D

Decision Making Theory

► Integrating Rational Choice and Other Theories

Defiance and Motivational Postures

Valerie Braithwaite
Regulatory Institutions Network (RegNet),
The Australian National University,
Acton, ACT, Australia

Overview

Motivational postures are the signals that people send to authorities, including criminal justice authorities, to indicate their liking for that authority and their willingness to defer to the authority's rules and processes. These signals change in response to the actions of authority. Five motivational postures describe the way in which individuals and groups position themselves in relation to authority. The posture of commitment represents belief in the authority, its goals and purpose. The posture of capitulation involves acquiescing to authority because it brings the least trouble. Resistance is a posture of protest and anger about how an authority operates. Disengagement is a posture of withdrawal, of severing the relationship with authority to the point where the authority is irrelevant. Game playing challenges authority by circumventing rules and laws while appearing to do what is expected. These postures are openly shared and coexist. They combine to form a complex signaling system that authorities can read and respond to in an emotionally intelligent way to form a more effective criminal justice system.

Fundamentals

What Are Motivational Postures?

When authorities take action to create order or enforce law, individuals display a range of responses. The response of prime concern in criminology is whether people obey the law. Yet other responses also play a role in the success of an authority in crime control and prevention. These responses include whether individuals trust authority, believe authority is legitimate, or cooperate with authority. Motivational postures are related to this class of responses.

Motivational postures are socially shared thoughts and feelings that become organized into well-crafted signals to authority about different kinds of approval of and deference to authority (Braithwaite et al. 1994; Braithwaite 1995, 2003, 2009). The distinctiveness of the motivational postures concept revolves around their multifaceted nature, providing individuals with a suite of responses. They can be used within a single encounter with authority, across several encounters, or even across a range of authorities. Because motivational postures are outward displays of

G. Bruinsma, D. Weisburd (eds.), *Encyclopedia of Criminology and Criminal Justice*,
DOI 10.1007/978-1-4614-5690-2,

approval or deference, they can be used to communicate about the quality of relationships and negotiate new relationships with authority.

Social Distancing

At the heart of motivational posturing theory is the notion that people choose how much social distance they place between themselves and an authority, just as they choose how much social distance they place between themselves and another person (Bogardus 1928). Sometimes people may be prepared to approach an authority, to listen and be open to its message. Other times they may keep their distance, being wary of the authority's purpose and turning a deaf ear. As people change their social distance, they are adopting positions that best protect them from an authority's power. In Harris's terms, individuals distance themselves from authority in such a way as to protect or enhance their ethical identity (Harris 2011). At the same time, authorities want individuals to position themselves closely, particularly when an authority wants the public to be responsive to its message (e.g., during a natural disaster or security crisis). Motivational postures are the signals that are sent by individuals and groups; they can be read by authorities and can be used to establish better working relationships.

Some motivational postures signal alignment with authority; others are oppositional. Within most of us, postures coexist. Early socialization teaches us that authority has the power to both help and hinder as we find our way in the world. Authority maps out paths of safety and success. Authority also hinders through rules and their enforcement. Authority in this sense always poses a potential threat to our freedom to act as we want. In response to that threat, we feel ambivalence. We see value in positioning ourselves closely to signal we are in accord with the authority. Then, as we observe the authority acting in ways with which we disagree, we increase our distance as we become less certain that our thoughts and actions are at one with the authority.

Five Motivational Postures

Five motivational postures have been identified empirically with some consistency across the domains of different authorities. *Commitment* and *capitulation* are postures that represent willingness to go along with authority. *Commitment* conveys a belief that the authority's purpose is sound and that, in principle, the authority and its goals should be valued and supported by everyone. Commitment is a posture that enables individuals and groups to go beyond compliance, to do more than an authority expects or asks in the interests of furthering the accomplishment of shared goals.

Capitulation is the posture of doing what is asked, without necessarily understanding or caring about purpose and goals. Similar to McBarnet's (2003) usage, capitulation reflects acquiescence to the powers that be. Capitulation in the posturing context also incorporates acceptance of the idea that authority has superior knowledge and knows what is best.

Commitment and capitulation are postures that signal alignment with authority and openness to cooperation, be it through shared beliefs or acquiescence. These are the most common postures in a healthy democracy. At the same time, however, people may respond with anger or annoyance when they are directly affected by the decisions and actions of authorities. *Resistance* is an expression of hostility toward an authority. When displeased with how an authority is performing, the posture of resistance communicates grievance, usually that the authority treats people in a manner that is unreasonable and unfair. Resistance is a posture that is less likely to take issue with the broader purpose of the authority and more likely to focus on the way the authority uses its power. Acceptance of the authority's purpose gives rise to hope that the authority will mend its ways in the face of criticism.

Once hope is abandoned, the more socially distant postures emerge. Individuals conclude that change will not occur and that their interests – or those of others – would be better served without the authority. Through the posture of *disengagement*, individuals withdraw from having any relationship with the authority. They take no notice of what the authority says or does. The posture of disengagement communicates rejection of the authority's

goals and processes. Disengagement may be a posture of anomie (Durkheim 1897[1952]) where people have lost meaningful connection with the norms and values of the authority – and the authority with them. As a result, individuals live their lives apart from the authority and remain impervious to its powers.

While disengagement has a degree of fatalism about it (e.g., the authority will do what it will do and I am not going to have any part in it), *game playing* has a combative agenda. The objective is to outsmart the authority and assert independence over the authority while technically playing within the rules. It is the posture that gives rise to creative compliance (McBarnet 2003). The posture of game playing, while paying attention to the letter of law, shows little respect for the spirit of the law. In adopting the posture of game playing, individuals cleverly sidestep deference to the authority.

Using Motivational Postures and Signaling Defiance

We may move from one posture to another in response to what an authority says or does, or we may simultaneously embrace two or more postures. We may feel committed to the goals of an authority (e.g., preventing terrorism) but at the same time disapprove of how the authority enforces the law (e.g., abusing human rights). We may be prepared to capitulate to an authority in which we have little interest (e.g., paying tax) but fall into a pattern of game playing because it is the norm among our peers (e.g., signing up to tax avoidance schemes).

As postures wax and wane in response to circumstance, it is nevertheless the case that postures overall will vary in their salience. Authorities govern effectively in democracies because the accommodating postures of commitment and capitulation tend to be more salient and dominate the postures of resistance, disengagement, and game playing, at least for most people most of the time. On occasions, however, the oppositional postures of resistance, disengagement, and game playing may assume dominance over the accommodating postures of commitment and capitulation. Oppositional posturing represents defiance (Braithwaite 2009). Defiance sends a message to authority that communicates a state of overriding emotional and rational rejection of authority. Authority sees defiance commonly as one unified whole that needs to be squashed. Motivational posturing theory allows us to approach defiance with a more analytic and discriminating lens. Defiance can be resistant or dismissive. Importantly, the message of each type of defiance is different, as is the most appropriate response.

When the posture of resistance dominates postures of commitment and capitulation, resistant defiance is displayed (Braithwaite 2009: 113–115, 262–269). Resistant defiance is driven by grievance about how the system is working. Responding to grievance and improving the integrity of the system, particularly through procedural reforms such as dealing with people more fairly, respectfully, and openly, reduce levels of resistant defiance.

When the postures of disengagement and game playing take hold, a different kind of defiance emerges – dismissive defiance (Braithwaite 2009: 113–115, 262–269). Dismissive defiance is driven by the belief that the authority is blocking opportunity illegitimately. Dismissive defiance is not readily turned around because the authority is seen as unable to redeem itself. People who are dismissively defiant question the power held by authorities. Most notably, they question why an authority should exist. Dismissive defiance is not reduced by displays of procedural justice, as occurs with resistant defiance. Such displays are either likely to be ignored or considered insincere by the dismissively defiant.

Why Do Motivational Postures Matter?

There are three ways in which motivational postures add value to the array of social science concepts on which criminologists draw. First, motivational postures provide a more nuanced appreciation of an authority's influence than singular concepts such as trust or legitimacy or cooperation. Seldom do people have only one posture. Individuals learn to adapt to institutional life through experiences in schools, religious groups, families, work, and leisure. In the process, individuals acquire the full suite of motivational postures.

The multidimensionality of postures is a reminder to authorities to consider which postures they wish to engage as they go about their business of crime prevention and control. What is more, a person's suite of postures will not necessarily be in perfect alignment. Lack of alignment creates cognitive dissonance and room for deliberation. Engaging with the full range of motivational postures provides opportunity for resolution of differences through persuasion and dialogue. Responsive regulation (Ayres and Braithwaite 1992) and restorative justice (Braithwaite 2002) are approaches to crime control that recognize and respond to multiple motivational postures in those being policed or regulated. Through being able to read the posturing of individuals and groups and understand how postures can be shifted, authorities improve the likelihood that they will elicit cooperation and improve their effectiveness (Braithwaite et al. 2007a, b). They may avoid the escalation of unnecessary conflict, for example, through responding constructively to resistance while strengthening bonds forged through commitment and capitulation. Or an authority may increase its legitimacy and compliance with laws through refining its mission and improving its enforcement regime, thereby tackling game playing while building commitment to new shared goals.

Some critics may say that government is neither flexible nor insightful enough to be responsive to motivational posturing in the way described here. In an era of regulatory capitalism, however, agents of government are doing more of the regulating and enforcement and can invest in understanding how to read posturing, build commitment, deal responsively with resistance, and find ways of reining in disengagement and game playing. Restorative justice is an approach that operates on such principles.

Postures provide useful social cues for improving communication between authorities (or their agents) and individuals at the micro level. They are also useful at the macro level. Motivational postures add value through providing insight into how well an authority is engaging with the public. Motivational postures can be measured and aggregated to give an indication of the social distance between an authority and the communities it serves. An analysis of an authority's motivational posturing profile within and between different communities provides insight into how an authority is falling short in its bid to win public confidence. Authorities that have good relations with the communities they serve and are achieving valued social goals should attract high levels of commitment and capitulation, some level of resistance (an appropriate countermeasure to an authority's power in a democracy), and low levels of disengagement and game playing. Problems in engaging with communities can be identified through departures from this profile. High resistance signals that authorities need to listen more and be responsive to concerns. Unusually high levels of disengagement or game playing signal a more fundamental problem in which authorities have become disconnected from the norms and values of a substantial segment of the community. Reversing the situation involves critically assessing the authority's mission, the moral purpose that underpins it, the laws and rules that supposedly reflect purpose, and the authority's capacity to enforce laws and rules in a respectful way.

The third insight provided by motivational posturing theory is recognition of the two types of defiance: resistant defiance which does not challenge the purpose of the authority but expresses grievance over the way in which the authority carries out its duties and dismissive defiance which challenges the existence of the authority and undermines its effectiveness. If authorities are to show emotional intelligence in how they deal with lawbreakers, it is important that they don't misread the defiance they are dealing with and its consequences, particularly mistaking the more benign form of resistant defiance for dismissive defiance. Resistant defiance is argumentative and annoying for authorities, but it can be turned around if the authority is prepared to listen to grievances and respond in a way that makes the system work more fairly and reasonably. Authorities that work at maintaining their public integrity will manage resistance routinely, allowing them to dedicate more resources to the challenges created by disengagement and game playing. It is the dismissive defiance

associated with disengagement and game playing that is most difficult to address constructively because dismissive defiance places lawbreakers psychologically beyond the reach of influence of authority. Dismissive defiance in both taxation and occupational health and safety is associated most strongly with individuals failing to comply with the law (Braithwaite 2009: 270–272, 2011; Braithwaite et al. 2007b).

Background

Motivational postures were discovered in the context of regulation (Braithwaite et al. 1994; Braithwaite 1995, 2003, 2009). Regulation to secure compliance with criminal law (e.g., policing) is a subset of regulation as it has been studied in the motivational posturing context.

As governments introduced laws to improve the standard of care in nursing homes, those working in the nursing home industry were preoccupied with making sense of how they would be affected and treated under the new laws. Change was inevitable, but with it came uncertainty and threat that they would not meet the standards and not receive the approval of the new authority. They adopted positions for dealing with the new authority – motivational postures.

Embracing the new laws involved accommodation to the new standards and processes and in principle commitment to quality care. Becoming reconciled to change and new inspection regimes that hopefully would be benign for those who tried to do the right thing manifested as capitulation. This posture reflected a sense of having no escape from the authority. Feelings of grievance and resentment over enforced change found expression as resistance to authority and the way inspectors used their power. Despair and dismissiveness toward the authority and its goals for better quality of care brought disengagement, most notably among those who felt their business would be untenable in the future. Subsequent research revealed that these postures were present in other regulatory contexts as well. Research in the field of taxation a decade later revealed game playing as a fifth posture at work. It has been replicated in other contexts.

The coherence and regularity of the five postures was identified through factor analyzing the responses of individuals to a motivational posturing questionnaire. These clusters of beliefs, attitudes, preferences, interests, and feelings have been assessed through some 30 statements rated on a five-point strongly disagree-strongly agree Likert scale. Postures have been studied in this way across a range of contexts including taxation (Braithwaite 2003, 2009; Hartner et al. 2008), environmental preservation (Bartel and Barclay 2011), occupational health and safety (Braithwaite 2011), policing (Murphy and Cherney 2012), and child protection (Ivec et al. 2011) as well as nursing home regulation (Braithwaite et al. 1994; Braithwaite et al. 2007a). Observational and qualitative analyses of posturing have been undertaken in research on small business taxation compliance (Harris and McCrae 2005), agricultural reform (Cartwright 2011), war making and peace building in Indonesia (Braithwaite et al. 2010), child protection (Harris 2012), and the resettlement of South Sudanese refugees (Losoncz 2011).

Research has shown that motivational postures are in part shaped by a person's values, norms, and expectations and in part by the actions taken by authorities (Braithwaite 2009). The term "postures" was chosen to capture the socially shared and acceptable nature of these signals – they were not deep dark secrets. The term "motivational" captures their purpose of protecting the individual from the potential threat that authority poses to his or her freedom. Postures are part of the psychological armory that individuals put on to allay any fears they might have in their next encounter with authority.

Applications of Motivational Posturing Theory

A study by Bartel and Barclay (2011) used motivational postures to demonstrate how interventions with farmers who were not complying with environmental laws could be designed in a more considered, targeted, respectful, and responsive way.

Bartel and Barclay identified qualitative differences in the responses of Australian farmers to more stringent regulation and criminalization in environmental protection. One cluster of farmers aligned themselves with authority through postures of commitment and capitulation. They not only complied with the basic legal requirements but also went "beyond compliance" in furthering the objectives of the regulations and the authority. They felt positively toward government intervention, were involved in newer industries, and better positioned economically to cope with change in the management of their smaller properties.

Other farmers clustered into the defiant groups of resistance, disengagement, and game playing. Consistent with motivational posturing theory, the game playing cluster of farmers was the least likely to preserve land for conservation purposes, opposed change, opposed government, and the law. They were older, well-established farmers of their district, and less educated. The resistant farmers and disengaged farmers were more ambivalent in how they approached regulation. Bartel and Barclay (2011) observed defiance as being linked to jurisdiction and industry. The resistant and disengaged farmers tended to be crop growers, dealing with drought and pests and a threatened livelihood. For Bartel and Barclay, motivational postures provided a fine-grained analysis of types of opposition to environmental laws and reasons for that opposition, as well as spotting where entrenched forms of defiance may lie.

In the context of policing, the motivational posture of disengagement has been used to explain the limits of procedural justice in dealings with ethnic minorities (Murphy and Cherney 2012). Murphy and Cherney found that when ethnic minorities believed that laws were not legitimate, procedural justice was counterproductive in eliciting cooperation. Disengagement from authority was the factor that explained why procedural justice was proving counterproductive as a means of increasing cooperation. Disengagement by ethnic minorities from the police meant that police were unable to find a foothold to start building a cooperative relationship through procedural justice.

Tyler and colleagues have shown convincingly that procedural justice can build legitimacy of laws and elicit cooperation with authorities (Tyler 1997). However, when no relationship is in place, as reflected in a posture of disengagement, procedural justice provides an opportunity for game playing with authorities rather than cooperating.

Harris (2012) used motivational posturing theory to better understand the responses of parents to their first visit from a child protection officer. Child protection authorities placed importance on their officers using an assessment framework which was expected to deliver consistency in decision making and "court readiness" should there be need to remove the child from the family. Harris observed parents expressing greater defiance over assessment procedures than over the visit from child protection authorities. Harris' thesis was that where assessment was experienced as being intrusive, good will on the part of parents to cooperate with the investigation was lost, and this was evident through motivational posturing.

Within Harris's (2012) sample of parents receiving their first visit, some were predominantly positive about the assessment process. These parents were more likely to see benefits in the intervention by child protection authorities (commitment posture). Others were less comfortable with the assessment measures but accepted that the investigation would occur and were resigned to putting up with it (capitulation posture). A third group took offense at the assessment process and responded with criticism and anger over the investigation (resistance posture). A fourth group was equally critical, but instead of fighting, withdrew from the authority, expressing no hope that the investigation or anything flowing from it could help them or their child in any way (disengagement posture). Whether parents dealt with what they saw as unreasonable intrusiveness through actively resisting assessment requests or feigning cooperation, they successfully increased the workload of overly stretched child protection officers and undermined effectiveness in protecting children.

Bartel and Barclay (2011) and Harris (2012) identified groups who were experiencing threat from authority at one point in time. Efforts to capture the dynamic possibilities of motivational posturing have been made by Robinson and

McNeill (2008). They propose a model to examine formal, substantive, and long-term compliance with community penalties. Robinson and McNeill argue that when one form of compliance is privileged over others (e.g., formal compliance such as attending scheduled appointments), those serving community sentences may be more likely to engage in a form of posturing (e.g., capitulation to the system of surveillance) that does little to engender commitment to abiding by conditions of community sentencing (substantive compliance) or being law abiding in the longer term.

Robinson and McNeill (2008) propose that offenders with community sentences move between these different levels of compliance. Changes in levels are hypothesized as reflections of their interaction with various actors in the correctional system; different postures come to the fore in response to different experiences and treatment.

Some evidence in support of Robinson and McNeill (2008) proposition comes from qualitative research on how Australian tobacco farmers responded to government closure of their industry in a small rural community. The government turned its back on the tobacco growing industry as it struggled to compete internationally and as it fell into disrepute with the rise of the anti-smoking lobby. Government removed protective tariffs and imposed heavy taxes on the tobacco growers, seriously threatening their livelihood. Recognizing discontent in the region, buyers operating an illicit "chop-chop" market saw an opportunity to move into the community, offering to buy tobacco for a very attractive price while circumventing excise tax.

As government authorities cracked down on the chop-chop market, farmers were incensed that government could be so merciless in its treatment of them. Defiant postures emerged in response to the intrusive and tactical maneuvers of the authorities. The result was that growers distanced themselves from government officials and closed ranks in silence over what they knew about the chop-chop market. The chop-chop industry flourished. Cartwright (2011) documented the postures of resistance and disengagement of the community in her interviews and heard stories of game playing by growers who had thrown their lot in with the chop-chop buyers. She also observed widespread tempering of defiance with capitulation. Growers wanted to stay on the right side of the law and feared the coercive measures taken by chop-chop buyers who demanded a steady supply of tobacco from their suppliers.

In spite of police, customs, and tax officer presence in the community, authorities were unable to gain support. Their strategy was narrowly focused on threatening and catching growers and buyers of chop-chop. There was less evidence of authorities acknowledging the law-abiding postures of growers and helping ensure that those who were trying to stay within the law could keep their farms viable in the future. Interestingly, Cartwright (2011) could find no evidence of commitment to government authority or policy among any of her participants.

The dynamic of how postures can change over time as a result of how authorities are seen to be acting also has been illustrated in the context of tax compliance (Braithwaite 2009, Chapter 8). Perceptions of deterrence and procedural justice (integrity) have been shown to affect defiant posturing. The trajectory of change for resistant defiance is different from dismissive defiance.

An analysis of how resistant defiance gathered momentum over time identified two pathways that competed for dominance over a three-year period. One was a consistent barrier to defiance, a pathway of moral obligation that was theorized as representing a "moral" self – a self that was law abiding and that had nothing to fear from authority. Competing with the moral self was a pathway hypothesized as representing the democratic collective self. The democratic collective self saw unfairness in the system and expressed grievances, including disillusionment with the democracy and dissatisfaction over the tax paid given the goods and services that government provided. The democratic collective self wanted improvements to the system and was prepared to protest in spite of the restraints urged by the moral self. This was how resistant defiance evolved over a three-year period.

Within these dynamics, deterrence played a dual role. It fuelled grievance in the early period, thereby boosting postures of resistance; but over time deterrence strengthened the value of being

law abiding, boosting commitment, and capitulation. Procedural justice similarly was shown to have a mixed fate. Those who had distanced themselves from government and felt oppressed by the system were unlikely to acknowledge procedural justice (integrity) in the system. But once acknowledged, postures changed in a more cooperative direction. Over time, resistant defiance was lowered by perceptions that the authority honored its commitment to procedural justice.

A similar but simpler story emerged for dismissive defiance. A weak pathway represented the moral self that upheld ideals of doing the right thing and meeting tax-paying obligations. The stronger pathway fuelling dismissive defiance represented a desire to achieve, particularly aspirations for social standing or wealth. This is referred to as the status-seeking self. At first, the status-seeking self was high on grievance, but within a short time span, this turned into interest to find ways of avoiding tax without breaking the law. Aggressive advisers became the ideal "alternative authority" for the status-seeking self. Only two avenues emerged for keeping dismissiveness in check. First was the weakened moral self. The second was deterrence. Again, deterrence initially fuelled grievance, and only in the longer term did it reduce dismissive defiance and bolster the moral self.

For tax authorities, this study emphasized the importance of system-wide coverage of procedural justice because it is a way of ensuring responsiveness to the posture of resistance and to resistant defiance. Just as important, however, is critical scrutiny of the law, enforcement strategies, and penalties that are necessary for dealing with the less common, but more substantial problems that arise when postures of disengagement and game playing assume dominance and manifest as dismissive defiance. Authorities need to be watchful of the postures developing in the community and adjust their enforcement strategies to preserve a "firm but fair" regime.

Key Issues and Controversies

Motivational posturing is purposely an amalgam of more basic psychological concepts such as attitudes, beliefs, norms, expectations, and needs. The advantage of an amalgam of more basic concepts is that it is more accessible to practitioners. Motivational postures provide a scientifically sound yet practically useful tool for criminologists and others involved in taking communities with them in developing or implementing policy. Yet there remain questions around theorizing change and potential controversy about the normative aspects of posturing.

Theorizing Change

Research has shown how postures can be shaped by individual characteristics such as values, worldviews, and circumstances (Bartel and Barclay 2011; Braithwaite 2009). They are also responsive to the actions of authorities – the rules they enforce and how they enforce them (Braithwaite 2009; Braithwaite et al. 1994). The actions of authorities are shaped, in turn, by jurisdictional laws, administrative instruments, and enforcement cultures (Bartel and Barclay 2011). All three – individual characteristics, actions of authorities, and the regulatory system – vary across domains, for example, policing, peacekeeping, nursing homes, taxation, occupational health and safety, and the environment. While postures have been shown to operate across all these domains, they do so differently. Understanding changes in how postures operate across domains, systems, and cultures requires further theoretical development.

Normative Value of Postures

If postures are responsive to characteristics of the person, the authority, and the jurisdiction, what then constitutes a desirable posture? From the perspective of an authority, commitment and capitulation are the signals authorities want to receive, while resistance, disengagement, and game playing are less welcome. An argument has already been made for why authorities should embrace resistance as essential feedback on performance. But the important normative questions to address are the following: Are postures of disengagement and game playing necessarily undesirable? Are there not occasions when dismissive defiance is necessary to overcome oppression and bad government?

Why Do Authorities Fear Postures That Promote Dismissive Defiance?

In the broader context of establishing law and order from local neighborhood policing through international peacekeeping, it is important to recognize that authorities create defiance through their very existence. Defiance is not just something that offenders experience and display because they are overly emotional or lack self-control. When authorities make their presence felt in the lives of individuals, individuals must manage the experience of intrusion on their freedom – they must manage the fact that they are expected to obey the authority even if they don't want to or believe what is being asked of them is morally wrong.

Low trust in governments, rising individualism, plural societies, and/or cultural heterogeneity (LaFree 1998) mean that it is easy to deal with the threat posed by authorities through demonizing them as not worthy of respect and rationalizing defiance as the only way forward. Whether the problem is child abuse, domestic violence, gang violence, street protests, tax evasion, or financial crimes on Wall Street, defiance, once widespread, undermines the morale of law enforcers and eventually the authority's public legitimacy. The global financial crisis revealed how the financial crimes of the powerful can be engineered with such defiance that neither government nor their regulatory authorities could or would effectively challenge practices and avert disaster. Authorities have reason to fear defiance. And there is good reason for thinking that dismissive defiance may not serve the public well.

The Other Side of the Argument: Dismissive Defiance Brings Change

But defiance, even dismissive defiance, in itself is not necessarily undesirable. Healthy democracies value and embrace defiance among their citizenry: Failure to do so undermines democracy (Durkheim 1961). Defiance may lead to injustice being challenged through formal processes including the courts. Or it may lead to a challenge on the streets, involving lawbreaking and criminal arrests (Lovell 2009). Suffragettes were arrested as they battled for equal rights for women. Rosa Parkes spearheaded the US Civil Rights Movement when she was arrested for refusing to obey a bus driver who directed her to give up her seat for a white passenger. Nelson Mandela and the anti-apartheid activists too were punished harshly by authorities for their defiance in fighting for freedom and justice. The Arab Spring has revealed the courage and defiance of ordinary people fighting for democracy. Lovell (2009) has provided a sympathetic analysis of "seasoned activists who are willing to transgress the law in the pursuit of social justice" (xi) and who carve out their niche at the poorly researched intersection of politics and criminology.

We can only conclude that the normative value of defiance depends on its manifestation and its purpose and the change that it brings to society. As Arendt (2000) points out, we may hope that change will be for the better, but it is not something of which any of us can ever be sure.

Future Directions

An important way of extending motivational posturing theory is to consider how these ideas operate at a group (community, national, or corporate) level as opposed to individual level and how posturing is used by authorities to control, rightly or wrongly, the communities that come under their influence. Such developments show promise in relation to three ongoing strands of work. Planned social change on a large scale requires the marshaling of collective hope (Braithwaite 2004). Marshaling the hopes of a few (the likes of Nelson Mandela or Osama Bin Laden) into collective hope involves building support around shared goals, engendering collective confidence that these shared goals can be achieved, and finding the pathways to progress the agenda such that individual efforts are coordinated to produce an outcome that is more than the sum of its parts (Braithwaite 2004). Such processes unfurl amidst numerous setbacks and hiccups, and arguably most fail. Motivational postures provide a framework for tracking the journey of collective hope in terms of its ascendancy, threats to ascendancy, or demise.

A second strand of work involves regulatory ritualism and understanding how organizations and nation-states can replace ritualism with more productive and authentic action (Braithwaite et al. 2007a). According to Merton (1968), ritualism means acceptance of institutionalized means for securing social goals while losing all focus on achieving the goals or outcomes themselves. In nursing homes, for example, new policies and in-service training programs may be introduced to respond to problems of noncompliance, but the new policy may never be implemented and the in-service training program may not address inspector's concerns. Regulation has spawned many rituals of comfort but not of compliance for good outcomes (Braithwaite et al. 2007a). A regulatory regime that settles for postures of capitulation to the neglect of postures of commitment will be at risk of regulatory ritualism. When postures of capitulation dominate at the expense of postures of resistance, the likelihood of self-learning and self-initiated change becomes impossible (Braithwaite et al. 2007a). The dialogue and conversations that can unsettle conventional practice and move people out of their comfort zone do not take place, and ritualism prevails.

In a similar way, regulatory ritualism threatens the effectiveness of many international "social justice" initiatives particularly in developing countries. Such countries are urged to become signatories to international agreements for human rights or equity in education or health care, sometimes with rewards from donors once agreements are signed. Capitulation without commitment means that meaningful change is unlikely, replaced by token gestures and empty rituals (Charlesworth 2011).

A third strand of work focuses on the contribution of motivational posturing to the goal of crime prevention through strengthening communities' efficacy and resilience. Within this agenda, sparks of individual defiance need to be harnessed and redirected toward goals that benefit communities. In such contexts, the posture of commitment within communities comes to the fore. Commitment gives rules and laws a sense of meaning and purpose beyond the situation people find themselves in. Commitment to ideas can drive out anomie. Anomie often prevails in communities that have declared "war" on authorities. Kolodziej (2011) has shown empirically that in Poland where tax authorities and taxes are held in low regard, commitment to taxation is related to being knowledgeable about the way the economy works. This understanding of the bigger picture enables individuals to adopt a more positive attitude to having a tax system and move beyond the complaints associated with a system that is not currently meeting public expectations.

Theoretical developments around hope, ritualism, and resilience require a synthesis of micro and macro social processes, the individual psychology of cooperative engagement and the mobilization of meaning and purpose on a larger scale. It remains to be seen how well motivational posturing theory can contribute to this synthesis. It also remains to be seen whether low crime societies manifest a politics of hope and commitment. Do they eschew the cynicism of ritualism, the lure of opportunity (Shover 2007) and gaming of law? That is a much bigger research agenda, one that we have yet to grasp.

Related Entries

- ▶ Procedural Justice, Legitimacy, and Policing
- ▶ Psychology of Procedural Justice and Cooperation
- ▶ Shame in Criminological Theory

Recommended Reading and References

Arendt H (2000) "What remains? The language remains": a conversation with Günter Gaus. In: Baehr P (ed) The portable Hannah Arendt. Penguin, New York

Ayres I, Braithwaite J (1992) Responsive regulation: transcending the deregulation debate. Oxford University Press, New York

Bartel R, Barclay E (2011) Motivational postures and compliance with environmental law in Australian agriculture. J Rural Stud XXX:1–18

Bogardus ES (1928) Immigration and race attitudes. D. C. Heath, Boston

Braithwaite V (1995) Games of engagement: postures within the regulatory community. Law Policy 17(3): 225–255

Braithwaite J (2002) Restorative justice and responsive regulation. Oxford University Press, New York

Braithwaite V (2003) Dancing with tax authorities: motivational postures and non-compliant actions. In: Braithwaite V (ed) Taxing democracy: understanding tax avoidance and evasion. Ashgate, Aldershot, pp 15–40
Braithwaite V (2004) The hope process and social inclusion. Ann Am Acad Pol Soc Sci 592:128–151
Braithwaite V (2009) Defiance in taxation and governance: resisting and dismissing authority in a democracy. Edward Elgar, Cheltenham/Northampton
Braithwaite V (2011) Motivations, attitudes, perceptions and skills: pathways to safe work. Safe Work Australia, Australian Government, Canberra
Braithwaite V, Braithwaite J, Gibson D, Makkai T (1994) Regulatory styles, motivational postures and nursing home compliance. Law Policy 16(4):363–394
Braithwaite J, Makkai T, Braithwaite V (2007a) Regulating aged care: ritualism and the new pyramid. Edward Elgar, Cheltenham/Northampton
Braithwaite V, Murphy K, Reinhart M (2007b) Taxation threat, motivational postures, and responsive regulation. Law Policy 29(1):137–158
Braithwaite J, Braithwaite V, Cookson M, Dunn L (2010) Anomie and violence: non-truth and reconciliation in Indonesian peacebuilding. ANU E-Press, Canberra
Cartwright S (2011) Policy, obligation and justice in Australia's tobacco growing industry. PhD dissertation, Australian National University, Canberra
Charlesworth H (2011) Swimming to Cambodia: justice and ritual in human rights after conflict. ALTA law research series (ALRS), 20 Sept. http://www.austlii.edu.au/au/journals/ALRS/2011/4.html
Durkheim E (1897[1952]) Suicide: a study in sociology. Routledge & Kegan Paul, London
Durkheim E (1961) Moral education. Free Press, New York
Harris N (2011) Shame, ethical identity and justice interventions: lessons from research on the psychology of influence. In: Karstedt S, Loader I, Strang H (eds) Emotions, crime and justice. Hart, Oxford
Harris N (2012) Assessment: when does it help and when does it hinder? Parent's experiences of the assessment process. Child Fam Soc Work 17(2):180–191
Harris N, McCrae J (2005) Perceptions of tax and participation in the cash economy: examining the role of motivational postures in small businesses. Australian National University, Canberra. Regulatory Institutions Network Working Paper 80
Hartner M, Rechberger S, Kirchler E, Schabmann A (2008) Procedural fairness and tax compliance. Econ Anal Policy 38(1):137–152
Ivec M, Braithwaite V, Reinhart M (2011) A national survey on perceptions of how child protection authorities work 2010: the perspective of third parties – preliminary findings. Australian National University, Canberra. Regulatory Institutions Network Occasional Paper 16
Kolodziej S (2011) The role of education in forming voluntary tax compliance. Gen Prof Educ 1:22–25
LaFree G (1998) Losing legitimacy: street crime and the decline of social institutions in America. Westview, Boulder
Losoncz I (2011) Blocked opportunity and threatened identity: understanding experiences of disrespect in South Sudanese Australians. Aust Rev Afr Stud 32:121–145
Lovell JS (2009) Crimes of dissent: civil disobedience, criminal justice, and the politics of conscience. New York University Press, New York
McBarnet D (2003) When compliance is not the solution but the problem: from changes in the law to changes in attitude. In: Braithwaite V (ed) Taxing democracy: understanding tax avoidance and evasion. Ashgate, Aldershot, pp 229–244
Merton RK (1968) Social theory and social structure. Free Press, New York
Murphy K, Cherney A (2012) Understanding cooperation with police in a diverse society. Br J Criminol 52:181–201
Robinson G, McNeill F (2008) Exploring the dynamics of compliance with community penalties. Theor Criminol 12(4):431–449
Shover N (2007) Generative worlds of white-collar crime. In: Pontell HN, Geis G (eds) International handbook of white-collar and corporate crime. Springer, New York, pp 81–97
Tyler TR (1997) The psychology of legitimacy: a relational perspective on voluntary deference to authorities. Person Soc Psychol Rev 1(4):323–345

Defiance Theory

Leana A. Bouffard[1] and
Lawrence W. Sherman[2,3]
[1]College of Criminal Justice, Sam Houston State University, Huntsville, TX, USA
[2]Institute of Criminology, Cambridge University, Cambridge, UK
[3]University of Maryland, Baltimore, MD, USA

Overview

In 1993, Lawrence W. Sherman proposed defiance theory, which laid out the factors leading to defiant or deterrent reactions to punishment. His original theory may be interpreted as a counter to deterrence/retribution approaches to achieving compliance with the law in that it proposed that while deterrent effects may result from sanctions,

a combination of factors may also produce defiant effects for that same sanction. Criminological theory and criminal justice policy have long focused on the relationship between sanctions and criminal behavior. Deterrence and labeling, in particular, are two major theoretical traditions that emphasize sanctions as a key explanatory factor, providing contradictory predictions about the impact of those sanctions on behavior. Deterrence theorists predict that sanctions, especially those which are swift, certain, and proportionally severe, will deter or reduce further criminal behavior. Additionally, criminal justice policy is often predicated on the assumption that sanctions deter offenders. Labeling theory, on the other hand, predicts that sanctions will stigmatize the offender, producing increased offending (i.e., secondary deviance) in the future.

Defiance Theory (1993)

A large body of research examining the deterrence and labeling theoretical perspectives has produced inconclusive results. Research suggests that, in varying instances, sanctions may either deter or increase future offending. Recognizing this diversity in the effect of sanctions, Lawrence W. Sherman (1993) argued that the pattern of sanction effects observed in existing research exhibits two themes. First, the impact of sanctions appears to depend on perceptions of fairness, in that sanctions viewed as unfair are more likely to increase offending. Second, sanctions appear to increase crime among out-groups while deterring crime among in-groups. Suggesting that existing theory is incapable of accounting for these patterns, Sherman proposed *defiance theory* to explain the conditions under which sanctions will increase criminal activity versus deterring offending.

The starting point in Sherman's original defiance theory is the concept of defiance, which he defined as "the net increase in the prevalence, incidence, or seriousness of future offending against a sanctioning community caused by a proud, shameless reaction to the administration of a criminal sanction" (p.459). Defiance may take several forms. Similar to deterrence, defiance may be either specific (i.e., the reaction of an individual to his or her own punishment) or general (i.e., the reaction of a group to the punishment of a group member). Additionally, individuals may exhibit either direct defiance, reacting against the sanctioning agent, or indirect defiance, reacting against another individual who vicariously represents the sanctioning agent. Sherman provides examples of the different types of defiance. For example, an individual who assaults a police officer during an arrest is exhibiting specific, direct defiance. An individual who assaults his or her spouse following a domestic violence arrest is exhibiting specific, indirect defiance. General, direct defiance is illustrated by the South African ambush killings of police officers, who were viewed as the tools of oppression during the apartheid era, a perception that has lingered. The 1992 Los Angeles riots following the acquittal of the four police officers who beat Rodney King during a traffic stop provide an example of general, indirect defiance.

Conditions for Defiance

Sherman argues that defiance is likely to occur when the sanctioned individual views his or her punishment as unfair or illegitimate, is poorly bonded, and denies the shame of the punishment. In contrast, sanctions are expected to produce deterrence when the sanctioned individual is well bonded, views the sanction as legitimate, and accepts the shame he or she feels. In this case, the individual remains proud of his or her connection to the community, recognizes the harm that his or her actions have caused, and attempts to repair that bond. Sanctions are irrelevant to future offending when these factors are evenly balanced. For example, when a well-bonded offender denies the shame associated with the unfair sanction, the expected outcome will likely be irrelevance, not deterrence. In this instance, the perceived unfairness of the sanction and the failure to accept the shame that accompanies the sanction will nullify any deterrent effect produced by the strong social bond. With this discussion, Sherman is able to theoretically account for the mixed effects of sanctions apparent in the research literature.

Defiance theory focuses on explaining the defiant reaction to a sanction. Specifically, there are four necessary conditions for defiance to occur: (1) the sanction must be defined by the offender as unfair, (2) the offender must be poorly bonded to society, (3) the sanction must be viewed by the offender as stigmatizing, and (4) the offender must refuse to acknowledge the shame produced by the sanction. In proposing his theory and identifying these four conditions, Sherman has borrowed from John Braithwaite's theory of reintegrative shaming, Tom Tyler's concept of procedural justice, and Thomas Scheff and Suzanne Retzinger's discussion of the role of shame and rage in destructive conflicts.

Perceptions of Fairness

According to Sherman, one key theme in understanding whether a sanction produces defiance or deterrence is the perceived fairness or legitimacy of the sanction or sanctioning agent. Unfairness may be related to disrespect by the sanctioning agent or a perception that the punishment is arbitrary, discriminatory, or otherwise unjust. Whether the unfairness of a punishment is substantive or perceptual, unfair or unjust sanctions may not have their intended deterrent effect. According to Tyler's procedural justice perspective, sanctions that are perceived as unfair reduce the legitimacy of law enforcement or the criminal justice system, which reduces the likelihood of compliance. If an individual perceives a punishment as unjust, he or she may begin to question the law itself and feel justified in disregarding it. Scheff and Retzinger contend that societal disapproval, when expressed disrespectfully or to a person with weak social bonds, may evoke anger. Reintegrative shaming theory likewise argues that sanctions that stigmatize and label the offender may weaken existing social bonds and produce increased offending. Thus, perceptions of the fairness of a sanction and the experience of being stigmatized by a sanction may interact with an individual's social bonds to produce defiant effects.

Social Bonding

Procedural justice argues that when the legitimacy of formal sanctions breaks down, social sanctions are expected to take their place. Thus, social bonding also plays a large role in defiance theory. In particular, Sherman relies on some elements of reintegrative shaming theory to explain the connection between unfair or stigmatizing sanctions and social bonds. For Braithwaite, individuals who have strong social bonds (i.e., interdependency) may be more likely to experience reintegrative sanctions, which are rejecting of the act but avoid applying a label to the individual. Thus, reintegrative sanctions are likely to be viewed as fair and to produce deterrence. Disintegrative sanctions, however, are rejecting of both the act and actor, stigmatizing the sanctioned individual. These sanctions are more likely to be viewed as unfair and disrespectful. Sherman likewise recognizes the potential criminogenic effect of stigmatizing sanctions, especially among individuals with weak social bonds. He argues that individuals with strong social bonds will not react defiantly to a punishment perceived as unfair so as not to jeopardize those bonds. On the other hand, individuals with weak social bonds are more likely to deny the shame of being sanctioned and respond with indignation and anger. This angry, prideful reaction sets the stage for defiance and increased offending.

Experiencing Shame

An individual's reaction to the shame of a sanction is the final link in the explanation of defiance. Sherman highlights the role of shame, pointing both to reintegrative shaming theory and to Scheff and Retzinger's work on the master emotions of shame and pride. Both Braithwaite and Scheff and Retzinger argue that individual reactions to the shame of a sanction will vary depending on an individual's level of social bonding. Similarly, labeling theory suggests that secondary deviance (i.e., defiant effects for Sherman) occurs as a reaction to the experience of being sanctioned, in that individuals may perceive their punishment as an attack and may act defiantly as a defense to society's disapproval.

