Toward a systematic foundation for identifying evidence-based criminal justice sanctions and their relative effectiveness

Daniel P. Mears a,⁎, J.C. Barnes b

a Florida State University, College of Criminology and Criminal Justice, 634 West Call Street, Tallahassee, FL 32306-1127, United States
b Criminology Program, The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, TX 75080, United States

A B S T R A C T

Nationally, there have been increased calls for evidence-based criminal justice policy. Despite considerable progress toward that objective, there still is no systematic, comparative foundation for assessing the relative effectiveness of diverse sanctions in achieving any of a range of goals. In this article, the importance of evidence-based policy and the critical research gaps that must be filled were discussed, as well as the next steps that must be taken to place criminal justice sanctioning on a solid, evidence-based foundation. Concluding remarks focused on the implications of current research gaps and several strategies for addressing them.

© 2010 Elsevier Ltd. All rights reserved.

Introduction

The past several decades have been witness to two prominent developments in criminal justice—a dramatic increase in correctional populations and costs and a corresponding increase in calls for evidence-based policy. For example, between 1980 and 2008, the number of individuals in state and federal correctional systems more than quadrupled, rising from 1.8 million to 7.3 million (Bureau of Justice Statistics, 2010a). Costs increased as well: from 1980 to 2006, the latest year for which national data are available, criminal justice expenditures increased from $19 billion to $99 billion (Bureau of Justice Statistics, 2010b).

At the same time, the mantra of accountability and evidence-based policy or practice surfaced in nearly all areas of government (Bardach, 1997; Tilley, 2009; Wells, 2009). The prominence of this shift can be seen in part by the change, in 2004, in the name of the U.S. General Accounting Office to the Government Accountability Office. As federal, state, and local governments have had to grapple with budget shortfalls due to a declining economy, the focus has assumed, if anything, greater urgency in recent years among those involved in designing, evaluating, overseeing, or implementing criminal justice policy (Drake et al., 2009; Lawrence, 2009; Listwan, Jonson, Cullen, & Latessa, 2008; Sherman, 2009; Spelman, 2009; Tonry, 2009; Welsh & Harris, 2008).

Policymakers frequently cite the high rate of recidivism among sanctioned offenders as a critical reason for why more effective (i.e., evidence-based) ways to combat crime are needed. The estimates vary according to the particular correctional population, area, or program, but most paint a dismal picture. Typical, for example, is Langan and Levin’s (2002) 15-state national study, which showed that 68 percent of released prisoners were rearrested within three years. Most such estimates only capture offenses that come to the attention of the criminal justice system, and thus understate the true scope of the problem. For this reason, the dramatic rise in the prisoner reentry problem alone—over 735,000 inmates are released from state and federal prisons annually (Sabol, West, & Cooper, 2009)—constitutes a significant cause for concern since over two-thirds of these released individuals can be expected to recidivate. The concern, however, extends more broadly to jail and community supervision populations and any other group subject to the control, authority, or supervision of the criminal justice system.

Outcomes other than recidivism have also been identified as a problem for which evidence-based policies are needed. Here, again, prisoner reentry highlights why. Research has underscored, for example, the numerous challenges that inmates face upon release, such as unemployment, homelessness, drug abuse, physical and mental illness, and family dysfunction (Barnes, Dukes, Tewksbury, & De Troye, 2009; Clear, 2007; Foster & Hagan, 2009; Petersilia, 2003; Travis, 2005; Travis & Visher, 2005). Nonetheless, although there has been much progress in criminal justice evaluation research and in the development of effective interventions (Cullen, 2005; Drake et al., 2009; MacKenzie, 2006; Mears, 2007, 2010), there remains no systematic or comprehensive foundation to which policymakers can turn for credible, “evidence-based” information about the relative effectiveness of any of a range of criminal justice policies aimed at improving recidivism or a range of other outcomes. As but one
illustration of the problem, consider that the large-scale increase in meta-analyses and systematic reviews over the past decade include—as Todd Clear (2010), in his Presidential Address to the American Society of Criminology, recently emphasized—“language that bemoans a weak empirical base” (p. 6).

The goal of this article is to highlight this problem and to contribute to efforts to improve criminal justice policy. To this end, critical research gaps that preclude placing criminal justice sanctioning on a solid evidence-based foundation are identified. Sanctioning, as defined here, refers to any punishments, interventions, programs, services, and treatments—whether in isolation or conjointly—that are used to punish, control, manage, help, or treat convicted offenders. This definition stems from the recognition that sanctions almost invariably include more than “just” punishment; they frequently include required or offered services or treatments. The current focus is restricted to sanctioning rather than criminal justice policy more broadly constructed to allow for a more focused discussion of the problems that face the central, or core, activity of the criminal justice system. As discussed in the conclusion, however, these problems extend not only to sanctioning but also to a range of other criminal justice endeavors, including proactive crime prevention programs and prison management policies and practices.

The discussion begins by describing the emergence, in recent years, of evidence-based practice as a central focus of criminal justice policy discussions and research. It then turns to a description of evidence-based sanctioning policies. Next, the article presents a conceptual framework for highlighting the critical research gaps that must be addressed to provide a solid foundation for making criminal justice sanctioning more evidence-based. The critical next steps that should be undertaken to create this foundation are then discussed. The article concludes by identifying the directions in which research and policy should be directed to increase both knowledge about and the use of evidence-based criminal justice sanctioning.

