lessons encouraging youth to avoid negative peer influences might not be sufficient to overcome these influences to achieve the intended outcome." Yet this same low dosage seems to produce significant effects on risk-seeking behaviors and on such attitudes as altruism and collective efficacy. The larger question is why adolescent choices about friends would be more resistant to change and how the gains in gang membership reduction could be accomplished without modifying these choices.

Esbensen et al. (2013) acknowledge the low effect sizes at multiple points in the article, along with other critiques of the initial program and the first evaluation. Notwithstanding substantial changes in the program content and delivery, and addressing threats to validity from attrition of subjects in the initial evaluation, effect sizes remain low, on both the primary outcomes and the hypothesized mediating attitudes and behaviors. Despite meeting other requirements of the Blueprints Series (i.e., rigorous methods and replication of sustained effects across multiple sites), the modest differences in gang membership between treatment and controls apparently prevented the Blueprint managers from anointing G.R.E.A.T. as "promising," much less, "effective."[1] Esbensen et al. ask the critics: "what effect size is reasonable to expect given the low dosage and the general audience targeted by this program, and how large must the effects be to justify the use of a program requiring such limited investment?" Answering only for myself, I would say that any positive and statistically significant result on the rates of joining street gangs is a good one, and this seems to be a cost-effective program. However, on the matter of the audience to be targeted, I believe the study provides implications not addressed by Esbensen et al.

1. Presumably, G.R.E.A.T. was not under consideration for other outcomes targeted by Blueprints, such as violence, delinquency, or drug use.

The supplemental analysis with the sample partitioned by a pretreatment gang risk indicator shows no significant main effects. This finding suggests that the program is effective for youth who are unlikely to join gangs in any case, as well as for those at higher risk. However, the primary outcomes of gang membership, delinquency (frequency and variety), and—marginally—attitudes toward police as teachers—interact with risk over time. The pattern is evident in gang membership (see Figure 2 of Esbensen et al., 2013), where program effects are marked in the high-risk group through Wave 4 (2 years posttreatment), but there seems to be little difference in effect among the low-risk youth. Esbensen et al. "see suggestions of a beneficial program effect gradually emerging so that at Wave 6, the rate of gang membership was only half as high in the treatment group as the control group," but they are referring to a comparison of approximately 1% of gang joiners among low-risk G.R.E.A.T. participants and 2% of joiners among low-risk controls. The proportion of low-risk youth that identify as gang members does not exceed 4% at any point in the 4-year assessment period. These data suggest that the positive G.R.E.A.T. effect was achieved primarily in youth at a high risk for membership, and this was sustained for 2 years after the program was delivered. Three (i.e., peer pressure, negative peer commitment, and delinquent peers) of the five peer variables also produced significant three-way interactions. It might be that G.R.E.A.T. alters peer selection or relationships among high-risk youth, which produces the effect on gang membership and delinquency. This would be interesting to explore. Because the researchers partitioned risk groups at the top quartile of risk, by definition, 75% of the sample was considered unlikely to join gangs based on pretreatment characteristics—and more than 95% of them did not.

Although this universal prevention program has few negative effects and these surfaced only in one site, still I argue that it is advisable to place programs where they are needed the most and to target those individuals most in need. The effects sizes might have been substantial if more targeting based on gang risk had transpired. I wondered whether perhaps Albuquerque, Philadelphia, and Portland—the only individual sites with significant gang reductions—generated youth samples at higher pretreatment risk relative to other sites?

Appropriate targeting at the place and individual levels is challenging. For school-based programs such as G.R.E.A.T., selecting appropriate sites might involve an assessment of community factors such as neighborhood gang activity or drug use or indicators of poverty and structural disadvantage (Thornberry, Krohn, Lizotte, Smith, and Tobin, 2003). However, the research data on community-level gang risk indicators are not extensive. In contrast, there is a solid research foundation on risk factors for gang joining measured at the individual level. No validated gang risk-assessment instrument is available, although we are testing one currently in Los Angeles. It is possible that such a tool could be used to select targets for G.R.E.A.T. It is not unusual for educational institutions to identify students for special programming, although labeling effects would be a concern.

Our experience in Los Angeles has revealed the challenges of implementing gang risk assessments for program entry (Hennigan, Maxson, Sloane, Kolnick, and Vindel,

2013; Maxson, 2011). Some practitioners resist the strategy of favoring actuarial methods over their clinical judgment. In the early months of our efforts, program providers were frustrated that a large proportion of difficult-to-obtain youth referrals did not meet eligibility requirements. This issue can be minimized with better training of outreach workers and likely referral sources on recognition of gang risk factors, but there is a strong tendency to assume that youth are at a much higher risk of joining gangs than they are (Melde, Gavazzi, McGarrell, and Bynum, 2011). This work is difficult for both researchers and practitioners, so I can appreciate the attraction of universal gang-prevention programs that can avoid the issue of target selection entirely. Perhaps with a low-cost program such as G.R.E.A.T., providing services to all youth to include the relatively few that are likely to join gangs is reasonable, but more intensive and targeted strategies might be more cost-effective because of their potential for larger effect sizes.

We have not yet identified effective gang-prevention programs of this type, but Terence P. Thornberry is leading an effort that shows great promise (Thornberry and Gottfredson, 2013). His team is adapting three treatment-intensive Blueprints model programs, which have been shown by rigorous evaluations to be effective in reducing youth violence, drug use, delinquency, or all of the above for gang prevention. These adaptations are informed by gang research expertise and close coordination with program developers. Randomized clinical trials will be performed to assess the adapted programs' effects on gang and other outcomes. Training of providers for the first adaptation, Functional Family Therapy-Gang, occurred in early fall of 2013, and the new program is being implemented now. It will be several years before we know the outcome of this evaluation and, if successful, whether it can be replicated in another site. Gang adaptations of Multi-Systemic Therapy and Multidimensional Treatment Foster Care are planned for the future. This represents a major investment in establishing a base of evidence that can guide gang-prevention practice.

In the meantime, many more communities will likely adopt G.R.E.A.T. This program manifests a solid logic model that rests on gang scholarship, can be delivered consistently with a high degree of fidelity, and can be delivered in a manner that reflects the best practices of classroom management and active learning. The thoughtful, conservative analyses and interpretation provided by Esbensen et al. (2013) suggests that G.R.E.A.T. is a good addition to the small toolbox of strategies that might be recommended to communities concerned about gang issues.

References

Egley, Arlen, Jr. and James C. Howell. 2013. *Highlights of the 2011 National Youth Gang Survey*. Washington, DC: Office of Juvenile Justice and Delinquency Prevention.

Esbensen, Finn-Aage, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Dena C. Carson. 2013. Short- and long-term outcome results from a multisite evaluation of the G.R.E.A.T. program. *Criminology & Public Policy*, 12: 375–411.

Gravel, Jason, Martin Bouchard, Karine Descormiers, Jennifer S. Wong, and Carlo Morselli. 2013. Keeping promises: A systematic review and a new classification of gang control strategies. *Journal of Criminal Justice*, 41: 228–242.

Hennigan, Karen A. and Cheryl L. Maxson. 2012. New directions for street gang prevention for youth: The Los Angeles experience. In (Joshua Sides, ed.), *Post-Ghettto: Reimagining South Los Angeles*. Berkeley: University of California Press.

Hennigan, Karen A., Cheryl L. Maxson, David C. Sloane, Kathy A. Kolnick, and Flor Vindel. 2013. Identifying high-risk youth for secondary gang prevention. *Journal of Crime and Justice*. E-pub ahead of print.

Hough, Mike and Stefano Maffei. 2013. Thinking about legitimacy. *Criminology in Europe: Newsletter of the European Society of Criminology*, 2: 4–10.

Klein, Malcolm W. and Cheryl L. Maxson. 2006. *Street Gang Patterns and Policies*. New York: Oxford University Press.

Maxson, Cheryl L. 2011. Street gangs. In (James Q. Wilson and Joan Petersilia, eds.), *Crime: Public Policies for Crime Control*. New York: Oxford University Press.

Melde, Chris, Stephen Gavazzi, Edmund McGarrell, and Timothy Bynum. 2011. On the efficacy of targeted gang interventions: Can we identify those most at risk? *Youth Violence and Juvenile Justice*, 9: 279–294.

Papachristos, Andrew V., Tracy L. Meares, and Jeffrey Fagan. 2012. Why do criminals obey the law? The influence of legitimacy and social networks on active gun offenders. *Journal of Criminal Law and Criminology*, 102: 397–440.

Thornberry, Terence P. and Denise C. Gottfredson. 2013. *Blueprints for Gang Prevention*. Final Report to the Office of Juvenile Justice and Delinquency Prevention, U.S. Department of Justice. College Park, MD: Department of Criminology and Criminal Justice, University of Maryland.

Thornberry, Terence P., Marvin D. Krohn, Alan J. Lizotte, Carolyn A. Smith, and Kimberly Tobin. 2003. *Gangs and Delinquency in Developmental Perspective*. Cambridge, U.K.: Cambridge University Press.

Wong, Jennifer S., Jason Gravel, Martin Bouchard, Carlo Morselli, and Karine Descormiers. 2011. *Effectiveness of Street Gang Control Strategies: A Systematic Review and Meta-Analysis of Evaluation Studies*. Ottawa, Ontario: Law Enforcement and Policy Branch, Public Safety Canada.

Cheryl L. Maxson is an associate professor in the Department of Criminology, Law and Society at the University of California's Irvine campus. Her research focuses on patterns of street gang participation and responses to gangs in the United States and Europe. Other publications concern status offenders, youth violence, policing, and community treatment of juvenile offenders. She is a Fellow of the Western Society of Criminology, where she has also been honored with the Paul Tappan and Joseph Lohman awards.

EVALUATION OF THE G.R.E.A.T. PROGRAM

Gangs, Criminal Offending, and an Inconvenient Truth:

Considerations for Gang Prevention and Intervention in the Lives of Youth

David C. Pyrooz
Sam Houston State University

For a theoretically driven school-based program, there is much to like about the findings from the 1- and 4-year evaluations of the Gang Resistance Education and Training (G.R.E.A.T.) program provided by Esbensen, Osgood, Peterson, Taylor, and Carson (2013, this issue). Relative to the control group, students who received the 13 lessons of the G.R.E.A.T. curriculum had (a) more positive scores across a range of attitudinal measures central to several theories of criminal behavior; (b) improved police–youth relationships and lower odds of gang membership, satisfying two of the three primary goals of the program; and (c) long-term positive effects on many of the attitudinal outcomes, as well as police–youth relationships and gang membership. The findings for gang membership are important because they stand in contrast to a large body of work, including the first national evaluation of G.R.E.A.T. (Esbensen, Osgood, Taylor, and Peterson, 2001), making this the first study with rigorous randomized control trial evaluation to have a demonstrable impact on decreasing the rates of gang membership. Despite being a "gang" spinoff of the ineffective Drug Abuse Resistance Education (D.A.R.E.) program, the aggregate results suggest that the more inviting and optimistic acronym is indeed fitting for the G.R.E.A.T. program.

But the picture is not all rosy. The third primary program goal to "prevent violent and criminal activity" was not achieved despite positive program effects on gang membership and numerous mediating mechanisms. This finding flies in the face of the logic of gang

The author would like to thank Scott Decker, Gary Sweeten, and Vince Webb for their comments on this policy essay. Direct correspondence to David C. Pyrooz, Department of Criminal Justice and Criminology, Sam Houston State University, 1806 Avenue J, Huntsville, TX 77341 (e-mail: David.Pyrooz@shsu.edu).

DOI:10.1111/1745-9133.12053

prevention and intervention programming. After all, what is the value of a program that causes some kids to disavow gang membership but does not reduce criminal offending? Furthermore, Esbensen et al. (2013) assessed whether the positive attributes of G.R.E.A.T. were consistent across seven cities. They were not. It is hard to reconcile this mismatch in the program effects across the sites on mediating mechanisms and primary program goals. For example, Albuquerque and Philadelphia drove the 4-year aggregate findings on gang membership, but the former site saw numerous program effects on the mediating mechanisms, whereas the latter site saw next to none. Neither site saw changes in criminal offending.

This policy essay at once recognizes the importance and significance of G.R.E.A.T. in the current state of knowledge on the prevention of gang membership while using this opportunity to consider the null program effect on criminal offending by drawing from contemporary research on gangs. A large body of high-quality scholarship has emerged from the first and second evaluations of G.R.E.A.T., including nearly 100 journal articles, theses and dissertations, books, and book chapters, many of which cover issues related to gangs and gang membership. It is not an overstatement to say that the gang literature would read much differently today in the absence of G.R.E.A.T., with less conceptual and empirical sophistication. This literature has helped shape the current thinking of this essay.

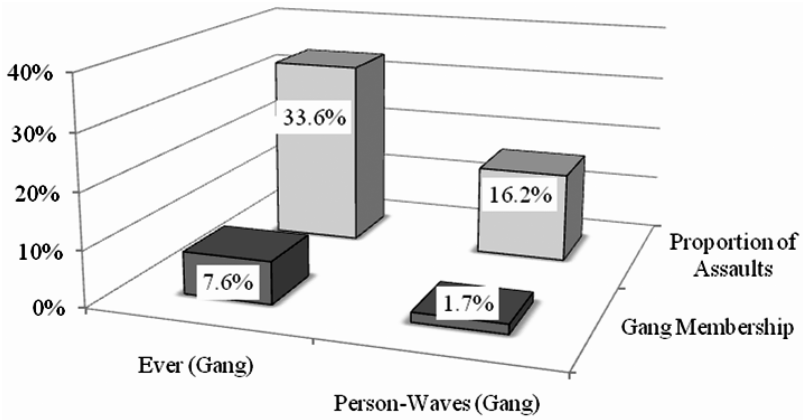
The Logic of Gang Prevention and Intervention and an Inconvenient Truth

The logic for gang prevention is straightforward: Stopping youth from joining gangs will avert a range of consequences linked to gang membership. This logic also extends to gang intervention, where the consequences of gang membership will diminish by shortening the duration of involvement. Effective efforts to reduce the aggregate amount of gang membership—prevalence and frequency—should pay dividends: (a) Individuals will commit fewer criminal acts, experience fewer incidents of victimization, and function better in educational, economic, and family domains (Krohn, Ward, Thornberry, Lizotte, and Chu, 2011; Pyrooz, 2013a; Thornberry, Krohn, Lizotte, Smith, and Tobin, 2003); (b) the rosters of gangs will shrink and the group processes of gangs will weaken, as youth turn to alternative outlets for excitement and protection (Ayling, 2011; Densley, 2012); and (c) there will be fewer "innocent bystanders," less community fear, safer schools, and a reallocation of precious resources to other pressing matters, so the story goes.

Following Wolfgang, Figlio, and Sellin (1972) and the "chronic 6%" of offenders, gang studies have shown the disproportionate contribution of gang membership to crime and delinquency. The numbers are alarming. Thornberry et al. (2003: 50) observed that 31% of youth in the Rochester (NY) study were gang involved but were responsible for 82% of all violent acts. Based on the cross-sectional 11-city data from the first evaluation of G.R.E.A.T., Esbensen, Peterson, Taylor, and Freng (2010: 82) found that the 9% of youth reporting active gang membership committed 36% of the total acts of general violence. These findings

FIGURE 1

Proportion of Self-Reported Incidents of Assault by Two Classifications of Gang Membership in the NLSY97



Note. Assault and gang membership are reported over wave 1 to 5 for the age 12 and 13 cohorts.

on "ever" and "active" gang membership, however, are not limited to high-risk Rochester youth or school-attending youth in the G.R.E.A.T. sites. Figure 1 illustrates this point using national data from the National Longitudinal Survey of Youth 1997 (NLSY97) and ever and active gang membership (see Pyrooz, 2013b) over the first five waves among the 12 and 13 age cohorts (similar to the data reported by Esbensen et al., 2013). Although 8% of respondents who were ever gang *members* were responsible for 34% of the total self-reported incidents of assault, 1.7% of the person-waves of gang *membership* accounted for 16% of all assaults.

Whereas the "gang/crime relationship is the single most robust finding over many decades of research" (Maxson, 2012: 166), gang members and the chronic offenders are not one in the same. That is, not all criminals are gang members and not all gang members are criminals. As Figure 1 illustrates, something unique about gang membership elevates criminal activity, not simply the individuals who join gangs. It is held that group processes and gang socialization are responsible for increasing the probability of engaging in criminal and delinquent behavior (Klein and Maxson, 2006; Thornberry et al., 2003). The disproportionality findings are the principal argument for establishing prevention and intervention efforts in the lives of future or current gang youth and young adults. Therefore, when reviewing Esbensen et al.'s (2013) results, much optimism surrounds the anticipated criminal returns to a 39% and 24% 1-year and 4-year reduction in the odds of gang membership, respectively.

But the inconvenient truth is that G.R.E.A.T. had no effect on the 14-item inventory of delinquency frequency and variety scores. Nor did G.R.E.A.T. have an effect on the 3-item inventory of violent delinquency frequency and variety scores, which included an item asking whether respondents had been "involved in gang fights." The violence findings are particularly curious, as the proximal consequences of gang membership have a firm theoretical and empirical foundation (Decker, 1996; Melde and Esbensen, 2013; Papachristos, 2009). In their now decade-old observations of the literature, Katz and Jackson-Jacobs (2004: 94) called the lack of establishing the causal significance of gangs and gang membership for individuals and communities a "defect in the corpus of gang research." Gang research post-2000 has satisfied much of the causal concerns, using modeling strategies that shield gang effects from observable and unobservable sources of selection (Curry, Decker, and Pyrooz, 2013; Krohn and Thornberry, 2008). Despite the consistency of the findings across sample site, type, demographic group, and crime measure, endogeneity remains a nagging concern in this voluminous literature. After all, gang membership cannot be randomized. But G.R.E.A.T. is exogenous, which raises the question: Why did a reduction in gang membership not correspond with a difference in criminal offending?

Criminologists may offer up several reasons, but the answer must lie in the wrong kind of effect sizes, the wrong kind of people, or the wrong kind of mechanisms. In terms of the *wrong kind of effect sizes*, was the base rate of gang membership too low for statistically and substantively meaningful differences in criminal offending to emerge? A closer inspection of the program effect sizes on gang membership is necessary because gang membership is a rare event. Using the 1-year estimates, there were 206 and 143 pooled (post-test and year-one waves) self-reports of gang membership for control and treatment groups (Esbensen, Peterson, Taylor, and Osgood, 2012: 139). These values indicate that the prevalence of gang membership was 6.4% and 3.8% for the control and treatment groups, respectively, which is a difference of 2.6 percentage points.[1]

Such a difference is numerically small but substantively large enough to rule out the *wrong kind of effect sizes* explanation. When Esbensen et al. (2013) rightfully ask—"what effect size is reasonable to expect given the low dosage and the general audience targeted by this program, and how large must the effects be to justify the use of a program requiring such limited investment?"—a net reduction of 2.6 percentage points in gang membership seems to be a pretty reasonable answer because of the anticipated criminal returns associated with such a reduction. Indeed, the disproportionality findings presented previously show that preventing 63 (pooled) youth from joining a gang should translate in the prevention of *several hundred* criminal offenses.

1. This result is approximate and unadjusted from the regression estimates, assuming equal attrition in person/wave cases across treatment (53.7%) and control (46.3%) groups and no missing data on gang membership. The odds ratio I come up with is slightly larger (41.7% = 100 × [1 – (143 / 3587) / (206 / 3012)]) than reported in the current study.

Did G.R.E.A.T. target the *wrong kind of people* by preventing gang joining or by inducing gang leaving? At the pretest interview, there were 168 gang members, comprising approximately 5% of the sample. This consideration is important because if it is the latter, then an emerging finding on the consequences of gang membership is that the effects of leaving a gang are not symmetrical to the effects of joining. In other words, if G.R.E.A.T. was more effective in encouraging disengaging from gangs than joining gangs, then it could have decreased the program effect size on criminal offending because active gang members are being compared with former gang members, many of whom have residual ties to the gang or are recovering from the cascading consequences of gang membership (Krohn et al., 2011; Melde and Esbensen, 2013; Pyrooz, Decker, and Webb, 2010).

The remainder of this essay considers a blend of the *wrong kind of people* and the *wrong kind of mechanisms* explanations that centers on treating gang membership as a dichotomy in research and practice. It may be the case that G.R.E.A.T. impacted gang membership "in name only" and only among a select group of youth who offered only limited criminal returns. That is, the program only influenced attitudes toward identifying as gang members but not the processes and behaviors surrounding gang membership, and the program effects were heterogeneous, only picking off "low-hanging fruit." It is held that in the context of program outputs in gang prevention and intervention programming, it is necessary to expand beyond dichotomies.

Beyond Dichotomies in Gang Prevention and Intervention

Gang membership is not a behavior, an act, or an event; rather, it is a status, a social category, or identity. For nearly 25 years, gang membership has been treated as a dichotomy. It is convenient, for research and practice, to lump people into "gang member" and "not a gang member" categories. For researchers, the one item does not crowd survey instruments. It avoids the definitional debate surrounding gangs by allowing the respondent to self-identify as a gang member. For practitioners, it is a neat classification scheme that avoids complications in defining and sanctioning those at the different social positions of influence in the gang (e.g., at the fringe or the core). Moreover, there is support for the reliability and validity of this classification scheme: Official and self-report data identify similar groups of gang and nongang youth (Curry, 2000) and self-identifying as a gang member is linked to a range of attitudes and behaviors theoretically associated with gang membership (Esbensen, Winfree, He, and Taylor, 2001; Thornberry et al., 2003). Indeed, the use of self-nomination has advanced the gang literature forward in important and significant ways over the last two decades and will (and should) continue to do so for years to come.

But simply because the measure "works" does not mean it is the best or only way to understand gang membership, especially in the context of gang prevention and intervention efforts. A complex reality surrounds the status and identity as a "gang member," one that is portrayed throughout the history of gang research (Klein, 1971; Miller, 2011; Papachristos, 2006; Thrasher, 1927; Yablonsky, 1962). Viewing gang membership as static

or homogenous ignores this reality because there is dynamic heterogeneity between and within individuals in their levels of immersion in and around gangs. Many have discussed this (e.g., Hughes, 2013; Klein, 1971; McGloin, 2007), but I will illustrate my main points with studies I have been involved in. What we (Pyrooz, Sweeten, and Piquero, 2013) have termed "gang embeddedness" treats individual immersion within gangs as a latent continuous construct based on five items: contact with the gang, position in the gang, importance of the gang, the balance of nongang to gang peers, and participation in gang assaults. Gang embeddedness is not restricted to individuals who self-identify as gang members. One can imagine how a future gang member ramps up his or her level of embeddedness in a gang prior to joining; likewise, one can imagine how a former gang member has residual or nonzero levels of gang embeddedness after leaving. Differences should exist in gang embeddedness between a child who grows up in a third-generation gang family in a gang-active neighborhood and a child who grows up in a middle-class family in a gang-free suburb. Whether someone is a future, an active, or a former gang member, these levels matter because individuals will be differentially susceptible to gang-related influences and group processes.

Gang prevention and intervention, as well as the evaluation of such programs, should begin not with dichotomies, but with the recognition that individuals are embedded within gangs differentially. A sole reliance or overreliance on the self-nomination of gang membership leaves much to be desired when evaluating the effectiveness of a program. For example, the results show that G.R.E.A.T. corresponded with more negative attitudes toward gangs. Could these effects also have shifted attitudes toward identifying as a gang member, but not actually influenced levels of gang embeddedness? Perhaps so. Garot (2010) likened gang identity to something that is performed among youth—"used" as opposed to being "affixed" to an individual. We know when someone is married or employed or a parent, official documents support these statuses, but the status of "gang member" can be hard to pin down sometimes. How else can we explain such high rates of intermittent gang careers? (e.g., Pyrooz, 2013b; Thornberry et al., 2003). If interpreted as moderate and understandable shifts in the continuous scale of embeddedness, rather than as more jarring active–inactive–active gang status changes, then these findings begin to make more sense. Moving beyond dichotomies of gang membership in prevention and intervention means three things, as follows.

First, complementing existing program outputs with additional indicators allows for multiple ways to demonstrate program success or understand program failure. Rather than solely reducing the aggregate self-reports of gang membership, this allows programs also to aim to reduce levels of gang embeddedness and its components (or other gang indicators). Instead of relying on one indicator for program success, there should be multiple ways to demonstrate that a program steers youth and young adults away from gangs and gang activity. For G.R.E.A.T., the statistically significant program effect for "attitudes about gangs" is consistent with this logic. Of course, using measures other than gang membership is a harder output to sell; it is more palatable to report a 2.6-percentage-point difference

between treatment and control groups or preventing 63 instances of gang membership than some vague 1 standard deviation reduction in a latent construct. But program managers and researchers can work around this in several ways, particularly by drawing attention to behavioral over attitudinal measures (e.g., decreasing time spent hanging out with the gang), which leads to the next point.

Second, moving beyond dichotomies recognizes that nothing is inherently criminal about gang membership. The social mechanisms linked to group processes, not simply gang status, are responsible for heightened levels of violence and criminal activity within gangs. As Sweeten, Pyrooz, and Piquero (2013) demonstrated, gang embeddedness matters as much for reducing levels of criminal activity as does de-identifying as a gang member. Based on their findings, they suggested an alternative strategy to promote desistance from crime, one that attempts to weaken levels of gang embeddedness as opposed to getting youth out of gangs. In other words, worry less about status and more about mechanisms. Programs that only prevent or stop individuals from identifying themselves as gang members but not stunt or decrease levels of gang embeddedness are unlikely to see reductions in criminal activity. Likewise, picking off those weakly embedded or unlikely to become deeply embedded may produce smaller criminal returns to the program. Both the *wrong kind of people* and the *wrong kind of mechanisms* might explain the null criminal offending effect for G.R.E.A.T., but it is impossible to tell without additional gang-related outputs.

Third, using additional measures such as embeddedness within gangs will aid in identifying subjects for appropriate programming. The concern that G.R.E.A.T. missed its target audience was raised originally by Klein and Maxson (2006: 101). For prevention efforts, this understanding can help the identification of youth for primary and secondary prevention programming, especially when supplemented with known risk factors for gang membership (Hennigan, Maxson, Sloane, Kolnick, and Vindel, 2013; Maxson, 2012). For intervention efforts, this understanding allows practitioners to steer participants toward the right programs. How programs approach an active gang member deeply embedded in his or her gang will differ from an active gang member with weak emotional and social ties. The risk heterogeneity models reported by Esbensen et al. (2013) follow this premise, but the question is whether the program works equally for prevention and intervention as well as across the spectrum of risks internal and external to gangs.

Concluding Remarks

At the Eurogang X workshop in Neustadt an der Weinstrasse, Germany, Terence P. Thornberry (2010) presented a blank slide listing the evidence-based gang prevention and intervention programs. The slide was blank because no programs met the criteria with rigorous evaluation and significant program effects. Esbensen et al.'s (2013) results move the criminological literature into a new phase of "something works" to prevent gang membership. Indeed, the rigorous randomized controlled trial design in combination with statistically and

substantively meaningful 39% and 24% 1-year and 4-year lower odds in gang membership linked to G.R.E.A.T. are unchartered territory for gang prevention.

No matter how much we would like to declare a prevention victory over gangs, it is important that we do not gloss over the details. By disaggregating the results by site, Esbensen et al. (2013) confront a much more complicated reality. The upside of these findings is that the 1-year program effects on gang membership were reproduced at three sites (Albuquerque, Philadelphia, and Portland), and the 4-year findings were reproduced at two sites. But the downsides are plenty, in particular that G.R.E.A.T. did not accomplish its primary goal to prevent gang membership in at least four sites. When combined with the inconvenient truth that lower odds of gang membership were unrelated to criminal offending, even in the disaggregated results, these findings should give pause to policy makers and administrators before aiming to roll out D.A.R.E.-like global expansion with aspirations to graduate millions of students annually. Low program costs just are not enough when program effects are only evident on attitudes and not behaviors.

What is needed is to get a better handle on the mechanisms underlying when G.R.E.A.T. works, and when it does not, and to balance those mechanisms with how the program accomplishes its primary goals of fostering positive police–youth relationships, preventing gang membership, and reducing criminal offending and violent behavior. This is particularly important for the null effects on criminal and violent offending and the uneven program effects across sites, especially as G.R.E.A.T.'s national and international footprint continues to grow (see great-online.org/) and the emerging positive evidence disseminates across the social scientific landscape. With its limited instructional time, low costs, and short- and long-term program effects, there is no doubt that G.R.E.A.T. is a program that "holds promise" (Esbensen et al., 2013)—the next step is ensuring that the program is keeping its promises.

References

Ayling, Julie. 2011. Gang change and evolutionary theory. *Crime, Law, and Social Change*, 56: 1–26.

Curry, G. David. 2000. Self-reported gang involvement and officially recorded delinquency. *Criminology*, 38: 1253–1274.

Curry, G. David, Scott H. Decker, and David C. Pyrooz. 2013. *Confronting Gangs: Crime and Community*, 3rd Edition. New York: Oxford University Press.

Decker, Scott H. 1996. Collective and normative features of gang violence. *Justice Quarterly*, 13: 243–264.

Densley, James A. 2012. *How Gangs Work: An Ethnography of Youth Violence*. London, U.K.: Palgrave.

Esbensen, Finn-Aage, D. Wayne Osgood, Dana Peterson, Terrance J. Taylor, and Dena C. Carson. 2013. Short- and long-term outcome results from a multisite evaluation of the G.R.E.A.T. program. *Criminology & Public Policy*, 12: 375–411.

Esbensen, Finn-Aage, D. Wayne Osgood, Terrance J. Taylor, and Dana Peterson. 2001. How great is G.R.E.A.T.? Results from a longitudinal quasi-experimental design. *Criminology & Public Policy*, 1: 87–118.

Esbensen, Finn-Aage, Dana Peterson, Terrance J. Taylor, and Adrienne Freng. 2010. *Youth Violence: Sex and Race Differences in Offending, Victimization, and Gang Membership*. Philadelphia, PA: Temple University Press.

Esbensen, Finn-Aage, Dana Peterson, Terrance J. Taylor, and D. Wayne Osgood. 2012. Results from a multi-site evaluation of the G.R.E.A.T. program. *Justice Quarterly*, 29: 125–151.

Esbensen, Finn-Aage, L. Thomas Winfree, Jr., Ni He, and Terrance J. Taylor. 2001. Youth gangs and definitional issues: When is a gang a gang, and why does it matter? *Crime & Delinquency*, 47: 105–130.

Garot, Robert. 2010. *Who You Claim: Performing Gang Identity in Schools*. New York: New York University Press.

Hennigan, Karen M., Cheryl L. Maxson, David C. Sloane, Kathy A. Kolnick, and Flor Vindel. 2013. Identifying high-risk youth for secondary gang prevention. *Journal of Crime and Justice*. E-pub ahead of print. DOI: 10.1080/0735648X.2013.831208.

Hughes, Lorine A. 2013. Group cohesiveness, gang member prestige, and delinquency and violence in Chicago, 1959–1962. *Criminology*. E-pub ahead of print. DOI: 10.1111/1745-9125.12020.

Katz, Jack and Curtis Jackson-Jacobs. 2004. The criminologists gang. In (Colin Sumner, ed.), *The Blackwell Companion to Criminology*. London, U.K.: Blackwell.

Klein, Malcolm W. 1971. *Street Gangs and Street Workers*. Englewood Cliffs, NJ: Prentice-Hall.

Klein, Malcolm W. and Cheryl L. Maxson. 2006. *Street Gang Patterns and Policies*. New York: Oxford University Press.

Krohn, Marvin D. and Terence P. Thornberry. 2008. Longitudinal perspectives on adolescent street gangs. In (Akiva M. Liberman, ed.), *The Long View of Crime: A Synthesis of Longitudinal Research*. Washington, DC: National Institute of Justice.

Krohn, Marvin D., Jeffrey T. Ward, Terence P. Thornberry, Alan J. Lizotte, and Rebekah Chu. 2011. The cascading effects of adolescent gang involvement across the life course. *Criminology*, 49: 991–1028.

Maxson, Cheryl L. 2012. Street gangs. In (James Q. Wilson and Joan Petersilia, eds.), *Crime and Public Policy*. New York: Oxford University Press.

McGloin, Jean M. 2007. The continued relevance of gang membership. *Criminology & Public Policy*, 6: 801–811.

Melde, Chris and Finn-Aage Esbensen. 2013. Gangs and violence: Disentangling the impact of gang membership on the level and nature of offending. *Journal of Quantitative Criminology*, 29: 143–166.

Miller, Walter B. 2011. *City Gangs*. Retrieved from gangresearch.asu.edu/walter_miller_library/walter-b.-miller-book/city-gangs-book.

Papachristos, Andrew V. 2006. Social network analysis and gang research: Theory and methods. In (James F. Short, Jr. and Lorine A. Hughes, eds.), *Studying Youth Gangs*. Lanham, MD: AltaMira.

Papachristos, Andrew V. 2009. Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology*, 115: 74–128.

Pyrooz, David C. 2013a. From colors and guns to caps and gowns? The effects of gang membership on educational attainment. *Journal of Research in Crime and Delinquency*. E-pub ahead of print. DOI: 10.1177/0022427813484316.

Pyrooz, David C. 2013b. "From your first cigarette to your last dyin' day": The patterning of gang membership in the life-course. *Journal of Quantitative Criminology*. E-pub ahead of print. DOI: 10.1007/s10940-013-9206-1.

Pyrooz, David C., Scott H. Decker, and Vincent J. Webb. 2010. The ties that bind: Desistance from gangs. *Crime & Delinquency*. E-pub ahead of print. DOI: 10.1177/0011128710372191.

Pyrooz, David C., Gary Sweeten, and Alex R. Piquero. 2013. Continuity and change in gang membership and gang embeddedness. *Journal of Research in Crime and Delinquency*, 50: 239–271.

Sweeten, Gary, David C. Pyrooz, and Alex R. Piquero. 2013. Disengaging from gangs and desisting from crime. *Justice Quarterly*, 30: 469–500.

Thornberry, Terence P. 2010. *A strategy for developing evidence-based gang intervention programs*. Paper presented at the 10th Eurogang Workshop, Neustadt an der Weinstrasse, Germany.

Thornberry, Terence P., Marvin D. Krohn, Alan J. Lizotte, Carolyn A. Smith, and Kimberly Tobin. 2003. *Gangs and Delinquency in Developmental Perspective*. Cambridge, U.K.: Cambridge University Press.

Thrasher, Frederic M. 1927. *The Gang: A Study of 1,313 Gangs in Chicago*. Chicago, IL: University of Chicago Press.

Wolfgang, Marvin E., Robert M. Figlio, and Thorsten Sellin. 1972. *Delinquency in a Birth Cohort*. Chicago, IL: University of Chicago Press.

Yablonsky, Lewis. 1962. *The Violent Gang*. New York: Macmillan.

David C. Pyrooz is an assistant professor in the Department of Criminal Justice and Criminology at Sam Houston State University. His main research interests are in the areas of gangs and deviant networks, violence, and developmental and life-course criminology. He is the recipient of a Graduate Research Fellowship from the National Institute of Justice and is the author of *Confronting Gangs: Crime and Community* (2013, with G. David Curry and Scott H. Decker).

VICTIM-CENTERED PROSECUTORIAL POLICIES

Empowering Women across Systems

The Impact of Intimate Partner Violence Intervention

Jill Theresa Messing
Arizona State University

This week, in Phoenix, where I live, Carol Sanders, her 14-year-old daughter Audra, and her brother-in-law Tom, were killed by Carol's estranged husband/Audra's father.[1] Carol had been to the police; she had both left her abusive husband and filed for divorce; she had told the authorities that her estranged husband had threatened her life, her daughter's life, and his own life; and she had obtained an order of protection. The morning of her death, Carol bravely faced her husband in court because he had contested the order of protection that she had received. Carol's murderer never surrendered any of the eight firearms he owned as he was instructed to do by the court; instead, he used one of them to murder his family and then to kill himself. Risk factors for homicide and homicide-suicide (Campbell et al., 2003; Koziol-McLain et al., 2006) are conspicuous throughout the news stories: gun ownership, threats with a weapon, threats to kill, threats directed at Audra, Carol's belief that he might kill her, suicide threats, recent separation, and controlling behavior are evident thus far (Woodfill and Madden, 2013). As one journalist aptly commented, "the best that we could do was to send Carol Sanders out to meet a bullet . . . armed with a piece of paper" (Roberts, 2013). This story, along with the more than four women killed by an intimate partner each day in the United States and the many more beaten and injured (Black et al., 2011; Catalano, Smith, Snyder, and Rand, 2009), raise questions about the interventions that we have in place to combat intimate partner violence and homicide.

Mary A. Finn (2013, this issue) responds to one of these questions in her article: What prosecution strategy (evidence-based prosecution or victim-centered prosecution) is more likely to reduce future abuse and violence, increase court empowerment, and enhance

Direct correspondence to Jill Theresa Messing, School of Social Work, Arizona State University, 411 N. Central Avenue, Suite 800, Phoenix, AZ 85004 (e-mail: jill.messing@asu.edu).

1. I am making a conscious decision to use the names of the victims and not the name of the perpetrator. Although it is important to be clear that the perpetrator is responsible for the violence he committed, I prefer only to memorialize the names of his victims.

DOI:10.1111/1745-9133.12062

survivor perceptions of safety? She finds that the victim-centered prosecution strategy that she examined was more likely to reduce both psychological abuse and physical violence in the 6 months after case disposition. Neither prosecution strategy had a significant effect on court empowerment or on survivor perceptions of safety. This study—and these findings—raise several questions about the prosecution of domestic violence crimes; many of these are addressed in the policy essays that follow the research article.

Buzawa and Buzawa (2013, this issue) pull from their and others' research to contextualize the move toward evidence-based prosecution and to raise important questions about survivor empowerment, the cost–benefit balance of evidence-based versus victim-centered prosecution, and deterrence theories. Peterson (2013, this issue) identifies findings (including his own) that conflict with those presented by Finn, and he questions the relationship between the naming of a particular prosecution strategy and actual prosecutorial practice. He stresses the importance of victim engagement that cuts across the various prosecutorial positions and policies. Similarly, Flannigan (2013, this issue) points out that there is not a fundamental conflict between evidence-based or pro-prosecution policies and survivor input into the criminal justice process. Although each author attends to empowerment slightly differently, this concept is woven throughout the research and discussion in such a way that it is apparent that providing survivors with a voice in the criminal justice process is both a great challenge and an important goal for policy makers and researchers.

The proportion of women seeking police assistance for domestic violence has increased to such an extent that, when intimate partner violence is identified as a crime, it is reported at rates similar to the reporting of other crimes (Felson, Messner, Hoskin, and Deane, 2002; Rennison and Welchans, 2000). Across published research, women use the criminal justice system as an intervention for intimate partner violence more than they use domestic violence services, including domestic violence shelters (see Messing et al., in press). Although most women report that social services are helpful or make the situation better (Goodkind, Sullivan, and Bybee, 2004; Goodman, Dutton, Weinfurt, and Cook, 2003) and shelter services were shown to be most effective in reducing severe and moderate reassault in one prospective study (Campbell, O'Sullivan, Roehl, and Webster, 2005), research on the criminal justice system tends to be more mixed. This is the case in terms of both survivor satisfaction (Goodkind et al., 2004; Goodman et al., 2003) and criminal justice outcomes (see the literature review in Finn, 2013; see also Campbell et al., 2003; Maxwell, Garner, and Fagan, 2001).

Although the criminal justice system has become the primary intervention for domestic violence crimes, it is not equipped to manage the relational and ongoing nature of intimate partner violence (Messing, 2011). By design, the criminal justice system focuses on a single criminal incident. Yet intimate partner violence is the systematic and ongoing use of power and control in an intimate relationship that is only sometimes reinforced through the use of physical violence (Stark, 2007). The criminal justice system emphasis on offender accountability is important and necessary; yet the system also cannot attend to the many

and diverse needs of survivors. Some interventions (e.g., coordinated community response, high-risk teams, and the Lethality Assessment Program) are working to engage actors across criminal justice and social service systems. However, these interventions are not broad enough in scope and are implemented in limited jurisdictions. The research conducted by Mary A. Finn (2013)—to examine the differential outcomes of various prosecution strategies with a focus on survivor empowerment—is a necessary step in understanding the impact of the criminal justice response on survivors' lives. It is particularly noteworthy that Finn examines reports of repeat abuse and violence *as reported by survivors*. Using survivor reports as opposed to reports of reoffense by the criminal justice system marks an important shift from institutional concerns ("Will the offender reenter the criminal justice system?") to women's lived experience ("Will violence and abuse continue?"). Yet, an emphasis on prosecution remains a small focus within a single system. Public policy surrounding intimate partner violence includes the criminal justice and social service systems, as well as health care, education, financial opportunities, and a multitude of other social systems that trap women and children in violent environments. Resources that support the empowerment of women within and outside of the criminal justice system are necessary to combat intimate partner violence. Until this is a priority, families will continue to live in violence and fear.

References

Black, Michele C., Kathleen C. Basile, Matthew J. Breiding, Sharon G. Smith, Mikel L. Walters, and Melissa T. Merrick, 2011. *National Intimate Partner and Sexual Violence Survey*. Atlanta, GA: Centers for Disease Control and Prevention.

Buzawa, Eve S. and Aaron D. Buzawa. 2013. Evidence-based prosecution: Is it worth the cost? *Criminology & Public Policy*, 12: 491–505.

Campbell, Jacquelyn C., Chris S. O'Sullivan, C. Janice Roehl, and Daniel W. Webster. 2005. *Intimate Partner Violence Risk Assessment Validation Study: The RAVE Study*. Final Report to the National Institute of Justice. NCJ 209731-209732. Retrieved November 17, 2013 from ncjrs.org/pdffiles1/nij/grants/209731.pdf.

Campbell, Jacquelyn C., Daniel W. Webster, Jane Koziol-McLain, Carolyn R. Block, Doris Campbell, Mary Ann Curry, et al. 2003. Risk factors for femicide in abusive relationships: Results from a multi-site case control study. *American Journal of Public Health*, 9: 1089–1097.

Catalano, Shannan, Erica Smith, Howard Snyder, and Michael Rand. 2009. *Female Victims of Violence*. NCJ 228356. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Felson, R. B., S. F. Messner, A. W. Hoskin, and G. Deane. 2002. Reasons for reporting and not reporting domestic violence to the police. *Criminology*, 40: 617–648.

Finn, Mary A. 2013. Evidence-based and victim-centered prosecutorial policies: Examination of deterrent and therapeutic jurisprudence effects on domestic violence. *Criminology & Public Policy*, 12: 443–472.

Flannigan, Krista R. 2013. The importance of prosecution policies in domestic violence cases. *Criminology & Public Policy*, 12: 481–490.

Goodkind, Jessica R., Cris M. Sullivan, and Deborah I. Bybee. 2004. A contextual analysis of battered women's safety planning. *Violence Against Women*, 10: 514–533.