Scheff and Retzinger criticize early versions of labeling theory for failing to take emotions,

especially shame, into account. For these authors, societal disapproval is described as a threat to the sanctioned individual's social bonds. If the individual accepts the shame that he or she feels and recognizes the harm he or she has caused, the individual may seek to avoid that behavior in the future (i.e., deterrence). Braithwaite's theory of reintegrative shaming presents a similar argument. On the other hand, if the person refuses to acknowledge or rejects that shame, he or she may respond with self-righteous anger. Scheff and Retzinger describe a shame/rage spiral, in which rage is a protective measure against shame, a way of rejecting the shame, and a defense against a perceived attack. Thus, this shame/rage spiral occurs when a person's bond is threatened, the shame is not acknowledged, and behavior is interpreted as an attack. This produces violence, hatred, and resentment which may lead to defiance. For Sherman, shame, or the refusal to acknowledge shame, is the primary causal mechanism in explaining defiant effects.

Empirical Evidence for Defiance Theory

Sherman concludes his theoretical formulation by noting that "until recently, the science of sanction effects has been short on facts and even shorter on theory. Now, it seems, the available theory has gotten ahead of the facts" (p. 468). Despite the promise of defiance theory in explaining variation in sanction effects, there have been no complete tests of the theory since its development. Most of the evidence that can be marshaled in support of the theory is derived from studies not originally designed to examine its propositions. Some research supports the notion that perceptions of unfairness, either to the law being imposed or to the sanction itself, are likely to lead to more criminal offending (i.e., defiant effects). Sherman highlights research suggesting that previously sanctioned individuals are less likely to be deterred. It may be that, because few people are formally sanctioned for offending, those who do receive a punishment perceive their treatment as comparatively unfair and respond defiantly by engaging in further delinquency.

Additional research examining police-citizen encounters indirectly tests some of the propositions articulated by defiance theory. These studies primarily focus on the offender's (or citizen's) perceptions of fair treatment by police officers in their encounters. Raymond Paternoster, Robert Brame, Ronet Bachman, and Lawrence Sherman (1997) examined the effect of arrest on the likelihood of engaging in subsequent domestic assaults and found that the offender's perceptions of fair treatment by police were important determinants of future offending. Other studies of police-citizen interactions support the premise that individuals who feel that they are unfairly treated by police are more likely to be resistant. In other words, the perceived legitimacy of a police officer's action is an important predictor of citizen compliance or resistance. When the police are perceived to be respectful to citizens, compliance is more likely. Confrontational and physical actions on the part of police, on the other hand, are more likely to produce resistance, possibly because the actions are interpreted as unfair and stigmatizing. This body of research supports Sherman's argument that defiant effects are more likely to occur when sanctions are perceived as unfair. While these results are suggestive, they do not address the key to defiance theory, which is an individual's perception of the sanction.

A more recent study examined the perceptual nature of defiance theory and the impact of those perceptions on future offending more closely and has provided the most complete test of the theory to date. Leana Bouffard and Nicole Leeper Piquero (2010) found that individuals who perceived a sanction as unfair and were poorly bonded had higher rates of future offending. While much existing research demonstrates that perceptions of unfairness and social bonding have a strong connection to offending, the role of shame in producing defiance remained unclear in this study. Other research, however, has supported the role of shame in offending.

In 2006, David Brownfield examined the propositions of defiance theory as they relate to gang membership and found that all elements of defiance theory were linked to gang membership. More specifically, perceptions of legitimacy and pride were

important. Gang members reported being happy about the possibility of being arrested. A study of recidivist drunk driving in Australia likewise found evidence that shame experienced as a result of the sanction may produce a sense of self-righteous anger, which leads individuals to question the legitimacy of the sanctioning agent (Freeman et al. 2006). These studies have also noted the importance of delinquent peers, who might provide justification for defiance and a social reaction that enhances the likelihood of a prideful response. Unfortunately, the existing research generally provides only piecemeal support for defiance theory. Studies specifically designed to link the theory's propositions together are necessary.

Recent Developments in Defiance Theory

While empirical evidence provides some support for the propositions of defiance theory (Sherman 1993), the explanation is limited in its position as an explanation of lawbreaking, especially in response to sanctions. In 2010, Sherman expanded the original conceptualization of defiance to position it as the independent variable in a broad, general theory of criminology that encompasses "lawmaking, law-breaking, and responses to law-breaking" (Sherman 2010, p. 360). Sherman comments that Sutherland's original definition of criminology included all three of these elements. No theory has yet been proposed, however, to unify all three in one causal framework. Thus, this reconceptualization is an ambitious project designed to provide a more comprehensive theoretical perspective that applies to all three components of criminology.

This reconceptualization may be viewed as an elaboration of the 1993 version of defiance theory, broadening and extending the original through the process of consilience, in which predictions from one domain of criminology (i.e., lawbreaking) are applied to other domains (i.e., lawmaking and responses to lawbreaking). As Sherman contends, observations from all three domains point to the same causal force, the sense of moral obligation to resist the status quo, what Sherman refers to as defiance.

The major elaborations of the original theory include broadening the definition of defiance as the moral sense of "obligation or justification to defy the status quo" (p. 364). Sherman recognizes a continuous contest between defiant and deterrent actors that encompasses both lawful and unlawful efforts to create or change law, the use of power by agents of law, and confrontations between lawbreakers and law-enforcers (e.g., police-citizen interactions). Because the general defiance theory incorporates all elements of criminology and reciprocal causation, this is the major restatement that allows the more general theory to cover all domains within criminology.

Sherman's formal restatement proposes that "criminal laws are more likely to be made, broken, obeyed and enforced when people intuitively feel that their actions are morally founded" (p. 366). He defines defiant effects in a similar framework as deterrent effects, in that a general reduction in crime is assumed to be a deterrent effect without evidence that the deterrence process (i.e., the sanction imparts a fear of future sanctions) is operating. Likewise, a defiant effect would be defined as "active resistance to a countervailing force" (p. 369) or resistance to the status quo. These are a broad class of observable behaviors reflecting decisions to make law or policy, break laws, enforce laws, or a choice not to enforce laws. Defiance then is an "individual or collective sense of moral obligation, including indignation and empathy" (p. 369). Individuals selectively perceive harm caused to themselves or others, and those perceptions of harm may produce moral intuitions of indignation, empathy, or both. These facilitate a definition of the status quo as illegitimate and produce an obligation to defy that status quo.

Sherman provides a number of specific predictions or hypotheses based on his general defiance theory that encompass the three domains of criminology:

1. Conduct will be criminalized to the extent that opponents of the conduct promote a feeling of moral obligation to oppose that conduct as illegitimate.
2. Laws will be enforced to the extent that enforcement agencies collectively and individually

accept the legitimacy of the moral obligation to do so.
3. Laws will be broken to the extent that potential offenders feel a moral obligation to defy agents making and enforcing laws as illegitimate.
4. Morally infused campaigns of law-breaking can lead to changes in law, to the extent that the violations promote a feeling of moral obligation to allow them.
5. Law-makers and enforcers will also break laws, or re-make them, to the extent that they feel moral indignation, empathy, or a moral obligation to do so.
6. Changes in moral indignation or empathy can be fostered by events causing an epiphany revealing the harm caused by some conduct, law or punishment.

(Sherman 2010, p. 375)

Types of Defiance

As in the original defiance theory, Sherman (2010) distinguishes a number of potential defiant effects. He retains a sense of specific as opposed to general defiance that refers to the individual or collective response to an encounter with the status quo, either its content or the violation or enforcement of law (see p. 370). In the original version of defiance theory, Sherman defined direct and indirect defiance. Here, he retains the initial meaning but renames these forms as direct and displaced defiance, referring to behavior aimed either at those who have committed a moral wrong (i.e., direct) or at those not directly known to have committed the harm but belonging to and representing the same group (i.e., displaced).

A distinction between concealed and identifiable defiance refers to efforts by the actor to hide their actions and is related to how much they are willing to sacrifice with their defiance. For example, those individuals who flee the scene of the crime to avoid capture are engaging in concealed defiance. Those who openly identify themselves and seek attention for their actions engage in identifiable defiance. For identifiable defiance, Sherman gives the example of suicide bombers who create video- or audiotape recordings prior to committing the act. Defiance may also be categorized as active (i.e., requiring action) or passive (i.e., requiring the withholding of action). In terms of law-breaking or offending, commission of most crimes would be considered active; however, failure to wear seat belts or helmets would be examples of passive defiance. For example, despite helmet laws, some motorcyclists may choose not to wear helmets as a defiant withholding of action related to perceptions that the government has overstepped their authority in regulating behavior. In terms of law-enforcing, Sherman describes the use of police discretion (e.g., under-enforcement or choosing not to arrest in certain situations) as passive defiance. On the other hand, excessive use of force by police may be defined as active defiant law-enforcing. Finally, Sherman distinguishes resistant and dismissive defiance based on perceptions about the moral authority of the wrongdoing agent or agency. Resistant defiance recognizes the moral authority of the agent but advocates for change, while dismissive defiance rejects the agent altogether, disabling or eliminating the authority.

Defiance and Other Theoretical Perspectives

In his 2010 restatement, Sherman retains a focus on the legitimacy of sanctioning agents or other agents of authority. In contrast to the earlier version, which viewed the acknowledgement of shame as the key predictor of defiance, the 2010 version seems to place legitimacy as the main element in understanding defiant effects. Thus, Tyler's work on procedural justice and other research exploring issues of fairness retains a place of importance within the General Defiance Theory. Sherman also positions General Defiance Theory as a complement to subcultural and conflict theories in that it recognizes both competing systems of morality within society and the use of the power and material resources of authorities to shape moral obligation through such forces as moral panics. Additionally, the theory borrows from techniques of neutralization in recognizing the justifications offered for defiance or resistance to the moral force.

Future Directions

Researchers have highlighted the links between Sherman's (1993) original defiance theory and other explanatory mechanisms. For example, within psychology, the personality construct of grandiosity (i.e., exaggerated perceptions of self-worth) may inform the path to defiance, in that grandiose or self-centered individuals may be more likely to reject the sanctioning agent and the shame associated with being sanctioned, resulting in a defiant response. In criminology, research also suggests that individuals with low levels of self-control are more likely to perceive sanctions as unfair and to respond with anger (Piquero et al. 2004). Though not specifically addressed by Sherman or necessarily intended, one advantage of defiance theory is that in linking concepts, like emotion and social bonding, it offers an integrative perspective that accounts for either defiance or deterrence.

From the life-course and criminal career perspective, defiance may also be seen as an explanation of continuity in and desistance from offending. The theory provides an explanation for continuity, which is the defiant response of a poorly bonded offender who defines his or her sanction as unfair and stigmatizing and refuses to acknowledge the shame he or she feels. These individuals may continue or escalate their offending, becoming involved in secondary deviance (i.e., persistence). The original theory can also explain desistance by arguing that if an individual defines a sanction as unfair and stigmatizing but has strong social bonds, that person may accept the shame that he or she feels or be unwilling to jeopardize his or her bonds through a defiant reaction. According to Sherman, these individuals will be deterred from future offending (i.e., they will desist). In the 2010 restatement, Sherman extends this argument by suggesting that desistance may be thought of as an individual defying their own prior behavior. In other words, those who have previously engaged in lawbreaking may eventually recognize the harm that they have caused to themselves and others, producing a moral obligation to resist their own previous behavior (i.e., a turning point in the language of the life-course perspective). Thus, exploring defiance theory from a longitudinal, life-course framework is another promising avenue for research.

James Unnever and Shaun Gabiddon (2011) also highlight the role of defiance in providing an understanding of the overrepresentation of African-Americans, especially men, in the criminal justice system. In their Theory of African-American Offending, these authors highlight perceptions of the legitimacy of criminal justice agencies as a key explanatory factor. Due to a history of racial injustices at the hands of the criminal justice system, African-Americans view law enforcement and other criminal justice actors and agencies as having little legitimacy, perceiving that the law and the criminal justice system are racist, disrespectful, and unfair to African-Americans. Greater levels of experienced and/or vicarious racial injustice interfere with bonding to predominantly white social institutions like school and work. Thus, offending by African-Americans is viewed as a defiant response to racial injustice. Sherman (2010) likewise points to the history of race and law in the United States as an illustration of various aspects of his General Defiance Theory. For example, in terms of lawmaking, Sherman argues that his theory is capable of accounting for both the justification/rise and the demise of slavery within the USA by positioning these decisions within a consideration of the shifting public sentiments from the moral justification of slavery to a moral obligation to oppose the enslavements of human beings. Both of these perspectives suggest the relevance of the concept of defiance and of General Defiance Theory in exploring the relationship between race and crime, criminal law, and the criminal justice system.

Conclusion

It is difficult to truly assess the value of Sherman's original defiance theory with the paucity of studies that directly test its propositions. Rather, the existing research provides suggestive evidence supporting some elements of the theory, particularly perceptions of fairness and social bonding. Additionally, the 2010

reconceptualization into a General Defiance Theory of criminology adds layers of complexity that make testing its propositions more difficult. As a new theory, no researchers have yet attempted to assess the propositions set forth in Sherman's (2010) restatement. What remains is to explicitly design studies that link the elements of defiance together as proposed by the theory and to explore the connections between this and other theories at both the specific level of the original version (i.e., in predicting lawbreaking) and at the more general level of the restatement (i.e., in also predicting lawmaking and law-enforcing). At this point, the theory is relevant to understanding the importance of the interplay between perceptions of fairness and legitimacy, social bonding, and the experience of shame, but future research is necessary to fully explore these relationships and to assess how well the concepts of defiance, moral obligation, etc., extend to all domains within criminology.

Related Entries

- ▶ Compliance and Corporate Crime Control
- ▶ Defiance and Motivational Postures
- ▶ Shame in Criminological Theory

Recommended Reading and References

Bouffard LA, Piquero NL (2010) Defiance theory and life course explanations of persistent offending. Crime Delinq 56:227–252

Braithwaite J (1989) Crime, shame, and reintegration. Cambridge University Press, Cambridge

Brownfield D (2006) A defiance theory of sanctions and gang membership. J Gang Res 13:31–43

Freeman J, Liossis P, David N (2006) Deterrence, defiance and deviance: an investigation into a group of recidivist drunk drivers' self-reported offending behaviours. Aust N Z J Criminol 39:1–19

Paternoster R, Brame R, Bachman R, Sherman LW (1997) Do fair procedures matter? The effects of procedural justice on spouse assault. Law Soc Rev 31:163–204

Piquero AR, Gomez-Smith Z, Langton L (2004) Discerning unfairness where others may not: low self-control and unfair sanction perceptions. Criminology 42:699–734

Scheff TJ, Retzinger SM (1991) Emotions and violence: shame and rage in destructive conflicts. Lexington Books, Lexington

Sherman LW (1993) Defiance, deterrence, and irrelevance: a theory of the criminal sanction. J Res Crime Delinq 30:445–473

Sherman LW (2010) Defiance, compliance, and consilience: a general theory of criminology. In: McLaughlin E, Newburn T (eds) The Sage handbook of criminological theory. Sage, Thousand Oaks.

Tyler TR (1990) Why people obey the law. Yale University Press, New Haven

Unnever JD, Gabbidon SL (2011) A theory of African-American offending: race, racism, and crime. Routledge, New York

Defining Disorder

Bernard E. Harcourt[1] and Danielle Wallace[2]
[1]Law School, University of Chicago, Chicago, IL, USA
[2]School of Criminology and Criminal Justice, Arizona State University, Phoenix, AZ, USA

Synonyms

Incivilities

Overview

There's a lot of order in disorder and a lot of disorder in order. The problem, it seems, is in the definition – in defining disorder, or order, that is. Prominent scholars have remarked that disorder is a "slippery" concept (Skogan 1990), and a few quick observations confirm that. Red-light districts – areas that have traditionally been associated with the notion of disorder – are often more orderly than expected (Skogan 1990, pp. 61–63). Neighborhoods that are governed by organized crime or other tight social networks often turn out to be some of the safer neighborhoods in an urban setting (Suttles 1972). Conventional indices of neighborhood disorder, for instance, people loitering at the corner, often turn out to be forms of social control that ensure safety in the community (Patillo 1998). And perceptions of

disorder, it seems, are often the product of the racial composition of the neighborhood, not the level of danger: individuals often perceive Black and Hispanic neighborhoods as far more disorderly than any objective measurement would establish (Sampson and Raudenbush 2004).

This makes it extremely difficult to define disorder – in fact, it may make it impossible. Disorder, it turns out, is in the eye of the beholder. It is a normative rather than purely descriptive category. It functions, most often, as a statement of preference. A good example is New York City under mayor Rudolph Giuliani and former police commissioner William Bratton: during the first 3 years of broken windows policing, from 1994 to 1996, the rate of robbery victimization fell about 60 %, and many observers claimed victory for order, but during the same period, allegations of police misconduct rose by about 60 % (Harcourt 2001, pp. 167–170). Whether to describe that period as orderly or disorderly tells us more about, say, our preferences and normative values, than it does about the scientific measure of disorder. But perhaps the most telling illustration of the difficulty of defining disorder is the multiplicity of definitions employed by scholars: even though there has been extensive work trying to define both disorder and order (see, e.g., Ross and Mirowsky 1999), as Kurbin (2008) notes "variability in how disorder is understood and conceptualized across studies is the rule rather than the exception" (p. 205). Indeed, researchers generally do not even agree on whether to call the phenomenon "disorder" or "incivilities" (Kubrin 2008). In this entry, we catalogue some of the main issues surrounding the definition of disorder.

Important Controversies

Creating Disorderly Subjects

Generally, disorder is thought of as containing two dimensions: physical and social disorder. Physical disorder pertains to the decaying urban environment – for example, abandoned buildings, rundown sidewalks, decaying storefronts, or broken windows; social disorder embodies any public behavior of individuals that can be seen as threatening (Sampson and Raudenbush 1999; Harcourt and Ludwig 2006). Cues in the neighborhood environment that are viewed as social disorder tend to result in the categorization of individuals as orderly or disorderly. Orderly or law-abiding people are generally understood to be those residents who, when confronted with disorder, retreat into their homes to avoid criminals and victimization. Disorderly people, on the other hand, are those who see disorder as an opportunity to commit crime (see Harcourt 2001 for a discussion on how broken windows theory creates these categories). Unfortunately, the line between the orderly and disorderly is often hard to define, and these categories become especially troubling when we consider that residents may tend to see disorder through the lens of race and ethnicity.

This point is illustrated well through example. In *Off the Books*, Sudhir Venkatesh (2000) shows how residents of a disadvantaged community in Chicago earn an income through illicit activities. Illicit income may come from activities such as drug selling, repairing cars in alleyways, or selling stolen goods from a van near a neighborhood park. Clearly, these activities are illegal and would draw police attention. But do the residents of the neighborhood see it that way? These individuals offer goods; some of those goods can be seen as necessities, like socks or beauty supplies, at prices that enable the residents to buy them where they otherwise could not. Are the residents seeing disorder? Most likely not. Are the police seeing disorder? Most definitely, it is these kinds of activities that broken windows policing explicitly targets, with the expectation that taking disorderly people off the streets will prevent serious crime from happening (Wilson and Kelling 1982). But in this case, the individuals selling stolen goods, while clearly engaged in illegal activities, are merely making money; additionally, they are providing low-cost goods to residents. Thus, rather than being disorderly, they are viewed as a positive economic force in the neighborhood. Again, the line between disorderly and orderly blurs.

One of the unique aspects of the broken windows theory is its temporal structure – disorder breeds serious crime. Apply this to individuals (i.e., social disorder), and the theory would

suggest that disorderly individuals will either themselves will commit serious crime or attract other criminals. Policing disorderly individuals, in theory, effectively removes serious criminals from the street. In the process, disorder becomes a degree of crime: breaking a window, littering, and jumping a turnstile become grades along a spectrum that leads to homicide. And since disorder theory and its implementation defines who is orderly and who is disorderly, this, combined with its temporal aspects, has a deleterious, long-standing stigmatizing effect on individuals (Harcourt 2001, p. 150).

This is especially problematic in communities were disorder does not represent "criminal activity" the way the police or nonresidents would think of it, specifically in communities where illegal activities (i.e., drug selling, prostitution, or panhandling) are a means to an end for daily subsistence (St. Jean 2007; Venkatesh 2006) or when neighborhood social networks are interwoven with individuals who are criminally active (Patillo 1998; Pattilo-McCoy 1999). The policing of disorder removes breadwinners and the protective agents of the community – furthering the dire economic conditions in the community and making the illegal economic activities that spurn policing more steadfast in the neighborhood (Rose and Clear 1998). Here, policing disorder and not acknowledging what disorder means to community residents redefine residents in the eyes of law enforcement and even mainstream society. Residents in these communities, therefore, are looked upon as hard-core criminals in the making, in need of policing and punishment. Regardless of individual variation in perceptions, policing disorder redefines individuals in disorder communities as disorderly individuals, rather than residents. Creating disorderly subjects, coupled with variation in individuals' perceptions of disorder, becomes even more nefarious because policing disorder in these communities does not reduce disorder; rather, it reinforces the need for policing.

Seeing Disorder

One of the primary reasons why theories of disorder have gained so much traction is their simplicity; generally, the presence of disorder in neighborhood signals criminal opportunities and a lack of social control (Wilson and Kelling 1982). The common sense appeal of disorder theories is evident. We have all been in "bad" neighborhoods, looked at our surroundings, and felt certain of the potential for crime and scared for our safety. It is this simple association between the presence of disorder and subsequent knowledge of neighborhood conditions in the minds of individuals that keeps disorder theory at the forefront of criminology and urban sociology. As a result, disorder theories are able to use "disorder" as a proxy for informal social control, neighborhood withdrawal, or other neighborhood characteristics related to crime or disorganization.

Yet, an important ingredient in the association between disorder, neighborhood outcomes, and crime is often assumed: residents must perceive, interpret, and act on disorder similarly in order to affect the neighborhood-level outcomes disorder is associated with. This assumption is made not by individuals but by academics. Disorder is at the core of neighborhood social dynamics because it represents deviance from prescribed norms about neighborhood like. Burisk and Grasmick (1993) suggest that all that is needed for neighborhood social control is the shared notion of a life that is free of personal harm. Disorder, it is assumed via its definition, is the visual reminder within a neighborhood that safety is not present. But does disorder uniformly represent this to individuals?

The broken windows theory will be used as an example of this, though it is certainly not the only disorder theory that assumes this process. The broken windows theory suggests that disorder leads to crime in a neighborhood because of resident withdrawal (Wilson and Kelling 1982). That withdrawal is based in residents perceiving their community as unsafe or that no one cares about what happens in the neighborhood and, perhaps most importantly, no one cares about what happens to *them* in the neighborhood. Because of what disorder signals, residents are afraid of potential victimization and left uncertain of help in a crisis; therefore, they restrict their activities in a neighborhood (Wilson and Kelling 1982).

Once withdrawal happens, informal social control in a neighborhood is diminished; it is this process that the broken windows theory suggests lead to crime.

Rested in the above process are three components related to individuals, though which are rarely discussed and often assumed: (1) perceptions of disorder cues, (2) how individuals interpret disorder cues, and (3) how people act on their interpretation of disorder cues. Each of these steps is crucial to the link between disorder and crime, though, to date, they have been left relatively unaccounted for by disorder theory. Additionally, each of these steps builds upon each other; therefore, we begin with perceptions.

For residents to ultimately restrict their involvement within a neighborhood, they first need to "see" the disorder around them. In much disorder research, there is an assumption that people see only the disorder that is present; more specifically, people only perceive the objective levels of disorder in their environment (Ross and Mirowsky 1999). Most researchers report that there is little within-neighborhood variation of disorder perceptions; in other words, people in the same neighborhood see disorder similarly (Skogan 1990). Yet, given the results of some recent studies, this is clearly not the case (Hipp 2010; Sampson and Raudenbush 1999, 2004; Wallace, in press). Sampson and Raudenbush (1999) report a 0.6 (approximately) correlation between subjective and objective measures of disorder; in another study, their results show within-neighborhood variation in disorder perceptions (Sampson and Raudenbush 2004). Even when constraining the ecological unit to about the size of a city block, effectively controlling for exposure to difference sections of a neighborhood, Hipp (2010) also found variations in disorder perceptions. In essence, research has clearly demonstrated that disorder perceptions vary. Researchers, with few exceptions (Taylor 2001), have yet to fully theorize why variation in disorder perceptions is so problematic for disorder theory: when people see disorder differently, it suggests that disorder is interpreted differently.

Just as with perceptions, for disorder to negatively affect residents' behavior within their neighborhood, to be in accordance with the broken windows theory, they first must see disorder and then interpret it as threatening. A great deal of research suggests that the presence of disorder in a neighborhood makes residents feel uneasy and at risk for victimization (Kelling and Coles 1996; Skogan 1990; Wilson and Kelling 1982). Hunter (1978) suggests that individuals interpret disorder to mean that both citizen groups and public agencies within and outside of the neighborhood cannot maintain satisfactory conditions within the neighborhood and subsequently residents feel vulnerable. When disorder is employed in sociopsychological studies, most studies suggest that individuals interpret disorder as the potential for crime and victimization, which generates fear and stress thereby causing negative personal outcomes, such as reduced health and well-being or increased depression and mistrust (Ross 1993; Ross and Jang 2000; Ross and Mirowsky 2001). While the link between disorder and fear of crime is established, it is unclear if individuals' interpretations is what generates fear, especially given that fact that individuals often are unable to distinguish disorder and crime (Gau and Pratt 2008) and researchers often conflate the two in disorder measurements (Harcourt 2001). As research shows that perceptions of disorder are not universal, it is unreasonable to assume that *interpretations* of disorder are universal as well.

Finally, we reach the third step in the process where perceptions and interpretations of disorder combine to impact individuals' behavior. If individuals perceive disorder and interpret it in a threatening way, the broken windows theory, as well as other theories, suggest they begin to withdraw from neighborhood life and restrict their territoriality (Skogan 1990; Wilson and Kelling 1982). Of course, this is predicated on an interpretation of disorder that will elicit such behavior. Should disorder be perceived and interpreted by individuals in any other light, their assumed behavioral reaction does not emerge. Indeed, there are several urban ethnographies that show that people do not necessarily withdraw from disorder. For example, in *Black Picket Fences*, Mary Patillo shows that neighborhood residents interact and work with local gang members selling

drugs to arrange places and times in the neighborhood were drugs can be sold that do not effect neighborhood life, such as basketball games or use of a playground. In *Off the Books*, residents often use services that in other neighborhoods would be considered disorderly, such as fixing cars in alleys or panhandling, because the services are economical and a means to help other residents get by (Venkatesh 2006). In sum, once variation in perceptions or interpretations of disorder does not cause individuals to withdraw from their neighborhood or be fearful, the link between disorder and neighborhood outcomes, such as crime or informal social control, is broken.

From the delineation of the above steps, it is easy to see how *any* kink in the three steps is a serious threat to validity for the various theories of disorder. There have been calls for research to address the above points (Taylor 2001), but only recently has scholarship begun to study them. To date, current research has established that there are variations in disorder perceptions, though has had little success in explaining why they occur. At best, we know that variations in perceptions of disorder are not primarily a result of neighborhood exposure or use (i.e., one's routine activities) but perhaps partially due to how individuals employ racial stereotypes that are attached to place (Sampson and Raudenbush 2004). Unfortunately, this explanation does not cover all the variation we see; for example, women, those highly educated, and married perceive more disorder, while older adults and non-Whites perceive less (Hipp 2010; Sampson and Raudenbush 2004; Wallace, in press). What we are left with is variation in perceptions without explanations. In essence, until we know how people see disorder, we are left uncertain of the effects disorder has on individuals and neighborhoods.

Race and Disorder

As noted in the previous section, research has shown that disorder perceptions are imbued with race (Sampson and Raudenbush 2004). The stereotype of Black Americans as violent and criminal is both well established and common (Eberhardt et al. 2004), and it should be no surprise that this stereotype has bled into the perception of place (Wacquant 1993) given segregation and slavery (Loury 2002). As a result, individuals' perceptions of disorder are filtered through "stigmatized groups and disreputable areas" (Sampson and Raudenbush 2004).

There is both direct and indirect evidence of this. First, when testing whether neighborhood racial characteristics influence disorder perceptions, Sampson and Raudenbush (2004) show that even when taking objective levels of neighborhood disorder into account, there is a positive association between disorder perceptions and percent of Black in a neighborhood. In essence, the more Black faces people see in their community, the more problematic they perceive disorder. As a result, they suggest that "residents supplement their knowledge (of disorder) with prior beliefs informed by the racial stigmatization of modern urban ghettos" (p. 336). As a consequence, Sampson and Raudenbush (2004) express concern about using the broken windows theory to inform disorder reduction strategies: if disorder perceptions and fear are racially motivated, reducing disorder will not reduce fear given that fear is being produced by another visual cue – race.

We also see evidence that race is involved in disorder perceptions via policing literature. The broken windows theory has motivated a type of policing strategy called order-maintenance policing that targets disorder cues, particularly social disorder cues, as a way to reduce serious crime in the neighborhood. The reasoning will not be repeated here as it is part of early sections, but there is an inherent problem with policing disorder, namely, the act of subject creation. When policing disorder, we run the risk of moving away from thinking that the act of loitering, for example, is a crime and toward thinking that the loiter is criminal. Harcourt (2001) expands on this as such: "Rather than judging the act of loitering, we may attribute to the person who is loitering certain propensities – certain tastes, attitudes, and values" (p. 163). Couple the issue of subject creation with the evidence of the racialization of disorder perceptions and making policing disorder in a race neutral way becomes difficult, if not impossible. The broken windows policing creates racial animus: "The prominence of race in the decision to stop citizens may not rise to the threshold of racial profiling,

but it does seem to create a racial classification of 'suspicion'" (Fagan and Davies 2001). Bringing about order to a "disorderly" neighborhood becomes impossible when the neighborhood has a minority racial composition; in effect, removing disorder will not remove the association of disorder with race. So long as black and brown faces remain, the neighborhood will be seen as disorderly regardless of the amount of disorder present.

The Paradox of Disorder

One aspect of disorder that has been continuously demonstrated by urban ethnographers and systematically ignored by disorder theorists is when residents see disorder as a positive force, or rather, as *order*. As discussed above, in disadvantaged neighborhoods, disorder has become an adaptation while still remaining disorder in the eyes of authority. Making order from disorder is a process aligned with the durability of social isolation and segregation. Generations of residents of disadvantaged neighborhoods are exposed to deleterious neighborhood conditions (Wilson 1987, 1996). It is this exposure that enables disorder to be viewed in a different light – rather than being a neighborhood problem, it is a form of daily survival (Anderson 1999) or a means of making money (St. Jean 2007; Venkatesh 2006).

A ripe example of this process is how some components of disorder were normalized in the Robert Taylor Homes housing project in Chicago (Venkatesh 2000). In *American Project*, Sudhir Venkatesh details life in a housing project in Chicago. Venkatesh shows how the identity of the housing project shifted over the course of 40 years. At the beginning, residents of housing projects saw their living environment as a move away from the decay of the surrounding areas. Where they once were sharing kitchens, now, residents had their own, self-contained apartments. Residents worked toward betterment of the community and the project by forming formal and informal groups that dealt with issues of safety, quality of life, and amenities in the projects. Yet, as time passed, the city and the Chicago Housing Authority (CHA) slowed the resources available to the projects. The efficacy residents once had in dealing with institutional forces faded – any solution to their housing problems had to come from their hands. With less social power, the solutions residents offered would not pull them out of poverty or their environment but only offer to make that environment safer or more negotiable. For instance, a group of powerful women in the Robert Taylor Homes created a law enforcement body of men who were staying with other residents and were not listed on the lease. (As a condition of their lease, residents were not allowed to have any individuals stay with them that were not listed on the lease.) In order not to be reported to the CHA, these women would often call upon the men to deal with problems, like domestic violence or snitching to the police (Venkatesh 2000). Soon, even this efficacy faded. Hustling – or making money through illicit means – became prominent and soon dominated public space. And as soon as hustling gained a foothold, residents stopped seeing disorder, like loitering or panhandling, and started to see hustling. This process was a direct result of opportunities and resources being pulled from the housing projects, leaving residents to cope with their situation in any way they can.

In disadvantaged communities, the economic situation so depressed and isolated that illicit economies provide one of the few reliable sources of income (Anderson 1999; Venkatesh 2000, 2006; Wilson 1996). Often conducted in public space, illicit economies often take the outward shape of social disorder. When occurring in public, illicit economies are not simply a panhandler or a drug dealer; they are richer and include more than the simple notions of what the classic definition of disorder leads us to imagine. Venkatesh (2006) paints an in-depth picture of an illicit economy working in a neighborhood park on the south side of Chicago:

> Pimps brought their sex workers to an abandoned building near the park. Carliss, a car mechanic, moved his outdoor 'Oil and Tire Change' operation to the alley next to the park's basketball court. Two gun brokers come to a nearby abandoned building once a week to see handguns and pistols. A few men sold stolen car stereos, guns, and other electronic equipment from the back of two beat-up beige vans that were always stationed at the park entrance. Mo-Town, the local hot dog vendor, and Charlie, who sold stolen cigarettes and beauty products, set up their respective carts at the edge

> of the park. And now the drug sales, as Big Cat had promised, round the clock. All of this was secured by placement of Big Cat's rank and file around the area: all were armed, they physically searched and harassed passersby, and they drank and smoked marijuana until the early morning hours with loud music blaring from their stereos. They also charged a fee to each entrepreneur based in and around the park. (Venkatesh 2006, p. 71)

Here we see economic opportunities come in many forms, forms often associated with disorder. In disadvantaged neighborhoods, the illicit economy is pervasive and often condoned by residents, regardless of whether they are in or out of the licit economy (Venkatesh 2006). Condoning does not mean acceptance of the illicit economy for many residents, though they understand the necessity of it. When working on public and accessible public space in the neighborhood, residents understand attacking the illicit economy means undoing someone's income: "residents may weigh delinquent activity that has an economic dimension differently than, say, crimes of passion like domestic violence and assault. This does not mean that all underground activity is tolerated. But of the activity generates income, any ethical dilemmas it creates must also be judged in terms of how the activity supports a household and ever the wider community" (Venkatesh 2006, p. 74). People turn a blind eye to the drug trade in inner-city communities: "people residing in the drug-infested, depressed inner-city community may understand the economic need for the drug trade" (Anderson 1999, p. 132). People only become concerned with the drug trade when it results in violence – then their intolerance comes to the surface. Now, illicit economic endeavors that are, to an outsider, disorderly are not seen in that light for neighborhood residents. Even when these activities are not welcomed, they are seen more as a means to an end than as disorder. The outward publicness of illicit economy shifts the disorderly aspects of its workings to a more complex problem of economic survival of individuals and communities. Hence, disorder is not linked to fear of crime, neighborhood decay, or crime; rather, it is rather the opposite: livelihood.

The handful of concerns about defining disorder that we have been able to address in this limited space – and there are many more – should make scholars wary of using the concept of disorder in its current form. Disorder is long overdue for reconceptualization, and careful scholarly work should aim to do just that.

Recommended Reading and References

Anderson E (1999) Code of the street: decency, violence, and the moral life of the inner city. W. W. Norton, New York

Eberhardt JL, Goff PA, Purdie VJ, Davies PG (2004) Seeing black: race, crime, and visual processing. J Personal Soc Psychol 87(6):876–893

Fagan J, Davies G (2001) Street stops and broken windows: terry, race, and disorder in New York City. Fordham Urban Law J 28:457–504

Gau JM, Pratt TC (2008) Broken windows or window dressing citizens' (in) ability to tell the difference between disorder and crime. Criminol Public Policy 7:163–194

Harcourt BE (2001) Illusion of order: the false promise of broken windows policing. Harvard University Press, Cambridge, MA

Harcourt BE, Ludwig J (2006) Broken windows: new evidence from New York City and a five-city social experiment. Univ Chic Law Rev 73(1):271–320

Hipp JR (2010) Resident perceptions of crime and disorder: how much is "bias", and how much is social environment differences? Criminology 48:475–508

St. Jean PKB (2007) Pockets of crime: broken windows, collective efficacy, and the criminal point of view. University of Chicago Press, Chicago

Kelling GL, Coles CM (1996) Fixing broken windows: restoring order and reducing crime in our communities. Simon and Schuster, New York

Kubrin CE (2008) Making order of disorder: a call for conceptual clarity. Criminol Public Policy 7(2):203–214

Loury GC (2002) The anatomy of racial inequality. Harvard University Press, Cambridge, MA

Patillo ME (1998) Sweet mothers and gangbangers: managing crime in a black middle-class neighborhood. Soc Forces 76(3):747–774

Pattilo-McCoy M (1999) Black picket fences: privilege and peril among the black middle class. University of Chicago Press, Chicago

Ross CE (1993) Fear of victimization and health. J Quant Criminol 9(2):159–175

Ross CE, Jang SJ (2000) Neighborhood disorder, fear and mistrust: the buffering role of social ties with neighbors. Am J Commun Psychol 28(4):401–420

Ross CE, Mirowsky J (1999) Disorder and decay: the concept and measurement of perceived neighborhood disorder. Urban Aff Rev 34(3):412–432

Ross CE, Mirowsky J (2001) Neighborhood disadvantage, disorder, and health. J Heal Soc Behav 42(3):258–276

Rose DR, Clear TR (1998) Incarceration, social capital, and crime: implications for social disorganization theory. Criminology 36:441–480

Sampson RJ, Raudenbush SW (1999) Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. Am J Sociol 105(3):603–651
Sampson RJ, Raudenbush SW (2004) Seeing disorder: neighborhood stigma and the social construction of "broken windows". Soc Psychol Q 67(4):319–342
Skogan WG (1990) Disorder and decline: crime and the spiral of decay in American neighborhoods. University of California Press, Berkeley
Suttles GD (1972) The social construction of communities. University of Chicago Press, Chicago
Taylor RB (2001) Breaking away from broken windows: Baltimore neighborhoods and the nationwide fight against crime, grime, fear and decline. Westview Press, Boulder
Venkatesh S (2000) American project: the rise and fall of a modern ghetto. Harvard University Press, Cambridge
Venkatesh S (2006) Off the books: the underground economy of the urban poor. Harvard University Press, Boston
Wacquant LJD (1993) Urban outcasts: stigma and division in the black American ghetto and the French urban periphery. Int J Urban Reg Res 17:366–383
Wallace D (in press) A test of routine activities and neighborhood attachment explanations for bias in disorder perceptions. Crime Delinq
Wilson WJ (1987) The truly disadvantaged: the Inner City, the underclass, and public policy. The University of Chicago Press, Chicago
Wilson WJ (1996) When work disappears. Knopf, New York
Wilson JQ, Kelling GL (1982) Broken windows: the police and neighborhood safety. Atl Mon 127:29–38

Delinquency

► Network Analysis in Criminology

Democratic Policing

Peter K. Manning, V. H. Elmer and Eileen M. Brooks
College of Criminal Justice, Northeastern University, Boston, MA, USA

Synonyms

Cause of police legitimacy; Concept of the police; Comparative police systems; Organizational change and police legitimacy; Origins of modern policing; Policing in developing democracies; Role and function of the police

Overview

The study of democratic policing might well be seen as the answer to the question "What are the police good for?" That is, what good are they expected to do in a democratic state? It surely cannot be dismissed by arguing that this required good is solely or even primarily that it must produce "crime control." Police are a fundamental force in a democratic society in which in the course of their practice they are obligated, as the Patten Report states, to "perceive their job as the protection of human rights." Furthermore, their obligation is collective and shared and cannot be judged in that sense by ad hoc criteria that reflect narrow interests and political positions. They reflect, at best, what is valued, or regulate social obligations we have to one another. The role of the police in democracy is variable over time and culture, not a constant. Police may enhance democracy by sustaining employment, security, and democratic procedures such as voting and demonstrations; they may decrease its viability by corrupt practices and collusion to destroy democratic competition; and, ideally, they may be models of decorum, propriety, and restraint. As a conservative force in general, of course, it is most likely that their role at best may be neutrality on behalf of the state. Why and how they contribute to these necessary functions has not been systematically addressed. Because such questions have not been raised in regard to stable democratic states, and the role of the police in nurturing democracy has rarely been explored, much preliminary ground work is needed. The primary impediment to serious scholarship is the addiction to the false idea that above all else police must "control" crime and be assessed by that toxic criterion. To discuss democratic policing, it is necessary to trace briefly the development of the field of police studies and its failures to address what is democratic about policing in a democracy. This will permit a definition of police, the identification of the features of Anglo-American policing, the application of criteria for democratic policing derived from the work of John Rawls, and the discussion of some issues requiring future research.