Calls for evidence-based criminal justice policy

During the 1980s and 1990s, calls from Republicans and Democrats for greater government accountability became increasingly prominent (Hatry, 2006; Marion & Oliver, 2006; Rossi, Lipsey, & Freeman, 2004; Welsh & Harris, 2008). The focus stemmed from many concerns, but not least was the view that government agencies operated inefficiently and ineffectively. Accordingly, “performance monitoring” entered into mainstream discussions of how to increase the efficiency and effectiveness of government. By measuring critical governmental operations and activities, as well as the outcomes that these activities were intended to achieve, one could gauge agency performance and, in so doing, hold government accountable. At the same time, “evidence-based” policy and practice increasingly surfaced in political discourse as a way of signifying whether an intervention rested on credible empirical research. The central underlying argument centered on the view that, for government to be accountable, it should implement policies and practices that research has shown effectively produces intended outcomes.

Against this backdrop, the nation experienced a dramatic rise in violent crime, one that began in the late 1980s and continued until the early to mid-1990s before then steadily declining (Blumstein & Wallman, 2006). Both the federal government and states responded by substantially increasing criminal justice system funding, which led to large-scale expansion of the courts, jails and prisons, and probation and parolee populations (Marion & Oliver, 2006; Spelman, 2009; Tonry, 2004; Travis & Visher, 2005; Useem & Piehl, 2008). This expansion, and the fact that it occurred at a time when accountability ascended into prominence, naturally led to investigation of whether such growth was needed or effective and, more generally, to an emphasis on identifying evidence-based policies and practices.

Perhaps the most prominent illustration of this change in focus is the federally funded review, Preventing Crime: What Works, What Doesn’t, What’s Promising, conducted by Lawrence Sherman and his colleagues (Sherman et al., 1997). This effort involved the systematic evaluation of a large body of research on a wide range of programs, using a ranking system to assess the methodological rigor and overall quality of the research. In keeping with the review’s title, the authors concluded: “This report found that some prevention programs work, some do not, some are promising, and some have not been adequately tested” (p. v).

Other reviews and discussions, including a National Academy of Sciences study (Lipsey, Adams, Gottfredson, Pepper, & Weisburd, 2005) and reports from the Crime and Justice Group of the Campbell Collaboration (www.campbellcollaboration.org), have arrived at assessments that are significantly more pessimistic about the state of science underlying many if not most criminal justice policies, programs, and practices (see, e.g., Eskridge, 2005; Farabee, 2005; Farrington & Welsh, 2005; Forst, 2004; Goldkamp, 2008; Mears, 2007; Warren, 2008; cf. Lipsey & Cullen, 2007). As Chris Eskridge (2005:305) observed, for instance, “by and large, the crime prevention programs . . . in the United States have not been systematically evaluated” (p. 305) and, although some progress has been made, “[criminology has] yet to attain the status of a mature, evidence-based and evidence-driven science” (p. 303) (see also Clear, 2010).

To be certain, many examples of well-evaluated policies, programs, and practices exist, as these reviews attest. The more general assessment, however, is that much of what occurs in criminal justice remains unevaluated and that the methodological rigor of criminal justice evaluations frequently suffers from considerable problems. In addition, and of particular relevance here, there are few studies that compare a wide range of criminal justice approaches across a wide range of relevant outcomes. This situation makes it difficult to arrive at any definitive assessments about the wisdom of one policy, program, or practice as against another.

Criminal justice sanctioning options

One way to highlight the complexity and challenges of creating an effective, evidence-based criminal justice system is to focus on sanctioning, arguably the central or focal activity of this system. Consider the seemingly simple question: When faced with a convicted felon, what sanction—including not just a particular punishment (e.g., probation, intermediate sanction, jail, or prison) but also any of a range of possible services or treatments or participation in specific programs or interventions—should the court impose?

Many judges will want to assign a sanction that: (1) most effectively reduces offending; (2) improves other outcomes (e.g., drug abuse, unemployment); and, (3) does so for the least cost. The list of viable options to most judges may be quite extensive, especially in larger, metropolitan areas. Non-incarcерative sanctions alone, for example, include many possibilities, such as probation or supervision, drug testing, boot camps, electronic monitoring, curfews, fines, employment requirements, and community service, all of which can vary in kind or intensity (Lipsey & Cullen, 2007; Spelman, 1995; Tonry, 1998). The variation in kind or intensity alone means that two or more “similar” interventions may differ appreciably. For example, supervision that involves weekly check-ins arguably differs substantially from supervision that requires monthly check-ins. At the same time, incarceration is not a simple or uniform policy. For example, many different sentence lengths can be assigned for any given offense, raising the question of which lengths produce the greatest returns in affecting recidivism (Nagin, Cullen, & Jonson, 2009) and in achieving retribution and other outcomes.

If local context and constraints played no role in judges’ calculations, they would want a list of all possibilities, relative to costs. The task then
would be to select the one that, based on available research, or the “evidence base,” would achieve the goal of maximizing impacts across a range of outcomes for the least cost. In reality, local context and constraints typically do matter (Weidner, Frase, & Schultz, 2005). Some jurisdictions may have drug treatment programs, for example, and others may not. For that reason, judges may want a list specific to their jurisdiction that details the relative impacts and costs, across various outcomes, for the constellation of possibilities available locally. This list would be essential even if sentencing guidelines existed that set parameters on sanctioning since guidelines typically allow the courts latitude in assigning non-incarcerative sanctions (Ashworth, 2005; Morris & Tony, 1990).