Goodman, Lisa, Mary Ann Dutton, Kevin Weinfurt, and Sarah Cook. 2003. The intimate partner violence strategies index development and application. *Violence Against Women*, 9: 163–186.

Koziol-McLain, Jane, Daniel W. Webster, Judith McFarlane, Carolyn Rebecca Block, Yvonne Ulrich, Nancy Glass, et al. 2006. Risk factors for femicide-suicide in abusive relationships: Results from a multisite case control study. *Violence & Victims*, 21: 3–21.

Maxwell, Christopher D., Joel H. Garner, and Jeffrey A. Fagan. 2001. *The Effects of Arrest on Intimate Partner Violence: New Evidence from the Spouse Assault Replication Program*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Messing, Jill Theresa. 2011. The social control of family violence. *Affilia: The Journal of Women and Social Work*, 26: 154–168.

Messing, Jill Theresa, J. C. Campbell, S. Brown, B. Patchell, D. Androff, and J. Wilson. In press. The association between protective actions and homicide risk: Findings from the Oklahoma Lethality Assessment Study. *Violence & Victims*.

Peterson, Richard R. 2013. Victim engagement in the prosecution of domestic violence cases. *Criminology & Public Policy*, 12: 473–480.

Rennison, Callie Marie and Sarah Welchans. 2000. *Intimate Partner Violence*. Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Roberts, Laurie. 2013. Hours before her death, Carol Sanders tells judge, "He was going to kill me." *The Arizona Republic*. November 16. Retrieved November 17, 2013 from azcentral.com/insiders/laurieroberts/2013/11/15/hours-before-her-death-carol-sanders-tells-judge-he-was-going-to-kill-me/.

Stark, Evan. 2007. *Coercive Control: How Men Entrap Women in Personal Life*. New York: Oxford University Press.

Woodfill, D. S. and Haley Madden. 2013. Records: Phoenix man who killed family had terrorized them. *The Arizona Republic*. November 13. Retrieved November 17, 2013 from azcentral.com/community/nephoenix/articles/20131113phoenix-police-release-ids-triple-homicide-abrk.html.

Jill Theresa Messing, MSW, PhD, is an assistant professor in the School of Social Work at Arizona State University. Her interest areas are intimate partner violence, domestic homicide/femicide, risk assessment, criminal justice-social service collaborations, and evidence-based practice. Dr. Messing specializes in intervention research, and is the Principal Investigator on the National Institute of Justice funded *Oklahoma Lethality Assessment Study* (#2008-WG-BX-0002) and a co-Investigator on the National Institute of Mental Health funded study *The Use of Computerized Safety Decision Aids with Victims of Intimate Partner Violence* (#R01 MH085641).

EXECUTIVE SUMMARY

VICTIM-CENTERED PROSECUTORIAL POLICIES

Overview of: "Evidence-Based and Victim-Centered Prosecutorial Policies: Examination of Deterrent and Therapeutic Jurisprudence Effects on Domestic Violence"

Mary A. Finn
Georgia State University

Research Summary

Differences in outcomes for domestic violence cases were compared across two court jurisdictions, one that employed victim-centered prosecutorial policies and one that employed evidence-based prosecutorial policies. Evidence-based prosecutorial policies argue that the reoccurrence of violence is deterred through the certain, swift, and severe punishment of offenders, whereas victim-centered prosecutorial policies claim that the reoccurrence of violence declines when victims interact with court officials who provide them with the opportunity to participate actively and provide input into the court's actions. Overall, 170 victims were interviewed at three time points (intake, disposition, and 6 months after disposition) to assess levels of court empowerment, reoccurrence of physical violence and psychological aggression, and perception of safety reported by victims. The results indicate that cases in the evidence-based policy jurisdiction, compared with the victim-centered policy jurisdiction, were significantly more likely to report reoccurrence of physical violence and psychological aggression. Victims who experienced physical violence during the 6 months after case disposition perceived themselves as less safe (i.e., they reported that physical violence was more likely to occur in the future).

Policy Implications

Interest in the positive and negative effects of prosecutorial policies on the lives of domestic violence victims involved in the justice process has been growing. Currently, the dual aims of the justice process are to assure offender accountability and to enhance victim safety, and two distinct policy approaches have emerged (mandatory prosecution and

DOI:10.1111/1745-9133.12046

victim-centered prosecution) to accomplish these aims. The current study examines the influence of each policy on revictimization and perceptions of safety of domestic violence victims rather than official measures of offender recidivism, thus informing policy makers of the broader impact of prosecutorial policies on the lives of victims. The results suggest that victim-centered polices yield better outcomes for domestic violence victims than evidence-based policies. This finding has implications for jurisdictions considering whether to adopt evidence-based policies, and it suggests that careful consideration be given to their implementation if their effect is to regard victims primarily as witnesses to a crime and they do not make efforts to encourage, educate, and support victims throughout the court process. As victim-centered prosecutorial policies are rooted in the theory of therapeutic jurisprudence, our findings suggest that justice professionals be encouraged to think more broadly about how involvement with the justice process can foster the improved well-being of victims. Although the current study was conducted in traditional courts, the number of specialty courts that addresses domestic violence is growing nationally, and the findings suggest this is a positive development.

Keywords

prosecution, *domestic violence*, *intimate partner violence*, *therapeutic jurisprudence*, *and prosecutorial policy*

Evidence-Based and Victim-Centered Prosecutorial Policies

Examination of Deterrent and Therapeutic Jurisprudence Effects on Domestic Violence

Mary A. Finn
Georgia State University

Many innovations designed to enhance victim safety and ensure offender accountability in domestic violence cases have occurred over the past three decades. In addition to proarrest policies enacted by law enforcement, prosecutors have played central roles in the adoption of mandatory prosecution, specialized domestic violence courts, and community coordinated responses (Worrall, 2008: 235). These innovations were undertaken to assure better justice outcomes for victims and communities by preventing prosecutors from routinely screening out domestic violence cases because of victim reluctance to proceed (Goodman, Bennett, and Dutton, 1999) or perceptions of such offenses as trivial (Hart, 1993). Nearly three quarters of prosecutors surveyed in more than 200 domestic violence courts reported that they often or always proceeded with a case regardless of the victim's willingness to support prosecution (Labriola, Bradley, O'Sullivan, Rempel, and Moore, 2009: 42). Thus, regardless of whether domestic violence cases were processed through traditional or specialized courts, prosecutors remained powerful actors in the justice process (Krug, 2002), serving as the key decision makers in selection and screening of cases for adjudication (Jacoby, 1980: 107). Despite their important role, prosecutors have not garnered the attention of scholars (Dempsey, 2009; Worrall, Ross, and McCord, 2006).

Advanced in the wake of research indicating that arrests deterred future abuse (Maxwell, Garner, and Fagan, 2001; Sherman and Berk, 1984), mandatory prosecution policies

Data for this research were obtained from a project funded under Grant 1999-WT-VX-0008 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Direct all correspondence to Mary A. Finn, Department of Criminal Justice & Criminology, Andrew Young School of Policy Studies, Georgia State University, 140 Decatur Street, Atlanta, GA 30302-4018 (e-mail: mfinn@gsu.edu).

DOI:10.1111/1745-9133.12049

(often also referred to as no-drop, evidence-based, or victim-less policies) were reportedly in widespread use in prosecutors' offices in large urban counties by the mid-1990s (Corsilles, 1994; Rebovich, 1996). Indeed, several scholars linked the adoption of mandatory prosecution policies specifically to higher prosecution and conviction rates and lower diversion and deferred adjudication rates for domestic violence cases across the United States (Klein, 2009; Smith, Durose, and Langan, 2008). Mandatory prosecution policies convey an institutional commitment to treat domestic violence as a serious crime and, thus, may serve as a powerful specific deterrent to batterers (Ferraro and Pope, 1993) and a general deterrent to potential domestic violence perpetrators (Goolkasian, 1986; Hanna, 1996; Lerman, 1981). Such policies were applauded because they clarified that domestic violence is a crime against the society and not just against the battered victim (McCord, 1992; Wills, 1997). As such, the state, not the victim, was the aggrieved party and the prosecutor controlled the direction of the prosecution (Corsilles, 1994; Wills, 1997). Some prosecutors and advocates reported that adoption of mandatory prosecution reduced batterer intimidation of victims (Wills, 1997) and increased the batterer's guilty pleas (Goolkasian, 1986). Given that such policies prevented prosecutors from dropping domestic violence cases as long as sufficient evidence existed to prove a crime occurred (Epstein, 1999), case attrition rates were lower (Corsilles, 1994).

However, critics of mandatory prosecution charged that resources were wasted by requiring that the prosecution proceed with cases that, without victim cooperation, had little chance of resulting in convictions (Corsilles, 1994), thus contributing to overcrowded court dockets (Davis, Smith, and Nickles, 1998). In addition, critics believed that such policies undercut efforts at empowering victims of abuse, further eroding their self-esteem and sense of control (Epstein, 1999; Ford, 1991; Goodman and Epstein, 2007; Goolkasian, 1986; Han, 2003; Zorza, 2010). Moreover, it was suggested that such policies further victimized women if enforcement of mandatory prosecution led to punitive actions against them (see Dempsey, 2009; Esptein, 1999; Waites, 1985), increased the risk of batterer retaliation, and perhaps most importantly, discouraged victims from calling the police if violence reoccurred (Bell, Goodman, and Dutton, 2007; Mills, 1998). Indeed, it was partly because of concerns about both the waste of resources and the potential harmful effects of coercing victims to move forward in cases that some jurisdictions adopted victim-centered policies (Davis, O'Sullivan, Farole, and Rempel, 2008).

On three separate occasions over the previous 8 years, *Criminology & Public Policy* has published articles that examined the influence of adoption of mandatory prosecutorial policies in cases of domestic violence on several important outcomes, including extent and length of court oversight and convictions (Peterson and Dixon, 2005); pleas, sentences, and time to disposition (Davis, Smith, and Taylor, 2003); and recidivism, victim safety, victim satisfaction, and empowerment (Davis et al., 2008) in criminal courts in New York City and in Milwaukee, WI (Davis et al., 2003). Policy essays accompanying the articles consistently noted the importance of understanding how prosecutorial decisions that do not

consider victim preferences affect victim empowerment and, ultimately, victim safety. In their essay, Belknap and Potter (2005) acknowledged the importance of understanding how policy affects traditional outcomes (convictions and recidivism), but then they called for an expansion of the outcomes examined by researchers to include whether victims regarded such policies as reflective of their wishes. They noted, "Victim empowerment is probably the most important focal point of implementation or continuation of [domestic violence] DV policies" (Belknap and Potter, 2005: 561). Buzawa and Buzawa's (2008) policy essay in response to Davis et al.'s (2008) work reiterated the importance of examining how prosecutorial policies may empower victims.

The current study advances understanding of prosecutorial policy on victims by examining court empowerment and the reoccurrence of violence over the course of domestic violence case processing (at court intake, at disposition, and at 6 months after disposition) across two metro-Atlanta misdemeanor court jurisdictions, one that employed an evidence-based prosecutorial policy and another that employed a victim-centered prosecutorial policy. This study advances prior research by measuring court empowerment and reoccurrence of violence reported from victims directly rather than relying on official reports of offender recidivism. It is the first study to compare the effects of evidence-based prosecutorial policy (a specific subtype of mandatory prosecutorial policies) and victim-centered prosecutorial policies in two jurisdictions in the same urban area within a single state. Both policies seek to hold the offender accountable and to enhance victim safety, albeit through different causal mechanisms: mandatory prosecution advancing deterrence and victim-centered policies advancing therapeutic jurisprudence.

Literature Review

Foundations of Prosecutorial Policies: Deterrence versus Therapeutic Jurisprudence

Dempsey (2009: 4) in *Prosecuting Domestic Violence* described the "radical transformation" of prosecutorial policies to address domestic violence over the past 30 years as they moved from a traditional approach, wherein cases were typically dismissed unless a victim insisted on prosecution, to a pro-prosecution approach, wherein a victim's request for dismissal was disregarded by prosecutors. She distinguished four types of prosecutorial policies: *victim-led* prosecution, wherein cases were always discontinued pursuant to a victim's request to dismiss; *victim-oriented* prosecution, wherein cases were generally discontinued pursuant to a victim's request to dismiss unless strong reasons weighed in favor of continuing prosecution; *pro-prosecution*, wherein cases were generally not discontinued because of a victim's request to dismiss; and *mandatory* prosecution, wherein a victim's request to dismiss was wholly irrelevant to prosecutorial decisions. These policies differed fundamentally on the degree to which a victim's desire to dismiss the case or to support actively or participate in the prosecutorial process influenced prosecutorial actions. If conceived of as a continuum, then victim influence is strongest under *victim-led* prosecution policies and weakest under *mandatory* prosecution policies.

"Mandatory prosecution" was the term originally used to describe policies/actions whereby prosecutors charged defendants when credible evidence existed, without consideration given to the degree of cooperation or support of the victim (Corsilles, 1994; Davis et al., 2008; Epstein, 1999; Peterson, 2002, 2003; Peterson and Dixon, 2005). As more studies on prosecutors were conducted, additional terms were added to the lexicon, such as "no-drop," "evidence-based," or "victim-less," and the terms have subsequently been used interchangeably, masking important distinctions that may exist. Whereas all of these terms described a trend in policy development that consistently shifted responsibility for the charging decision squarely onto the prosecutor and away from the victim, they seemed to vary in the degree to which the *strength of evidence* drove this charging decision. In several jurisdictions, consideration of the evidence was at the forefront of prosecutorial decisions, where strengthening the amount and quality of evidence collected in domestic violence cases to assure successful prosecution was central. Referred to as evidence-based policies, this unique subtype of mandatory prosecution policies described partnerships between prosecution and law enforcement that strengthened and expanded evidence gathered to support conviction, such as medical reports, photographs, witness statements, and 911 transcripts (Davis et al., 2003; Gewirtz, Weidner, Miller, and Zehm, 2006; Han, 2003; Messing, 2010). Evidence-based prosecutorial policies have become even more relevant in light of recent U.S. Supreme Court decisions (e.g., *Crawford v. Washington,* 2004; *Davis v. Washington,* 2006; *Michigan v. Bryant*, 2011) that placed restrictions on the admission at trial of testimonial statements obtained by police at crime scenes.[1]

Not all jurisdictions adopted mandatory prosecutorial polices; instead, some opted to permit, and even encouraged, victims' input into the prosecutorial decision of whether to charge the accused. Referred to variously as victim-centered, victim-led, victim-oriented, victim-informed, and victim-empowered policies (Cattaneo and Goodman, 2010; Davis et al., 2008; Han, 2003), the prosecution was unlikely to proceed unless the victim clearly supported such action. Theoretically, such policies are best understood from the conceptual framework of therapeutic jurisprudence, which studies the law's impact on the physical and psychological well-being of individuals who come into contact with it (see Wexler, 1990; Winick, 1996, 1997; Winick and Wexler, 2003). Giving the victim the opportunity to participate actively in the prosecutorial process enhances his or her well-being (Erez and Hartley, 2003; Gover, Brank, and McDonald, 2007; Wemmers, 2008; Wemmers and Cyr, 2005) and reinforces that he or she is competent and can exercise autonomy. Furthermore, the absence of an active role and real influence in criminal proceedings may be harmful to victims' well-being (Diesen, 2011). Empowerment of the victim is the first

1. Although the *Crawford* decision was lauded as ringing the death knell of witness-less prosecution in domestic violence cases (Fine, 2006), recent reviews suggest that many trial courts continued to allow the admission of prior accusations at trial in place of live testimony (Mosteller, 2005; Ross, 2007), and as evidenced in both the *Davis* and *Bryant* decisions, hearsay exceptions to the Confrontation Clause are likely to expand (Badawy, 2011).

step to recovery from trauma (Herman, 1997). In contrast to deterrence-based policies that focus on the importance of offender accountability to prevent future violence, decisions within the therapeutic jurisprudence framework are made with consideration of future ramifications for individuals, relationships, and society long after a person's contact with the criminal justice system has ceased (Slate, 2003: 15). Thus, the two theoretical frameworks that underlie prosecutorial policies, deterrence and therapeutic jurisprudence, both seek a reduction in the reoccurrence of violence. However, they differ in that deterrence focuses exclusively on how strengthening legal sanctions can reduce offenders' recidivism, whereas therapeutic jurisprudence focuses on how the legal system can enhance victims' well-being by considering and reflecting on the wishes of the victim in its decisions and responses to violence.

Effects of Prosecutorial Policies

Several works examined the deterrent effects of judicial outcomes on the reoccurrence of violence in domestic violence cases, which are typically measured as revictimization or defendant's rearrest (Davis et al., 1998; Ford and Regoli, 1993a, 1993b; McFarlane, Willson, Lemmey, and Malecha, 2000; Mears, Carlson, Holden, and Harris, 2001; Thistlethwaite, Wooldredge, and Gibbs, 1998; Tolman and Weisz, 1995). With the exception of Thistlethwaite et al. (1998), who reported that type of sanction imposed was negatively related to likelihood of rearrest 1 year after case closure, all others reported that neither the case disposition nor the type of sanction influenced the reoccurrence of violence (Davis et al., 1998; Ford and Regoli, 1993a, 1993b; McFarlane et al., 2000; Mears et al., 2001; Tolman and Weisz, 1995). None of these works controlled for the effects of prosecutorial policies on court outcomes or on the reoccurrence of violence.

Largely concentrated in a few large, urban judicial districts (e.g., Boston, MA; Indianapolis, IN; Milwaukee, WI; New York, NY; and St. Paul, MN), more recent research examined the influence of mandatory and victim-centered prosecutorial policies on important outcomes, including dispositions (Peterson and Dixon, 2005), sanctions (Davis et al., 2003), offender recidivism, victim safety, victim satisfaction, and victim empowerment (Buzawa, Hotaling, Klein, and Byrne, 1999; Davis et al., 2008; Ford and Regoli, 1993a, 1993b). One justification for the use of mandatory prosecutorial polices was that even though a greater percentage of cases ultimately may be dismissed prior to trial, charging all cases referred by police permitted courts greater oversight over more cases and, thus, arguably greater protection of victims from subsequent intimidation or threat (Belknap et al., 2000). Peterson and Dixon (2005) tested this assumption relying on misdemeanor domestic violence cases drawn from the first quarter of 2001 in the Bronx and Brooklyn. Their findings indicated that filing rates were higher and that the length of court oversight from filing to disposition was longer when universal filing policies operated. This study did not directly examine how policy type affected victim safety or the reoccurrence of violence as perpetrated by the defendant or as experienced by the victim.

Taking advantage of a natural experimental design, Davis et al. (2003) examined the differences in pleas, sentences, and time to disposition when the Milwaukee prosecutors' office shifted in 1995 from a policy requiring victims to attend a charging conference for prosecutorial charges to be filed to a policy requiring that the decision to charge be based on the seriousness of the incident, legal merit, and history (e.g., evidence-based policy). They found that despite the change in policy, victim attendance at the charging conference remained a significant factor in the decision to prosecute, suggesting that despite a commitment to move forward based exclusively on evidence, prosecutors continued to defer to victims. Evidence-based charging did not produce significantly better justice outcomes and ultimately contributed to lower levels of victim satisfaction and victim safety. They cautioned that "to ignore victims' wishes as an important piece of data in deciding whether to prosecute invites a caseload of unwinnable cases, disgruntled victims, and (potentially) prosecution of innocent defendants" (Davis et al., 2003: 279–280). However, without a control group, it is difficult to determine whether empirically these cautions are merited.

Only one published study to date examined how prosecutorial policies influenced victim safety, victim satisfaction, and victim access to resources.[2] Returning to the Bronx and Brooklyn court settings, Davis et al. (2008) examined whether it was better, in terms of victim safety, victim satisfaction, and victim access to resources, to prosecute all arrests made by police (universal filing policy) or to respect the wishes of the victim and prosecute only those in which victims supported prosecution and were willing to participate (selective filing policy). Their findings indicated that regardless of the charging policy operating, prosecution was more likely if the victim signed the Domestic Incident Report filed by police and if the offender had a history of domestic violence. Recidivism, operationalized as the rearrest for assault, menacing, or harassment at 6 months, did not differ across the two sites. In addition, the universal filing policy did not delay the onset of recidivism. Data collected from individual interviews with 23 victims and a group interview with 12 victims indicated all victims opposed prosecution at the time of arrest, but they did not state a clear preference for the selective filing policy that would allow their cases to be dropped. Importantly, prosecution policy (universal vs. selective) had no effect on the intent of victims to report future crime or to support prosecution of reported crimes, but the small number of victims interviewed suggests that such conclusions may be premature without additional confirmation.

2. In *Evaluation of Efforts to Implement No-Drop Policies*, Smith, Davies, Nickles, and Davies (2001) presented results from interviews with victims from four jurisdictions that employed no-drop policies in efforts to learn more about their experiences with prosecutors. Sixty-four percent indicated that they were satisfied with the prosecution's handling of their cases. Most reported that since the case was disposed, the defendant had not threatened to harm the victim (86%), had not damaged the victim's property (92%), had not been physically violent toward the victim (91%), or had not been verbally abusive toward the victim (63%). Overall, 85% of victims indicated that they thought it was a good idea that the case was prosecuted.

Empowerment and Court Outcomes

At a fundamental level, empowerment is about gaining power or influence in the context of social interactions (Catteneo and Chapman, 2010). It is a term used within many fields, including community psychology, public health, nursing, organizational management, social work, industrial psychology, education, and criminal justice (Chronister and McWhirter, 2003; Mills, 1996; Wood and Middleman, 1992). It has been regarded simultaneously as a process and as an outcome (Andrews, Guadalupe, and Bolden, 2003; van Uden-Kraan et al., 2008). According to Zimmerman (1995), a leading empowerment scholar, empowerment takes different forms in different people (age, race, and gender), in different contexts (work, family, recreation, and school), and at different times (persons experience both empowering and disempowering processes over their life course) (see also Akey, Marquis, and Ross, 2000). Given its characteristics, Zimmerman (1995) has cautioned against the development of a universal or global measure of empowerment and has argued that it is important in conceptualization and measurement of empowerment to distinguish between empowering *processes* and empowered *outcomes* (see Zimmerman [1995, 2000] for a more in-depth discussion).

The concern that mandatory prosecution policies had a negative influence on victims' empowerment and the reoccurrence of violence was first identified as an important factor by Ford and Regoli (1993a, 1993b). They found that, among cases entering the court via victim-initiated complaints, women who were permitted to drop charges but chose to pursue them reported lower rates of violence in the 6 months after case settlement. In contrast, victims who were permitted to drop the charges, and did, reported the highest rates of violence during the 6 months after case settlement. However, why choice led some women to escape revictimization, and some women not, remains unanswered. Ford and Regoli (1993a: 74) noted that prosecutors played an important role by "securing arrangements to minimize the chance of violence, by affirming the legitimacy of her criminal complaints and by respecting her decisions on what is best for her unique circumstances, even if contrary to the prosecutor's administrative wishes."

Subsequently, scholars attempted to understand more clearly the role of empowerment in reducing violent revictimization. Unfortunately, much of the research on the empowerment of victims has lacked clear definition and measurement (see McDermott and Garafalo [2003] for more in-depth discussion). Building on the foundation of Ford and Regoli (1993a, 1993b), Hartley (2003) and Han (2003) argued that for prosecutors to empower victims, they should allow victims shared power in decisions and a voice in crafting solutions. More recently, the work of Cattaneo and colleagues advanced and refined the understanding of empowerment in the court context (Cattaneo and Chapman, 2010; Cattaneo and Goodman, 2010). Cattaneo and Goodman (2010) found that empowering experiences with the court were not related to victims' level of fear but did lead to greater improvements in victims' well-being (depression and quality of life), above and beyond

their experiences with repeated abuse, expectations about the court, and the case. They recommended that future research employ more refined measures of empowerment and examine the role that prosecutorial policy and quality of interactions with prosecutors and victim-witness advocates might have on victims' well-being.

Current Study

This study examines the influence of prosecutorial policies on court empowerment, physical violence, psychological aggression, and perceptions of safety reported by victims over the course of domestic violence case processing in two jurisdictions: one that employs an evidence-based prosecutorial policy and one that employs a victim-centered prosecutorial policy. Evidence-based prosecutorial policies argue that reoccurrence of violence is deterred through the certain, swift, and severe punishment of offenders, and as a result, victims will report feeling safer. Specifically, in the evidence-based jurisdiction, compared with the victim-centered jurisdiction, a significantly greater number of cases will be formally disposed, the length of time from arrest to disposition of cases will be significantly shorter, and more severe dispositions and more severe punishments will be imposed. Furthermore, the reoccurrence of intimate partner violence during court processing will be significantly lower. Ultimately, more swift, certain, and severe punishment should result in decreases in the occurrence of psychological aggression and physical violence after disposition, and victims' perceptions of safety should be enhanced.

In contrast, victim-centered policies claim to advance victims' safety and reduce the reoccurrence of violence, largely through providing victims empowering experiences with the court, which is accomplished when prosecutors act to seek victim input and to reflect victim input into the handling of the criminal case. Based on therapeutic jurisprudence, it is expected that court empowerment will be greater in the victim-centered jurisdiction than in the evidence-based jurisdiction. Based on the findings of Ford and Regoli (1993a), it is predicted that the therapeutic effects of court interaction will decrease both the likelihood that intimate partner violence will reoccur during court processing and the likelihood that psychological aggression or physical violence will reoccur after disposition and will increase victims' perceptions of safety.

The current study builds on prior research in several ways. First, it employs a measure of court empowerment validated in prior research (Cattaneo and Goodman, 2010) but not yet examined in jurisdictions employing different prosecutorial policies. Second, by including two jurisdictions that employ distinct prosecutorial policies, the effect of such policies on the elements of deterrence (e.g., certainty, swiftness, and severity) and on the elements of therapeutic jurisprudence (e.g., court empowerment) will be examined. Third, in addition to their differential influences on court empowerment and deterrence measures, evidence-based and victim-centered policies both claim to improve victim safety and reduce the reoccurrence of violence. Prior research has relied almost exclusively on official reports of

recidivism to measure the reoccurrence of violence, which likely underestimates the extent of violent or abusive behavior engaged in by the offender. Rather than rely on official reports of defendants' rearrest, victims' reports of violent and aggressive behavior during the 6 months after disposition and victims' perception of safety in the future are assessed. Finally, it measures the occurrence of violence as reported by victims at three distinct time points: prior to court intake, during the pretrial stage (between court intake and court disposition), and 6 months after court disposition.

Method

Research Sites

Census profiles. Two prosecutors' offices located in metro-Atlanta, Georgia counties (herein referred to as metro county A [evidence-based jurisdiction] and metro county B [victim-centered jurisdiction]) were selected as the research sites based on the type of prosecutorial policy in operation. A review of the U.S. Census Bureau (2000) data for the two counties indicated that the two counties were similar in many respects. Metro county B had a slightly larger population ($N = 665,865$ residents) than metro county A ($N = 588,448$ residents). With regard to the sex and age composition of residents, approximately half of the residents in both counties were female, and the median age for metro county B was 32.3 years and the median age for metro county A was 32.5 years. The educational attainment of residents was similar across the two counties; 87.3% of residents in metro county A had achieved a high-school degree or higher and 34.1% reported attainment of a bachelor degree or higher, whereas in metro county B, 85.1% reported being a high-school graduate or higher and 36.3% reported earning a bachelor degree or higher. The employment status of residents in the civilian labor force was similar across the two sites; nearly 72% in metro county A and 67% of residents in metro county B reported being employed.

However, two key differences in the counties were evident. Metro county A was more affluent and less racially diverse than metro county B. Residents in metro county A reported a higher per-capita income ($25,006) and higher median household income ($66,693) compared with residents of metro county B ($23,968 and $54,018, respectively). Furthermore, a smaller proportion of metro county A residents were living in poverty than in metro county B (3.8% compared with 7.8%), and the proportion of female-headed households was slightly smaller in metro county A versus metro county B (17.6% and 10.0%, respectively). The racial composition of the residents also differed significantly, with Blacks/African Americans comprising at least half of the population of metro county B compared with 14% of the population in metro county A. The ethnic composition of the two counties was similar; residents of Hispanic or Latino origin totaled 7.9% in metro county B and 10.9% in metro county A.

Prosecutorial Policies

Evidence-based policy. Metro county A employed an evidence-based prosecutorial policy in which all cases were referred to a specialized unit that consisted of three prosecutors (one serving as supervisor), two victim/witness advocates, and a domestic violence investigator. Each prosecutor reviewed a case independently and sent it forward for accusation to state court if it reached legal sufficiency (e.g., probable cause). No separate court calendar or specialized court for domestic violence cases operated. Diversion, abeyance, or mediation was specifically prohibited. Victims were provided referrals to community resources, but local service providers were not physically present in the court. Victims were primarily regarded as witnesses to the crime, and the ability of the victim to control any decisions regarding how the case was handled was minimized. Victims were informed from the outset that if a prima facie case could be made against the defendant without victim testimony, then the case would go forward regardless of their wishes. The decision whether to proceed with prosecution was made solely by the prosecutor and not by the victim. The charges against the accused were to be prosecuted without victim cooperation. A victim who failed to appear in court after being lawfully served with a subpoena would be treated like any other witness who failed to appear to an authorized and legally served subpoena. The prosecutor would not offer or agree to reduce a domestic violence charge to a lesser offense unless, after investigation and discussion with the victim, doing such served the ends of justice. Once an accusation was filed, the prosecutor would not move to dismiss or nolle presequi the charges simply because the victim was reluctant to cooperate.

In metro county A, domestic violence cases were classified as high priority if the victim received injuries that resulted in serious bleeding or serious bruising, or required medical treatment, or if the accused had a history of violence against anyone. Victim services would attempt to contact the victim and schedule an interview within 24 hours of receipt of the police incident report. The purpose of the victim services unit was to inform victims of their rights, to secure victim cooperation, to gather evidence necessary or helpful for criminal prosecution, and to be an ongoing contact for the victim until final disposition of the case. In deciding whether to accuse or dismiss a case, prosecutors were to consider the extent or seriousness of injuries, the use of weapon, defendant's criminal history, history of violence, existence of relevant court orders, status of defendant's arrest, victim cooperation, and presence of independent evidence. Independent evidence included the following: (a) witnesses of injuries or crime commission; (b) medical reports of injuries; (c) 911 tape with statements of victim, witnesses, defendant, or all of the above; (d) presence of physical evidence; (e) admissions by defendant; and (f) photographic evidence. Prosecution of the case can proceed without independent evidence if a victim indicates that he or she will cooperate with the prosecution.

Victim-centered policy. Metro county B employed a victim-centered prosecutorial policy in which all cases were referred to a specialized unit that consisted of supervisor (victim-witness coordinator), two victim-witness advocates, and a domestic violence investigator. Victims were contacted within 1–5 days of the incident by a victim-witness advocate, after review of all police arrest incident reports from local law enforcement. In addition to phone calls and subsequent visits by a domestic violence investigator to the victim's residence, if phone calls were not fruitful, each victim was sent a letter and brochure with information on the status of the case, his or her rights and responsibilities, and information on local domestic violence resources. Cases in which a defendant was arrested and detained were prioritized for victim contact and assessment of risk status before drawing up the accusation and making recommendations for acceptance of a plea of guilt or disposition. The prosecutor participated in the initial screening of cases in which the defendant was detained in jail and expressed willingness to plead, with the supervisor conducting the initial screening of all remaining cases.

All cases were screened to identify those appropriate for the domestic violence calendar, which resulted in their diversion from criminal prosecution for 3 months pending the defendant's successful completion of required treatment. High-risk cases not considered suitable for diversion included those where the defendant had a prior record, defied court stay away orders, killed animals or pets, used weapons, objectified his or her partner, threatened or fantasized about homicide/suicide, or destroyed the victim's property or substantially injured the victim. Cases diverted from prosecution were required to complete special programs (Men Stopping Violence, Women's Resource Center, Child Impact Classes, Narcotics Anonymous, or Alcoholics Anonymous) and to abide by certain conditions. If the defendant successfully completed the special program at the end of the 3 months, then the victim was consulted and the case was considered for possible dismissal. Representatives from the two largest service providers/advocacy groups for domestic violence were involved in precourt meetings with victims to explain the court process and information on support groups, hotlines, and shelters; to meet and evaluate victims and defendants; and to make recommendations on how to proceed with a case. Cases not diverted were sent to a prosecutor who considered prosecution in the state court. If a victim wanted to prosecute or the state requested a hearing, then a preliminary hearing was held and the judge determined whether the case moved forward for arraignment. A case might not move forward after arraignment at the request of the victim and at the discretion of the solicitor.

Data Sources

Interview data. All adult female victims of misdemeanor acts of intimate partner violence perpetrated by an adult male partner (spouse, ex-spouse, boyfriend, former boyfriend, and co-parents) were informed about the study by victim-witness advocates assigned to the

solicitor's office in each metro county.[3] During the time frame of this study (March 2000 to June 2002), a total of 1,611 cases entered the court systems (684 in county A and 927 in county B), and valid contact information was available for 58% or 933 of the cases.

Interviewers telephoned all eligible victims where contact information was available to solicit their participation. Potential participants were informed that participation involved an in-person interview, lasting approximately 1 hour, conducted in a convenient and safe setting. Compensation and a list of community resources were provided at the interview's conclusion. After completion of an initial interview, the case was monitored using the court's tracking database until it was disposed (dismissed, diverted, or adjudicated), at which point victims were contacted to schedule a second interview (e.g., disposition interview). The final interview occurred 6 months after the initial disposition of the case. Appropriate methods were used to assure confidentiality of victims' identity and their data.[4] A total of 286 women (31%) agreed to participate in the study. Given the longitudinal nature of our research design, case attrition rates were calculated. For 93 cases, no interview with the victim occurred at disposition or at 6-month follow-up, and for 25 cases, no 6-month follow-up interview was completed. Thus, for a total of 170 cases (59%), victim interview data were collected at all three time points, but this represents only 10.5% of the cases that entered into both jurisdictions in the time frame of the study.

3. We chose to interview only adult female victims because recent victimization surveys indicate that approximately 85% of all adult victims of intimate partner violence are women (Rennison and Welchans, 2000). In both counties, family violence cases were identified by the type of charge and type of relationship between the parties involved. Criminal charges that fell within the domestic violence designation included (a) battery, (b) criminal trespass, (c) harassment or verbal threats, (d) interference with custody, (e) pointing a pistol at another, (f) sexual battery, (g) simple assault, (h) simple battery, (i) theft by taking, and (j) violation of civil protection orders. Relationships that could result in domestic violence designation were the following: (a) spouses, (b) ex-spouses, (c) boyfriend/girlfriend, (d) former boyfriend/girlfriend, (e) homosexual relationships, (f) older relatives, (g) parents who shared a child, (h) abuse of a minor by a relative, (i) siblings living together, (j) parent/child, and (k) grandparent/grandchild.

4. Participants were assigned an ID number placed in a master file with their names, contact information, and batterer's court-assigned case number. This master file was stored in a separate location in a secure filing cabinet away from the completed interview questionnaires. Interview questionnaires contained the participant's ID numbers and no other identifying information. Over the course of the research, three interviews were conducted: at case screening, at initial case disposition, and at 6 months after initial case disposition. At the beginning of each interview, participants were provided with a consent form, which was read to them. Each consent form indicated the approximate length of the interview, the types of information asked, the benefits and risks of participating in the study, the local resources available to the victim, the participant's right to stop the interview at any time without penalty, assurances of confidentiality of the information provided, contact information on the principal investigator and Georgia State University's Research Office, and the amount of compensation for the interview. All interviews were conducted with a trained adult interviewer in a safe location selected by the victim. An office was available on the Georgia State University campus for interviewing. For safety and practicality (to provide childcare for the victim's children if needed), two-person teams were used to conduct interviews when they took place away from the university. At the termination of each interview, the participant was given reimbursement and asked to sign a receipt acknowledging such.

The assessment for sample selection bias consisted of comparing within each county the criminal incidents and criminal arrest histories of participants and a random sample of nonparticipant cases (county A = 260 nonparticipants, county B = 204 nonparticipants). The mean differences in the defendants' number of prior arrests and the defendants' number of criminal charges in the current criminal incident for participant and nonparticipant cases within each county were calculated. This analysis revealed that in county A, the mean number of arrests did not differ significantly for cases of participants and nonparticipants. However, in county B, the cases of participants had a significantly higher number of arrests ($M = 6.55$) than the nonparticipants ($M = 4.60$); $t\ (370) = -2.74$, $p < .01$). With regard to the current criminal incident, the mean number of charges for the participants and nonparticipants did not differ significantly; in county A, participants $M = 3.25$ and nonparticipants $M = 2.88$, and in county B, participants $M = 3.83$ and nonparticipants $M = 3.98$. Overall, this analysis suggests that the defendants in the cases interviewed in county B had more extensive arrest histories and thus may not adequately represent the population of cases that were processed through the court. However, given that the cases of participants contained more serious and not less serious offenders, this yields a conservative test of the differential policy effects. Furthermore, it suggests the importance of controlling for defendants' prior arrests in statistical analyses.

Solicitor's Office/Court Data

The solicitors' offices provided access to information in their court tracking database on the criminal defendants arrested for victimizing study participants. Each case that entered the solicitors' offices was assigned a case number. Knowledge of this number, coupled with access to the court's databases, allowed researchers to monitor the progress of the case through the court system, to determine when it was initially disposed, and to determine what actions were taken by the prosecutor and by the judge (if the case was adjudicated). Specifically, information on the number and type of criminal charges filed per incident, the disposition of the case, and the punishment/sanction imposed were collected. In addition, both offices provided access to case files on each criminal incident. The types of documents typically found in the case files included police incident reports, victim-witness intake sheets, and affidavits of arrest warrant. The police incident reports were standardized within each county but not across the two counties. Furthermore, the information on the victim-witness contact sheets was not uniform across the two counties.

Measures

Reoccurrence of violence in 6 months after disposition. Self-reports of victims' experiences with psychological aggression and violence during the 6 months after case disposition were obtained from their completion of selected scales from the Revised Conflict Tactics Scale (CTS2) (Straus, Hamby, Boney-McCoy, and Sugarman, 1996). Specifically, items from the

Psychological Aggression,[5] Physical Assault,[6] Sexual Coercion,[7] and Physical Injury[8] scales were used. Internal reliability for each scale was psychological aggression (alpha = 0.90), physical assault (alpha = 0.87), sexual coercion (alpha = 0.78), and injury (alpha = 0.78). This allowed for the measurement of the prevalence of abuse or violence within the 6 months after case disposition. The original coding for the item was as follows: 0 = never happened, 1 = once in 6 months, 2 = twice in past 6 months, 4 = 3–5 times in past 6 months, 8 = 6–10 times in past 6 months, 15 = 11–20 times in past 6 months, 25 = more than 20 times in the past 6 months, and 7 = not in past 6 months but happened before. This was further collapsed into a dichotomous measure, 1 = did happen in the past 6 months and 0 = did not happen in the past 6 months. Prevalence rates for the sample were as follows: 42.4% experienced psychological aggression, 13.5% experienced a physical assault, 10.6% experienced sexual coercion, and 6.5% experienced physical injury. For analytical purposes, a single variable, *Experienced Violence in 6-Month Follow-up*, was hierarchically coded as follows: physical assault, sexual coercion, or physical injury = 3; psychological aggression = 2; and no abuse or violence = 1.

Perception of safety. At the 6-month interview, victims were asked, "How likely is it that your partner will physically hurt you in the next 6 months?" Responses were coded using a Likert scale: 3 = very likely, 2 = somewhat likely, and 1 = not at all likely. Nearly three quarters (72.9%) indicated that physical violence was not at all likely, 21.2% indicated that physical violence was somewhat likely, and 5.9% indicated that physical violence was very likely.

5. The Psychological Aggression subscale contained the following eight items: (a) my partner insulted or swore at me; (b) my partner shouted or yelled at me; (c) my partner stomped out of the room or house or yard during a disagreement; (d) my partner said something to spite me; (e) my partner called me fat or ugly; (f) my partner destroyed something that belonged to me; (g) my partner accused me of being a lousy lover; and (h) my partner threatened to hit me or throw something at me.

6. The Physical Assault subscale contained the following ten items: (a) my partner pushed or shoved me; (b) my partner grabbed me; (c) my partner slapped me; (d) my partner used a knife or gun on me; (e) my partner punched or hit me with something that could hurt; (f) my partner choked me; (g) my partner slammed me against the wall; (h) my partner beat me up; (i) my partner burned or scalded me on purpose; and (j) my partner kicked me.

7. The Sexual Coercion subscale contained the following seven items: (a) my partner made me have sex without a condom; (b) my partner insisted on sex when I did not want to (but did not use physical force); (c) my partner insisted that I have oral or anal sex (but did not use physical force); (d) my partner used force (like hitting, holding down, or using a weapon) to make me have oral or anal sex with him; (e) my partner used force (like hitting, holding down, or using a weapon) to make me have sex; (f) my partner used threats to make me have oral or anal sex with him; and (g) my partner used threats to make me have sex.

8. The Physical Injury subscale contained the following six items: (a) I had a sprain, bruises, or small cuts because of a fight with my partner; (b) I felt physical pain the next day because of a fight with my partner; (c) I passed out from being hit in the head by my partner in a fight; (d) I went to the doctor because I had a fight with my partner; (e) I needed to see a doctor because of a fight with my partner, but I didn't; and (f) I had a broken bone from a fight with my partner.

Court empowerment. Court empowerment was measured at the disposition interview using modified items from Cattaneo and Goodman's (2010) measure of court empowerment. The items included (a) "The court considered my rights and wishes just as important as my partner's rights and wishes" and (b) "The court treated me fairly and listened to my side of the story." These two items were measured on a four-point Likert-scale (1 = never, 2 = rarely, 3 = sometimes, and 4 = often). One additional item, measured on a five-point scale (1= not at all to 5 = completely), assessed the degree to which the court outcome satisfied the victim: "How satisfied were you with this outcome (with what happened to your spouse/partner?)" A reliability analysis indicated that these three items formed a reliable scale of court empowerment ($M = 9.21$; standard deviation [SD] = 3.21; Cronbach's alpha = 0.770).

Prosecutorial policy. The type of prosecutorial policy was coded as evidence based = 0 (metro county A) or victim centered = 1 (metro county B). More than half of the cases (55.3%) were processed in the victim-centered jurisdiction, and 44.7% were processed in the evidence-based jurisdiction.

Length of case processing. From the official court documents, the number of days between arrest and case disposition was calculated ($M = 160.61$; SD = 98.80). This variable measures swiftness of punishment, which is one attribute of deterrence.