The Development of the Field of Police Studies

The systematic study of policing in North America began in the early part of the twentieth century as a result of the pioneering work of Raymond Fosdick, Bruce Smith, and later, Augustus Vollmer. They approached the topic as reformers, as students of law and public administration, and were pragmatic men of action. Later, O.W. Wilson, who approached the study of policing as a student of public administration, served as Commissioner of Police in Chicago. He later wrote the first book on police administration, thus inventing the term demarcating the field. It is generally accepted that the sociological study of policing was initiated by the fieldwork-based study of policing in Gary, Indiana, in the '50s by a University of Chicago graduate student, William Westley. Westley later published this study as *Violence and the Police: A Sociological Study of Law, Custom and Morality* (1971). Some 20 years after Westley' s initial field work, the results of which were known as a result of a series of published brilliant papers, books written by social scientists Michael Banton, James Q. Wilson, Jerome Skolnick, and Arthur Niederhoffer set out the questions still shaping the field. They were in many ways looking at policing from the bottom, from the perspective of the patrol officer or constable. A book of lasting importance, the *Democratic Policeman* by George Berkeley (1969), was overlooked perhaps as a result of the growing concern in the late '60s with riots, disorder, and crime. This book, based on comparative analysis of policing in several European countries, was the first to address the principles of democracy that should ideally be reflected in police administration.

At this time, systematic social science addressed at least some of the lasting concerns of political philosophers from the time of Plato: What are the requirements for a fair, just, even-handed police? Who watches the watchers? From the Greeks onward, this question was couched implicitly at least in the context of democracy as a compelling system of governance. The question raised by Plato and Aristotle was again vibrant and commanding. Here, we encounter an oddity: as lasting as this concern has been, no tight definition of "democracy," "police," or "democratic policing" has been accepted. Political scientists, economists, and conservative apologists bypassed this thorny issue, and moved directly to impassioned advocacy of the idea. Here, policy and advocacy collide with empirical analysis and open-minded exploration of data. It has been known, for example, that some democratic policing practices can be found in nondemocratic societies, and that nondemocratic practices can be found in democracies, especially in times defined as crises. The assumption has been that police are neutral on behalf of the state, committed only to "law enforcement," and loyal. In a stable, economically sound society, the following questions might be considered: Loyalty to what principles? Action on behalf of any government? Commitment to any law, regardless of its impact? Available in any crisis, including a revolution? Preserve and protect what? Whose rights, property, and vested interests? In many respects, policing grew within large European cities as a kind of order-maintaining force, a source of steady employment, and an unpretentious and often violent governmental resource.

In part, this blindness to the broader parameters of policing is due ironically to the success of the political efforts of Vollmer and his students between the two great world wars. They sought diligently to encourage and persuade the public to accept the police as a nascent profession based on science (social, biological, and chemical), carried out by well-educated officers and experts, and focused diligently and relentlessly on chasing and jailing criminals. "Crime" rose to the surface as the "problem" for which they were designed. Ironically, given his later infamy, J. Edgar Hoover was among the brilliant innovators who shaped most "professional policing" by lobbying for new federal laws, opposing corruption, developing a statistical capacity within the FBI, and tightening the administrative controls over agents. The standardized appearance, conservative suit, white shirt, dark tie, black shoes, and hat were no small part of his ingenious impression management. This leadership at the federal level was important as a symbolic brand and image for policing in general. As such, the organization argued successfully that they were deserving of respect, honor, and decent

wages consistent with their skills and duties. This position was augmented some 30 years later by the US government's creation of the Law Enforcement Assistance Administration (LEAA), later to become the National Institute of Justice within the Justice Department, a source of research funding, technical information support to police departments, and expert knowledge. With the publication of the report of the President's Crime Commission (PCC), several ideas emerged that were to have considerable influence and staying power over the coming years. The commission argued that the police, courts, prisons and jails, and probation and parole were not disparate competing units with diverse interests, training, aims, and tactics, but rather were part of an integrated, coherent, meaningful, and orderly *criminal justice system*. The assumption they made that it was just and produced justice was unchallenged. The commission viewed this "system" as flawed, but nevertheless it was created and implemented by just and honest people; it only required a bit of reform. It took a legal realist position that the system produced legal outcomes on a case by case basis. Since legal processes were the source of justice, the system produced justice of a sort. This system, they argued, had failed for lack of resources and needed an infusion of money, material, especially scientifically based technology, and personnel. The reformers hoped that this system could be reshaped by the application of science and technology, and they placed special emphasis on the role of technology in fighting crime. An entire volume of the final report was devoted to the imagined role of science and technology. This was social engineering applied to the problem of crime defined by officially sanctioned, gathered, and presented data. If reformed, they felt, this system could manage and control crime. Again, "crime" became the featured player in the drama. The report argued for enormous investment in human capital and technologies, and the government of the day was prepared to invest. Finally, the Commission viewed education and training as needed to provide "criminal justice professionals." As a result of the report of the commission, legislators created federally funded colleges within several large state Universities that granted degrees in a new field called "criminal justice." In due course, Ph.D. programs were begun, and the field of criminal justice, highlighted by the field of police studies, emerged.

Broadening the Scope of the Field

The subsequent years saw a rise in officially recorded crime (ORC), protest, disorder, and rioting that in part were caused and exacerbated by violent police interventions in traffic, domestic disorder, and demonstrations. It was widely thought that policing was too soft, and the recommendations of the PCC and reforms of the Johnson and Nixon administration were failures. When later in the 1980s the concept of community policing arose as a panacea, its aims were local, parochial, and vague. They at best were tactical and adopted in spite of no evidence of their efficacy. It was not until the early 1990s that policing again emerged as a powerful force believed to serve, protect, and reduce the specter of crime. As a result of the claims of Commissioner Bratton of the New York City Police and a book written by his advisor and colleague George Kelling, a new view of policing emerged. The media validated the claim that "smart management" focused on arrests connected to "disorder" rationalized the process of crime control and resulted in reduced crime. Policing rode the headlines as a positive, rational scientific organization devoted to the general good. There was no question of its democratic qualities, procedures, or negative effects. This was to come later. In this media frenzy, the claims of the broken windows perspective and crime control were wedded and dominated research for almost 10 years. By the time that it was well established that the dip in crime could not be separated from other factors, including manipulating the standards for reporting, recording, and investigating crime, and the research revealed the empty claims of the broken windows perspective, the field had moved on to other concerns, other issues, and other sources of funding. There is no question that heavy concentration of

officers in targeted areas can reduce crime in the short term of 6 months to a year, but this comes at costs in regard to staggering misdemeanor arrests, stops and jailing of minority youths, and questions of fairness and race profiling.

Concern with policing in the context of democracy and development began to escalate in the 1990s with mini-wars in the Balkans, the Middle East, and Somalia to which armed United Nations Forces responded. As David Bayley (2005, Bayley and Perito 2010) has argued, such low-intensity conflicts required policing of a new sort – policing that may serve as a model for other new nations; policing that is peace-keeping, somewhere between crime-based prevention and low-intensity conflicts; policing as an export commodity to be provided variously by cooperating nation states, private security companies such as Halliburton, Blackwater, and others; policing as a kind of nation-building instrumentality; policing as a basis for civic governance. Some claimed this was a moment of exporting a commodity, and policing was merely the next version of Coca-Cola and Mickey Mouse (Brogden and Nijhar 2005). These police-like operations were carried out experimentally in the sense that each new venture was undertaken by different nation states as suppliers of force; each economy and culture encountered was in uneven state of development; and there were no models for such operations. The closest analogy was not "peace-keeping," but suppression of rebellions against colonial powers in Malaya, Palestine, Kenya, and earlier in Ireland. Furthermore, there was a sense in which the policing mission as it unfolded was open-ended. New innovations had to be developed to cope with unexpected contingencies. In addition, sociopolitical changes captured the imagination of scholars and refocused interest in policing and new forms of policing. These included the growth of the European Union and its cooperative policing arrangements, emergent police force; globalization of crime, especially drug-related crime, and regional policing responses to this in Latin America and the Caribbean; growing police cooperation around globally significant meetings and occasions such as international economic summits, Olympics, World Cup soccer matches, and large-scale rioting; and not of least importance, inexpensive networked communications; travel and the growth of global transnational corporations. The insular and parochial view of the United States, of the United Nations, international law, and international cooperation did not reduce the movement of "police peace-keeping": the claim was that such activities were bringing democracy to the world via some new and malleable forms of policing.

Democratic Policing as a Problem Emerges

At about the same time scholars, including notably Professors Jerome Skolnick and David Bayley who were employed by the State Department, began to reflect, to consider, and reconsider the nature of "democratic policing." It was a new enterprise insofar as the actions of policing preceded the concern for rationalizing and defining the concept. Since the early '70s, however, scholars such as David Bayley, David Sklansky, Tim Newburn and Trevor Jones, Clifford Shearing, Hsi Liang, and George Berkeley were characterizing democratic policing. They were focused on programmatic and prescriptive features, rather than empirical features that could be assessed. There was a new and an abiding concern with justice and democracy in the context of policing. There was also a growing awareness that past scholarship on policing had been restricted to anthropological, sociological, or historical studies of policing, often crime or order based, that had *assumed* the democratic nature of policing. The field of what was to become police studies was an odd amalgam of Anglo-American ideas, somehow connected to the idea that "policing" was essentially an invention of Sir Robert Peel.

This idea ignored several powerful ideas of great importance. Peel also invented the most powerful, vicious, and violent gendarmerie in Anglo-American society, the Peace Presentation Force, an Irish Constabulary that morphed into the RIC and RUC. This innovation gave rise to the varieties of policing, including the Garda, various state

policing systems in Australia as well as the Australian Federal Police, and the Royal Canadian Mounted Police (RCMP). As Sinclair (2006) has cogently argued, this was the model for all colonial policing throughout the British Empire. These forces had little concern with "prevention," and they wrought destructive effects of colonial, post-colonial societies, and emerging nations. Finally, these gendarmerie-like forces had enormous and insidious effects on the differential means of police accountability that arose in the Anglo-American world. These were ignored in conventional treatments of "policing." In summary, the assumption that "democratic policing" is unequivocally the case in North America obscured the varieties of policing that developed from the Peel model, the negative and powerful effects of this form of policing, and characterized nominally preventive policing mistakenly as the only consequence of his innovations. The textbook version of modern Anglo-American policing as a direct consequence of Peel's ideas is misleading and partial.

Also unexamined, except in crude asides like the claim that it is obvious that the primary role of the police is to combat crime, is the matter of studying policing's achievements beyond their struggles to manage crime. Criminologists have accepted the claims of the police and their professionalism and have focused largely on their crime management efforts. What else do they or can they do that might sustain or improve the quality of life in democratic societies? The prior question remains: why study police in a democratic context, or as a democratic institution? The police are one organization in a network of institutions that shape social life, such as religion, family, economy, education, and are but one mode of formal social control. They do not stand alone, nor can they control any aspect of social life unilaterally. They deal with failures; they are failure-processors who cope with the vagaries of other forms of social control. However, it is clear that they perform a variety of functions of significance. They increase the life chances and possibilities of the advantaged classes, the middle class and above, by overlooking, reducing, or cooperating to minimize their crimes; they increase the negatives that are lodged against the minorities, the poor, the disadvantaged of all sorts. This is the case, not because police are "racists," elitists, unsympathetic, or incompetent; it is rather as Bittner (1970) says that they are designated to cope with and manage society's difficulties, or its "dirty work" as Everett Hughes (1958) writes. Furthermore, police capacity to control crime is highly limited and reductions and rises are revealed in reports showing brief dips in official statistics. These are partial, misleading, and necessary. The mandate of the police in a democratic society cannot be reduced to crime control, although it remains central; their obligations extend well beyond this, and their role cannot be reduced to visible patrol functions.

To address the mandate, it is necessary to review events of the last 20 years. They suggest the following:

- Crime in North America is declining, and there is no consensus on the explanation for this.
- Policing is rising in status, in part because of the drop in crime and in part because its top command are articulate and well-educated. Their claims to professionalism have been validated.
- Terrorism, globalization, and international police cooperation are increasing awareness of the isolation of American policing from international trends of cooperation, exchange of information and personnel, and increased and more systematic training.
- High levels of immigration and migration within the United States make clear it is no longer an isolated island in the world.
- There are concentrated levels of violence within ghettoized sections of large American cities, and inequality and poverty are rising to the highest levels in 50 years.
- The drop in crime combined with increased inequality is inconsistent with conventional criminological theories.

These issues have brought to mind the need to reflect on the mandate of the police in the United States. The fads – community policing, broken windows policing, and policing as an export commodity – are exhausted, and the sponsored research field is quite dazed.

Police Defined and Policing as an Organization Outlined

The previous discussion of the putative functions of democratic policing assumes the need for a fairly precise definition. Unfortunately, most definitions of policing in Anglo-American societies are loose or misleading, as has been argued, insofar as they focus on crime control to the exclusion of other functions. Furthermore, this begs the questions of the negative consequences of reduced crime and/or the failure to carry out other functions. What indeed are the negative consequences of policing as it is represented currently? It is important that a definition contains their violence potential, the organized character of their practices, their local character, and their focus on order and ordering. Both public and private police exist and operate under the canopy of the law. Finally, police require compliance, within their organization and from citizens, and in a democracy this is essentially limited by the notion that civil rights are universal and procedural guarantees present. Ironically, as Egon Bittner has argued, police in Anglo-American societies are given an enormous range of freedom to deal with the unexpected matters arising and others matters they judge may escalate and cause even greater damage to the social fabric. Manning (2010: 68) has proposed this definition of policing:

> Police organizations in Anglo-American societies, constituted of many diverse agencies, are authoritatively coordinated and legitimate. They stand ready to apply force up to and including fatal force in politically defined territories. They seek to sustain politically defined order and ordering via tracking, surveillance, and arrest. As such, they require compliance to command from lower personnel and citizens, and the ability to proceed by exception.

From this general definition, it is necessary to confine subsequent discussions to the public police. Private police are anomalous insofar as they are not paid to be fair. It is not a small matter that public police are permitted to proceed by exception because their mandate is elastic, negotiated, and local in large part. The mandate is the source of legitimacy and it is a kind of negotiated contract between the several publics they serve, not a reified single "public," and their own definition of the nature of the work and its obligations. It therefore expands and contracts over time and changes in response to public demands and legal constraints. The mandate in a democracy is fraught with contradictions because of the various expectations of police – service, crime control, order management, control of traffic, and demonstrations – and the shifting scope of their practices. That is, the police in a democracy are subject to changes in public opinion, law, political decisions, and media, and their elasticity is essential. In part because of this the democratic police have been most successful when they have convinced the middle class "respectable public" that their primary role is to control or at least manage officially recorded crime. The crimes of concern are of course what might be called "decent nineteenth century street crime": homicide, burglary, robbery, and assault. And in the late twentieth century, this includes policing "drugs." In many respects, this is code word for the patrolling the life styles of the poor in large cities. In many respects, this is an impossible mandate because of the limits of official modes of crime control. In fact, public trust in the police and maintenance of the public trust are probably the most significant "product" of democratic police. They assess trustworthiness in others; they monitor sources of distrust, both in groups and individuals; they are a repository of trust by citizens in their symbolic expressive role. It is for this reason that corruption, crime, and self-serving actions of police are judged more harshly than the actions of other citizens. Finally, policing in Anglo-American societies is less violent and more restrained in spite of periodic eruptions and killings. This suggests at least a rethinking of the Bittnerian definition of policing (Brodeur 2010).

Eleven Features of Anglo-American Policing

Given this definition and outline of the essential features of democratic policing, there still remains what might be called the constraining

functions of democratic policing in any society. By stating the problem in this way, the scope of this discussion can be widened to include problems associated with the transformation of police systems to democratic modes of policing. Here is a brief list of proposed features (Manning 2010: 65–66) that can be used to judge police practices:

1. Police are a public entity and are accountable directly to the people.
2. Individual citizens, their demands, or requests cannot alone guide the organization because its obligations are collective and general, not individual. Callers are not "customers" or clients because they cannot choose to refuse the "service" or opt out of paying for it. There is no free market in police services.
3. Police cannot eschew their symbolic role as representatives of governance and governing. This is central to the belief in fair practice within a democracy.
4. Police violence is generally cautious and limited. Media amplification serves to confuse viewers and produce periodic crises in large cities.
5. Political territory of policing is problematic and cannot be restricted in advance to the limits of a legally defined territory. This ambiguity has increased as a result of international police cooperation; concerns about terrorism and its global dimensions and nonnational origins; policing under the guise of foreign aid and rapid inexpensive travel.
6. Police manage order: they do not produce it and may destroy it.
7. Police are not a neutral political force; they act on behalf of the state in crises.
8. Compliance with command and loyalty in the face of danger are essential to consistent policing, but they are always problematic.
9. Police are highly respected for carrying out dangerous work and given wide tolerance for their actions.
10. Police tracking and surveillance of citizens are increasing and largely unknown and unmonitored. This is concentrated in lower class areas and on lower class groups, but not restricted to them.

In summary of these considerations surrounding a definition of democratic policing, one can conclude the following. The police organization as presently constituted in Anglo-American societies has evolved historically and these patterns are reflected in the structure, function, and image of the police. The development of the role on the ground is a function of the emerging concern for managing risky situations that transcend in meaning the current on-view encounters and entail some form of risk management (Ericson and Haggerty 1997). In the police world, this means gathering "intelligence," information gathered prior to an event, mining data bases, compiling records from other organizations such as insurance companies and schools, and developing even more forms of information gathering and analysis. This means that the organization has evolved a heavily loaded observational role for the patrol officer, granted a wide range of choice for the officer in the event, and has gathered the general support of tradition and courts for trust in the observational and intuitive skills of such officers. Variations in the pattern of sanctioning are a function of the sanctions available as well as the targets seen as sources of uncertainty. These are the shifting targets of police when they act proactively. As nation states develop, sources of conflicts move from local to national and transnational, and these require a kind of reasoned neutrality in practice that may violate local norms and expectations for conflict resolution. The police in a democracy are double coded as both the source of violence and a protection against it. This is complicated by transcendental rules and norms propagated by the nation state. Police are at the same time no longer obligated to kinsmen and their local norms and practices and obligated to other orders – feudal loyalty, state, or legal loyalty. If we think of policing as encoding uncertainty, that is, responding to what it implies for order, then the job is to render manageable uncertainty in terms of continuous procedures of some kind. This, in turn, engenders trust in the police. In addition, as historical studies suggest, continuity in the job as a full-time paid and responsible agent, associated with continuity of procedures, also

implies continuity in the office or role and its functions. These functions are in modern times associated strongly with the Weber (1947) idea of a rational bureaucracy, even though functionally police are punishment centered, not rational-legal in their operation. The connection of public police with some form of accountability and direction in the interests of the "people" seems a tacit expectation of democratic police, but not of policing in general, nor policing in fact. This question of accountability raises another one – how do the police sustain their role as a neutral arbitrator of conflicts? In other ways, the police are transducers or means of converting one sort of fact into another; they process facts into information: facts placed within a context. In some sense, they must display authority whilst being rooted in the everyday lives of citizens. The police mediate and manage conflicts in terms inconsistent with the ways in which they are conceived by citizens.

These ideas are putative and descriptive; they do not say what is actually done, why, or for what reason. Let us now consider democratic policing more specifically.

Policing in Accord with the Difference Principle

It is necessary to step back from the historical development of police, the faddish nature of police studies, and the drive for police professionalism to ask what analytic principles might guide judgment of police practices. It is police practices, not public attitudes, "transparency," or "accountability" that should command attention. Monitoring and legal constraints do not alone change police practices. The most important statement of political philosophy in this century is the work of John Rawls as found in his *Theory of Justice* (1970 TJ) and a condensed version of the argument entitled *Justice as Fairness* (2001 JF). These works presented two fundamental principles of justice. The first claims that all positions in general, and in this case the police, should be open to all on the basis of competition and the second claims that whatever passes for policy should be based on the difference principle. The difference principle can be summarized as stating that any implicit or explicit policy affecting extant inequalities should be to the benefit of the least-advantaged (JF: 42–43). Since it is unlikely that the police can actually reduce inequality, and in fact have no obligation to do so, a political philosophy with policing in mind and based on the justice as fairness idea requires some modification. One might argue that the justice as fairness principle in regard to policing should be refocused in a manner consistent with the Hippocratic Oath: the police should strive to minimize harm. The working version of this abstraction in regard to policing means that any action, planned stated, or enacted, should not increase inequalities. How can this grand working principle be grasped as a set of objectives or guide lines? Expectations of policing are questions of function: if the below principles or rules of thumb are observed, one might expect of (domestic) democratic policing that it function:

- Constrained in dealing with citizens and fair in procedure. These dealings should entail a degree of civility in interactions and in police practices. This excludes under virtually all conditions, torture, mass detentions, "round ups" based on political beliefs, not behaviors, and lengthy suspensions of habeas corpus for citizens.
- Largely reactive to citizens' complaints – reticent rather than sporadic – and not given to frequent secret proactive interventions, crackdowns, sweeps, and militaristic "operations."
- Equal in its application of coercion to populations defined spatially and temporally. The level of coercion is based on minimalistic criteria, much as counter insurgency tactics, rather than a mechanistic "use of force continuum."
- Fair in hiring, internal evaluation, promotion and demotion, transfers, and disciplinary treatment of employees, officers, and civilians.
- Competitive in an environment which includes private police, vigilante groups, posses, ad hoc policing under the guise of "self-help" and revenge. It may include the National Guard

and the armed services (army, navy, coast guard, and the air force). This implies formal and informal modes of cooperation rather than unified and unrelenting actions.
- Accountable and responsible for their actions individually and organizationally.

Finally, such broadly based ideas are inconsistent with the echoes of free market ideology that the police should be efficient, and effective. If efficient means the best usage of resources within the organization, it is clear that police use modest budgeting tools, are locked into invariant strategies (answering 911, random patrol, and investigating crime) and tactics (ways of delivering such services), and in fact have an open-ended remit with regard to overtime pay and resources whenever a crisis occurs. Add to this the facts that sick leave, disability, and days off are contractual, budgets are determined by nonmarket-based criteria, and not in the hands of police supervisory ranks. It is virtually impossible to fire a police officer short of criminal conviction. They cannot be efficient in the nature of the mandate. Effectiveness assumes a criterion against which functions and costs can be judged. If this is defined, for example, as the percentage of known crime for which police make arrests or charges, and rewards were made for more arrests, the consequences would be unacceptable in a democracy as being unfair in regard to risks to the less-advantaged classes given current laws and regulations.

Conclusion

Democratic policing is a concept in need of specification. Police studies throughout the twentieth century remained focused on American or at least Anglo-American police organizations. The onset of peace-keeping in low-intensity conflicts created some reflection on the nature of policing and its mandate and raised questions about the conventional concern for crime management as a paramount function. In this entry, police organizations are defined and their features enumerated. These led us to examine the governing ideas of democratic policing. These are predicated on the Rawlsian theory of justice with emphasis upon the difference principle. Finally, a list of six expectations was outlined by which policing practices could be judged. Efficiency and effectiveness are not among them.

Recommended Reading and References

Bayley D (2005) Changing the guard. Oxford University Press, New York
Bayley D, Perito R (2010) The police in war. Lynne Reiner, Boulder/London
Berkeley G (1969) The democratic policeman. Beacon, Boston
Bittner E (1970) The functions of police in modern society. USGPO, Washington, DC
Brodeur J-P (2010) The policing web. Oxford University Press, New York
Brogden M, Nijhar P (2005) Community policing: national and international approaches. Willan, Devon
Ericson RV, Haggerty KD (1997) Policing the risk society. University of Toronto Press, Toronto
Hughes EC (1958) Men and their work. Free Press, Glencoe
Liang H (1992) The rise of the police and the European state system from Metternich to the second world war. Cambridge University Press, Cambridge
Manning PK (2010) Democratic policing in a changing world. Paradigm, Boulder
Sinclair G (2006) At the end of the line: colonial policing and the imperial endgame, 1945–1980. Manchester University Press, Manchester
Weber, M (1947) The theory of economic and social organization. The Free Press, Glencoe
Westley W (1971) Violence and the police: a sociological study of law, custom and morality. MIT Press, Cambridge, MA

Design

▶ Innovation and Crime Prevention

Design Out Crime

▶ Designing Products Against Crime

Designing Out Crime

▶ Designing Products Against Crime

Designing Products Against Crime

Paul Ekblom
Design Against Crime Research Centre, Central Saint Martins College of Arts and Design, University of the Arts London, London, UK

Synonyms

Design Out Crime; Designing out crime; Hot products

Overview

Much practice and research exists that addresses how the design of the built environment increases, or decreases, the risk of crime, particularly through the architectural approach of ▶ Crime Prevention Through Environmental Design (CPTED) and place-based ▶ situational crime prevention (SCP). But this entry covers the domain of *products*, essentially two- or three-dimensional objects that have been designed and manufactured in some way and which may be portable (e.g., laptops), mobile (e.g., cars), movable (e.g., home cinema TV sets), incorporated in another product (e.g., a tamper-evident lid for a medicine container), or installed in a place (e.g., a cash machine). There are however some transferrable concepts with architectural approaches: for example, certain products such as handbags or vehicles can be considered "enclosures" with equivalent "access control" issues to buildings. However, such transfer is often difficult because CPTED has developed its own terminological traditions which are not always clear or well connected with other crime prevention or design approaches.

The first section of this entry begins by defining key terms relevant to design against crime, such as risk and risk factors. It then reviews how the latter feature in situational crime prevention notably via the phenomenon of hot products, the underlying causes of elevated risk, and the risk life cycle of products. The second part covers the response to elevated risk, notably via intervention through design, covering both content and process; anticipation of future risks; and evidence of effectiveness. The important role of businesses in creating or reducing crime opportunities in manufactured products, and the difficulties of influencing their "design decision-making" to give some weight to security, is covered only briefly; more is in Ekblom (2012a) and Hardie and Hobbs (2005). The creativity and ▶ innovativeness of criminals themselves is well addressed in Cropley et al. (2010). The role of government in incentivizing and otherwise leading on design is discussed in Clarke and Newman (2005) and was exemplified in the UK Home Office's Design and Technology Alliance (see www.designcouncil.org.uk/our-work/challenges/security/design-out-crime/ for useful case studies).

Fundamentals of Risk and Risk Factors

Crime, unlike mercy, doesn't fall like a gentle rain evenly covering the land – it gathers in pools. Risk of crime is concentrated in particular ▶ places, on particular ▶ victims, and on particular products, the focus of this entry. This concentration has two kinds of implication. On the one hand, it gives strong clues about the causation of criminal events, whether concerning the targets or tools of those crimes or the insecurity of their immediate situation; on the other, it guides the kind of ▶ situational crime prevention strategy that can be adopted. That strategy can be developed either in reaction to an established pattern of risk or in anticipation, but in either case the underlying rationale is the same. If you – as policy maker, police officer, designer, manufacturer, or consumer – can identify the targets and tools at elevated risk of featuring in crime, then you can

respectively concentrate your preventive policies and practices, direct your costly operational resources, design and incorporate elevated security performance in particular products, and choose the make and model of product you buy according to security ratings, as happens, say, with the UK car theft index (Laycock 2004).

Definitions

The term "risk" is used loosely in SCP, normally covering probability alone, and implicitly the risk of *harm to offender*. But risk can be decomposed into possibility (the nature of the undesired events), probability, and harm (Ekblom 2012b), and prevention can address each. Eliminating the *possibility* includes, say, replacement of tempered beer glasses with toughened or plastic ones so they cannot be misused as weapons (see, e.g., www.designcouncil.org.uk/our-work/challenges/Security/Design-out-crime/Alcohol-related-crime/). Reducing the *probability* includes providing clips at tables in bars to combat theft of customers' bags (Ekblom et al. 2012, and see www.designcouncil.org.uk/our-work/challenges/Security/Design-out-crime/Case-studies1/Stop-Thief-Chair-and-Grippa-Clips/). Reducing the *harm* includes designing an easy-to-use backup system for numbers stored on mobile phones (the numbers on stolen phones would otherwise be lost). Crime prevention approaches have until recently underemphasized harm reduction, focusing instead on "cutting the numbers of incidents." Harm is both something for designers to avoid, reduce, or mitigate and a consideration in setting priorities within the design process. Harm can be further divided: With a stolen purse harm could befall the user (if assaulted during a snatch), the product (the bag handle ripped), contents taken or damaged (e.g., phone), and a whole new range of victims through "propagated" offenses (e.g., identity theft from bank cards or burglary using stolen keys).

"Risk factors" is a concept derived from health (e.g., risk factors for heart disease). It is one of the few concepts used right across the crime prevention field. Coverage ranges from risk (and protective) factors for offending (such as poor parenting) through to criminogenic properties and features of places (e.g., the Burgess scale (Armitage 2006)) covering the immediate environment of houses (such as whether they are sited on a corner plot) and through to products (e.g., their value-to-weight ratio – Cohen and Felson 1979).

Risk Factor Approaches in Situational Crime Prevention

Classes of items at elevated risk of crime have been dubbed "hot products" (Clarke 1999), covered below. Products may be hot by virtue of their intrinsic material value (such as jewelry or bronze statuary), their manufactured-in value (such as a mobile phone), or some combination. In either case, this "reward" value (using ▶ rational choice perspective terms – Cornish and Clarke 1986) is often accompanied by some wider elements of opportunity, enabling the product to be taken with relatively little effort or risk to the offender. Of course, risk and effort may partly reside in the nature of the environment in which the products are typically found, such as whether guardians of targets (Cohen and Felson 1979) or other kinds of crime preventer (Ekblom 2011) are present, capable, and motivated. But much of that opportunity may reside in the rewarding and/or vulnerable design of the product itself; and even if the design is not obviously "culpable" (e.g., an easy-to-steal car or a provocative poster), design solutions may be the most reliable and/or cost-effective remedy.

The risk factor approach to products and crime was pioneered, like much else, by Cohen and Felson (1979) and Clarke (1999). Cohen and Felson generated the first of many acronyms, in the shape of VIVA. Value, inertia (i.e., weight), visibility, and access are a set of risk factors empirically backed by, among other things, correlating the decreasing weight of particular products (like TV sets) in the Sears catalogue with increasing crime risk. Clarke extended the analysis in a report using target-of-crime data from, for example, the British Crime Survey, to generate the more widely used CRAVED acronym:

- **C**oncealable
- **R**emovable
- **A**vailable
- **V**aluable
- **E**njoyable
- **D**isposable

CRAVED focuses more clearly on how the properties of particular classes of product connect to criminal opportunity from the offender's perspective, both as rewarding ends in themselves (valuable, enjoyable) and as means to achieving those ends via reduced risk and effort (concealable, removable, available, disposable). Examples of hot products include mobile/cell phones and cash.

Other risk factor acronyms have since emerged to demonstrate the versatility of this approach including AT CUT PRICES (Gill and Clarke 2012). This characterizes fast-moving consumer goods like batteries:

- **A**ffordable
- **T**ransportable
- **C**oncealable
- **U**ntraceable
- **T**radeable
- **P**rofitable
- **R**eputable
- **I**mperishable
- **C**onsumable
- **E**valuable
- **S**hiftable

A complementary "protective factors" approach was developed by Whitehead et al. (2008) who summarized the crime-resisting properties to design into mobile/cell phones. IN SAFE HANDS describes phones with these characteristics:

- **I**dentifiable – by owner, for example, through marking.
- **N**eutral – anti-theft design features should not adversely affect user's experience or elevate risk of other crimes.
- **S**een – to be protected – deterrence.
- **A**ttached – mechanical/electronic links to its owner.
- **F**indable – lost/stolen product can be tracked and found.
- **E**xecutable – can be deactivated if lost/stolen.
- **H**idden – for example, about the person, and used covertly.
- **A**utomatic – security built-in/automated.
- **N**ecessary – to be the owner and to be able to use a product, for example, via mechanical keys, codes, and biometrics.
- **D**etectable – make it obvious that product is being/has been stolen, for example, via alarm.
- **S**ecure – protection itself should not be easily removable or hackable.

Most of these factors describe the causal mechanisms whereby risk is reduced. Neutral and automatic relate to avoiding interference with other design requirements for the user, and secure describes self-protection of the security function. The latter connects with Ekblom's (2012b) distinction between a security feature on a product being "in function" (i.e., delivering the protection as intended) and "as object" (being a target of an attack intended to disable the security, steal the product for its valuable material, or vandalize it).

The risk factors approach, or at least how it is realized in practice, can be criticized. For example, concealable has a very different role depending on *when* in the theft *script* (Cornish 1994) the concealment occurs (Ekblom and Sidebottom 2008): when the thief is seeking a target (concealed in owner's pocket) versus when the thief is making off with a target (concealed in the thief's pocket). This has significant practical implications, introducing a design conflict between making the product concealable for users but not for thieves. Resolutions may involve active discrimination, for example, with computers which "report themselves" to owners or recovery companies such as Immobilise when separated from their registered place of use or cash-in-transit boxes which churn out smoke when stolen.

Risk factor lists are also somewhat ad hoc. To make the approach more systematic and to augment the capacity for exploring empirical risk patterns and generating new lists, an abstraction can help. Ekblom (e.g., 2008) devised the Misdeeds and Security framework to characterize very generic crime risks (and corresponding prevention opportunities) originally for assessing the criminogenic/criminocclusive impact of innovations in science and technology. The misdeeds are broad ways whereby products can feature in crime: a camera phone, say, can be:

- **M**isappropriated or stolen
- **M**istreated or deliberately damaged

- **M**ishandled, for example, smuggled
- **M**isbegotten or counterfeited
- **M**isused as a tool for crime, for example, in anonymous drug deals
- **M**isbehaved with, for example, cyber-bullying

Even broader is to treat the product as a *target* of crime (the first 4) or *contributor* to crime (the last 2). The contributor concept connects with the "crime facilitators" – tools or weapons – of ▶ SCP and "resources for offending" in the Conjunction of Criminal Opportunity framework (Ekblom 2011). A product could, of course, alternatively act as a *resource for crime preventers* as discussed below.

Underlying Causes of Risk

Risk factors are *correlates* of heightened possibility, probability, and harm attending particular products. What causal mechanisms underlie them? This is important for connecting the nature and design of products with underlying *theory*. Theory confers the capacity to generate variety in candidate modifications and innovations which are a priori plausible rather than "shots in the dark."

Sometimes elevated risk comes predominantly from *exposure* factors – some products tend to be left unattended (e.g., cars) or worn or used in risky environments (bling jewelry in clubs, mobile/cell phones on late-night streets). More localized exposure factors include being close to markets for stolen goods, being in locations where drug addicts steal to fuel their habit, or being the subject of aggressive and incautious sales techniques that leave valuable items accessible on the retailer's shelves.

Otherwise, predominant causes center on products themselves. *Value* is obviously important here – even if a satnav is concealed in a compartment in a car, offenders may seek it out. Value of course can change, as with steep fluctuations in commodity prices (Sidebottom et al. 2011). Other motivating causes engendered by products include acting as *precipitators* in the crime situation. These include *prompts* (drawing attention to a stealable object, such as a flashy new bike) and *provocations* (e.g., a coin-operated drink dispenser that swallows money and doesn't deliver).

In yet more cases, some sort of inherent *vulnerability* attends the product. The term vulnerability has been used variously, but Ekblom and Sidebottom (2008), attempting to define a consistent suite of product security concepts, suggest it be confined to covering all criminogenic properties (those enhancing probability of crime) of products *except* the motivation they engender. These normally relate to being seen and taken by the offender. In the case of harm, the product can be inherently *susceptible* to the actions of the offender – easily damaged, tampered with, etc.

These causes often reinforce one another. For example, properties that make phones inherently attractive to offenders, such as small size and portability, also make for ease of theft. Perhaps the most generic cause of elevated risk of mass-produced products is what might be called their *promiscuity* – they can be bought by anybody, sold on by anybody, and used by anybody, and virtually identical copies may be found throughout the community.

Vulnerability and susceptibility are not absolutes but depend on the resources the offender can bring to bear in taking, damaging, or manipulating the target product. These include other, misused products (e.g., portable cutting tools) and strength and dexterity (e.g., breaking anchorages and picking locks). Reflecting this understanding, insurance specifications for secure products like vehicles nowadays are stated as *performance criteria* (e.g., "resist attack by currently available tools for a minimum of 5 min") rather than technical construction (e.g., "lock must be made of manganese steel").

The Risk Life Cycle of Products

Felson (1997) observed that, besides a life cycle of legitimate use, products have a criminal one too:

1. Product does not exist.
2. Product exists, but few consumers know how to use it.
3. Product spreads, gains interest, worth stealing.
4. Product everywhere, no longer worth stealing.

This should produce an "inverted U curve" over time, of accelerating then decelerating risk. However, the "criminal nirvana" of saturation rarely occurs in reality. Both fashion and

marketing/manufacturing tricks to get people to buy the latest model (not to mention true obsolescence and product unreliability) continue to drive both legitimate consumption and theft long after everyone possesses their first mobile phone.

Fundamentals of the Response to Elevated Risk

There are several broad strategies for responding to elevated risk of crime associated with particular products. The products can be:

- Safe – kept in a guarded or locked environment, like bullion.
- Secured after sale – protected by a dedicated security device (e.g., a "crooklock" linking steering wheel and brake pedal of unattended cars).
- Secured in production – incorporating specialized security components like anti-counterfeiting stickers.
- Security adapted – where design features explicitly reduce vulnerability (like anti-picklocks) or lower value (like the folding Puma bike whose diagonal down-tube is replaced by a tensioned steel cable, which unlocks at one end to wrap around the bike stand and which, if cut, renders it unrideable and unsaleable – see www.designagainstcrime.com/projects/puma-bike/).
- Inherently secure – by virtue of weight or bulk, for example, home cinema TV sets are currently awkward to carry off (but future technologies may see roll-up versions).

Applying the Misdeeds and Security framework cited above (fuller treatment is in Ekblom 2008), products can be:

- Secured against misappropriation – for example, vehicles with built-in immobilizers
- Safeguarded against mistreatment – for example, street signs that avoid couching regulations in provocative, confrontational terms
- Scam-proofed against mishandling and misbegetting – for example, fold-over airline baggage labels concealing holidaymakers' addresses from burglars' touts; or anti-copying functions within DVDs
- Shielded against misuse and "sivilized" against misbehavior – for example, "once-only" syringes and waste bins that reveal their contents – including hidden bombs (Lulham et al. 2012, and see www.designcouncil.org.uk/our-work/challenges/Security/Design-out-crime/Case-studies1/An-anti-terrorist-rubbish-bin/) or metro station seating shaped to discourage rough sleeping

Design and the Design Process

Links with Situational Crime Prevention

Product design connects closely with, and applies, many of the 25 techniques of ▶ situational crime prevention (see www.popcenter.org/25techniques/). It can even extend this list, for example, with the concept of "target softening" – with, say, the lock whose bolt can swivel in its housing, causing hacksaw blades to slip.

More theoretically, product design engages with the risk, effort, and reward (as encountered or perceived by the offender) of the Rational Choice perspective (Cornish and Clarke 1986). The Karrysafe handbag (www.designagainstcrime.com/projects/karrysafe/) has a Velcro fastening which increases the risk of the owner hearing or feeling the thief's action. An anchor cable for securing laptop to table leg increases the effort and resources (a cutter is needed to release it). Ink tags clipped to expensive clothing spoil the reward when shoplifters try to remove them.

SCP also seeks to manipulate *crime precipitators*, which add emotional/motivational/perceptual influences to the opportunity to act out the emotion or realize the criminal goal thus awakened. A frustrating door entry system can provoke damage from "machine rage"; a stylish new mobile phone can prompt thoughts of theft.

Even if the design is not obviously "culpable" for a product's elevated crime risk, design solutions may be the most reliable and/or cost-effective remedy, for several reasons:

- Design potentially removes the burden of effort from guardian of target products: with central locking, for example, car owners no longer must remember to lock all the individual doors of their vehicle.

- Mass production potentially enables incorporation of security into a huge proportion of product classes and individual items, covering many individual crime situations.
- Such mass coverage can supply the "herd immunity" needed for impact – when the proportion of a particular kind of items protected is high enough for thieves to give up on the whole category.

Designers can confer criminocclusive or harm-reducing properties on their products by intervening in the materials (e.g., resistant to damage), structural features (e.g., concealable), or functionality of a product (incorporating security or securing functions). Even the products' packaging can be recruited for security (Segato 2012). Criticisms of the design approach center on "paranoid products" (Gamman and Thorpe 2007) where the security focus is excessive, fear arousing, ugly, or inconvenient. But these apply to *poor* designs and clunky "engineering" solutions. Done properly, design can resolve a range of contradictions or "troublesome tradeoffs" (Ekblom 2008, 2012a, b), for example, between security and aesthetics, cost, convenience, and social inclusion (e.g., locks usable by elderly/disabled people).

The important thing is that designers capture all these requirements in the design *process*, which should begin with a penetrating analysis of the diverse stakeholders' interests. This should provide the basis for an effort to maximize the meeting of the often-contradictory requirements – not by seeking compromises but by applying ingenuity and creativity in a process of iterative development and testing (see Ekblom 2012c). The end result should be simultaneously user friendly while abuser unfriendly (Ekblom 2005). Further accounts of the design against crime process are in Thorpe et al. (2009) and www.designagainstcrime.com/methodology-resources/design-methodology/#users-abusers; and on the UK Design Council website. See the "double-diamond" model of design at www.designcouncil.org.uk/designprocess and the guide to designing out crime at www.designcouncil.org.uk/our-work/challenges/Security/Design-out-crime/Design-out-crime-guide/.