Observe that without such a list, any decision about particular sanctions would necessarily reflect non-evidence-based considerations, such as personal punishment philosophy, a preference for certain types of interventions, or the ability of some programs to market their offerings better than others. That is not to say that court officials arbitrarily mete out sanctions. They typically weigh and balance many factors, such as the likelihood that convicted individuals will reoffend, the perceived quality of certain intermediate sanctions, and the availability of specialized treatments or services (e.g., drug abuse or mental health counseling, employment training) (Ashworth, 2005; Forst, 2004; Spohn, 2009). Regardless, without empirical research that compares the relative impacts of these various sanctioning options across a range of outcomes, decisions to assign specific sanctions perforce are not evidence-based. That is, they are not based on rigorous empirical studies documenting that, say, one program is markedly better at reducing recidivism and drug abuse, for a given cost, relative to other programs that could be considered.

Research gaps that inhibit evidence-based sanctioning

The question thus arises: What is the evidence base for court sanctioning options in general or for specific jurisdictions? Evidence along both dimensions is critical. For example, drug courts in general may be effective (Wilson, Mitchell, & MacKenzie, 2006; cf. Merrall & Bird, 2009), but that fact is irrelevant if, in a specific jurisdiction, a drug court does not exist (and so is not available as an option) or if it is poorly implemented. In the latter case, evidence from, say, a meta-analysis of drug court effectiveness would not provide a safe platform from which to make assumptions about how well a local drug court reduces recidivism relative to other available sanctioning options. If quality implementation could be assured, safer grounds for making such an assumption would exist. Unfortunately, poor or insufficient implementation appears to be far more the statistical norm rather than the exception (Lipsey et al., 2005; Mears, 2010).

Below, the importance of research that compares the relative effectiveness of multiple sanctioning options is considered first. Second, the importance of similar research that examines effectiveness along a range of relevant outcomes will be discussed. This discussion will draw attention to the necessity of research that compares multiple sanctions along multiple outcomes for identifying evidence-based sanctioning decisions and policies. Such research should, as is discussed, involve evaluations of various sanctioning options in general and of these options as implemented in specific jurisdictions. Although sanctioning is the main focus of the article, the conclusion highlights that the observations extend equally well to other system activities, including those aimed at reducing crime (e.g., community policing, situational crime prevention programs, offender registries) and managing correctional system populations (e.g., administrative, classification, security, or programming efforts).

Comparative research on multiple sanctioning options

In deciding how to reduce recidivism and improve other offender outcomes, to say nothing of reducing crime, a vast array of sanctioning options exist in most places. Any rational approach to sanctioning requires carefully selecting those options that are most effective for the system, context, population, or individual being targeted. Unfortunately, no bank of knowledge or empirical research exists that systematically compares multiple sanctioning options either in general or in specific jurisdictions. As a result, it is impossible for most if not all criminal courts, or the jurisdictions or states within which they reside, to claim legitimately that they use evidence-based sanctioning. Indeed, they typically have no empirical foundation from which to conclude whether a sanction is evidence-based.

To be certain, many sanctions have been evaluated. Indeed, the literature on the effectiveness of many of them (e.g., fines, community service, probation, death penalty) is vast. There is also a growing body of evidence surrounding the most effective correctional interventions strategies (see, e.g., Andrews et al., 1990; Cullen & Gendreau, 2001). The bulk of this work, however, does not compare more than one sanction at a time. For example, a study might evaluate whether boot camp participants recidivate less than those who do not (it appears that there is no difference—see, e.g., Farabee, 2005). To this end, it might use an experimental design involving random assignment to the boot camp and a control condition, enabling scholars to trust more that any identified effects, or lack thereof, are real.

Consider, however, the problem that a judge would face in allowing the results of the study to inform his or her sanctioning decisions or, more generally, that local and state policymakers would face in devising sanctioning policies. Specifically, any results would speak to comparisons between the boot camp and the “average” offender’s “typical” experience. That experience encompasses a range of possibilities. As a result, any difference between the boot camp and control group depends entirely on the constellation of possibilities available locally and their relative effectiveness. (Assignment to nothing would not be appropriate because, in many jurisdictions, nothing is typically not an alternative. Rather, individuals would receive some type of sanction.) These options may vary greatly from one jurisdiction to another, in turn affecting whether a significant treatment effect surfaces. Put differently, if a treatment effect emerges, it is entirely relative to the range of sanctions that served as the basis for the comparison [i.e., the control condition]. It bears emphasizing that many sanctions are intended for sub-sets of the offender population and so, in an experiment, randomized assignment to control groups would come from these sub-sets. The corollary, however, still would hold—that is, the individuals randomly assigned to control groups would typically experience a range of sanctions, not just one.

An illustration can highlight the problem. Consider a study that shows that an entrée one orders at a restaurant tastes better than all other entrées combined (or that it does not). The problem with the study is that the comparison is to a mish-mash of dishes, some of which might be awful and some of which might be quite good. In addition, whether the dish emerges as significantly better or worse than the comparison would depend on the range and quality of the other dishes. More helpful, therefore, would be to have an absolute ranking of the taste of each separate entrée that then could be used to compare the taste of each separate entrée that then could be used to make comparisons of the entrée with specific dishes at the restaurant or with dishes at other establishments.