Disposition. Disposition refers to the action taken by the court and was hierarchically coded from the most to least severe based on information from the official court documents. The least severe action taken, no action beyond arrest (coded 1), comprised 18.2% of the cases and included dismissals or not guilty findings (8.2%), no accusation filed (4.1%), and nolle prose (5.9%). The next most severe action taken, nolo contendere or no contest plea (coded 2), is not a legal admission of guilt but has similar consequences to a guilty plea (29.4%). The next type of disposition, diversion from prosecution (coded as 3), describes when formal action is held in abeyance dependent on whether the defendant successfully completes conditions outlined by the prosecution/court (8.8%). The final and most severe disposition (coded as 4) was guilty plea or a finding of guilt by bench or jury trial (43.5%). This variable measures the deterrent attribute—certainty of punishment.

Severity of punishment. Similarly, numerous punishments may be imposed at conviction. To capture the most severe sanction imposed from official court documents, this variable was hierarchically coded from most to least severe: incarceration in jail/prison = 4 (34.6%), probation = 3 (37.3%), mandated treatment intervention = 2 (18.3%), or dismissals or a not guilty finding = 1 (9.8%). Cases ($n = 17$ or 10%) that were nolle prose or not accused were excluded from analysis. Twelve of the cases that were either nolle prose or not accused came from metro county B (the victim-centered jurisdiction) and five cases that were either nolle prose or not accused came from metro county A (the evidence-based jurisdiction).

Additional Control Variables

Victim's race/ethnicity was coded as Black/African American = 1 (62.9%) and non-Black = 0 (37.1%). The non-Black category comprised 51 Caucasians, 3 Asians, 6 Latinos, and 3 other. *Defendant's prior arrests* was obtained from their criminal histories ($M = 3.84$; SD = 5.74).[9] *Victim lived with abuser throughout follow-up* was obtained by asking at 6-month follow-up if the victim maintained a relationship with the defendant: yes = 1 (49.4%) and no = 0 (50.6%). *Act of intimate partner violence during pretrial processing* captured violence that occurred after the defendant was arrested but prior to disposition of the case. At the disposition interview, each woman was asked: "During the court processing of this case (from arrest to case settlement), did your spouse/partner attempt, threaten, or complete an act of family violence against you?" The responses were coded yes = 1 (21.8%) or no = 0 (78.2%).

Results

Description of Sample

A total of 170 cases comprised the final sample where interview data were available at all three time points, at court intake, at disposition of the case, and at 6 months after disposition; 44.7% of the cases ($n = 76$) were in county A (evidence-based prosecutorial policy operated), and the remainder 55.3% or 94 cases were in county B (victim-centered prosecutorial strategy operated). More than half (62.9%) of the victims were Black, 30% were Caucasian, and the remainder were of Latino or Asian descent. The only significant difference in demographic characteristics across the two jurisdictions was victim's race. Black participants comprised a significantly larger percentage of cases in the victim-centered jurisdiction than in the evidence-based jurisdiction (77.6% compared with 22.4%) ([chi square (1, $N = 170$) = 57.96, $p = .001$, as was the Fisher's exact test, $p = .001$]). Therefore, a control variable for race will be included in the analyses. The average age of victims was 33.3 years (SD = 9.19). Only 11.2% reported not having earned a high-school education or its equivalent. Most women (88.8%) reported having been employed either full or part time in the previous 12 months. Single women who had never been married comprised 37.6% of the cases, and slightly less than one third (31.8%) indicated they were currently married. Close to half (49.4%) of the women reported that they shared children in common with their abusive partner.

9. The skewness and kurtosis of the distribution of number of prior arrests was examined to assure that it did not violate normality assumptions. The skewness was 18.76, and kurtosis (40.18) was a significant departure from acceptable symmetry and peak, requiring transformation prior to analyses in regression models. The transformed version of the variable yielded skew of 1.3 and kurtosis of 2.8.

Bivariate Analyses

Deterrence and therapeutic jurisprudence attributes by prosecutorial policy type.

Length of case processing. Prior research has noted that mandatory prosecution policies result in higher filing rates and longer court oversight (Peterson, 2003; Peterson and Dixon, 2005). On the one hand, one might argue that longer court oversight translates into greater deterrence, as defendants may refrain from criminal activity while under the watchful eye of the court. However, on the other hand, deterrence also posits that punishment is more effective if it is swiftly imposed, as delay in the imposition of punishment dilutes its deterrent effect. We examined difference in the length of case processing in the two jurisdictions studied. The length of time between arrest and disposition in the evidence-based jurisdiction ranged from 4 to 478 days, with an average case processing time of 170.95 days, whereas in the victim-centered jurisdiction, case processing time ranged from 27 to 452 days, with the average slightly lower, 152.25 days. The average number of days between arrest and disposition was compared across the two jurisdictions.[10] The results of an independent-samples *t* test with equal variance not assumed indicated that the length of time between arrest and disposition did not differ significantly across the jurisdictions ($t\,(169) = 1.19, p < .15$).

One rationale for support of evidence-based policies is that such policies permit closer monitoring of defendant's conduct during case processing (Belknap et al., 2000; Peterson and Dixon, 2005; Wills, 1997). Therefore, it is expected that victims will experience fewer threats/acts of physical violence from their partners during pretrial case processing if their cases are handled in an evidence-based jurisdiction compared with a victim-centered jurisdiction. Overall, in approximately one fifth of the cases ($n = 37$), the victim reported that her partner threatened or committed an act of violence toward her prior to disposition during the pretrial phase. However, the differences by jurisdiction were not statistically significant; specifically, 17.1% of cases in the evidence-based jurisdiction and 25.5% in the victim-centered jurisdiction experienced violence during case processing, chi square ($1, N = 170) = 1.75, p < .10$.

Disposition and sanctions. According to prior research (Klein, 2009; Smith et al., 2008), evidence-based jurisdictions will dispose of a larger percentage of cases through formal disposition; in other words, cases are less likely to be dismissed. The end result is more certainty that the incident will be formally disposed of by the court. Indeed, in the evidence-based jurisdiction, few (6.6%) cases were disposed of by taking no action beyond arrest meaning the cases had no accusation filed, were nolle prose, were dismissed, or a not guilty verdict was rendered. Furthermore, in only one case (1.3%) was diversion recommended. The most common disposition was nolo contendere or no contest (51.3%), followed by

10. The skewness and kurtosis of the distribution of number of days between arrest and disposition was examined to assure that it did not violate normality assumptions and the *F* ratio (Levene's test for equality of variances) was significant ($F(1, 170) = 5.23, p < .05$).

guilty plea/finding of guilt (40.8%). Thus, over 9 of 10 cases ended in a disposition involving a plea or finding by the court. Comparatively, in the victim-centered jurisdiction, the most common disposition was the guilty plea/finding of guilt (45.7%), followed by no action beyond arrest (27.7%). Diversion from disposition was used in 14.9% of the cases and few (11.7%) were resolved by nolo contendere/no contest pleas. Chi square tests of significance indicated that the dispositions differed significantly by jurisdiction (chi square (3, $N = 170$) $= 41.68$, $p < .001$).

Theoretically, given their deterrent aim, evidence-based jurisdictions should impose more severe sanctions than victim-centered jurisdictions. Punishments across the two jurisdictions varied significantly (chi square (3, $N = 153$) $= 78.53$, $p < .001$). In the evidence-based jurisdiction, 71.8% of the cases were punished with a probation term, 26.8% received a term of jail incarceration, one case (1.4%) was referred for treatment intervention, and none were dismissed at trial or found not guilty. In the victim-centered jurisdiction, comparatively, 41.5% received jail incarceration, 32.9% were referred to domestic violence treatment intervention, 18.3% were dismissed at trial or found not guilty, and 7.3% were referred to probation.

Court empowerment. Prior research has posited that court empowerment will be higher at court disposition in victim-centered prosecutorial policy jurisdictions compared with evidence-based jurisdictions (Buzawa et al., 1999; Cattaneo and Goodman, 2010; Davis et al., 2008; Ford and Regoli, 1993a, 1993b). Our analysis examined the differences in the level of court empowerment at disposition across the evidence-based and victim-centered jurisdiction. The results of independent-samples t test indicated that empowerment levels were not higher in the jurisdiction employing victim-centered policies ($M = 9.38$; SD $=$ 3.15) compared with the jurisdiction using evidence-based policies ($M = 8.99$; SD $= 3.30$; $t\,(143) = -0.799$, $p < .10$).

Multivariate Analyses

Predicting reoccurrence of violence and victims' perceptions of safety. Both evidence-based policies and victim-centered policies claim that victim safety is enhanced through either the deterrent effects of sanctions or the therapeutic effects of court interaction. Overall, nearly half (44.7%) of victims experienced a reoccurrence of violence or abuse in the 6 months immediately after case disposition, and slightly more than one quarter (27.1%) reported at the final interview that it was somewhat ($n = 36$) or very likely ($n = 10$) that their partner would physically hurt them in the next 6 months.

Multivariate analyses were employed to conduct a more rigorous assessment of the independent effects of therapeutic and deterrent attributes of court action, as well as prosecutorial policy type, on the reoccurrence of violence and victims' perceptions of safety. For the dependent variable, reoccurrence of violence, multinomial logistic regression analyses were used to yield a single model containing the results for two comparisons: the likelihood of a victim experiencing psychological aggression versus no abuse and the likelihood of a

victim experiencing physical violence versus no abuse in the 6 months after disposition. The second dependent variable of interest, victims' perception of safety, was analyzed using logistic regression, which predicted the likelihood of a victim reporting that violence was somewhat or very likely to occur versus a victim reporting that violence was not at all likely to occur.[11] The use of multinomial and logistic regression models permitted an examination of the effects of several theoretically relevant independent variables and control variables on each outcome. Each model included the following independent variables: swiftness, certainty, and severity of punishment (elements of deterrence); court empowerment (element of therapeutic jurisprudence); and victim-centered or evidence-based prosecutorial policy; as well as four control variables: whether intimate partner violence occurred during the pretrial stage of the process (after arrest and before disposition), defendants' prior arrests, race of victim, and relationship status of victim and abuser. Each model yielded coefficients and change in odds for variables that permit assessment of their relative effects on the outcomes of interest—reoccurrence of violence and perceptions of safety.

Reoccurrence of violence in 6 months after disposition. The results of the first model predicting the likelihood of a victim experiencing a reoccurrence of violence are presented in Table 1. For this model, reoccurrence of violence was hierarchically coded into three categories: victim reported physical violence, victim reported psychological aggression, or victim reported no violence or aggression, with the last category serving as the reference category. The multinomial logistic regression analysis produced two sets of comparisons. Comparison 1 reports the coefficient estimates (*B*) and odds ratios for the independent variables and the control variables predicting the likelihood of psychological aggression occurring versus no violence. Comparison 2 reports the coefficient estimates (*B*) and odds ratios for the independent variables and the control variables predicting the likelihood of physical violence occurring versus no violence. The odds ratios are more intuitive to interpret than the coefficients. An odds ratio larger than 1 indicates that, as the independent variable increases, the likelihood that a victim reported the occurrence of physical violence during follow-up (comparison) versus no violence (reference) also increases. Conversely, an odds ratio of less than 1 indicates that, as the independent variable decreases, the likelihood that a victim reported the occurrence of physical violence during follow-up (comparison) versus no violence (reference) also decreases.

With regard to comparison 1 that predicts the report of psychological aggression versus no violence in the 6 months after disposition, the results indicate that the operation of an evidence-based prosecutorial policy significantly affected the outcome. Victims whose cases were prosecuted in the jurisdiction employing evidence-based prosecutorial policies,

11. The second dependent variable of interest, victims' perception of safety, was measured on a scale that would have permitted multinomial regression, but only 5.9% of the cases indicated that physical violence was very likely, and this number was too few to support such analyses so logistic regression was used instead.

TABLE 1

Coefficients for Multinomial Model Predicting Victims' Self-Reports of Psychological Aggression and Physical Violence

	Comparison 1: Victims' Self-Report of Psychological vs. No Abuse in 6 Months After Disposition			Comparison 2: Victims' Self-Report of Physical Violence vs. No Abuse in 6 Months After Disposition		
Variables	***B* (SE)**	**Odds Ratio (Confidence Interval)**	**Wald Statistic**	***B* (SE)**	**Odds Ratio (Confidence interval)**	**Wald Statistic**
Length of time from arrest to disposition	−0.004 (0.002)	0.996 (0.991–1.00)	3.29†	−0.001 (0.003)	0.999 (0.994–1.00)	0.170
Disposed by guilty plea/finding	−0.408 (0.430)	0.665 (0.286–1.55)	0.900	−0.432 (0.509)	0.649 (0.239–1.76)	0.721
Punished with incarceration	0.576 (0.485)	1.78 (0.688–4.59)	1.41	−0.289 (0.526)	0.749 (0.267–2.10)	0.301
Court empowerment at disposition	−0.050 (0.060)	0.951 (0.845–1.07)	0.686	0.066 (0.074)	1.068 (0.924–1.23)	0.790
Evidence-based prosecutorial policy	1.33 (0.551)	3.76 (1.28–11.07)	5.80*	1.97 (0.700)	7.17 (1.82–28.28)	7.92**
Physical violence during pretrial	−0.901 (0.477)	0.406 (0.160–1.03)	3.58†	−1.17 (0.535)	0.310 (0.108–0.884)	4.80*
Prior arrests of defendant[a]	.0042 (0.175)	1.04 (0.740–1.47)	0.058	0.220 (0.199)	1.25 (0.844–1.84)	1.22
Victim lived with abuser after disposition	−0.070 (0.394)	0.932 (0.431–2.02)	0.032	−0.199 (0.464)	0.820 (0.330–2.04)	0.184
Black victim	0.970 (0.557)	2.64 (0.887–7.85)	3.04†	1.76 (0.677)	5.80 (1.54–21.80)	6.77**
Constant	−1.16 (0.913)			−0.515 (1.005)		
n	170					
Model chi square	34.32*					
McFadden	0.103					
df	18					

Notes. df = degrees of freedom; SE = standard error.

[a]The transformed version yielded skew of 1.3 and kurtosis of 2.80. Prior to analysis, this variable was transformed to address its positive skew (3.49) and kurtosis (14.69) by taking its square root. The transformed version yielded skew of 1.3 and kurtosis of 2.80.

†$p < .10$, *$p < .05$, **$p < .01$.

compared with the victim-centered policy jurisdiction, were nearly four times more likely to report that psychological aggression reoccurred (adjusted odds ratio = 3.76, $p < .05$). None of the elements of deterrence (the swiftness, certainty, or severity of punishment imposed by the court) or therapeutic jurisprudence (court empowerment) significantly predicted the reoccurrence of psychological aggression. None of the control variables (whether an act of intimate partner violence occurred during pretrial, the number of prior arrests of the defendant, whether the victim and abuser maintained a relationship after disposition, or the race of the victim) was significantly related to reports of the occurrence of psychological aggression versus no violence.

Considering comparison 2 that predicts victims' report of physical violence (i.e., physical assault, sexual coercion, or physical injury) versus no violence in the 6 months after disposition, the results again indicate a significant effect for type of prosecutorial policy. Victims whose cases were prosecuted in the jurisdiction with the evidence-based prosecutorial policy in place, compared with the victim-centered policy jurisdiction, were more than seven times more likely to report that physical violence reoccurred than no violence took place (adjusted odds ratio = 7.17, $p < .01$). Consistent with the previous comparison, none of the deterrence measures or the therapeutic jurisprudence measures affected victims' reports of the reoccurrence of physical violence. Two control variables significantly affected this outcome. First, victims who reported having experienced the threat, attempt, or completion of an act of intimate partner violence during pretrial were less likely to report that physical violence had occurred in the 6 months after disposition of the case than to report no violence. The change in relative risk was substantial (adjusted odds ratio = 0.310, $p < .05$). Second, being Black or African American significantly increased the likelihood that a victim would experience physical violence rather than no violence (adjusted odds ratio = 5.80; $p < .001$).

Victims' perception of safety in the future. Table 2 presents the results of the logistic regression model predicting victims' perceptions of safety in the future (in the 6 months after the final interview). The findings indicate that neither attributes of deterrence or therapeutic jurisprudence, nor the type of prosecutorial policy, had significant effects on victims' perceptions of safety. Indeed, only one variable, whether a victim reported having experienced physical violence (physical assault, sexual coercion, or physical injury) during the follow-up period, was statistically significant. Those who experienced physical violence during the 6 months after disposition were more than five times more likely (adjusted odds ratio = 5.07, $p < .001$) to report that physical violence was somewhat or very likely to occur in the 6 months after the final interview, which was conducted 6 months after case disposition.

Discussion

This study examined the influence of prosecutorial policy on the reoccurrence of violence and perceptions of future safety in the lives of domestic violence victims after their

TABLE 2

Coefficients for Logistic Regression Model Predicting Victims' Perception of Safety in 6 Months After Final Interview

	B (SE)	Odds Ratio (Confidence Interval)	Wald Statistic
Length of time from arrest to disposition	−0.001 (0.002)	0.999 (0.995–1.00)	0.264
Disposed by guilty plea/finding	0.240 (0.404)	1.27 (0.576–2.80)	0.353
Punished with incarceration	0.085 (0.444)	1.09 (0.456–2.59)	0.036
Court empowerment at disposition	0.012 (0.059)	1.01 (0.902–1.14)	0.044
Evidence-based prosecutorial policy	0.354 (0.474)	1.43 (0.563–3.61)	0.559
Physical violence during 6-month follow-up	1.62 (0.455)	5.07 (2.08–12.39)	12.73***
Number of prior arrests of defendant[a]	−0.152 (0.179)	0.859 (0.605–1.22)	0.723
Victim lived with abuser after disposition	−0.121 (0.379)	0.886 (0.421–1.86)	0.102
Black victim	−0.535 (0.464)	0.585 (0.236–1.45)	1.33
Constant	−0.996 (0.917)		
n	170		
−2Log likelihood	179.34		
Model chi square	19.16*		
df	9		

Notes. *df* = degrees of freedom; SE = standard error.

[a]Prior to analysis, this variable was transformed to address its positive skew (3.49) and kurtosis (14.69) by taking its square root. The transformed version yielded skew of 1.3 and kurtosis of 2.80.

*$p < .05$, ***$p < .001$.

involvement with the criminal court. First, the analysis examined the degree to which the operation of evidence-based or victim-centered policies produced differences in key elements of deterrence and therapeutic jurisprudence. It was expected that the evidence-based jurisdiction would display attributes characteristic of deterrence, whereas the victim-centered jurisdiction would display attributes characteristic of therapeutic jurisprudence. However, findings were mixed with regard to whether the evidence-based jurisdiction displayed more deterrent attributes than the victim-centered jurisdiction. Although the findings yielded no evidence that the jurisdiction employing an evidence-based policy pursued justice more swiftly than the jurisdiction employing a victim-centered policy, significant differences emerged when the certainty and severity of punishment were examined. The evidence-based jurisdiction was far more likely than the victim-centered jurisdiction to dispose of cases formally through an action that required judicial involvement in the case and opportunity for subsequent punishment of the defendant, either through the (a) entering of a plea of guilt or no contest or (b) a finding of guilt. Furthermore, in the evidence-based jurisdiction, it was rare for cases to be dismissed (less than 1 in 10) and referral to treatment in lieu of prosecution was rarely used. Comparatively, dismissal of charges and referrals of batterers to domestic violence treatment intervention were far more common in the victim-centered policy jurisdiction. Both findings coincide with prior research evidence that victims often

request no further criminal actions be taken by the court, and hence, victim-centered jurisdictions may consider and honor such requests (Davis et al., 2008). Somewhat unexpected was the finding that incarceration was used more frequently in the victim-centered policy jurisdiction, whereas probation was used more frequently in the evidence-based jurisdiction. Neither of these findings was predicted based on the assumptions that efforts at deterrence would foster greater use, rather than less use, of incarceration as a punishment.

Although we expected that victim-centered policies would enhance court empowerment, whereas evidence-based policies would not, the results found no differences in court empowerment after case disposition by type of prosecutorial policy. This finding, coupled with our consistent findings in multivariate analyses that cases processed in victim-centered jurisdictions reported better outcomes for victims (i.e., they were less likely to report the reoccurrence of violence or psychological aggression after court involvement), suggests that the adoption of policy alone may not translate into predictable and direct effects on court empowerment as reported by victims. It further suggests that how policy translates and shapes interactions between victims and the court, especially with prosecutors and victim-witness advocates, warrants further study. Future research should examine the nature of interactions between prosecutors and victims, and how the quality and quantity of such interactions may differ based on the type of prosecutorial policy in effect.

Cattaneo and Goodman (2010) suggested that it is important to consider what the victim hopes to accomplish in her interactions with the court and the degree to which a victim's aim coincides with that of prosecutors. It is reasonable to expect that when the goals of a victim and a prosecutor align in that both seek the same outcome, whether that be toward conviction and punishment of the batterer or toward avoiding criminal penalties and instead seeking domestic violence services, that the victim should experience a higher level of court empowerment than when her goals do not align with those of the prosecutor. Our current study did not include such measures, but we recommend that future researchers consider their inclusion.

The measure of court empowerment employed in this study, although having been used in prior research and demonstrating moderate reliability, might have been inadequate. Recent work by Cattaneo, Dunn, and Chapman (2013), published after data collection and completion of this article, reported the results of the pilot test of a new measure of the impact of the court on empowerment that seems promising in its ability to capture more clearly the concept of court empowerment. Future researchers are encouraged to consider its incorporation.

Our analyses shed important light on the influences of evidence-based policy, theoretically aligned with deterrence theory, and victim-centered policy, theoretically linked to therapeutic jurisprudence, on the reoccurrence of violence in the lives of victims. Such analyses permitted an examination of the influence of these policies and key elements of deterrence (swiftness, certainty, and severity of punishment) and therapeutic jurisprudence (court empowerment) in both jurisdictions while controlling for factors identified in prior

research as related to the reoccurrence of violence. Overall, the results yielded no support for the deterrence model, confirming the findings of prior research that found neither case disposition nor type of punishment imposed influences offenders' recidivism (Davis et al., 1998; Ford and Regoli, 1993a, 1993b; McFarlane et al., 2000; Mears et al., 2001; Tolman and Weisz, 1995). The analyses tested the policy effects on specific deterrence and not general deterrence, however, and our results cannot speak to the argument that mandatory prosecution policies may be beneficial in deterring potential batterers.

Perhaps most importantly, the results suggest that the aims of therapeutic jurisprudence, accomplished through the adoption of victim-centered prosecutorial policies, yield better outcomes for victims of intimate partner violence. This finding reinforces those of prior scholars (Buzawa et al., 1999; Cattaneo and Goodman, 2010; Ford and Regoli, 1993a, 1993b) and suggests that efforts to improve the therapeutic nature of the courts may have potential long-term influences on reducing revictimization and their beliefs about their future safety.

The current research has several limitations that warrant consideration in interpreting the findings, including the sample selection and size and a short observation period. First, the sample, which was drawn from two jurisdictions within a single state, has the advantage of similar state laws and criminal procedures operating, but it has the disadvantage in that the generalizability of findings to other locations is uncertain. Replication of the findings in additional jurisdictions, and if possible with a more representative sample, is warranted to build greater confidence in the findings. Second, the sample of victims who volunteered for interviews might differ in fundamental ways from those who did not volunteer. Although we compared the samples drawn from the two jurisdictions on several key attributes, we could not ascertain whether those who volunteered for interview differed demographically or in their victimization histories from those who did not volunteer. To assess such selection bias in the future, we encourage prosecutors' offices to collect accurate demographic information on victims on intake forms. Third, our measures of the reoccurrence of violence and victims' perception of likelihood of future harm were based exclusively from reports provided by the victims at various stages of the process (intake and disposition) and for 6 months after disposition. Although this is a significant improvement over prior research that focused almost exclusively on official reports of rearrests of defendants, the current study would have benefited from cross-validation of victims' reports of reoccurrence of violence with reports of such violence to law enforcement and court authorities.

In closing, whereas the current research was conducted in traditional court settings and not in specialty court settings, the findings do lend support to current efforts to expand the use of specialty courts to address intimate partner violence. Although more than 200 domestic violence courts operated in 32 states in 2009 (more than half were concentrated in the states of California and New York), to date their adoption has been piecemeal (Labriola et al., 2009). Recent research on the effectiveness of domestic violence courts in the state of New York found that domestic violence courts that prioritized deterrence through

implementation of policies that (a) sanction offenders' noncompliance while under court supervision and (b) address the needs of victims are the most effective at reducing recidivism (Cissner, Labriola, and Rempel, 2013). Indeed, in the state of Georgia currently, a few local courts have domestic violence calendars or court dockets, but no stand-alone domestic violence courts operate. Importantly, the jurisdiction that adopted the victim-centered policy described herein operated a domestic violence calendar. Elements of this calendar aligned with some characteristics of domestic violence courts, especially the integration into the court of domestic violence service provision for victims and batterer intervention programs for defendants. However, sanctioning noncompliance of offenders' cases assigned to this calendar, an important tool for assuring offender accountability and linked to decreased recidivism, was not examined in the current study but should be in future research. Recently, the Administrative Office of the Courts in the state of Georgia has created a statewide Task Force on Domestic Violence Courts focused on establishing domestic violence courts based on "victim-centered, best-practices approach to holding offenders accountable for their actions" (Administrative Office of the Courts, 2013). Our findings suggest that this step is encouraging for the state to undertake as it holds promise for improving the safety and well-being of victims of domestic violence.

References

Administrative Office of the Courts. 2013. Task Force on Domestic Violence Courts. Retrieved July 15, 2013 from w2.georgiacourts.gov/dvtaskforce/index.php?option=com_content&view=article&id=50&Itemid=55.

Akey, Theresa M., Janet G. Marquis, and Margaret E. Ross. 2000. Validation of scores on the psychological empowerment scale: A measure of empowerment for parents of children with a disability. *Educational and Psychological Measurement*, 60: 419–438.

Andrews, Arlene Bowers, José L. Guadalupe, and Errol Bolden. 2003. Faith, hope, and mutual support: Paths to empowerment as perceived by women in poverty. *Journal of Social Work Research*, 4: 5–18.

Badawy, Rami S. 2011. *The Supreme Court Clarifies the Primary Purpose Test Michigan v. Bryant, 562 U.S. (2011)*. Alexandria, VA: National District Attorneys Association's National Center for Prosecution of Child Abuse.

Belknap, Joanne, Dee L. R. Graham, Jennifer Hartman, Victoria Lippen, P. Gail Allen, and Jennifer Sutherland. 2000. *Factors Related to Domestic Violence Court Dispositions in a Large Urban Area: The Role of Victim/Witness Reluctance and Other Variables*. Washington, DC: National Institute of Justice.

Belknap, Joanne and Hillary Potter. 2005. The trials of measuring the "success" of domestic violence policies. *Criminology & Public Policy*, 4: 559–566.

Bell, Margret E., Lisa A. Goodman, and Mary Ann Dutton. 2007. The dynamics of staying and leaving: Implications for battered women's emotional well-being and experiences of violence at the end of a year. *Journal of Family Violence*, 22: 413–428.

Buzawa, Eve S. and Aaron D. Buzawa. 2008. Courting domestic violence victims: A tale of two cities. *Criminology & Public Policy*, 7: 671–685.

Buzawa, Eve S., Gerald T. Hotaling, Andrew Klein, and James Byrne. 1999. *Response to Domestic Violence in a Pro-Active Court Setting: Final Report*. Washington, DC: National Institute of Justice.

Cattaneo, Lauren Bennett and Aliya R. Chapman. 2010. The process of empowerment: A model for use in research and practice. *American Psychologist*, 65: 646–659.

Cattaneo, Lauren Bennett, Jessica L. Dunn, and Aliya R. Chapman. 2013. The court impact scale: A tool for evaluating IPV victims' experience in court. *Journal of Interpersonal Violence*, 28: 1088–1108.

Cattaneo, Lauren Bennett and Lisa A. Goodman. 2010. Through the lens of therapeutic jurisprudence: The relationship between empowerment in the court system and well-being for intimate partner violence victims. *Journal of Interpersonal Violence*, 25: 481–502.

Chronsiter, Krista M. and Ellen Hawley McWhirter. 2003. Applying social cognitive career theory to the empowerment of battered women. *Journal of Counseling & Development*, 81: 418–425.

Cissner, Amanda B., Melissa Labriola, and Michael Rempel. 2013. *Testing the Effects of New York's Domestic Violence Courts: A State-Wide Impact Evaluation*. New York: Center for Court Innovation.

Corsilles, Angela. 1994. No-drop cases in the prosecution of domestic violence cases: Guarantee to action or dangerous solution. *Fordham Law Review*, 63: 853–881.

Davis, Robert C., Chris S. O'Sullivan, Donald J. Farole, Jr., and Michael Rempel. 2008. A comparison of two prosecution policies in cases of intimate partner violence: Mandatory case filing versus following the victim's lead. *Criminology & Public Policy*, 7: 633–662.

Davis, Robert C., Barbara E. Smith, and Laura B. Nickles. 1998. The deterrent effect of prosecuting domestic violence misdemeanors. *Crime & Delinquency*, 44: 434–442.

Davis, Robert C., Barbara E. Smith, and Bruce Taylor. 2003. Increasing the proportion of domestic violence arrests that are prosecuted: A natural experiment in Milwaukee. *Criminology & Public Policy*, 2: 263–282.

Dempsey, Michelle Madden. 2009. *Prosecuting Domestic Violence: A Philosophical Analysis.* New York: Oxford University Press.

Diesen, Christian. 2011. Therapeutic jurisprudence and the victim of crime. *Progression in Forensic Psychiatry.* Retrieved July 15, 2013 from scribd.com/doc/124602238/Therapeutic-Jurisprudence-and-the-Victim-of-Crime.

Epstein, Deborah. 1999. Effective intervention in domestic violence cases: Rethinking the roles of prosecutors, judges, and the court system. *Yale Journal of Law & Feminism*, 11: 3–50.

Erez, Edna and Carolyn Copps Hartley. 2003. Battered immigrant women and the legal system: A therapeutic jurisprudence perspective. *Western Criminology Review*, 4: 155–169.

Ferraro, Kathleen J. and Lucille Pope. 1993. Irreconcilable differences: Battered women, police, and the law. In (N. Zoe Hilton, ed.), *Legal Responses to Wife Assault: Current Trends and Evaluation*. Newbury Park, CA: Sage.

Fine, Andrew C. 2006. Refining *Crawford*: The Confrontation Clause after *Davis v. Washington* and *Hammon v. Indiana*. *Michigan Law Review First Impressions*, 105: 11–15.

Ford, David A. 1991. Prosecution as a victim power resource: A note on empowering women in violent conjugal relationships. *Law & Society Review*, 25: 313–334.

Ford, David A. and Mary Jean Regoli. 1993a. *The Indianapolis Domestic Violence Prosecution Experiment: Final Report*. Washington, DC: National Institute of Justice.

Ford, David A. and Mary Jean Regoli. 1993b. The criminal prosecution of wife assaulters. In (N. Zoe Hilton, ed.), *Legal Responses to Wife Assault: Current Trends and Evaluation*. Newbury Park, CA: Sage.

Gewirtz, Abigail, Robert R. Weidner, Holly Miller, and Keri Zehm. 2006. Domestic violence cases involving children: Effects of an evidence-based prosecution approach. *Violence and Victims*, 21: 213–229.

Goodman, Lisa A., Lauren Bennett, and Mary Ann Dutton. 1999. Obstacles to victims' cooperation with criminal prosecution of their abusers: The role of social support. *Violence and Victims*, 14: 427–444.

Goodman, Lisa A. and Deborah Epstein. 2007. *Listening to Battered Women: A Survivor Centered Approach to Advocacy, Mental Health and Justice*. Washington, DC: American Psychological Association.

Goolkasian, Gail A. 1986. *Confronting Domestic Violence: A Guide for Criminal Justice Agencies*. Washington, DC: National Institute of Justice.

Gover, Angela R., Even M. Brank, and John M. MacDonald. 2007. A specialized domestic violence court in South Carolina: An example of procedural justice for victims and defendants. *Violence Against Women*, 13: 603–626.

Han, Erin L. 2003. Mandatory arrest and no-drop policies: Victim empowerment in domestic violence cases. *Boston College Third World Law Journal*, 23: 159–191.

Hanna, Cheryl. 1996. No right to choose: Mandated victim participation in domestic violence prosecutions. *Harvard Law Review*, 109: 1849–1910.

Hart, Barbara J. 1993. Battered women and the criminal justice system. *American Behavioral Scientist*, 36: 624–638.

Hartley, Carolyn Copps. 2003. A therapeutic jurisprudence approach to the trial process in domestic violence felony trials. *Violence Against Women*, 9: 410–437.

Herman, Judith L. 1997. *Trauma and Recovery.* New York: Basic Books.

Jacoby, Joan E. 1980. *The American Prosecutor: A Search for Identity*. New York: Lexington Books.

Klein, Andrew R. 2009. *Practical Implications of Current Domestic Violence Research: For Law Enforcement, Prosecutors, and Judges*. Washington, DC: U.S. Department of Justice.

Krug, Peter. 2002. Prosecutorial discretion and its limits. *American Journal of Comparative Law*, 50: 643–664.

Labriola, Melissa, Sarah Bradley, Chris S. O'Sullivan, Michael Rempel, and Samantha Moore. 2009. *A National Portrait of Domestic Violence Courts*. New York: Center for Court Innovation.

Lerman, Lisa G. 1981. *Prosecution of Spouse Abuse: Innovations in Criminal Justice Responses*. Washington, DC: Center for Women Policy Studies.

Maxwell, Christopher D., Joel H. Garner, and Jeffrey A. Fagan. 2001. *Effects of Arrest on Intimate Partner Violence: New Evidence from the Spouse Assault Replication Program*. Washington, DC: National Institute of Justice.

McCord, Joan. 1992. Deterrence of domestic violence: A critical review of research. *Journal of Research in Crime & Delinquency*, 29: 229–239.

McDermott, M. Joan and James Garafalo. 2004. When advocacy for domestic violence victims backfires. *Violence Against Women*, 10: 1245–1266.

McFarlane, Judith, Pam Willson, Dorothy Lemmey, and Ann Malecha. 2000. Women filing assault charges on an intimate partner: Criminal justice outcome and future violence experienced. *Violence Against Women*, 6: 396–408.

Mears, Daniel P., Matthew J. Carlson, George W. Holden, and Susan D. Harris. 2001. Reducing domestic violence revictimization: The effects of individual and contextual factors and type of legal intervention. *Journal of Interpersonal Violence*, 16: 1260–1283.

Messing, Jill Theresa. 2010. Evidence-based prosecution of intimate partner violence in the post-Crawford era: A single-city study of the factors leading to prosecution. *Crime & Delinquency*. E-pub ahead of print. DOI: 10:1177/0011128710362056.

Mills, Linda G. 1996. Empowering battered women transnationally: The case for postmodern interventions. *Social Work*, 41: 261–268.

Mills, Linda G. 1998. Mandatory arrest and prosecution policies for domestic violence: A critical literature review and the case for more research to test victim empowerment approaches. *Criminal Justice and Behavior*, 25: 306–318.

Mosteller, Robert P. 2005. *Crawford v. Washington* encouraging and ensuring the confrontation of witnesses. *University of Richmond Law Review*, 39: 511–626.

Peterson, Richard R. 2002. *Cross-Borough Differences in the Processing of Domestic Violence Cases in New York City Criminal Courts*. New York: New York City Criminal Justice Agency.

Peterson, Richard R. 2003. *The Impact of Case Processing on Re-Arrests Among Domestic Violence Offenders in New York City*. New York: New York City Criminal Justice Agency.

Peterson, Richard R. and Jo Dixon. 2005. Court oversight and conviction under mandatory and nonmandatory domestic violence case filing policies. *Criminology & Public Policy*, 4: 535–558.

Rebovich, Donald J. 1996. Prosecution response to domestic violence: Results of a survey of large jurisdictions. In (Eve S. Buzawa and Carl G. Buzawa, eds.), *Do Arrests and Restraining Orders Work?* Thousand Oaks, CA: Sage.

Rennison, Callie Marie and Sarah Welchans. 2000. *Intimate Partner Violence*. Washington, DC: U.S. Department of Justice.

Ross, Josephine. 2007. *Crawford's* short-lived revolution: How *Davis v Washington* reins in *Crawford's* reach. *North Dakota Law Review*, 83: 387–462.

Sherman, Lawrence W. and Richard A. Berk. 1984. The specific deterrent effects of arrest for domestic assault. *American Sociological Review*, 49: 261–272.

Slate, Risdon N. 2003. From jailhouse to Capitol Hill: Impacting mental health court legislation and defining what constitutes a mental health court. *Crime & Delinquency*, 49: 6–29.

Smith, Barbara E., Robert C. Davies, Laura B. Nickles, and Heather J. Davies. 2001. *Evaluation of Efforts to Implement No-Drop Policies: Two Central Values in Conflict, Final Report*. Washington, DC: U.S. Department of Justice.

Smith, Erica L., Matthew R. Durose, and Patrick A. Langan. 2008. *State Court Processing of Domestic Violence Cases*. Washington, DC: Bureau of Justice Statistics.

Straus, Murray A., Sherry L. Hamby, Sue Boney-McCoy, and David B. Sugarman. 1996. The revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues*, 17: 283–316.

Thistlethwaite, Amy, John Wooldredge, and David Gibbs. 1998. Severity of dispositions and domestic violence recidivism. *Crime & Delinquency*, 44: 388–398.

Tolman, Richard M. and Arlene Weisz. 1995. Coordinated community intervention for domestic violence: The effects of arrest and prosecution on recidivism of woman abuse perpetrators. *Crime & Delinquency*, 41: 481–495.

U.S. Census Bureau. 2000. Census Bureau Summary File 1 and 3. Retrieved July 15, 2013 from factfinder.census.gov.

van Uden-Kraan, Cornelia F., Constance H. C. Drossaert, Erik Taal, Bret R. Shaw, Erwin R. Seydel, and Mart A. F. J. van de Laar. 2008. Empowering processes and outcomes of participation in online support groups for patients with breast cancer, arthritis, or fibromyalgia. *Qualitative Health Research*, 18: 405–417.

Waites, Kathleen. 1985. The criminal justice systems response to battering: Understanding the problem, forging the solutions. *Washington Law Review*, 60: 267–329.

Wemmers, Jo-Anne. 2008. Victim participation and therapeutic jurisprudence. *Victims and Offenders*, 3: 165–191.

Wemmers, Jo-Anne and Katie Cyr. 2005. Can mediation be therapeutic for crime victims? An evaluation of victims' experiences in mediation with young offenders. *Canadian Journal of Criminology and Criminal Justice*, 47: 527–544.

Wexler, David B. (ed.). 1990. *Therapeutic Jurisprudence: The Law as a Therapeutic Agent.* Durham, NC: Carolina Academic Press.

Wills, Donna. 1997. Domestic violence: The case for aggressive prosecution. *UCLA Women's Law Journal*, 7: 173–182.

Winick, Bruce J. 1996. The jurisprudence of therapeutic jurisprudence. In (David B. Wexler and Bruce J. Winick, eds.), *Law in a Therapeutic Key: Developments in Therapeutic Jurisprudence*. Durham, NC: Carolina Academic Press.

Winick, Bruce J. 1997. The jurisprudence of therapeutic jurisprudence. *Psychology, Public Policy & Law*, 3: 184–206.

Winick, Bruce J. and David B. Wexler (eds.). 2003. *Judging in a Therapeutic Key: Therapeutic Jurisprudence and the Courts.* Durham, NC: Carolina Academic Press.

Wood, Gale Goldberg and Ruth R. Middleman. 1992. Groups to empower battered women. *Affilia*, 7: 82–95.

Worrall, John L. 2008. Prosecutors in problem-solving courts. In (John L. Worrall and M. Elaine Nugent-Borakov, eds.), *The Changing Role of the American Prosecutor.* Albany: State University of New York Press.

Worrall, John L., Jay W. Ross, and Eric S. McCord. 2006. Modeling prosecutors' charging decisions in domestic violence cases. *Crime & Delinquency*, 52: 472–503.

Zimmerman, Marc A. 1995. Psychological empowerment: Issues and illustrations. *American Journal of Community Psychology*, 23: 581–599.

Zimmerman, Marc A. 2000. Empowerment theory: Psychological, organizational and community levels of analysis. In (Julian Rappaport and Edward Seidman, eds.), *Handbook of Community Psychology*. New York: Kluwer Academic Plenum.

Zorza, Joan. 2010. Empowering battered women, expanding their options, honoring their choices. *Family & Intimate Partner Violence Quarterly*, 3: 109–121.

Court Cases Cited

Crawford v. Washington, 541 U.S. 36 (2004).

Davis v. Washington, 126 S. Ct. 2266 (2006).

Michigan v. Bryant, 562 U.S. (2011).

Mary A. Finn is a professor in the Department of Criminal Justice & Criminology, Andrew Young School of Policy Studies, Georgia State University. Her research interests include crime policy and evaluation of justice system responses to crime, primarily violence against women and children. Her most recent publications appear in *Crime & Delinquency* and *Journal of Interpersonal Violence*.

VICTIM-CENTERED PROSECUTORIAL POLICIES

Victim Engagement in the Prosecution of Domestic Violence Cases

Richard R. Peterson
New York City Criminal Justice Agency

Policies and practices for prosecuting domestic violence (DV) cases vary widely. Some District Attorney's (DA's) offices file charges in most DV cases (a policy sometimes described as "mandatory filing" or "universal filing"), whereas others file charges only in cases that have strong evidence for conviction. Once they file charges, some prosecute all cases (sometimes described as "mandatory prosecution" or a "no-drop" policy); others drop charges if it becomes clear that conviction is unlikely. Some DA's offices divert cases from criminal prosecution while a defendant participates in a mandated intervention program, whereas others do not use diversion. Among those whose policies allow discretion in filing charges, dropping charges, or diverting cases, some allow victims to participate in and influence these decisions (sometimes described as a "victim-centered" policy), whereas others minimize victim participation.

Some DA's offices subpoena reluctant victims to testify in DV cases, whereas others do not. Regardless of their policy on subpoenas, some use "evidence-based" prosecution to pursue cases when the victim does not testify. DA's offices also vary in the extent to which they use victim advocates to provide information about the status of the case, to encourage victims to testify, or to link victims to support services. DA's offices implement these policies and practices in a variety of contexts. Some have specialized prosecution units, some prosecute their cases in specialized DV courts, and some operate in jurisdictions with a coordinated community response.

Because so many policies, practices, and contexts exist for prosecuting DV cases, characterizing the approach taken by any particular jurisdiction is difficult. Comparing jurisdictions also can be problematic because many features of each jurisdiction can differ. Interpreting studies that compare prosecutorial policies requires a detailed understanding

Direct correspondence to Richard R. Peterson, New York City Criminal Justice Agency, 52 Duane Street, New York, NY 10007 (e-mail: rpeterson@nycja.org).