To support a more efficient, and potentially more effective, design effort, several preconditions must be established:

- First, designers require a "think thief" (or "think terrorism") mindset (Ekblom 2005), which may not come naturally if they assume all those who will encounter their product are legitimate users and honest, well-behaved citizens.
- Second, they need the simultaneous guidance and constraint offered by theoretical principles, so they reliably come up with plausible ideas as a starting point (though we should preserve the "wonky thinking" and importation of "foreign" ideas that generate true novelty).
- Third, they need "design freedom" to innovate, which is best served by performance-based requirements (as described) and tested theoretical principles rather than detailed technical specifications or specific exemplars of products that have successfully resisted crime – though particular security features can be innovatively recombined and tweaked to adapt to new products or new contexts of use.
- Finally, developing a clear design rationale (Ekblom 2012c) is important both to sharpen thinking and communicate with other designers and clients. One such rationale is the Security Function framework (Ekblom 2012a). This systematically describes products in terms of:
  1. *Purpose* (what are they for?)
  2. *Security niche* (how do they fit with the "ecology of security" – are they, for example, security products, securing products like "Stop Thief" chairs for cafes with notches to securely hang bags behind one's knees (see www.designcouncil.org.uk/our-work/challenges/Security/Design-out-crime/Case-studies1/Stop-Thief-Chair-and-Grippa-Clips/), or inherently secure products?)
  3. *Mechanism* (how do they work in cause-effect/theoretical terms?)
  4. *Technicality* (how are they constructed and how do they operate?)

Retrospective descriptions of products using this framework include the Grippa clip for securing customers' bags to bar tables (Ekblom et al. 2012); Meyer and Ekblom (2011) use it to provide a prospective specification for an explosion-resistant railway carriage.

Some of the above preconditions may arguably militate against individualistic deployment by talented designers of sheer intuitive genius. But their establishment enables society to build a broader design against crime capacity, one that enables more of the "field of emerging products" to be covered on a more routine basis. We thus raise the overall ground level of design fitness rather than achieving a few spectacular, but isolated, peaks.

Anticipating Risk Through Design

Every new product design is a bet on the future, whether concerning market success or undesired side effects like crime. Continual arrival on the market of new, naively insecure products generates what Pease (2001) calls *crime harvests*, followed by hasty retrospective efforts to cope with the crime and clumsily patch the damage by remedial design. The classic example has been with mobile/cell phones. While older "phone cloning" leaks are now plugged with the switch from analogue to digital systems, arguably the early vulnerabilities enabled the establishment of a crime market, with a persistent corpus of criminal expertise, criminal service providers, and criminal networks.

Advances in technology also produce a steady stream of new resources for crime, like cordless drills or pocketable 12 V batteries (which can be misused to energize car door locks). Previously secure items become vulnerable overnight. In fact, adaptive, entrepreneurial offenders and a changing technological backdrop set the scene for arms races between offenders and preventers, especially designers. The classic case is the evolution of the safe. The pace of the race is boosted by dissemination through the Internet – there are, for example, many lock-picking forums. Such arms races are described in Ekblom (2005). The basic strategy for handling them is (1) to develop ways to anticipate offender moves and countermoves and (2) to build designers' capacity (and manufacturers' will) to out-innovate the offenders and more generally to design variety and upgradeability into products.

The risk factor approach naturally primes anticipation. An ambitious attempt to develop a theft-proofing approach for personal electronic products was the EU-funded Project MARC (Armitage 2012). The basic plan was to devise a system for (1) determining the anticipated risk of theft attending some new product exemplar and then (2) incorporating, at the design stage, a commensurate level of security (obviously, products judged to be potentially at elevated risk of theft from their own properties and/or context of use should be given correspondingly higher security specifications). An attempt was made to try this out with a sample of existing mobile personal electronic products, rated by diverse experts, but various difficulties arose. For one thing, it was judged that a security checklist approach would impose an artificial ceiling on the exercise of ingenuity and skill in crime preventive design. This could lead designers to design-down to the level of security required by the checklist and militate against innovation. The security checklist also understates the degree to which security is specific to product type, which rather removes the justification for standardization. Nevertheless, this first serious attempt at crime proofing of products is unlikely to be the end of the story. The original lead researcher, in a recent reprise (Armitage 2012), makes a strong case, with practical suggestions, for taking a modified approach forward.

Horizon-scanning and foresight approaches (Department of Trade and Industry 2000) acknowledge the need to entertain diverse possibilities when making products robust to the future: specific predictions will likely be wrong. One such case was the TV set-top box, designed to allow analogue TV sets to receive digital signals. This seemed a likely hot product, until the service providers changed their marketing strategies from selling the boxes at cost to heavily subsidizing the price and recouping revenue from the additional service. Given such uncertainties,

designers could cope by incorporating into their products some flexibility and upgradeability of the security function.

Mobilizing Design of Products Against Crime

Designs are often intended to work with human crime preventers. The Grippa clip (Ekblom et al. 2012), for example, requires bar customers to fasten their bag to it. Others work against unintentional crime promoters. The M-shaped "caMden" bike stand (designed by Adam Thorpe (Thorpe et al. 2009) on the basis of extensive research into locking behavior and secure parking configurations) nudges cyclists to lock their bike securely (both wheels and the frame) rather than relying on a single lock in the middle of the crossbar, which leaves the wheels removable and the frame liable to misuse as a tool for its own theft – serving as a lever to snap the lock.

Mobilization of preventers is challenging. Ekblom (2012d) describes how various mobilization failures among bar customers, bar staff, management, and senior company executives of one company left the Grippa clips unused (though not so in other companies' venues).

Governments could play various roles in modifying criminogenic products (Clarke and Newman 2005). Policy justifications for governments to mobilize companies center on "polluter pays" principles where a company that generates crime opportunities that fall as "externalities" on other victims or taxpayers is required to modify their products or to compensate in some other way (e.g., Newman 2012; Roman and Farrell 2002). Mobilization can be motivated through incentives including tax breaks, regulations, and "naming and shaming" of criminogenic designs. A useful review was undertaken by the UK Home Office (2006). An example of incentivization of vehicle manufacturers to improve security through awakening market demand is the UK Home Office's Car Theft Index (latest version covering 2005–2006 at http://data.gov.uk/dataset/car-theft-index-2004-2006). This encourages buyers to take theft rates of particular makes and models into account when choosing which to purchase. (Learmount (2005) makes the wider case for demand-led influence in a review of the field.) Similar pressures from insurance companies have also been effective (Hardie and Hobbs 2005).

Does Design Against Crime Work?

Assessment and feedback from workshop tests, field trials, user and service engineer experience, and ultimately sales, profitability, and market leadership are inherent to the iterative process of directed improvement that is product design (Thorpe et al. 2009). In impact evaluation and cost-effectiveness terms normally applied to crime prevention, however, there is unfortunately little hard evidence that relates to product design as opposed to "target hardening" and other situational approaches in general. Such evidence as exists is often characterized by weak research designs; formally evaluated products were summarized in Clarke and Newman (2005, Table 4), and few such studies have emerged since.

One reason is that prototypes are expensive to produce and test in sufficient quantities to support an impact evaluation of sufficient statistical power (Bowers et al. 2009). Another is the timescale for developing a product then evaluating it within a typical research funder's time frame.

Circumstantial, correlational evidence points to the contribution of vehicle security technology towards the substantial and sustained reduction of theft of cars in the UK in recent years, following implementation of a European Directive on compulsory factory fitting of immobilizers from 1998 (Webb 2005). British Crime Survey figures (Home Office 2007, Table 2.01) show theft of vehicles reduced by 65 % from 1995 to 2006–2007 following the design of improved security into the vehicle. None of the case studies commissioned by the Design Council for the Home Office have been formally evaluated.

Other evidence is more anecdotal but almost entirely self-evident (Clarke and Newman 2005). An example is the fabric curtain between certain London Underground train carriages, retrospectively fitted to stop boys riding the couplings. A glance reveals nowhere left to stand. But self-evidence cannot be taken for granted and gives no information on comparative cost-effectiveness.

A recent study (Sidebottom et al. 2009) of attempts to reduce bike theft by installing

advisory stickers on the bike stands has yielded reliable intermediate outcome evidence, important where behavioral change of people acting as crime preventers or promoters (Ekblom 2011) is sought. The stickers were designed after systematic observation of bike-locking behavior and analysis of perpetrator techniques. The simple advice – lock both wheels and frame to the stand – yielded significant and substantial reduction (from 62 % to 48 % of observations) in the proportion of bikes locked insecurely (available funding did not, however, cover evaluation of impact on theft).

Conclusions and Future Research

The practice of systematically designing products against crime and the research into that practice have barely left their infancy. Concepts and frameworks are beginning to emerge (and will need further integration with one another and with other crime-oriented disciplines like SCP or schools of practice like CPTED). They have only superficially tapped the wealth of knowledge and experience that is the field of design.

The emerging domain faces a significant challenge in motivating designers, producers, and consumers to press for, and to use, secure designs. And any knowledge that does accumulate must be considered a "wasting asset," vulnerable to becoming out of date with social and technical change and adaptive offenders. Finding ways to incentivize designers and their clients, and developing and building innovative capacity among designers, is the only way to keep ahead in the long run.

We still lack a sufficient range of rich and rigorous case studies to build on. In particular, the effort to find hard evidence of the cost-effectiveness of product design against crime must continue. Only then will design against crime fare better in obtaining sustained funding and attention from the government. The evidence may also help convince consumers to favor products so designed and manufacturers to routinely include security in their requirements capture.

Related Entries

- ▶ Crime Prevention Through Environmental Design
- ▶ Hot Spots and Place-Based Policing
- ▶ Innovation and Crime Prevention
- ▶ Motor Vehicle Theft
- ▶ Naming and Shaming of Corporate Offenders
- ▶ Problem-Oriented Policing
- ▶ Rational Choice Theory
- ▶ Repeat Victimization
- ▶ Routine Activities Approach
- ▶ Situational Crime Prevention
- ▶ Theories for Situational and Environmental Crime Prevention

Recommended Reading and References

Armitage R (2006) Predicting and preventing: developing a risk assessment mechanism for residential housing. Crime Prev Community Saf: Int J 8: 137–149

Armitage R (2012) Making a brave transition from research to reality. In: Ekblom P (ed) Design against crime: crime proofing everyday objects, vol 27, Crime prevention studies. Lynne Rienner, Boulder

Bowers K, Sidebottom A, Ekblom P (2009) CRITIC: a prospective planning tool for crime prevention evaluation designs. Crime Prev Community Saf 11:48–70

Clarke R (1999) Hot products: understanding, anticipating and reducing demand for stolen goods, vol 112, Police Research Series. Home Office, London

Clarke R, Newman G (2005) Modifying criminogenic products – what role for government? In: Clarke R, Newman G (eds) Designing out crime from products and systems, vol 18, Crime prevention studies. Willan, Cullompton

Cohen L, Felson M (1979) Social change and crime rate changes: a routine activities approach. Am Sociol Rev 44:588–608

Cornish D (1994) The Procedural analysis of offending and its relevance for situational prevention. In: Clarke R (ed) Crime Prevention Studies, vol 3. Criminal Justice Press, Monsey

Cornish D, Clarke R (1986) The reasoning criminal. Springer, New York

Cropley D, Kaufman J, Cropley A, Runco M (2010) Creativity: the dark side. Cambridge University Press, Cambridge, UK

Department of Trade and Industry (2000) Turning the corner. Report of Foresight Programme's Crime

Prevention Panel. Department of Trade and Industry, London
Ekblom P (2005) Designing products against crime. In: Tilley N (ed) Handbook of crime prevention and community safety. Willan, Cullompton
Ekblom P (2008) Designing products against crime. In: Wortley R, Mazerolle L (eds) Environmental criminology and crime analysis. Willan, Cullompton
Ekblom P (2011) The 5Is framework for crime prevention, security and community safety. Palgrave Macmillan, Basingstoke
Ekblom P (2012a) The private sector and designing products against crime'. In: Welsh B, Farrington D (eds) The oxford handbook on crime prevention. OUP, Oxford, pp 384–403
Ekblom P (2012b) The security function framework. In: Ekblom P (ed) Design against crime: crime proofing everyday objects, vol 27, Crime Prevention Studies. Lynne Rienner, Boulder
Ekblom P (2012c) Happy returns: Ideas brought back from situational crime prevention's exploration of design against crime. In: Farrell G, Tilley N (eds) The reasoning criminologist: essays in Honour of Ronald V. Clarke, Crime Science series. Willan, Cullompton, pp 163–198
Ekblom P (2012d) Citizen participation in crime prevention: capturing practice knowledge through the 5Is framework. In: Coester M, Marks E (eds) International perspectives of crime prevention 4. Contributions from the 4th and the 5th Annual International Forum 2010 and 2011 within the German Congress on Crime Prevention. Forum Verlag Godesberg GmbH, Mönchengladbach
Ekblom P, Sidebottom A (2008) What do you mean, 'is it secure?' Redesigning language to be fit for the task of assessing the security of domestic and personal electronic goods. Eur J Crim Policy Res 14(61–87):20
Ekblom P, Bowers K, Gamman L, Sidebottom A, Thomas C, Thorpe A, Willcocks M (2012) Reducing handbag theft in bars. In: Ekblom P (ed) Design against crime: crime proofing everyday objects, vol 27, Crime Prevention Studies. Lynne Rienner, Boulder
Felson M (1997) Technology, business and crime. In: Felson M, Clarke R (eds) Business and crime prevention. Criminal Justice Press, Monsey
Gamman L, Thorpe A (2007) Profit from paranoia: design against 'paranoid' products. European Academy of Design conference 07: dancing with disorder: design, discourse, disaster. Izmir, Turkey. www.bikeoff.org/2007/04/30/profit-from-paranoia-design-against-paranoid-products/
Gill M, Clarke R (2012) Slowing thefts of fast-moving goods. In: Ekblom P (ed) Design against crime: crime proofing everyday products, vol 27, Crime Prevention Studies. Lynne Rienner, Boulder
Hardie J, Hobbs B (2005) Partners against crime: the role of the corporate sector in tackling crime. In: Clarke R, Newman G (eds) Designing out crime from products and systems, vol 18, Crime Prevention Studies. Willan, Cullompton
Home Office (2006) Changing behaviour to prevent crime: an incentives-based approach. Online report 05/06. Home Office, London www.homeoffice.gov.uk/rds/pdfs06/rdsolr0506.pdf
Home Office (2007) Crime in England and wales 2006/2007. Statistical Bulletin 11/07. Home Office, London
Laycock G (2004) The UK car theft index: An example of government leverage. In: Maxfield M, Clarke R (eds) Understanding and Preventing Car Theft, Crime Prevention Studies 17. Monsey, NY, Criminal Justice Press
Learmount S (2005) Design against crime. In: Clarke R, Newman G (eds) Designing out crime from products and systems, vol 18, Crime prevention studies. Willan, Cullompton
Lulham R, Duarte O, Dorst K, Kaldor L (2012) Designing a counterterrorism trash bin. In: Ekblom P (ed) Design against crime: crime proofing everyday objects, vol 27, Crime Prevention Studies. Lynne Rienner, Boulder
Meyer S, Ekblom P (2011) Specifying the explosion-resistant railway carriage: a desktop test of the security function framework. J Transp Sec 5:69–85
Newman G (2012) A market approach to crime prevention. In: Ekblom P (ed) Design against crime: crime proofing everyday objects, vol 27, Crime prevention studies. Lynne Rienner, Boulder
Pease K (2001) Cracking crime through design. Design Council, London
Roman J, Farrell G (2002) Cost-benefit analysis for crime prevention: opportunity costs, routine savings and crime externalities, vol 14, Crime prevention studies. Willan, Cullompton
Segato L (2012) Packaging against counterfeiting. In: Ekblom P (ed) Design against crime: crime proofing everyday objects, vol 27, Crime prevention studies. Lynne Rienner, Boulder
Shepherd J (1994) Preventing injuries from bar glasses. Br Med J 308:932–933
Sidebottom A, Johnson S, Thorpe A (2009) Using targeted publicity to reduce opportunities for bicycle theft: a demonstration and replication. Eur J Criminol 6:267–286
Sidebottom A, Belur J, Bowers K, Tompson L, Johnson SD (2011) Theft in price-volatile markets: On the relationship between copper price and copper theft. Journal of Research in Crime and Delinquency 48:396–418
Thorpe A, Gamman L, Ekblom P, Johnson S, Sidebottom A (2009) Bike Off 2 – catalysing anti-theft bike, bike parking and information design for the 21st century: an open innovation research approach. In: Inns T (ed) Designing for the 21st century, vol 2, Interdisciplinary methods and findings. Farnham, Gower
Webb B (2005) Preventing vehicle crime. In: Tilley N (ed) Handbook of crime prevention and community safety. Cullompton, Willan
Whitehead S, Mailley J, Storer I, McCardle J, Torrens G, Farrell G (2008) IN SAFE HANDS: a review of mobile phone anti-theft designs. Eur J Crim Policy Res 14:39–60

Desistance

▶ Desistance and Supervision

Desistance and Supervision

Fergus McNeill[1], Stephen Farrall[2], Claire Lightowler[3] and Shadd Maruna[4]
[1]Scottish Centre for Crime and Justice Research, University of Glasgow, Glasgow, UK
[2]University of Sheffield, Sheffield, UK
[3]Institute for Research and Innovation in Social Services, Glasgow, UK
[4]Queen's University Belfast, Belfast, UK

Synonyms

Corrections; Desistance; Parole; Probation; Rehabilitation; Supervision

Overview

This entry explores the origins and development of arguments about the use of theories of and evidence about desistance from crime as a basis for developing approaches to offender rehabilitation and supervision. We begin by outlining arguments about the relative merits of offense-focused and desistance-focused approaches before going on to review the case for a "desistance paradigm" for correctional practice. In the final section, we outline the process and (some of the) outcomes of a recent project which aimed to further these debates by actively engaging ex-offenders, their families and supporters, frontline practitioners, managers, and policymakers in coproducing a set of recommendations for criminal justice reform, informed by their experiences of desistance and supporting desistance.

Introduction

Desistance theories and research seek to understand and explain how and why people stop offending – and stay stopped. The notion that such studies might provide a framework for offender supervision, and even for criminal justice interventions more widely conceived, has a long history – dating back at least to the work of the Gluecks (1937 [1966]). If, as they argued, desistance is about maturing out of criminal conduct, what could be done through criminal justice interventions to "force the plant" or to accelerate the maturational process? The question is, of course, a very good one, not least in the context of contemporary preoccupations with the economic, human, and social costs of reoffending by ex-prisoners (see, e.g., Ministry of Justice 2010) and with the broader challenges of ex-prisoner reentry (see Petersilia, this volume).

And yet, between the 1930s and the end of the twentieth century, hardly any use of desistance research to inform sentencing and correctional policy and practice is discernible. Instead, the story of this era is the familiar one of the rise and fall and rise again of rehabilitative interventions, a historical cycle linked to but not fully explained by debates about their effectiveness or ineffectiveness. Though these two topics – the process of desistance from crime and the effectiveness of rehabilitative interventions – are obviously linked in several ways, the connections between them did not begin to be properly explored until the turn of the century.

Today, debates and discussions about desistance and how to support it through criminal justice interventions seem to be bubbling up all around the world of corrections, not just in jurisdictions with deep cultural and historical connections, like the UK and the USA, but also in places as diverse as Norway and Singapore.

This entry does not aim to explain this upsurge of interest in desistance nor does it engage with the important and interesting question of when desistance research is relevant (and irrelevant) to criminal justice (see McNeill and Weaver 2010). Suffice it to say that unless criminal justice is concerned on some level with rehabilitation and reducing reoffending, desistance theory and research is unlikely to have much purchase. But to the extent that sentencing and correctional systems, and more specifically to the structure and

practice of supervision, are concerned with these outcomes, understanding how and why people stop offending (with or without help or hindrance from the justice system) has obvious appeal. Throughout the entry, we use the US convention of referring to "sentencing and corrections," meaning the end of the justice process where sanctions are decided and then delivered. Our particular focus is on supervisory sanctions, that is, those sanctions or elements of sanctions, like probation and parole, which involve the supervision of the sentenced person in the community.

Rather than seeking to review theories of and evidence about desistance itself, this entry has the more modest aim of charting the emergence and development of the arguments advanced over the last 12 about the implications of this body of work for offender supervision.

Offense-Focused or Desistance-Focused Supervision

The emergence and development (or perhaps the revival) of debates about how desistance research could and should inform the development of supervision owes a great deal to the work of Stephen Farrall (2002) and of Shadd Maruna (2001). Both books were based in research projects which drew on and developed earlier work by Ros Burnett; her *Dynamics of Recidivism* study (Burnett 1992) was critical in generating new interest in desistance research in the UK. Another important early foray into the study of "assisted desistance" (as opposed to spontaneous or unaided desistance) was undertaken by Sue Rex (1999), who argued explicitly that:

> The knowledge we are beginning to acquire about the type of probation services which are more likely to succeed could surely be enhanced by an understanding of the personal and social changes and developments associated with desistance from crime. (Rex 1999: 366)

Thus, even while Farrall's and Maruna's research projects were ongoing, publications (like Rex's) had begun to emerge which engaged directly with the question of how desistance theory and research might inform supervision. An interesting early example was an edition of *Offender Programs Report* (volume 4, issue 1) which, among several interesting short articles, included a paper from Maruna (2000) in which he argued for a marrying of the desistance and "What Works?" literatures, taking from the former its analyses of the "micro-mechanisms of change" at the individual level and from the latter an appreciation of the general principles of effective rehabilitative intervention.

That marriage however looked ill fated when Farrall's (2002) book was published. It challenged the somewhat narrow and managerialized interpretations of "What Works?" research which, at that time, dominated correctional policy and practice in the UK. As well as developing a searching methodological critique of the "What Works?" research, Farrall's study (based on a qualitative longitudinal study of 199 probationers and their supervising officers) presented findings which suggested that motivation and social context were more clearly associated with desistance than probation supervision and that the focus of supervision (on risk factors and "criminogenic needs") neglected the crucial roles of relationships and social capital in the desistance process. Farrall's (2002) related proposition was that supervision should focus not solely on "offence-related factors" (or "criminogenic needs") but also on "desistance-related needs." The nature of the difference between the two approaches is perhaps best captured by one of the probationers in his study, in response to a question about what would prevent him from reoffending:

> Something to do with self progression. Something to show people what they are capable of doing. I thought that was what [my Officer] should be about. It's finding people's abilities and nourishing and making them work for those things. Not very consistent with going back on what they have done wrong and trying to work out why – 'cause it's all going around on what's *happened* – what you've already been punished for – why not go forward into something. . . For instance, you might be good at writing – push that forward, progress that, rather than saying 'well look, why did you kick that bloke's head in? Do you think we should go back into anger management courses?' when all you want to do is be a writer. Does that make any sense to you at all? *Yeah, yeah. To sum it up, you're saying you should look forwards not back.*

Desistance and Supervision, Table 1 Ideal-type contrasts: offense-focused and desistance-focused practice

	Offense-focused practice	Desistance-focused practice
Orientation	Retrospective	Prospective
Problem locus	Individual attitudes and behaviors	Individual problems and behaviors in social context
Practice focus	Individual attitudes and behaviors	Personal strengths and social resources for overcoming obstacles to change
Medium for effective practice	Rehabilitative programs (to which offenders are assigned on the basis of risk/needs assessment instruments)	Individual processes and relationships
Worker's Roles	Risk/needs assessor, program provider, case manager	Risk/needs/strengths assessor, advocate, facilitator, case manager
Intended outputs	Enhanced motivation	Enhanced motivation
	Pro-social attitudinal change	Changes in narrative/self-concept
	Capacity/skills development	Development of inclusion opportunities
Intended outcomes	Reduced re-offending	Reduced re-offending
		Enhanced social inclusion

> Yeah. I know that you have to look back to a certain extent to make sure that you don't end up like that [again]. The whole order seems to be about going back and back and back. There doesn't seem to be much 'forward'. (Farrall 2002: 225)

McNeill (2003), McNeill and Batchelor 2004), drawing not just on the work of Farrall and Maruna but on a wider range of desistance studies (*as well as* on "What Works?" research), sought to further elaborate what "desistance-focused probation practice" might look like. He argued that such practice would require thoroughly individualized assessment, focused on the interrelationships between desistance factors (linked to age and maturation, social bonds, and shifts in narrative identity), which build towards clear plans to support change. It would also require engaging in active and participative relationships characterized by optimism, trust, and loyalty, as well as interventions targeted at those aspects of each individual's motivation, attitudes, thinking, and values which might help or hinder progress towards desistance. Crucially, in McNeill's (2003) assessment, it would require work not just to develop personal capabilities but also to access and support opportunities for change, for example, around accommodation and employment. Finally, such practice would require approaches to evaluation which were themselves engaging, since such approaches would be vital in learning more from those involved about what persuaded them to desist and about the support that they needed to see their decisions through.

Beyond these practical prescriptions, and inspired by Farrall (2002), McNeill and Batchelor (2004: 66) went on to further elaborate the shift in practice dispositions or perspectives that desistance research seemed to suggest:

Table 1 above contrasts two notional "ideal types" of practice. McNeill and Batchelor (2004) were clear that this was intended only as a heuristic device; arguably neither of these approaches could or should exist in a "pure" form. Rather, the challenge, they argued, was to combine elements of both approaches in a case-sensitive manner. Hence, an offense focus must, of course, be necessary and appropriate given that, within any justice context, it is offending which occasions and justifies state intervention. However, being only or overly offense-focused might in some senses tend to accentuate precisely those aspects of a person's history, behavior, and attitudes which intervention aims to diminish. It may also, they suggested, tend towards misidentifying the central problem as one of individual "malfunctioning":

> Being desistance-focussed, by contrast, implies a focus on the purpose and aspiration of the intervention rather than on the 'problem' that precipitates it. It also tends towards recognising the broader social contexts and conditions required to

> support change. Thus, where being offence-focussed encourages practice to be retrospective and individualised, being desistance-focussed allows practice to become prospective and contextualised. (McNeill and Batchelor 2004: 67)

Although Maruna's (2001) book engaged less directly with the implications of his study for supervision, his ideas (conceived and elaborated along with his colleague and coauthor Tom LeBel) about "strengths-based" approaches to reentry and corrections were already developing in similar directions, informed both by desistance research and by a wider range of influences (Maruna and LeBel 2003, 2009). In essence, Maruna and LeBel (2003) exposed the limitations and problems associated with both risk-based and support- (or need-) based narratives for reentry. The former, they argued, casts the offender ultimately as a threat to be managed, the latter as a deficient to be remedied by the application of professional expertise. By contrast, "[s]trengths-based or restorative approaches ask not what a person's deficits are, but rather what positive contribution the person can make" (Maruna and LeBel 2003: 97).

Drawing on his *Liverpool Desistance Study* (LDS), Maruna et al. (2004) engaged more directly with the implications of the LDS for supervision, but reached similar conclusions. Probation discourse, they suggested, should move away from risks and needs and towards strengths, seeking to support and encourage redemptive and generative processes, such as those involved in constructive service or voluntary activities. Such a discursive shift could signal the positive potential of probationers not just to them but equally importantly to their communities. Although they supported Farrall's (2002) call for a more explicitly prospective or future-oriented form of supervisory practice, they also recognized the need for people to make sense of their pasts and therefore suggested the need for rehabilitative practices to support a reconstruction of the person's personal narrative, one which recognized and repaired wrongdoing but which refused to define or delimit the person by their previous (mis)conduct.

A Desistance Paradigm

By the middle of the first decade of this century, debate about the implications of desistance theory and research had developed to the point where "A Desistance Paradigm for Offender Management" was proposed (McNeill 2006). The "desistance paradigm" was written in the context of a peculiarly British debate about how probation practice should be reframed in the light of both changing evidence and normative arguments. As such, it engaged with two preceding paradigm-defining papers, the first of which (at the height of the "Nothing Works" era) argued for a "Non-Treatment Paradigm for Probation Practice" (Bottoms and McWilliams 1979) and the second deploying emerging evidence about effective intervention approaches to propose a "Revised Paradigm" (Raynor and Vanstone 1994). McNeill (2006) used both desistance research and normative arguments to seek to displace not their earlier paradigms but what he perceived as the misappropriation and misinterpretation of evidence in a managerialized and reductionist "What Works" paradigm that dominated probation policy and practice at that time (Table 2). He summed up the four paradigms as follows:

McNeill (2006: 56–57) summed up his central argument as follows:

> Unlike the earlier paradigms, the desistance paradigm forefronts processes of change rather than modes of intervention. Practice under the desistance paradigm would certainly accommodate intervention to meet needs, reduce risks and (especially) to develop and exploit strengths, but whatever these forms might be they would be subordinated to a more broadly conceived role in working out, on an individual basis, how the desistance process might best be prompted and supported. This would require the worker to act as an advocate providing a conduit to social capital as well as a 'treatment' provider building human capital. Moreover, rather than being about the technical management of programmes and the disciplinary management of orders, as the current term [in England and Wales] 'offender manager' unhelpfully implies, the forms of engagement required by the paradigm would re-instate and place a high premium on collaboration and involvement in the process of co-designing interventions. Critically, such interventions would not

Desistance and Supervision, Table 2 Probation practice in four paradigms

The non-treatment paradigm	The revised paradigm	A what works paradigm	A desistance paradigm
Treatment becomes help	Help consistent with a commitment to the reduction of harm	Intervention required to reduce reoffending and protect the public	Help in navigating towards desistance to reduce harm and make good to offenders and victims
Diagnoses becomes shared assessment	Explicit dialogue and negotiation offering opportunities for consensual change	"Professional" assessment of risk and need governed by structured assessment instruments	Explicit dialogue and negotiation assessing risks, needs, strengths and resources and offering opportunities to make good
Client's dependent need as the basis for action becomes collaboratively defined task as the basis for action	Collaboratively defined task relevant to criminogenic need and potentially effective in meeting them	Compulsory engagement in structured programs and case management processes as required elements of legal orders imposed irrespective of consent	Collaboratively defined tasks which tackle risks, needs and obstacles to desistance by using and developing the offender's human and social capital

> be concerned solely with the prevention of further offending; they would be equally concerned with constructively addressing the harms caused by crime by encouraging offenders to make good through restorative processes and community service (in the broadest sense). But, as a morally and practically necessary corollary, they would be no less preoccupied with making good to offenders by enabling them to achieve inclusion and participation in society (and with it the progressive and positive reframing of their identities required to sustain desistance).

McNeill's argument had been developed partly as the result of undertaking a literature review which aimed to explore the key skills required of supervisors charged with reducing reoffending (McNeill et al. 2005). Importantly, both that review and the paradigm paper included and attempted to integrate findings from desistance studies, "What Works?" correctional research, and the wider literature on the characteristics of effective psychosocial interventions more generally. Despite different methodologies and disciplinary orientations, some similar findings were emerging in desistance and "What Works?" research, for example, about the importance of the worker/client relationship in supporting change and about the need for "brokerage" of access to wider services to address practical needs.

More recently, Maruna and LeBel (2010) have engaged directly with the promise of developing and employing a desistance paradigm for correctional practice. Like McNeill (2006), they begin with the recognition that evaluation evidence (about what works to produce particular outcomes) is not the only form of evidence that matters in developing evidence-based practice (EBP). To focus exclusively on evaluation evidence in guiding intervention choices is to miss the importance of evidence about the very processes that interventions exist to support. Echoing Lewis (1990), they suggest a shift in focus "from programmes to lives," eschewing a correctional or medical model of change. In a more recent paper, McNeill et al. (2012a) have argued that it is simply wrong to treat desistance as the outcome of an intervention; interventions can contribute to change, but they do not "produce" it in any simple sense; desistance can and does exist "before, beyond, and behind" interventions. By focusing on and better understanding the change process, Maruna and LeBel (2010) argue, we may be able to better adapt supervision to contribute to (but not to "produce") the process. It follows that a desistance paradigm must place the person changing and *their* change process (and not the program) center stage, a message that finds support among prisoners and probationers themselves, who are often resistant to being processed through more programmatic interventions (Harris 2005).

Although the views and voices of "offenders" may not carry much weight with political or public audiences of correctional work, given the well-documented problems of supervision violation and program dropout, correctional practitioners would do well to heed the views and voices of those they are seeking to influence. Desistance research, in many respects, is one way (and only one way) in which these views and voices can be heeded.

Maruna and LeBel (2010: 72) sum up their argument for a desistance paradigm as follows:

> the desistance paradigm argues that the search for 'what works' should not begin with existing expert models of crime reduction, but rather should begin with an understanding of the organic or normative processes that seem to impact offending patterns over the life course. That is, if turning 30 is the 'most effective crime-fighting tool' (Von Drehle 2010), then we should seek to learn as much as we can about that process and see if we can model these dynamics in our own interventions.

However, recognizing that the practical implications of such a perspective remain seriously underdeveloped (see also Porporino 2010); Maruna and LeBel (2010) go on to articulate one form of desistance-supporting intervention. Drawing on labeling theory, and on their earlier work (cited above), they make the case for an approach to corrections which stresses pro-social labeling; as well as avoiding negative labeling, this requires practices and systems that expect, invite, and facilitate positive contributions and activities from people subject to supervision and that then certify and celebrate redemption or rehabilitation – de-labeling and de-stigmatizing the reformed offender.

McNeill (2012) has similarly argued, further to his original arguments for a desistance paradigm, that the field of corrections needs its own "Copernican Correction" – one in which supervision and support services revolve around the individual change process, rather than requiring offenders' lives to revolve around programs and interventions. Moreover, he argues for a shift away from seeing the "offender" as the target of the intervention (the "thing" to be fixed) to seeing the broken relationships between individuals, communities, and the state as the breach in the social fabric (or breach of the social contract) that requires repair. Importantly, this casts correctional agencies less as agents of "correction" and more as mediators of social conflicts. The objective becomes not the correction of the deviant so much as the restoration of the citizen to a position where she/he can both honor the obligations and enjoy the rights of citizenship.

From Paradigms to Practices or Not

The arguments reviewed in the previous section are important, but they are arguably more concerned with *reframing* supervision than with *redesigning* it in practice. The clearest attempts at tackling the latter challenge are perhaps to be found in reports prepared by McNeill (2009) and McNeill and Weaver (2010) for the Scottish Government and the National Offender Management Service of England and Wales, respectively. Both of these reports represent attempts to respond to policymaker and practitioner requests for a more explicit articulation of the practical implications of desistance research for supervision practice; both share a similar reticence in responding to these requests. As McNeill and Weaver (2010: 6) explain:

> One of the 'problems' with desistance research is that it is not readily translated into straightforward prescriptions for practice… As Porporino (2010: 61) has recently suggested: 'Desistance theory and research, rich in descriptive analysis of the forces and influences that can underpin offender change, unfortunately lacks any sort of organised practice framework.'
>
> However, though this is a practical problem, it is not necessarily a weakness. Even if we wished that there was a 'desistance manual' that could be prescribed for practitioners, there is not… [D]esistance research itself makes clear that offenders are heterogeneous, their needs are complex and their pathways to desistance are individualised. Overly generalised approaches to interventions therefore are themselves inconsistent with desistance research. It follows that evidence based practice can only really emerge from practitioners' reflective engagement and continual dialogue with those individuals with whom they work, and with the research that should inform how they work.

With this in mind, rather than offering a desistance manual or a desistance program, McNeill's (2009) paper *Towards Effective Practice in Offender Supervision* is an attempt to summarize evidence about both desistance and "What Works?" in order to inform evidence-informed reflective practice. Instead of offering a predesigned and therefore homogenized intervention, he attempts to articulate the range of issues and questions with which a reflective practitioner would have to engage on a case-by-case basis in seeking to support desistance. The "offender supervision spine" that he delineates includes explicit suggestions about how to approach the preparatory, relationship-building stage of supervision, as well as assessment, planning, intervention, and evaluation. McNeill's (2009) suggestion is that a supervisor working their way along this spine (in partnership with the supervisee) must be continually developing and testing evidence-informed "theories of change," seeking to work out together "why and how we think that doing what we propose to do will bring about the results that we seek" and then to implement that plan and to evaluate its progress.

Beyond such emergent models of the supervision process, a number of broader practical implications of desistance research have been identified in the literature. In a very recent overview, for example, McNeill et al. (2012b) identify eight broad principles, the genesis of which is apparent above:

1. Desistance, for people who have been involved in persistent offending, is a difficult and complex process, and one that is likely to involve lapses and relapses. Criminal justice supervision must be realistic about these difficulties and find ways to manage setbacks and difficulties constructively. It may take considerable time for supervision and support to exercise a positive effect.
2. Since desistance is an inherently individualised and subjective process, approaches to supervision must accommodate and exploit issues of identity and diversity. One-size-fits-all interventions will not work.
3. The development and maintenance not just of motivation but also of hope become key tasks for supervisors.
4. Desistance can only be understood within the context of human relationships; not just relationships between supervisors and offenders (though these matter a great deal) but also between offenders and those who matter to them.
5. Although the focus is often on offenders' risks and needs, they also have strengths and resources that they can use to overcome obstacles to desistance – both personal strengths and resources, and strengths and resources in their social networks. Supporting and developing these capacities can be a useful dimension of supervision.
6. Since desistance is in part about discovering self-efficacy or agency, interventions are most likely to be effective where they encourage and respect self-determination; this means working *with* offenders not *on* them.
7. Interventions based only on developing the capacities and skills of people who have offended (human capital) will not be enough. Probation also needs to work on developing social capital, opportunities to apply these skills, or to practice newly forming identities (such as 'worker' or 'father').
8. The language of practice should strive to more clearly recognize positive potential and development, and should seek to avoid identifying people with the behaviors we want them to leave behind.

This summary was produced as part of an ongoing project which aimed to develop a much more comprehensive and innovative response to the challenge to "operationalize" desistance (explained more fully in McNeill et al. 2012a). Rejecting the prospect of an often futile, counterproductive, and disrespectfully one-sided conversation between research and practice, the collaborators in this project have sought to foster a dialogue about supporting desistance involving academics, policymakers, managers, practitioners, ex/offenders, and their families and supporters.

The project is entitled "Discovering Desistance" and is funded by the UK Economic and Social Research Council (award no. RES-189-25-0258). Partnered in the USA by

Prof. Faye Taxman's Center for Advancing Correctional Excellence (http://www.gmuace.org/), it aims to explore the experience and knowledge of these different stakeholders in relation to desistance from crime and how correctional supervision in the community can best support it. To this end, the project involves three key elements:

1. Developing, with key stakeholders, user-friendly methods of disseminating existing research about desistance from crime and about supporting desistance in offender supervision. This involved working with ex/offenders, practitioners, and an independent film production company to make a documentary film about desistance (The Road from Crime; http://www.iriss.org.uk/resources/the-road-from-crime) which was released in July 2012.
2. Facilitating dialogue through the "Discovering Desistance" blog site (see http://blogs.iriss.org.uk/discoveringdesistance/) where academics, ex/offenders, and practitioners have all made key contributions to developing the discussion.
3. Running a series of stakeholder workshops (across four jurisdictions) which aim at coproducing a set of clear recommendations about the further development of "practice for desistance" and beginning to delineate the features of an organized framework for offender supervision practice to support desistance. This is the focus of the ongoing final stage of the project.

Though it is too soon to report the findings of these workshops here, it is interesting to note that the ideas and proposals that they have generated extend far beyond supervision practice and into much broader aspects of reentry and reintegration. Some of the recurring "propositions" suggested in the UK-based workshops include the following:

1. **Make greater use of reformed offenders**: Participants called for more meaningful service-user (i.e., supervisee) involvement in the design, delivery, and improvement of policies and provision across the criminal justice system, involvement that could be part of clear career routes and would include developing ways of recognizing and rewarding skills. Participants argued that greater involvement of ex/offenders in mentoring schemes should be a key part of this involvement.
2. **Reduce the reliance on imprisonment**: Participants argued that there is a need to reduce the prison population (especially women, black men, those with mental health issues, and those on short sentences), with freed-up money reinvested in community justice.
3. **Reorientate the philosophy of probation**: They called for a rethinking of probation, making it more "holistic" or "humanized," more focused on the service user's wants, strengths, and aspirations, as well as aiming for more community involvement and a greater degree of flexibility and creativity.
4. **Reconnect probation to local communities**: Participants argued that, in the future, probation offices and officers need to become better connected with local communities.
5. **Mobilize wider support networks**: All of society needs to take on a responsibility for helping people stop offending (organizations, families, and individuals).
6. **Focus on the positive, not the negative or risks**: Participants suggested that criminal justice needed to focus more readily on the positives and what people have achieved *and can achieve* in the future.
7. **Suggestions for supervision, release, and reintegration**: Community supervision needs to work to challenge inequality and promote equality, equalizing life chances and contributing to social justice (pursuing both substantive equality and equality in the criminal justice process).
8. **Redraft the [UK] Rehabilitation of Offenders Act**: Participants suggested the need to redraft the legislation around criminal records, so as to really encourage and recognize rehabilitation, not stand in the way of it. Reformed offenders should have the opportunity to have their record "spent" much earlier than is currently the case.