This same logic applies to sanctioning. An experiment may show that a given sanction is better than some other. The same experiment with the same sanction, however, may identify no effect in a different jurisdiction, depending on the types of sanctions available to serve as the control condition. In addition, even if a significant sanction effect is identified, all that can be concluded is that the sanction is better than the jurisdiction’s mish-mash of “business-as-usual” sanctions. That information is of limited use because it obscures the fact that some of the particular sanctions that form the “mish-mash” might be much better or much worse than the sanction of interest. Here, again, it would be far better to have an absolute ranking system. That is,
ideally a common scale or metric about the absolute effectiveness of each sanctioning option would be available. This ranking in turn could be used to judge the relative effectiveness of one type of sanction as against another.

Table 1 illustrates the point. It presents, in rows, some of the more common types of sanctions available to the courts and, as discussed in the next section, it presents some of the prominent criminal justice outcomes that might be used to compare the effectiveness of these options (i.e., the columns). The sanctions are considered first.

Some offenders may receive no sanction. That policy is, essentially, an implicit one. Most jurisdictions operate with no precise estimate of the true prevalence of crime and they all capture only a sub-set of actual offenders (Blumstein & Beck, 1999). Those that allocate more funding to law enforcement typically will capture more offenders and so may sanction a greater percentage of actual offenders as compared to those that allocate proportionately less funding to law enforcement. Of course, some arrested offenders receive no sanction; the precise percentage will vary across jurisdictions and implicitly constitutes a policy itself. Among the sanctioned offenders, some may receive a form of community supervision, with the intensity of the supervision (e.g., low, medium, or high) varying depending on their assessed risk. Fines, community service, treatment services, specialized courts, and electronic monitoring are among the more common community-based sanctions, separately or in combination, which can be imposed as sanctions. The extent to which each such sanction is used—that is, the relative distribution of cases that receive each type of sanction—also constitutes a policy decision.

Boot camp, jail, and prison are all additional sanctioning possibilities, with each entailing a different type of incarcerative experience that varies both in the duration and in the type of environment, programming, services, and treatment available. Notably, changing the length or the intensity of a sanction can result in a qualitatively different punishment. Indeed, that assumption underlies the very idea of applying differing durations for some sentences. The assumption is that, say, a ten-year sentence achieves more retribution or a greater reduction in crime or recidivism as compared with a five-year sentence. Whether that assumption is true is, of course, an empirical matter (see Nagin et al., 2009). Related to that observation is the fact that different approaches to structuring sentencing decisions exist. For example, states can allow indeterminate sentencing, where judges or parole boards have the ability to determine when a term of incarceration should end, or determinate sentencing, where guidelines largely dictate the sentence length. Such approaches themselves constitute types of sanctioning policies, and so the question is salient for them as well: to wit, what is the evidence base for asserting their effectiveness?

Decisions about appropriate incarceration rates—which constitute sanctioning policies at a macro level—vary greatly across states, as reflected in the dramatic variation in rates across the country (Sabol et al., 2009). The variation itself illustrates the fact that, net of crime rates, states use incarceration in different amounts (Gottschalk, 2006; Greenberg & West, 2001). Some believe that incarcerating 500 or 600 offenders per 100,000 residents is appropriate or effective, while others believe that incarcerating 100 or 200 per 100,000 residents is the better way to proceed. The critical policy question, however, is whether higher rates create greater reductions in crime and whether they do so more than other policies would. To date, the literature on this question is mixed. As a general matter, it appears that a greater use of incarceration leads to less crime; it is not, however, clear that it achieves this outcome more so, or more cost-effectively, than other sanctioning approaches (see, e.g., Spelman, 2009).

Sanctioning policies extend to the post-incarceration experience as well. For example, some states allow for unconditional releases from prison, where ex-prisoners have no supervision. Others have release policies that mandate supervision; still others allow correctional systems to exercise discretion about who receives supervision and about the nature and intensity of supervision. In many states, drug-testing is a critical part of parole, but in others it is not (Travis & Lawrence, 2005). Here, again, the different approaches constitute distinct sanctioning options, ones that ideally would be selected based on empirical evidence of effectiveness.

To assess the relative or the absolute impacts of these different sanctioning options, their effects on different outcomes must be examined. The columns present those that arguably stand as among the most critical outcomes relevant for assessing the effectiveness of sanctions. In the next section, these specific outcomes are discussed.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Framework for comparing criminal justice sanctions and their relative effectiveness across a range of outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Types of Sanctions</strong></td>
<td><strong>Crime (1-5)</strong></td>
</tr>
</tbody>
</table>

* 1 = harmful, 2 = no effect, 3 = small positive effect, 4 = moderate positive effect, 5 = large positive effect.
Here, the focus is on the fact that most extant evaluations compare only one of these sanctions either with all others (e.g., specialized court versus “business-as-usual” sanctioning) or with one other type of sanction. They do not, as reflected by the series of question marks within each column, systematically compare the effectiveness of the full range of sanctions to one another.