DOI:10.1111/1745-9133.12054

not only of the policies being compared but also of numerous other policies, practices, and contexts. Applying labels like "evidence-based" and "victim-centered" might be necessary and appropriate to describe and compare jurisdictions. However, we should view these labels as summary descriptions of particular features of a policy or practice. Although we might describe a particular policy as "evidence-based" or "victim-centered," these terms might not describe accurately the broader approach taken by the jurisdiction. For example, jurisdictions that use "evidence-based" prosecution also might implement a broad array of policies and practices that are "victim centered."

Comparing Prosecutorial Policies in Metro Atlanta DV Cases

Finn (2013, this issue) conducted research comparing a jurisdiction with "victim-centered" prosecution policies and practices with a jurisdiction with "evidence-based" prosecution policies and practices. As Finn notes, the terminology for describing these prosecutorial policies varies across studies, and the same term might have different meanings in different studies. This increases the importance of focusing on the description of the policies and practices to understand what we can learn from each study. Finn describes the policies and practices used in two counties in the metropolitan area of Atlanta, Georgia. Both counties had a specialized unit for DV cases; each unit was staffed by victim/witness advocates and a DV investigator.

The specialized prosecution unit in metro county A had three prosecutors, one of whom served as the unit supervisor. Prosecutors filed charges in all legally sufficient cases whether or not the victim wished the case to go forward. Once filed, prosecutors minimized victim involvement in decisions about how to handle the case. They viewed victims primarily as witnesses and sometimes issued subpoenas to compel their appearance. Finn does not say how often prosecutors issued subpoenas nor how aggressively they pursued this strategy.[1] Prosecutors used independent evidence (e.g., medical reports, 911 tapes, photos, etc.) to build a case if the victim did not participate in the prosecution. Although they did not dismiss or nolle prose cases simply because victims were not participating, prosecutors did consult victims when considering whether to reduce charges to a lesser offense. The victim services unit provided victims with information about their rights and the progress of the case, and they encouraged victim participation with the case. The unit contacted victims within 24 hours of receiving a police report to attempt to schedule an interview.

Metro county B's specialized prosecution unit used a two-step screening process for DV cases. When a defendant was detained and expressed interest in a plea, then a prosecutor screened the case. (Based on the description provided, it seems that prosecutors were not part of the specialized unit but were brought in to handle cases when necessary.) The

1. Victim/witnesses who failed to appear "would be treated like any other witness who failed to appear" (Finn, 2013), but we do not know what that treatment was. Prosecutors vary considerably in how they handle subpoenaed witnesses.

prosecutor decided how to proceed with these cases based on a risk assessment and, when possible, on consultation with the victim. However, it is not clear how often the prosecutor consulted victims or what influence victims had on these decisions. The unit supervisor, described as the "victim-witness coordinator" (apparently not a prosecutor), screened all remaining cases. In this second step, the supervisor identified cases appropriate for diversion to a 3-month treatment program while criminal prosecution was suspended. Finn (2013) does not indicate whether victims had any input into the diversion decision. The supervisor referred high-risk cases and other cases not suitable for diversion to a prosecutor for possible action in state court. If the victim wished to pursue the case, then the prosecutor filed charges; however, if the victim did not want to move forward and the prosecutor agreed, then the case was dropped. Although it is not clear how often the prosecutor acceded to the victim's wishes, it seems that victims had a strong influence on the decision.

In metro B, victim/witness advocates called all victims within 1 to 5 days of the incident, and a DV investigator visited victims not reached by phone. The advocates also mailed to victims a letter and brochure with information about the case and about local DV resources. When the unit supervisor diverted cases, two service provider/advocacy groups met with victims to explain court procedures and to provide information about accessing counseling, shelter, and other services.

Based on these differences, Finn (2013) describes metro A as having an evidence-based prosecutorial policy and metro B as having a victim-centered prosecutorial policy. Although there were some decisions in which victim influence seemed inconsistent with these characterizations or information about victim influence was not provided, these labels generally seem appropriate. They reflect primarily each jurisdiction's approach to deciding whether to file charges in court after a DV arrest and to deciding whether to drop charges after they were filed. In the evidence-based jurisdiction, prosecutors based these decisions on the sufficiency and strength of evidence; if the victim did not wish to participate in the prosecution, then the case went forward with other evidence. In the victim-centered jurisdiction, with some exceptions, victims had a strong influence on decisions about filing charges and dropping cases. The greater influence of victims in metro B had a significant impact on case processing—27.7% of cases had no action beyond arrest compared with only 6.6% of cases in metro A (Finn, 2013), where cases were prosecuted with or without victim participation.

Finn (2013) compares the impact of these prosecutorial policies on deterrence measures (certainty, swiftness, and severity of disposition), victims' "court empowerment," victims' reports of reoccurrence of violence, and victims' perceptions of safety. Formal dispositions were more certain and more severe under the evidence-based policy than the victim-centered policy, but jail sentences were unexpectedly more common under the victim-centered policy. No difference was found in time to disposition or court empowerment. Multivariate analyses showed that victims' reports of the reoccurrence of both psychological and physical violence in the 6 months after case disposition were significantly higher under the evidence-based

policy than under the victim-based policy. However, although they reported significantly more violence, victims' perception of their risk of future violence (in the next 6 months) was no higher in the evidence-based jurisdiction than in the victim-centered jurisdiction.

Implications for Policy and Practice

Finn's (2013) main conclusion regarding prosecutorial policies is that "victim-centered policies yield better outcomes for domestic violence victims than evidence-based policies." Finn based this conclusion primarily on multivariate analyses showing higher levels of psychological and physical violence after case disposition under the evidence-based policy. Finn suggests also that these findings support efforts to establish specialized DV courts, which address victims' needs while holding defendants accountable. However, Finn acknowledges that both Atlanta jurisdictions use mixed-docket courts to hear DV cases, and that the findings provide only a weak basis for this conclusion.[2] Other research addresses more directly the benefits of specialized DV courts (e.g., Cissner, Labriola, and Rempel, 2013; Labriola, Bradley, O'Sullivan, Rempel, and Moore, 2009).

Regarding Finn's (2013) main conclusion that adoption of victim-centered prosecutorial policies yields better outcomes for victims, I believe the current study does not provide sufficient evidence to characterize them as better than the outcomes under evidence-based policies.

First, the victim's role in prosecutorial decisions might not be the key difference between the evidence-based (metro A) and victim-centered (metro B) jurisdictions in Atlanta. Based on the description of victim outreach practices, it seems that metro B not only gave victims a greater role in the charging decision but also invested more time and resources in connecting victims to counseling, shelter, and other support services. If the victim services unit in metro A had made a similar effort, rather than narrowly focusing on encouraging victim participation in the prosecution, then victim outcomes might have been similar to those in metro A. As will be discussed, some jurisdictions with evidence-based policies use strong victim-engagement strategies.

Second, jurisdictions vary considerably in how they implement particular prosecutorial policies. This might explain why Finn's (2013) findings are inconsistent with two prior studies that found no impact of victim-centered versus evidence-based prosecutorial policies on postdisposition recidivism in DV cases, and another study that found only conditional effects. Peterson (2003) found no significant difference between postdisposition DV rearrest rates in the Bronx, which used a victim-centered filing policy, and in Brooklyn, which used an evidence-based filing policy. Davis, O'Sullivan, Farole, and Rempel (2008) compared postdisposition rearrests in Bronx cases where the victim did not file charges with prosecuted Brooklyn cases where the victim did not participate with the prosecution. They

2. Metro B has a specialized DV calendar for the 15% of cases diverted for treatment. However, the study did not examine the impact of this DV calendar on victim outcomes.

found no difference in the rearrest rates for assault, menacing, or harassment (a proxy for DV recidivism), and they concluded that their study "did not produce a definitive picture of which prosecution policy is superior" (Davis et al., 2008: 658). Ford and Regoli (1993) used an experimental design to assign DV cases in Indianapolis to a "drop-permitted" policy (allowing victims to drop charges) or a "no-drop" policy. Although DV recidivism was lower under the "drop-permitted" policy, this reduction only occurred when the victim chose not to drop charges. When victims dropped charges (which they did in nearly half the "drop-permitted" cases), DV recidivism was *higher* than it was under the "no-drop" policy.

The Atlanta study is the only one to find lower postdisposition DV recidivism in a jurisdiction that prosecutes all DV cases using a victim-centered prosecutorial policy versus a jurisdiction that uses an evidence-based policy. However, the implementation of victim-centered and evidence-based prosecutorial policies varies considerably across jurisdictions and might account for the different findings. Unlike metro A, some DA's offices that use evidence-based policies do not attempt to compel victim participation with the case. While a case is pending in Brooklyn, victims might decide to withdraw or to participate, and prosecutors seldom seek court orders to compel victim testimony (Peterson, 2012).[3] Unlike metro B, some jurisdictions that use victim-centered policies might severely limit the nature and scope of victim input. For example, although the Bronx rarely files charges in misdemeanor DV cases if the victim does not sign the complaint, victims must sign it within the 24-hour time limit for arraigning an arrested defendant. Once a prosecutor files charges, the case remains on the court docket even if the victim requests later that it be dropped (Peterson and Dixon, 2005). This policy is different from the policy tested in Indianapolis, where the drop-permitted policy applied only after charges were filed and did not apply to cases initiated by on-scene arrest. Although victims influenced some decisions in the Bronx and in Indianapolis, the scope of victim involvement was much more limited than in metro B.

Other variations in practice among jurisdictions implementing a particular policy also might affect outcomes. For example, using an evidence-based policy, prosecutors filed charges in 99% of Brooklyn DV cases (Peterson and Dixon, 2005), 69% of Milwaukee cases (Davis, Smith, and Taylor, 2003), and 92% of metro A cases (Finn, 2013). Using a victim-centered policy, prosecutors filed charges in 83% of Bronx cases (Peterson and Dixon, 2005), 30% of Milwaukee cases (Davis et al., 2003), and 72% of metro B cases (Finn, 2013).[4] Under its victim-centered policy, the Bronx filed charges more often than Milwaukee did under its evidence-based policy. These variations in charging practices might

3. When victims want to testify but fear retaliation, Brooklyn prosecutors might issue a subpoena as a strategy to convince the defendant that the testimony was not voluntary.

4. To calculate the Atlanta percentages, nolle prose cases, cases with no accusation filed, and cases diverted for treatment were counted as not charged.

affect case and victim outcomes. If so, then it is difficult to draw broad conclusions about *all* victim-centered or *all* evidence-based prosecutorial policies.

Taking a Broader Approach to Victim Empowerment

Because prosecutors can implement victim-centered and evidenced-based policies in various ways, we should focus on the implications of Finn's (2013) study for victim engagement more broadly, rather than on the narrower question of which of the two prosecutorial policies examined is "better." Although recidivism in Atlanta was lower under victim-centered policies, this outcome was the result of the full range of policies and practices as implemented in metro B versus metro A. The more extensive victim outreach efforts in metro B might be sufficient to produce the lower levels of recidivism even without the victim-centered prosecutorial policy.

Victim engagement, rather than victim-centered prosecutorial policies, might be the key to successful outcomes. Finn (2013) emphasizes the importance of victim engagement throughout the article, and I believe this is the most important policy implication based on the study's findings. Although Finn argues that victim-centered prosecution policies are better, Finn also recommends that jurisdictions considering adopting evidence-based policies should "make efforts to encourage, educate, and support victims throughout the court process." The available research suggests a variety of strategies that successfully engage victims.

Many communities have Family Justice Centers, which locate prosecutors, police, civil legal advocates, and community-based advocates in one place to provide comprehensive services to victims of domestic violence (EMT Associates, 2013; Peterson, 2013a). Family Justice Centers empower victims to respond to abuse by providing victims access to a variety of services customized to their needs and by allowing all victims to use these services, even if they are not participating in the prosecution.

Engaging victims early in the process also can be an important empowerment tool. For example, the Early Victim Engagement (EVE) Project in Brooklyn contacts victims of intimate partner violence by telephone immediately after a defendant is arraigned in Criminal Court. EVE reaches 80% of victims and provides them with information about the case, the defendant's release status, the order of protection and how to enforce it, safety planning, and the availability of services at the Family Justice Center (Peterson, 2013a, 2013b). EVE staff members ask victims to come in to the District Attorney's office for an intake appointment. The intake appointment is used to gather and provide information about the criminal case and to connect victims to services available at the Family Justice Center on the same floor.

The Triage Project in Denver also uses victim advocates to contact victims shortly after arrest to assess their needs and to make appropriate service referrals (DePrince, Belknap, Labus, Buckingham, and Gover, 2012). A multidisciplinary review team, composed of law enforcement, criminal justice, and community-based agencies, assesses the most pressing

needs and designates an appropriate agency to contact the victim and offer services. Trained victim advocates from the designated agencies reach approximately 77% of victims and offer appropriate referrals to their own and other agencies.

These victim-engagement programs can have powerful effects on victim empowerment, providing victims with options for meeting their emotional, financial, child care, and housing needs. If we define victim empowerment to include a wide range of victim goals and needs, not just involvement in prosecution decisions, then broadly conceived victim outreach programs might be more important than policies empowering victims to make prosecution decisions. Whether the criminal case is prosecuted might not be a victim's most important concern. Using the criminal case as a way to link victims to services might improve victim outcomes regardless of the case outcome. In addition, there might be an ancillary impact on criminal justice outcomes. Both the EVE Project and the Triage Project increased victim participation in the prosecution, and the EVE Project also increased the conviction rate in cases of intimate partner violence (DePrince et al., 2012; Peterson, 2013a, 2013b).

I would caution practitioners not to conclude that a victim-centered prosecutorial policy (allowing victims to influence decisions about filing charges or dropping them after they are filed) is necessarily better than an evidence-based policy. That conclusion might apply to the narrow comparison of metro A and metro B as they implemented these policies, but it is not necessarily applicable to other settings. Some jurisdictions implement victim-centered prosecution policies in ways that might reduce or eliminate their positive impacts on victim outcomes. Similarly, some jurisdictions implement evidence-based prosecution policies in ways that encourage victim engagement and might enhance positive outcomes for victims. Taking a broad approach to victim engagement is likely to be more important than deciding how much influence victims should have on prosecutorial decisions. As Finn (2013) notes, prosecutors and advocates have made several arguments for and against victim-centered and evidence-based prosecutorial policies. If metro A and other jurisdictions do not want to drop their evidence-based prosecution policies, then other strategies are available to engage victims. Finn's study provides valuable support for the argument that all District Attorneys, whether they use a victim-centered or evidence-based prosecutorial policy, should consider using a broad, victim-centered approach that empowers victims to address their needs.

References

Cissner, Amanda, Melissa Labriola, and Michael Rempel. 2013. *Testing the Effects of New York's Domestic Violence Courts.* New York: Center for Court Innovation.

Davis, Robert C., Chris S. O'Sullivan, Donald J. Farole, Jr., and Michael Rempel. 2008. A comparison of two prosecution policies in cases of intimate partner violence: Mandatory case filing vs. following the victim's lead. *Criminology & Public Policy*, 7: 633–662.

Davis, Robert C., Barbara E. Smith, and Bruce Taylor. 2003. Increasing the proportion of domestic violence arrests that are prosecuted: A natural experiment in Milwaukee. *Criminology & Public Policy*, 2: 263–282.

DePrince, Anne P., Joanne Belknap, Jennifer S. Labus, Susan E. Buckingham, and Angela R. Gover. 2012. The impact of victim-focused outreach on criminal legal system outcomes following police-reported intimate partner abuse. *Violence Against Women*, 18: 861–881.

EMT Associates. 2013. *Final Evaluation Results: Phase II California Family Justice Initiative Statewide Evaluation*. Burbank, CA: Author.

Finn, Mary A. 2013. Evidence-based and victim-centered prosecutorial policies: Examination of deterrent and therapeutic jurisprudence effects on domestic violence. *Criminology & Public Policy*, 12: 443–472.

Ford, David A. and Mary Jean Regoli. 1993. The criminal prosecution of wife assaulters. In (N. Zoe Hilton, ed.), *Legal Responses to Wife Assault: Current Trends and Evaluation*. Newbury Park, CA: Sage.

Labriola, Melissa, Sarah Bradley, Chris S. O'Sullivan, Michael Rempel, and Samantha Moore. 2009. *A National Portrait of Domestic Violence Courts*. New York: Center for Court Innovation.

Peterson, Richard R. 2003. *The Impact of Case Processing on Re-Arrests Among Domestic Violence Offenders in New York City*. New York: New York City Criminal Justice Agency, Inc.

Peterson, Richard R. 2012. *The Kings County District Attorney's Video Statement Program for Domestic Violence Cases*. New York: New York City Criminal Justice Agency, Inc.

Peterson, Richard R. 2013a. *Early Victim Engagement in Domestic Violence Cases*. New York: New York City Criminal Justice Agency, Inc.

Peterson, Richard R. 2013b. *The EVE Project*. New York: New York City Criminal Justice Agency, Inc.

Peterson, Richard R. and Jo Dixon. 2005. Court oversight and conviction under mandatory and nonmandatory domestic violence case filing policies. *Criminology & Public Policy*, 4: 535–558.

Richard R. Peterson is the director of research for the New York City Criminal Justice Agency. His current research interests include prosecutorial strategies in domestic violence cases, specialized domestic violence courts, prevention of intimate partner homicides, and pretrial misconduct. He has published articles in *Criminology & Public Policy*, *American Sociological Review*, *Journal of Marriage and the Family,* and *Social Forces.* He is also the author of *Women, Work and Divorce* (SUNY Press, 1989).

VICTIM-CENTERED PROSECUTORIAL POLICIES

The Importance of Prosecution Policies in Domestic Violence Cases

Krista R. Flannigan
Florida State University

Traditionally, police and prosecutors have viewed battering as a private, family problem not appropriate for government intervention (Buzawa and Buzawa, 1996). The criminal justice system seemed to shield abusive spouses from the public in the belief that the parties should be left to work out their "differences" privately. If any intervention was deemed appropriate, then counseling was preferred over prosecution (McGuire, 1999).

In the 1970s, the women's movement and resulting political pressure led to the creation of more aggressive domestic violence laws such as mandatory arrest policies and hard line investigation (Davis and Smith, 1995). By the 1980s, prosecutors in some jurisdictions had initiated special programs for domestic violence cases (McGuire, 1999). In the decade after the change in policy, the U.S. Department of Justice reported that the rate of violence by intimate partners fell 42–49% (Bureau of Justice Statistics, 2013). These figures could be read to suggest that systemic changes in the way the criminal justice system views domestic violence have had a positive impact on this serious social problem (Fulkerson and Patterson, 2006).

Understanding of the impact of domestic violence on individual victims, families, and communities has expanded greatly over the past 20 years (Hamel, 2013). A plethora of research has been conducted on a variety of domestic violence issues, including physical abuse; victimization; perpetration, risk factors, and risk assessment; emotional abuse and control; abuse in ethnic minority and lesbian, gay, bisexual, and transgender (LGBT) populations; impact of parental violence and conflict on children; impact of abuse on partners; motives for abuse perpetration; retraining orders; prevention; victim services and perpetrator treatment; the criminal justice response in the context of gender and ethnicity; and the effectiveness of criminal justice sanctions (Hamel, 2013). Research results have

Direct correspondence to Krista R. Flannigan, Florida State University, College of Criminology and Criminal Justice, 145 Convocation Way, Tallahassee, FL 32306 (e-mail: kflannigan@fsu.edu).

DOI:10.1111/1745-9133.12058

provided guidance in developing intervention and prevention policies and practices such as mandatory arrest and no-drop prosecution policies.

Research to date regarding the effectiveness of criminal justice system intervention and sanctions has been reflected in recidivism rates and has not addressed other outcomes or possible underlying causes. Furthermore, the research has not collected information that distinguishes victim empowerment or other potential mechanisms that might affect future behavior (Maxwell, Garner, and Fagan, 2001). Nor has the research studied prosecution and conviction as it affects domestic violence. Therefore, Finn's (2013, this issue) research of prosecutorial polices and domestic violence is the beginning of study on a significant gap in domestic violence study.

In this policy essay, I will assess Finn's (2013) findings and in the conclusion will provide actionable recommendations to policy makers and practitioners about how to study and create more effective domestic violence prosecution and comprehensive victim services for all victims of crime.

Evidence-Based and Victim-Centered Prosecutorial Policies

Finn (2013) studied court jurisdictions that employed the two different prosecution strategies. The purpose was to assess and compare levels of court empowerment, reoccurrence of physical violence and psychological aggression, length of case processing, dispositions and sanctions, and perception of safety reported by victims in each jurisdiction. Metro county A practiced evidence-based prosecution, and metro county B used a victim-centered approach.

Evidence-Based Prosecution

Evidence-based prosecution (sometimes termed "victimless prosecution") refers to a collection of techniques used by prosecutors in domestic violence cases to convict abusers without victim participation (Viswantathan, 2003). The practice of evidence-based prosecution was developed in response to the fact that battered women often cannot or are unwilling to cooperate with the prosecution of their batterer (Fulkerson and Patterson, 2006). This type of prosecution is called evidence based because it relies on physical evidence and on the testimony of third parties to support the charges against the defendant (Fulkerson and Patterson, 2006). Victimless prosecution typically works in conjunction with "no-drop" policies whereby prosecutors refuse to dismiss domestic violence cases at the request of the victim (Jaros, 2005). Arguments in favor of this approach includes society's interest in ending abusive relationships (Jones, 2000), safety needs of the victims (Mills, 1999), and victim empowerment through participation in the process of the prosecution (Robbins, 1999).

In 2004, the U.S. Supreme Court ruled in *Crawford v. Washington* that the admission of certain statements by victims into evidence violates the 6th Amendment Confrontation Clause unless the defendant has the opportunity to cross-examine the victim. This ruling

limited evidence-based prosecution in that if a victim is not participating, then his or her statements to law enforcement during the investigation are prohibited as evidence unless the victim testifies.

Although *Crawford* restricts the evidence that can be used in victimless prosecution, the case has not eliminated all tools available in the prosecution of domestic violence cases (Fulkerson and Patterson, 2006). Medical evidence is still permitted and other court rulings have allowed statements from 911 calls (*Washington v. Davis,* 2005) and some crime scene statements (*Hammon v. Indiana,* 2005). These rulings promote the use of reliable evidence without forcing the victim to testify and protect the constitutional rights of the accused to confront witnesses (Fulkerson and Patterson, 2006).

A common criticism of the evidence-based model is that it erodes what little self-esteem and control a victim might believe they have (Han, 2003). Domestic violence is a crime of power and control of the perpetrator over the victim (Buel, 1999). It is argued that evidence-based prosecution only transfers power from one controlling entity to another (Mills, 1999). Such policies might even further victimize the women if enforcement of the mandatory prosecution leads to forced testimony from victims. For example, when victims are subpoenaed to testify, it might seem that they are being punished, even though most likely inadvertently. They also have little incentive to report future violence to law enforcement or cooperate with prosecutors (Buel, 1999).

Victim-Centered Prosecution

Victim-centered prosecution, as described by Finn (2013), is when prosecutors seek input from victims regarding the handling of a criminal case. This practice is commonly referred to as a victim empowerment approach to prosecution (Kanter and Enos, 2001).

The empowerment model recognizes that some choices presented in domestic violence cases are simply too important and too difficult to be made by a third party (Kanter and Enos, 2001). Legal options cannot guarantee, and in some instances might actually jeopardize, safety. For example, women are often in the most danger when they leave or attempt to leave their abuser (Adams, 1989). Statistics show that a woman is at 75% greater risk of being killed when she leaves her abuser (Barnes, 2006). The victim must be informatively and supportively empowered to make decisions for herself (Kanter and Enos, 2001).

Victim empowerment prosecutions follow the client-centered models of a lawyer–client relationship where the client makes the decisions regarding their case (Binder, Bergman, and Price, 1991). Critics of this approach will argue that victims are not prosecutors' clients. The National Prosecution Standards (1–1.2, 3rd Edition) clearly states a prosecutor's duty of the public over the individual interest:

> A prosecutor should zealously protect the rights of individuals, but without representing any individual as a client. A prosecutor should put the rights and

> interests of society in a paramount position in exercising prosecutorial discretion in individual cases . . . Societal interests rather than individual or group interests should also be paramount in a prosecutor's efforts to seek reform of criminal laws. (National District Attorney's Association, n.d.)

However, as discussed later in this essay, prosecutors can represent society's greater interests while seeking input from victims on individual cases.

Prosecutorial Polices and Domestic Violence Study Results

Finn's (2013) results showed no statistical difference between metro county A (evidence-based practice) and metro county B (victim-empowerment model) related to the length of case processing, sanctions and dispositions, levels of court empowerment, and victim perception of safety prior to disposition. The results did indicate, however, that cases in the evidence-based policy jurisdiction, compared with the victim-centered policy jurisdiction, were significantly more likely to report reoccurrence of physical violence and psychological aggression. Understandably, victims who experienced physical violence during the 6 months after case disposition perceived themselves as less safe. However, none of the elements of deterrence (swiftness, certainty, or severity of punishment) or court empowerment significantly predicted the reoccurrence of psychological aggression or physical violence.

Finn (2013) expected that victim-centered policies would enhance court empowerment and that evidence-based policies would not. But the results found no differences in court empowerment after case disposition by type of prosecutorial policy. However, Finn notes that the measure of court empowerment employed might have been inadequate and further evaluation could be done using other tools that have been developed since completion of this study.

Existing Guidelines for Domestic Violence Prosecution

The Guidelines for Prosecution of Domestic Violence Cases were established in 2001 through a report funded from the Violence Against Women Act. Although the guidelines strongly encourage aggressive prosecution, they also caution prosecutors that domestic violence perpetrators seek to control victims without regard to the victims' welfare and victim safety is a priority. With this in mind, the pro-prosecution policy makes it clear to the perpetrators that the prosecutors, not the victims, are responsible for decisions regarding criminal prosecution. "By relying primarily on the evidence collected by law enforcement rather than solely on the victim's testimony, the prosecutor may be able to reduce the risk of retaliation by the perpetrator against the victim and increase the likelihood of a successful prosecution" (Alabama Coalition against Domestic Violence, 2004: 6).

Among these guidelines is the recommendation of "vertical prosecution." Vertical prosecution is a management policy that designates the same specialized prosecutors and

victim–witness staff to handle all aspects of a domestic violence case. Vertical prosecution increases communication between the victim and the prosecution office thereby enhancing the ability of the prosecutor to address victim concerns effectively. In addition, the recommendations include specialized training for domestic violence prosecutors. Specialization often results in higher conviction rates of domestic violence offenses. Domestic violence cases could be frustrating for prosecutors who do not understand the dynamics of domestic violence and the reasons for the victim's reluctance to participate in criminal prosecution. Specialization and training will enhance the ability of prosecutors to understand and therefore safely respond to victim behavior (Prosecution Guidelines, 2001).

Vertical prosecution is a tool that is used in jurisdictions with specialized domestic violence units, which is also a recommendation in the Guidelines (Prosecution Guidelines, 2001). The challenge of implementation, however, is resources. To create such a unit requires specially trained and dedicated prosecutors. Unfortunately, many offices do not have the luxury of committing lawyers to a specific type of prosecution.

It seems that the evidence-based prosecutors' offices studied by Finn (2013) did not adhere to all components of the established guidelines but took a more restrictive, almost victim-adversarial approach. One cannot help but wonder if they had followed more true to the guidelines, then would the outcomes have been different?

Furthermore, the American Bar Association Criminal Justice Standards (3–3.2 (h)) states:

> Where practical, the prosecutor should seek to insure that victims of serious crimes or their representatives are given an opportunity to consult with and to provide information to the prosecutor prior to the decision whether or not to prosecute, to pursue a disposition by plea, or to dismiss the charges.

Victim Rights Amendment

Thirty-three states have enacted a crime victims' rights amendment (VRA). With variations among the jurisdictions, they all include mandatory rights to be "heard" at certain stages of the criminal justice process and to be informed of when proceedings will occur (NCVLI, 2011). Victim status is commonly determined by those who are harmed by another criminal's conduct in certain types of offenses, which include domestic violence or assault. Usually, states follow federal statutes defining victims including the Crime Victims' Rights Act (2004), Mandatory Victim Restitution Act (1996), and Victim and Witness Protection Act (1982). The right to be heard refers to the right to make an oral or written statement to the court at a criminal justice proceeding. Most statutory and constitutional rights to be heard are drafted in mandatory terms, leaving judges no discretion whether to allow crime victims to make a statement at sentencing. Depending on the jurisdiction, victims have the right to be heard at bond release, plea, sentencing, and parole. Focusing on the critical stages of plea and sentence, at least 12 states provide for the right to be heard by the court

prior to the acceptance of any proposed plea agreement,[1] and 33 states provide for the right to be heard by the prosecutor prior to the presentation of the plea agreement to the court.[2] A few states provide for the victim to be heard by both the prosecutor and the court prior to acceptance of a plea agreement.[3] At least 39 states provide crime victims with a constitutional or statutory right to be heard at sentencing.[4] Whereas the right to be heard at plea and sentencing often are written as separate statutes, the right to be heard at sentencing "implicitly includes the right to be heard at plea because the right to be heard at sentencing may only be meaningful if exercised prior to a defendant's plea" (Beloof, 2003).

The VRA's mandate requires most prosecutors in the country to consider the opinion of victims in the disposition of their cases. Although prosecutors are not required to abide by the victims' desires for outcome, their opinions are to be at least considered. In some jurisdictions, prosecutors must inform the court of their consultation with the victim and the victim's agreement or lack thereof.

The VRA, in addition to the prosecution guidelines for domestic violence cases, proposes a "morphing" of the evidence-based and victim-centered models discussed by Finn

1. See Ariz. Const. art. 2, § 2.1(A) (6); or Const. art. I, § 42(1)(f); S.C. Const. art. I, § 4(A)(7); Ala. Code § 15–23–64; Ark. Code Ann. § 16–21–106(b); Colo. Rev. Stat. § 24–4.1–302.5(e); Del. Code Ann. tit. 11, § 9405; Fla. Stat. § 960.001(g); Ga. Code.

2. See Ariz. Const. art. 2, § 2.1(A) (6); Or. Const. art. I, § 42(1)(f); S.C. Const. art. I, § 4(A)(7); Ala. Code § 15–23–64; Ark. Code Ann. § 16–21–106(b); Colo. Rev. Stat. § 24–4.1–302.5(e); Del. Code Ann. tit. 11, § 9405; Fla. Stat. § 960.001(g); Ga. Code Ann. § 17–17–11; Haw. Rev. Stat. § 801D-4(a)(1); Ind. Code Ann. § 35–40–3(b)(3); Ky. Rev. Stat. Ann. § 421.500(6); Me. Rev. Stat. Ann. tit. 17-A, § 1173; Mich. Stat. Ann. § 780.756(3); Miss. Code Ann. §§99–43–11, -27; Mo. Rev. Stat. § 595.209(4); Mont. Code Ann. § 46–24–104(3); Neb. Rev. Stat. § 29–120; N.H. Rev. Stat. Ann. § 21-M:8-k(II)(f); N.J. Stat. Ann. § 52:4B-44(b)(2); N.Y. Exec. Law § 642(1); N.C. Gen. Stat. § 15A-832(f); N.D. Cent. Code §12.1–34–02(13); Ohio Rev. Code § 2930.06(A); Pa. Const. Stat. §§ 11.201(4), 11.213(b); S.D. Codified Laws § 23A-28C-1(5) (limited to written input); Tenn. Code Ann. § 40–38–114(a); Tex. Code Crim. Proc. Ann. art. 56.02(a)(13); Utah Code Ann. § 77–38–2(5)(d); Vt. Stat. Ann. tit.13, § 5321(e); Va. Code Ann. § 19.2–11.01(4)(d); W. Va. Code § 61–11A-6(5); Wis. Stat. § 971.095(2).

3. See Ariz. Const. art. 2, §§ 2.1(A)(4), (6); S.C. Const. art. I, § 24(A)(5), (7); Colo. Rev. Stat. § 24–4.1–302.5(d), (e); Me. Rev. Stat. Ann. tit. 17-A, §1173; Mo. Rev. Stat. § 595.209(4); Tex. Code Crim. Proc. Ann. art. § 56.02(a)(13); Utah Code Ann. §§77–38–4(1), 77–38–2(5)(d).

4. See Ala. Const. amend. 557, Ala. Code § 15–23–74; Alaska Const. art. 2, § 24; Ariz. Const. art. 2, § 2.1(A)(4); Cal. Penal Code § 679.02(a)(3); Colo. Const. art. II, § 16a, Colo. Rev. Stat. § 24–4.1–302.5(10(g); Conn. Const. art. 1, §(8)(b)(8); Fla. Const. art. I, § 16, Fla. Stat. § 960.01(1)(k); Idaho Const. art. 1, § 22(6); Ill. Const. art. 1, § 8.1(a)(4); Ind. Code Ann. § 35–40–5–5; Iowa Code § 915.21(1)(b); Kan. Const. art. 15, § 15(a); La. Const. art. I,§ 25, La. Rev. Stat. Ann. § 1842(2); Me. Rev. Stat. Ann. tit. 17-A, § 1174(1)(A); Md. Const. Decl. of Rights, art. 7(b), Md. Code Ann., Crim. Proc. § 11–403; Mass. Gen. Laws ch. 258B, § 3(p); Mich. Const. art. I, § 24(1); Minn. Stat. § 611A.038(a); Miss. Const. art. 3, § 26A(1), Miss. Code Ann. § 99–43–33; Mo. Const. art. I, § 32(1)(2); Neb. Const. art. I, §28(1); Nev. Const. art. 1, § 8(2)(c); N.H. Rev. Stat. Ann. § 21-M:8-k(I)(p); N.J. Stat. Ann. § 52:4B-36(n); N.M. Const. art. II, § 24(A)(7); N.C. Const. art. 1, §37(1)(b); Ohio Rev. Code § 2930.14(A); Okla. Const. art. II, § 34(A); Pa. Const. Stat. § 11.201(5); R.I. Gen. Laws § 12–28–3(11); S.C. Const. art. I, § 24(A)(5); S.D. Codified Laws § 23A-28C-1(8); Utah Const. art. I, § 28(1)(b), Utah Code Ann. § 77–38–4(1); Vt. Stat. Ann. tit. 13, § 5321(a)(2); Va. Const. art. I, § 8-A(3); Wash. Const. art. 2, § 35; Wis. Const. art. I, § 9(m); Wyo. Stat. Ann. § 14–6–502(a)(xvii).

(2013). Little research has examined their success in recidivism or victim empowerment, but that should certainly be the next step.

Victim Centered—Another View

Victim-centered prosecution most often is used in sexual assault and human trafficking cases and is not mutually exclusive of evidence-based prosecution. Using the victim-centered approach, prosecutors, although not victims' attorneys, can advocate for victims' rights and proactively address victims' concerns. Prosecutors can assist, or empower, victims to overcome the myriad of their common concerns of many victims. For example, prosecutors can help victims by orienting them to the criminal justice system, providing waiting areas that are separate from offenders, and working with advocates to help meet victims' emotional needs. Prosecutors also can seek no-contact orders as conditions of bail or release of offenders on their own recognizance; pursue defendants who harass, threaten, or intimidate victims; work with civil attorneys to assist victims with landlord, employer, educator, and creditor issues when needed; incorporate victims' views in bail arguments, continuances, plea negotiations, dismissals, sentencing, and restitution; arrange prompt return of victims' property when it is no longer needed as evidence; and keep the same prosecutor throughout the criminal justice process (vertical prosecution) (Office for Violence Against Women, 2012). Much of this approach follows the pro-prosecution guidelines established in 2001.

Conclusions

Some jurisdictions have been successful at merging the evidence-based and victim-empowerment model practices, whereby the cases are given priority and victims are informed and consulted, but the ultimate decision to move forward with a prosecution is the onus of the prosecutor. The approach is victim centered while deferring to the expertise of the prosecutor.

In addition, some jurisdictions are allowing for crime victims to have an attorney represent them as individuals to assure their rights are being upheld in the criminal justice system. This approach follows more literally the client-centered approach as discussed by Finn (2013) and is worthy of research on its effectiveness.

Finally, in states that have a Victim Rights Amendment, the victim's voice in the process and the outcome is mandated. VRAs practically require that prosecution be victim centered. Perhaps if other states adopted constitutional amendments with similar language, then victim empowerment, at least in the judicial process, will be implicit in the prosecution of cases.

Although Finn's (2013) study is a start in filling the research gap of measuring victim empowerment, it does not take into account the more common prosecution approach of promoting victim input while staying true to the role of the prosecutor as a representative of the state as a whole. In addition, as with much of the research regarding criminal justice response, the study isolates one part of the system and does not evaluate the whole system

response. Isolating one intervention and attempting to link it to future battering or victim empowerment fails to take into account that the observed results might have been caused not by the studied intervention but by the rest of the system's actions or lack of actions.

Recommendations to advance the research and policy development as it relates to the empowerment of victims of domestic violence and the accountability of perpetrators are as follows:

- Assess the effectiveness of the victim-centered alternative view as discussed and explore whether it might be useful in domestic violence cases.
- Evaluate the empowerment of victims on a broad scope in the criminal justice system.
- Review the prosecution policies to determine whether they remain relevant post-*Crawford* and whether they can be enhanced to facilitate further victim safety and autonomy.

Beyond Domestic Violence Prosecution

Victim services have evolved over the last 30 years in (and out of) the criminal justice system. One could argue that overall services have improved—there are certainly more of them. However, providing more services to more victims does not mean services are useful or even needed. Crime victimization also is changing; for example, cyber-crimes and incidents of mass violence. As a result, the body of knowledge about victimization needs to be expanded (Office for Victims of Crime, 2012). The issues regarding the impact of victimization and the type of support and intervention needed are different than they were 30 years ago.

Furthermore, although quantitative data are collected regarding types of services offered, categories of victims served, and to some extent the use and nonuse of services, no empirical research exists regarding policy influence and service quality. Providers, funders, and policy makers are operating merely in good faith that services provided and laws enacted to protect victims and their rights are appropriate, useful, and effective. With resources becoming more and more restricted, it is incumbent on those in the field to ensure that the laws passed and the services offered are of value, are beneficial, and truly are victim centered.

References

Adams, David M. (ed.). 1989. Identifying the assaultive husband in court: You be the judge. *Boston Bar Journal*, July/August: 23, 24.

Alabama Coalition against Domestic Violence. 2004. *Guidelines for the Prosecution of Domestic Violence Cases*. Montgomery: Author.

American Bar Association. 2013. Criminal Justice Standards, Prosecution Standards, 3–3.2(h).

Barnes, Kirsten J. 2006. *Domestic Violence: Women's Risk Increase After Breakup*. Montgomery, AL: Montgomery Advertiser.

Beloof, Douglas E. 2003. Constitutional implications of crime victims as participants. *Cornell Law Review*, 88: 282, 286, 290.

Binder, David A., Paul Bergman, and Susan C. Price. 1991. *Lawyers as Counselors: A Client Centered Approach*. St. Paul, MN: West.

Buel, Sarah M. 1999. Domestic violence and the law: an impassioned exploration for family peace. *Family Law Quarterly*, 33: 719–744.

Bureau of Justice Statistics. 2013. *Intimate Partner Violence in the U.S.: Victim Characteristics*. Retrieved September 19, 2013 from bjs.gov/content/intimate/victims.cfm.

Buzawa, Eve S. and Carl G. Buzawa. 1996. *Domestic Violence: The Criminal Justice Response*, 2nd Edition. Thousand Oaks, CA: Sage.

Davis, Robert C. and Barbara Smith. 1995. Domestic violence reforms: Empty promises or fulfilled expectations? *Crime & Delinquency* 41: 541–552.

Finn, Mary A. 2013. Evidence-based and victim-centered prosecutorial policies: Examination of deterrent and therapeutic jurisprudence effects on domestic violence. *Criminology & Public Policy*, 12: 443–472.

Fulkerson, Andrew and Shelley L. Patterson. 2006. Victimless prosecution of domestic violence in the wake of *Crawford v. Washington*. Urbana, IL: Forum on Public Policy.

Hamel, John. 2013. *Partner Abuse State of Knowledge Project, Journal Partner Abuse*. New York: Springer Publishing.

Han, Erin L. 2003. Mandatory arrest and no-drop policies: Victim empowerment in domestic violence cases. *Boston College Third World Law Journal*, 23: 159–191.

Jaros, David. 2005. The lessons of *People v. Moscat*: Confronting judicial bias in domestic violence cases interpreting *Crawford v. Washington. American Criminal Law Review*, 42: 995–1010.

Jones, Ruth. 2000. Guardianship for coercively controlled battered women: Breaking the control of the abuser. *Georgetown Law Journal*, 88: 605, 621–622.

Kanter, Lois H. and V. Pualani Enos. 2001. *Domestic Violence Manual*. Unpublished manuscript. Domestic Violence Institute at Northeastern University School of Law, Boston, MA.

Maxwell, Christopher D., Joel H. Garner, and Jeffrey A. Fagan. 2001. *Effects of Arrest on Intimate Partner Violence: New Evidence from the Spouse Assault Replication Program*. Research in Brief. Washington, DC: National Institute of Justice.

McGuire, Linda A. 1999. *Criminal Prosecution of Domestic Violence*. Minneapolis: Minnesota Center Against Domestic Violence.

Mills, Linda G. 1999. Killing her softly: Intimate abuse and the violence of state intervention. *Harvard Law Review*, 113: 550, 557.

National Crime Victim Law Institute (NCVLA). 2011, November. *Law Bulletin*. Portland, OR: Author.

National District Attorney's Association. n.d. *National Prosecution Standards*, 3rd Edition. Alexandria, VA: Author.

Office for Violence Against Women. 2012. *How to Begin a SART*. Washington, DC: Author.

Office for Victims of Crime. 2012. *Vision 21 Executive Summary*. Washington, DC: Author.

Robbins, Kalyani, 1999. No-drop prosecution of domestic violence: Just good policy, or equal protection mandate? *Stanford Law Review*, 52(1): 205–233

Viswantathan, Hema. 2003. *Evidence Based Prosecution of Domestic Violence: The Significance of New Domestic Violence Hearsay Exceptions*. Minneapolis, MN: The Domestic Violence Project.

Statues Cited

Crime Victims' Rights Act, 18 U.S.C. § 3771 (2004).

Mandatory Victim Restitution Act, 18 U.S.C. §3663 (1996).

Victim and Witness Protection Act, Pub. L. No. 97–291, 96 Stat. 1248 (1982).

Cases Cited

Crawford v. Washington, 541 U.S. 36 (2004).

Hammon v. Indiana, 829 N.E.2d 444 (Ind. 2005), *cert. granted*, 126 S.Ct. 552 (U.S. Oct. 31, 2005) (No. 05–5705).

Washington v. Davis, 154 Wn.2d 291, 111 P.3d 844 (Wash. 2005), cert. Granted 126 S.Ct. 547, 2005 U.S. LEXIS 7859 (October 31, 2005).