9. **Educate the general public about the processes of desistance**: They suggested that there is a need to better educate the general public about the process of leaving crime behind and the lives of current and former service users in order to break down the "them" and "us" mentality; this would ensure that there is better public understanding that people are capable of change and that we all have a part to play in supporting change. Criminal justice agencies ought to demonstrate that positive change is possible and show that it is common.
10. **Give people hope**; **show them they have a future**: The criminal justice system needs to become more acquainted with hope and less transfixed with risk, pessimism, and failure.

Conclusions

Though, as this entry demonstrates, many academics have worked hard in recent years to conduct and to disseminate desistance research, their voice in the debate about criminal justice reform is and should be just one among many. Others have different kinds of expertise to bring to this discussion. The Discovering Desistance project has been more about harnessing different forms of expertise than privileging or prioritizing one perspective. Perhaps its most important contribution has been to demonstrate the potential of "coproduction" in supporting change – at both the personal and the systemic levels (Weaver 2011).

As such, the impact on criminal justice – and more narrowly on correctional supervision – of engaging with desistance research is beyond the control of scholars and researchers; those working in and living with the correctional system have already started to talk and think about how and why people build new lives; the "desistance genie" is well and truly out of the bottle. In some respects, the work of academics in trying to stimulate new ways of thinking about supporting change in the criminal justice system and in the practice of supervision is complete. In other respects, it has barely begun. But while academics and researchers have plenty more work to do developing a robust, evidence-informed understanding of these processes and of what supports them, arguments over language, social attitudes, policy developments, and practice processes should not and cannot wait for research to provide "answers." Rather, all of the stakeholders engaged with supervision – policymakers, practitioners, parolees and probationers, and families – need to press on with the urgent business of working out what to do to continue discovering and supporting desistance together.

Related Entries

- History of Corrections
- Penal Justice and Social Injustice
- Penal Philosophy and Sentencing Theory
- Punishment and the State
- Rehabilitation
- Sentencing Research
- Theories of Punishment

Recommended Reading and References

Bottoms A, McWilliams W (1979) A non-treatment paradigm for probation practice. B J of Soc Work 9(2):160–201

Burnett R (1992) The dynamics of recidivism. University of Oxford Centre for Criminological Research, Oxford

Farrall S (2002) Rethinking what works with offenders: probation, social context and desistance from crime. Willan Publishing, Cullompton

Glueck S, Glueck E (1937 [1966]) Later criminal careers. Kraus, New York

Lewis DA (1990) From programs to lives: a comment. Am J Community Psychol 18:923–926

Maruna S (2000) Desistance from crime and offender rehabilitation: A tale of Two research literatures. Offender Program Rep 4(1):1–13

Maruna S (2001) Making good. American Psychological Association, Washington, DC

Maruna S, LeBel T (2003) Welcome home? Examining the "re-entry court" concept from a strengths-based perspective. West Criminol Rev 4(2):91–107

Maruna S, LeBel T (2009) Strengths-based approaches to reentry: extra mileage toward reintegration and destigmatization. Jpn J Sociol Criminol 34:58–80

Maruna S, LeBel T (2010) The desistance paradigm in correctional practice: from programmes to lives.

In: McNeill F, Raynor P, Trotter C (eds) Offender supervision: new directions in theory, research and practice. Willan, Cullompton
Maruna S, Porter L, Carvalho I (2004) The liverpool desistance study and probation practice: opening the dialogue. Probat J 51(3):221–232
McNeill F (2003) Desistance-focused probation practice. In: Chui W-H, Nellis M (eds) Moving probation forward: evidence, arguments and practice. Pearson Education, Harlow
McNeill F (2006) A desistance paradigm for offender management. Criminol Crim Justice 6(1):39–62
McNeill F (2009) Towards effective practice in offender supervision. Scottish Centre for Crime and Justice Research, Glasgow. Available online at: http://www.sccjr.ac.uk/documents/McNeil_Towards.pdf
McNeill F (2012) Counterblast: a Copernican correction for community sentences. Howard J Crim Justice 51(1):94–99
McNeill F, Batchelor S (2004) Persistent offending by young people: developing practice, issues in community and criminal justice, Monograph No. 3. National Association of Probation Officers, London
McNeill F, Weaver B (2010) Changing lives? Desistance research and offender management. Scottish Centre for Crime and Justice Research, Glasgow. Available online at: http://www.sccjr.ac.uk/documents/Report%202010_03%20-%20Changing%20Lives.pdf
McNeill F, Batchelor S, Burnett R, Knox J (2005) 21st century social work reducing reoffending: key skills. Scottish Executive, Edinburgh
McNeill F, Farrall S, Lightowler C, Maruna S (2012a) Reexamining "evidence-based practice" in community corrections: beyond "a confined view" of what works. Justice Res Policy 14(1):35–60
McNeill F, Farrall S, Lightowler C, Maruna S (2012b) How and why people stop offending: discovering desistance. Insights: evidence summaries to support social services in Scotland, No. 15, April 2012, Institute for Research and Innovation in Social Services. Available online at: http://www.iriss.org.uk/sites/default/files/iriss-insight-15.pdf
Ministry of Justice (2010) Breaking the cycle. Effective punishment, rehabilitation and sentencing of offenders. Ministry of Justice, London
Porporino F (2010) Bringing sense and sensitivity to corrections: from programmes to 'fix' offenders to services to support desistance. In: Brayford J, Cowe F, Deering J (eds) What else works? Creative work with offenders. Willan, Cullompton
Raynor P, Vanstone M (1994) Probation practice, effectiveness and the non-treatment paradigm. B J Soc Work 24(4):387–404
Rex S (1999) Desistance from offending: experiences of probation. Howard J Crim Justice 36(4):366–383
Weaver B (2011) Co-producing community justice: the transformative potential of personalisation for penal sanctions. B J Social Work 41(6):1038–1057

Desistance from Crime

Shawn D. Bushway[1] and Raymond Paternoster[2]
[1]School of Criminal Justice, University at Albany, State University of New York, Albany, NY, USA
[2]Department of Criminology and Criminal Justice, University of Maryland, College Park, MD, USA

Synonyms

Criminal career; Desistors; Termination

Overview

Early understandings of the cessation from crime considered desistance to be the event of moving from a state of committing crime to a state of not committing crime. Gradually, however, scholars have begun to understand desistance not as an event but as a process. Fagan (1989) was the first to recognize this, differentiating the process of desistance, defined as the reduction in the frequency and severity of offending, from the event of quitting crime. Le Blanc and Fréchette (1989) also referred to desistance as a set of processes that leads to the cessation of crime, using the term deceleration to refer to a reduction in the frequency of offending prior to cessation. Continuing the dialog, Laub and Sampson (2001) explicitly separated the process of desistance from the termination of offending, which they viewed as the outcome of desistance. There are currently in the literature several excellent reviews of possible theoretical explanations for desistance – most notably Laub and Sampson (2003). This essay describes desistance by explicitly mapping the different processes of desistance to different stochastic time series models.

Describing Desistance

In its most basic form, a time series of individual offending can be *described* by the following

equation, which is known as an autoregressive time series:

$$Y_t = \alpha + \rho Y_{t-1} + e_t \quad (1)$$

where e_t is a time series of uncorrelated shocks. A key assumption of time series analysis is that the process is *stationary*, which simply means that the parameters of the model are stable throughout the time period. This kind of stationary process cannot create a long-term desistance trajectory that declines to zero over time. The path described by Eq. 1 will move to an equilibrium level and then stay flat with short-term variation around the equilibrium line – the very formulation means that there can be no meaningful shift in the level of offending. State dependence (past offending causes current offending) and individual heterogeneity (differences between individuals in the stable propensity to commit crime) as captured in the lagged Y term in Eq. 1 ***cannot*** explain desistance. Therefore, *desistance is inherently a nonstationary process*.

There are four broad classes of nonstationary time series. The first is a *series with a trend*. This trend is based on age. The trend predetermines the path. With a trend, Eq. 1 becomes Eq. 2

$$Y_t = \alpha_t + \rho Y_{t-1} + e_t \quad (2)$$

This elementary model does not try to explain the existence of the trend, except in the most basic or general terms. The best example in criminology for a desistance theory that appeals to a basic trend is Gottfredson and Hirschi's (1990) theory of self-control. In this theory, any change in an individual's time series trend in offending over time is simply attributed to the "inexorable aging of the organism" (1990: 141). Since age is the time marker in this time series, saying age explains desistance is simply the same thing as saying that there is an undefined trend that mimics the trend in age (Bushway et al. 2001). Glueck and Glueck's (1974) maturational theory of the decline in crime over time is but one small step removed from Gottfredson and Hirschi's assertion about age (1990). They are careful to explicitly distinguish age from maturation – which means that the maturational process need not occur at the same age for everyone. However, all this statement does is to extend Eq. 2 to say that there is an unknown distribution of time trends in the population – there is a definitive time trend but everyone does not have the same time trend. This claim leaves open the possibility that this maturational process is preprogrammed and deterministic. Laub and Sampson (2003: 33), for example, characterize these kinds of developmental theories as being preprogrammed – essentially fixed trends: "[d]evelopmental accounts ... focus on regular or lawlike individual development over the life span."

The second type of time series, a *cointegrated time series*, captures the counterargument to Sampson and Laub's characterization of the developmental path. Here, Eq. 3 is developed by adding a time-varying covariate X_t. The coefficient on X_t is time constant. This variable trends in the same way as criminal propensity.

$$Y_t = \alpha + \rho Y_{t-1} + \delta X_t + e_t \quad (3)$$

This basic model in which time-varying covariates can explain the long-term pattern of desistance fits with the class of theoretical models in which theorists simply extended existing theories of the onset of crime to account for its desistance. For example, Agnew (2005) argued that the bulk of offenders desist from crime simply because the strains that they experienced as adolescents that launched them into crime in the first place (school, relationship, and job strains) diminish over time, and the ability to adapt in a conventional way to existing strains increases as they entered adulthood. The movement into adulthood, then, comes with both fewer and/or less intense strains and/or an increased capability to adapt to strain in a nondeviant way. Similarly, Akers (1998: 164) argued that the most important predictor of all dimensions of offending, including desistance, is involvement with delinquent peers: "... the single best predictor of the onset, continuation, or desistance of delinquency is differential association with law-violating or norm-violating peers." Most existing theories of

crime responded to the new conceptual terrain brought about by the criminal career perspective, then, by simply insisting that they could just as easily explain desistance as they could onset or other dimensions of offending.

Developmental or maturational theories of crime can also be thought of as describing a cointegrated time series rather than a deterministic trend to the extent to which the theorist describes a variable or process that explains the change in propensity over the life course. For example, Gove (1985) posits that there are biological and psychological factors over time that peak and decline in the same manner as offending propensity. These factors are plausibly cointegrated with offending propensity.

Although they are skeptical about whether this can be done, Gottfredson and Hirschi (1990) acknowledge the possibility that time-varying covariates can explain long-term change. On the empirical side, Osgood (2005) advocates inserting time-varying covariates with time constant parameters into growth curve models in an attempt to explain the age crime curve. Within the growth curve framework, Osgood (2005) suggests testing to see if the time-varying covariates can detrend the data. This basic approach has been applied by Nieuwbeerta and Blokland (2005) where they look to see how much marriage and employment can explain the age crime curve. It is also seen in Sweelen et al. (2013) in which they look to see how much a set of time-varying covariates can explain the divergence between those who desist from and those who persist in crime. In each case, the researchers are looking to see if the time-varying covariates can make a nonstationary time series stationary – with time constant parameters, the only way this is possible is if the covariates themselves trend or track in the same manner over time as offending propensity.

The third type of time series process that can explain or accommodate nonstationarity is a *time series with a structural break*. A structural break implies that there are two or more sets of parameters, meaning that the causal process is different across different time periods.

$$Y_t = \begin{cases} \alpha_a + \rho_a Y_{t-1} + e_{at} \text{ if } & t<T \\ \alpha_b + \rho_b Y_{t-1} + e_{bt} \text{ if } & t>=T \end{cases} \quad (4)$$

Of course, there can be more than one structural break. There are elements of structural breaks that are harmonious with several desistance theories. For example, the notion of age-graded causal factors is entirely consistent with the idea that the value of coefficients on some time-varying variable changes over time.

A more general way of thinking about structural breaks is that some relatively time-stable component of an individual, such as self-control, changes over time. This is only relevant if life events and social context interact with self-control to affect behavior. In Thornberry's interactional model, for example, the exact nature of state dependence depends in meaningful ways on the individual's relatively stable characteristics (Thornberry 1987). In Thornberry's interactional theory, those individuals who are heavily embedded in crime are less "dynamic," in that, they are less responsive to changes in their environment, and, therefore, are also less state dependent (Thornberry and Krohn 2005). Nagin and Paternoster built on this idea in their own version of an interactional theory when they posited that the impact of sanctions depended in meaningful ways on the person's level of self-control (Nagin and Paternoster 1994). Subsequent empirical work by Wright and colleagues (2004), as well as by Hay and Forrest (2008), have all found evidence for an interaction between life events and stable individual characteristics such as self-control. If this basic preference function shifts over time in purposeful ways, as suggested by Hay and Forrest (2008) and Giordano et al. (2007), then the same inputs and opportunities lead to different behaviors at different times – and state dependent processes can start to head people in a different direction. This situation, where a person experiences different causal processes depending on changes in their underlying personal preferences, extends interactional theories to accommodate a structural break and strengthens the ability of these types of theories to explain long-term changes in offending propensity.

Another type of desistance theory that accommodates a structural break are theories which anchors a change in crime to an offender's change in personal identity (Giordano et al. 2002, 2007; Paternoster and Bushway 2009). The importance of identity theories from this perspective is that they provide an explanation for how fundamental individual characteristics such as self-control can change from one time period to another manifesting itself as a structural break. Changes in identity can trigger fundamental shifts in how people value the future (time discounting) or value their social contacts. Simply saying that preferences change is easy – explaining the mechanism by which they change is both important and difficult (see Akerlof and Kranton 2010). Identity theorists like Giordano and her colleagues (Giordano et al. 2002, 2007) and Maruna and Farrall (Maruna 2001; Farrall 2005) offer social psychological theories of desistance which revolve around structural breaks in the process that generates crime. Basing their views on a symbolic interactionist foundation, Giordano et al. (2002) argue that desistance requires substantial cognitive transformations or "upfront" cognitive work such as the development of a general openness to change, receptivity to "hooks for change," and consistent support from social others. In a later revisiting of this view, Giordano et al. (2007) developed a desistance theory that relies much more heavily on external, social processes – the regulation of emotions and the emotional identity (an "anger identity") of ex-offenders as they struggle with getting out of crime. Maruna (2001) also adopted a theory of desistance that relies on notions of the actor's identity, though not one premised on a change in identity. For Maruna (2001), "making good" does not so much involve an intentional change in the desister's identity from bad to good as it does a reinterpretation of one's criminal past to make it consistent with their current pro-social identity.

The fourth major type of nonstationary time series is a *random walk*, a well-known form that has been found occurring in many contexts, including the stock market price of a company and the financial status of a gambler. Random walks have a unit root:

$$Y_t = \alpha + Y_{t-1} + e_t \tag{5}$$

According to Eq. 5, behavior in a given period is simply where you were in the previous period, plus a constant and a shock. The series has an infinite memory, since any shock is permanently incorporated into the time series. Random walks do not, therefore, return to any mean. The same formula can generate flat, increasing, decreasing, or U-shaped curves, depending entirely on the time series of uncorrelated shocks e_t.

This description of a random walk is consistent with Laub and Sampson's (2003: 34) characterization of life course theories of desistance as the result of a series of random events or "macro-level shocks largely beyond the pale of individual choice (for example, war, depression, natural disasters, revolutions, plant closings, industrial restructuring)." Random walks are inherently unpredictable, and as described by Laub and Sampson (2003: 33–34), this lack of predictability is the key factor which distinguishes life course trajectories from predetermined developmental trajectories:

> Developmental accounts...focus on regular or lawlike individual development over the lifespan. Implicit in developmental approaches are the notions of stages, progressions, growth and evolution... with the imagery being one of the execution of a program written at an earlier point in time. ... In contrast, life-course approaches ... emphasize variability and *exogenous* influences on the course of development over time that cannot be predicted by focusing solely on enduring individual traits. (emphasis added)

Another way to discuss the time series properties of life course theories is to consider the key life course assertion that the impact of life events depends on when they occur in a person's life. This is the notion that "timing matters." To the extent to which this timing dependence is predictable, it is consistent with time series models with structural breaks because the implication of timing dependence is that there are simply different models for different time periods. If there are a small number of changes, and these changes are tied to observable changes in identity, then this age-gradedness should be

both predictable and identifiable. But if there are many structural breaks, and these breaks are tied to malleable social contexts, the age-gradedness becomes much more unpredictable. Indeed, a random walk can be characterized as a time series with N structural breaks, where N converges to the total number of periods in the time series.

The main difference between life course theories (random walks) and identity theories (structural breaks) is the number of breaks. In a world with many breaks, predicting long-term change is difficult. The result is a change in focus to explaining change in any given period, which is driven by these relatively exogenous life events. This conclusion is consistent with empirical practice – if a time series is a true random walk, with no trend and no cointegrated time series, the only feasible strategy is to explain period-to-period, that is, short-term, change. With a time series characterized by random walks, it is simply not possible to explain any long-term pattern because that long-term pattern is driven by random shocks. Ironically, this interpretation of life course theory implies that it is neither possible nor even interesting to study a life course "trajectory" since only period-to-period change contains interesting information – there is no meaning to a *life* course.

All theories of desistance must fit into one of the four basic categories of nonstationary time series models – trends, cointegrated series, series with a structural break, and random walks. Given the distinct empirical character of each of these four basic types of time series, a serious examination of individual time series characteristics should be a fruitful avenue for future empirical research. Further explication of theories within the framework provided by the extensive literature on time series processes should also help to clarify and formalize theories of desistance. Readers interested in seeing empirical examples of this approach should peruse Paternoster and Bushway (2009; Bushway and Paternoster 2012), where some basic illustrations are provided using data from the Cambridge Study in Delinquency Development (CSDD) data (Farrington et al. 2006).

Key Issues/Controversies

Bushway and Paternoster (2012) argue that the evidence shows that desistance is not inevitable as suggested by Sampson and Laub (1993; Laub and Sampson 2003) nor is it a simple part of the biological aging process as suggested by Gottfredson and Hirschi (1990). Moreover, they believe the evidence suggests that individuals stop committing crime not because they cannot physically commit crime anymore, but because they *choose* not to. Some choose to exit before others who wait until much later in their lives to quit crime, and as a result, there is a long right hand tail in the age distribution of offending. This long right hand tail also casts doubt on a strictly structural version of desistance which attributes the initial thrust into conformity to an acquiring of "turning points" or pro-social roles like jobs and marriages. While there is a convincing body of research that documents the ability of marriage and work to decrease crime, this work frequently does not speak very directly or clearly to the *causal mechanism* by which this effect occurs (Sampson et al. 2006). If the explanation is entirely or immediately structural, desistance would be highly correlated with the arrival rates of first marriages and stable employment during the 20s and into the 30s as people move into adulthood. And, indeed, a large portion of desistance clearly does occur between the ages of 20 and 40. *But, it cannot be ignored that employment and marriage have been available states for 20 years by age 40*. A simple matching or sorting story in which people desist when matched to jobs and spouses should not require more than 20 years before it reveals itself.

Of course, it is possible that work (and potentially marriage) has a differential impact depending on age such that work is involved in the desistance process but only when offenders reach a certain age. But this explanation would imply that something about the individual or her set of circumstances has changed with age, their identity and preferences, for example, and this change in turn leads to different choices by the individual. The typical interpretation is that the effect of these variables is age-graded, but this

term simply describes what has to be explained. Another interpretation is that these factors (such as jobs and marriages) have a different impact on different kinds of people, and different kinds of people select into marriage and employment at different ages. Research on employment and crime is now increasingly showing that the established "fact" that employment is bad for youth (but good for adults) is entirely an artifact of selection. Strong controls for selection show that employment has the same modest *negative* impact on crime for youth as it does for adults. Entering into pro-social roles may have a role to play in desistance, but perhaps the acquisition of such roles is only part of the picture and comes later in the desistance process when other obstacles have first been overcome.

The facts of desistance state loudly and clearly that desistance cannot be explained either by strictly biological or structural explanations. If not biology and if not the immediate acquisition of pro-social roles, what then? One possible explanation lies in a person's identity and the corresponding changes this brings in how they weigh the inputs of their decision making, their preferences, and how they make choices (Akerlof and Kranton 2010).

Theoretical Frontiers: Identity Theory

There is a long intellectual tradition in sociology and social psychology which emphasizes the importance of one's identity (Stryker 1968). In recent years, economists have also argued that the preferences people have and ultimately the decisions that they make are influenced by who they think they are or who they want to become (Akerlof and Kranton 2010). Identity is important for numerous reasons, the most important for our concerns is that it motivates and provides a direction for behavior (Stryker 1968). A person's actions are seen as expressions of their self-identity – people intentionally behave in ways that are consistent with who they think they are. In interaction with others, therefore, people project an identity of who they are, and a primary vehicle for communicating to others who "one is" is through one's behavior.

Identities or selves vary in terms of their temporal orientation. Some selves are oriented toward the present as the *working self* (Markus 1977, 1983). The working self is that component of the self that can be accessed at the moment and is based upon the individual's here-and-now experience. In addition to a sense of who and what one is at the moment, or a self that is fixed on the present, people also have a sense of self that is directed toward the future. This future oriented self is defined positively as the self they would like to become and negatively defined as the self they would not want to become or fear that they might become. Markus and Nurius (1987) have defined this future orientation of the self as a *possible self*. The possible selves "are conceptions of the self in future states" (Markus and Nurius 1987: 157) and consist of goals, aspirations, anxieties, and fears that the individual has as to what he could become. While the working self is aware of what skills the person has and does not have and what the person can and cannot do in the present, the possible self is directed toward the future and what it is possible to be and what the person would not like to be. A person may, for example, see herself currently (the working self) as a thief, drug user, poor father, unskilled worker, but may see herself in the future as working in a job (though perhaps for minimum wage), legitimately buying things for her family, owning a used car, and ceasing to use drugs and commit crime. A person may, however, also fear that she may turn out to be a burned-out addict, riddled with disease, homeless, childless, jobless, and destined to die alone.

An important consequence of a possible self is that it provides directed motivation for one's behavior (Markus and Nurius 1987). Possible selves, both positive and negative, therefore, not only contain satisfying images of what the person would like to be or desperately fears becoming, they can also provide a specific and realistic set of instructions or a "road map" directing what one can do to achieve the positive future self and avoid the negative possible self. This is referred to as the self-regulating component of the possible self. The self is self-regulating because, among other things, it compares the past and

current working self with the possible self and provides specific directions, strategies, or plans for narrowing any discrepancy between the two, thereby connecting the present with the future. Motivation is generated and is more likely to be successful, then, when the person not only has a goal of self-improvement but specific and realistic means to reach that goal. While the positive possible self is frequently a longer-term goal, an initial movement out of a deviant identity is more likely to be based on a motivation to avoid a feared self than it is a desired to achieve a positive self.

Though stable, identities clearly can and do change. A working identity as a criminal offender can change to a more conventional identity when the person thinks of a conventional identity as a positive possible self and an identity of a burned-out ex-con with no friends or possessions as a negative possible self or feared self. Contemplation of a possible self that does not include criminal offending in turn occurs when the working identity of criminal is perceived to be unsatisfying or disappointing. As one begins to find less success and satisfaction with the criminal identity, it is likely to conjure up negative possible selves – long terms in prison with young hoodlums, a violent death during a crime, small payoffs from criminal enterprises. These negative possible selves and the activation of positive selves – a working person, a person with a good spouse, a giving father, a law abider – can provide both the motivation and direction for change. Before one is willing to give up his working identity as a law breaker, then, one must begin to perceive it as unsatisfying, thus weakening one's commitment to it. This weakening of one's commitment to a criminal identity does not come about quickly, nor does it come about in response to one or two failures, but only gradually and only as the result of linking together of many failures and the attribution of those linked failures to one's identity and life as a criminal.

The process of desisting from crime first requires an offender to recognize that their working identity of offender is no longer satisfactory and their attachment to this identity must be weakened. The weakening of a criminal identity comes about gradually and comes about as a result of a growing sense of dissatisfaction with crime and a criminal lifestyle. The dissatisfaction with crime is more likely to lead to a conventional possible self when failures or dissatisfactions with many aspects of one's life are linked together and attributed to the criminal identity itself. It is not just that one has experienced failures but that diverse kinds of failures in one's life become interconnected as part of a coherent whole which leads the person to feel a more general kind of life dissatisfaction, the kind of life dissatisfaction that can lead to intentional identity change.

It is such a new understanding of one's life that leads to the effort to intentionally change it, or as Shover (1996: 132) put it: "[t]his new perspective symbolizes a watershed in their lives. . .[t] hey decide that their earlier identity and behavior are of limited value for constructing the future." The importance of this is that one consequence of this linking together of diverse failures in life, or what Baumeister (1994) calls the crystallization of discontent, is that after this occurs, the dissatisfactions that one has experienced now has implications for the future. Events that seemed atypical and isolated that have been linked are now seen as interrelated and therefore both less easily dismissed and seen as likely to continue to occur in the future. The projection into the future of continued life dissatisfaction leads the person to begin to seek changes.

Kiecolt (1994: 56) has argued that intentional self-change is unlikely to be successful without what she calls "structural supports" for change. These supports "provide individuals with means and opportunities for effecting self change" and include self-help groups and professional changers such as psychiatrists and social workers. As a separate condition for successful self-change, Kiecolt includes the assistance of social supports such as friends, family members, and spouses and partners.

Obviously if successful self-change is going to occur, the benefits of a new identity must outweigh the costs of leaving the old one.

However economically marginal a life of crime is, criminal offenders, particularly those with official records of arrest, conviction, and incarceration, find legitimate employment opportunities, even in the secondary labor market, very restricted. Some opportunity to secure a conventional job must be available for criminal offenders to desist, no matter how strong the motivation to change their identities and selves. Generally, anyone exiting one role or identity needs access to alternative sources of employment – nuns leaving religious orders no less than prostitutes leaving "the trade" must find outside employment. Without these kinds of structural supports, identity change becomes difficult. Social supports, whether in the form of friends, spouses/partners, jobs, or professional help, are important in self-change because they provide the one in the throes of a crystallization of discontent with an alternative existence or identity.

In an identity theory of desistance, changes in friendship networks and the securing of alternative jobs and vocations are important because they help maintain or bolster a fledging changed identity. To be clear, securing jobs, attracting new partners, and involvement with new friends come about *after* a change in identity has occurred. The change in identity has already occurred in the mind of the person; he has weighed the costs and benefits of the exiting identity and alternatives and is behaving in ways that conform to the new possible self.

Empirical Frontiers of Desistance Research: Long-Term Hazard Models

Bushway et al. (2009) show convincingly that the main types of growth curve models largely discard as noise information about change from the individual trajectories. This finding should be particularly troubling for desistance scholars, who are fundamentally interested in studying change. But how can researchers study change if individual trajectories are too imprecise and long-term trajectory models essentially ignore the very change that desistance scholars are interested in studying? Another possibility for examining desistance processes would be to turn to a study of recidivism. Thirty years ago, recidivism and desistance were complementary measures. Those who failed after a certain period were recidivists, and those who did not were desistors.

This static approach to thinking about recidivism and desistance has been effectively rejected. Now, cutting edge recidivism studies focus on hazard rates of offending over time and cutting edge desistance studies focus on measuring trajectories of offending rates over time. But, it is a well-known fact in statistics and quantitative criminology that these two models (hazards and trajectory-type models) are actually measuring the same concept, with hazard rate models focusing on *short-term* change in the propensity to offend and trajectory models focusing on *long-term* change in the propensity to offend. For example, having noted that the hazard rate focuses on the hazard of involvement in a given criminal event, Hagan and Palloni (1988) observe that

> (T)he expected number of criminal events during the age interval being examined is a unique function of these hazards. This expected number of criminal events is what Blumstein et al. are estimating when they calculate lambda (offending rate). So, lambda is a summary of the combined hazards of criminal events of various orders over a period time. (Hagan and Palloni 1988: 97)

As a result, the use of trajectories of rates to study desistance has brought the study of desistance conceptually very close to the study of recidivism.

In their article, Hagan and Palloni (1988) present arguments for focusing on the causal nature of the events, rather than on the rate of offending. At the time they made their argument, however, empirical methods only allowed for the estimation of time-stable rates for individuals. The ability to capture time variation in offending rates while controlling for individual heterogeneity, combined with the new emphasis on the *process* of desistance, provides a persuasive

counterargument for a focus on the more long-term perspective.

The potential productivity of using hazard models with long-term data was highlighted by Barnett et al. (1989), who applied their insight about desistance and trajectories of offending to an analysis of recidivism using a hazard model. Barnett et al. (1989) examined the risk of recidivism until the 30th birthday among a small group of 88 offenders who had at least two convictions before their 25th birthday. Each offender was given a probability of a new offense as well as a desistance parameter that indicated the probability of instantaneously desisting after each event. Thus after each criminal event, the offender had the choice of continuing to offend at the given rate (λ) or desisting. By dividing the offenders into two groups, "frequents" (annual $\mu = 1.14$ or a 1 in 320 daily chance of offending) and "occasionals" (annual $\mu = 0.4$ or a 1 in 913 daily chance of offending), they were able to quite reliably predict future patterns of recidivism. The only complication in their models was a small group of "frequent" offenders who had appeared to desist from crime according to their predictions, but actually resumed a criminal career later in life. It was this small group of offenders they deemed "intermittent" for which their basic models were not adequate. They therefore called for "more elaborate models to incorporate the concept of intermittency, whereby offenders go into remission for several years and then resume their criminal careers" (p. 384).

Kurlychek et al. (2012) have attempted to learn about desistance in the short term by using survival models which can be tied to different models of desistance. Research on survival starts with a group of active offenders and then follows them for a period of time to model the risk of recidivism as well as the time (t) to recidivism. A hazard ratio is then estimated for each time period (t) as follows:

$$H(tg) = \frac{\#\text{ arrested time } t}{\#\text{ survived through time } t-1}$$

Those who have not failed by the end of the follow-up period may be assumed to have desisted from crime. However, it is also possible that they would have recidivated if they had been followed for a longer period of time, meaning that the observation was merely right censored. While much current recidivism research utilize the semi-parametric Cox regression strategy which does not force a functional form on the data over time (e.g., the models are more interested in explaining the effect of covariates over time), Kurlychek et al. (2010) suggest that the use of parametric methods might be more informative if one is attempting to explain the *actual form or time pattern of offending*.

This approach was first introduced to criminology by Maltz (1984) and extended by Schmidt and Witte (1988). For example, Schmidt and Witte (1988) applied a variety of functional forms to two cohorts of releases from the North Carolina prison system and were unsatisfied with the fit of any of the basic models. They identified the problem to be the basic assumption that everybody in the sample will fail if only followed up for a sufficiently long period of time. To address this issue, the authors then turn to what is known as a "split-population" or mixture model which allows for the fact that everyone does not fail. That is, some people do desist.

Split-population models include an extra parameter, often referred to by biostatisticians as the "cure" factor, which estimates the portion of the risk set that will never experience a failure (in other words, they will be "cured"). The cure factor is evidence of instantaneous desistance, or a structural break, particularly for individuals who have substantial rates of offending before the current offense. In this instance, individuals somehow decide (perhaps because they have changed their identity) to quit crime immediately. When applying split-population models to their data, Schmidt and Witte found that all split-population models outperformed their non-split model counterparts. However, Schmidt and Witte (1989) only follow their subjects for 5–7 years, not long enough to fully conclude that there has been desistance.

Kurlychek et al. (2012) estimated similar models using data with 18 years of follow-up from Essex County, NJ. They find that the two-parameter split-population exponential model fits the data almost as well as the more complex three-parameter lognormal counterpart and, in fact, out-performs this model in the later years of the data. It is striking how well this simple model can explain the observed behavior. Like the split-population lognormal model, the split-population exponential model assumes that there are two groups of offenders – those who have desisted at the beginning of the follow-up period and those who remain active. They find support for instantaneous desistance with the split-population lognormal and exponential model actually reaching quite similar conclusions about the size of the permanent desisting population at the outset of the follow-up period (the lognormal model is in the 20–23 % range while the exponential is 25–27 % range). This estimate is smaller than the estimates from Brame et al. (2003) looking at desistance after an arrest. However, it is still substantial. While the focus of most recidivism studies is on the high recidivism rates, the flip side here is that a full quarter of the sample of felony offenders desists after this conviction. Clearly, then, not all individuals are equally risky after a conviction. Indeed, because the exponential model assumes that the active offenders experience a constant risk of recidivism throughout the follow-up period, there is no evidence of declining hazard rates among the active offenders.

The length of the follow-up period in the Essex County dataset has a lot to do with the performance of the split-population exponential model. If the Essex County study had only followed offenders for 3, 4, or 5 years – typical follow-up periods for recidivism studies – the conclusions about the split-population lognormal and exponential models would have been different. Over this shorter window of time, the split-population lognormal model clearly performs better, but viewed over the entire 18-year follow-up period, the simpler, two-parameter split-population exponential model emerges as a formidable competitor. As more datasets with long follow-up periods are studied, it will be interesting to see how well the split-population exponential model performs, especially after the first few years of follow-up.

A final insight revolves around the concept of intermittency or reactivation of criminal careers after a period of dormancy or "temporary desistance" (Barnett et al. 1989). The concept of intermittency has been gaining ground in criminology in recent years and leads to certain theoretical and policy implications (e.g., the idea that desistance is always provisional). The Kurlychek et al. (2012) analysis is certainly consistent with the idea that a low rate offender can go for many years before committing a new offense. But intermittency is a particularly dynamic model of offending in which the offender goes from an active rate of offending to a zero rate of offending back to a fully active criminal career (what Laub and Sampson (2003) refer to as a "zigzag" criminal career). Barnett et al. (1989) moved to an intermittency explanation after they found evidence of a "fat" tail – higher rates of offending more than 5 years after the last offenses than could be explained by the exponential model. While Kurlychek et al. (2012) found support for their simple split-population exponential model, there was no fat tail even though they observed a more serious population over a longer follow-up period. As a result, they concluded that there is no evidence for intermittency, at least as described by Barnett et al. (1989). Replication and extension of these findings with other long-term datasets represents an important avenue for future research. In addition, there is a need to rejuvenate theoretical work in desistance. Previous efforts, while useful starting points, are not able to explain either the criminal patterns of contemporary offenders or the findings that have accumulated from recent empirical studies with different analytical models.

Related Entries

▶ Age-Crime Curve
▶ Cambridge-Somerville Youth Experiment

- Career Criminals and Criminological Theory
- Certainty, Severity, and Their Deterrent Effects
- Crime Specialization, Progression, and Sequencing
- Criminal Careers
- Criminal Careers and Public Policy Responses
- Desistance and Supervision
- Desistance from Gangs
- Examining the Effectiveness of Correctional Interventions
- Focused Deterrence and "Pulling Levers"
- Group-Based Trajectory Models
- Group-Based Trajectory Models and Developmental Change
- Growth Curve Models with Categorical Outcomes
- Identification Issues in Life Course Criminology
- Interactional Theory of Delinquency
- Interrupted Time Series Models
- Offender Change in Treatment
- Offender Decision Making and Behavioral Economics
- Parole and Prisoner Reentry in the United States
- Pathways to Delinquency
- Psychopathy and Offending
- Risk Factors for Prison Recidivism
- Transitional Work Programs for Ex-prisoners

Recommended Reading and References

Agnew R (2004) Why do criminals offend? A general theory of crime and delinquency. Roxbury Publishing, Los Angeles

Akerlof GA, Kranton RE (2010) Identity economics. Princeton University Press, Princeton

Akers RL (1998) Social learning and social structure: a general theory of crime and deviance. Northeastern University Press, Boston

Baumeister RF (1994) Meanings of life. Guilford Press, New York

Baumeister RF (1996) The crystallization of discontent. In: Heatherton TF, Weinberger JL (eds) Can personality change? American Psychological Association, Washington, DC, pp 281–297

Blokland A, Nieuwbeerta P (2005) The effects of life circumstances on longitudinal trajectories of offending. Criminology 43:1203–1240

Blokland A, Nagin D, Nieuwbeerta P (2005) Life span offending trajectories of a Dutch conviction cohort. Criminology 43:919–954

Bushway S, Paternoster R (2012) Understanding desistance: theory testing with formal empirical models. In: MacDonald J (ed) Measuring crime and criminality, advances in criminological theory, vol 17. Transaction Publishers, New Brunswick, pp 299–333

Bushway S, Piquero A, Broidy L, Cauffman E, Mazerolle P (2001) An empirical framework for studying desistance as a process. Criminology 39:491–516

Bushway S, Brame R, Paternoster R (2004) Connecting desistance and recidivism: measuring changes in criminality over the lifespan. In: Maruna S, Immarigeon R (eds) After crime and punishment: pathways to offender reintegration. Willan Publishing, Devon

Bushway SD, Sweeten G, Nieuwbeerta P (2009) Measuring long term individual trajectories of offending using multiple methods. J Quant Criminol 25(3):259–286

Doherty EE (2006) Self-control, social bonds, and desistance: a test of life-course interdependence. Criminology 44:807–834

Fagan J (1989) Cessation of family violence: deterrence and dissuasion. In: Ohlin L, Tonry M (eds) Crime and justice: an annual review of research, vol 11. University of Chicago Press, Chicago, pp 377–425

Farrall S (2005) On the existential aspects of desistance from crime. Symbol Int 28:367–386

Farrington DP, Coid JW, Harnett L, Jolliffe D, Soteriou N, Turner R, West DJ (2006) Criminal careers up to age 50 and life success Up to age 48: new findings from the Cambridge study in delinquency development. Home Office, London

Giordano PC, Cernkovich SA, Rudolph JL (2002) Gender, crime, and desistance: toward a theory of cognitive transformation. Am J Sociol 107:990–1064

Giordano PC, Cernkovich SA, Schroeder RD (2007) Emotions and crime over the life course: a neo-median perspective on criminal continuity and change. Am J Sociol 112:1603–1661

Glueck S, Glueck E (1974) Of delinquency and crime. Charles C. Thomas, Springfield

Gottfredson MR, Hirschi T (1990) A general theory of crime. Stanford University Press, Stanford

Gove WR (1985) The effect of age and gender on deviant behavior: a biopsychosocial perspective. In: Rossi AS (ed) Gender and the life course. Aldine de Gruyter, New York, pp 115–144

Hay C, Forrest W (2008) Self-control theory and the concept of opportunity: making the case for a more systematic union. Criminology 46: 1039–1072

Hirschi T, Gottfredson MR (1983) Age and the explanation of crime. Am J Sociol 89:552–584
Kiecolt KJ (1994) Stress and the decision to change oneself: a theoretical model. Soc Psychol Quart 57:49–63
Kurlychek M, Bushway S, Brame R (2010) Testing theories of desistance and intermittency by studying long-term survival models. Working paper
Laub JH, Sampson RJ (2001) Understanding desistance from crime. In: Tonry M (ed) Crime and justice: a review of research, vol 28. University of Chicago Press, Chicago
Laub JH, Sampson RJ (2003) Shared beginnings, divergent lives: delinquent boys to age 70. Harvard University Press, Cambridge, MA
Le Blanc M, Fréchette M (1989) Male criminal activity from childhood through youth. Springer, New York
Loeber R, LeBlanc M (1990) Toward a developmental criminology. In: Tonry M, Morris N (eds) Crime and justice: a review of research, vol 12. University of Chicago Press, Chicago, pp 375–473
Maltz M (1984) Recidivism. Academic, Orlando
Markus H (1977) Self-schemata and processing information about the self. J Pers Soc Psychol 35:63–78
Markus H (1983) Self-knowledge: an expanded view. J Pers 51:543–565
Markus H, Nurius P (1987) Possible selves: the interface between motivation and the self-concept. In: Yardley K, Honess T (eds) Self and identity: psychological perspectives. Wiley, New York
Maruna S (2001) Making good: how ex-convicts reform and build their lives. American Psychological Association Books, Washington, DC
Nagin DS, Paternoster R (1994) Personal capital and social control: the deterrence implications of a theory of individual differences in offending. Criminology 32:581–606
Osgood W (2005) Making sense of crime and the life course. The Annals 602:196–211
Paternoster R, Bushway S (2009) Desistance and the "feared self": toward an identity theory of criminal desistance. J Crim Law Criminol 99:1103–1155
Piquero AR (2008) Taking stock of developmental trajectories of criminal activity over the lifecourse. In: Lieberman A (ed) The long view of crime: a synthesis of longitudinal research. Springer, New York, pp 23–78
Piquero AR, Blumstein A, Brame R, Haapanen R, Mulvey EP, Nagin DS (2001) Assessing the impact of exposure time and incarceration on longitudinal trajectories of criminal offending. J Adolesc Res 16:54–74
Sampson RJ, Laub JH (1993) Crime in the making: pathways and turning points through life. Harvard University Press, Cambridge, MA
Sampson RJ, Laub JH, Wimer C (2006) Does marriage reduce crime? A counter-factual approach to within-individual causal effects. Criminology 44:465–508
Shover N (1996) Great pretenders: pursuits and careers of persistent thieves. Westview Press, Boulder
Steffensmeier D, Ulmer JT (2005) Confessions of a dying thief: understanding criminal careers and criminal enterprise. Transaction Aldine, New Brunswick
Stryker S (1968) Identity salience and role performance: the relevance of symbolic interaction theory for family research. J Marriage Family 4:558–564
Sweelen, Gary, Alex Pifuerot Caupensel Stemberg (2013) Age and the explanation of crime, revisited J Youth Adolescence 42(6):921–938
Thornberry TP (1987) Toward an interactional theory of delinquency. Criminology 25:863–891
Thornberry TP, Krohn MD (2005) Applying interactional theory in the explanation of continuity and change in antisocial behavior. In: Advances in criminological theory, vol 14. Transaction, New Brunswick, pp 183–210
Wright BRE, Caspi A, Moffitt TE, Paternoster R (2004) Does the perceived risk of punishment deter criminally-prone individuals? rational choice, self-control, and crime. J Res Crime Delinq 41(2):180–213

Desistance from Gangs

Katie Feavel[1] and David Pyrooz[2]
[1]School of Criminology and Criminal Justice, Arizona State University, Huntsville, TX, USA
[2]College of Criminal Justice, Sam Houston State University, Phoenix, AZ, USA

Overview

Gangs present serious challenges to social and criminal justice institutions. The proliferation of gangs, coupled with their high involvement in criminal activity, fuels concern among citizens and policymakers alike. Indeed, gangs account for approximately one of five homicides in large US cities, and their members experience homicide victimization at rates 100 times greater than the general public (Decker and Pyrooz 2010). Further, gangs are a driving force for much of the nation's gun crime as well as other pressing social issues (Howell 2007). The most recent law enforcement figures estimate that there are about 28,100 gangs and 731,000 gang members in the USA, demonstrating the scope of the problem (Egley and Howell 2011). Gangs continue to have a prominent presence in many neighborhoods, schools, communities, and families. As such, it is necessary to further understand both the characteristics and processes surrounding gangs and gang-related activity.