In addition, studies vary greatly in the measures and statistical procedures that they use to gauge effectiveness. As a result, it can be difficult to assess how the magnitude of difference in one study compares with that of another. Meta-analytic techniques can help address this problem (Lipsey & Cullen, 2007). Regardless of whether such techniques or others are used, however, any systematic and valid comparison of different studies requires use of a standardized scale or metric or, better yet, one that provides an absolute basis for ranking sanction effectiveness. Developing such a scale or metric is, admittedly, not easy. In the table, a “1 = harmful” to “5 = large positive” effect is suggested simply because it accords with the types of scales used in policy reviews (see, e.g., Sherman et al., 1997; see also Farrington, 2003). Even with such a scale, it is critical that evaluations compare multiple sanctions at the same time. Otherwise, the “entrée” problem surfaces—that is, comparisons that over-simplify and potentially distort assessments of the effectiveness of any of a range of sanctions will prevail.

If a completed table existed for these types of sanctions, then judges, as well as states, would have an evidence-based foundation from which to assign sanctions to offenders. Were such a table completed for the specific constellation of sanctions available in a particular judge’s jurisdiction, he or she would have an even stronger basis for making evidence-based sanctioning decisions because the evidence would concern the sanctions as actually implemented rather than “average” sanctions that were implemented elsewhere. There is, however, no body of studies as yet that allows this table to be completed.

**Comparative research on multiple outcomes**

It is a central axiom of evaluation research that the effectiveness of a policy or program, or in this case a sanction, be assessed with respect to all of its goals (Rossi et al., 2004). So, if a sanction—and any attendant restrictions, treatments, or services it entails—is intended to reduce recidivism, drug use, abuse, or addiction, homelessness, or mental illness, all intended outcomes should be examined and, ideally, weighted according to their priority. Otherwise, a highly incomplete assessment of effectiveness emerges. For example, if an evaluation shows that a drug court fails to reduce recidivism, it might be concluded that the drug court merits little additional support. It might well decrease drug addiction, however, and so possibly merit continued support despite the lack of impact on recidivism.

This consideration would matter little if all criminal justice programs and policies aimed to achieve the same goals, but they do not, as shown in Table 1. Certainly, many aim to reduce crime and recidivism. They typically aim, however, to achieve some level of not, as shown in Table 1. Certainly, many aim to reduce crime and recidivism and to recidivism. They typically aim, however, to achieve some level of not, as shown in Table 1. Certainly, many aim to reduce crime and recidivism. They typically aim, however, to achieve some level of

The importance of sanctioning impacts on diverse outcomes can be seen through inspection of the table. Ideally, scholars would be able to say which of the range of listed sanctions produce not only greater reductions in crime rates but also which ones achieve greater improvements in the other outcomes. In fact, few studies exist that compare, in relative or absolute terms, the impacts of diverse sanctions on crime rates, to say nothing of simultaneous assessments of crime rates and of other relevant outcomes. Of course, researchers have investigated the impacts of some sanctions on recidivism. This work is useful for clarifying what the potential impacts on recidivism may be, but it says little about their effects on crime rates (Rosenfeld, 2008). Even if the focus were restricted exclusively on recidivism as an outcome, the limitation discussed earlier remains—namely, few studies simultaneously compare the impacts of multiple sanctions. This situation creates a problem when the goal is to select a sanction that will produce the greatest improvement in, say, recidivism. The problem, however, is compounded by the fact that comparative studies of the impacts of diverse policies on other outcomes is lacking.

What about some of the other outcomes? Retribution stands as another critical outcome that should be considered in developing an evidence base for one policy as against others. When a crime occurs, society and victims typically expect a retributive response. The offender is supposed to suffer. How exactly and to what extent they are to suffer is not always clear. Indeed, considerable debate exists about the role of retribution in sentencing and how exactly it should be achieved (see, e.g., von Hirsch & Ashworth, 1992). Amplifying this problem, however, is a lack of evidence on the comparative severity of various sanctions. Generating such evidence is complicated by the fact that any assessment of retribution ultimately is impressionistic and the related fact that perceptions of punishment severity will vary. For example, one person may perceive a year in prison to be a terrible burden while another may see it as the cost of doing business (Spelman, 1995). Similarly, how offenders perceive punishments may differ from how the general public views them, and both groups may view them differently than do policymakers (Roberts, Stalans, Indermur, & Hough, 2003; Rossi & Berk, 1997). To illustrate, intermediate sanctions were billed as offering a clear middle ground between probation and prison, yet studies, including Spelman’s (1995), indicated that many sanctioned offenders view intermediate sanctions as more punitive than prison (May & Wood, 2005; Tonry, 1998).

Some sanctions, in short, clearly involve different amounts or types of retribution. Without comparative empirical studies—nation- ally or in specific states or jurisdictions—scholars are unable to credibly assert that one type of policy is more severe than another, much less that it is more severe and achieves greater reductions in recidivism or crime rates. The limited comparative research on retribution assumes particular importance in a historical context in which more severe punishment arguably has served as the central guiding feature of sanctioning policies in the United States, especially during the 1980s and 1990s (Eskridge, 2005; Gottschalk, 2006).

Among other things, it creates the risk that policies will be overly or insufficiently punitive if policymakers misread public opinion (Cullen, Fisher, & Applegate, 2000). No single accepted approach to quantifying sanction severity yet exists. Even so, because retribution is central to the criminal justice system and because many sanctions are assumed to be more severe than others, it is important that some ranking system be developed to compare them. The alternative approach, which arguably is more problematic, consists of allowing the relative retributiveness of different sanctions to be assumed.