Krista R. Flannigan, J.D., is an attorney and advocate experienced in emergency response and management, media relations, community collaboration and program development. She is also a criminal justice and victim issues educator. Krista has trained nationally on coordinated community responses for victims of mass tragedy and high profile trials as well as on the impact of mass tragedy on victims and communities. She also founded a domestic violence advocacy organization and worked as a domestic violence advocate for many years. Krista is currently the Director of the Institute for Crime Victim Research and Policy at Florida State University, College of Criminology. The goal of the Institute is to collaborate with victim services professionals and ultimately provide policy and practice recommendations that are based upon need and the best available research and evaluation evidence. She also teaches classes in the FSU College of Criminology. In addition, Krista provides consultation to the Office for Victims of Crime regarding a variety of crime victim related issues.

VICTIM-CENTERED PROSECUTORIAL POLICIES

Evidence-Based Prosecution

Is it Worth the Cost?

Eve S. Buzawa
University of Massachusetts—Lowell

Aaron D. Buzawa
United States Air Force

We believe that Finn's (2013, this issue) article is theoretically grounded and is an important contribution to the research measuring the effects of prosecution policies on several key aspects of victim reabuse. It seeks to examine the difference in outcomes between a victim-oriented approach to prosecuting domestic violence and an evidence-based prosecution policy. Outcomes were studied between two counties in suburban Atlanta, Georgia. "County A" had adopted an evidence-based/no-drop policy, and "county B" had adopted a policy requiring victim concurrence in prosecution absent other extenuating facts.

Study Findings

Finn (2013) uses sophisticated and methodologically appropriate bivariate and multivariate analyses to report key victim outcomes: self-reports of both *psychological* and *physical aggression* 6 months after the initial disposition. Finn's logistic regression model finds that victims in the evidence-based jurisdiction are significantly (3.76 times) more likely to report repeat psychological aggression. In fact, being in an evidence-based jurisdiction had a greater impact than case processing time or all the numerous control variables used, none of which turned out to be significant.[1]

The views expressed in this article are the authors' own and do not reflect the policy or views of the U.S. Government, or any entity therein, including the U.S. Air Force. Direct correspondence to Eve S. Buzawa, University of Massachusetts, Lowell, School of Criminology & Justice Studies, 150 Wilder Street, Lowell, MA. 01854 (e-mail: Eve_Buzawa@uml.edu).

1. Finn (2013) reports several possible alternative factors that might have accounted for a variance in outcomes including time from arrest to disposition, finding of a guilty plea, a sentence of incarceration,

DOI:10.1111/1745-9133.12056

In the case of *physical or sexual violence*, Finn's (2013) logistic regression model of these same factors finds that victims in the evidence-based jurisdiction are slightly more than seven times (odds ratio 7.17) more likely to report increased violence than in the victim-led county. In common with much of the existing literature, race (victim being Black/African American) was also a key factor, as Blacks/African Americans were almost six times more likely to report repeat physical or sexual violence. However, no other case processing or demographic variables are significant.

A second logistic regression model is used to predict victims' perceptions of future safety. No differences based on the county where the victim resided are found. The only variable of significance is that, understandably, a victim who reported an act of physical or sexual violence during the 6-month follow-up period is a little more than five times more likely (odds ratio 5.07) to express concerns for future safety.[2]

Analysis

Overall, the data presented give significant policy support to the use of a victim-centered approach to prosecution. Most reported studies have assessed the impact of an intervention by focusing on the reduction in official reported rates of re-abuse and whether victims were reassaulted or subjected to psychological abuse during the critical 6-month period after the initial assault. This study, however, measures directly the reoccurrence of abuse as reported by victims, rather than relying on official reports. Data based on actual victim experiences rather than on official statistics are less likely to be impacted by low reporting rates, especially in cases where the victim in the initial incident did not want an arrest and failed to disclose subsequent abuse. Previous research has shown, for example, that approximately half of victims report new offenses occurring within 1 year after the original offense (Buzawa and Hotaling, 2007).

The study's use of two counties in the same state with similar (but not identical) demographic profiles minimized the effects of either statutory difference or major differences between specific demographic variables often cited as key predictors of re-abuse.[3] Overall, the

and elements of therapeutic jurisprudence defined as court empowerment at disposition. Finn uses control variables that include prior arrests of defendant, victim living with abuser after disposition, and race.

2. This finding is interesting in that victims who reported having experienced a threat or act of violence during the pretrial period were significantly less likely to report the recurrence of physical violence 6 months after disposition (odds ratio = 0.310).

3. County A, the county using the more aggressive evidence-based approach, differed in one major factor from county B in that county B had a far higher percent of African Americans in its population (more than 50% vs. 14%) and ultimately in the victims participating in the study. As a result of historic criminal justice agency practices toward minorities, the racial dimension might be significant in exacerbating the negative effects if a victim felt "disempowered" by the system and worse if she was subpoenaed to testify.

data presented give significant policy support to a victim-centered approach to prosecution, with less physical and psychological abuse reported in county B.

We also note that whereas Finn (2013) did not stress this measurement in her conclusions, it seems that county A spent more on domestic violence prosecution cases than county B, both in aggregate and in proportion to cases brought to the system. County A was 12% smaller in population than county B, was more affluent (predictive of fewer overall cases of domestic violence), had far fewer minorities (again predictive of less domestic violence), and had placed more targeted resources on domestic violence cases: three dedicated prosecutors and two victim advocates compared with no dedicated prosecutors but one victim witness coordinator-supervisor and two victim witness advocates. Despite what seems to be far more resources per incident, case disposition took longer in county A than in county B. Time to final disposition has, in addition to higher systemic costs and probable inefficiency, been cited as a predictor of future domestic violence (Buzawa, Hotaling, Klein, and Byrne, 1999; Klein and Tobin, 2008).

We also note that the researchers do not report any significant differences between jurisdictions in a series of questions styled as predicting victim empowerment. We are aware that a diverse range of strategies is available for measuring victim empowerment and that it is important to differentiate empowerment at various stages of the process (Nichols, 2013a, 2013b).

In this case, it is unclear that empowerment as we might preferably define it has been measured. As Finn (2013) notes, the questions asked were related to "process" rather than to "outcome." No direct questions asked victim preferences or whether those preferences were followed. For example, in contrast to the victim-led prosecutions used in county B, in county A, recalcitrant victims would be required to testify if the prosecutor wanted them to do so. If victims refused, then they were subject to subpoena. As Finn notes, "A victim who failed to appear in court after being lawfully served with a subpoena would be treated like any other witness who failed to appear to an authorized and legally served subpoena" (e.g., presumably subject to arrest for contempt of court).

The extent to which actions such as a threat of subpoena, issuance of a subpoena, or contempt of court orders, along with the percent of cases where the victims' wishes to proceed/not proceed were ignored, would seem to be far more direct measures of victim disempowerment rather than the questions used, such as, "The court considered my rights and wishes just as important as my partner's rights and wishes," and "The court treated me fairly and listened to my side of the story." The problems with measurements used are twofold. First, the victim might be satisfied with her treatment at a certain stage, for example, by the "court" if the case was heard by a trained, courteous, and attentive judge, but she might remain very dissatisfied with the charging decisions of the prosecutor. It also is a possibility that the "court" considered both the victim and offender preferences but decided to ignore the preferences of both parties—which we do not believe to be "victim

empowering." Thus, the measure, although styled as "victim empowerment," seems really to be the victim's perceptions of the court's "procedural fairness," which is not the same.

Prosecutorial Trends

During the last several decades, all aspects of society's response to domestic violence have changed dramatically. The importance of addressing victim safety and their other needs is now considered by police in their actions, by legislatures in enacting pro-arrest and mandatory arrest statutes, by courts in trying to ensure batterer accountability and victim safety, and by society in funding numerous victim shelters and support services. It is now generally recognized that domestic violence victims face a greater risk of revictimization by the same perpetrator than do victims of most other criminal offenses.

As a result, police in most jurisdictions no longer habitually ignore such crimes, and more cases are being forwarded for potential prosecution. Prosecutors, in turn, have responded to the organizational challenge of increased domestic violence caseloads and awareness of the need to prosecute in two primary ways: those that prosecute if the evidence supports a conviction, sometimes known as evidence-based prosecution or no-drop policies, and those that will not move forward without a victim's willingness to testify or, at a minimum, support prosecution.

Without a change in policy, high levels of prosecutorial and judicial indifference would have negated the increase in law enforcement arrest policies and unintentionally communicated a lack of societal care for such problems. Most empirical research has shown that simple arrest without any prosecutorial or judicial follow-up has failed to achieve long-term effects on a subset of hard-core offenders (Buzawa, Buzawa, & Stark, 2012; Buzawa et al., 1999; Klein, 2009; Klein, Wilson, Crowe and DeMichele, 2005) and that rapid disposition of cases positively impacts victim safety (Klein, 2009).

The aggressive prosecution of all domestic violence cases is difficult because the evidentiary burden for a successful prosecution is difficult to meet without a willing victim, especially in light of the U.S. Supreme Court's decision in *Crawford v. Washington* (2004). In *Crawford*, the Court held that allowing testimonial statements to be entered into evidence when a witness was unavailable to testify violated the 6th Amendment right to confrontation of accusers, unless the defendant had prior opportunity to cross-examine the witness.

Crawford was initially viewed as a potentially debilitating blow to evidence-based prosecution as prosecutors were not sure what would constitute a "testimonial statement."[4] Although prosecutors have found various ingenious ways to attempt to admit the statements of noncooperating victims into court, it is clear that this remains a significant burden to

4. The Supreme Court provided some illumination in two later cases, *Davis v. Washington* (2006) and *Hammon v. Indiana* (2006), where the Court attempted to clarify what constituted a testimonial statement in the context of domestic violence. The Supreme Court noted that not every statement made to the police will be construed as testimonial in nature and, thus, excluded from court if the victim does not cooperate in any subsequent prosecution.

prosecution requiring both police and prosecutors to develop additional strategies to collect additional evidence as a victim's continued cooperation cannot be assumed.

Therefore, even after district attorneys moved beyond any historic tendency of neglecting domestic violence, they were faced with key evidentiary problems and the need to determine how to allocate scarce resources appropriately (Garner and Maxwell, 2008). Although there are variations, several competing approaches to prosecution have developed, which are outlined as follows.

Evidenced-Based Approach

An evidence-based approach often is conflated with "no-drop prosecution," although the distinction is in their application. An evidence-based approach calls on the prosecutor to evaluate all the evidence and move forward with prosecution only if the prosecutor believes that prosecution will be successful, regardless of victim cooperation. A no-drop policy, in its most extreme form, calls for a prosecutor to move forward with prosecution despite evidentiary difficulties. Under an evidence-based approach, a victim's direct testimony is only one more piece of evidence and her desires with regard to moving forward with prosecution is important only as it affects a prosecutor's ability to garner a successful conviction. If the prosecutor feels that the case can proceed without her testimony, or if faced with a recalcitrant victim, then the victim is either subpoenaed to testify or the case will be brought to court without the victim.

An evidence-based approach has intrinsic appeal for several reasons. First, it expressly recognizes that prosecutors, although they need to protect the rights of individuals, must place the rights of society in a paramount position in exercising their discretion [ABA Standards 3–1.2(b)]. The state has a heavy interest in punishing and preventing the occurrence of violent crime. In this manner, prosecution is an offender-oriented approach aimed at specific deterrence of an identified offender. One principal tenet of deterrence theory is that speed and certainty of prosecution will limit future reoffending behavior (Zimring, 1974). Indeed, several highly publicized "empirical" research studies have suggested that specific deterrence of the offender in question might be enhanced by the certainty of arrest (Sherman and Berk, 1984) and prosecution (Paternoster, Saltzman, Waldo, and Chiricos, 1983). These studies, however, are not without controversy (Maxwell, Garner, and Fagan, 2001; Paternoster, 1987).

Likewise, some have posited that a high likelihood of prosecution will have a general deterrent effect on other potential offenders. Advocates of such an approach believe that, in common with most crimes, the state has an overriding interest in general deterrence. The prospect of a reasonably certain conviction of an offender after a crime becomes known to the community and domestic violence rates decline when potential batterers realize that they are likely to be convicted. The importance of deterrence in the context of evidence-based policies cannot be overstated. Advocates of evidence-based prosecution usually implicitly conclude that the primary overall value of prosecution is to punish and thereby deter both

the offender and future potential perpetrators, thus justifying overriding any specific victim's desire not to prosecute. Without such a state interest, the responsibility of the prosecutor should logically be to support victim-legitimate interests in restorative justice, as long as these are largely consistent with proportional treatment of equally guilty offenders.

As an ancillary matter, evidence-based practices can be potential solutions to several vexing organizational problems—such as the historic tendency to underfund domestic violence prevention and prosecution (Garner and Maxwell, 2008). By forcing agencies to expend resources and prosecute the majority of cases, many have anticipated that the organizational culture will shift among prosecutors, favoring awareness and responsiveness to such criminal activity. Such a policy also might assist the court in preventing victims of domestic violence from being coerced into not testifying, or being psychologically unprepared to face their abuser in court by collecting and analyzing evidence without consideration of the victim having to testify, pressure is taken off of her to cooperate and the prosecutor does not have to worry about a victim changing her mind on the eve of trial.

Victim-Centered Approach

By contrast, a victim-centric approach to prosecution recognizes that although domestic violence affects society as a whole, the victim has the most interest in the outcome of any court case, except the offender (Bennett Cattaneo and Goodman, 2010; Han, 2003; O'Sullivan, Davis, Farole, and Rempel, 2008).

Prosecutors using this approach first educate the victim of legal rights and encourage the victim to follow the case through to conviction (or a suitable diversionary program), but ultimately they allow most victims to decide whether prosecution is warranted. Even in jurisdictions like county B, exceptions to victim preference are made when prosecutors are presented with a case of serious injury, repeat violence, or clear inability of the victim to make a rational choice. Overall, proponents of this policy claim that empowering victims enhances their well-being, making them more self-reliant and less likely to tolerate abuse (Diesen, 2011; Winick and Wexler, 2003).

Furthermore, studies have shown that giving power to the victim to influence case conduct promotes specific deterrence because the offender is, to a degree, dependent on the victim's forbearance (Ford and Regoli, 1993). It is increasingly recognized that prosecution itself might result in significant costs to victims in terms of time, fear of loss of income, risk of offender retaliation, possible effects on child custody, and possible negative reaction from her community or family. Although the societal interest in prosecuting a habitual offender is undoubtedly more important than these potential costs for certain habitual or serious offenders, it can be argued that victims, especially in cases of minor assaults, should be allowed to make this decision.

The importance of the empowerment model is a recognition that in most cases, domestic abusers are not committing violence merely to hurt or cause injury, but instead they want to dominate and gain control in a relationship by inflicting pain and instilling fear

(Stark, 2007). The victims, who may be traumatized, remain aware of their risks and can best predict future violence (Campbell, 2012; Campbell, Glass, Sharps, Laughon, and Bloom, 2007; Campbell, Webster, and Glass, 2009). Although some victims possibly cannot express their needs or desires because they are under coercive control, in most interventions, victims can articulately express their intents and fairly appraise the consequences of alternative courses of action, including prosecution of the case. Thus, although many criminal justice and victim advocates advise victims on what they *should* do (i.e., prosecute and leave the abusive environment), often a victim is in the best position to know whether this is realistic. Although it is certainly not optimal from a societal point of view, if a victim knows that she is likely to return to her accuser, then it might not be in her best interest to prosecute. Unfortunately, this is true with the most serious offenders who often abuse while enraged, with little regard to future or previous prosecution.

Victim-centric approaches allow victims to regain power and return a sense of control to their lives by giving them a role in the prosecution process. Serving as the key decision maker might empower victims and facilitate their emotional and physical recovery (Belknap and Potter, 2005). Advocates of this approach argue that in cases where the victim, rather than society, suffers the greater harm, a primary goal should be victim empowerment. In this regard, scarce resources should be given to ensure all victims have (a) knowledge of their legal remedies, (b) available alternatives including shelters, (c) prosecution through conviction, (d) diversion programs, and (e) the power to use such laws and institutions as resources to get key needs met, including safety.

From this perspective, automatic case prosecution might vitiate an opportunity to empower a victim while preventing perfectly acceptable alternatives. In reviewing the relative merits of these theories, and to place the studied research in the proper context, we have to examine existing empirical data on both general and specific deterrence in the context of prosecution policies.

General Deterrence

No definitive research is available on how prosecutorial policy outcomes (independent of police and judicial dispositions) in domestic violence cases impact domestic violence. Partially this is because such research is difficult to conduct. Many elements of the criminal justice system and society's response to domestic violence have changed partly because of legislative and policy mandates. Meanwhile, there has been an overall reduction in the rate of domestic violence over the past 20 years, which allows an inference to be drawn that many potential offenders were deterred by something that has changed during the last 20 years.

But what are the relevant factors? Clearly, the media regularly reports on domestic violence, whether in the context of an overall societal issue, the glare of an arrest of a public figure, or the shame of a murderous rampage. This observation does not independently support prosecution for minor domestic violence because the arrest, not subsequent court disposition, garners press attention. Changes in arrest practices have an impact because they

are highly visible events; often are published or become known to a circle of witnesses, friends of the family, or acquaintances; or appear in local media reports. However, the only court dispositions publicized typically involve significant felony murders, serious injuries, or in rare cases a publicly known offender, not a low-level miscreant. Therefore, we do not anticipate that the general deterrence created by aggressive case prosecution would be a major factor in general deterrence of domestic violence.

Specific Deterrence

During the last several decades, substantial empirical data have been accumulated to suggest that actual court outcomes—whether dismissal, acquittal, conviction, or even incarceration—do not impact the likelihood of future recidivism substantially (Klein, 2009; Klein and Tobin, 2008; Mears, Carlson, Holden, and Harris, 2001). This finding is not surprising because most acts of domestic abuse are not carefully preplanned. In this context, specific deterrence's underlying premise that an offender carefully considers the utility of future misconduct is unlikely. Even if concern for future repercussions might deter some, for most the deterrent value is the likely intervention by the police—not the added dimension of certain prosecution. In fact, the most serious, and likely repeat, offenders are those with a substantial history of violent crimes, precisely those least likely to be deterred by yet another prosecution for a relatively minor assault (Buzawa et al., 1999; Klein and Tobin, 2008).

Only two factors have been shown to decrease recidivism reliably. First, cases in which the offender is consistently monitored closely by the judicial disposition, including probation and batterer intervention programs (Rempel, 2009), seem to be promising. The offenders, through repeat interactions with law enforcement and court personnel, are aware that they are under the surveillance of the court and should understand that future assaults will bring immediate negative consequences. It should be noted that this outcome can be achieved not only by conviction but also through case deferral conditional on acceptance and completion of a batterer intervention program (BIP) or supervised probation.[5] Unfortunately, as discussed, this option is unlikely to stop the most serious cohort of offenders from reoffending.

The second is when the victim has been empowered to dismiss a case. Most offenders who would be deterred by actual prosecution also might be deterred by credible threats of future prosecution in which commission of the next offense results in prosecution for multiple offenses. Victim control over an offender's criminal prosecution might inherently change the balance of power in a relationship, predictive of lower rates of recidivism (Ford and Regoli, 1993). In this context, the reported study by Finn (2013) does seem to offer confirmation that in the two most important criteria for success or failure, "violence" and "intimidation," victim involvement in prosecutorial decisions does work.

5. In this regard, county A did not report any "deferrals," whereas county B made full use of this disposition.

Public Policy Implications

Cost Differentials Do Not Seem Warranted

Finn's (2013) reported findings that an evidence-based approach leads to more, not less, future violence and psychological intimidation is an important result. Furthermore, it must be placed in a public policy context. Although numerous articles address the difficulty in prosecuting victimless domestic violence cases in the wake of *Crawford*, less has been written about the cost in pursuing these cases.

Given the fiscal situation faced by many police departments and district attorney offices throughout the country, budget and manpower constraints must be an important consideration when evaluating an evidence-based approach to prosecution. Unless the state is willing to force a victim to testify via subpoena, and enforce it, any prosecution that does not take a victim's desires into account at the outset is bound to use far more resources. Either prosecutions will have to be dropped on the eve of court or prosecutors will have to threaten contempt of court by subpoenaing victims and moving for the court to issue contempt citations if they remain uncooperative. By contrast, a victim-centric approach should reduce this—if the police and prosecutor work with the victim from the onset, then they will have a much better idea whether the victim is willing to cooperate and be less likely to have a victim who backs out immediately preceding trial.

Targeting of Resources Might Be Achieved by a Victim-Centric Approach

A victim-oriented approach allows prosecutors to target limited resources toward cases with the greatest likelihood of conviction. This can be supplemented on a case-by-case basis, overriding victim preferences if it is reasonably apparent that (a) a victim will not make informed decisions regarding her safety, (b) children are impacted, or (c) the offender poses a high risk or is a habitual violent offender. This approach is the antithesis to assembly line justice in which all cases may be prosecuted and many convictions result, but because of the massive volume of cases, sentences rarely reflect criminal history (Klein, 2009).

In contrast, as this study by Finn (2013) confirms, an evidence-based approach, as in the contrasting study of the practices in Brooklyn and the Bronx, showed a slower time to trial despite the increased resources (Davis, Farole, O'Sullivan, & Remple, 2008), which in turn not only drives systemic costs but also increases prospects for recidivism during the critical period after the first reported assault (Buzawa and Hotaling, 2007; Klein, 2009).

Likely Impact on the Victim Experience with the Criminal Justice System

Although Finn (2013) did not report major differences in her process-orientated measurement of victim empowerment, overall most research has reported lower victim empowerment in evidence-based jurisdictions both because victims are likely to be more physically abused or psychologically terrorized and because disempowering the victim negatively affects victim satisfaction with the intervention, a factor in turn associated with lower rates of reporting future abuse (Buzawa and Hotaling, 2007; Belknap et al., 2000; Weisz,

Canales-Portalatin, and Nahan, 2001). Ironically, the failure to report further abuse might be misinterpreted to convince policy makers mistakenly that recidivism rates are reducing faster than they actually are (Belknap et al., 2000; Buzawa and Hotaling, 2007; Weisz, Canales-Portalatin, and Nahan, 2001).

Research Implications

Future Victim Behavior

Finn (2013) examined one dimension of the impact of prosecutorial policy: deterrence of further psychological or physical abuse by the *offender*. The same methodology could be expanded to determine whether an evidence-based or a victim-centered policy affected the *victim's* likely future behavior. Future research should therefore expressly ask victims whether they would re-report an incident and separately whether they believe their current experiences made them more or less likely to report.

Our prediction would be that overall, a lower percentage of victims would likely report in the evidence-based jurisdiction. We also predict that victims at greatest risk would be more likely not to report future abuse because the criminal justice system failed to protect them, especially in jurisdictions like county A where enhanced prosecutions do not weed out low-risk offenders via pretrial diversions into batterer treatment programs nor target limited resources on key offenders.

Need for Further Research

This research by Finn (2013), although it is significant, is based on the response of 170 victims of 1,600 cases. Despite reasonable efforts to determine whether nonreporters differed from those who report, the low completion rate makes it unclear whether the substantial number of nonresponders did so because of increased dissatisfaction with the criminal justice response. Thus, Finn's research might be complemented by merging the study's victim interviews with official data from the larger population—similar to how the comparative study of the Bronx and Brooklyn prosecution practices used official data on recidivism and studied victim satisfaction within a small cohort (O'Sullivan et al., 2007).

Controlling Influence of Demographic and Other Factors

More than 50% of the residents in county B were African American compared with 14% in county A. This difference was apparent in the percentage of cases, as 77.6% of victims in county B were African American compared with 22.4% in county A. Although we believe that a victim-centered prosecution policy is preferable in all jurisdictions, we acknowledge that a victim-centered policy might be more effective in the African American community, where the historic tensions between their community and the police and courts make mutual distrust more likely. This instrument could beneficially be repeated in jurisdictions that are more comparable.

County A, which employed the evidence-based prosecution approach, did not seem to use diversion into BIPs, unlike in county B. It would be useful to determine the role played by diversionary programs in victim-oriented approaches. This is important because prior research found (Rempel, 2009) that a BIP with supervision did increase offender accountability and increase victim safety. In other words, the key factor favoring county B's outcomes might be the increased assurance of offender accountability via BIP as much as the victim's ability to determine prosecution.[6]

County A also seemed to sentence a larger percentage of offenders to incarceration, whereas the offenders studied in county B had more extensive criminal records. This artifact could affect statistics on violence reoccurrence independently for the simple reason that if someone is incarcerated or on highly supervised parole, then it is less likely that he or she will reoffend during the 6-month study period.

Replication in a Specialized Domestic Violence Court

As Finn (2013) notes, specialized domestic violence courts might be better able to support the needs of a victim in cases in which an evidence-based or "no-drop" prosecution policy is followed. Perhaps this might alleviate the problems observed in county A.[7] The necessary coordination and buy-in from the judiciary, typically observed in such specialized courts, might be a requisite for prosecutors planning to adopt an evidence-based system prospectively. Alternatively, such courts, when staffed properly with competent, knowledgeable personnel, might work even better using a victim-centered prosecution policy because these officials would be far more likely to remind victims of their rights, have available resources, and display a willingness to divert most suitable offenders into the right diversionary programs while targeting the more hardcore subset of violent offenders.

Conclusion

Advocates for a victim-centered approach have long asserted that prosecution has a "cost" for a victim. Finn (2013) demonstrates one aspect of the "cost"—increased risks of physical abuse and psychological intimidation. Mounting empirical evidence suggests that a no-drop policy negatively impacts specific deterrence, whereas claims of general deterrence of potential batterers remain, as before, highly speculative.

6. Similarly, even if an offender is sentenced to BIP, it may be as effective, if not more effective, prior to prosecution as after prosecution. Some might argue that offenders might be more willing to attend simply because it was not court mandated and they could refuse, whereas others would argue that court-forced attendance is more important. Regardless, county B made extensive use of this, whereas the other county did not, adding that ideally, a potential artifact that might be controlled or otherwise limited.

7. However, the research regarding the effectiveness of these courts on recidivism is unclear. Some recent work has suggested they are no more effective at preventing reoffending but are more effective at increasing victim satisfaction (which may impact re-reporting by victims).

One remaining issue is somewhat more open to judgment and not answerable by empirical data alone. Should the societal and thus criminal justice goal be to match the punishment to the seriousness of the offense, or should it be to respond to what is in a victim's best interests (increasing personal safety, empowerment, and satisfaction)? This question is not trivial. Empowering victims to be key decision makers results in variation of case disposition for comparable offenses. Although we are satisfied with this trade-off, it could be argued that such a process is inherently unfair. One overarching goal of the criminal justice system is to ensure the equitable treatment of offenders and that punishments, when meted out, are not arbitrary. This type of victim-led prosecution, would, for example, be viewed as fundamentally unfair if applied unilaterally to many other offenders.

Finally, we recognize that a victim-centered approach has the potential to make achieving another emerging goal more difficult: the immediate identification of habitual violent offenders. Research has now recognized that the most hardcore offenders capable and likely to commit serious injuries are "generally violent" (Holtzworth-Munroe and Meehan, 2004; Johnson and Ferraro, 2000) and are likely to both re-abuse the same victim or target new victims of violent assaults unless they are identified and held accountable (Buzawa, Hotaling, Klein, and Byrne, 1999; Klein, 2009). The failure to prosecute such offenders at an earlier stage might increase the potential for the future victimization of the victim or someone else. Therefore, we want any victim-centered approach to retain the capability of targeting these most serious offenders, with or without victim concurrence.

References

Belknap, Joanne, Dee L. R. Graham, Jennifer Hartman, Victoria Lippen, P. Gail Allen, and Jennifer Sutherland. 2000. *Factors Related to Domestic Violence Court Dispositions in a Large Urban Area: The Role of Victim/Witness Reluctance and Other Variables*. Washington, DC: National Institute of Justice. NCJ 184112. Retrieved from cjrs.gov/App/Publications/abstract.aspx?ID = 184112.

Belknap, Joanne and Hillary Potter. 2005. The trials of measuring the "success" of domestic violence policies. *Criminology & Public Policy*, 4: 559–566.

Buzawa, Eve S. and Aaron D. Buzawa. 2008. Courting domestic violence victims: A tale of two cities. *Criminology & Public Policy*, 7: 671–685.

Buzawa, Eve S., Carl G. Buzawa, and Evan D. Stark. 2012. *Responding to Domestic Violence: The Criminal Justice and Societal Response*. Thousand Oaks, CA: Sage.

Buzawa, Eve S. and Gerald T. Hotaling. 2007. Understanding the impact of prior abuse and prior victimization on the decision to forego criminal justice assistance in domestic violence incidents: A lifecourse perspective. *Brief Treatment & Crisis Intervention*, 7: 55–76.

Buzawa, Eve S., Gerald T. Hotaling, Andrew Klein, and James Byrne. 1999. *Response to Domestic Violence in a Pro-Active Court Setting: Final Report*. Washington, DC: National Institute of Justice.

Campbell, Jacquelyn C. 2012. Risk factors for intimate partner homicide: The importance of Margo Wilson's foundational research. *Homicide Studies*, 16: 438–444.

Campbell, Jacquelyn C., Nancy Glass, Phyllis W. Sharps, Kathryn Laughon, and Tina Bloom. 2007. Intimate partner homicide: Review and implications of research and policy. *Trauma, Violence, & Abuse*, 8: 246–269.

Campbell, Jacquelyn C., Daniel W. Webster, and Nancy Glass. 2009. The danger assessment: Validation of a lethality risk assessment instrument for intimate partner femicide. *Journal of Interpersonal Violence*, 24: 653–674.

Cattaneo, Lauren Bennett and L. Goodman. 2010. Through the lens of therapeutic jurisprudence: The relationship between empowerment in the court system and wellbeing for intimate partner violence victims. *Journal of Interpersonal Violence*, 25: 481–502.

Davis, Robert C., Donald J. Farole, Jr., Chris S. O'Sullivan and Michael Rempel. 2008. A comparison of two prosecution policies in cases of intimate partner violence. *Criminology and Public Policy*, 7(4): 633–662.

Diesen, Christian. 2011. Therapeutic jurisprudence and the victim of crime. In (Lernestedt and Tham, Eds.) Brottsoffret och kriminalpolitiken. Malmö, Sweden: Liber. Retrieved October 29, 2013 from law.arizona.edu/depts/upr-intj/pdf/Therapeutic-Jurisprudence-and-the-Victim-of-Crime.pdf.

Finn, Mary A. 2013. Evidence-based and victim-centered prosecutorial policies: Examination of deterrent and therapeutic jurisprudence effects on domestic violence. *Criminology & Public Policy*, 12: 443–472.

Ford, David A. and Mary Jean Regoli. 1993. The criminal prosecution of wife assaulters. In (N. Zoe Hilton, ed.), *Legal Responses to Wife Assault: Current Trends and Evaluation*. Newbury Park, CA: Sage.

Garner, Joel H. and Christopher D. Maxwell. 2008. Coordinated community responses to intimate partner violence in the 20th and 21st centuries. *Criminology & Public Policy*, 7(4): 301–311.

Han, Erin L. 2003. Mandatory arrest and no-drop policies: Victim empowerment in domestic violence cases. *Boston College Third World Law Journal*, 23: 151–192.

Holtzworth-Munroe, Amy and Jeffrey C. Meehan. 2004. Typologies of men who are maritally violent: Scientific and clinical implications. *Journal of Interpersonal Violence*, 19: 1369–1389.

Johnson, Michael and Kathleen Ferraro. 2000. Research on domestic violence in the 1990's: Making distinctions. Journal of Marriage and the Family, 62: 948–963.

Klein, Andrew R. 2009. *Special Report: Practical Implications of Current Domestic Violence Research*. Washington, DC: U.S. Department of Justice, National Institute of Justice.

Klein, Andrew R. and Terri Tobin. 2008. Longitudinal study of arrested batterers, 1995–2005: Career criminals. *Violence Against Women*, 14: 136–157.

Klein, Andrew R., Douglas Wilson, Ann H. Crowe, and Matthew T. DeMichele. 2005. *Evaluation of the Rhode Island Probation Specialized Domestic Violence Supervision Unit*. Final Report submitted to the National Institute of Justice (American Probation and Parole Association and BOTEC Analysis Corporation).

Maxwell, Christopher D., Joel H. Garner, and Jeffrey A. Fagan. 2001. *Effects of Arrest on Intimate Partner Violence: New Evidence from the Spouse Assault Replication Program*. Washington, DC: National Institute of Justice.

Mears, Daniel P., Matthew J. Carlson, George W. Holden, and Susan D. Harris. 2001. Reducing domestic violence revictimization: The effects of individual and contextual factors and type of legal intervention. *Journal of Interpersonal Violence*, 16: 1260–1283.

Nichols, Andrea J. 2013a. Meaning-making and domestic violence victim advocacy: An examination of feminist identities, ideologies, and practices. *Feminist Criminology*, 8: 177–201.

Nichols, Andrea J. 2013b. Survivor-defined practices to mitigate revictimization of battered women in the protective order process. *Journal of Interpersonal Violence*, 28: 1403–1423.

O'Sullivan, Christ S., Robert C. Davis, Donald J. Farole, Jr., and Michael Rempel. 2007. *Comparison of Two Prosecution Policies in Cases of Intimate Partner Violence*. New York: Center for Court Innovation.

Paternoster, Raymond. 1987. The deterrent effect of the perceived certainty and severity of punishment: A review of the evidence and issues. *Justice Quarterly*, 4: 173–217.

Paternoster, Raymond, Linda E. Saltzman, Gordon P. Waldo, and Theodore G. Chiricos. 1983. Perceived risk and social control: Do sanctions really deter? *Law & Society Review* 17: 457–479.

Rempel, Michael. 2009. Batterer programs and beyond. In (Eve Stark and Eve S. Buzawa, eds.), *Violence Against Women in Families and Relationships, Volume Three: Criminal Justice and the Law*. Santa Barbara, CA: Praeger.

Sherman, Lawrence W. and Richard A. Berk. 1984. The specific deterrent effects of arrest for domestic assault. *American Sociological Review*, 49: 261–272.

Stark, Evan. 2007. *Coercive Control: How Men Entrap Women in Personal Life*. New York: Oxford University Press.

Weisz, Arlene N., David Canales-Portalatin, and Neva Nahan. 2001. *Evaluation of Victim Advocacy Within a Team Approach*. Final report for National Institute for Justice, NCJ 187107. Washington, DC: U.S. Department of Justice, National Institute of Justice. Retrieved October 27, 2013 from ncjrs.gov/App/Publications/abstract.aspx?ID = 187107.

Winick, Bruce J. and D. B. Wexler (Eds.). 2003. *Judging in a Therapeutic Key*. Durham, NC: Carolina Academic Press.

Zimring, Franklin E. 1974. Threat of punishment as an instrument of crime control. *Proceedings of the American Philosophical Society*, 118: 231.

Court Cases Cited

Crawford v. Washington, 541 U.S. 36 (2004).

Davis v. Washington, 126 S. Ct. 2266 (2006).

Hammon v. Indiana, 547 U.S. 813 (2006).

Eve S. Buzawa director and professor of the School of Criminal Justice and Criminology at the University of Massachusetts—Lowell. Dr. Buzawa's research interests and publications encompass a wide range of issues pertaining to domestic violence, policing, and violence against women. She has authored and edited numerous books and monographs. Recent books include *Responding to Domestic Violence: The Integration of Criminal Justice & Human Services* (co-authors are Carl G. Buzawa and Evan Stark) and *Violence against Women in Families and Relationships: Making and Breaking Connections*, a 4-volume set, (co-editor with Evan Stark, 2009). She has also served as a Principal Investigator on several federally funded research projects as well as directing numerous state funded research and training projects.

Aaron Buzawa is a Captain in the U.S. Air Force. He is currently assigned as a Special Victims' Counsel at Keesler AFB, Mississippi. He has previously served as an assistant staff judge advocate. His areas of research interest include interventions and services for victims of crime and underserved populations. He is a member of the Florida bar. The views expressed in this article are the author's own and do not reflect the official policy or position of the United States Government, or any entity therein, including the Air Force or the Department of Defense.

Machine Learning Approaches as a Tool for Effective Offender Risk Prediction

William Rhodes
Abt Associates

The prediction of criminal behavior plays an instrumental role in criminal justice administration (Gottfredson and Moriarty, 2006). Social workers provide high-risk youth with mentoring, police intensify patrol activity in high-risk neighborhoods, prison administrators segregate high-risk offenders from low-risk ones, and community corrections administrators concentrate controlling and correctional resources on supervisees at an elevated risk of recidivism. Risk assessment is an essential ingredient of evidence-based criminal justice administration, and good risk assessment is better than bad risk assessment.

Berk and Bleich (2013, this issue) advocate for wider use of machine learning approaches for deriving predictions. This advocacy appears as a gentle introduction intended to convince readers that machine learning works, to motivate a deeper reading into a technical literature for why it works (Berk, 2012), and eventually to induce widespread application. Their argument has two principal components.

Berk and Bleich (2013) assert that predication is difficult and uncertain using conventional regression-based methods. They reference a complex decision boundary, by which they mean that good prediction can depend on many variables, which might require transformations, interactions, and nonlinear manipulation of data. They are skeptical of regression-based adaptations for identifying complex decision boundaries, and even if in theory regression-based procedures could be hammered into a suitable form, the training and experience of most criminal justice researchers provide inadequate statistical carpentry. Do not take a chance on conventional regression-based procedures, they exhort; turn to machine learning algorithms that can detect complex patterns.

The second component of Berk and Bleich's (2013) argument is that regression-based approaches use an inadequate loss function that fails to weight some outcomes as more serious than others (Berk, 2011). For example, predictions placing greater weight on future

Direct correspondence to William Rhodes, Abt Associates, 55 Wheeler Street, Cambridge, MA 02138 (e-mail: bill_rhodes@abtassoc.com).

DOI:10.1111/1745-9133.12060

homicide than on shoplifting are more valuable than predictions of undifferentiated recidivism. Their argument's second component has different standing than its first component because nonsymmetric loss functions could be incorporated into conventional regression-based approaches, although there might be an advantage to incorporating a loss function directly into the estimation routine rather than solely into the prediction procedure.

Some criminal justice researchers and practitioners might not find Berk and Bleich's (2013) arguments compelling. Criminal justice risk assessment has moved through four generations. In the first generation, risk assessment was based on professional judgment, which holds considerable sway in the practitioner community. In the second generation, prediction was placed on a statistical foundation. Berk and Bleich's approach falls solidly within this second-generation framework. Third- and fourth-generation risk assessment recognizes that predictions should change dynamically over the course of supervision and that prediction is necessarily intertwined with program interventions. Berk and Bleich's approach does not address concerns that motivate third- and fourth-generation risk prediction, which is a point conceded by Berk and Bleich.

Likewise, criminal justice program evaluators might not find Berk and Bleich's (2013) arguments convincing. A program evaluator, especially one working with observational data, where the objective is disentangling causation from selection bias, is unlikely to gravitate toward the machine learning approach (Bushway and Smith, 2007; Rhodes, 2011). Berk and Bleich concede this point as well: The machine learning approaches provide empirical profiles, adequate for prediction in steady-state environments but generally inadequate for explanation or adaptation in environments that change because of policy interventions.

The preceding discussion is not a criticism of Berk and Bleich (2013). They are not promoting a jackknife suitable for solving all problems but rather a single blade solution to a difficult and important public policy concern. They are inarguably correct that flexible computer-driven search procedures, backed up by validation, can improve predictions. Criminal justice researchers should take heed.

The four scholars commenting on Berk and Bleich's (2013) article join me as a sympathetic audience. Shawn D. Bushway (2013, this issue), a distinguished user of structural equation models, nevertheless concludes that machine learning is providing strong and growing competition to more familiar regression-based tools. He recommends that criminal justice methodologists include machine learning in their bag of standard methodological tools.

Greg Ridgeway (2013, this issue) agrees that criminal justice researchers should grasp machine learning principles, but his endorsement is more circumspect. Echoing my comments regarding third- and fourth-generation prediction tools, Ridgeway observes that prediction often needs to be dynamic; for example, in the domain of community supervision, risk assessment is updated continuously based on recent probationer performance. He observes that analysis ignoring underlying structural equations "conflate[s] intervention activities with the underlying crime phenomena." Developing dynamic risk assessments and

distinguishing offender behavior from public activities to control that behavior are not the strengths of machine learning's black box.

As contributors to machine learning methods in criminal justice, Tim Brennan and William Oliver (2013) are, like the other commentators, sympathetic to Berk and Bleich's (2013) arguments, but not uncritically. They contest the view that criminal justice practitioners need not understand the workings within the black box. Citing Tata (2002), Brennan and Oliver conclude that "judges, and other decision makers, must reach some understanding of their cases to design effective sentencing components and to often justify their reasoning."

First-round reviewers of Berk and Bleich's (2013) article took issue with contradictory evidence that machine learning leads to better predictions. Although they find Berk and Bleich's evidence "compelling," Brennan and Oliver (2013) repeat the criticism of first-round reviewers: "[W]e suggest that a prudent approach-—given the current dearth of comparative studies of ML forecasting accuracy in criminal justice—is to wait for subsequent systematic evaluations." Hopefully Berk and Bleich's introduction will motivate these requested studies. I would especially like to read studies that contrast machine learning with rigorously applied linear probability models and other statistical methodology such as survival analysis. How, for example, does the machine learning approach answer questions about timing or recidivism, and how does it deal with censored data and competing events?

My colleagues and I use supervised machine learning to assemble and analyze correctional statistics, so as a fan, I join others in commending Berk and Bleich (2013) for introducing a broader audience to machine learning applications suitable for prediction. I am also a fan of the Boston Red Sox but that does not preclude a healthy skepticism about their winning the World Series every year.

References

Berk, Richard A. 2011. Asymmetric loss functions for forecasting in criminal justice settings. *Journal of Quantitative Criminology*, 27: 107–123.

Berk, Richard A. 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. New York: Springer.

Berk, Richard A. and Justin Bleich. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, 12: 513–544.

Brennan, Tim and William Oliver. 2013. The emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions. *Criminology & Public Policy*, 12: 551–562.

Bushway, Shawn D. 2013. Is there any logic to using logit: Finding the right tool for the increasingly important job of risk prediction. *Criminology & Public Policy*, 12: 563–567.

Bushway, Shawn D. and Jeffrey Smith. 2007. Sentencing using statistical treatment rules: What we don't know can hurt us. *Journal of Quantitative Criminology*, 23: 377–387.

Gottfredson, Stephen D. and Laura J. Moriarty. 2006. Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, 52: 178–200.

Rhodes, William. 2011. Predicting criminal recidivism: A research note. *Journal of Experimental Criminology*, 7: 57–71.

Ridgeway, Greg. 2013. Linking prediction and prevention. *Criminology & Public Policy*, 12: 545–550

Tata, Cyrus. 2002. Accountability for the sentencing process—Towards a new understanding. In (Cyrus Tata and Neil Hutton, eds.), *Sentencing and Society*. Aldershot, U.K.: Ashgate.