Despite the commonly held perception that gang membership is a lifelong commitment, *most individuals that join gangs also leave gangs*. In fact, involvement in gangs is typically short lived, with the majority of gang members remaining in their gang for 2 years or less (Krohn and Thornberry 2008; Pyrooz et al 2012). Involvement in gangs follows specific patterns: youth join gangs, they persist in gangs and participate in gang activities over some period of time, and then, more often than not, they leave their gang. That said, most research on gang membership has concentrated on either (1) the "ramping up" period of gang membership (i.e., when individuals enter into the gang trajectory) or (2) the persistence period of gang membership, focusing on the (mostly delinquent) activities of gang members. However, gang desistance – the process of exiting the gang – is an equally important aspect of gang membership. Indeed, hastening periods of gang membership will result in reduced rates of criminal offending and serious victimization.

This entry examines the process of *gang desistance*. It begins by answering: what is gang desistance? In doing so, it illustrates and characterizes the concept of gang desistance, drawing from the larger debates and issues experienced in the life-course criminology literature. Next, it discusses the major parameters of gang membership: onset (joining the gang), duration (time spent within the gang), intermittency (the rejoining of gangs), and termination (de-identifying with the gang). It applies key concepts from life-course criminology – trajectories, turning points, and transitions – to the gang context, which helps bring meaning to the gang desistance process. This is followed by examining the primary dimensions of gang leaving, including the *motives* for leaving the gang, the *methods* by which gang members execute leaving, the *abruptness* or rate at which this process transpires, the role of continued social and emotional *ties* that are retained despite having left the gang, and the process of shifting identities. This entry is concluded with a brief sketch of life after the gang, followed by a discussion on the importance of gang desistance research, the implications and benefits of understanding the many facets of gang desistance, and suggestions for future study, based on current gaps in the gang desistance literature.

Fundamentals

Gang desistance refers to the "declining probability of gang membership – the reduction from peak to trivial levels of gang membership" (Pyrooz and Decker 2011, p. 419). The component parts of this definition are derived from the life-course criminological literature (Kazemian 2007; Massoglia 2006), which decomposes desistance into two parts: (1) a reduction in the severity or frequency of participation in criminal activity and (2) an eventual, permanent end, or "true desistance" (Bushway et al. 2001, p. 492). Bushway et al. (2001, p. 500) stated that true desistance occurs when an individual's rate of offending is "indistinguishable from zero" – in other words, when there is no empirical difference from non-offenders.

Gang desistance, however, differs from crime desistance in important ways. First, gang membership is a *state*, while offending is an *event* (Pyrooz et al. 2010). Granted the state of gang membership is comprised of a host of events that display group allegiance, similarly, the confluence criminal events could be conceived as criminal states. Yet, it is clear that gang membership implies at least some degree of connection to a group as opposed to events. To be sure, desistance from gang membership refers to leaving groups, while desistance from crime means disengaging from criminal behaviors. For instance, desisting from the gang does not necessarily mean desisting from crime, nor does the termination of offending indicate a disengagement from the gang. Groups have unique structures, a set of rules, defined roles, and activities; they also have "social cement," a sense of cohesion within that specific group. In this sense, gang desistance is realized as individuals lessen their participation in group activities (e.g., meetings, social events, criminal events) as well as reducing the weight they place on the prominence of the gang and their identification with the gang, which leads to the second key difference between crime and gang desistance.

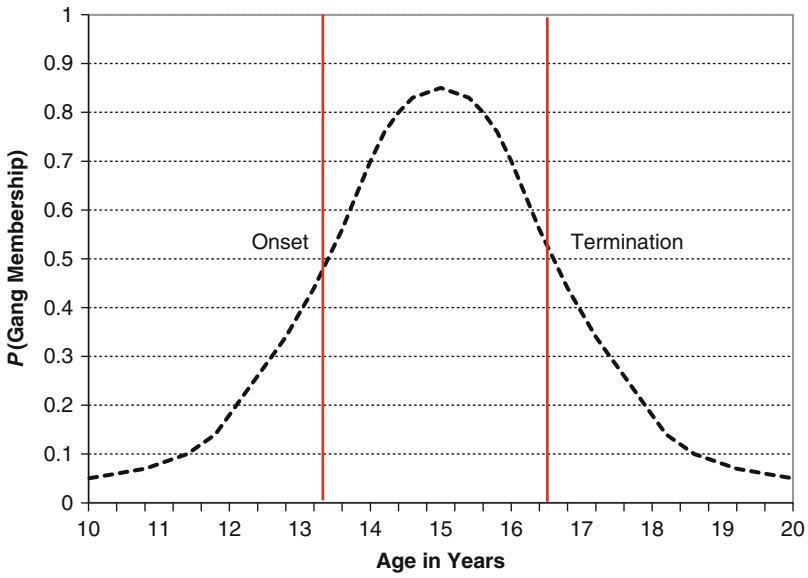
Operationally, gang membership is determined by self-nomination among study participants. In survey research, this typically takes on the following form: "Are you currently in a gang?" This measure has been found to be a "robust measure capable of distinguishing gang from non-gang youth" (Esbensen et al. 2001, p. 124). Those who have left the gang would have reported being a current gang member at some point in time, but no longer self-report gang membership. This is different from crime desistance because it is not necessary for "ex-criminals" to identify as having been a criminal. In other words, gang leaving is cognitive oriented in that it is determined by individual de-identification, while crime leaving is action oriented in that it is determined by individual behavior. This means that an individual can participate regularly in gang-oriented activities – flashing gang signs, recruiting and initiating new members, and selling drugs – but not be recorded as a gang member.

Criminal justice agencies, alternatively, determine gang membership differently than the research literature described above. Personnel in policing and correctional agencies focus on whether an individual meets a certain list of criteria to be entered into official gang databases. Self-reports are one of many measures that can exist for an individual to be entered into a gang database (Barrows and Huff 2009). In addition to self-report, it is common for agencies to consider gang-related tattoos, association with known gang members, flashing gang signs (observed directly or in pictures), the possession of gang paraphernalia, and information from informants. Those that meet two of the criteria are typically considered associates and those that meet three of the criteria are typically considered gang members. The number of necessary indicators and what attributes are included in the criteria vary between states (Barrows and Huff 2009). Leaving the gang, or getting removed from the gang database, is problematic because more emphasis has been placed on including rather than removing individuals in/from gang databases. As such, the regulation of gang databases and removing ex-gang members from such databases is a problem that criminal justice agencies will have to confront (Jacobs 2009). Regardless of the method of determining gang membership, there is movement both into and out of the gang. This movement is detailed in the following sections.

By defining gang desistance as a reduction in the probability of gang membership, it provides a broader view of the life course of gang membership in adolescence and young adulthood. In this sense, gang membership can be thought in terms of a distribution, with age along the x-axis and the probability of being in a gang along the y-axis. This is detailed in Fig. 1 using (somewhat) hypothetical examples of the intercept and slopes – or the points at which onset and termination occur. Gang desistance takes place from the peak level of involvement, the zenith of the curve around age 15, and slowly lowers to trivial level of membership around the late teens.

Couched with this hypothetical curve are two key parameters of gang membership: onset and termination. *Onset* refers to the first self-reported instance of identification with the gang. *Termination* refers to the first self-reported instance of de-identification with the gang. In relation to Fig. 1, these events boost or reduce the probability of gang membership above or below the 50 % threshold that would consider an individual a gang member.

In relation to life-course theory and research (Sampson and Laub 1993), onset and termination of gang membership take on added significance because they act as life-course *transitions*. Transitions are important events dotted throughout the life course that bring meaning to lives; events such as graduating high school, moving away to college, or having a baby are examples of significant events. Joining and leaving a gang are transitions because they are likely to constitute an important event in the life course. Moreover, these events are often formalized with getting "jumped into" or "blessed out" of the gang. Life events known as turning points, though, are key to understanding larger changes in the life course. Laub et al (2006, p. 314) noted: "turning points may modify trajectories in ways that cannot be predicted from earlier events." These events



Desistance from Gangs, Fig. 1 Key gang membership parameters

redirect life trajectories in significant ways. Thornberry et al. (2003) argued that gang membership acts as a turning point in the life course. Melde and Esbensen (2011) recently demonstrated this empirically using a sample of youth from multiple cities in the USA. They found that not only did gang joining have a significant impact on criminal offending, but it also influenced the routine activities of gang joining youth. Further, gang membership had a negative impact on the attitudes, emotions, and social bonds of these youth, which led Melde and Esbensen to conclude that gang joining does indeed act as a turning point. These effects, however, were not reversed once one left the gang, so it remains an open question if gang leaving can put lives back on track.

Duration refers to periods between onset and termination. In other words, duration answers the question: how long do gang members remain in gangs? In relation to life-course research, gang membership can be thought of as a trajectory because it is sustained over a period of time. Sampson and Laub (1993, p. 8) defined a trajectory as a "pathway or line of development over the life span." The gang trajectory is marked by formal periods of onset and termination, as noted in Fig. 1, but there are antecedent and ensuing connections to the gang – the latter is an integral dimension of the gang desistance process – which are denoted in areas below the curve yet above the x-axis.

To examine these time trends in the gang trajectory, it is necessary to have longitudinal data that systematically documents patterns of gang membership. Pyrooz et al (2012) reviewed a series of studies that examined the descriptive characteristics of continuity in gang membership, including the Gang Resistance Education and Training, Pittsburgh Youth Study, Rochester Youth Developmental Study, and Seattle Social Developmental Project (Thornberry et al. 2003). All of the aforementioned studies were carried out for at least 4 years and documented patterns of gang membership in Pittsburgh, Rochester, Seattle, and the 5-site G.R.E.A.T. study. In addition, Pyrooz et al. (2012) documented patterns of gang membership using longitudinal data collected in Philadelphia and Phoenix, based on the Pathways to Desistance study. As a whole, most youth did not remain in gangs for significant periods of time. The majority (48–69 %) remained involved with the gang for 1 year or less. Despite this, there was considerable variability around these figures, with 17–48 % reporting 2 years of membership, 6–27 % reporting 3 years of membership, and 3–5 % reporting 4 or more years of duration within the gang.

Previous research, however, only documented the descriptive characteristics of gang membership duration. Pyrooz et al.'s study modeled this relationship to determine what factors impact continuity and change in gang membership. They found males, minorities (blacks and Hispanics), Phoenix gang members, individuals with poor self-control, and those more deeply embedded within their gang persisted over longer time periods. In fact, that one standard deviation higher in gang embeddedness – which includes for involvement, importance, status, position, and activities in the gang – was associated with at least 1 more year of gang membership. These are important factors to consider as it is well established that gang membership is strongly related to violent offending and victimization.

Intermittency refers to the leaving and subsequent rejoining of a gang. Two studies have examined intermittency – Pyrooz et al. (2012) and Thornberry et al. (2003) – finding that of the individuals that reported gang membership for multiple waves, about 57–66 % were intermittent gang members. In other words, these individuals would have joined a gang at least two times and left the gang at least one time. The extent to which this is an artifact of measuring gang membership or a reality in the streets is unknown. However, it is safe to say that intermittency poses hurdles for understanding the key parameters of gang membership and gang desistance research more broadly. The fact that intermittency exists is a testament to the view of gang membership as a dynamic process.

Kazemian (2007) believes it is reasonable to hypothesize that *all* criminal careers experience some level of intermittency across the life course, making the issue even more conceptually applicable to gang desistance. Intermittent participation in gangs may lead to an illusion of desistance, since ex-gang members may rejoin the gang at some later point in time or have varying levels of participation throughout the life course without entirely leaving. Most available data sets lack the ability – for criminal involvement and gang involvement – to rule out later rejoining of the gang; ultimately, measuring desistance based on non-offending simply does not suffice. Studies that have been done on desistance followed up on individuals within a relatively limited time period, thus possibly generating cases of false desistance that may have led to inaccurate conclusions about the desistance process thus far. However, the length of follow-up required to study *true* desistance is problematic, as studies would need to extend longitudinally throughout the entire life course in order to control for intermittency.

Thus far, gang desistance has been defined and identified by the key parameters (onset, termination, duration, intermittency) of gang membership. This section focuses on gang desistance processes, that is, key factors that characterize moving an individual from current to former gang membership status. It points out that this is a dynamic and evolving process that can be characterized by motives for leaving, methods for leaving, abruptness of the departure, the residual social and emotional ties that persist despite having left the gang, and changes in identities (see Pyrooz and Decker 2011). This section refers to these concepts as the dimensions of gang leaving.

Motives for leaving refer to reasons that influenced a gang member to exit the gang, and this dimension can be thought of as the subjective component in this process. Furthermore, motives for desistance can be organized in terms of push and pull factors. Push motives are factors internal to the gang that make persistence in gang membership undesirable. Typical push motives include "getting tired of the gang lifestyle," "wanting to avoid trouble and violence," and "getting tired of always having to watch my back." The factors inspire one to seek out and select into other social arenas. Pull motives are factors external to the gang that steer or "yank" gang members away from the group. Typical pull motives include girlfriends, jobs, or children as the motivation for exiting the gang. Pyrooz and Decker found that about two out of every three gang members left their gang due to factors internal to the gang, or push factors. In other words, gang members didn't leave the gang because of social interventions or jobs, but because of factors internal to the gang itself.

Methods for leaving refer to modes employed in order to exit the gang. In other words, how one

was able to exit the gang and whether that exit was met with resistance. It is commonly believed that gang departures are synonymous with "blood in, blood out" – shedding blood to enter and exit the gang. In fact, it is often believed that the only way to leave the gang is by getting murdered or murdering one's mother (Decker and Van Winkle 1996). While that might be an extreme example, it is not uncommon to hear that it is necessary to get "jumped out" of the gang. This process involves the "exiter" to have to endure punches, kicks, and other forms of interpersonal violence for a delimited time period. After these ceremonial actions, the individual is then "free to be on their way" having paid their debt to the gang. Another example of a departure method requires the exiter to commit a crime or a mission against a rival gang member. Many of these methods for leaving are similar to the methods for joining the gang (e.g., getting jumped in, going on a "mission").

Pyrooz and Decker (2011) described the above departures as "hostile" in that an individual was forced to engage in some type of behavior to formalize leaving. Based on interviews with over 80 former gang members in Phoenix, they found that most gang members walk away from the gang without any repercussions. In fact, only 20 % of former gang members reported being met with some type of resistance. Importantly, especially for practitioners, is that the methods for departure are often conditioned by the motive for leaving. Pyrooz and Decker found that of the individuals that left the gang due to pull motives (e.g., pregnant girlfriend, job), *none experienced a hostile method of leaving the gang*. As a whole, hostile methods are more of a myth than a reality, which is not uncommon when studying gangs and gang-related behavior (Howell 2007).

Abruptness of departure refers to the rate at which the gang desistance slope declines. In other words, how long does it take to exit the gang after reaching peak levels of involvement? Further, once an individual decides to leave the gang, how long does it take until that process comes full circle? There are two main categories for describing the abruptness of this process: "knifing off" (Laub and Sampson, 1993) and "drifting off" (Decker and Lauritsen 2002). Knifing off is characterized by suddenly leaving the gang, which can be fueled by a significant event, such as the death of a friend as a result of gang violence. Drifting off takes places gradually as a result of shifting ties and social connections to the gang over a longer period of time. In essence, the desistance slope declines a slower rate for "drifters" and at a faster rate for the "knifers." That said, Decker and Lauritsen (2002) found that it is more common for people to drift away from the gang rather than leaving quickly. Much of this is likely due to the natural progression from adolescence to adulthood where adolescents and young adults began to enter into new social arenas, continuing their education, taking jobs, and entering into relationships. These changes slowly disrupt social networks.

Continued ties persist despite having de-identified as being a gang member (Decker and Lauritsen 2002; Pyrooz et al. 2010). These ties to the gang can be both social and emotional; Decker and Lauritsen (1996) identified this as a "gray area" of leaving the gang. For example, some ex-gang members reported that they continued to hang out with the gang, participating in social activities such as drinking, smoking, and other social arrangement. Other ex-gang members reported that if their former gang was disrespected, they would respond to the disrespect. Similarly, if someone from their former gang network was assaulted, they would retaliate across rival gangs. Pyrooz et al. (2010) found that irrespective of how long an individual had left the gang, continued ties to the gang were associated with increased rates of violent victimization. In other words, the well-established pernicious effects of gangs persisted even if someone had renounced their allegiance.

Identity plays a rather obscure role in the desistance process, as it is not an easily observable dimension. However, gang members do use a number of outward symbols to identify themselves with their gang: hand signals, colors, tattoos, etc. The *name* of the gang also plays an important role in the gang member's identity construct (Bjerregaard 2002). Desisting from a gang often means shedding these symbols to

which the individual identifies so closely. Not only do ex-gang members have to go through both cognitive and identity restructuring processes, but they must also navigate through changes in their daily routine activities. As desisters move away from the gang, they establish new social ties in the greater society, and the activities that take up their time tend to shift from gang-related to what would be considered more normative outlets, such as school or sports. Much of the continued tie to the gang pertains to the labels that are attached to gang membership. That is to say, gang membership is in many ways a "master status" that is difficult to shake (Decker and Lauritsen 2002). There are key players involved in assigning that label, including (1) the self, (2) the gang, (3) the neighborhood, (4) the family, and (5) agents of the justice system. Note that throughout the progression from the self to the justice system, there is greater distance in awareness leaving the gang. In other words, while this decision begins with player #1, the remaining players can seriously impact – both positively and negatively – the decision to leave. Even if one has denounced their allegiance to the gang, that decision is formalized by the gang, and one is no longer socially and emotionally tied to the gang; this does not mean that their identity shifts are consistent with the views of the neighborhood or criminal justice agencies.

What about life after the gang? Two seminal contributions to the gang literature – Thrasher (1927) and Decker and Van Winkle (1996) – focus mostly on *life in the gang*, but what about *life after the gang*? Stated another way, what are the life circumstances of gang members 5, 10, or even 20 years after joining a gang? After exiting the gang, do the disruptive activities that characterize gang membership cease and give way to conventional lifestyles putting lives back "on track"? Or, do the disadvantages that accumulate during periods of gang membership have a lasting impact, posing additional difficulties for ex-gang members? Malcolm Klein (1971) noted in his book *Street Gangs and Street Workers* that "[a]lthough the need is great, there has been no truly careful study of gang members as they move on into adult status." Fortunately, in the last 40 years since Klein's critique, several studies have been able to address these issues.

Moore (1978) and Hagedorn (1998) reported on the adult lives of current and former gang members in East Los Angeles and Milwaukee. Both studies followed up on earlier ethnographic research (Hagedorn 1998; Moore 1978) and found that gang members did not fare well in their adult life. Moore found that while some male and female gang members settled down, started families, and pursued conventional employment, this was not the norm. She reported that high rates of early parenthood, unemployment, literacy barriers, and failed relationships made the transition from adolescence to adulthood difficult for many individuals. Hagedorn found that his subjects had dismal high school graduation rates, high rates of unemployment, a reliance on illicit underground markets for income, a dependence on state welfare, and had children at young ages; nearly nine of ten female gang members were mothers in their early twenties. Both researchers relied heavily on Wilson's (1987) hypothesis regarding the impact of deindustrialization and concentrated disadvantage. They argued that their subjects, unlike earlier generations, were shut off from blue-collar job opportunities and thus relied on gang membership, illegitimate labor markets, and entry- and service-level employment for support.

Levitt and Venkatesh (2001) studied gangs operating in Chicago housing projects in 1991. They followed up on the members of their sample, which included gang and non-gang members, in 2000 to examine a host of outcomes and related changes that occurred over that 9-year period. Based on the 2000 interviews, Levitt and Venkatesh examined the effect of gang membership on nine life outcomes, including high school graduation; employment; current incarceration; ever incarceration; annual total, legal, and illegal income if not incarcerated; number of times shot; and housing project resident. The results of their analyses revealed that the negative outcomes tied to gang membership were associated with crime (incarceration, incidence of gunshot victimization, illegal income) rather than other social domains (high school graduate, current employment, public

housing residence). This distinction between involvement in prosocial activities and involvement in crime is important for youth policy in general and gang policy in particular. The results suggest that gang membership has little impact on participation in prosocial activities, but long-term negative effects on involvement in crime and victimization.

Thornberry et al. (2003) and Krohn et al. (2011) followed a sample of about 1,000 at-risk youth attending middle schools in Rochester, NY, until their late twenties and early thirties. They argued that gang members would be less successful in accomplishing normative transitions from adolescence to adulthood, such as graduating high school and attending college, than youth who did not join a gang because of their involvement in "precocious" behaviors and risky activities. From their perspective, gang membership can be viewed as cutting off or limiting possibilities for youth, particularly in the key areas of education and employment. The outcomes these studies examined included high school dropout, teenage parenthood, early nest leaving, adult unemployment, welfare dependence, interpersonal problems in the household, cohabitation, adult offending, and adult arrest. Gang membership influenced all of these outcomes positively, increasing the likelihood of their occurrence. These findings, however, were stronger for male gang members and persistent gang members, compared to their female and transient counterparts.

Finally, Pyrooz (2012) explored the educational, economic, and employment trajectories of youth, focusing on the impact that adolescent gang membership had on these trajectories in early adulthood. He found that joining a gang had an immediate impact on educational attainment, particularly for graduating from high school and matriculating to college. While the differences in high school graduation lessened over time, gang joiners were less likely to attend college and earn a 4-year college degree. In summary, the net effect of gang membership on educational attainment was one-half year. That might sound minimal, but that was the difference between graduating from high school or not, as gang joiners completed 11.5 years of education compared to the 12 years completed by their similarly situated counterparts. Educational attainment had consequences for employment, where Pyrooz observed that gang members were jobless for longer periods throughout the year. While gang members, when working, earned as much and worked as many hours as non-gang individuals, over a 6-year period, they made $14,000 less because of their inconsistent patterns of joblessness.

In summary, this research paints a gloomy picture for the adult lives of adolescent gang members. This implies that the well-established consequences of gang membership are not entirely limited to the short term. That is, gang membership has lasting consequences that (1) extend across the life course and (2) cascade into other significant life domains, such as marriage, family, employment, and education. These consequences could impact not only the rates at which gang members desist from the gang but also complicate the motives and methods for desistance as well as the residual ties that persist after leaving.

Future Directions

This entry has explored a series of issues pertaining to leaving gangs. It has provided both conceptual and operational definitions of gang desistance, identifying key issues in the research literature. It explored the parameters of gang membership – onset, termination, duration, and intermittency – and, in doing so, placed these parameters in the broader context of the life course. The key portion of this entry detailed the multiple dimensions of leaving the gang. These dimensions include the motives for leaving, the methods for leaving, the abruptness of desistance patterns, the continued ties after leaving, and the process of identity reconstruction for ex-gang members. Based on what has been discussed above, there are several key conclusions that may help provide directions for future research.

First, leaving the gang is a dynamic and evolving process, one that takes time to realize fully.

This process involves factors that push and pull the gang member away from the nucleus of the gang – van Gemert and Fleisher (2005) refer to it as the "grip of the group" – helping to shred the social and emotional ties to the gang. However, notification of leaving the gang travels across key social players and paths at different rates: from the individual, to the gang, to the neighborhood, to the family, and finally, to the police. As this information transfers across each of these paths, there is the possibility for resistance and pushes and pulls *back into* the gang. For example, if a gang member on the desistance pathway was arrested in suspicion of committing a crime, he or she would be housed in a county jail facility – both jails and prisons are typically hotbeds for gang activity. This gang member would likely be segregated out of the general population into a pod consisting of non-rival gangs and likely fellow gang members, which could result in strengthened ties and increased embeddedness. Further, past antagonisms between former rival gangs may persist, and victimization may push one back into the gang lifestyle. To be sure, it may take several attempts before a gang member is able to break free from the group processes associated with gangs. Identifying intervention points where gang members' ties to their gang are weak, during violent episodes or after someone has "snitched," may be key to influencing decisions to leave, especially since internal factors appear integral to leaving the gang. Understanding these processes – especially with regard to complicating leaving the gang – should be a central task of gang researchers.

Second, leaving the gang might not be associated with benefits or virtues symmetrical to the impact of joining the gang. That is, while joining a gang can be viewed as a turning point in the life course due to wide-ranging negative effects, leaving the gang does not appear to necessarily return lives to the previous "unblemished" state. In this sense, the gang environment "takes on a history of its own" (Laub and Sampson 1993, p. 320) in that it changes lives in important ways. One might expect that while escaping the expectations of the group should decrease rates of criminal offending and victimization, the deficits accumulated during periods of gang membership may overwhelm any gain achieved from leaving. Nevertheless, given the robust overrepresentation of gang members in self-report and officially recorded rates of offending and victimization (Curry et al. 2002; Thornberry et al. 2003), it is probably no coincidence that the age-crime curve and the age-gang membership curve are tapering off at comparable or parallel rates. Despite the possibility that leaving the gang may not be associated with the drop in crime that is equivalent to gang joining, hastening periods of gang membership should remain a high priority given what is known about the serious consequences of gang membership in individual lives and gangs in communities.

Third, "desistance" is a process not unique to the gang context. The gang literature has drawn heavily from life-course criminology to aid in the development of gang desistance research. Yet, as Ebaugh (1988) noted, role exits contain a great deal of continuity across important life states, including retiring or switching careers to changing genders to leaving the religious convent. In other words, whether one is studying differences between exiting conventional and deviant networks or exiting within deviant network types, there are likely to be universal factors that characterize the exit. To be sure, leaving groups and roles is a global phenomenon. Ebaugh refers to four parts in the exit process, including (1) initial doubts, (2) seeking role alternatives, (3) turning points, and (4) establishing an "ex" identity. Understanding variability in these patterns within and across groups would bring a greater understanding to the difficulties in these processes and, ultimately, assist policies to help or hinder such exits. In the gang context, for example, are there organizational structural characteristics – such cohesion, hierarchies, or collective action (Decker and Pyrooz 2011) – that promote longer periods of gang membership? Or, are there characteristics of the gang that meet attempts to leave with serious resistance? Informal social controls in the group context will likely lend considerable insight into exit and desistance processes.

In the attempt to further develop an understanding of gang desistance, this entry has

summarized and expanded on the current state of the gang desistance literature. In doing so, it has identified key concepts in the process of desistance and, in some cases, comparing these processes to research on crime desistance in the life course. Leaving the gang is a dynamic process that is not only characterized by motives and methods for leaving but also by broader factors that take place long before and after exiting the gang. This process has not received attention from the research community equal to the process of joining a gang, despite the equally important implications. Future research should further develop, both theoretically and empirically, the concepts and findings discussed in this entry.

Related Entries

- ▶ Desistance from Crime
- ▶ Gangs and Social Networks

Recommended Reading and References

Barrows J, Huff CR (2009) Gangs and public policy. Criminol Public Policy 8:675–703

Bjerregaard B (2002) Self-definitions of gang membership and involvement in delinquent activities. Youth Soc 34:31–54

Bushway SD, Piquero AR, Broidy LM, Cauffman E, Mazerolle P (2001) An empirical framework for studying desistance as a process. Criminology 39: 491–516

Curry GD, Decker SH, Egley A (2002) Gang involvement and delinquency in a middle school population. Justice Q 19:275–292

Decker SH, Lauritsen J (2002) Leaving the gang. In: Ronald Huff C (ed) Gangs in America, 3rd edn. Sage, Thousand Oaks

Decker SH, Pyrooz DC (2011) Gangs: another form of organized crime? In: Paoli's L (ed) Oxford handbook of organized crime. Oxford University Press, New York

Decker SH, Van Winkle B (1996) Life in the gang: family, friends, and violence. Cambridge University Press, Cambridge, UK

Ebaugh HRF (1988) Becoming an ex: the process of role exit. University of Chicago Press, Chicago

Egley A Jr, Howell JC (2011) Highlights of the 2009 National Youth Gang Survey. U.S. Department of Justice, Office of Juvenile Justice and Delinquency Prevention, Washington, DC

Esbensen FA, Winfree LT, He N, Taylor TJ (2001) Youth gangs and definitional issues: when is a gang a gang, and why does it matter? Crime Delinq 47:105–130

Hagedorn JM (1998) People and folks: gangs, crime and the underclass in a rustbelt city, 2nd edn. Lakeview Press, Chicago

Howell JC (2007) Menacing or mimicking? Realities of youth gangs. Juv Fam Court J 58(2):39–50

Jacobs JB (2009) Gang databases: context and questions. Criminol Public Policy 8:705–709

Kazemian L (2007) Desistance from crime: theoretical, empirical, methodological, and policy considerations. J Contemp Crim Justice 23:5–27

Klein MW (1971) Street gangs and street workers. Prentice-Hall, Englewood Cliffs

Krohn MD, Thornberry TP (2008) Longitudinal perspectives on adolescent street gangs. In: Liberman AM (ed) The long view of crime: a synthesis of longitudinal research. National Institute Of Justice, Washington, DC

Krohn MD, Scmidt NM, Lizotte AJ, Baldwin JM (2011) The impact of multiple marginality on gang members and delinquent behavior for Hispanic, African American, and White Male Adolescents. J Contemp Crim Justice 27:18–42

Laub JH, Sampson RJ, Sweeten G (2006) Assessing Sampson and Laub's life-course theory of crime. In: Cullen FT, Wright JP, Blevins K (eds) Taking stock: the status of criminological theory, vol 15, Advances in criminological theory. Transaction, New Brunswick

Levitt SD, Venkatesh SA (2001) Growing up in the projects: the economic lives of men who came of age in Chicago public housing. Am Econ Rev 91:79–84

Massoglia M (2006) Desistance or displacement? The changing patterns of offending from adolescence to young adulthood. J Quant Criminol 22:215–239

Melde C, Esbensen FA (2011) Gang membership as a turning point in the life course. Criminology 49:513–552

Moore JW (1978) Homeboys – gangs, drugs, and prison in the barrios of Los Angeles (CA). Temple University Press, Pennsylvania

Pyrooz DC (2012) The non-criminal consequences of gang membership: impacts on education and employment in the life-course. Doctoral dissertation, School of Criminology and Criminal Justice, Arizona State University

Pyrooz DC, Decker SH (2011) Motives and methods for leaving the gang: understanding the process of gang desistance. J Crim Justice 39:417–425

Pyrooz DC, Decker SH, Webb VJ (2010) The ties that bind: desistance from gangs. Crime Delinq. Online First. doi:10.117/0011128710372191

Pyrooz DC, Sweeten G, Piquero AR (2012) Continuity and change in gang membership and gang embeddedness. J Res Crime Delinq. Online First. doi:10.1177/0022427811434830

Sampson RJ, Laub JH (1993) Crime in the making: pathways and turning points through life. Harvard University Press, Cambridge

Thornberry T, Krohn MD, Lizotte AJ, Smith CA, Tobin K (2003) Gangs and delinquency in developmental perspective. Cambridge University Press, Cambridge
Thrasher FM (1927) The gang: a study of 1,313 gangs in Chicago. University of Chicago Press, Chicago
van Gemert F, Fleisher M (2005) In the grip of the group. In: Decker SH, Weerman FM (eds) European street gangs and troublesome youth groups. AltaMira, Lanham, pp 11–30
Wilson WJ (1987) The truly disadvantaged: the inner city, the underclass, and public policy. University of Chicago Press, Chicago

Desistors

▶ Desistance from Crime

Detecting Deception with fMRI

George T. Monteleone
Center for Cognitive and Social Neuroscience, The University of Chicago, Chicago, IL, USA

Synonyms

Detection of deception; Lie detection

Overview

The application of contemporary neuroimaging technologies to the detection of deception has garnered popular attention in recent years. Members of the scientific community have proposed that functional magnetic resonance imaging (fMRI) can be employed as signal detectors to predict behavior and cognitive states. This endeavor is often discussed with a tone of hopeful optimism, but it must be considered with adequate scientific rigor, proper understanding of the limitations of the tool being used, and good social responsibility. In the formative years of the field of criminology in the late nineteenth century, attempts to unify imaging techniques and physiological data for the purpose of human classification yielded questionable results and undesirable social influences. It is necessary not to repeat past mistakes in the excitement over a novel technique.

Imaging, Data, and Criminal Classification

The intersection of imaging, behavioral science, classification, and criminology has origins that date back to the late nineteenth century, when Sir Francis Galton described methods of composite photographic portraiture, a technical innovation at the forefront of imaging technology in its time (Galton 1879). Galton hoped to apply this tool to "elicit the principal criminal types by methods of optical superimposition of [. . .] portraits," in other words, to identify through composite imaging the physical appearance of a typical criminal. These experiments, though embraced with optimism by the scientific community, proved to be a failure. Galton ultimately observed that ". . .the features of the composites are much better looking that those of the components. The special villainous irregularities in the latter have disappeared, and the common humanity that underlies them has prevailed." While Galton's technical developments are impressive in their detail and rigor, his broader perspective on the applications of his techniques was flawed. Nearly a century later, scientist and historian Stephen Jay Gould would appropriately label Galton an "apostle of quantification," suggesting that his tireless measurements of human physical features in the attempt to achieve classification were simply too narrow in scope to serve as reliable models for the natural world (Gould 1981). Companions to Galton's experiments with composite portraiture were the notions of physiognomy and eugenics – ideas that suggested that human characteristics were biologically determined and socially controllable. These hypotheses not only failed to achieve any scientific credence in the twentieth century but also demonstrated the socially dangers political malleability of unsubstantial scientific ideas in later decades through their misappropriation.

The well-known nineteenth-century criminologist Cesare Lombroso was also no stranger to physiognomy and eugenics (Lombroso 1876–1897). A proponent of biological determinism and physiological classification, Lombroso's early scientific hypotheses extended from phrenology, a practice (now widely discredited) that attempted to predict and classify human characteristics through measurements of the shape of the skull (Gray 2004). This putative science is a distinct example of a sound fundamental theory (an existing relationship between the brain, personality, emotion, and behavior) coupled with a thoroughly unsound notion of how it could be measured and applied. Instead of being treated as a scientifically falsifiable hypothesis, phrenology was practiced under the assumption of its validity. Lombroso's foundational text also addressed numerous other external features that could be observed, quantified, and described in relation to the criminal type, such as tattoo markings on the body, linguistic and emotional behaviors, and nuances of handwriting. It is no surprise, given his interest in the indexical tabulation of observable features, that Lombroso dedicated a chapter of the 1884 edition of his text to "Photographs of Born Criminals," proposing that images could be employed to profile a criminal type.

Late nineteenth-century science was no doubt catalyzed by rapidly evolving complementary technological innovations such as advances in photographic techniques and printing that had, to that point, undergone several decades of development and refinement. The period was characterized by great eclecticism and imagination, though these positivist pursuits were not always tempered by healthy social skepticism; Galton and Lombroso were relatively unconfined to particular specializations, and their research practices branched into some peculiar explorations. Indeed, Lombroso's final project (published posthumously) was a positivist inquiry into hypnotism and spiritual phenomena, complete with photographic "evidence" that photographic technology enabled the visualization of ghosts, phantasms, apparitions, and ectoplasms (Violi 2004). Lombroso had been wary of such investigations earlier in his career, changing his position in the early 1890s after taking to attending séances. His reputation as a scientist ultimately overcame any skepticism about his association with the field of paranormal investigation, his authority instead lending credence to the field instead of being debased by it. Though not completely estranged from what was scientifically acceptable at the time, paranormal studies, like physiognomy and eugenics, have remained scientifically unsubstantiated over a century hence (Porter 2003).

Galton and Lombroso's particular attempts to unify physiognomy, photography, statistics, and criminology, though ambitious, were alloyed by bunk assumptions and demonstrably misguided, regardless of any lasting contributions they may have contributed in other fields. Galton, particularly, was a brilliant technical innovator, but failed to recognize that a useful tool is not tantamount to a useful model of human nature. Both eugenics and physiognomy, despite a general enthusiasm about their potential among the scientific communities of their times, have been demonstrated to be at best untenable, and at worse socially dangerous as evidenced by their appropriation in the false sciences of the Nazi party as propaganda to justify monstrous ideologies and acts of mass murder (Gray 2004). Such unsettling potential branches of the ideals of scientific positivism underscore the crucial importance of responsible social vision, good sense, and integrity in any applied science, despite any promises it may appear to offer for the immediate future.

As easy as it may seem to dismiss bad ideas from century-old science as antiquated and irrelevant in the present time, it is important not to fall into similarly narrow interpretations clouded by enthusiasm about current technological prospects. Functional magnetic resonance imaging (fMRI), the most contemporary imaging technology employed at the intersection of physics, physiology, psychology, and neuroscience, offers great promise to future developments in scientific understanding. However, improper application and misunderstanding of this tool undermine its potential; its technical complexity can grant it a false credibility that risks to be passed off to the public under the authoritative moniker of science.

Clarifying a Complex Tool

Commercial firms offering for-profit fMRI lie detection services are optimistic about its promise and technical merits. fMRI has been described in promotional material as a "direct measure of truth verification," an "unbiased method for the detection of deception and other information stored in the brain," a means to investigate "the science behind the truth" and "provide independent, scientific validation that someone is telling the truth." While the marketing decision to shift language from "lie detection" to "the science behind the truth" may be a clever one, it distracts from the important fact that a technology or procedure, no matter how sophisticated, cannot justify a model of nature that is fundamentally inaccurate. MRI is, simply, a tool, like a camera or a microscope, which can offer insights into the workings of the mind and brain when properly applied and reasonably interpreted in conjunction with other forms of inquiry. It is important to distinguish the use of fMRI to detect a signal, as in the case of deception detection, from the practice of scientific research, which attempts to test and refine abstract models of nature by repeatedly testing falsifiable hypotheses; using fMRI for signal detection is, rather, a kind of engineering, an attempt to develop a means to perform a desired function. An elephant in the room in fMRI diagnostics is a simple question: Is the putative signal, in fact, what it is assumed to be? Another elephant, perhaps the next room over, might follow with: Even if so, and even if it can be detected, is the proposed application an appropriate use of this resource? In any case, a clearer understanding of fMRI technology is in order in order to consider its potential, its limitations, and how it ought best (or ought not) be used.