If the focus is extended beyond crime rates, recidivism, and retribution, still other outcomes merit attention. For example, research on prisoner reentry documents that ex-prisoners have many problems that, if unaddressed, may increase the likelihood that they not only will recidivate but also will develop problems that ultimately harm them and society. Petersilia (2005) and others (e.g.,
Critical steps toward identifying evidence-based sanctions

In developing an evidence-based approach to sanctioning, a series of studies are needed. First and foremost, and as highlighted above, studies that simultaneously evaluate multiple sanctions and, at the same time, their impacts on multiple outcomes are needed. That step, however, is insufficient on its own. Additional steps are needed to enable policymakers and practitioners to make strides toward implementing evidence-based sanctions. These include, but are not limited to, the following steps.

Second, Table 1 ideally would be completed using an empirically-based assessment of not only the impacts of various sanctions but also their cost-effectiveness. Information about cost is essential. Consider, for example, that many jurisdictions may not be able to afford a particular intervention but they may well have a problem that they wish to address. Ideally, jurisdictions would be able to compare different criminal justice sanctions and select those that achieve a maximum impact at the lowest cost. The larger goal, however, would be to have cost-benefit rather than cost-effectiveness assessments. Cost-effectiveness analyses allow one to make comparisons between policies that have the same goals. As discussed above, however, many criminal justice sanctions have non-overlapping goals. For example, one may focus on providing employment assistance while another may focus on providing housing assistance. Both may aim to reduce recidivism, but they differ with respect to these other goals. Cost-benefit analyses address this situation by monetizing all benefits (as well as costs), which creates a common metric for making comparisons (Kee, 2004). Given the range of goals and impacts associated with many criminal justice sanctions, to say nothing of criminal justice policies more generally (Cohen, 2005), cost-benefit analyses are needed if policymakers are to have a credible basis by which to select one sanction over another based on efficiency considerations.

Third, Table 1, and any accompanying cost-benefit table, ideally would be completed not only for policies in general but also for policies as they are implemented in specific locales. The former focus is essential for documenting whether particular efforts, such as a particular type of intensive probation, can be more effective than others. The latter focus is essential for documenting whether such efforts indeed—that is, in practice—are more effective than other options available in local jurisdictions. For example, it is entirely conceivable for research to find that a policy of some type, such as community policing, achieves expected outcomes in one area but that implementation or other problems inhibit its effectiveness in others (Innes, Abbott, Lowe, & Roberts, 2009; Worrall & Kovandzic, 2007). In addition, in some communities, highly effective crime prevention policies may exist. In such cases, it would be harder than in other communities to detect an impact of community policing on crime relative to that associated with other local efforts to prevent crime. The same is true of any of a range of sanctions, especially those that involve multiple features (e.g., intensive supervision, urine analysis tests, drug treatment).

Fourth, criteria must be developed for defining a sanction; otherwise, it is not possible to make valid comparisons. This step is critical because considerable variation exists in the types of sanctions that exist. Drug courts, for example, have become increasingly popular—in two decades, beginning in 1989, over 2,000 drug courts were created (Huddleston, Marlowe, & Casebolt, 2008)—yet many of them differ greatly in the target population they serve, the setting in which they operate, the services and treatments they provide, and the amount and quality of supervision, drug testing, and court involvement (Butts & Roman, 2004). The problem is not simply that relatively few rigorous impact evaluations of these courts have been conducted (Merrall & Bird, 2009). Rather, it is that the generalizability of any one study is highly circumscribed because of what may be unique features of the evaluated court.

Fifth, criteria must be developed for determining what constitutes a significant improvement in a given outcome; the risk otherwise is that arbitrary determinations are made about the worth of a given sanction. For example, a five percentage point reduction in the crime rate might be deemed significant in some areas of the country, but not so in others. Statistical analyses provide little assistance here. They can provide insight into whether an observed difference between one sanction and some alternative is not purely an artifact of the samples used. Assuming that statistically significant differences are identified, the challenge of arriving at some standard for adjudicating when an improvement of some type is noteworthy still remains.

Sixth, a standardized set of outcome measures is needed to make valid cross-study comparisons of different sanctions and their effectiveness. To date, evaluations of sanctions employ a wide range of different measures rather than ones that are standardized. Consider the possible ways to operationalize recidivism. One can use different time periods (e.g., 1 year, 2 years, or 3 years) in which to follow individuals after their sanction is complete. A binary measure (recidivated or not) could be used. Alternatively, a count (X crimes) or a rate (e.g., X crimes per year) could be used. One also might examine offending trajectories, such as the way in which patterns of offending change over time (Piquero, Farrington, & Blumstein, 2003). Crimes could be combined to create an index; many approaches, some more defensible than others, exist for doing so (Osgood, McMorris, & Potenza, 2002). Alternatively, crime categories (e.g., violent, property, drug) or specific crimes (e.g., rape, arson, terrorism, assault) could be examined. These and other possibilities populate the sanction evaluation landscape, and so make it difficult to arrive at valid comparisons. Use of a common set of measures and approaches to combining them would help address that problem. At the same time, it would allow for a better test of the impacts of different sanctions. It is, for example, possible that a particular sanction may reduce the frequency of offending but not its absolute occurrence. An evaluation that employed a comparison group of some type would fail to detect this possibility if it used only a binary measure of effectiveness (e.g., recidivated or not).