William Rhodes is a principal scientist at Abt Associates Inc., a public policy consulting firm located in Cambridge, MA. An econometrician and program evaluation expert within Abt's domestic health division, his work is concentrated in behavioral health, especially crime and drug abuse. Current clients include the federal Office of Probation and Pretrial Services, the Office of Drug Control Policy, the Bureau of Justice Statistics, the National Institute of Justice, and the Agency for Healthcare Research and Quality. A Ph.D. economist (Minnesota), Rhodes was formerly an assistant professor at Florida State University, senior economist at the Institute for Law and Social Research, and director of research at the U.S. Sentencing Commission.

Overview of: "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment"

Richard A. Berk

Justin Bleich

University of Pennsylvania

Research Summary

A substantial and powerful literature in statistics and computer science has clearly demonstrated that modern machine learning procedures can forecast more accurately than conventional parametric statistical models such as logistic regression. Yet, several recent studies have claimed that for criminal justice applications, forecasting accuracy is about the same. In this article, we address the apparent contradiction. Forecasting accuracy will depend on the complexity of the decision boundary. When that boundary is simple, most forecasting tools will have similar accuracy. When that boundary is complex, procedures such as machine learning, which proceed adaptively from the data, will improve forecasting accuracy, sometimes dramatically. Machine learning has other benefits as well, and effective software is readily available.

Policy Implications

The complexity of the decision boundary will in practice be unknown, and there can be substantial risks to gambling on simplicity. Criminal justice decision makers and other stakeholders can be seriously misled with rippling effects going well beyond the immediate offender. There seems to be no reason for continuing to rely on traditional forecasting tools such as logistic regression.

Keywords

forecasting, *machine learning*, *recidivism*, *logistic regression*

DOI:10.1111/1745-9133.12044

Statistical Procedures for Forecasting Criminal Behavior

A Comparative Assessment

Richard A. Berk

Justin Bleich

University of Pennsylvania

Forecasts of recidivism have been widely used in the United States to inform parole decisions since the 1920s (Borden, 1928; Burgess, 1928). Of late, such forecasts are being proposed for a much wider range of criminal justice decisions. One important example is recent calls for predictions of "future dangerousness" to help shape sentencing (Casey, Warren, and Elek, 2011; Pew Center of the States, 2011). The recommendations build on related risk-assessment tools already operational in many jurisdictions, some mandated by legislation (Hyatt, Chanenson, and Bergstrom, 2011; Kleiman, Ostrom, and Cheeman, 2007; Oregon Youth Authority, 2011; Skeem and Monahan, 2011; Turner, Hess, and Jannetta, 2009). In Pennsylvania, for instance, a key section of a recent statute reads as follows:

42 Pa.C.S.A.§2154.7. Adoption of risk assessment instrument.

(a) General rule. – The commission shall adopt a sentence risk assessment instrument for the sentencing court to use to help determine the appropriate sentence within the limits established by law for defendants who plead guilty or nolo contendere to, or who were found guilty of, felonies and misdemeanors. The risk assessment instrument may be used as an aide in evaluating the relative risk that an offender will reoffend and be a threat to public safety.

(b) Sentencing guidelines. – The risk assessment instrument may be incorporated into the sentencing guidelines under section 2154 (relating to adoption of guidelines for sentencing).

Thanks go to Bill Rhodes and three anonymous reviewers for many helpful comments on this article. Direct correspondence to Richard A. Berk, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (e-mail: berkr@wharton.upenn.edu).

Figures 1–4 are available in color in the online version of the article.

DOI:10.1111/1745-9133.12047

(c) Pre-sentencing investigation report. – Subject to the provisions of the Pennsylvania Rules of Criminal Procedure, the sentencing court may use the risk assessment instrument to determine whether a more thorough assessment is necessary and to order a pre-sentence investigation report.

(d) Alternative sentencing. – Subject to the eligibility requirements of each program, the risk assessment instrument may be an aide to help determine appropriate candidates for alternative sentencing, including the recidivism risk reduction incentive, State and county intermediate punishment programs and State motivational boot camps.

(e) Definition. – As used in this section, the term risk assessment instrument means an empirically based worksheet which uses factors that are relevant in predicting recidivism.

With such widespread enthusiasm and very high stakes, one might assume forecasting accuracy has been properly evaluated and determined to be good. In fact, competent evaluations can be difficult to find for a wide variety of criminal justice decisions. Some of the problems have a long history (Ohlin and Duncan, 1949; Ohlin and Lawrence, 1952; Reiss, 1951). For example, it is relatively rare for evaluations to be based on "test data" that were not used to construct the forecasting procedures. The danger is grossly overoptimistic assessments. More recent commentaries have documented several other problems, sometimes including no evaluation at all (Berk, 2012; Farrington and Tarling, 2003; Gottfredson and Moriarty, 2006).

The need for thorough and thoughtful evaluations has become even more important over the past decade because in addition to calls for a more routine use of crime forecasts, new forecasting tools from computer science and statistics have been developed. Often supported by formal proofs, simulations, and comparative applications across many different data sets, these tools promise improved accuracy in principle (Breiman, 1996, 2001a; Breiman, Friedman, Olshen, and Stone, 1984; Chipman, George, and McCulloch, 2010; Friedman, 2002; Vapnick, 1998).[1] For example, Breiman (2001a) provided a formal treatment of random forests and its comparative performance across 20 different data sets. Several instructive criminal justice applications are in print as well (Berk, 2012).

Yet, several recent articles have claimed that for criminal justice applications, the new tools perform no better than the old tools (Liu, Yang, Ramsay, Li, and Coid, 2011; Tollenaar and van der Heijden, 2013; Yang, Liu, and Coid, 2010). Logistic regression (Berkson, 1951) is a favorite conventional approach. The conclusion seems to be "why bother?" For criminal justice forecasting applications, the new procedures are mostly hype:

> The conclusion is that using selected modern statistical, data mining and machine learning models provides no real advantage over logistic regression and

1. Very accessible treatments can be found in several textbooks (Berk, 2008; Bishop, 2006; Hastie, Tibshirani, and Friedman, 2009).

> LDA. If variables are suitably transformed and included in the model, there seems to be no additional predictive performance by searching for intricate interactions and/or non-linear relationships. (Tollenaar and van der Heijden, 2013: 582)[2]

How can the proofs, simulations, and many applications provided by statisticians and computer scientists be so wrong? How can it be that statistical procedures being rapidly adopted by private firms such as Google and Microsoft and by government agencies such as the Department of Homeland security and the Federal Bureau of Investigation are no better than regression methods that have been readily available for more than 50 years? Why would the kinds of new analysis procedures being developed for analyzing a variety of data sets with hundreds of thousands of cases (Dumbill, 2013; National Research Council, 2013: Ch. 7) not be especially effective for a criminal justice data set of similar size?

A careful reading of the technical literature and recent criminal justice applications suggests that there can be a substantial disconnect between that technical literature and the applications favored by many criminal justice researchers. Statisticians and computer scientists sometimes do not distinguish between forecasting performance in principle and forecasting performance in practice. Criminal justice researchers too often proceed as if the new procedures are just minor revisions of the generalized linear model. In fact, the conceptual framework and actual procedures can be very different and require a substantial change in data analysis craft lore. Without a proper appreciation of how the new methods differ from the old, there can be serious operational and interpretative mistakes.

In this article, we try to improve the scientific discourse by providing an accessible discussion of some especially visible, modern forecasting tools that can usefully inform criminal justice decision making. Machine learning is used as the primary illustration. The discussion is an introduction to material addressed far more deeply in *Criminal Justice Forecasts of Risk: A Machine Learning Approach* (Berk, 2012). We also try to provide honest, "apples-to-apples" performance comparisons between the newer forecasting methods and more traditional approaches.

For some readers, it may be useful to make clear what this article is not about. As one would expect, there have been jurisprudential concerns about "actuarial methods" dating from at least the time when sentencing guidelines first became popular (Feeley and Simon, 1994; Messinger and Berk, 1987), and more recent discussions about the role of race have introduced an important overlay (Berk, 2009; Harcourt, 2007). The issues are difficult and real, but they are not addressed in this article. Our concerns are more immediate. Forecasts of future dangerousness are being developed and used. Real decisions are being made affecting real people. At the very least, those decisions should be informed by the

2. "LDA" stands for linear discriminant analysis.

best information available. And that information depends significantly on the forecasting procedures deployed.

Proper Criminal Justice Forecasting Comparisons

The conceptual foundation for criminal justice forecasting can easily be misconstrued (Ridgeway, 2013). We begin, therefore, with a fundamental conceptual point that some readers may at first find counterintuitive. As a formal matter, one does not have to understand the future to forecast it with useful accuracy. Accurate forecasting requires that the future be substantially like the past. If this holds, and one has an accurate description of the past, then one has an accurate forecast of the future. That description does not have to explain why the future takes a particular form and certainly does not require a causal interpretation. Readers comfortable with traditional time-series analysis (Box and Jenkins, 1970) should have no problem with this reasoning.

It follows that a key distinction between forecasting and explanation has been badly conflated in some accounts (Andrews, Bonta, and Wormith, 2006). Understanding a phenomena may lead to improved forecasting accuracy, or it may not, but forecasting and explanation are different enterprises that can work at cross-purposes. For example, explanatory models should be relatively simple and provide instructive interpretations. Such models can leave out a large number of weak predictors that one by one do not enlighten but in the aggregate dramatically improve forecasting accuracy. Common practice implicitly folds such variables into the disturbance term. Alternatively, such predictors, often called "nuisance variables" in limited information structural models, are associated "nuisance parameters" and given "minimal attention" (Cameron and Trivedi, 2005: 36). Similar issues arise if simple, easily interpretable functional forms (e.g., linear) are used when complex functional forms might fit the data somewhat better.[3]

The approach we take is to maximize forecasting accuracy, and that is the premise on which the underlying mathematics depend. We take this approach because it leads to clear performance criteria and various proofs of optimal forecasting accuracy for a given data set. Such clarity is an undeniable virtue about which more will be said shortly.

Equally important, there are a wide variety of decisions made by criminal justice officials in which a necessary condition is the best possible forecasting accuracy. Consider a judge's decision to sentence an offender to either incarceration or probation. Pennsylvania's statute states that a "risk assessment instrument may be used as an aide in evaluating the relative risk that an offender will reoffend and be a threat to public safety." Presumably, accuracy really matters. Imagine the ethical and legal implications of using a particular risk tool to

3. Some differences in jargon can be instructive. In machine learning, a "predictor" is often called an "input," and a response or dependent variable is often called a "target."

justify a long incarceration when more accurate risk tools exist from which a sentence of probation could be more appropriate.

Also, the legislation contains no requirement that a judge understand why an individual is high or low risk. Indeed, it is not even clear what a judge would do with such information.[4] Other examples include pretrial decisions to release defendants on bail or decisions by parole boards to release under supervision inmates who have not served their full terms. One also could imagine forecasts of future dangerousness helping to determine charging decisions by prosecutors.

Thus, there is no formal concern in this article with why certain predictors improve forecasting accuracy and no attempt is made to interpret them as explanations for the forecasted behavior. For example, if other things equal, shoe size is a useful predictor of recidivism, then it can be included as a predictor. Why shoe size matters is immaterial. In short, we are not seeking to identify risk factors that may or may not make any subject-matter sense. That can be a useful enterprise, but it is a different enterprise.

Indeed, if the enterprise really is explanation, then some form of structural equation modeling may be called for. An extensive and largely unrebutted literature has been highly critical of structural equation modeling in general. An excellent, accessible, and technically sound treatment can be found in David A. Freedman's textbook *Statistical Models: Theory and Practice* (2005). We cannot rehash the issues in this article except to stress that machine learning is not a form of structural equation modeling and should never be interpreted as such.[5] Moreover, if the goal is to use one or more risk factors to design and test interventions, then many would argue that the only sound approach is randomized experiments or very strong quasi-experiments.

Some Common-Sense Requirements for Fair Forecasting Comparisons

If one intends to compare the forecasting performance of different forecasting tools, then there are several basic, common-sense requirements. These provide the following ground rules:

1. One must be clear on what features of forecasting procedures are being compared. As we explain below, "black box" forecasting methods may forecast with remarkable accuracy and provide decision makers with tools that can be enormously helpful (Breiman, 2001b). But black box forecasting methods may have little to say about which risk

4. In the special case when there are clear indications of substance dependency or psychological problems, a judge might order treatment along with the sentence. But such conditions are not necessarily risk factors for many kinds of crime, and indications of need can be sufficient.

5. A structural equation model is an algebraic theory of how nature generated the data and, as such, can be right or wrong. Machine learning employs algorithms that seek some well-defined empirical goal, such as maximizing forecasting accuracy. There is no structural model. Concerns about whether the model is correct are irrelevant. What matters is how well the algorithm performs.

factors matter most. If the goal is to compare different procedures by their forecasting accuracy, then forecasting accuracy should be the benchmark.

2. Forecasting comparisons must be based on data not used to construct the competing forecasting procedures. Such data are often called "test data," and accuracy is often called "out-of-sample performance." Data used to build the forecasting procedures can be called "training data." If training data are also used as test data, then all comparisons risk contamination through overfitting (Hastie et al., 2009: 219–226). As already noted, this point has been appreciated for more than 50 years, but it is often ignored.
3. Proper performance criteria must be used that are the same across competing methods. For example, measures of fit are not appropriate if the competition claims to be testing forecasting accuracy. In addition, there are many different measures of forecasting performance (Hastie et al., 2009: Ch. 7), and the same measure should be used for all of the competitors. For example, the area under a receiver operating characteristic curve (ROC) provides very different information from that available through direct estimates of generalization (forecasting) error (Hastie et al., 2009: 314–317).
4. All of the forecasting competitors should be accurately characterized if comparisons are to be properly understood. For example, forecasting procedures are sometimes represented as state-of-the-art that actually are not. Certain forecasting procedures are sometimes characterized as machine learning that actually are not. Classification trees (CART), for instance (Breiman et al., 1984), is neither state-of-the-art nor a machine learning technique. AdaBoost (Freund and Schapire, 1997) is a machine learning procedure, but it was state-of-the-art 15 years ago. Bayesian additive regression trees (Chipman et al., 2010) can be considered state-of-the-art, but it is not formally within machine learning traditions. Random Forests (Breiman, 2001a) is state-of-the-art and a machine learning procedure.[6]
5. Many of the popular forecasting procedures have tuning parameters that researchers can use to improve forecasting accuracy.[7] In addition, sometimes researchers do not understand that in their effort to maximize forecasting accuracy they are implicitly

6. What qualifies as state-of-the-art can certainly be debated, but within sensible boundaries, there can be remarkable consensus. For example, random forests is certainly not the newest machine learning procedure, but for a wide range of applications, nothing else seems to perform better consistently. Likewise, sharp distinctions between machine learning, statistical learning, and a variety of other related procedures are increasingly difficult to defend and probably are not worth quarreling over (National Research Council, 2013: 61). Nevertheless, within somewhat fuzzy boundaries, there can be widespread agreement.

7. Tuning parameters can be set at particular values to improve the performance of a given statistical procedure (National Research Council, 2013: 70–73). In the estimation of a logistic regression, for instance, the convergence threshold of the iteratively reweighted least-squares algorithm is a tuning parameter. It needs to small enough to produce a close approximation to a maximum likelihood estimate but not so small that unnecessary iterations are performed. Another example is a decision in stepwise regression to fix the number of predictors that can be included in the final model. In forecasting settings, tuning parameters usually are chosen in service of forecasting accuracy.

tuning their procedure. Fair comparisons require that all competitors are tuned in a comparable fashion. This can be difficult because the tuning is often based on principles that can depend on the particular forecasting procedure being used.

6. All forecasting competitions are necessarily data dependent and can vary across different applications. Forecasting competitions do not reveal fundamental and invariant forecasting truths. To take a simple example, a procedure that performs poorly in small samples may be a star in large samples because its best properties only materialize asymptotically. Appropriate caveats should be attached to the results of all forecasting comparisons.
7. Performance differences across competing forecasting procedures must be thoughtfully evaluated. This will often mean a careful consideration of how a forecasting procedure will be used. A small difference in forecasting accuracy can translate into a difference of hundreds of crimes. Academic researchers may not care. But stakeholders surely do. Also, an equally important matter is taking uncertainty into account. Some apparent differences wash out in new realizations of the data. They are just chance artifacts.
8. It should go without saying, but all forecasting procedures must be implemented correctly. Ample evidence suggests that too often this is not the case (Berk, 2012).

Some Conceptual Fundamentals

We turn now to a conceptual overview of classification and forecasting. The intent is to provide a very accessible, didactic overview that can apply to a broad range of forecasting procedures used previously in criminal justice applications. Readers interested in a technical discussion should consult the references cited.

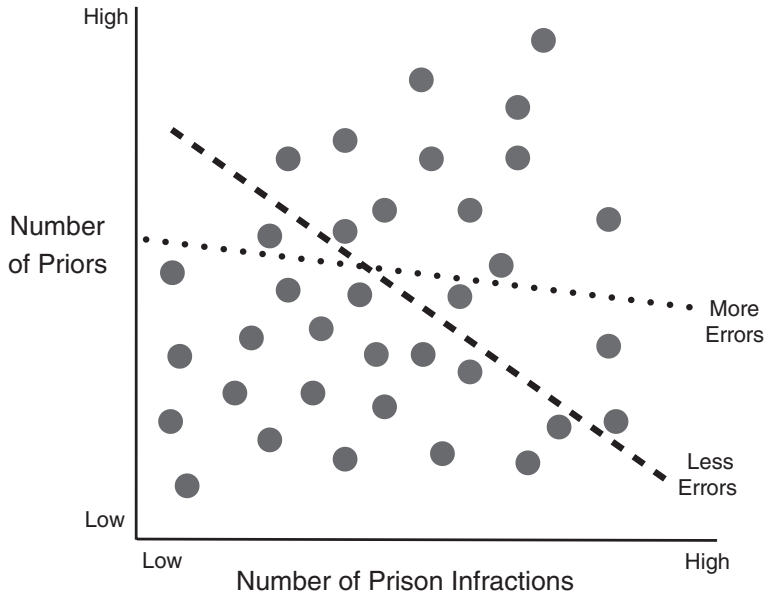
Consider the decision of whether to release an individual on parole. Since the 1920s, such decisions have often been informed by forecasts of whether a given inmate will be arrested for a new crime soon after release. The forecasts are shaped by actuarial procedures applied to information from inmates who had been released in the past. In effect, profiles are developed that can classify inmates by whether they succeeded or failed on parole. These profiles are used to forecast parole outcomes when they are not yet known. In the next few pages, we provide a basic, nontechnical overview of how this can be done. We build on a prior treatment written for criminal justice researchers (Berk, 2012) and on more formal textbook discussions as needed (Bishop, 2006; Hastie et al., 2009).

The Basic Account

Figure 1 is a simplified and initial plot illustrating how classification and forecasting can be undertaken. The red circles represent individuals who have failed on parole in the past. The blue circles represent individuals who have succeeded on parole in the past. There are two predictors in this illustration. One predictor is the number of prior arrests. The other predictor is the number of rule infractions during the most recent incarceration. Both

FIGURE 1

Two Linear Decision Boundaries in Two-Dimensional Predictor Space



can be considered "dynamic" predictors, but "static" predictors would have not materially changed the discussion. Figure 1 can be seen as a three-dimensional scatterplot.[8]

The statistical task is to impose a "decision boundary" on the two-dimensional predictor space that can be used to define two classes: those who fail and those who do not. The term "decision boundary" is used because the intent is to inform actual decisions directly.[9] Statistical procedures that partition the data into different groupings are often called "classifiers." In this instance, the partitioning should result in the fewest classification errors possible. For Figure 1, there will necessarily be two regions defined, one for failures and one for successes. Ideally, the failure region has no successes, and the success region has no failures. Usually, one has to settle for less.

8. The meanings of "dynamic predictors" and "static predictors" can depend on the context and the decision to be informed by the forecast. For example, the difference between statistic and dynamic predictors plays a key role in the fairness of parole decisions. Is it appropriate to use static predictors already employed at sentencing when later parole decisions are made? Is there a risk of unfair "double counting?" Thus, the crime that sent an individual to prison is static. Should it be also used to help inform parole decisions? In contrast, time in a prison secure housing unit (SHU) is in this context dynamic. There would be no concerns about double counting if it were employed by a parole board.

9. The underlying mathematics is shaped by the same goal.

The dotted line is one possible linear decision boundary. In the region above the dotted line, failures predominate by a count of 13 to 2. So, that region is assigned the class of "failure." In the region below the dotted line, successes predominate by a count of 17 to 5. So, that region is assigned the class of "success."

The assigned classes can be used for forecasting. When a new case is found for which a forecast is needed, that case is placed in one region or the other depending on its values for the two predictors. For example, a case with a very large number of priors and a very large number of prison infractions would be placed in the "failure" region to the upper right, and a forecast of failure would be made. A decision to impose a stiff prison sentence could follow.

The dotted decision boundary results in several classification errors. There are 2 (blue) successes classified as failures, and 3 (red) failures classified as successes. Overall, there are 5 errors for 30 cases, which means that the classification procedure is right approximately 75% of the time. In real applications, this would be considered very good performance.

The dashed line is another attempt to separate accurately the successes from the failures. Above this alternative linear decision boundary, the majority of cases once again are failures. Therefore, the class of "failure" is assigned to that region of the figure. Below the alternative linear decision boundary, the majority of cases are successes. Therefore, the class of "success" is assigned to that region of the figure. Now there are only five misclassified cases: Two blue circles are in the red region, and three red circles are in the blue region. The new boundary produces correct classifications approximately 85% of the time, and on those grounds, it is likely to be preferred to the old boundary.

As before, any cases with predictor values that place them above the decision boundary, but whose outcomes are not yet known, are forecasted to be failures. Similarly, any cases with predictor values that place them below the decision boundary, but whose outcomes are not known, are forecasted to be successes. From a classification exercise comes a forecasting procedure. The forecasts, in turn, are used to inform parole decisions.

How might one arrive at the best linear decision boundary? If the two outcomes are coded as 1 or 0, and conventional linear regression is applied using the two predictors as regressors, then one important kind of optimal linear decision boundary can be imposed on the predictor space. That line is defined by fitted values of .50. Cases with regression fitted values greater than .50 are assigned one class, and cases with regression fitted values equal to or less than .50 are assigned the other class. By minimizing the sum of squared residuals and imposing a fitted value threshold at .50, one is also minimizing the sum of the classification errors (Hastie et al., 2009: 20–22).

Alternatively, one can apply logistic regression. The same basic reasoning works. When the response is represented as the log of the odds of the category coded as 1, then there is again a linear decision boundary in "logit" units. The threshold is a logit of 0.0 (Hastie et al., 2009: 102), which in a probability metric is .50. Forecasting accuracy may be better or worse than for linear regression. Linear regression assumes that in the metric of the

1/0 outcome, relationships with the predictors are linear. Logistic regression assumes that in the metric of the 1/0 outcome, relationships with the predictors are S-shaped (i.e., the cumulative logistic function). Which of these leads to better forecasts in a given setting will usually be an empirical matter. Both functions are typically arbitrary because there will rarely be compelling subject-matter theory requiring one or the other.[10]

Building in Differential Forecasting Error Costs

To this point, all classification errors are given equal weight. A success classified as a failure counts the same as a failure classified as a success. This is why the least-squares regression minimizes the number of forecasting errors. In many criminal justice settings, the assumption of equal weights is not responsive to the preferences of stakeholders. For example, the consequences of forecasting a parole success for an individual who will fail can be far more serious than forecasting a parole failure for an individual who will be a success. The parole failure may entail a heinous crime. Failing to release an individual who would be crime free leads to increased time behind bars. Both forecasting errors are costly, but for many stakeholders, the costs to victims of a heinous crime are far greater than the costs of extra prison time. Despite whether these relative costs generally hold, an assumption that all forecasting errors have equal costs is likely to be unrealistic.[11]

And costs matter for forecasts meant to inform real decisions. Figure 2 shows why. Using the broken line as the decision boundary, two successes are incorrectly classified as failures. For this illustration, suppose that stakeholders think that the costs of "over-incarceration" are greater than the costs of crimes committed while on parole. There are reasons, therefore, to upweight the blue mistakes relative to the red mistakes. We show this in Figure 2 by making the two blue mistakes much larger. A new linear decision boundary results. Least-squares regression can be used as before. But the decision boundary shifts toward the upper right with perhaps also a change in the slope.

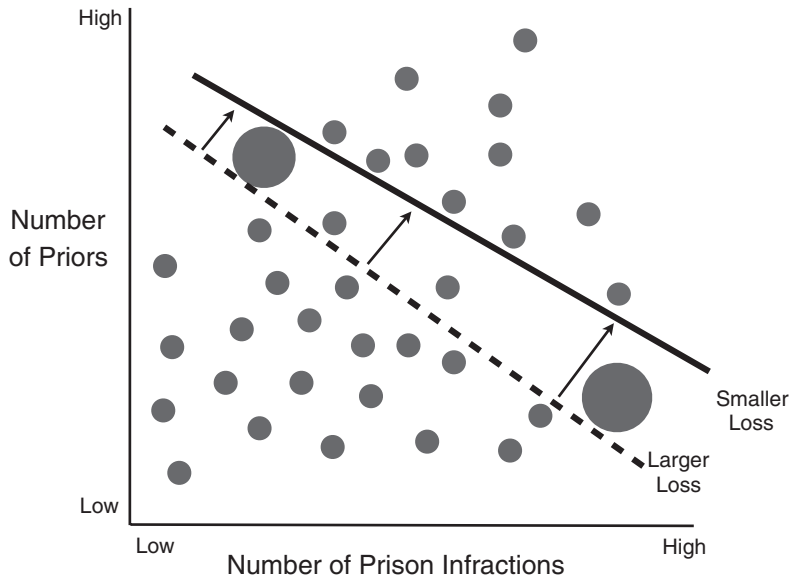
The two blue mistakes are now accurately classified as successes. They no longer count as errors. But in trade, there are now five rather than three misclassified red circles. It looks like a wash—two fewer successes are classified as failures, and two more failures are classified as successes. But it is not a wash. The new decision boundary is to be preferred because the original two blue mistakes were much more costly than the two new red mistakes.

10. Linear and quadratic discriminant function analysis has much in common with logistic regression and has been used in criminal justice risk assessments. We do not consider linear or quadratic discriminant function analysis because one must assume that the predictors have a multivariate normal distribution (Hastie et al., 2009: section 4.3). This is unrealistic for most predictors in criminal justice settings, especially when any of the predictors are categorical.

11. A more complete discussion about the role of asymmetric costs is beyond the scope of this article. An excellent treatment can be found in a special issue of the *Albany Law Review*, edited by Shawn D. Bushway (2011).

FIGURE 2

Impact of Asymmetric Costs in Two-Dimensional Predictor Space



If the new decision boundary is preferred, then many forecasts can change. In this example, cases to be forecasted as failures will need a greater number of priors and a greater number of prison infractions than previously. The increase will be larger for the number of prison infractions because the new decision boundary was shifted outward more for the infractions predictor.

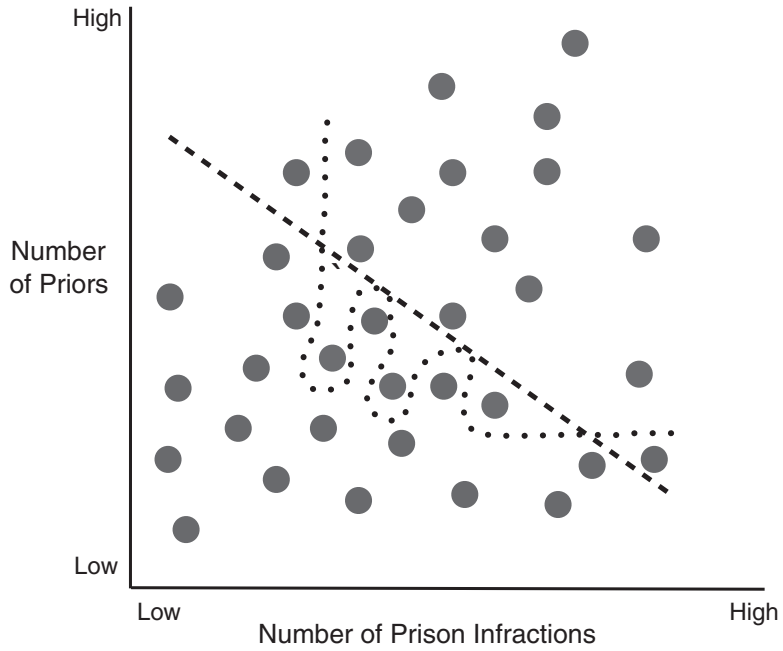
The point is that not all forecasting errors are created equal, and the relative costs of different kinds of forecasting errors should be built into any classification/forecasting procedure. To ignore this issue is to assume equal costs. And if equal costs are not consistent with stakeholder preferences, then the forecasts will not be properly responsive. Misleading forecasts can result.

Nonlinear Decision Boundaries

Why be limited to linear decision boundaries? Nonlinear boundaries can, in principle, perform better. In Figure 3, we reproduce much of Figure 1 but now with a nonlinear decision boundary shown by the dotted line. No red circles fall below the nonlinear decision boundary, and no blue circles fall above the nonlinear decision boundary. Classification is perfect. The prospects for forecasting accuracy look very promising indeed.

FIGURE 3

A Linear and Nonlinear Decision Boundary in Two-Dimensional Predictor Space



The linear decision boundary is far less complex than the nonlinear decision boundary.[12] The price for greater simplicity is more classification errors. Clearly, one's ability to classify accurately is enhanced when the decision boundary can be more complex. It is easier for the nonlinear decision boundary to respond to complicated data structures. A sensible statistical aim, therefore, can be to use predictors in a manner that allows for nonlinear decision boundaries as needed. There can be two related approaches (National Research Council, 2013: 63). For parametric procedures such as logistic regression, greater complexity can be addressed in principle by including a larger number of predictors. Transformations of predictors can help. For instance, one might include not only the age of an inmate but also some polynomial function of age. One might even break up age into a set of binary dummy variables. Statistical interactions might also be captured with products of variables. The point is that the capacity to address greater complexity needs to be built in from the

12. There seems to be no consensus on how best to define the amount of complexity. One popular approach is the degrees of freedom used to construct the decision boundary. In this example, the nonlinear decision boundary would use many more degrees of freedom than the linear decision boundary. A closely related approach is to link complexity to the "effective dimension" of the statistical procedure or, in some cases, the data itself (National Research Council, 2013: 70).

beginning or determined later in a set of very effective exploratory procedures. Also required is that the requisite predictors are included in the data set. Many would argue that these requirements cannot be met in practice.

For nonparametric procedures such as smoothing splines (Hastie et al., 2009: section 5.4), one may include as many predictors as possible, along with promising transformations, but the procedure attempts to determine the decision boundary complexity needed. At one extreme, the fitted values are a hyperplane (just as in conventional linear regression). At the other extreme, the fitted values are an interpolation between all data points. The former is much less complex than the latter. In practice, some result between these extremes is typical. In contrast to parametric methods like logistic regression, an adaptive process is used to arrive at a decision boundary—the procedure exploits information in the data to determine both the shape and the location of a decision boundary.[13] Unless a researcher is close to prescient and has the data rich enough to respond constructively, adaptive procedures start with a substantial forecasting advantage.[14]

But there is a downside to adaptively determined decision boundaries. As a greater number of degrees of freedom is used up for a given sample size, there is the real risk of increased instability in the results. Less information is available per procedure parameter. In addition, there can be overfitting in which the procedure responds to idiosyncratic features of the data. Because forecasting involves new data, not the data used to develop the decision boundary, forecasting accuracy can be disappointing. The procedure does not generalize well to new data, which is precisely what forecasting entails.

For example, an individual with a large number of priors and a large number of prison misconducts may have a high large probability of failure on parole. But a high large probability is not a certainty. If that individual does not fail, then a complex decision boundary would try to classify accurately that individual as a success. As a result, an anomalous case inconsistent with most of the data would help shape the decision boundary. When that decision boundary is then used for forecasting with data in which such anomalous cases were absent, the decision boundary would not perform as well. It would be unnecessarily complex and risk an increase in forecasting errors. Looking back at Figure 3, if any one of

13. Stepwise regression is an example of a very simple adaptive procedure within a conventional regression framework. But again, distinctions may not be sharp. When researchers respecify their models after looking at the results, the final model is shaped by data-informed induction. Some would say that the difference is that the model selection process is not built into the data analysis algorithm itself.

14. If resources allow, a parametric brute force approach may help to level the playing field. With thousands of observations and hundreds of predictors, one can in addition construct a priori many nonlinear transformations and interaction variables. In effect, the researcher tries to anticipate how an effective adaptive procedure could respond. All of the original predictors and new transformations can then be included in a single "kitchen sink" regression. The regression will likely be uninterpretable. The complexity and multicollinearity alone could be toxic. If model selection procedures are applied to simplify, then one is doing a seat-of-the-pants adaptive modeling with all of its attendant problems (Berk, Brown, and Zhao, 2010). Why settle for a brute force approximation to the desired procedure? An example can be found in the recent paper by Tollenaar and van der Heijden (2013).

the three red circles had as little as one or two more prison infractions or priors, then the red circle would have fallen above the linear decision boundary, and one of the fingers in the nonlinear decision boundary would not have been constructed.

Useful responses to overfitting often are called "shrinkage" or "regularization." The intent is to reduce the instability. With smoothing splines, for instance, the fitting function is penalized for increases in complexity (Hastie et al., 2009: section 5.4). In a least-squares context, the residual sum of squares is increased so that what might be the smallest sum of squared residuals no longer is the smallest. A residual sum of squares that starts out being larger, but has a smaller penalty because of less complexity, can be the preferred minimizer. In other words, a price is put on complexity that does not substantially improve the fit.

Another approach, called bagging (Breiman, 1996), capitalizes on a large number of random samples with replacement from the data on hand. A classification procedure is applied to each sample, and the results are averaged across samples. One important consequence is that idiosyncratic results tend to cancel out.

Finally, in this illustration, the two predictors have substantive interpretations. In general, parolees with a great number of prior arrests and a greater number of prison infractions are more likely to fail on parole. However, any substantive insights are a bonus. The primary goal is to classify accurately because that can lead to the most accurate forecasts. With respect to that goal, the two predictors could as well be longitude and latitude. This allows for the possibility of using "black box" classification procedures, for which no apologies need be made. One does not have to rely a "structural model" when forecasting is the primary motive. Indeed, the requirement of a structural model can undercut forecasting accuracy (Breiman, 2001b). Two different masters are being served.

In summary, when forecasting accuracy is the primary goal, parametric approaches such as logistic regression can, in principle, perform as well as nonparametric approaches when the best decision boundary is relatively simple, and when the predictors required by the correct model are available in their proper form. When the best decision boundary is complex and/or the requisite predictors are not all available, nonparametric procedures will forecast more accurately, often substantially more accurately.

Enter Machine Learning

Where does machine learning come in? Machine learning, sometimes called "statistical learning," can be viewed as a special form of nonparametric regression. The goal can be to find the "right model." But when machine learning is used strictly as a forecasting procedure, the connections to conventional regression models become very distant indeed.

As will soon be discussed in more detail, there is no structural model even in principle. The transition to machine learning can confer a number of important benefits, some of which are not readily available otherwise. The benefits are as follows:

1. One is not limited to classifiers able to forecast one of two outcome categories. In some recent applications, for instance, parole outcomes are forecasted for three classes: an arrest for a violent crime, an arrest for a crime that is not violent, and no arrest (Berk, Barnes, Ahlman, and Kurtz, 2010). Increasingly, criminal justice agencies want to forecast more than the binary outcome of any arrest versus no arrest (Berk, 2012). The kind of arrest really matters. In particular, arrests for crimes of violence are distinguished from other kinds of arrests.
2. Forecasting errors that do not have equal costs can be introduced into the procedure at the beginning so that all of the results properly represent the preferences of stakeholders (Berk, 2011).
3. Regularization is often built directly into the procedure to increase forecasting accuracy (Hastie et al., 2009: Ch. 5, section 8.7).
4. Highly unbalanced distributions for the classes to be forecasted create no special problems as long as the rare outcomes are important enough to be given extra weight in the analysis. For example, in some recent work for individuals primarily on probation, the outcome classes to be forecasted included a class for homicide or attempted homicide, which represented only approximately 2% of the outcomes (Berk, Sherman, Barnes, Kurtz, and Ahlman, 2009; Berk, 2009).
5. Some procedures work well and in a principled manner with an enormous number of predictors and even when there are more predictors than cases (Hastie et al., 2009: Ch. 15).

The Forecasting Contestants

We will compare the forecasting performance of three different classifiers: logistic regression, random forests, and stochastic gradient boosting. Logistic regression has represented business as usual over the past 50 years. It is a special case of the generalized linear model and is very familiar to criminal justice researchers. Random forests (Breiman, 2001a) and stochastic gradient boosting (Friedman, 2002) represent true machine learning procedures based on ensembles of classification trees. Both are nonparametric, rest on solid mathematical foundations, and have been widely battle tested. All of the evidence to date indicates that they can perform well in criminal justice applications (Berk, 2013). All three are worthy competitors.[15]

15. There are other worthy competitors such as Bayesian neural nets (Hastie et al., 2009: section 11.9) and support vector machines (Hastie et al., 2009: Ch. 12). They are not considered in this article for lack of space and the need to introduce a substantial amount of new technical material. Suffice it to say that they too are well equipped to address complex decision boundaries and should have foresting skill roughly comparable to random forests and stochastic gradient boosting. But comparisons are difficult because a new suite of tuning parameters is introduced.

Forecasting Class Membership with Logistic Regression

Logistic regression, sometimes called binomial regression, is a special case of the generalized linear model. As such, it is meant to represent how nature generated the data—it is an algebraic translation of subject-matter theory. In that sense, it is a "structural model," and forecasting can be little more than an afterthought. Nevertheless, if the theory is correct and its algebraic representation is consistent with the theory, then accurate forecasting can result.

Forecasting is undertaken through the regression's fitted values. These can either be in logit (i.e., log odds) units or probability units. Researchers typically use the probabilities when forecasting. To get from the probabilities to a forecasted class, a single threshold must be applied. For example, it is common to use a threshold of .50. Probabilities greater than .50 are assigned one outcome class (e.g., failed on parole). Probabilities less than or equal to .50 are assigned the other outcome class (e.g., succeeded on parole). The threshold of .50 implies that the costs of false negatives and false positives are the same. As already noted, they are usually not the same. Suppose a "positive" is a person who commits a violent crime. Suppose a "negative" is a person who does not commit a violent crime. It follows that if false negatives are three times more costly than false positives, then one should use a threshold of .25. Cases with predicted probabilities greater than .25 are forecasted to be violent offenders. Cases with predicted probabilities equal to or less than .25 are forecasted to be nonviolent offenders. It is three times easier for a person to be forecasted a violent offender than a nonviolent offender or no offender at all ($.75 / .25 = 3$).

Altering the threshold only affects the step from probabilities to classes. All of the other logistic regression output is computed under the assumption that false negatives have the same costs as false positives. In particular, the logistic regression coefficients would almost surely be different had the actual relative costs of false negatives and false positives been properly taken into account. It can be a serious error, more generally, to use the regression coefficients as weights for constructing risk assessment instruments.[16]

Finally, logistic regression can only be used for binary outcomes. These days, criminal justice stakeholders often want much more—they want to forecast different kinds of crimes. As already noted, in some applications, the intent is to work with three crime categories: arrests for violent crimes, arrests for crimes that are not violent, and no arrest at all (Berk, Barnes, et al., 2010). In the context of probation supervision, one motivation is to move

16. They are in logits units, not probability units. If one follows the common practice of exponentiating the regression coefficients and intercept, one is now working in odds units. In addition, the regression coefficients and intercept are then multipliers and do not represent additive weights. If the intent is to obtain risk factor weights in probability units, one must go back to the original nonlinear logistic model. But because of the nonlinear functional form, there is not one weight for each risk factor—there is a limitless number. So that strategy fails too. It is also possible to ignore the regression coefficients and weight risk factors by simply "assigning weights or 'points'" (VanNostrand and Rose, 2009: 9). The statistical foundation for that approach is obscure.

supervisory resources from individuals who do not threaten public safety to individuals who do, a strategy that has been shown to work well (Berk, Barnes, et al., 2010). When there are more than two outcome classes, multinomial logistic regression may be an option, but there are a number of unresolved issues about how best to go from predicted probabilities for each class to the classes themselves.

Random Forests

A random forest is an ensemble of classification trees. The classification trees are an intermediate product used because they fit the data adaptively. They have no stand-alone role, and in the end, they are effectively invisible. They disappear into a machine learning black box through the following algorithm:

1. A random sample of size N is drawn with replacement from a "training" data set. Observations not selected are retained as the "out-of-bag" (OOB) data to later serve as "test data." On average, approximately one third of the data will be OOB. The growing process for the first classification tree then begins.
2. A small sample of predictors is randomly drawn (e.g., three predictors).
3. After selecting the best split as usual from among the randomly selected predictors, the first partition is determined. Then two subsets of the data together maximize the improvement in the Gini index.
4. Steps 2 and 3 are repeated for all later partitions until the fit does not improve or the observations are spread too thinly over terminal nodes.
5. The Bayes classifier is applied to each terminal node to assign a class. The class for each terminal node is determined by the class in the node that has the largest number of cases.
6. The OOB data are "dropped down" the classification tree. Each observation is labeled with the class assigned to the terminal node in which it lands. The result is the predicted class for each observation in the OOB data for that tree.
7. Steps 1 through 6 are repeated many times to produce a large number of classification trees. There are often 500 trees or more.
8. For each observation, the class assigned is determined by "vote" over all trees in which that observation is OOB. The class with the most votes is chosen. That class can be used for forecasting when the predictor values are known but the outcome class is not.

The adaptive nature of classification trees helps to reduce bias. In addition to the predictors used as inputs, "derived" predictors are constructed as needed. The sampling of training data and predictors serves as a form of regularization that can improve the stability of class assignments and help make those assignments more independent over trees. Averaging over trees enhances both results. Finally, the use of OOB data helps to prevent overtfitting. In the end, random forests does not overfit as the number of trees in the random forest

increases. A formal proof can be found in Breiman's (2001a) seminal paper on random forests.

There are several ways to introduce asymmetric costs. Perhaps the best way is to employ stratified sampling in step 1. There is one stratum for each outcome class. The sample sizes for each stratum are determined so that some outcome classes are oversampled and some are undersampled. In effect, the oversampled classes are given more weight as each tree is grown, which in turn will affect the balance of false negatives to false positives. That balance captures relative costs. For example, if there are ten false positives for every false negative, then false negatives are necessarily ten times more costly than false positives.