In the interest of demystifying a complex and sophisticated technology, it can be considered in more familiar terms. An MRI brain image is similar in ways to a commonplace photograph. Indeed, some MRI technicians, when communicating to lay participants, refer to MRI image acquisition with the familiar language of "taking pictures." This metaphor is apt; MRI is, much like a camera, a means to record an index of a space within a given field of view (this technical term, field of view, is used in both photography and MRI imaging). In the case of photography, the field of view is determined by optics and perspective; an image of a three-dimensional object is projected onto a two-dimensional plane and recorded through some technical means. The photographic image is an index of a visual space, a record of the phenomenon of light, though an incomplete or slightly distorted version of it due to its optical projection. A two-dimensional photograph can never, despite any level of technical or optical sophistication, fully represent the three-dimensional space from which it originated (Arnheim 1954/1974). Furthermore, as simple as it may seem, it is important to distinguish the constitution of the image from its content: the image is often described as "the reality," but it is, in fact, an image and an index, not the reality itself. This consideration is important in relationship to the claim that an imaging technology self-validates by bringing the viewer "closer to the source." The translation of this problem to fMRI is elementary, but easy to overlook. Technology may afford us the ability to represent ("visualize") what was not possible in the past, but it is important not to confuse a model with what it represents. Scientific models, and even scientific images, will always contain some level abstraction, and in MRI this is absolutely the case. Images analyzed in fMRI studies are parts of a model: a complex, multilayered index of brain activity, and throughout the analysis process, this index becomes ever more abstracted. This abstraction is extremely useful for testing and interpreting results and formulating theories according to the most reasonable interpretations, but this image evidence, by the very structure of the scientific method, is not simplistic, hard, nor immutable. It is, at best, a means by which to formulate theories of what is likely to be happening in a system that is not directly observable.

Some basic misconceptions that surface in the popular discussion of fMRI can be done away with through a better basic understanding of the technology. For one, it does not index

"blood flow," as many descriptions claim (including those published in certain commercial fMRI detection promotional material). Functional MRI signals are indices of shifts in local blood oxygenation levels, which in turn are interpreted as indices of neural activity due to the consumption of energy (via oxygen) during neural activity. The particular relationship between the immediate shift in local blood oxygenation levels and neural activity is not very well understood, but at the very least, the measure should be identified the same way it is by those who study it: the BOLD (blood oxygen level dependent) signal. A more specific model for the brain activity presumed to underlie the BOLD signal is at the cellular level, where neurons (brain cells) consisting of an axon tail extending from the cell body send electrical potentials (propagating voltages) along the axonal membrane. At the terminus of the axon, the signals influence neurotransmitter releases across intercellular space to receptors on the dendrites of other cells (receptors), leading to the buildup and propagation of subsequent electrical pulses by the recipient cell. This activity, compounded on the order of millions in cellular groups (ganglia), requires energy for fundamental physiological processes: some that are typical to cellular operation – such as metabolism – as well as some functions specialized to the neuron, such as the operation of "pumps" that move positive and negative charges across the cell membranes to propagate electrical signals.

The fundamental physics of MRI involve the scanner's sensitivity to differential alignment of atomic nuclei, which are manipulated by pulsing radiofrequency signals through an object in extremely strong magnetic field. Extending the photographic analogy, while a photograph is a record of light (photo: light; graph: drawing) and an MRI image is like a kind of "atomograph," a means of "drawing" an image of the nuclei of atoms in the magnetic field. Indices of blood oxygenation levels collected in fMRI are indices of shifts in the shape of hemoglobin, a macromolecule in the blood that carries oxygen to supply cells with energy. Hemoglobin takes different forms depending on whether or not it is carrying oxygen (it is either "oxygenated" or "deoxygenated"), and the MRI scanner, with its atomic recording properties, can be tuned to detect the differences in the molecule in these respective states. In order to create three-dimensional images, the scanner collects a sequence of two-dimensional slice images in sequence that are subsequently stacked into a three-dimensional volume, wherein the two-dimensional pixels constituting each slice become three-dimensional voxels according to the spacing parameters of the slices. The "functional" term (the "f" in fMRI) involves the incorporation of a time factor – the scanner collects a sequence of brain volume images, which are later reconstructed as a time series of three-dimensional volumes, just as a film is simply a sequence of still photographs. Functional MRI studies typically use two different types of image acquisition. A high-resolution anatomical image is acquired, which is of relatively precise detail, with cubes of around 1 mm per side, but requires several minutes to acquire, like a long exposure photograph. The functional images are acquired rapidly, with an entire brain volume (usually 30–40 slices) collected in 2–3 s. Because of the need for rapid acquisition, the images are much coarser, with the brain volume segmented into cubes around 3–4 mm per side. This trade-off of image quality for time analogous to photography and motion pictures; anyone who has operated digital still and video cameras has probably noticed the difference in image quality between the two formats, with reductions necessary in order to stream video through a limited bandwidth.

Though the BOLD signal is the starting point in fMRI analysis, these signals undergo a complex series of analytical processes and, ultimately, are translated to three-dimensional statistical maps distributed through a model of the brain. Data processing is typically described in two general stages: preprocessing and analysis. Preprocessing is a set of computations that are performed on the data time series to correct for temporal and spatial errors that might have occurred during data collection, improve signal-to-noise ratio in the BOLD signal, and to spatially normalize ("warp") data to a physiological template from which specific neuroanatomical

regions can be estimated. Such steps may include (but not be limited to) slice-time correction to account for offsets in the serial acquisition of the slices comprising each three-dimensional image brain image in each temporal sample (think of it as a "frame" in a three-dimensional "brain movie"); co-registration of each three-dimensional map in the time series (an alignment of each "frame" of the "brain movie"); temporal filtering or linear de-trending of the time series (removal of "drift" artifacts that are generated by the scanner but not a product of any physiological activity); application of a Gaussian spatial smoothing to increase signal-to-noise ratio in regions of the brain presumed to be active at a scale larger than a single voxel; spatial normalization (a computational warping) to a documented anatomical coordinate system such as the Talairach-Tournoux or Montreal Neurological Institute template, resampling each voxel's native dimension to an isometric space of voxels of equal dimension; and finally any normalization of the signal to a standard measurement scale such as percent signal change. After preprocessing is completed, a statistical map of response to experimental conditions is obtained, typically by applying the general linear model (GLM) to estimate parameters given the particular design model of the stimulus timing presentation. The GLM is the application of a parametric model which accounts for the different factors in the study design by modeling a hypothesized response (referred to as the hemodynamic response or impulse response function) which is a set of statistical values determined by the fit between empirical data and the model. This series of fit values (beta values) is then mapped into the voxel space of the brain image, providing a statistical parametric map of the activity in each region. These beta values are then entered into subsequent calculations in order to test various contrasts between conditions. Subsequent statistical tests are carried out by averaging data using regions spatially defined according to a priori criteria, or tests are carried out independently within each voxel in the cortex, the results of which must be statistically controlled for the tens of thousands of repeated measures to obtain a corrected probability of error in the result (Nichols 2012).

As is the case in contemporary science, fMRI research involves the compounding of layers upon layers of theoretical models and can be misleading to characterize any of this as direct measures of reality. Indeed, in science we can hope to identify what is most likely and expect that we might always be wrong. A margin for error is always part of any quantitative model. It is not uncommon for a thorough research group to require several months to design and administer an fMRI study, refine the analyses, and develop a reasonable interpretation of the results prior submitting them to a peer review panel for publication, after which of course they would be open to discussion in the broader scientific community.

The administration of a complete fMRI study is by no means as simple as administering a traditional polygraph test, and even the fundamental task of data collection is wrought with caveats. Slight head movements can cause degradation of the signal or compromise the integrity of the spatial mapping of BOLD signal. It is not unusual for respondents with head motion greater than 8 mm (approximately one-third of an inch) in any direction to be discarded from a study due to loss of data integrity, despite the considerable cost to the researcher, a challenge heightened by the confined space of the magnet bore in which respondents must remain motionless during scanning. Thus, it is important to maximize the relative comfort of respondents in an otherwise uncomfortable circumstance; claustrophobia is a disqualifying factor in the selection of participants. Hearing protection should be worn, as acoustic noise can reach 120 dB(A), a sound pressure level adequate to cause hearing damage. Nevertheless, despite the discomfort of complete immobility in an enclosed space and extreme levels of auditory noise, participants can fall asleep during scanning, yielding useless (but expensive) data. The magnet, typically 3 Tesla in strength, emits a powerful magnetic field. As such, any ferromagnetic material in proximity to the magnet is extremely dangerous, both to the individual being scanned and the administrators of the scan. At risk of serious

injury and/or death, it is impossible to scan any individual with metal implants, microscopic ferromagnetic remnants that might be present in the skin or mucous membranes (a concern for anyone who has spent appreciable time in metalworking or industrial settings), or even large tattoos which can contain trace amounts of iron oxides which can cause serious dermal burns (followers of Lombroso who would wish to use fMRI technology for criminal classification could only hope his suggested relationship between body markings and criminal behavior is specious, as his tattooed criminals would not qualify for interrogation.) Other risks are the generation of body heat caused by electromagnetic radiofrequency excitation and the risk of toxic cryogenic supercooling materials that could be released in the event of a technical malfunction of the magnet. Typically, fMRI magnets are maintained by institutions (such as hospitals) that keep a support staff of physicists, radiologists, and skilled technicians, supporting costs by renting scanning time to researchers and physicians with a considerable maintenance budget. A dedicated in-house fMRI lie detector would be wasteful and impractical, as the best model for such a resource-heavy device is shared use by a broad set of researchers and physicians with adequate budgets to support its administration costs.

Signal and Specificity

The complexity of collecting and analyzing functional MRI data raises a question: Why would such a technology be suited for the detection of deception? While brain imaging may appear to be a direct measure of cognitive activity, it would be better considered as an index of an extremely complex neurophysiological system from which cognition originates. Despite the expense and complexity of the technology and its power for making inferences about physiological activity from studies of groups of individuals, it has yet to be demonstrated as a reliable signal detector at the level of individuals, and its use in individual diagnosis still faces numerous challenges.

Any signal detection model has three possible outcomes: a *Hit*, in which the signal is present (the respondent is lying) and is properly identified (the detector registers a lie); a *Miss*, in which the signal (lie) is present but is not identified (the detector registers as "not-lie"); and a *False Alarm*, in which the signal is not present (the respondent is telling the truth) but is improperly identified (the detector registers a lie). In the extrapolation of such a model to the interrogation of suspects, the imperative is clear that such a signal detector would have to be perfect to avoid wrongful convictions (False Alarms).

Furthermore, a reliable signal detector needs to satisfy four criteria. First, it must be sensitive to the signal that it attempts to detect. Second, it must be specific enough to detect the signal that it purports to detect and should not detect a more abstract corollary of that signal. Third, it must be generalizable across individuals and contexts – it must not be a function of a given circumstance or sample of the population but must work in a variety of scenarios. Fourth, it must be robust enough to resist attempts to intervene on the signal – in the case of lie detection, this could be noise introduced into the signal or attempts by the participant to "deceive" the signal by enacting countermeasures, be they cognitive or physiological.[9]

It is worthwhile to consider some examples from the extant research literature on functional imaging and classification. Numerous studies have endeavored to identify the network of neural activity involved in truth telling and lying, and there is converging evidence that such a network can be identified as the likely set of neural components active during the cognitive state of lying. Research suggests that a network involving the ventrolateral prefrontal cortex, dorsolateral prefrontal cortex, dorsomedial prefrontal cortex, and anterior cingulated cortex is active during intentional deception (Phan et al. 2005). These regions, or a subset of them, have been discussed as possible target regions from which individual lies can be identified (Abe et al. 2007; Mohamed et al. 2006). At this point, it is necessary to clarify the practice of group classification from the individual. Of the voluminous body of functional

MRI research, the vast majority of studies involve group-level studies to identify brain regions active in a given task. The aforementioned results provide evidence that these regions, identified at the group-level, may be active during lying, but do not attempt classification at the individual level. Abe et al. (2007) did report lie classification at a 92 % success rate, but in fact used the traditional physiological polygraph to obtain this result (with a small sample of only 6 subjects), and simply reported aggregate results of the accompanying fMRI study without classification, underscoring the challenge fMRI faces to improve on such results and satisfy its promise of meeting the necessary standards of reliability, generalizability, and robustness.

Individual-level classification has been carried out by several researchers using sophisticated within-subjects paradigms that creatively extend the traditional group-level analysis methods employed in most functional MRI research. Kozel and colleagues reported an individual classification success rate of 90–93 %, and rates between 71 % and 85 % have been reported in other research (Langleben et al. 2005; Monteleone et al. 2009). These results suggest that fMRI is indeed well better than chance but remains imperfect. Furthermore, it is only with the utmost rigor that the network of regions employed in classification be distinguished from its role in other processes; the set of regions used to identify lies are just as likely active in a variety of other social and cognitive processes.

At the individual level, even if fMRI signal detectors are far from perfect, they offer the promise to satisfy the criterion of sensitivity. This, however, could also be said of the traditional polygraph operated by an expert, as classification rates above 90 % have also been reported using this method (Abe et al. 2007). The inability of the traditional polygraph to meet the requirements of generalizability and robustness is clear (Cacioppo 2003), and fMRI methods, despite their sophistication, still face these same challenges. According to the best of our limited understanding of the nature of cortical networks and function, fMRI could just as well prove a failure in terms of specificity. While the aggregate analysis of fMRI data is scientifically sound as a way to consider evidence that could converge on a better understanding of neural function through repeated studies and aggregates of results from research from other levels of modeling neural function (such as cellular-level recordings), the appropriation of specific task-related aggregate brain images to individual classification could risk to be no more justified than Galton's attempts to use aggregate photography to classify criminals. Over a century after Galton's failed imaging classification experiments, a surprisingly similar problem appears: the classifier must consider an aggregate image and make a reasonable conclusion about what it is and what it is not. One must consider that individual variability may not be accurately predictable in every case from an aggregate.

The difficulty in reliably identifying individuals from a putative network suggests that the network is not, in fact, specific to delivering false information, and alternative networks may be active during the presumed lie state. This is sensible, as the construct of "lying" is, in fact, a nonspecific term itself. Though adequate to describe a particular behavior, the colloquial idea of a lie must be reconsidered according to cognitive neuroscience models' neural and mental function. For example, lying could be considered a complex cognitive state composed of processes including (but not limited to) language, memory, and higher-order reasoning. Furthermore, the network of regions described by researchers as the "lie network" is by no means specific; these regions have been demonstrated to be active in a variety of cognitive tasks. For example, a useful counterpoint to any enthusiasm over the specificity of the prefrontal cortex loses momentum when we consider the other possibilities of the complex range of functions the prefrontal cortex plays in neural and mental function, from well-documented cases of its role in emotional processing (Damasio, 1994) to its function even in the "blank-slate" default network of brain function that is active when no cognitive task is at hand (Raichle and Snyder 2007).

The specificity problem is perhaps the most often overlooked amidst the enthusiasm about diagnostic fMRI. Given the complexity of brain

activity, the limited understanding of it, and the index-of-an-index structure of the fMRI model, it would be foolhardy to presume that any signal collected in the brain is, a priori, specific to a particular behavior taking place when the signal was recorded. The theory that cortical function is not fixed, but instead is dynamic, modular, and adaptable, is not a new one. Over a half-century ago, neurologist Karl Lashley dedicated decades of research to the search for "the engram," a cortical trace of memory that could be defined in rodents. After years of carefully controlled experiments failed to detect a locus for a particular memory, even through the methodical excision of very part of the rodent cortex, he concluded that "there is no demonstrable localization of a memory trace" (Lashley 1950). Since Lashley's work, cognitive neuroscientists have embraced distributed processing models of brain activity, which propose that the brain may best be understood as indices of subcomponents of cognitive processes, instead of representing specific behaviors, thoughts, or cognitive states (Rumelhart et al. 1986). While regions participating in a network may be more reliable in their specificity of certain components, it could be reductive, limiting, or simply inaccurate to claim that a given region is tied to a particular module of a cognitive process.

Indeed, cognitive neuroscientists face the daunting challenge of synthesizing models of neural function from a variety of levels of inquiry: molecular, unicellular, multicellular, physiological, electrical, and computational, in the attempt to reconcile some sense of understanding of how a mass of trillions of electrically charged biological units give rise to the phenomena of cognition, emotion, and behavior. While estimates of the number of cells in the human brain are on the order of a hundred billion, the consideration of the interconnectivity of transmitted signals at the subcellular level is much larger, on the order of the hundred trillions. Given that these connections at the subcellular level are hypothesized to code neural information, the spatial precision fMRI signal, on the order of the size of a pea, is considerably coarse in comparison to the microscopic scale of individual neural connections. In light of this, functional MRI is but one imperfect tool in an extremely complex puzzle pursuing converging lines of evidence to best suggest the relationship between the physiology of the brain and the abstract functions of the mind. To reduce this endeavor to the simple detection of deception would be not only reductive but would propagate an oversimplification that would be counterproductive to scientific understanding.

Broader Cognitive and Social Caveats

Given the questionable scientific substance of lie detection, a reliable lie detector remains an idea of fiction. Presuming, however, that such a detector *could* even be devised lends little support for whether it would be prudent to implement it. Lombroso's ideas of biological determinism are reflected in other outmoded ideas from early twentieth-century scientific theories such as behaviorism, which still hold influence on some models of personality and social behavior. There exists a risk to overestimate the "nature" side of a dichotomous nature/nurture model for human behavior. It may be tempting to presume that physiology and the brain should be compartmentalized simply as a phenomenon of nature, but the brain is not a static entity. As a functional organ, a system of complex, time-dependent network signals, the brain appears to be anything but a fixed, deterministic system. Though brain *structures* may be essentially fixed, that does not mean brain *function* cannot be molded and modulated through experience and subject social influences. Furthermore, even consciousness and memory themselves are unreliable, as brilliantly demonstrated in the work of Elizabeth Loftus (Loftus 2003). Even if a sophisticated signal detector were to exist to trace the neural register of conscious memory, what a subject deems to be "the truth" cannot always be trusted.

Returning to the subject of fiction, literature has long warned of the dangerous malleability of "the truth" in a technocracy. This idea is brilliantly described in the dystrophic vision of George Orwell's classic, *Nineteen Eighty-Four*.

The Orwellian "Ministry of Truth" is, of course, an institution for the ironic promulgation of blatant falsehoods. Orwell, who was born at the end of Galton and Lombroso's generation and wrote his most visionary work during the turbulence of World War II (a decade replete with Nazi pseudoscience), was prescient in his consideration of the misappropriation of science and its potential for social injustice. But few authors could be considered more apropos to the particular discussion of technological classification than Philip K. Dick. Dick imagines the horrific social potential of misappropriated technological classification in several works: downloaded brain data are used to subject individuals to life-or-death trials (Dick 1966); electrophysiological prints are employed as classifiers in an oppressive eugenic society where lack of identity results in a lifetime of forced labor (Dick 1974); individuals are incarcerated for "precrimes" predicted by mysterious employees of a deterministic police state (Dick 1956); and a portable machine containing a battery of physiological measures is employed to test for humanity – failure results in classification as a rogue android, resulting in termination (Dick 1968). In any of the imaginary dystopias, an fMRI lie detector would seem, disturbingly, to be a perfect fit.

In his 1953 novella, *The Variable Man*, Dick presents an astute and more optimistic synecdoche of applied science in society: set in deterministic technocracy where wars are waged based on the quantitative predictions of a computer algorithm, the story describes a scenario eerily similar to the institutional application of a signal detection classification system. The titular "variable man," a nineteenth-century tinker who has been transported to the future through an unexpected accident, serves as an unpredictable parameter whose unique circumstance causes him to disrupt the computer's ability to make reliable classifications. Dick proposes that even the most sophisticated device may fail to predict with absolute accuracy the breadth of individual variability, as even a single outlier can undermine a system that is presumed – and required – to function perfectly. This simple yet viable suggestion, so easily overlooked by eager Galtonian and Lombrosian hypotheses of classification, is a crucial consideration in light of the broad variability observed in the human sciences – a variability that ought by all means be reflected in individual differences in brain activity. Dick's accompanying vision of the future, not unlike our own present, is wrought with hyperspecialization that begets a dangerous technological dependency: individuals in different areas of even the same general field are completely unable to understand the details of one another's professional expertise, and consequently, no individual is fully able to comprehend the workings of the very tools upon which they rely to make life-altering decisions. The social application of the thoroughly complex technology of fMRI could easily create a similar dilemma; indeed, even the relatively simple polygraph has been widely misapplied and misinterpreted during its controversial tenure as a putative classification tool (Cacioppo 2004). Dick's variable man possesses a unique savant-like intuition for technological devices amidst the otherwise narrow-minded, hyper-specialized future society in which he finds himself. As such, he is contracted by the government, which is engaged in a prolonged intergalactic conflict, to build a powerful weapon of war. But along with his technical genius, he possesses a simple but important perspective on not just how to innovate technically but how to best deploy his inventions. In lieu of the weapon which he is contracted to build, he creates a device with which he alters the past, circumventing the war altogether. It is such broad thinking that an institution of power, which has the resources and authority to steer the application of our technologies and resources, would do well to adopt.

The demands of functional MRI are extensive, and its reduction to a simple "lie detector" would be a narrow application of its capabilities, further complicated by the risk of misappropriation in light of achieving a particular goal. Furthermore, if fMRI signal detection could be refined to the point of reliable application, it would be more valuable applied as a means to diagnose and treat medical and psychological disorders. Errors in such a diagnostic circumstance, though

hopefully minimized, would at worst yield a null result. An fMRI lie detector, by contrast, through its sheer complexity, expense, and technological impressiveness, is at risk to be easily misinterpreted, oversimplified, and abused as an invalid means of social control.

Related Entries

- ▶ Crime Science
- ▶ Eysenck's Personality Theory
- ▶ False Memories
- ▶ History of the Statistics of Crime and Criminal Justice
- ▶ Identification Technologies in Policing and Proof
- ▶ Marxist Criminology
- ▶ Synthesizing Biological and Social Theorizing

Recommended Reading and References

Abe N, Suzuki M, Mori E, Itoh M, Fujii T (2007) Deceiving others: distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. J Cogn Neurosci 19(2):287–295

Arnheim R (1954/1974) Art and visual perception: a psychology of the creative eye. University of California Press, Berkeley/Los Angeles

Cacioppo JT (2003) Committee to review the scientific evidence on the polygraph. The polygraph and lie detection. National Academy Press, Washington, DC

Damasio A (1995) Descartes' error: emotion, reason, and the human brain. Harper Perennial, New York

Dick PK (1966) The unteleported man/Lies, Inc. Ace Books, New York

Dick PK (1968) Do androids dream of electric sheep? Doubleday, Garden City

Dick PK (1974) Flow my tears, the Policeman said. Doubleday, Garden City

Dick PK (1953) The variable man. Space Science Fiction 2(2)

Dick PK (1956) The minority report. Fantastic Universe

Galton F (1879) Composite portraits. J Anthropol Inst Great Brit Ireland 8:132–144

Gould SJ (1981) The mismeasure of man. W.W. Norton, New York

Gray RT (2004) About face: German physiognomic thought from Lavater to Auschwitz. Wayne State University Press, Detroit

Hakun JG, Ruparel K, Seelig D, Busch E, Loughead JW, Gur RC, Langleben DD (2009) Towards clinical trials of lie detection with fMRI. Soc Neurosci 4(6): 518–527

Kozel FA, Johnson KA, Mu Q, Grenesko EL, Laken SJ, George MS (2005) Detection deception using functional magnetic resonance imaging. Biol Psychol 58:605–613

Langleben DD, Loughead JW, Bilker WB, Ruparel K, Childress AR, Busch SI, Gur RC (2005) Telling truth from lie in individual subjects with fast event-related fMRI. Hum Brain Mapp 26(4):262–272

Lashley KS (1950) In search of the engram. Soc Exp Biol Symp 4:454–482

Loftus E (2003) Make-believe memories. Am Psychol 58:867–873

Lombroso C (1876–1897) Criminal Man. Duke University Press, Durham

Mohamed FB, Faro SH, Gordon NJ, Platek SM, Ahmad H, Williams JM (2006) Brain mapping of deception and truth telling about an ecologically valid situation: functional MR imaging and polygraph investigation-initial experience. Radiology 238(2):679–688

Monteleone GT, Phan KL, Nusbaum HC, Fistgerald D, Hawkley LC, Irick JS, Fienberg SE, Cacioppo JT (2009) Detection of deception using functional magnetic resonance imaging: well above chance, though well below perfection. Social Neuroscience

Nichols TE (2012) Multiple testing corrections, nonparametric methods, and random field theory. Neuroimage 62(2):811–815

Orwell G (1949) Nineteen eighty-four. Secker and Warburg, London

Phan KL, Magalhaes A, Ziemlewiz TJ, Fitzgerald D, Green C, Smith W (2005) Neurl correlates of telling lies: a functional magnetic resonance imaging study at 4 Tesla. Acad Radiol 12(2):164–172

Porter R (2003) "Marginalized practices", The Cambridge history of science: eighteenth-century science, The Cambridge history of science, vol 4. Cambridge University Press, pp. 495–497

Raichle ME, Snyder AZ (2007) A default mode of brain function: a brief history of an evolving idea. Neuroimage 27(4):1083–1090

Rumelhart DE, McClelland JL, the PDP Research Group (1986) Parallel distributed processing: explorations in the microstructure of cognition, vol 1 and 2. MIT Press, Cambridge, MA

Violi A (2004) Lombroso e i Fantasmi Della Scienza. In: Forum Avanzato di Ricerca. Universitaria Multimediale (rivista elettronica)

Detection of Deception

- ▶ Detecting Deception with fMRI

Deterrence

Robert Apel[1] and Daniel S. Nagin[2]
[1]School of Criminal Justice, Rutgers University, Newark, NJ, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA

Overview

The criminal justice system dispenses justice by apprehending, prosecuting, and punishing individuals who break the law. These activities may also prevent crime by three distinct mechanisms – incapacitation, specific deterrence, and general deterrence. Convicted offenders are often punished with imprisonment. Incapacitation refers to the crimes averted by their physical isolation during the period of their incarceration. Specific deterrence and general deterrence involve possible behavioral responses. Specific deterrence refers to the reduction in reoffending that is presumed to follow from the *experience* of actually being punished. However, there are many sound reasons for suspecting that the experience of punishment might instead increase reoffending. The *threat* of punishment might also discourage potential and actual criminals in the general public from committing crime. This effect is known as general deterrence and is the subject of this entry.

Key Concepts of Deterrence

Deterrence is a theory of choice in which would-be offenders balance the benefits and costs of crime. Benefits may be pecuniary in the case of property crime but may also involve intangibles such as defending one's honor, expressing outrage, demonstrating dominance, cementing a reputation, or seeking a thrill. The potential costs of crime are comparably varied. For example, crime can entail personal risk if the victim resists, and it may also invoke pangs of conscience or shame. The theory of deterrence is predicated on the idea that if state-imposed sanction costs are sufficiently severe, criminal activity will be discouraged, at least for some. Thus, one of the key concepts of deterrence is the severity of punishment. In this entry, the review of severity effects focuses on research findings concerning imprisonment.

Severity alone, however, cannot deter. There must also be some possibility that the sanction will be incurred if the crime is committed. For that to happen, the offender must be apprehended, usually by the police. He must next be charged and successfully prosecuted, and finally sentenced by the judiciary. None of these successive stages in processing through the criminal justice system are certain. Thus, another key concept in deterrence theory is the certainty of punishment. In this regard, the most important set of actors are the police – absent detection and apprehension, there is no possibility of conviction or punishment. For this reason, the present entry separately considers what is known about the deterrent effect of police.

One of the key conclusions that emerged from the 1960s- and 1970s-era deterrence literature was that the certainty of punishment was a more powerful deterrent than the severity of punishment. The analyses of this era generally used cross-sectional data on states and involved testing the effects on the statewide crime rate of the certainty and severity of punishment, along with other demographic and socioeconomic control variables. The certainty of punishment was measured by the ratio of prison admissions to the number of reported crimes, while the severity of punishment was measured by median time served of recent prison releases. The basis for the "certainty not severity" deterrence conclusion was that punishment certainty was consistently found to have a negative and significant association with the crime rate, whereas punishment severity generally had no significant association.

This conclusion at the time was probably based on faulty statistical inference. Two primary criticisms were leveled. The first was that the negative association between the certainty measure and crime rate was an artifact of the number of crimes appearing in the denominator of the

certainty measure and the numerator of the crime rate. It can be mathematically demonstrated that errors in the measurement of number of crimes, of which there are many, will force a negative, deterrent-like association between the crime rate and certainty even if, in fact, the certainty of punishment had no deterrent effect on crime. The second involved the use of theoretically indefensible statistical methods for parsing out the cause-effect relationship between sanction levels and the crime rate. After all, sanctions may deter crime, but crime may also affect sanction levels. For example, perhaps overcrowded prisons might reduce the chances of newly caught offenders going to prison. However, subsequent findings from the so-called perceptual deterrence literature and economic studies of the effects of contact with the criminal justice system on access to legal labor markets provide a far firmer empirical and theoretical basis for the "certainty principle." Due to space constraints, this entry will not cover these research traditions.

The Deterrent Effect of Imprisonment

There have been two distinct waves of studies of the deterrent effect of imprisonment. As already noted, studies in the 1960s and 1970s examined the relationship of the crime rate to the certainty of punishment, measured by the ratio of prison admissions to reported crimes, and the severity of punishment as measured by median time served. These studies suffered from a number of serious statistical flaws that are detailed in Blumstein et al. (1978). In response to these deficiencies, a second generation of studies emerged in the 1990s. Unlike the first-generation studies, which primarily involved cross-sectional analyses of states, second-generation studies had a longitudinal component in which data were analyzed not only across states but also over time. Another important difference is that the second-generation studies did not attempt to estimate certainty and severity effects separately. Instead, they examined the relationship between the crime rate and rate of imprisonment as measured by prisoners per capita.

A review by Donohue (2007) identifies six such studies. All find statistically significant negative associations between imprisonment rates and crimes rates, implying a crime-prevention effect of imprisonment. However, the magnitude of the estimate varied widely – from nil for a study that allowed for the possibility of diminishing returns to an elasticity of −0.4. (By an elasticity of −0.4, it is meant that 10 % growth in the imprisonment rate reduced the crime rate by 4 %.) It is important to note that these studies are actually measuring a combination of deterrent and incapacitation effects. Thus, it is impossible to decipher the degree to which crime prevention is occurring because of a behavioral response by the population at large or because of the physical isolation of crime-prone people.

Donohue (2007) goes on to show that the small elasticity estimates imply that the current imprisonment rate is too large, while the high-end estimates imply the rate is too small. He lists a variety of technical shortcomings of these studies that make it impossible to distinguish among the widely varying effect size estimates. The most important is the degree to which the studies were successful in separating cause from effect. While imprisonment prevents crime through a combination of deterrence and incapacitation, crime also generates the prison population. This is an example of what is called the "simultaneity problem," whereby one wants to ascertain the effect of one variable (the imprisonment rate) on another variable (the crime rate) in a circumstance where it is known or suspected that reverse causation is also present, namely, that the crime rate simultaneously affects the imprisonment rate. Thus, statistical isolation of the crime-prevention effect requires properly accounting for the effect of crime on imprisonment. The Levitt (1996) study is arguably the most successful in this regard. It uses court-ordered prison releases as an instrument for untangling the cause-and-effect relationship. However, even the Levitt analysis suffers from many of the technical limitations detailed by Donohue.

More fundamentally, this literature suffers from more than just technical shortcomings that

future research might strive to correct. It also suffers from important conceptual flaws that limit its usefulness in devising crime-control policy. Prison population is not a policy variable; rather, it is an outcome of sanction policies dictating who goes to prison and for how long, namely, the certainty and severity of punishment. In all incentive-based theories of criminal behavior, the deterrence response to sanction threats is posed in terms of the certainty and severity of punishment, not in terms of the imprisonment rate. Therefore, to predict how changes in certainty and severity might affect the crime rate requires knowledge of the relationship of the crime rate to certainty and severity as separate entities, which is not provided by the literature that analyzes the relationship of the crime rate to the imprisonment rate. The studies are also conducted at too global a level. There are good reasons for predicting differences in the crime-reduction effects of different types of sanctions (e.g., mandatory minimums for repeat offenders vs. prison diversion programs for first-time offenders). Obvious sources of heterogeneity in offender response include factors such as prior contact with the criminal justice system, demographic characteristics, and the mechanism by which sanction threats are communicated to their intended audience.

Three studies nicely illustrate heterogeneity in the deterrence response to the threat of imprisonment: The Weisburd et al. (2008) study on the use of imprisonment to enforce fine payment finds a substantial deterrent effect; the Helland and Tabarrok (2007) analysis of the deterrent effect of California's third-strike provision finds only a modest deterrent effect; and the Lee and McCrary (2009) examination of the heightened threat of imprisonment that attends coming under the jurisdiction of the adult courts at the age of majority finds no deterrent effect. These three important studies are considered in more detail below.

Weisburd et al. (2008) report on a randomized field trial of alternative strategies for incentivizing the payment of court-ordered fines. The most salient finding involves the "miracle of the cells," namely, that the imminent threat of incarceration is a powerful incentive for paying delinquent fines. The miracle of the cells provides a valuable vantage point for considering the oft-repeated conclusion from the deterrence literature that the certainty rather the severity of punishment is the more powerful deterrent. Consistent with the "certainty principle," the common feature of treatment conditions involving incarceration was a high certainty of imprisonment for failure to pay the fine. However, the fact that Weisburd and colleagues label the response the "miracle of the cells" and not the "miracle of certainty" is telling. Their choice of label is a reminder that certainty must result in a distasteful consequence, namely, incarceration in this experiment, in order for it to be a deterrent. The consequences need not be draconian, just sufficiently costly to deter proscribed behavior.

Helland and Tabarrok (2007) examine whether California's "Three Strikes and You're Out" law deters offending among individuals previously convicted of strike-eligible offenses. The future offending of individuals convicted of two previous strike offenses was compared with that of individuals who had been convicted of only one strike offense but who, in addition, had been tried for a second-strike offense but were ultimately convicted of a non-strike offense. The study demonstrates that these two groups of individuals were comparable on many characteristics such as age, race, and time in prison. Even so, it finds that arrest rates were about 20 % lower for the group with convictions for two strike offenses. The authors attribute this reduction to the greatly enhanced sentence that would have accompanied conviction for a third-strike offense.

For most crimes, the certainty and severity of punishment increases discontinuously upon reaching the age of majority, when jurisdiction for criminal wrongdoing shifts from the juvenile to the adult court. In an extraordinarily careful analysis of individual-level crime histories from Florida, Lee and McCrary (2009) attempt to identify a discontinuous decline in the hazard of offending at age 18, the age of majority in Florida. Their point estimate of the discontinuous change is negative as predicted, but minute in

magnitude and not even remotely close to achieving statistical significance.

In combination, these three studies nicely illustrate that the deterrent effect of the threat of punishment is context-specific and that debates about whether deterrence works or not are ill posed. Instead, the discussion should be in terms of whether the specific sanction deters or not and if it does, whether the benefits of crime reduction are sufficient to justify the costs of imposing the sanction. To illustrate, while Helland and Tabarrok (2007) conclude that the third-strike effect in California is a deterrent, they also conclude, based on a cost-benefit analysis, that the crime-saving benefits are likely far smaller than the increased costs of incarceration. The Helland and Tabarrok study is an exemplar of the approach that should be taken in evaluating different sanctioning regimes.

The Deterrent Effect of Police

The police may prevent crime through many possible mechanisms. Apprehension of active offenders is a necessary first step for their conviction and punishment. If the sanction involves imprisonment, crime may be prevented by the incapacitation of the apprehended offender. The apprehension of active offenders may also deter would-be criminals by increasing their perception of the risk of apprehension and thereby the certainty of punishment. Many police tactics such as rapid response to calls for service at crime scenes or post-crime investigation are intended not only to capture the offender but to deter others by projecting a tangible threat of apprehension. Police may, however, deter without actually apprehending criminals because their very presence projects a threat of apprehension if a crime were to be committed. Indeed, some of the most compelling evidence of deterrence involve instances where there is complete or near-complete collapse of police presence. In September 1944, German soldiers occupying Denmark arrested the entire Danish police force. According to an account by Andeneas (1974), crime rates rose immediately but not uniformly. The frequency of street crimes like robbery, whose control depends heavily upon visible police presence, rose sharply. By contrast, crimes like fraud were less affected.

The Andenaes anecdote illustrates two important points. First, sanction threats (or the absence thereof) may not uniformly affect all types of crime and more generally all types of people. Second, it draws attention to the difference between absolute and marginal deterrence. Absolute deterrence refers to the difference in the crime rate between the status quo level of sanction threat and a complete (or near) absence of sanction threat. The Andenaes anecdote is a compelling demonstration that the absolute deterrent effect is large. However, from a policy perspective, the important question is whether, on the margin, crime deterrence can be affected by incrementally manipulating sanction threats.

Research on the marginal deterrent effect of police has evolved in two distinct literatures. One has focused on the deterrent effect of the aggregate police presence measured, for example, by the relationship between police per capita and crime rates. The other has focused on the crime-prevention effectiveness of different strategies for deploying police. These two literatures are reviewed separately.

Aggregate Police Presence and Crime

Studies of police hiring and crime rates have been plagued by a number of impediments to causal inference. Among these are cross-jurisdictional differences in the recording of crime, feedback effects from crime rates to police hiring, the confounding of deterrence with incapacitation, and aggregation of police manpower effects across heterogeneous units, among others. Yet the challenge that has received the most attention in empirical applications is the simultaneity problem referred to in the previous section – in the present case, the feedback from crime rates to police hiring.

The two studies of police manpower by Marvell and Moody (1996) and Levitt (1997) are notable for their different identification strategies. The Marvell and Moody (1996) study is based on an analysis of two panel data sets: one composed of 49 states for the years 1968–1993

and the other of 56 large cities for the years 1971–1992. To untangle the simultaneous causation problem, they regress the current crime rate on lags of the crime rate as well as lags of police manpower. If the lagged police measures are jointly significant, they are said to "Granger cause" crime. The strongest evidence for an impact of police hiring on total crime rates comes from the city-level analysis, with an estimated elasticity of −0.3, meaning that 10 % growth in police manpower produces a 3 % decline in the crime rate the following year.

However, regression analyses of this type do not generally provide a valid basis for making causal claims. But other forms of analysis can provide such a basis – one is instrumental variables regression. Levitt (1997) performs an instrumental variables (IV) analysis from a panel of 59 large cities for the years 1970–1992. Reasoning that political incumbents have incentives to devote resources to increasing the size of the police force in anticipation of upcoming elections, he uses election cycles to help untangle the cause-effect relationship between crime rates and police manpower. Levitt's model produces elasticities of about −1.0 for the violent crime rate and −0.3 for the property crime rate. Following Levitt's use of the electoral cycle as an instrument for the number of sworn police officers, other studies have employed alternative instrumental variables and reported comparable elasticities.

In recent years, a number of more targeted tests of the police-crime relationship have appeared. These studies investigate the impact on the crime rate of reductions in police presence and productivity as a result of massive budget cuts or lawsuits following racial profiling scandals. Each of these studies concludes that increases (decreases) in police presence and activity substantially decrease (increase) crime. By way of example, Shi (2009) studies the fallout from an incident in Cincinnati in which a white police officer shot and killed an unarmed African-American suspect. The incident was followed by 3 days of rioting, heavy media attention, the filing of a class action lawsuit, a federal civil rights investigation, and the indictment of the officer in question. These events created an unofficial incentive for officers from the Cincinnati Police Department to curtail their use of arrest for misdemeanor crimes, especially in communities with higher proportional representation of African-Americans out of concern for allegations of racial profiling. Shi demonstrates measurable declines in police productivity in the aftermath of the riot and also documents a substantial increase in criminal activity. The estimated elasticities of crime to policing based on her approach were −0.5 for violent crime and −0.3 for property crime.

The ongoing threat of terrorism has also provided a number of unique opportunities to study the impact of police resource allocation in cities around the world. The study by Klick and Tabarrok (2005) examines the effect on crime of the color-coded alert system devised by the US Department of Homeland Security (in the aftermath of the September 11, 2001, terrorist attack) to denote the terrorism threat level. Its purpose was to signal federal, state, and local law enforcement agencies to occasions when it might be prudent to divert resources to sensitive locations. Klick and Tabarrok (2005) use daily police reports of crime for the period March 2002 to July 2003, during which time the terrorism alert level rose from "elevated" (yellow) to "high" (orange) and back down to "elevated" on four occasions. During high alerts, anecdotal evidence suggested that police presence increased by 50 %. Their estimate of the elasticity of total crime to changes in police presence as the alert level rose and fell was −0.3.

To summarize, aggregate studies of police presence conducted since the mid-1990s consistently find that putting more police officers on the street – either by hiring new officers or by allocating existing officers in ways that put them on the street in larger numbers or for longer periods of time – has a substantial deterrent effect on serious crime. There is also consistency with respect to the size of the effect. Most estimates reveal that a 10 % increase in police presence yields a reduction in total crime in the neighborhood of 3 %, although studies that consider violent crime tend to find reductions ranging

from 5 % to 10 %. Yet these police manpower studies speak only to the number and allocation of police officers and not to what police officers actually do on the street beyond making arrests. The next section proceeds from here by reviewing recent evaluations of deployment strategies used by police departments in order to control crime.