Finally, as a general matter, more rigorous impact evaluations of different criminal justice sanctions are needed. Many published evaluations suffer from critical flaws that substantially delimit their generalizability (Eskridge, 2005; Farrington, 2003; Farrington & Welsh, 2005; Goldkamp, 2008; Lipsey et al., 2005; Mears, 2007; Merrill & Bird, 2009;Sherman et al., 1997). Selection bias, samples that are too small to detect meaningful differences, and missing data are, for example, all too common in such evaluations. In addition, many impact evaluations generate findings that result from unusually complete or high-quality implementation of a given sanction or from insufficient implementation of it (Rossi et al., 2004). Accordingly, in developing the evidence base for different sanctions, criteria are needed for determining the extent to which one can credibly anticipate that similar results would occur if the sanction of interest were implemented in some other area. Experiments

Travis & Visher, 2005) have documented that ex-prisoners typically have spotty employment records that, when combined with a criminal record, make it difficult to obtain work, and they also typically are at high risk of homelessness, drug abuse, and suffer from a range of serious physical and mental illnesses. In addition, they frequently come from and return to families characterized by considerable dysfunction and communities in which disadvantage is concentrated. A range of other challenges, such as laws that preclude ex-prisoners from certain types of jobs or receiving welfare benefits, increase the likelihood that the transition back into society will fail on many different fronts. Precisely for this reason, many reentry programs aim to target these areas of need (see, e.g., Solomon, Waal, Van Ness, & Travis, 2004; Travis, 2005). Accordingly, any evidence-based policy focus would want comparative information about each reentry program and its impacts on these different outcome areas.
cannot always be conducted (Clear, 2010), and for that reason such criteria are especially needed for results emanating from quasi-experimental studies. These steps do not exhaust the possible ones needed to create an evidence base for identifying sanctions that are better than others in producing more improvement along more dimensions for the least cost. Nonetheless, they serve to create a foundation of evidence that to date is lacking as policymakers devise ever-new sets of strategies for sanctioning offenders.

Conclusion and implications

In the face of dramatic growth in America's criminal justice system and calls nationally for using evidence-based policies there stands an odd fact—precious little evidence exists to claim that the sanctions currently in use are effective (Clear, 2010; Eskridge, 2005; Farabee, 2005; Mears, 2007, 2010). This is not to say that criminal justice policy is in any obvious way less evidence-based than many other areas of social policy (e.g., welfare, public health, homeland security). Many medical interventions, for example, lack any clear foundation in scientific evaluations of impact (Gawande, 2007, 2009; Groopman, 2007; Sherman, 2003). Such an observation does not, however, vitiate the importance of undertaking efforts to improve the effectiveness and efficiency of the criminal justice system.

Studies certainly exist that show that one type of sanction or another is more effective than, say, the traditional approach to processing a certain type of offender (i.e., “business-as-usual” sanctioning). In addition, much basis for optimism can be found in a number of recent reviews of criminal justice policy (e.g., Cullen, 2005; Drake et al., 2009; Farrington, 2003; Howell, 2009; Mackenzie, 2006; Mears, 2007, 2010; Sherman et al., 2002; Tonry, 2009). As these same reviews have emphasized, however, and as Clear’s (2010) Presidential Address recently underscored, the research gaps that remain are considerable. Indeed, as was argued above, few studies to date provide insight into the relative effectiveness of one type of sanction versus the many other types of sanctions that exist in any given jurisdiction. In addition, the results of such studies typically should not be generalized because they rely on comparisons specific to the jurisdictions in which the studies were undertaken and because they do not compare all relevant outcomes. Put differently, even the most well-done experimental studies generate results that may provide insights of questionable policy relevance (Bardach, 2009; Clear, 2010; Heckman & Smith, 1995; Rossi et al., 2004).

To be sure, the task at hand is daunting. This article has outlined several steps, however, that, if used to guide future research, would help set criminal justice research on a path toward identifying evidence-based policy. (1) Studies must evaluate multiple sanctions and multiple outcomes simultaneously to provide a basis for saying whether one type of sanction is more effective than any of the range of options available. (2) Researchers should compile estimates not only of impacts on outcomes but also of associated costs; otherwise, little to no basis exists for identifying how much more or less efficient one sanction is versus another in achieving one or more outcomes for a given cost. (3) The effectiveness of specific sanctions in general and of their effectiveness in specific areas should be established; without the latter information, it is entirely possible for a jurisdiction to use a sanction that, for any of a variety of reasons (e.g., the amount or quality of resources available for implementing the sanction as designed), is not effective for the population or area in which it is being used. (4) Standards for defining a sanction must exist so that improper inferences are not drawn. Among the thousands of drug courts that now exist, for example, some vary greatly from others. Defining sanctions, or identifying sub-types of each (e.g., specific types of drug courts) is essential for identifying this fact and, in turn, delimiting generalizations from impact evaluations. (5) Criteria are needed for determining when a sanction's effect is, in any meaningful way, an improvement over some alternative. The absence of such criteria means that trivial differences may be viewed as significant and that potentially important differences are treated as insignificant. (6) Uniformity in outcome measures that are used is needed to facilitate appropriate comparisons of results across different evaluations. (7) Not least, more and better impact evaluations of sanctions—using measures that can overcome the limitations of extant data sources—is essential if an evidence base is to emerge that credibly identifies effective sentencing options.