In addition to "confusion tables" in which forecasted outcomes from OOB data are cross-tabulated with the observed outcomes, there are measures of the contribution to forecasting accuracy for each predictor and plots that show the way in which each predictor is related to the response, holding all other predictors constant. The details are beyond the scope of this article, but some examples are provided later. (See, for example, Berk, 2008, for the details.)

Stochastic Gradient Boosting

Stochastic gradient boosting proceeds by applying a "weak learner" repeatedly to the data. After each pass through the data, all observations are reweighted, giving more weight to observations that were more difficult to classify accurately. The fitted values from each pass are used to update earlier fitted values. The weak learner is "boosted" to perform as a strong learner. The following is an outline of the algorithm for a binary outcome coded numerically as "1" for failure on parole or "0" for success on parole:

1. The algorithm is initialized with fitted values for the binary outcome. The overall proportion of cases that fail is a popular choice.
2. A random sample without replacement is drawn from the training data with a sample size of about half the sample size of the training data.[17]
3. The "negative gradient" (sometimes called the "pseudo-residuals") is computed. Just like with usual residuals, each fitted value is subtracted from its corresponding observed value of 1 or 0. The residual is a quantitative outcome variable within the algorithm: $(1 - p)$ or $-p$, where p is the overall proportion coded as "1."
4. Using the randomly selected observations, a regression tree is grown to fit the pseudo-residuals.[18]
5. The conditional mean in each terminal node is the estimate of the probability of failure.

17. The goal is much the same as the sampling with replacement used in random forests. A smaller sample is adequate because when sampling without replacement, no case is selected more than once; there are no "duplicates."

18. The procedure is much the same as for classification trees, but the fitting criterion is the error sum of squares or a closely related measure of fit.

6. The fitted values are updated by adding to the existing fitted values the new fitted values weighted to get the best fit.
7. Steps 3 through 6 are repeated until the fitted values no longer improve by a meaningful amount. The number of passes can in practice be quite large (e.g., 10,000), but unlike random forests, stochastic gradient boosting can overfit. Some care is needed because there is formally no convergence.
8. The fitted probability estimates can be transformed into outcome classes just as they were for logistic regression.
9. When forecasts are needed for new cases, they are constructed from the aggregated fitted values and their relationships with the predictors.

Like random forests, stochastic gradient boosting capitalizes on random samples of the training data, adaptive fitting tree by tree, and aggregation over trees. However, asymmetric costs can only be introduced at the end when probabilities are transformed into classes. Experience to date suggests that it can forecast about as well as random forests.

A Simulation

Logistic regression can forecast well when it is able to capture the data structure. However, logistic regression is not adaptive and depends on the researcher to specify an effective model. Important nonlinearities and interaction effects must be anticipated and included using the available predictors. If the researcher lacks the requisite insight or data, then logistic regression will necessarily stumble. In contrast, adaptive procedures such as random forests or stochastic gradient boosting can shine because both algorithms are designed to search for structure with each pass through the data.

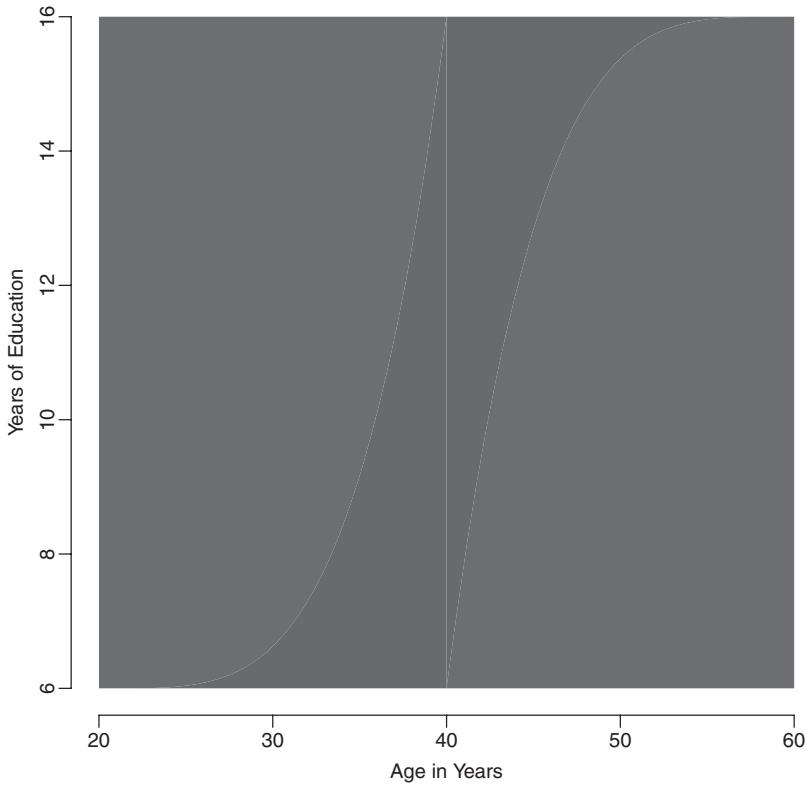
Figure 4 shows a fictitious data set constructed to illustrate when logistic regression will perform poorly and when random forests or stochastic gradient boosting will perform well.[19] It is by intent a worst-case scenario for logistic regression and is not meant to represent in general the relative merits of the forecasting competitors. We are trying to address why nonparametric methods can forecast better than parametric methods. The exercise is didactic.

There are 100,000 observations. The outcome is binary. Red is coded 1, and blue is coded 0. There are two predictors. The two-dimensional predictor space contains a blue area that is homogeneously successes and two red areas that are homogeneously failures. The graphical conventions are no different from those used for the earlier figures except that the colored circles for individual observations are replaced by solid colors for different regions. It is as if we have printed a very large number of overlapping red circles and a very

19. The lessons learned can be applied far beyond logistic regression to any parametric regression approach. The lessons also apply to a wide range of functions that have clear structures but are very difficult for parametric regression models to capture.

FIGURE 4

A Very Challenging Classification Example



large number of overlapping blue circles. However, the data structure is far more complex because the blue region has red regions to its left and its right. Complex data structures of this sort are routinely analyzed in the classification literature (Hastie et al., 2009) but usually with many more than two predictors so that visualizations such as Figure 4 are unavailable. Any researcher trying to arrive at the correct parametric model from an examination of a scatter plot would necessarily be flying blind.

The surface was built by first drawing one predictor from a uniform distribution. The second predictor was constructed as a power function of the first. Then the predictor space was partitioned to show an interaction effect: Both predictors had to be high or low for the area to be red. That is, there are nonlinear effects and an interaction effect. Because each of the three regions is perfectly homogeneous, the data provide a clear and compelling signal that a good classifier should be able to detect accurately.

After the fact, one might overlay the following subject-matter account. The outcome is whether a parolee finds employment. The blue area contains successes, and the red area

contains failures. On the horizontal axis is age in years. The young and the old do not do well. The vertical axis is years of education. The association is not strong, but parolees with a lot of education or very little do slightly better. In addition, when the educational level is higher, the best ages for finding work are older.[20]

Why might such patterns occur? The kinds of positions for which parolees apply and the kinds of employers who would hire them represent a very limited subset of all jobs. By and large, the positions will involve physical labor for which not much experience or skill is required. The pay will be low and the work will be hard. Younger parolees may not be inclined to seek such positions, and older parolees may be incapable of doing the work. Education may be largely irrelevant for most of the jobs a parolee will seek. But those who have very little education may correctly target their job search only for menial positions. Those with more education may correctly understand that they have a wider range of employment options. Finally, having more education may give some older workers, who would have difficulty working at demanding menial jobs, the chance to take entry-level white-collar positions (e.g., taking orders and making change at fast food restaurants).[21]

This post hoc account may well be wrong, perhaps very wrong. The intent is to provide a less abstract setting in which to think about each contestant's performance. By itself, the story has no impact whatsoever on how well a given classifier performs. Any good classifier should forecast with near perfect accuracy. Unlike in real data, there is no noise.

When logistic regression is used, both regression coefficients are virtually zero.[22] Logistic regression is unable to extract any useful information from the two predictors. All that remains is the intercept, which is effectively the logit of the outcome variable's proportion of reds (i.e., .80). The distribution of the predicted probabilities ranges from .7958 to .8022. The predicted probabilities have almost no variability.

For didactic purposes and with no important loss of generality, we assume that the costs of false negatives are the same as the costs of false positives. The corresponding threshold of .50 is applied. It follows that forecasting error is minimized by always predicting red. Twenty percent of the time the forecast would be wrong. The true reds would be forecasted with 100% accuracy, and the true blues would be forecasted with 0% accuracy. Table 1 shows the results.[23]

20. Plots of this sort may be unfamiliar and at first difficult to interpret. For the main effects, one has to do an eyeball integration over the variable whose role is not being described. For example, to gauge the marginal association between age and employment, one must consider vertical slices of the data and what fraction of each area is blue. Similar reasoning applies to years of education, but now the slices are horizontal.

21. Some preliminary analyses we are doing for the program "Ready, Willing & Able," supported by the Doe Fund, are consistent with this account.

22. The two regression coefficients are –.03 and –.01. Even with 100,000 observations, one cannot reject the null hypothesis of 0.0 for either. In an odds multiplier metric, both coefficients are very close to 1.0.

23. Other thresholds would not change the performance of logistic regression. A threshold a very little bit below .80 would allow some blues to be correctly forecasted. The price would be a commensurate

TABLE 1

Logistic Regression Confusion Table Using Simulated Test Data

	Predict Blue	Predict Red	Model Error
Actual Blue	0	20,078	1.0
Actual Red	0	79,922	0.0

TABLE 2

Logistic Regression with Interaction Confusion Table Using Simulated Test Data

	Predict Blue	Predict Red	Model Error
Actual Blue	0	20,078	1.0
Actual Red	0	79,922	0.0

Suppose a researcher is astute enough to include in advance the product of the two predictors to capture an interaction effect. Our reading of criminal justice forecasting applications is that such interactions are rarely used, but it is useful to see how logistic regression performs when given an especially good opportunity to deliver.

Table 2 shows the results. Although there are now nonzero regression coefficients for all three regressors, there are still no predicted probabilities smaller than .5. As before, forecasting error is minimized by always forecasting red. Nevertheless, there is some meaningful information in the predicted probabilities, and with cost ratios that weight forecasting errors for blue cases more heavily than for red cases, some blue cases will be correctly predicted.[24] For example, if a cost ratio of 4 to 1 is used, then actual blues and actual reds are both correctly forecasted approximately two thirds of the time. That may seem quite good, but for these data, the appropriate target is perfection.

How does an adaptive machine learning procedure perform? For illustrative purposes, we take random forests as our machine learning champion.[25] Table 3 shows the results for random forests assuming equal costs. With respect to the cost ratio, we are

increase in reds forecasted incorrectly. Virtually no predictive information from the predictors is being used. The predictors might as well be ignored.

24. The predicted probabilities now range from .5333 to .9384.

25. We used the procedure random Forest in R, originally written by Leo Breiman and Adele Cutler and ported to R by Andy Liaw and Matthew Wiener. To the best of our knowledge, there is no implementation of random forests in any of the popular statistical packages such as SPSS (SPSS Corporation, Chicago, IL), STATA (StataCorp, College Station, TX), or SAS (SAS Institute, Inc., Cary, NC). Salford Systems (Salford Systems, San Diego, CA) has a procedure they call random forests, but the source code is proprietary, and it is difficult to know exactly what is being done. Also, according to the current Salford Systems Web site, the available version of random forests will not run on a Mac computer.

TABLE 3

Random Forests Confusion Table Using Simulated Test Data

	Predict Blue	Predict Red	Model Error
Actual Blue	19,975	102	0.005
Actual Red	92	79,830	0.001

comparing apples to apples. The same two predictors are used, but there is no product variable for an interaction effect. The researcher using random forests is not allowed to be as clever as the researcher using logistic regression—random forests begins with a model specification disadvantage. Still, random forests is just about perfect. Given either outcome, random forests forecasts correctly more than 99% of the time. The failure to be literally perfect results from randomness in the random forests algorithm itself.

The implications of this forecasting contest are clear. When the data structure is complex, machine learning procedures can perform very well. An adaptive process that "learns" from data can be very effective. This is precisely what the large literature in statistics and computer science has said. Logistic regression and other parametric forecasting procedures will not perform as well unless the researcher can construct a parametric model that captures all of the significant features of the data structure. As already noted, this can be a daunting task.

An Empirical Example

We turn now to analyses of real data. The data set was selected to be typical of those recently used in parole or probation settings. Recall, however, that it is very difficult with real data to arrive at results that are broadly generalizable.

Forecasting Arrests for Serious Crimes

The data address how well parolees manage under supervision. There are 20,000 observations in the training data and 5,000 observations in the test data. We consider whether an individual is arrested for a serious crime within 2 years of release on probation. Serious crimes include murder, attempted murder, rape, aggravated assault, and arson. Approximately 13% fail by this definition. Such crimes are of widespread concern. Static and dynamic predictors include the following:

1. Date of birth
2. Number of violent priors as an adult
3. Earliest age for a charge as an adult
4. Total number of priors as an adult
5. Earliest age for a charge as a juvenile

TABLE 4

Logistic Regression Test Data Confusion Table for Serious Crime

	Predict Fail	Predict No Fail	Model Error
Actual Fail	378	302	0.444
Actual No Fail	1,385	2,935	0.321

TABLE 5

Random Forests Test Data Confusion Table for Serious Crime

	Predict Fail	Predict No Fail	Model Error
Actual Fail	427	253	0.372
Actual No Fail	1,196	3,124	0.277

6. Total number of priors as a juvenile
7. Number of charges for drug crimes as an adult
8. Number of sex crime priors as an adult

There is nothing special about these predictors. They represent the usual kinds of information that are routinely available on parolees when they begin their supervision. From experience, they can make important contributions to forecasting accuracy (Berk, 2012).

We first apply logistic regression to the training data. A threshold of .135 is imposed on the predicted probabilities to arrive empirically at a 5-to-1 cost ratio of false negatives to false positives. Table 4 is the confusion table that results when the model is applied to test data. From the column on the far right, approximately 44% of the true failures are misclassified and approximately 32% of the true successes are misclassified. The forecasting accuracy is within the range of recent studies with similar data (Berk, 2012) and could be useful for decision makers.

Table 5 is the confusion table for random forests using the test data. The procedure was tuned to also arrive at a cost ratio of approximately 5 to 1 for false negatives versus false positives. From the column on the far right, approximately 37% of those who actually fail are incorrectly identified and approximately 28% of those who actually do not fail are incorrectly identified. The forecasting accuracy for random forests seems to be superior.

Table 6 is the confusion table for stochastic gradient boosting using the test data.[26] A threshold of .13 was used on the predicted probabilities from the training data to arrive

26. We used the R procedure gbm, written by Greg Ridgeway. Several tuning parameters can make a difference, and we are not certain that the comparisons are fully fair. To the best of our knowledge, there is no implementation of stochastic gradient boosting in any of the popular statistical packages.

TABLE 6

Stochastic Gradient Boosting Test Data Confusion Table for Serious Crime

	Predict Fail	Predict No Fail	Model Error
Actual Fail	396	284	0.418
Actual No Fail	1,361	2,459	0.315

empirically at a cost ratio of approximately 5 to 1. From the column on the far right, approximately 42% of those who actually fail are incorrectly identified and approximately 32% of those who actually do not fail are incorrectly identified. Stochastic gradient boosting does a little better than logistic regression when forecasting failures but only slightly better when forecasting successes.

It seems that across the three tables, random forests performs better than logistic regression and stochastic gradient boosting. This is consistent with published studies (Berk, 2012) and a decision boundary that is relatively simple. But one must not overstate what is learned from the comparisons we report. It is difficult to guarantee that after tuning, one is necessarily comparing apples to apples. We have tried to ensure that for all practical purposes, the false-negative to false-positive cost ratios are the same for all three procedures. But the cost ratios are not identical, and it is essentially impossible to make them so. The test data and training data are different random splits of the available data set. Tuning done on the training data will carry over a bit differently to the test data, depending on the forecasting procedure. Moreover, each procedure was tuned with its own special set of tuning parameters. There is no guarantee that the results are fully comparable. Indeed, it is not even clear how to define such a thing.

Another important issue is whether the differences are large enough to matter. As already explained, that judgment depends on the application. For example, the agency from which these data were obtained supervises approximately 40,000 individuals on probation each year. Approximately 5,000 of these individuals are arrested for a serious crime within 24 months, most within less than a year. For failures, the difference of approximately 7% between the accuracy of logistic regression compared to random forests translates into approximately 350 serious crimes. Roughly 50 of those will be homicides or attempted homicides, the perpetrator of which could be identified in advance by random forests but not by logistic regression. In this instance, stakeholders found the practical difference in forecasting accuracy dramatic.

If one is looking for firm conclusions about forecasting accuracy from our results and others, it is almost certain that properly applied, random forests will always do at least as well as logistic regression and much of the time meaningfully better. Stochastic gradient boosting will do at least as well as logistic regression, but it is somewhat less likely to dominate it.

There are several other reasons why random forests should be the forecasting method of choice, given currently available alternatives. For this illustration, the success category included individuals who were arrested for crimes not defined locally as "serious" and individuals not arrested at all. This is, of course, less than ideal. In fact, one goal of the supervising agency was to identify low-risk offenders who could be supervised less intensively with no increased risk to public safety. Resources recaptured from the low-risk offenders could then be allocated to the high-risk offenders. To address this policy preference, random forests was applied using three outcome categories: an arrest for a serious crime, an arrest for a crime that was not serious, and no arrest at all. Three outcome classes are not an option for logistic regression. The forecasting accuracy for the low-risk offenders was very good, implying that approximately half of the agency's case load could be minimally supervised. A reorganization of the supervisory practices followed, and a subsequent evaluation showed that rearrest rates for the low-risk individuals were not higher than under the previous, more intensive supervision regimes (Berk, Barnes, et al., 2010).

Random forests also provides output that can help explain how the forecasting works in practice. Recall that logistic regression coefficients, for instance, are estimated under equal costs and can be misleading if the costs of false negatives and false positives differ. In place of regression coefficients, random forests provides estimates of each predictor's contribution to forecasting accuracy that incorporate asymmetric costs. How this is done is beyond the scope of the article, but it is explained in many published papers and texts (e.g., Breiman, 2001a). Figure 5 is an example of the output that easily can be obtained.

Date of birth makes the largest contribution to forecasting accuracy for those who are arrested for a violent crime. The value of a little over .08 means that if date of birth is not allowed to contribute to forecasting accuracy, model error increases from approximately .37 in Table 5 to .45. The contributions of all other variables are smaller, with sexual priors contributing little or nothing.

Stakeholders have found this kind of information very useful. However, forecasting accuracy does not identify risk factors in the usual sense. A given predictor will often be transformed in many different ways, including as a component of interaction effects. All of these roles are combined when contribution to forecasting accuracy is computed. One has in test data the net association between a given machine learning input and the outcome being forecasted. Currently, there is no way to represent each role separately.[27]

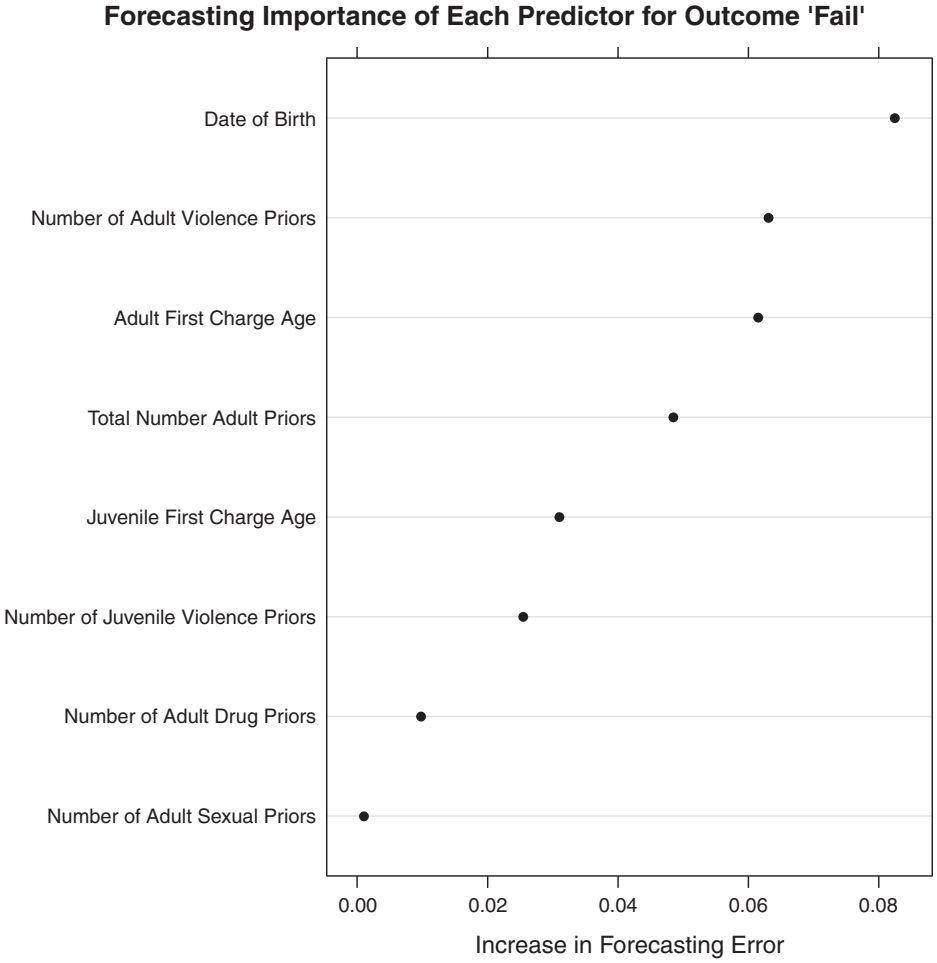
"Partial response plots" show how each predictor is related to each outcome class, with all other predictors held constant. Again, the details are beyond the scope of this article but easily found elsewhere (e.g., Berk, 2012). Figure 6 is an example.

The predictor is the age at which the first arrest as an adult occurred. The response is being subsequently arrested for a serious crime while on probation. Units on the vertical

27. Recall that each tree in the random forest can transform each predictor differently. If there are, for instance, 500 trees, then a given variable may be transformed in 500 different ways.

FIGURE 5

Random Forests Variable Importance Plot



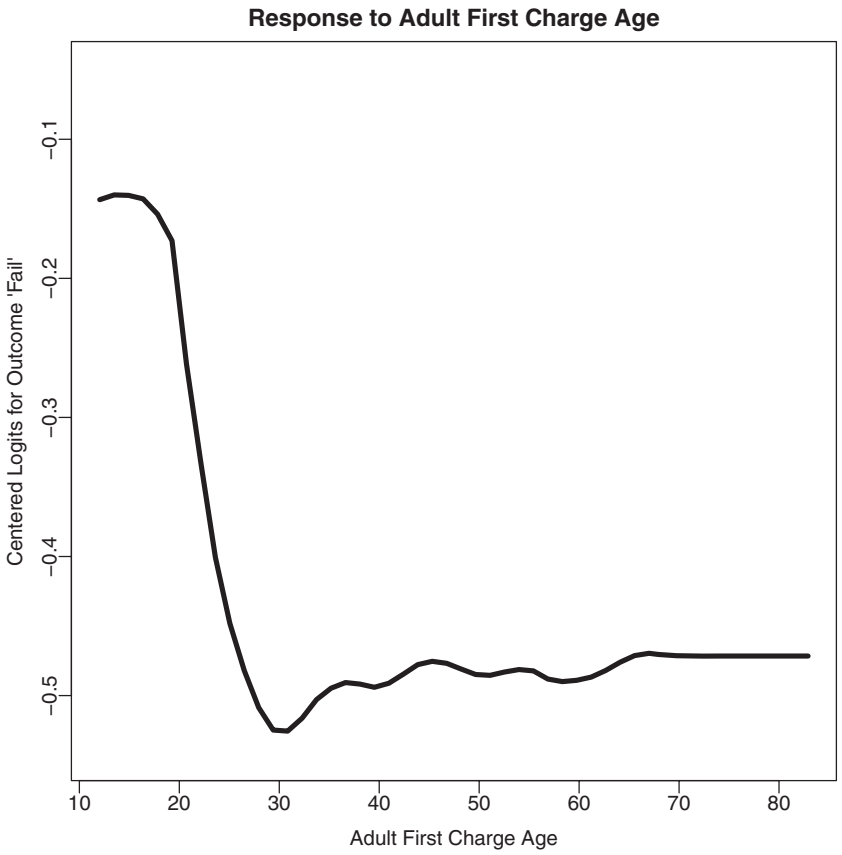
axis are centered logits. The details need not concern us here—movement in the vertical direction means that the probability of failure increases.

The figure shows that the chances of an arrest for a serious crime are high for parolees whose first arrest as an adult occurred at a very young age. Starting in the late teens, those chances decline rapidly. For parolees whose first arrest occurred after age 30, increases beyond that in age of first arrest do not matter. In random forests, partial response plots are available for all predictors. For categorical predictors, the plots are bar charts.

Just as with contributions to forecasting accuracy, partial plots also do not identify risk factors in the usual sense. Each plot captures an average across trees in the forest and

FIGURE 6

Random Forests Partial Response Plot for "Adult First Charge Age"



across each term in which that predictor is used. So age at first arrest is related to failure on parole in the manner shown in Figure 6, but all sorts of potentially important relationships involving the variables (e.g., interaction effects with gender) are masked.

Some Implications for Use

The forecasting output from machine learning classifiers is a forecast for given individuals and the sorts of descriptive output just discussed. Decision makers "drop" an individual's predictor values into an algorithm, and a forecast is computed in real time. There is no explicit use of risk factors, whether weighted or not. Thus, the algorithm must be live on some computer when forecasts are needed. Ideally, that computer is part of a network connected electronically to databases containing the predictor values. Then, a decision maker may need only to enter an individual's unique ID number for appropriate predictor

values to be properly downloaded into the machine learning algorithm. Experience to date indicates that such arrangements are well within the capabilities of information technology personnel in many criminal justice settings (Berk, 2012).

Conclusions

Complex decision boundaries pose a significant challenge for logistic regression or any other parametric classifier. To forecast well, a researcher must understand the nature of the complexity, be able to translate that knowledge properly into an algebraic expression, and then have the data to construct an appropriate model. These requirements are daunting for criminal justice applications.

In contrast, adaptive machine learning procedures have the capacity to discover empirically patterns in the data and construct suitably complex decision boundaries. The requirements are a conventional menu of predictors and a large enough sample to exploit them. The tree-based machine learning procedures we have reviewed can then perform well and have several other important assets that logistic regression lacks: the capacity for outcome categories with more than two classes, a natural way to build in the asymmetric costs of forecasting errors, and a variety of instructive output that builds in asymmetric costs.

In practice, performance differences between logistic regression and most machine learning procedures can be small if the true decision boundary is simple. But how would one know? If logistic regression is used because a simple decision boundary is incorrectly assumed, then substantial forecasting accuracy can be forfeited. In criminal justice settings where real lives can be at stake, the consequences could be significant. Why take the risk?

References

Andrews, Don A., James A. Bonta, and J. S. Wormith. 2006. The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, January: 7–24.

Berk, Richard A. 2008. *Statistical Learning from a Regression Perspective*. New York: Springer.

Berk, Richard A. 2009. The role of race in forecasts of violent crime. *Race and Social Problems*, 1: 231–242.

Berk, Richard A. 2011. Asymmetric loss functions for forecasting in criminal justice settings. *Journal of Quantitative Criminology*, 27: 107–123.

Berk, Richard A. 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. New York: Springer.

Berk, Richard A. 2013. Algorithmic criminology. *Security Informatics*, 2(5): 1–14.

Berk, Richard A., Geoffrey Barnes, Lindsay Ahlman, and Ellen Kurtz. 2010. When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, 6: 191–208.

Berk, Richard A., Lawrence Brown, and Linda Zhao. 2010. Statistical inference after model selection. *Journal of Quantitative Criminology*, 26: 217–236.

Berk, Richard A., Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz, and Lindsay Ahlman. 2009. Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *Journal of the Royal Statistics Society—Series A*, 172 (part I): 191–211.

Berkson, Joseph. 1951. Why I prefer logits to probits. *Biometrics*, 7: 327–339.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Borden, Howard G. 1928. Factors predicting parole success. *Journal of the American Institute of Criminal Law and Criminology*, 19: 328–336.

Box, George E. P. and Gwilym M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

Breiman, Leo. 1996. Bagging predictors. *Machine Learning*, 26: 123–140.

Breiman, Leo. 2001a. Random forests. *Machine Learning*, 45: 5–32.

Breiman, Leo. 2001b. Statistical modeling: The two cultures. *Statistical Science*, 16: 199–231.

Breiman, Leo, Jerome H. Friedman, R. A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth.

Burgess, E. W. 1928. Factors determining success or failure on parole. In (Andrew Alexander Bruce, Albert James Harno, Ernest Watson Burgess, and John Landesco, eds.), *The Working of the Indeterminate Sentence Law and the Parole System in Illinois*. Springfield, IL: State Board of Parole.

Bushway, Shawn D. 2011. Estimating empirical Blackstone ratios in two settings: Murder cases and hiring. *Albany Law Review*, 74: 1087–1104.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge, U.K.: Cambridge University Press.

Casey, Pamela M., Roger K. Warren, and Jennifer K. Elek. 2011. *Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group*. Williamsburg, VA: National Center for State Courts. Retrieved from ncsconline.org/.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4: 266–298.

Dumbill, Edd. 2013. Making sense of big data. *Big Data*, 1: 1–2.

Farrington, David P. and Roger Tarling. 2003. *Prediction in Criminology*. Albany: SUNY Press.

Feeley, Malcolm M. and Jonathan Simon. 1994. Actuarial justice: The emerging new criminal law. In (Dorothy Nelken, ed.), *The Futures of Criminology*. London, U.K.: Sage.

Freedman, David A. 2005. *Statistical Models: Theory and Practice*. Cambridge, U.K.: Cambridge University Press.

Freund, Yoav and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and system Sciences*, 55: 119–139.

Friedman, Jerome H. 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38: 367–378.

Gottfredson, Stephen D. and Laura J. Moriarty. 2006. Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, 52: 178–200.

Harcourt, Bernard E. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. New York: Springer.

Hyatt, Jordan M., Steven L. Chanenson, and Mark H. Bergstrom. 2011. Reform in motion: The promise and profiles of incorporating risk assessments and cost-benefit analysis into Pennsylvania sentencing. *Duquesne Law Review*, 49: 707–749.

Liu, Yuan Y., Min Yang, Malcolm Ramsay, Xiao S. Li, and Jeremy W. Coid. 2011. A comparison of logistic regression, classification and regression trees, and neutral networks model in predicting violent re-offending. *Journal of Quantitative Criminology*, 27: 547–573.

Kleiman, Matthew, Brian J. Ostrom, and Fred L. Cheeman II. 2007. Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, 53: 1–27.

Messinger, Sheldon L. and Richard A. Berk. 1987. Dangerous people: A review of the NAS report on career criminals. *Criminology*, 25: 767–781.

National Research Council. 2013. *Frontiers for Massive Data Analysis*. Washington, DC: National Academies Press.

Ohlin, Lloyd E. and Otis Dudley Duncan. 1949. The efficiency of prediction in criminology. *American Journal of Sociology*, 54: 441–452.

Ohlin, Lloyd E. and Richard A. Lawrence. 1952. A comparison of alternative methods of parole prediction. *American Sociological Review*, 17: 268–274.

Oregon Youth Authority. 2011. *OYA Recidivism Risk Assessment—Violent Crime (ORRA-V): Modeling Risk to Recidivate with a Violent Crime*. Salem: Oregon Youth Authority.

Pew Center of the States, Public Safety Performance Project. 2011. *Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders*. Washington, DC: The Pew Center of the States. Retrieved from pewcenteronthestates.org/publicsafety.

Reiss, Albert J., Jr. 1951. The accuracy, efficiency, and validity of a prediction instrument. *American Journal of Sociology*, 56: 552–561.

Ridgeway, Greg. 2013. The pitfalls of prediction. *NIJ Journal*, 271: 34–40.

Skeem, Jennifer L. and John Monahan. 2011. Current directions in violence risk assessment. *Current Directions in Psychological Science*, 21: 38–42.

Tollenaar, N. and P. G. M. van der Heijden. 2013. Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive methods. *Journal of the Royal Statistical Society, Series A*, 176 (part 2): 565–584.

Turner, Susan, James Hess, and Jesse Jannetta. 2009. *Development of the California Risk Assessment Instrument*. Irvine: Center for Evidence Based Corrections, University of California, Irvine.

VanNostrand, Marie and Ken J. Rose. 2009. *Pretrial Risk Assessment in Virginia*. St. Petersburg, FL: Luminosity.

Vapnick, Vladimir N. 1998. *Statistical Learning Theory*. New York: Wiley.

Yang, Min, Yuanyuan Liu, and Jeremy Coid. 2010. *Applying Neural Networks and Other Statistics Models to Classification of Serious Offenders and the Prediction of Recidivism*, Vol. 6/10. London, U.K.: Ministry of Justice.

Richard A. Berk is a professor in the Department of Statistics and Department of Criminology at the University of Pennsylvania. He is an elected fellow of the American Statistical Association and the American Association for the Advancement of Science.

Justin Bleich is a graduate student in the Department of Statistics at the University of Pennsylvania.

Linking Prediction and Prevention

Greg Ridgeway
National Institute of Justice

First, I want to congratulate Berk and Bleich (2013, this issue) for their article that advances the use of modern statistical prediction methodology in criminal justice. Opportunities abound in the criminal justice system to use statistical prediction methods to improve decision making. Success depends on the quality of the prediction models, the quality of the prevention efforts, and the quality of the link between prediction and prevention. This policy essay will focus on these three components, which are essential to the effective blend of humans and machines in making criminal justice decisions.

Prediction Models

The science of prediction continues to evolve. This evolution includes the development of new statistical theory and methods, instantiation of those methods in software, improvements in data collection, and evaluation of these new data and methods on practical criminal justice issues.

Berk and Bleich (2013) cover numerous foundational issues that should be well understood by quantitative criminal justice scholars. These issues include recognizing the distinction between prediction and explanation, the need for validating prediction models in terms of out-of-sample predictive performance, and that not all prediction errors carry the same cost. In addition, Berk and Bleich note that prediction methods have advanced greatly. No longer should traditional regression be the only prediction tool in the criminal justice analyst's toolbox. Tools such as random forests and boosting are readily (and freely) available thanks to open-source implementations, including my implementation of generalized boosted models that now has a large user and developer community (Ridgeway, 2013). The current criminal justice literature has sufficient examples of the use of these tools so that no mystery should remain around them.

. The findings and conclusions reported in this article are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice. Direct correspondence to Greg Ridgeway, National Institute of Justice, 810 Seventh Street NW, Washington, DC 20531 (e-mail: greg.ridgeway@usdoj.gov).

DOI:10.1111/1745-9133.12057

Statistical learning is a dynamic field with new methods proposed regularly. How do we know which method is right for our application? The short answer is that we cannot know until we put them to the test as Berk and Bleich (2013) did in comparing out-of-sample predictive performance of logistic regression, random forests, and boosting on a particular example. Comfort with the traditional regression methods should not prevent analysts from experimenting with machine learning methods. Machine learning is the discipline that studies how to make computers learn without explicitly programming them (Mitchell, 1997). Machine learning algorithms are methods that absorb experiences (e.g., data sets) and improve by some performance measure (e.g., misclassification rate) in conducting a task (e.g., prediction) as experiences accumulate. Classification trees and even regression are machine learning techniques as they both meet this definition, although newer algorithms are vastly superior in their ability to learn. Regardless of the origin of the method and the label attached to the method, criminal justice analysts need to stay abreast of emerging higher quality prediction methodologies so that criminal justice can reap the benefits of the new advancements.

We need to monitor three issues as prediction models become more prevalent (as I am assuming that they will): concept drift, adaptive predictions, and the effect of existing prediction models.

First, concept drift occurs when the definitions of features and outcomes and their relationships change over time. The probation risk-assessment model that might be working splendidly now will need an update in the future as the nature of crime and the criminal justice community changes. Changes in the economy, new technologies, new criminal opportunities, and demographic shifts can degrade the performance of existing prediction models. In practice, simply letting older data expire addresses this problem, but when to update and how much data to let expire can be challenging to determine.

Second, many criminal justice issues to which we could apply prediction models are dynamic. Modern probation case management systems, for example, can collect streams of data on offenders including attendance in substance abuse treatment programs, probation officer reports, and offender location tracking. Although numerous prediction methods exist for static problems, the methodology for prediction from dynamic data streams is less developed. Consider an example. A prediction model for a new probationer predicts a low risk of reoffense. However, within 1 week of beginning probation, the probationer fails to appear for the first substance abuse treatment session. In the second week, the probationer misses a curfew by 30 minutes. In week three, the probationer finds a part-time job. The assessed risk should adapt to the timing and nature of each of these events. However, this requires new methods of estimating and fielding prediction models.

Last, prediction models usually are developed on data that conflate intervention activities with the underlying crime phenomena. For example, before the deployment of any prediction model, probation officers evaluate and monitor their probationers. Even without a sophisticated prediction model, perhaps they can assess accurately the high-risk cases and

can monitor in such a way as to prevent probation violations. Perhaps they have a treatment program that is phenomenally effective at suppressing criminal opportunities for high-risk probationers. A prediction model constructed from such cases might learn the rule that genuinely high-risk cases present low risk because the data show that they rarely offend. However, the prediction model has simply learned to predict suppression by the treatment program of high-likelihood crime outcomes, rather than predicting which probationers are at genuine high risk for offending.

Fielding a prediction model can exacerbate this conflation by identifying clearly the high-risk cases and situations to which prevention strategies and resources are applied. For example, consider a scenario in which police saturate intersections that the prediction model predicts to be at high risk of violent crime. As a result of the saturation, no violent crime actually occurs at those intersections. An assessment of the prediction model will conclude that no crime occurs in those places the model had predicted crime would occur. Attempts to update the prediction model using new crime data will continue to conflate the previous prediction model, the prevention activities, and the suppressed crime patterns. Interventions based on prediction models will change the system under study.

This problem is not easily solved without leaving some cases out of the intervention activities (e.g., continue business as usual for a high-risk probationer or leave high-risk intersections at unsaturated levels). Creating such a set of control cases is uncomfortable for practitioners, especially when they are convinced that their models and prevention efforts are effective. However, without control cases, we cannot be sure that the new prediction model is actually an improvement. Just like in clinical trials, we need to consider the principle of equipoise; when we really do not know whether the new model improves or impairs our performance, we are on solid ethical grounds to withhold an intervention until we have solid evidence.

Whereas those deploying prediction models need to be cognizant of these issues, they also need to have prevention strategies at the ready to respond to the predictions.

Prevention Models

The best prediction models will be useless unless we have effective prevention strategies to implement in response to predictions. Predictive policing research largely has focused on developing algorithms for predicting where and when crime will occur. Responses to those crime predictions almost always involve reallocating police patrol resources to the times and places predicted to have crime. Naturally, these are not the only plausible interventions. For example, changes in the street design, street lighting, bus stop locations, and positioning of school personnel also could be reasonable and perhaps more durable responses to the predictions.

I am particularly intrigued by the prospect of prediction models prompting the development of creative prevention models. The Chicago Police Department has created "hot lists" of residents predicted to be at high risk of being victimized. A highly accurate list will

only be of value if the department can act on those predictions effectively. The *Chicago Tribune* reported local police paying visits to those individuals that top the list (Gorner, 2013). This intervention is relatively low cost, although its effectiveness is yet to be seen. Much of the interest surrounds the use of a prediction model to drive the intervention. However, it is possible that the program's success might have more to do with making lists than using prediction models. That is, simply making and acting on lists of residents could be the entirety of the intervention, and machines, detectives, or community members might be equally effective at drawing up the list of high-risk individuals. Other available interventions include extensive surveillance or relocation of those at high risk of victimization. Such interventions are more disruptive and much more expensive.

When working with the Los Angeles Police Department (LAPD), I set up a prediction model to help their recruiting efforts (Lim, Matthies, Ridgeway, and Gifford, 2009). Some candidates were very likely to join the police force, such as those with some college and living in Los Angeles County. Other candidates never joined, such as those living outside California and reporting problematic answers on a background questionnaire. Given a candidate predicted to have a high probability of joining the LAPD, numerous candidate strategies were available to prevent these candidates from dropping out of the process. A cost-neutral approach was to prioritize those candidates for interviews and background investigations. Other strategies included additional mentoring and invitations to special events. However, we never resolved whether these intervention strategies further increased the chances of these high-viability candidates joining. So although LAPD had a readily available recruit prediction model with prospective predictive performance that could be measured, the available interventions were untested even if they seemed plausibly effective.

As high-quality prediction models become more available, this should continue to prompt the development of new interventions, customized for the individual, time, place, and context of the predicted crime. These intervention strategies need at least as much attention as the prediction model.

Linking Prediction to Prevention

Although prediction modeling has greatly advanced and those models are prompting innovative responses, substantial barriers lie in the path to linking modern prediction methods with criminal justice interventions. In every field in which statistical prediction methods have made inroads, they face opposition from the "experts" in the field, discomforted by the idea that a black box could compete with their decision making. Guszcza and Lucker (2012: 11) noted, "[t]he biggest challenges of executing on analytics are often found where algorithmic indications should be integrated with human professional judgment." A classic example is MYCIN (Yu et al., 1979), which is a computerized medical expert system that would take in patient characteristics and recommend treatments. In a well-designed experiment, the system outperformed practitioners in selecting the right antibiotic treatments. If 1979 technology offered such performance gains, surely we should be having even

greater performance today. Even though MYCIN is just one in a series of computer-based prediction successes, these systems are not found in doctors' offices. Heathfield (1999) noted, "theoretical and technical limitations are not the major barriers to the successful implementation . . . but rather more complex professional and organizational issues are at stake."

Judges, police officers, correctional officers, probation officers, prosecutors, and other criminal justice actors are accustomed to using their own experiences and instincts in making decisions. As one attorney noted in regard to the Philadelphia Department of Probation and Parole's use of a prediction model to assess offender risk, "I'm comfortable with judges looking at a particular defendant during sentencing, looking that defendant in the eye, asking them questions, gauging their sincerity" (Fiedler, 2013). I would agree if the judge's predictive accuracy was at least as good as alternatives based on statistical risk-assessment tools. We should not be deciding based on comfort but on our criminal justice objectives: public safety effectiveness, efficiency, and fairness.

In rare cases, the experts are willing to go head to head against the machine as they did in the MYCIN example. For example, in Ruger, Kim, Martin, and Quinn (2004), a panel of 83 experts, all of whom were prominent Supreme Court scholars, competed against a statistical prediction model to predict how the U.S. Supreme Court would vote on the 2002 docket of 68 cases. They correctly predicted how the Supreme Court would vote on 59% of the cases. The statistical model correctly predicted the majority opinion in 75% of the cases, substantially outperforming the experts.