The research gaps discussed focused exclusively on sanctions, as well as the services and treatments that may be linked to specific sanctions. The problem is, however, considerably larger. This article did not discuss, for example, the full panoply of laws, policies, programs, rules, protocols, practices, and the like that collectively constitute the criminal justice system’s efforts to achieve its goals. The gaps identified here exist for almost all such efforts (see, generally, Clear, 2010). Consider community policing—despite the considerable funding and support for community policing, it remains unclear how well it reduces crime and, as importantly, how well it reduces crime as compared with other approaches to crime prevention (Worrall & Kovandzic, 2007). It also remains unclear how large an impact it has on community residents’ perceptions of safety and how any impacts compare to alternative approaches to achieving that goal (Greene, 2000). Similar observations can be made about many other activities within the criminal justice system, ranging from different approaches to crime prevention to a range of strategies for managing offenders in prison and after release. As with sanctioning policies, an evidence-based approach to criminal justice writ large requires the development of a body of research that systematically and simultaneously compares different approaches to achieving different goals.

It is readily apparent that creating the research foundation for systematically making comparisons about the relative impacts and costs of a wide range of criminal justice policies across diverse outcomes constitutes a tremendous challenge. The failure, however, to develop the relevant knowledge base means that the federal government, as well as state and local governments, must resort to arbitrary grounds for selecting certain sanctions or policies over others. By failing to select ones that produce the largest impact at the least cost, they inevitably must make mistakes that adversely affect society through increased crime, unemployment, homelessness, mental illness, and other outcomes.

Notwithstanding these observations, it is important to recognize that it may not be possible ever to achieve a state of knowledge in which the relative effectiveness, including cost-effectiveness, of a range of sanctions across a range of outcomes, is credibly established. Even so, clarity about the importance of this goal serves at least two important purposes. First, it helps to place into context what it means to refer to “evidence-based” policy (see, generally, Clear, 2010). For example, experimental studies are typically construed as providing highly credible evidence about whether a policy works. The generalizability of many experiments, however, is limited based on the comparison involved and the range of outcomes studied. That insight surfaces more readily when one tries to situate the results of any given experiment within a matrix that includes a range of potential policy interventions and a range of relevant outcomes. Second, it provides a clear framework for pointing to the types of research that might create greater advances in the needed evidence base for improving criminal justice policy on a systemic level. It highlights, for example, that any given policy always occurs within a larger policy context, one in which multiple other policies may exist and thus need to be included in any assessment of impact and also in which multiple outcomes are relevant (Clear, 2010). Accordingly, evaluations must take such context into account or risk creating highly circumscribed studies that produce results of little relevance to effecting large-scale improvements in the efficiency and effectiveness of the criminal justice system.
To be clear, it should not be viewed as unrealistic that, with proper funding and focus, substantial advances could be made toward identifying the relative impacts of different policies across a range of outcomes. In recent decades, as discussed further below, relatively little funding has gone toward evaluating criminal justice policies. Increased funding would create considerable room for undertaking more systematic and comprehensive evaluations.

In addition, the goal of identifying relative impacts of different interventions across diverse outcomes is one that has been pursued, however imperfectly, in other policy arenas—such as education (see, e.g., Means, Padilla, & Gallagher, 2010) and welfare (see, e.g., Alber & Gilbert, 2010)—suggesting that it may be possible to undertake a parallel effort in criminal justice. Over a decade ago, for example, McQuay and Moore (1997) described efforts in medicine to establish a credible comparative basis for identifying when a given intervention is effective and how well it works “in general and compared with placebo, no treatment, or other interventions that are currently in use” (p. 712). Other studies in the public health field have emphasized the importance of examining multiple outcomes, although clearly that ideal frequently may not be reached (see, e.g., Kraemer, Glick, & Klein, 1991). Thus, the comparison includes many different alternative interventions instead of faulty intervention implementation but also to the likelihood that the intervention under study will be implemented more systematically and comprehensively in practice.

Studies of criminal justice interventions across diverse outcomes is one that has been pursued, more systematic and comprehensive evaluations. Little funding has gone toward evaluating criminal justice policies. Funding alone will not be sufficient, however. Contribution from the scholarly community is critical. Notably, the American Society of Criminology and the Academy of Criminal Justice Sciences—the two most prominent criminological associations in the United States—have become actively involved in promoting policy evaluation research, and considerably more progress is needed (Clear, 2009, 2010). Perhaps more important is the creation of the infrastructure necessary for building a coherent knowledge base (Clear, 2009, 2010; Eskridge, 2005; Mears, 2007). Different possibilities exist, but the federal government likely needs to spearhead the development, funding, and organization of criminal justice research nationally. A national institute, regional, state, and local policy councils could be created as well that would work collectively to develop research efforts that would complement and build on one another (Blumstein, 1997, 2008; Blumstein & Petersilia, 1995; Clear, 2010; Mears, 2010; Petersilia, 1991). Regardless of the arrangement, efforts must be directed toward facilitating comparative studies along the lines outlined here. Absent such efforts, progress toward developing an evidence-based science of criminal justice policy and sanctioning will remain at a nascent stage of development.

References