Such examples show that prediction models can be valuable decision support tools in the justice system, but the legacy of MYCIN indicates that acceptance is a major barrier. That is, even with an accurate prediction model and a powerful set of intervention responses, a prediction model will be useless without a decision maker willing to deploy interventions in response to the predictions. As Berk and Bleich (2013) note, Pennsylvania is but one state passing legislation to promote the use of prediction models in criminal justice. The statute says that the instruments should "help," "may be incorporated," and "may be an aide" in decisions, leaving room for a professional in the decision-making process. Ultimately the success of these initiatives depends on the prediction model, the prevention strategies, and the interplay of humans and machines in making criminal justice decisions.

References

Berk, Richard A. and Justin Bleich. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, 12: 513–544.

Fiedler, Elizabeth. 2013. Risk management: New computer model helps Philly officers predict criminal potential. *Newsworks*. February 11. Retrieved October 15, 2013 from newsworks.org/index.php/health-science/item/50744-judicious-move-new-computer-model-helps-philly-officers-assess-criminal-risk.

Guszcza, James and John Lucker. 2012. A delicate balance: Organizational barriers to evidence-based management. *Deloitte Review*, 10: 5–21.

Gorner, Jeremy. 2013. Chicago police use "heat list" as strategy to prevent violence. *Chicago Tribune*. August 21.

Heathfield, Heather. 1999. The rise and "fall" of expert systems in medicine. *Expert Systems*, 16: 183–188.

Lim, Nelson, Carl F. Matthies, Greg Ridgeway, and Brian Gifford. 2009. *To Protect and to Serve: Enhancing the Efficiency of LAPD Recruiting*. Santa Monica, CA: RAND.

Mitchell, Tom. 1997. *Machine Learning*. New York: McGraw-Hill.

Ridgeway, Greg. 2013. *Generalized Boosted Models*. Software and code retrieved from code.google.com/p/gradientboostedmodels/.

Ruger, Theodore W., Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn. 2004. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decision making. *Columbia Law Review*, 104: 1150–1210.

Yu, Victor L., Lawrence M. Fagan, Sharon M. Wraith, William J. Clancey, A. Carlisle Scott, John Hannigan, et al. 1979. Antimicrobial selection by a computer: A blinded evaluation by infectious disease experts. *Journal of the American Medical Association*, 242: 1279–1282.

Greg Ridgeway is the acting director of the National Institute of Justice. His research spans both statistical methodology, including the development of statistical learning methods and open-source implementation of boosting, and empirical analysis in criminal justice, often using modern statistical prediction methods. He is a fellow of the American Statistical Association.

The Emergence of Machine Learning Techniques in Criminology

Implications of Complexity in our Data and in Research Questions

Tim Brennan
Northpointe, Inc., Georgia State University

William L. Oliver
Northpointe, Inc.

We view Berk and Bleich's (2013, this issue) article as potentially an important contribution to both criminology and practical criminal justice decision making. Although it is focused on new machine learning (ML) methods for forecasting, the article ramifies into several general problems that challenge the dominant, widely used standardized methods in criminology. The initial challenge raised by Berk and Bleich concerns the complexity of the data underlying many substantive and theoretical issues in criminology and criminal justice (e.g., multidimensionality, nonlinearity, complex interactions, feedback loops, and multimodality). Second, this challenge leads to the question of whether our standard parametric methods for predictive forecasting (e.g., logistic regression) are mismatched to the complexity and nonlinear decision boundaries found in many research and practical situations. Third, although ML methods seem to have substantial forecasting advantages in dealing with complex data, several current evaluative studies in criminal justice have suggested that ML methods have little predictive advantage over our standard parametric forecasting methods. Berk and Bleich explore reasons for these conclusions and conduct new comparative evaluations between two ML forecasting methods (random forests [RF] and stochastic gradient boosting) compared against logistic regression. In this policy essay, we examine some implications of the complexity issue for criminal justice method, theory, and practice, and we elaborate on several key issues raised by Berk and Bleich.

Direct correspondence to Tim Brennan, Northpointe Inc., 211 Old Town Road, Simpsonville, South Carolina 29681 (e-mail: tbrennan38@earthlink.net).

DOI:10.1111/1745-9133.12055

Two Minor Issues Regarding Berk and Bleich's (2013) Article

The Scope of ML: What Qualifies as ML?

We basically agree with Berk and Bleich (2013) on their categorization of what constitutes modern ML methods. Their selection of RF and stochastic gradient boosting for their comparative study is appropriate. However, we note the broad scope of ML and similar methods, and we agree that sharp distinctions between ML and related methods are probably not worth quarreling over. Thus, we describe briefly this broader range of methods—including unsupervised ML procedures—that has emerged over the last few decades in several disciplines concerned with pattern recognition and classification. Many of these methods were then implemented and steadily improved in computer science and ML laboratories.[1] We include unsupervised ML methods because of their history of uses in criminology although as yet they have minimal use in criminal justice practice. The two large categories of ML methods are as follows.

Supervised learning. This class of ML methods is well designed for the predictive tasks of classifying new cases into known preexisting categories (e.g., assigning probationers into preestablished risk levels). Typically, a large data set with known class membership is used to "learn" progressively a new predictive classification function that, after validation, can classify new unknown cases accurately. Predictive performance is assessed using various indicators of errors as demonstrated by Berk and Bleich (2013). We agree that RF and several other ML methods have great promise for supporting a wide range of criminal justice applications.

Unsupervised learning. Several families of unsupervised ML methods are designed for the basic research task of discovering novel or latent categories, new classifications, and patterns in complex data. The researcher is initially not aware of the existence and nature of any latent categories and the ML task is to "discover" them, verify their existence, and clarify their key characteristics (e.g., types of drunk drivers, varieties of sociopathic offenders, and offender developmental pathways). These discovery tasks use ML approaches such as *K*-means clustering, latent class mixture models, hierarchical clustering, neural networks, self-organizing maps, density estimation methods, and so on. Some emerged decades ago (e.g., *K*-means) and clearly would not qualify for Berk and Bleich's (2013) category of "modern ML" methods. However, most of these methods have had successive upgrades in the computer science and related ML literatures (Jain, 2010). Additionally, unsupervised ML has haltingly entered criminology with papers stretching back several decades (Brennan, 1987; Brennan and Breitenbach, 2009; Francis, Soothill, and Fligelstone, 2004; Harris and

1. For example, early versions of *K*-means emerged in psychology (Thorndike, 1953; Tryon, 1939). Biometrics pioneered hierarchical agglomerative methods (Sokal and Sneath, 1973). A history of the progressive development of many of these methods is offered by computer scientist Jain (2010), in his article "Data Clustering: 50 Years Beyond *K*-Means."

Jones, 1999; Hindelang and Weis, 1972; Jones and Harris, 1999; Megargee and Bohn, 1979, Raine et al., 2005).

Regarding this array of ML methods, we note the rapid software availability and ongoing refinement of both classic and newer ML methods. As noted by Berk and Bleich (2013), the *R* language for statistical computing provides good access to most recent ML methods in both supervised and unsupervised methods. For instance, we have used *R* extensively in our own recent research applying ML methods for developing and validating offender classification systems (Brennan, Breitenbach, and Dieterich, 2008; Brennan, Breitenbach, Dieterich, Salisbury, and Van Voorhis, 2013). We now review several key issues raised by Berk and Bleich and we consider some more general implications of ML developments for criminology.

On the Importance of Both Forecasting and Understanding in Criminal Justice

Berk and Bleich (2013) assert strongly that predictive forecasting using supervised ML methods is not expected or designed to produce an "understanding" of underlying causal processes. We agree that the fundamental role of these is to forecast the future with useful accuracy. Additionally, we agree with their warning that the distinction between forecasting and explanation is badly conflated in some quarters (Andrews, Bonta, and Wormith, 2006).

However, some of Berk and Bleich's (2013) remarks seem to downplay the necessity of reaching an "understanding" of individual cases. For example, they assert that:

> [T]he legislation contains no requirement that a judge understand why an individual is high or low risk. Indeed, it is not even clear what a judge would do with such information.

If this comment is focused purely on the role of risk forecasting, then it seems benign and consistent with Berk and Bleich's specific focus on ML forecasting methods and their omission of unsupervised learning. However, if this comment is meant to be more general, then we must disagree because we believe that judges, and other decision makers, must reach some understanding of their cases to design effective sentencing components and often to justify their reasoning (Tata, 2002).

Challenges to Criminology Emerging from Berk and Bleich's (2013) Article

Although Berk and Bleich (2013) focus specifically on the predictive accuracy of ML methods, their study brings us close to certain highly contested disputes in criminology that ramify into our methods, theories, and practical decision making. In our view, Berk and Bleich add an important and powerful voice to these debates. We comment on several challenges to criminology raised by Berk and Bleich, as well as on some implications of these issues.

Are Our Dominant Standardized Methods Often Mismatched to the Complexity of Our Data?

The complexity of criminological data. Initially, Berk and Bleich (2013) confront the data complexity challenges (nonlinear boundaries, complex interaction effects, etc.). This issue has long been controversial. Paul Meehl (1978) in his classic "Two Knights" paper reeled off 20 reasons for the extreme complexity of social science data as possible reasons for the slow progress of these disciplines, for example: divergent causality (multifinality and equifinality), feedback loops, nuisance variables, extreme multidimensionality, cultural influences, context influences, situational taxonomies, violation of parametric assumptions, and so on. Berk and Bleich remind us that our discipline has all of these problems (see also Lykken, 1991; Richters, 1997).

Are our standard widely used methods mismatched to the complexity of our data? There have been rumblings across several social science disciplines over the last two decades that the analytical and research methods of the softer social and biological sciences (psychology, sociology, developmental psychology, and criminology) are seriously mismatched to the complexity of their data and research issues (Cairns, Bergman, and Kagan, 1998; Lykken, 1991; Ragin, 1987, 2000; Richters, 1997). These researchers from different social and developmental science disciplines all discuss the potential benefits of their disciplines adopting an arguably more realistic open systems concept and its methods, versus staying with the currently dominant closed system paradigm and methods that social sciences inherited from the natural sciences. If the closed system option remains our choice, then it seems that for many complex research and policy issues, we will continue to face the constraints and deficiencies of this approach. This issue is being increasingly recognized as in Berk and Bleich's (2013) article.

All of the preceding publications argue that the closed system approach is inadequate for studying persons who are viewed as complex open systems exhibiting extreme heterogeneity, contingencies, equifinality, and multifinality in their developmental paths (Bergman, Cairns, Nillson, and Nystedt, 2000; Brennan et al., 2012; Granic and Patterson, 2006). Berk and Bleich (2013) are drawn into this debate mainly through their insightful recognition and demonstration of the complexity of criminological data and the simplifying assumptions required by parametric methods to cope with complexity. Berk and Bleich are not necessarily calling for us to abandon parametric methods but to broaden the range of our quantitative tools to include relatively modern nonparametric ML methods that might better match this complexity.

Potential gaps between parametric methods, complex data, and analytic practices. First, it is acknowledged openly that our data frequently fail to meet parametric assumptions (e.g., normal distributions, linearity, and so forth). This raises the question of whether the various "work-around" strategies typically used to meet these requirements are successful

(e.g., variable transformations, interaction terms, and so on). Berk and Bleich (2013) suggest that where violations are relatively minor, these strategies should work. However, serious problems can emerge as data sets are more complex, where the requirements of parametric methods require major transformations of several predictors, dummy variable coding, and the use of complex interactions using products of variables. Some brief comments from Berk and Bleich regarding this issue suggest caution in many research and practical situations: "Many would argue that these requirements cannot be met in practice."

Later, we find another important statement where Berk and Bleich (2013) discuss lessons from their simulation study to illustrate reasons why nonparametric methods can forecast better than parametric methods:

> The lessons learned can be applied far beyond logistic regression to any parametric regression approach. The lessons also apply to a wide range of functions that have clear structures but are very difficult for parametric regression models to capture.

Thus, with complex data, researchers might simply have to live with unknown and potentially substantial gaps among the data, assumptions, and methods. Berk and Bleich (2013) note also that in many situations, the decision boundaries might be relatively simple and little difference might exist in forecasting accuracy between logistic regression and ML methods, but then regarding these gaps, they ask, "How would one know? Because criminal justice decisions are so critical, their next question is, "Why take the risk?"

Higher forecasting accuracy of nonparametric ML methods with more complex data. Berk and Bleich's (2013) basic point is that parametric procedures will fail more often with highly complex data and nonlinear boundaries. To support this conclusion, they point to the formal proofs, simulations, and many compelling comparative studies in the computer science and ML literature; the positive results of their own comparative studies; and the clear visual demonstrations of how ML methods in the presence of nonlinear boundaries reduce classification errors. Although we find these to be compelling, we suggest that a prudent approach—given the current dearth of comparative studies of ML forecasting accuracy in criminal justice—is to wait for subsequent systematic evaluations.

Do ML methods offer more insight into the data structures than standard linear methods? Another important consideration is a potential gap between standard parametric and ML methods in their relative capacity to identify and describe complex or hidden data structures. This gap might be of substantial theoretical and practical importance (Bergman et al., 2000; Brennan and Breitenbach, 2009; Richters, 1997). Here, we note that unsupervised ML methods (e.g., self-organizing maps, latent class analysis, clustering, and bagged clustering) are designed explicitly to detect complex data structures adaptively (Brennan et al., 2008; Hennig, 2011; Hennig and Liao, 2010). RF and stochastic gradient methods are strongly adaptive and will search iteratively for complex structures in multiple passes through the data. In contrast, conventional parametric methods lack this adaptive character

and remain relatively insensitive to data structures involving nonlinearity, higher order interactions, multimodality, and non–global-type effects. Yet, as noted, these structures can be significant for understanding and interpreting criminological data, for example, in clarifying alternative developmental pathways to delinquency and crime, which have both theoretical and practical importance (Brennan et al., 2012; Granic and Patterson, 2006; Raine et al., 2005).

In a provocative analogy, Richters (1997) described the problems of the Hubble Space Telescope in which the initial mirror design of the telescope was faulty. These misdesigned mirrors were giving Hubble wrong prescription glasses and sending back only flawed and blurred images. When the corrected mirrors were installed (more than 3 years later), profound precision, clarity, and detail were revealed to open up the universe to the Hubble scientists. Criminology might have a parallel situation with our continuing strong emphasis on data analytic methods that are not designed to deal with highly complex issues, whereas a more advanced set of methods geared explicitly to address complexity has become available. To date, there are only halting movements toward the use of both unsupervised and supervised ML methods in criminology. Yet, Berk and Bleich's (2013) implication is that these new era analytical methods seem to be aligned more closely with the complexity of our data and research questions. Yet, there is some movement in the ML direction as observed in the choice of cluster analysis discovery methods by highly respected scholars in youth development and delinquency (Raine et al., 2005).

The Challenges of User Acceptance and Potential Resistance to ML Forecasting and Classification Methods

A realistic concern of Berk and Bleich (2013) is the possibility that ML methods might be overlooked, ignored, or undervalued by criminological researchers and practitioners for several reasons, which are explored subsequently. They pointedly ask the following:

> Why would the kinds of new analysis procedures being developed for analyzing a variety of data sets with hundreds of thousands of cases . . . not be especially effective for a criminal justice data set of similar size?

In our opinion, they have several causes to be concerned. This section discusses potential hurdles that could derail the successful diffusion and productive use of ML methods in criminology and criminal justice.

A history of weak adoption, poor implementation, and rejection of technical advances in risk assessment and classification in criminal justice agencies. A discouraging finding in the criminal justice implementation literature is that agencies are often highly resistant to change and to new or novel methods that require substantial new learning, that are unfamiliar, or that require abandoning familiar methods (Bonta, 2002; Brennan, 1999; Garrison, 2009; Harris and Smith, 1993; Wormith, 2001). A substantial proportion of the case studies in this implementation literature is focused on risk assessment, risk prediction,

and other classification tools. A potentially relevant example—somewhat analogous to RF diffusion—is the history of attempts to introduce quantitative divisive trees for prediction. Don M. Gottfredson introduced quantitative divisive trees for recidivism risk prediction 50 years ago (Gottfredson and Ballard, 1965). The method achieved minimal use beyond its initial introduction and was abandoned. Subsequently, newer quantitative divisive tree methods evolved during the next three decades with steadily upgraded methods (e.g., AID, CHAID, THAID, and CART). Despite the wide availability of user-friendly software for these newer methods, (e.g., SPSS [SPSS Corporation, Chicago, IL], JMP [SAS Institute, Inc., Cary, NC], and Salford Systems [Salford Systems, San Diego, CA]), none has achieved wide dissemination or use in applied criminal justice or criminology with only few studies entering our literature.

The challenge of achieving acceptance and use of novel risk-prediction and classification tools for criminal justice is shown also in surveys of correctional mental health staff to identify what risk-assessment tests they use (Bonta, 2002; Boothby and Clements, 2000). A key finding—at that time—was that staff overwhelmingly avoided the best validated correctional risk assessments (e.g., Psychopathy Checklist–Revised [Hare, 1990]; Level of Service Inventory–Revised [LSI-R; Andrews and Bonta, 1995]). These assessments were used only by 11% or less of the staff. In contrast, instruments reflecting their prior mental health or psychological training had far higher levels of use (e.g., Minnesota Multiphasic Personality Inventory, MMPI-2, 96%). Bonta (2002), a coauthor of the LSI, called these results "dishearteningly informative" (p. 357).

Acceptance depends partly on the perceived benefits of a new method. The acceptance of advanced ML methods can depend profoundly on whether they are viewed as having meaningful benefits to users. Thus, the few comparative studies of ML methods suggesting that they are no better than standard logistic regression might be devastating and could derail efforts to introduce ML methods into criminal justice agencies.[2] Berk and Bleich (2013) strongly challenge these negative conclusions.

Competency and training problems in ML methods. Clearly, Berk and Bleich (2013) are aware of competency problems in alluding to flawed prior studies and potential misuse of ML techniques because of the lack of experience and training. They also note that the conceptual framework and procedures of ML methods are different from the standard general linear model (GLM). They suggest that resolving these competency issues will require a "substantial change in data analysis craft lore" and in interpretative practices. Thus, training issues might be substantial for both researchers and practical users. For applied use in criminal justice agencies (courts, probation, and parole), we suggest that a

2. A recently published study of ML procedures used criminal justice data to predict violent arrests after prison release. This study offered support for the superior performance of RF under certain conditions compared with both other ML methods and logistic regression (Breitenbach, Dieterich, Brennan, and Fan, 2009). This study used several ML algorithms including random forests, support vector machines, gradient descent, neural networks, and ADTree, as well as logistic regression for comparisons.

broader and simpler form of user competence and procedural techniques will be needed for any wider use of ML methods. This will require careful planning and training to introduce ML methods into agency policies and procedures in ways that are comfortable and informative to CJ practitioners.

The black box problem and practical and political needs to justify criminal justice decisions. Another feature of some ML methods that might hurt acceptance among certain practitioners is that they are not designed to offer any clear logic or explanation for their forecasting decisions. Their logic can be inscrutable to human users, and they might be viewed negatively as a black box. This failure to offer explanatory tools might be awkward for decision makers who must provide justifications for their decisions (e.g., judges, probation officers, and parole boards) (Tata, 2002). However, Berk and Bleich (2013) are right to remind us of the differing functions of prediction and explanation and that the goal of ML techniques such as RF is limited to forecasting accuracy and this should be the benchmark and not the development of an explanation.

Inadequate and misleading evaluations of ML methods in criminal justice. Berk and Bleich (2013) seem to be on solid ground in their identification of the difficulties in achieving fair evaluations of ML methods in criminology and on the flaws in several prior comparative studies. They list several of these flaws usefully, as follows: failure to address follow-up testing samples, failure to assess accuracy appropriately, failure to specify carefully the features on which methods will be evaluated, confusion over what evaluative criteria to use, as well as weak implementation. To combat such studies, Berk and Bleich offer a useful list of design requirements to achieve fair "apples-to-apples" comparisons. This might be useful to help criminal justice researchers avoid adding flawed studies to this literature.

A brief comment on unsupervised ML methods. As noted, Berk and Bleich (2013) focus only on supervised ML methods (RF) for prediction with no comment on unsupervised methods. Space does not permit a similar discussion of evaluations of unsupervised methods—but the following can be reviewed for these methods generally (Arabie, Hubert, and DeSoete, 1996; Han and Kamber, 2001; Hastie, Tibshirani, and Friedman, 2008) and some applications in criminal justice (Brennan et al., 2008; Harris and Jones, 1999; Huizinga, Esbensen, and Weiher, 1991; Raine et al., 2005).

Comments on Technical Issues

In this section, we shift perspectives and move to several technical issues raised by Berk and Bleich (2013). In general, Berk and Bleich's discussion of technical issues is exceptionally clear and accurate. However, we believe the following issues might be important to consider.

Complex Data and Extreme Multidimensionality

One aspect of ML methods is that they are ideally suited for high-dimensional data that include hundreds, or even thousands, of predictors while having relatively few cases. This strength is not exploited by Berk and Bleich (2013) perhaps because the example data

set contained only eight predictor variables. This aspect, however, might be important in criminal justice because data for modeling recidivism often can include a large number of variables. This is particularly the case when the offenders are assessed with newer more comprehensive theory-guided instruments such as the Correctional Offender Management profiles for Alternative Sanctions (COMPAS) (Brennan, Dieterich, and Ehret, 2009), LSI (Andrews and Bonta, 1995), or with recently developed gender responsive instruments for women offenders (Van Voorhis, Wright, Salisbury, and Bauman, 2010). We suggest that for such data, the researcher will especially want to try the machine learning methods.

Although Berk and Bleich (2013) mention regularization methods for predictive modeling only briefly, these also could be especially useful for building models with high-dimensional data. Lasso regression for binary data incorporates shrinkage, or regularization, and from a practitioner's perspective, it is similar to logistic regression (Hastie et al., 2008). The lasso carries out variable selection so that one can include many predictor variables in the model but still arrive at a simple, parsimonious final model. Lasso regression and the RF model would provide a useful comparison between a simple regression model, involving the linear effects of only a few inputs, and a complex RF model, involving the linear and, possibly, nonlinear effects of all inputs.

Supporting the Interpretation of ML Models

As noted, the random forest and stochastic gradient boosting models are black box models to some degree. Although they provide measures of variable importance, these do not explain fully the relationship between the predictor variables and the outcome variable. For example, if there is an interaction with gender, then a variable could be positively related to the outcome for males and negatively related for females. As argued previously, an understanding of the complexity underlying criminal behavior could move the field forward theoretically and often is required by practitioners who must defend their decisions. We are not aware of sophisticated tools for identifying interactions between predictor variables for the random forest model. However, such tools exist for stochastic gradient boosting models (Elith, Lethwick, and Hastie, 2008). These tools are implemented in R software and offer the opportunity to understand the complex relationships that the machine learning method uses when making predictions.

Discussion

Perhaps the most important implication of the lead article is to alert criminology to the challenge posed in the Hubble story. To the degree that our standard widely used methods are ill suited to many of the more complex data sets and research questions that pervade criminological research, we might be handicapping our field. More accuracy in predictive techniques would obviously have an enormous practical impact, and there are constant calls from criminal justice practitioners (judges, probation officers, parole board members, and policy makers) for improved predictive accuracy. Yet, as noted, serious implementation

problems have occurred throughout criminal justice history, and this history must be taken into account with the introduction and diffusion of ML techniques. From a theoretical perspective, improved accuracy in predictive modeling would be important for theory testing. However, the abilities of unsupervised ML in improving our methodologies for detecting complex latent structures in our data will not be of lesser importance. We believe that Berk and Bleich (2013) might be valuable for criminology and criminal justice in providing a strong voice alerting us to the current deficits of our major research tools and perhaps opening the way for future developments regarding more effective tools.

References

Andrews, Don A. and James A. Bonta. 1995. *The Level of Service Inventory–Screening Version*. Toronto, Ontario, Canada: Multi-Health Systems, Inc.

Andrews, Don A., James A. Bonta, and J. Stephen Wormith. 2006. The recent past and near future of risk/need assessment. *Crime & Delinquency*, 52: 7–27.

Arabie, Phipps, Lawrence J. Hubert, and Geert De Soete. 1996. *Clustering and Classification*. Singapore: World Scientific.

Bergman, Lars R., Robert B. Cairns, Lars-Goran Nillson, and Lars Nystedt. 2000. *Developmental Science and the Holistic Approach*. Hillsdale, NJ: Lawrence Erlbaum.

Berk, Richard A. and Justin Bleich. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, 12: 513–544

Bonta, James A. 2002. Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior*, 29: 355–379.

Boothby, Jennifer L. and Carl B. Clements. 2000. A national survey of correctional psychologists. *Criminal Justice and Behavior*, 27: 715–731.

Breitenbach, Markus, William Dieterich, Tim Brennan, and Adrian Fan. 2009. Creating risk-scores in very imbalanced datasets. In (Yun Sing Koh and Nathan Rountree, eds.), *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection*. Hershey, PA: IGI Global.

Brennan, Tim. 1987. Classification: An overview of selected methodological issues. In (Don M. Gottfredson and Michael H. Tonry, eds.), *Prediction and Classification: Criminal Justice Decision Making*. Chicago, IL: University of Chicago Press.

Brennan, Tim. 1999. Implementing organizational change in criminal justice: Some lessons from jail classification systems. *Corrections Management Quarterly*, 3: 11–27.

Brennan, Tim and Markus Breitenbach. 2009. The taxonomic challenge to general theories of delinquency. In (Ozan Sahin and Joseph Maier, eds.), *Delinquency: Causes, Reduction and Prevention*. Hauppauge, NY: Nova Science.

Brennan, Tim, Markus Breitenbach, and William Dieterich. 2008. Towards an explanatory taxonomy of adolescent delinquents: Identifying several social-psychological profiles. *Journal of Quantitative Criminology*, 24: 179–203.

Brennan, Tim, Markus Breitenbach, William Dieterich, Emily J. Salisbury, and Patricia Van Voorhis. 2012. Women's pathways to serious and habitual crime: A person-centered analysis. *Criminal Justice and Behavior*, 39: 1481–1508.

Brennan, Tim, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS Risk and Needs Assessment System. *Criminal Justice and Behavior*, 36: 21–40.

Cairns, Robert B., Lars R. Bergman, and Jerome Kagan. 1998. *Methods and Models for Studying the Individual*. Thousand Oaks, CA: Sage.

Elith, J., Lethwick, J.R. and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77: 802–813.

Francis, Brian, Keith Soothill, and Rachel Fligelstone. 2004. Identifying patterns and pathways of offending behavior. *European Journal of Criminology*, 1: 47–87.

Garrison, Arthur H. 2009. The influence of research on criminal justice policy making. *Professional Issues in Criminal Justice*, 4: 9–21.

Granic, Isabela and Gerald R. Patterson. 2006. Toward a comprehensive model of antisocial development: A dynamic systems approach. *Psychological Review*, 113: 101–131.

Gottfredson, Don M. and Kelley B. Ballard. 1965. *The Validity of Two Prediction Scales: An Eight Year Follow Up Study*. Sacramento, CA: Institute for the Study of Crime and Delinquency.

Han, Jiawei and Micheline Kamber. 2000. *Data Mining: Concepts and Techniques*. New York: Morgan Kaufmann.

Hare, Robert D. 1990. *The Hare Psychopathy Checklist—Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Harris, Philip W. and Peter R. Jones. 1999. Differentiating delinquent youths for program planning and evaluation. *Criminal Justice and Behavior*, 26: 403–434.

Harris, Philip W. and S. H. Smith. 1993. *Developing Community Corrections: An Implementation Perspective*. In (A. Harland, ed.), *Choosing Correctional Options that Work: Defining the Demand and Evaluating the Supply*. Thousand Oaks, CA, Sage Publications.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

Hennig, Christian. 2011. *How to Find the Best Cluster Method*. London, U.K.: Department of Statistics, University College of London.

Hennig, Christian and Tim F. Liao. 2010. *Comparing Latent Class and Dissimilarity Based Clustering for Mixed Type Variables with Application to Social Stratification*. Research Report No. 308. London, U.K.: Department of Statistics, University College London.

Hindelang, Michael J. and Joseph G. Weis. 1972. Personality and self-reported delinquency: An application of cluster analysis. *Criminology*, 10: 268–294.

Huizinga, David, Finn-Aage Esbensen, and Anne Wylie Weiher. 1991. Are there multiple paths to delinquency? *Journal of Criminal Law & Criminology*, 82: 83–118.

Jain, Anil K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31: 651–666.

Jones, Peter R. and Philip W. Harris. 1999. Developing an empirically based typology of delinquent youths. *Journal of Quantitative Criminology*, 15: 251–276.

Lykken, David T. 1991. What's wrong with psychology anyway? In (Dante Cicchetti and William M. Grove, eds.), *Thinking Clearly about Psychology*, Volume 1. Minneapolis: University of Minnesota Press.

Meehl, Paul E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46: 806–834.

Megargee, Edwin Inglee and Martin J. Bohn. 1979. *Classifying Criminal Offenders: A New System Based on the MMPI*. Beverly Hills, CA: Sage.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago, IL: University of Chicago Press.

Raine, Adrian, Terrie E. Moffitt, Avshalom Caspi, Rolf Loeber, Magda Stouthamer-Loeber, and Don Lynam. 2005. Neurocognitive impairments on boys on the life-course persistent antisocial path. *Journal of Abnormal Psychology*, 114: 38–49.

Richters, John E. 1997. The Hubble hypothesis and the developmentalist's dilemma. *Development and Psychopathology*, 9: 193–229.

Sokal, Robert Reuven and Peter Henry Andrews Sneath. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco, CA: Freeman.

Tata, Cyrus. 2002. Accountability for the sentencing process—Towards a new understanding. In (Cyrus Tata and Neil Hutton, eds.), *Sentencing and Society*. Aldershot, U.K.: Ashgate.

Thorndike, Robert L. 1953. Who belongs in the family. *Psychometrika*, 18: 267–276.

Tryon, Robert C. 1939. *Cluster Analysis*. Oxford, U.K.: Edwards Bros.

Van Voorhis, Patricia, Emily M. Wright, Emily Salisbury, and Ashley Bauman. 2010. Criminal women's risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior*, 37: 261–288.

Wormith, J. Stephen. 2001. Assessing offender assessment: Contributing to effective correctional treatment. *ICCA Journal on Community Corrections*, July: 12–18.

Tim Brennan is chief scientist at Northpointe Institute, and a visiting scholar at Georgia State University. His interests are in taxonomy development and applications in criminal justice. He received the Warren-Palmer award in 2007 from the American Society of Criminology.

Bill Oliver is a research scientist at Northpointe, Inc. His interests are primarily in the areas of predictive modeling and statistics for high-dimensional data.

Is There Any Logic to Using Logit

Finding the Right Tool for the Increasingly Important Job of Risk Prediction

Shawn D. Bushway
University at Albany, State University of New York

Richard A. Berk and Justin Bleich's (2013, this issue) article is a direct response to an article in the *Journal of the Royal Statistical Society, Series A* on risk prediction (Tollenaar and van der Heijden, 2013).[1] In the article, Tollenaar and van der Heijden concluded that "using selected modern statistical, data mining and machine learning models provides no real advantage over logistic regression and LDA" (p. 582). Obviously, Berk and Bleich disagree. Their primary source of disagreement comes with the models chosen by Tollenaar and van der Heijden to represent "modern" techniques. Most notably, Tollenaar and van der Heijden did not include random forests, a technique championed by Berk, among others.[2] When Berk and Bleich repeat the exercise using their own data, logistic regression is outperformed by both random forests and stochastic gradient boosting. Although Berk and Bleich know that there are cases in which the performance differences between their preferred technique(s) and logistic regression may be small (when there is a simple decision boundary), they believe that, at the very least, machine learning techniques should be directly compared with logistic techniques before a decision is made to rely on the logistic techniques.

Direct correspondence to Shawn D. Bushway, 219 Draper Hall, School of Criminal Justice, University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222 (e-mail: sbushway@albany.edu).

1. As of January 2013, the author has been an associate editor of the *Journal of the Royal Statistical Society, Series A*. The author was not on the editorial board during the review process for the article by Tollenaar and van der Heijden (2013). The views and opinions expressed in this article are those of the author and do not necessarily reflect the official policy, position or opinions of the editorial board of the *Journal of the Royal Statistical Society, Series A*.
2. Readers concerned that Berk is merely annoyed because Tollenaar and van der Heijden (2013) ignored his preferred technique should read Ridgeway (2013), where acting director of the National Institute of Justice, Greg Ridgeway, talks about the strengths of random forests.

DOI:10.1111/1745-9133.12059

Where does Berk and Bleich's (2013) conclusion leave the reader of *Criminology & Public Policy*? Perhaps Tollenaar and van der Heijden (2013) have data with a simple decision boundary, and their model performs at least as well as machine learning tools. Alternatively, perhaps Berk and Bleich are correct, and Tollenaar and van der Heijden simply failed to use the latest and greatest techniques for risk prediction.[3] The only way to know for sure would be to test random forests on Tollenaar and van der Heijden's data, something that I hope Tollenaar and van der Heijden do, and eventually submit to this journal. Absent this step, what can be said about the relative value of these techniques?

Tollenaar and van der Heijden (2013), I believe, would argue that policy makers can safely adopt logistic regression methods, without going through the step of comparing them with machine learning. In my opinion, their article is clearly meant as a breakwater against the growing swell of opinion favoring machine learning techniques. This argument has two motivations. First, logistic regression techniques are widely understood and easy to implement. In contrast, machine learning techniques are not widely understood and are, at least at this time, more difficult to implement well.[4,5] Second, Tollenaar and van der Heijden believe that "the formulation of the model can always be translated into a set of equations in an [E]xcel spreadsheet so it can readily be used by a probation worker" (p. 582). In contrast, "black box" methods like random forests need to be generated live by a computer program that is literally a black box to the end user.[6]

It is not difficult to understand why Tollenaar and van der Heijden (2013) and many criminal justice researchers both inside and outside of government would prefer their conclusion over Berk and Bleich's (2013) conclusion favoring machine learning tools or a conclusion that requires a runoff between machine learning tools and logistic regression. But can their position be justified? The answer, I think, is a resounding no. I am reminded of the useless manmade breakwater I saw recently in Bar Harbor, Maine, abandoned by the

3. Tollenaar and van der Heijden's (2013) article contains no reference to Berk's work in risk prediction. It also does not use the term "random forests."

4. Berk and Bleich (2013) explain how the methods differ. See also a good description of the difference between regression and "black box" methods by Neyfakh (2011) in *The Boston Globe*.

5. Part of Berk and Bleich's (2013) complaint about Tollenaar and van der Heijden's (2013) article is that they do not believe that the machine learning approaches were properly tuned, which is a key part of the implementation.

6. Ritter (2013) did a nice job of describing the steps needed to design and implement a random forest forecaster in a real agency. This latter issue involves tension around forecast accuracy and simplicity/usability. Greg Ridgeway hit the nail on the head when he said that "(a) decent transparent model that is actually used will outperform a sophisticated system that predicts better but sits on a shelf (Ridgeway, 2013: 36). However, I believe that Tollenaar and van der Heijden (2013) oversell the transparency and usability of logistic-based forecasts. Although they can be programmed in Microsoft Excel, they are, in my opinion, no more transparent to the user than the types of "black boxes" described by Ritter (2013). Having personally tried to explain the logistic transformation to practitioners, I am convinced that only additive scales (which cannot be derived from logistic models) are truly "transparent" to the end user.

government after a decision was reached, too late, that it simply was not possible to protect the harbor from incoming waves. It is a bit forlorn and serves as a useful warning about the need to make prudent decisions prior to investing (or believing in) effective breakwaters.

As Tollenaar and van der Heijden (2013) recognized, and as both Berk and Bleich (2013) and Ridgeway (2013) repeated, no model will work best for all problems. Therefore, using logistic models, without direct comparison with other methods, easily could lead to results that are systematically and substantially worse than what might be achievable using machine learning tools. This is particularly the case if policy makers wish to take advantage of important features like asymmetric cost structures and multiple cut lines that are much easier to implement with machine learning tools than with logistic regression. In my opinion, articles like that by Tollenaar and van der Heijden will only delay the inevitable need to "tool up" on these nonlinear, nonparametric techniques. After reading Berk and Bleich (2013), I am more convinced than ever that when it comes to forecasting, the game has changed—and it is no longer being played exclusively, or even mainly, with regression approaches.

I come to this conclusion grudgingly. I am a regression-based researcher, and although recently I have begun working on forecasting issues, I have done so almost exclusively with regression-based approaches. Although random forests are simple conceptually, they are foreign territory and difficult to understand. To be honest, there is something inherently unnerving about Berk and Bleich's (2013) statement that, "One does not have to rely on a 'structural model' when forecasting is the primary motive." I understand this statement intellectually—but I like my structural models nonetheless. Placing my faith in a "black box" that is not easily interpreted is a difficult step.

But Berk and Bleich (2013) are right—if prediction is the game, then machine learning techniques are the high draft picks. And just like high draft picks, they need to be invested in before they will bear fruit. The field of criminal justice would be wiser to spend its time and energy investing in learning these new techniques than in relying on logistic regressions that hopefully, or at least in some circumstances, do at least as well.

This type of investment will not happen overnight. But it is worth thinking about the kinds of things that need to happen to make the consideration of these kinds of issues more common and straightforward in criminology. Without a doubt, a more focused study on forecasting/prediction needs to be conducted. Other fields in social science have whole groups of researchers devoted to the study of forecasting. Most econometric texts and policy books have whole sections on forecasting/prediction and are clear on how it is distinguished from causal modeling. Yet I am unaware of any course dedicated to the statistics of risk prediction in a graduate program of criminology, and it does not seem that there has been a risk-prediction workshop on workshop day at the annual American Society of Criminology meeting.[7] Most forecasting discussions in criminology center on time series (crime trends),

7. I did find one occasional seminar at Temple University by Kate Auerhahn on risk assessment. The focus seems to be on the issues surrounding the use of risk assessment rather than on the techniques

and although it is important, it is a different exercise than the kinds of forecasting discussed by Berk and Bleich (2013). A recent search of the American Society of Criminology 2013 conference program found no titles with the words "predict" or "forecast" or "machine learning." Approximately a dozen papers on risk assessment have been published, but a cursory review of the abstracts gives me little hope that the authors are systematically addressing the questions raised by Berk and Bleich. Instead, the focus is on whether these tools do better than clinical decision makers and whether (or how) "off-the-shelf" tools can be validated in other settings. Overall, the lack of attention to the important methodological issues around risk assessment must be overcome if criminology is to keep pace with and inform the state of art in criminal justice practice.

I recognize that the field of criminology has mixed feelings about the use of risk-prediction tools.[8] However, these mixed feelings are largely absent from the world of criminal justice, where practitioners are under pressure, often from lawsuits, to be more effective with fewer resources.[9] The use of risk-prediction tools is already substantial, and it is growing rapidly (Harcourt, 2007; Simon, 2005). Criminology as a field can choose to ignore this, or it can help build the science and practice of risk prediction.

One way this might happen is through the application of some of the lessons learned by criminal justice researchers facing the "nothing works" problem in the world of treatment programs. Over time, the scientific community developed a variety of mechanisms intended to help identify and then "translate" the research on program effectiveness such that practitioners could put this research into practice (see, for example, the Campbell Collaboration). Similar "translational" techniques could work also to help practitioners faced with mandates to use risk assessment as part of their everyday decision making.

Berk and Bleich (2013) start the conversation for this type of exercise by proposing some of the standards for what would prove to be a valuable and valid comparison between techniques. Simple tasks like creating credible lists of best available techniques could serve to minimize confusion and reduce search time by practitioners looking to identify prospective tools. This type of exercise also could make some progress on creating a consensus

themselves (temple.edu/cj/graduate/documents/AuerhahnRiskPredictionandClassification.pdf). I would be happy to learn about other examples.

8. For example, see Harcourt (2007) and Hannah-Moffat (2012). Both authors raised important questions about the potential negative side effects of using formal risk-assessment tools. However, these legitimate issues will not be addressed in practice if the people who care about them are not involved with the creation, implementation, and validation of these techniques.

9. Modern risk assessment started when correctional institutions faced the need to create better classification systems as a result of class-action lawsuits related to overcrowding. Courts viewed more objective classification systems as one remedy for the problem of overcrowding and put pressure on correctional systems to create transparent and testable systems of risk assessment. This pressure from the courts to create more effective classification systems later expanded to parole and probation systems facing lawsuits for releasing individuals who subsequently harmed people within the community (Clements, 1996).

about reasonable measures of "fit," a conversation often lost under the blizzard of different metrics.[10]

The evidence is clear. Actuarial risk tools are now a standard part of how criminal justice professionals make decisions. These tools are growing increasingly complex, with the best performing techniques often coming from families of techniques that are very different than the types of regression-based tools with which most criminologists are familiar. We have a choice—we can ignore this or we can engage with the new science. Individual criminologists, as well as institutions in criminology, whether they are departments, journals, institutes, or associations, need to become more actively involved in developing and helping to propagate best practices among criminal justice professionals, professionals who are using these tools thousands of times a day to make important and potentially life-changing decisions.

References

Berk, Richard A. and Justin Bleich. 2013. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, 12: 513–544.

Clements, Carl B. 1996. Offender classification: Two decades of progress. *Criminal Justice and Behavior*, 23: 121–143.

Hannah-Moffat, Kelly. 2012 Actuarial sentencing: An unsettled proposition. *Justice Quarterly*, 30: 270–296.

Harcourt, Bernard E. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.

Neyfakh, Leon. 2011. You will commit a crime in the future. Inside the new science of predicting violence. *The Boston Globe*. February 20.

Ridgeway, Greg. 2013. The pitfalls of prediction. *NIJ Journal*, 271: 34–40.

Ritter, Nancy. 2013. Predicting recidivism risk: New tool in Philadelphia shows great promise. *NIJ Journal*, 271: 4–13.

Simon, Jonathan. 2005. Reversal of fortune: The resurgence of individual risk assessment in criminal justice. *Annual Review of Law and Social Science*, 1: 397–421.

Tollenaar, Nikolaj and Peter van der Heijden. 2013. Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive methods. *Journal of the Royal Statistical Society, Series A*, 176 (part 2): 565–584.

Shawn D. Bushway is a professor in the School of Criminal Justice and the Rockefeller College of Public Affairs and Policy at the University of Albany. His research interests include criminal behavior over the life course (especially desistence), sentencing policy and its effects, and the process of reentry for people exiting the control of the criminal justice system.

10. Berk and Bleich's (2013) reliance on a two-by-two grid of forecast errors was refreshingly clear—and is an excellent place to start when trying to evaluate the power (and limits) of the given prediction tool.