Police Deployment and Crime

Much research has examined the crime-prevention effectiveness of alternative strategies for deploying police resources. This research has largely been conducted by criminologists and sociologists. Among this group of researchers, the preferred research designs are quasi-experiments involving before-and-after studies of the effect of targeted interventions as well as true randomized experiments. The discussion that follows draws heavily upon two excellent reviews of this research by Weisburd and Eck (2004) and Braga (2008). As a preface to this summary, the theoretical link between police deployment and the certainty and severity of punishment is clarified. For the most part, deployment strategies affect the certainty of punishment through its impact on the probability of apprehension. There are, however, notable examples where severity may also be affected.

One way to increase apprehension risk is to mobilize police in a fashion that increases the probability that an offender is arrested after committing a crime. Strong evidence of a deterrent as opposed to an incapacitation effect resulting from the apprehension of criminals is limited. Studies of the effect of rapid response to calls for service (Spelman and Brown 1981) find no evidence of a crime-prevention effect, but this may be because most calls for service occur well after the crime event, with the result that the perpetrator has fled the scene. Thus, it is doubtful that rapid response materially affects apprehension risk. Similarly, because most arrests result from the presence of witnesses or physical evidence, improved investigations are not likely to yield material deterrent effects because, again, apprehension risk is not likely to be affected. A series of randomized experiments were conducted to test the deterrent effect of mandatory arrest for domestic violence. The initial experiment conducted in Minneapolis by Sherman and Berk (1984) found that mandatory arrest was effective in reducing domestic violence reoffending. However, findings from follow-up replication studies (as part of the so-called Spouse Assault Replication Program, or SARP) were inconsistent.

The second source of deterrence from police activities involves averting crime in the first place. In this circumstance, there is no apprehension because there was no offense. This is the primary source of deterrence from the presence of police. If an occupied police car is parked outside a liquor store, a would-be robber of the store will likely be deterred because apprehension is all but certain. Thus, measures of apprehension risk based only on enforcement actions and crimes that actually occur, such as arrest per reported crime, are seriously incomplete because such measures do not capture the apprehension risk that attends criminal opportunities that were not acted upon by potential offenders because the risk was deemed too high.

Two examples of police deployment strategies that have been shown to be effective in averting crime in the first place are "hot spots" policing and problem-oriented policing. Weisburd and Eck (2004) propose a two-dimensional taxonomy of policing strategies. One dimension is "level of focus" and the other is "diversity of focus." Level of focus represents the degree to which police activities are targeted. Targeting can occur in variety of ways, but Weisburd and Eck give special attention to policing strategies that target police resources in small geographic areas (e.g., blocks or specific addresses) that have very high levels of criminal activity, so-called crime "hot spots." Just like in the liquor store example, the rationale for concentrating police in crime hot spots is to create a prohibitively high risk of apprehension and thereby to deter crime at the hot spot in the first place.

Braga's (2008) informative review of hot spots policing summarizes the findings from nine experimental or quasi-experimental evaluations. The studies were conducted in five large US cities and one suburb of Australia. Crime-incident reports and

citizen calls for service were used to evaluate impacts in and around the geographic area of the crime hot spot. The targets of the police actions varied. Some hot spots were generally high-crime locations, whereas others were characterized by specific crime problems like drug trafficking. All but two of the studies found evidence of significant reductions in crime. Further, no evidence was found of material crime displacement to immediately surrounding locations. On the contrary, some studies found evidence of crime reductions, not increases, in the surrounding locations – a "diffusion of crime-control benefits" to non-targeted locales.

The second dimension of the Weisburd and Eck taxonomy is diversity of approaches. This dimension concerns the variety of approaches that police use to impact public safety. Low diversity is associated with reliance on time-honored law enforcement strategies for affecting the threat of apprehension, for example, by dramatically increasing police presence. High diversity involves expanding beyond conventional practice to prevent crime. One example of a high-diversity approach is problem-oriented policing. Problem-oriented policy comes in so many different forms that it is regrettably hard to define.

One of the most visible examples of problem-oriented policing is Boston's Operation Cease Fire (Kennedy et al. 2001). The objective of the collaborative operation was to prevent inter-gang gun violence using two deterrence-based strategies. One was to target enforcement against weapons traffickers who were supplying weapons to Boston's violent youth gangs. The second involved a more innovative use of deterrence. The youth gangs themselves were assembled (and reassembled) to send the message that the response to any instance of serious violence would be "pulling every lever" legally available to punish gang members collectively. This included a salient severity-related dimension – vigorous prosecution for unrelated, nonviolent crime such as drug dealing. Thus, the aim of Operation Cease Fire was to deter violent crime by increasing the certainty and severity of punishment but only in targeted circumstances, namely, if the gang members were perpetrators of a violent crime. Just as important, Operation Cease Fire illustrates the potential for combining elements of both certainty and severity enhancement to generate a targeted deterrent effect. Further evaluations of the efficacy of this strategy should be a high priority.

Conclusions

This entry has reviewed the evidence on the general deterrent effect of sanctions. Evidence of a substantial effect is overwhelming. Just as important is the evidence that the effect is not uniform across different sanctions, jurisdictions, and individuals. Both conclusions are important to devising crime-control policies that make effective use of sanctions to prevent crime. The first conclusion implies that a well-balanced portfolio of strategies and programs to prevent crime must necessarily include deterrence-based policies. However, the second conclusion implies that not all deterrence policies will be effective in reducing crime or, if effective, that the crime-reduction benefits may fall short of the social and economic costs of the sanction.

Future research on sanction effects will be most useful for policy evaluation if it moves closer to a medical model. Medical research is not organized around the theme of whether medical care cures diseases, the analog to the question of whether sanctions prevent crime. Instead, medical researchers address far more specific questions. Is a specific drug or procedure effective in treating a specific disease? Does the drug or procedure have adverse side effects for certain types of people? Furthermore, most such research is comparative – is the specific drug or procedure more effective than the status quo alternative? The analogous questions for deterrence research are whether and in what circumstances are sanction threats effective, and which threats are more effective and in what circumstances.

Related Entries

- ▶ Certainty, Severity, and Their Deterrent Effects
- ▶ Deterrence: Actual Versus Perceived Risk of Punishment

- ▶ Deterrent Effect of Imprisonment
- ▶ Econometrics of Crime
- ▶ Focused Deterrence and "Pulling Levers"
- ▶ Rational Choice, Deterrence, and Crime: Sociological Contributions
- ▶ Situational Action Theory

Recommended Reading and References

Andenaes J (1974) Punishment and deterrence. University of Michigan Press, Ann Arbor

Blumstein A, Cohen J, Nagin D (1978) Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates. National Academies Press, Washington, DC

Braga AA (2008) Police enforcement strategies to prevent crime in hot spot areas. Office of Community Oriented Policing, U.S. Department of Justice, Washington, DC

Donohue J (2007) Assessing the relative benefits of incarceration: the overall change over the previous decades and the benefits on the margin. Working paper. Yale Law School, New Haven

Helland E, Tabarrok A (2007) Does three strike deter? A nonparametric estimation. J Hum Resour 42:309–330

Kennedy DM, Braga AA, Piehl AM, Waring EJ (2001) Reducing gun violence: the Boston gun project's operation ceasefire. U.S. National Institute of Justice, Washington, DC

Klick J, Tabarrok A (2005) Using terror alert levels to estimate the effect of police on crime. J Law Econ 48:267–279

Lee DS, McCrary J (2009) The deterrent effect of prison: dynamic theory and evidence. Unpublished manuscript. Princeton University, Princeton

Levitt SD (1996) The effect of prison population size on crime rates: evidence from prison overcrowding litigation. Q J Econ 111:319–351

Levitt SD (1997) Using electoral cycles in police hiring to estimate the effect of police on crime. Am Econ Rev 87:270–290

Marvell T, Moody C (1996) Specification problems, police levels, and crime rates. Criminology 34: 609–646

Sherman LW, Berk RA (1984) The specific deterrent effects of arrest for domestic assault. Am Sociol Rev 49:261–272

Shi L (2009) The limits of oversight in policing: evidence from the 2001 Cincinnati Riot. J Public Econ 93: 99–113

Spelman W, Brown DK (1981) Calling the police: a replication of the citizen reporting component of the kansas city response time analysis. Police Executive Research Forum, Washington, DC

Weisburd D, Eck J (2004) What can police do to reduce crime, disorder, and fear? Ann Am Acad Pol Soc Sci 593:42–65

Weisburd D, Einat T, Kowalski M (2008) The miracle of the cells: an experimental study of interventions to increase payment of court-ordered financial obligations. Criminol Public Policy 7:9–36

Deterrence of Tax Evasion

Christian Traxler
Department of Economics, University of Marburg, Marburg, Germany
Max Planck Institute for Research on Collective Goods, Bonn, Germany

Overview

Building on the logic of Gary Becker's rational choice model of crime, Allingham and Sandmo (1972) developed the theory of income tax evasion. Their model studies the choice of a taxpayer who rationally trades off the benefits from evading taxes and the risky costs from detection and incurring fines. By increasing auditing frequencies or the magnitude of fines and penalties, tax authorities can deter evasion. This survey discusses three selected aspects of the literature that studies the general deterrence hypothesis in the context of tax evasion. First, the entry will summarize several empirically studies that convincingly solved the key identification problem. The majority of these studies provide evidence that strongly support the deterrence hypothesis. In a second step, the present survey will address the role of taxpayers' imperfect information about the actual enforcement policy. This point, which is closely related to the field of perceptual deterrence research, has recently gained attention in empirical work on deterrence in the context of evasion. The third and last aspect covered in this survey concerns the role of standard deterrence incentives as compared to the tax morale and alternative noneconomic factors that shape tax compliance. The entry concludes with a brief summary and remarks on future directions for research on tax enforcement.

Introduction

Four years after Becker's seminal work on crime and punishment, Michael Allingham and Agnar Sandmo published their analysis of income tax evasion (Allingham and Sandmo 1972). They considered a risk-averse taxpayer who decides how much of his income to report to tax authorities and how much to conceal. Taxes on concealed incomes are evaded. With a certain probability, however, the taxpayer is audited and all the concealed income is detected. In this case, the true taxes plus a fine – which is increasing in the concealed income – have to be paid. Rational taxpayers choose the optimal level of underreported income, trading off the benefits (saving on taxes) with the costs (higher fines) from evasion. Similarly as in Becker's (1968) work, there is scope for general deterrence: by increasing the probability of detection and the magnitude of fines (and other penalties), evasion becomes less attractive. A rational taxpayer will therefore conceal less income. Along these lines, tax authorities can enforce a certain level of compliance with taxes.

The work by Allingham and Sandmo (1972), one of the most cited articles in public economics, triggered a wave of theoretical studies that quickly extend the original analysis. (For detailed surveys of this literature, see Cowell (1990), Andreoni et al. (1998), and Slemrod and Yitzhaki (2002).) The empirical assessment of the model's implications, however, evolved at a slower rate. In particular, the test of the deterrence hypothesis in the context of tax evasion turned out to be quite complicated. First of all, evasion and tax enforcement – i.e., the level of auditing determining the detection risk as well as the size of the fines – are endogenously determined: evasion responds to enforcement, but enforcement policies also react to evasion behavior (exactly as with the police-crime circularity). By now, only a handful studies managed to break this endogeneity. Section 2 (Causal Evidence on Deterrence) discusses several of these contributions. Except for one study, these contributions provide evidence suggesting that deterrence works.

A second issue in empirical studies on deterrence in the context of tax evasion emerges from concerns about the limited knowledge of taxpayers about the actual enforcement policies. While the imperfect information of decision makers about the likelihood and magnitude of sanctions is of general interest for the research on deterrence (Bebchuk and Kaplow 1992), this point appears particularly relevant for the case of tax enforcement. In contrast to visible indicators for enforcement such as police on the streets, there are few cues that taxpayers can use to assess tax enforcement policies. In addition, tax enforcement agencies do not reveal details about their auditing practice and often – at least in many continental European countries –operate quietly in the background. Section 3 (Tax Enforcement and Perceptual Deterrence) discusses a very recent strand of literature on general deterrence in the context of evasion that explicitly addresses this point. This literature, which closely related to the field of perceptual deterrence research (see the survey in Nagin 1998), points to the importance of the channels through which information on detection and punishment is spread.

A further complication in the empirical analysis of deterrence emerges in the operationalization of the theoretical concept of "the detection risk." The problem is best illustrated with the group of (not self-employed) labor income receivers. In modern economies, the evasion rate in this subpopulation is almost zero (Slemrod 2007). At the same time, the effective auditing frequency in this group is close to zero, too. The observation of this evasion and enforcement pattern has motivated many researchers to question the validity of the model of rational taxpayers: given the low auditing rate, the standard model would predict more evasion. To explain this puzzle, the importance of noneconomic motives for compliance – often subsumed among the label "tax morale" – has been put forward (Torgler 2007). Section 4 (Tax Morale and Third-Party Information) will briefly summarize this literature. The section will also highlight that the modern literature classifies the zero-evasion, zero-auditing observation quite differently. In practice, an individual's auditing risk is not a fixed parameter (as assumed in the simple benchmark model), but rather a function of her

evasion choice: the more income she conceals (given a certain level of true income), the higher is her auditing risk. Observing an effective auditing frequency that is close to zero corresponds to just one point of this auditing function (see, e.g., Kleven et al. 2011). Hence, full compliance can be a rational response to standard incentives from deterrence if small levels of underreporting would be detected with a high probability.

Causal Evidence on Deterrence

The identification of a deterrent effect from auditing and the severity of sanctions is complicated by several problems. First of all, it is obviously difficult to measure evasion (Slemrod and Weber 2012). Secondly, enforcement strategies are no exogenous: it is not only evasion, which should respond to enforcement; the choices of enforcement authorities are also determined by the (past, present, or expected future) level of evasion.

Early studies on the determinants of tax evasion were mainly concerned with the first issue (e.g., Clotfelter 1983; Feinstein 1991). Based on random auditing data from the US tax authority's Taxpayer Compliance Measurement Program (TCMP), these contributions derived individual level measures for evasion. In some studies, this dependent variable was regressed on regional measures of auditing frequencies or income type specific proxies (e.g., Dubin et al. 1990; Beron et al. 1992). In doing so, however, they did not cope with the second issue, the endogeneity of enforcement. Cross-sectional and intertemporal variation in regional auditing rates, for instance, might reflect variation in the unobserved determinants of the taxpayers' noncompliance that also drive evasion.

A more recent wave of contributions that studies the determinants of evasion behavior is based on randomized field experiments. These experiments provided truly exogenous variation in enforcement policies and therefore contributed to what Slemrod and Weber (2012) called the "credibility revolution in the empirical analysis of tax evasion." One of the first studies is based on the Minnesota Income Tax Compliance Experiment (Slemrod et al. 2001). In cooperation with a team of researchers, the Internal Revenue Service (IRS, the US tax authority) approached a randomly selected set of taxpayers from a predefined group of income earners with different types of letters. Next to a baseline letter that served as a control treatment, one treatment included a letter with the information that the tax return filed by the taxpayer was selected to be "closely examined." The comparison of the first-differences in reported incomes provided mixed results.

In line with the deterrence hypothesis, low- and middle-income taxpayers responded to the audit treatment with an increase in reported income. This effect was considerably stronger for those with more opportunities to evade income (e.g., self-employed income). For high-income taxpayers, however, the audit threat backfired: relative to the control group, it induced *lower* reported tax liabilities. One possible interpretation of this observation is that high-income taxpayers who know that they will be audited with certainty behaved strategically. By entering the bargaining-like interaction with the IRS with a lower reported income, these taxpayers might expect to end up with lower tax payments.

More clear evidence on deterrence is offered by a recent field experiment in Denmark. In cooperation with the Danish tax authorities, Kleven et al. (2011) gained access to a stratified and representative sample of more than 40,000 income tax filers. Half of the individuals were randomly assigned to similar audit threat treatments as in Slemrod et al. (2001), communicating either a 100 % or a 50 % chance of an audit. Relative to a control treatment that did not receive any letter, the audit threats triggered a significant increase in reported incomes, with a larger effect under the 100 % auditing risk (Kleven et al. 2011).

In contrast to the other contributions, the study by Fellner et al. (2013) does not rely on a random draw from a certain population but rather a selected sample of 50,000 potential evaders. They experimentally study the effect of various

mailing types on the evasion of TV license fees. Relative to a neutral mailing, an audit threat treatment that also stressed the severity of sanctions had a significantly positive impact on compliance. The evidence in Fellner et al. (2013) therefore shows that deterrence also works for the most important group – those that self-selected into evasion.

All these contributions focused on independent individual or household level choices of evasion. The field experiment by Pomeranz (2011) documents an interesting spillover effect from deterrence in the context of joint, interdependent evasion decisions. Based on a large-scale field experiment with more than 400,000 small Chilean firms, she finds that audit threat letters not only reduces the evasion of Value Added Taxes (VAT) of the targeted firms (i.e., the firm that receives the letter), but also for untreated suppliers, who are also involved in the evasion decision. The logic behind this observation is straightforward: the potential detection of the VAT evasion of the treated firm increases the detection risk of the untreated firm. Hence, in the context of interdependent evasion decisions – as it is the case with VAT evasion – the deterrence effect from tax enforcement is multiplied.

Beyond this strand of research that experimentally varied the threat from an audit, there is still little evidence on the specific impact of actual audits. A rare piece of causal evidence on specific deterrence in the domain of tax enforcement is again found in the Danish study from above (for an earlier studies, see Erard 1992). In addition to and independently of the letter treatments, Kleven et al. (2011) randomly selected a group of taxpayers that was actually audited. Experiencing an audit turned out to have a substantial positive effect on the level of reported incomes in the years following the audit. Quantitatively, this effect was much stronger than the impact of the threat letters (Kleven et al. 2011). While the mechanisms driving this observation are not yet identified, the evidence suggests that actual audits have a much stronger and more long-lasting effect than simple audit threats (on the longevity of deterrence effects, see also Fellner et al. 2013).

Tax Enforcement and Perceptual Deterrence

As Becker's model of crime, the rational model of evasion assumes that decision makers know the precise detection probability set by the enforcement authority. Perceptual deterrence research provides some evidence that "there presumably is a positive relationship between actual and perceived levels of enforcement"; however, it is considered "implausible that individuals' probability estimates are generally accurate, particularly when the probability is extremely low" (Bebchuk and Kaplow 1992, p. 366). While this observation is of general interest for research on deterrence, it appears particularly relevant for the case of tax enforcement. In contrast to visible cues such as the number of policemen on the streets, there is hardly any publicly accessible information that allows potential tax evaders to accurately predict their true detection risk. This raises the question, how changes in the enforcement policy translate into changes in compliance behavior.

The question was addressed in the field experiment in Fellner et al. (2013). The researchers confronted subjects with the different letter treatments, in particular, letters with and without an audit threat (see above). Subjects were then asked to indicate their perceptions about the detection risk and potential fines, etc. The results from the perception survey neatly complemented the evidence on the behavioral responses observed in the field experiment: the audit threat considerably increased the perceived detection risk. The evidence suggests that the threats shaped perceptions and that individuals rationally adjust their compliance behavior to these perceptions.

While Fellner et al. (2013) provide evidence on the impact of a targeted policy on individual perceptions and compliance behavior, the impact of "hidden" policies – i.e., generally unobservable enforcement activities – in a society remains unclear. A theoretical analysis of this issue is provided by Sah (1991), who models agents that update their perceived detection risk based on information obtained from neighbors and acquaintances. A higher number of detections among this

sample ceteris paribus result in an increased risk perception. In turn, the inclination to comply with the law increases. Implicitly, the model therefore highlights the crucial role of communication for the dispersion of information on enforcement activities.

A recent study that addressing behavioral responses to interpersonal communication in the context of tax enforcement is Alm et al. (2009), who study tax compliance in a laboratory experiment. They implemented a design in which there was no "official" information on auditing risks available; however, they allowed the experimental subjects to communicate. Alm et al. find that income reporting is sensitive to information on auditing frequencies that is spread via communication. More closely related to the model in Sah (1991) is the contribution by Rincke and Traxler (2011), who provide field evidence on a substantial deterrent effect that is mediated by word of mouth. They analyze the enforcement activities of field inspectors that approach potential evaders of TV license fees (compare Fellner et al. 2013) at their homes. These field inspectors unregularly visit different neighborhoods in different months. The presence of inspectors in an area is not announced; inspectors are not uniformed nor do they use any police-like cars. Hence, it is basically impossible to observe if inspectors are visiting an area.

Making use of monthly panel data on the activities of field inspectors in more than 1,000 small municipalities, Rincke and Traxler (2011) identify a substantial impact of these "hidden" enforcement activities. Following an instrumental variable approach, they find that three additional detected evaders induce one further evader (that was not approached by field inspectors) to switch to compliance. Making use of microlevel data, they provide further evidence that strongly suggests that the effect is driven by communication about detections among neighbors. In addition, they show that the enforcement spillover is concentrated in the close spatial proximity to the detected household.

The findings in Alm et al. (2009) and Rincke and Traxler (2011) provide two important insights. First and most importantly, the fact that enforcement activities are often "hidden" and that auditing rules are typically not announced by tax authorities does *not* imply that an increase in enforcement activities fails to deter evasion. Communication on experiences with enforcement authorities seems to be sufficient to convey the information about an increased detection risk. This implies – and that is the second point to take away – that patterns of communication and information dispersion become decisive for general deterrence.

Tax Morale and Third-Party Information

Third-Party Information

The rational model of evasion predicts that a sufficiently high detection probability should deter evasion. The comparison of evasion levels for different sources of incomes (Slemrod 2007), however, seems to provide a picture that first appears at odds with this prediction. In modern economies, the evasion of taxes on nonbusiness labor income is nearly zero. At the same time, the auditing rate of labor income receivers is close to zero, too. As noted earlier, this pattern has motivated many researchers to question the rational model of evasion: it has been argued that the standard model would predict "too much evasion." As an explanation for this "puzzle," noneconomic motives for compliance, often subsumed among the label tax morale, have been put forward (e.g., Torgler 2007). The modern taxation literature, however, provides a different explanation to the zero-evasion, zero-auditing observation: the presence of third-party income reporting, which makes many taxpayers unable to evade.

In its most simple version, the standard model of evasion captures the risk of detection in one single parameter, the auditing probability p. In practice, however, an audit does not necessarily result in the detection of all evasion. In addition, and more importantly, an individual's auditing risk is not given by a constant parameter, but rather a function of her underreporting and her true income. It appears plausible to assume that, for a given level of true income and a given

income source, an individual's risk of detection increases with a higher level of underreporting (several extensions of the basic model follow this approach; see the surveys in Andreoni et al. 1998; Cowell 1990). If one thinks about the observation of zero-auditing and zero-evasion for labor incomes along the lines of this more realistic model, one arrives at a straightforward explanation.

In economies with a modern tax enforcement system, labor income is typically "third-party reported." This means that not only the labor income receiving taxpayer declares her incomes to the tax authority, but also the firm who pays these incomes (and often withholds taxes) reports the payments. Tax authorities can therefore easily compare the taxpayer's declared income with the income that is reported by her firm. For a taxpayer who exclusively receives labor income, this implies that she is basically unable to conceal income (see, e.g., Kleven et al. 2011). While there is a zero-auditing risk under full compliance (which is observed on average), any noncompliance will get detected basically with certainty. Given this nonlinear detection risk, full compliance is a rational response to standard economic incentives – and perfectly in line with a more realistic extension of the baseline model of rational evaders.

This point appears trivial. In fact, it was well understood among policy-oriented researchers for a long time (see, e.g., Long and Swingen 1990). For the theoretically motivated research, however, it took more time until it was acknowledged that the model by Allingham and Sandmo (1972) is a simplified model of a taxpayer that fully self-reports her, e.g., self-employed income, rather than a model that describes the evasion of third-party reported income. (For a theoretical analysis of the case where employer and employee jointly evade taxes on third-party reported incomes, see Kleven et al. 2009.) In line with the baseline model, fully self-reported incomes – in contrast to third-party reported incomes – are actually concealed to a significant amount (Slemrod 2007).

The availability of third-party information provides a convincing explanation for the high compliance and the low auditing risk that is observed for labor incomes. Hence, the argument that this pattern can *only* be explained by tax morale or social norms for tax compliance is not justified. This does not mean, however, that tax morale – or any other behavioral incentive beyond those captured in the standard model – is irrelevant for tax compliance. It is therefore interesting to consider possible noneconomic incentives and their relevance for tax enforcement. The remainder of this section therefore discusses the role of tax morale. Models of bounded rational taxpayers, probability weighting, and loss aversion are beyond the scope of this survey (see, e.g., Dhami and al-Nowaihi 2007, and references therein).

Tax Morale

The modern literature uses the term tax morale quite broadly, typically subsuming the idea of a social norm for tax compliance, perceptions about civic duties as "honest taxpayer," but also the notion of stigmatization (for a survey, see Torgler 2007). Theoretical approaches typically model tax morale in form of an additional, noneconomic cost of evasion (an overview of different modeling strategies is provided by Traxler 2010). In these frameworks, concealing income triggers economic consequences (in case of detection) as well as psychic (e.g., shame) or social cost (e.g., from social exclusion) that are motivated by tax morale. These costs might differ between different individuals – some agents might not be affected by tax morale, others might even perceive a moral obligation to evade taxes. In addition, it is typically assumed that the noneconomic costs of evasion depend on the compliance behavior of "relevant others" (Traxler 2010). In a society, where everybody complies with taxes, the noneconomic costs of evasion might be large. Hence, individuals experience strong incentives to comply with taxes. However, if evasion would be widespread in this society, the same individuals might perceive hardly any noneconomic costs of cheating and might evade taxes. A tax morale that depends on the behavior of others therefore has an important implication: it gives scope for a multiplicity of equilibria (or, more generally, social interaction in tax compliance; see Galbiati and Zanella 2012). For a given level of tax enforcement, a society might coordinate on a state where

evasion is widespread or a state where there is a strong norm to comply with taxes (and, potentially, further equilibria between these extremes; see Traxler 2010).

These ideas on tax morale have several important implications for the effectiveness of deterrence. The first two are mainly of theoretical interest: changing the level of tax enforcement might eliminate the existence of one (out of several) possible equilibrium. Hence, minor changes in the enforcement policy could, in principle, produce large shifts in compliance behavior. Related to this point, one can think more generally about the interaction of the formal, legal and informal, private enforcement of compliance: if stricter tax enforcement is an expression of the societies' weight attached to compliance with this rule (an idea from expressive law), it might also crowd-in stronger psychic or social cost from tax evasion. Similar as with the private versus public provision of public goods, however, there might also be scope for crowding-out (Kube and Traxler 2011).

A third implication relates to the notion of the "relevant others." If the most relevant others (in terms of shaping perceptions about tax morale) are, for instance, politicians or high-profile members of the business world, the models highlight potential benefits from focusing enforcement policies on these groups: Enforcing a high level of compliance in a narrow group might contribute to a high tax morale among a broader population (Traxler 2010). Hence, there might be a positive spillover from deterrence. One should note, however, that stricter tax enforcement among these specific groups could also backfire. This is the case, since media reporting on detected evaders from moral reference groups could erode the tax morale in the general population.

A further interesting implication for deterrence is contained in models that consider tax morale and stigmatization (e.g., Kim 2003). These studies assume that stigma-related costs are only incurred in case of detection. Given that taxpayers want to avoid stigmatization and any further negative social consequence from being known a detected evader (e.g., ostracism, social exclusion), the fear of stigmatization will increase the overall level of (economic and non-economic) punishment and therefore renders an increase in the auditing frequency more effective. Relating the idea of stigmatization to the previous point, one might argue that the costs from stigmatization crucially depend on how many other taxpayers are detected evading. If stigmatization is more severe in a society where evasion is a rare phenomenon, this means that the deterrent effect from an increase in auditing is larger as compared to a society where evasion is widespread.

The discussion from above shows that tax morale has nontrivial implications for tax enforcement and deterrence. It is therefore worth asking, how important tax morale really is. Unfortunately, the empirical literature provides little causal evidence that allows assessing the role of tax morale. Beyond a huge number of correlational studies, there is only one very recent contribution that documents a causal impact of (survey measures of) tax morale on actual evasion behavior (Halla 2012). In addition, evidence from field experiments that tested approaches which rely on tax morale suggests that it is difficult to effectively apply this type of alternative enforcement strategies. The Minnesota Income Tax Compliance Experiment discussed above, tested – next to an audit threat letter – also different moral appeal letters. Relative to a control treatment, these moral letters turned out to be ineffective (Blumenthal et al. 2001). Similar evidence is provided in the field experiments of Fellner et al. (2013) and Pomeranz (2011). However, Fellner et al. (2013) found some evidence that an information letter, that emphasized the high level of compliance in the population, had a positive effect on compliance (relative to a neutral treatment which did not include this information).

Concluding Discussion and Future Research

This entry provided an overview of different strands of research on deterrence within the domain of tax evasion. Section 2 (Causal Evidence on Deterrence) summarized field studies that

randomly varied the auditing risk that is communicated to potential evaders. While some of these studies point to the particularities that distinguish tax evasion from the deterrence of noneconomic crimes (e.g., the evidence on responses from high-income group in Slemrod et al. 2001), the overall picture provides clear support for the deterrence hypothesis: a higher auditing risk reduces evasion. These randomized studies made a major contribution to the empirical analysis of deterrence in the domain of tax evasion, as they established credible, causal evidence on a conceptually important issue (Slemrod and Weber 2012). Nevertheless, there is still a lot left for future research.

First of all, there is little causal evidence on the effectiveness of actual audits. Existing results suggest that there is a substantial effect of specific deterrence (Kleven et al. 2011). Due to the lack of data, however, the robustness and intertemporal persistence of this observation is still unclear. To analyze these points and to better understand the impact of audits, more studies with exogenous variation in auditing are needed.

Second, the empirical literature offers only little evidence on the impact of different sanction levels. Moreover, there is basically no field evidence on the optimal mix of the two policy tools for deterrence – the detection risk and the magnitude of the punishment. One theoretical solution – an enforcement system with infinitely high sanctions and an infinitesimally small auditing risk ("hang taxpayers with zero probability," as suggested in an early contribution), which is cheap in terms of auditing resources – will most likely not be tested empirically. However, one can easily imagine a randomized study that varies both dimensions of the enforcement system – the detection risk and, e.g., the size of fines. Such an experiment would make a huge contribution to the important policy debate on the composition of an optimal enforcement system as well as to the theoretical literature on the trade-off between certainty and severity of sanctions.

Section 3 (Tax Enforcement and Perceptual Deterrence) of this entry discussed a recent literature which links perceptual deterrence research with the empirical analysis of tax enforcement. The first contributions in this domain suggest that there is a clear link between policies and perceptions, as well as perceptions and behavior (Fellner et al. 2013). In addition, the crucial role of communication about own experiences with auditing and detection is highlighted (Rincke and Traxler 2011). In the context of imperfect perceptions about the objective auditing risk, word of mouth about detections crucially shapes the effectiveness of enforcement strategies.

Given the difficulty to obtain precise microdata, this young literature has mainly focused on communication within neighborhoods and spatial distances. It would be important for future research to bring this analysis to further layers of social interaction and to future measures of "proximity" (such as friends in a social network). This would allow assessing, for instance, the role of communication at the workplace or among business partners for spreading information about auditing policies and auditing technologies ("what can easily be detected").

Naturally, this research agenda seems to be linked with the growing empirical literature on networks. Given that communication between taxpayers is crucial for their compliance behavior, there emerges an interesting research question: how should an optimal auditing policy allocate resources between different taxpayers within a (neighborhood, social, workplace) network? Who is the key evader to target? Empirical studies that address these questions would provide important insights for the design of optimal enforcement strategies and would make major contributions to the applied work on networks.

Section 4 (Tax Morale and Third-Party Information) addressed the role of third-party information reporting and tax morale for tax compliance. It is now widely acknowledged that the benchmark model of evasion is tailored to the case of fully self-reported incomes and that the availability of third-party information explains the observed auditing and evasion pattern for different occupational groups and income sources – in particular, self-employed and business versus nonbusiness, labor incomes. A straightforward policy implication is that an extension of the application of third-party information reporting – or more generally: policies that give tax authorities broader access to

matchable data – will render a tax enforcement system more efficient. In order to increase compliance, investing in clever information systems therefore appears to be an alternative to investments in stricter auditing.

Finally, several possible implications of tax morale for deterrence were discussed. Field experiments that tested moral appeals or other approaches to enforcement that are rooted in the notion of tax morale are generally found to be ineffective in turning evaders into compliant taxpayers. It would be misleading, however, to interpret this observation as an indicator for tax morale being irrelevant. Based on the available evidence, the question about the importance of tax morale for actual behavior remains unanswered. Again, this is an important issue to address in future research.

Related Entries

- ▶ Econometrics of Crime
- ▶ Certainty, Severity, and Their Deterrent Effects
- ▶ Deterrence: Actual Versus Perceived Risk of Punishment
- ▶ Rational Choice, Deterrence, and Crime: Sociological Contributions

Recommended Reading and References

Allingham M, Sandmo A (1972) Income tax evasion: a theoretical analysis. J Public Econ 1:323–338

Alm J, Jackson B, McKee M (2009) Getting the word out: enforcement information dissemination and compliance behavior. J Public Econ 93:392–402

Andreoni J, Erard B, Feinstein J (1998) Tax compliance. J Econ Lit 36:818–860

Bebchuk L, Kaplow L (1992) Optimal sanctions when individuals are imperfectly informed about the probability of apprehension. J Legal Stud 21:365–370

Becker G (1968) Crime and punishment: an economic analysis. J Polit Econ 78:169–217

Beron KJ, Tauchen HV, Witte AD (1992) The effect of audits and socioeconomic variables on compliance. In: Slemrod J (ed) Why people pay taxes. University of Michigan Press, Ann Arbor, pp 67–89

Blumenthal M, Christian C, Slemrod J (2001) Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota. Natl Tax J 54:125–138

Clotfelter C (1983) Tax evasion and tax rates: an analysis of individual returns. Rev Econ Stat 65(3):363–373

Cowell FA (1990) Cheating the government – the economics of evasion. MIT Press, Cambridge, MA

Dhami S, al-Nowaihi A (2007) Why do people pay taxes: expected utility versus prospect theory. J Econ Behav Organ 64:171–192

Dubin J, Graetz M, Wilde L (1990) The effect of audit rates on the federal individual income tax 1977–1986. Natl Tax J 43(4):395–409

Erard B (1992) The influence of tax audits on reporting behavior. In: Slemrod J (ed) Why people pay taxes. University of Michigan Press, Ann Arbor

Feinstein J (1991) An econometric analysis of income tax evasion and its detection. Rand J Econ 22(1):14–35

Fellner G, Sausgruber R, Traxler C (2013) Testing enforcement strategies in the field: threat, moral appeal and social information. J Eur Econ Assoc 11(3): 634–660

Galbiati R, Zanella G (2012) The tax evasion social multiplier: evidence from Italy. J Public Econ 96:484–495

Halla M (2012) Tax morale and compliance behavior: first evidence on a causal link, B.E. J Econ Anal Policy 12(1)

Kim Y (2003) Income distribution and equilibrium multiplicity in a stigma-based model of tax evasion. J Public Econ 87:1591–1616

Kleven H, Claus K, Emanuel S (2009) Why can modern governments tax so much? An agency model of firms as fiscal intermediaries. NBER Working Paper 15218

Kleven H, Knudsen M, Kreiner CT, Pedersen S, Saez E (2011) Unwilling or unable to cheat? evidence from a tax audit experiment in Denmark. Econometrica 79:651–692

Kube S, Traxler C (2011) The interaction of legal and social norm enforcement. J Public Econ Theory 13:639–660

Long S, Judyth S (1990) Third-party information reporting and compliance. In: Research conference report, how do we affect taxpayer behavior? Internal revenue service document 7302, Washington, DC, pp 3–91

Nagin D (1998) Criminal deterrence research at the outset of the 21st century. Crime Justice 23:1–42

Pomeranz D (2011) No taxation without information: deterrence and self-enforcement in the value added tax. Working Paper, Harvard Business School

Rincke J, Traxler C (2011) Enforcement spillovers. Review of economics and statistics 93:1224–1234

Sah R (1991) Social osmosis and patterns of crime. J Polit Econ 99(6):1272–1295

Slemrod J (2007) Cheating ourselves: the economics of tax evasion. J Econ Perspect 21:25–48

Slemrod J, Shlomo Y (2002) Tax avoidance, evasion, and administration. In: Auerbach AJ, Feldstein M (eds) Handbook of public economics, vol 3, Chapter 22, pp 1423–1470, Elsevier: Amsterdam

Slemrod J, Weber C (2012) Evidence of the invisible: toward a credibility revolution in the empirical analysis of tax evasion and the informal economy. Int Tax Public Finance 19(1):25–53
Slemrod J, Blumenthal M, Christian C (2001) Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota. J Public Econ 79(3):455–483
Torgler B (2007) Tax compliance and tax morale: a theoretical and empirical analysis. Edward Elgar Publishing, Cheltenham
Traxler C (2010) Social norms and conditional cooperative taxpayers. Eur J Polit Econ 26:89–103

Deterrence: Actual Versus Perceived Risk of Punishment

Gary Kleck
College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL, USA

Overview

Extant evidence indicates that individual perceptions of the certainty, severity, and swiftness of punishment have essentially no correlation with actual levels of those measures of risk prevailing in the area in which the individuals reside. This suggests that public policies that are designed to reduce crime by increasing the deterrent effect of punishment are unlikely to succeed because they are not likely, in general, to increase prospective offenders' perception of the legal risks of committing crime. This does not mean that there are no deterrent effects of the threat of punishment, but only that variations in objective levels of punishment may not affect the magnitude of any deterrent effects that do exist.

Key Issues

There is now an enormous body of scholarly research on the question of whether higher levels of punishment for crime reduce rates of criminal behavior. While there are other mechanisms by which greater punishment levels could reduce crime, such as the incapacitative effects of incarceration or moral reformation, perhaps the most thoroughly studied mechanism is that of deterrence, in which persons otherwise inclined to commit crimes refrain from doing so because they fear the prospect of legal punishment. A person might be deterred from crime as a result of their own personal experiences with punishment, an effect commonly labeled "special deterrence" because it is specific to persons who have been punished. Or a person might refrain from crime as a result of their awareness of others being punished, and thus of the possibility that they might likewise be punished, an effect called "general deterrence," because it can affect anyone in the general population, not just those punished (Zimring and Hawkins 1973, pp. 72–73). For this reason, general deterrent effects are potentially far more potent sources of social control than incapacitative effects or special deterrent effects.

Research on the general deterrent effect of punishment on criminal behavior has largely fallen within two broad categories: (1) macro-level research using official crime statistics to assess the links between objective levels of punishment, such as the ratio of arrests to offenses or the average length of prison sentences, to crime rates, and (2) individual-level "perceptual" research using survey methods to assess the links between perceptions of punishment (e.g., its certainty or severity) and self-reported criminal behavior (Paternoster 1987; Nagin 1998). Many early researchers testing deterrence propositions using macro-level data assumed that any negative associations observed between punishment levels and crime rates were due to deterrent effects. That is, the researchers commonly assumed, but did not demonstrate, links between actual punishment levels and perceptions of punishment.

The second variety of research typically used individual-level survey research to assess the effect of perceptions of punishment on (self-reported) criminal behavior. It addressed the scientific question of whether the former affects the latter, but did not directly address the policy issue that commonly lurks behind the

scientific debate: do higher levels of punishment imposed by the criminal justice system produce lower rates of criminal behavior? It is possible that, even if the individual-level research indicates that there is an effect of punishment perceptions on criminal behavior, higher levels of punishment still may not increase the deterrence of criminal behavior because punitive policy efforts fail to intensify perceptions of risk in the first place (Jacob 1979, p. 584; Nagin 1998). Complicating things still further, higher punishment levels may reduce crime rates, but not by increasing general deterrence. For example, crime may be reduced through incapacitation effects, i.e. the physically restraining effects of criminals being incarcerated, or a host of other mechanisms (see Andenaes 1952 for a classic listing). Incapacitation effects could well explain most of the past macro-level findings of an apparent crime-reducing effect of certainty of arrest, conviction, or imprisonment on crime rates.

There is a modest literature on the formation of sanction risk perceptions which has generally focused on the effects of (1) an individual's prior experience of sanctions, such as arrests or incarceration experiences, and of (2) the individual's prior criminal behavior on perceptions of risk. In general, researchers have concluded that the more crimes a person has committed, the more crimes they are likely to have committed without suffering punishment, which reduces perceived risks – a phenomenon described by some as an "experiential effect" (Paternoster et al. 1982). Conversely, others have claimed that people will adjust their estimates of the certainty of punishment upward when they personally experience sanctions such as arrests (Horney and Marshall 1992), though most research does not support this assertion (Pogarsky and Piquero 2003).

Unfortunately, there is little empirical evidence on exactly how individuals go about gathering information on punishment risks. Their perceptions of sanctions risks could be based on the experience of their criminal friends, as Cook (1980) speculated, or on the news media, or even on fictional sources like television and movies. Kleck et al. (2005) found that the frequency with which people watched television news programs was unrelated to their perceptions of risk of punishment, but it may be the specific content rather than frequency of exposure to news that shapes perceptions. Or perhaps there is more of a media impact among more criminally inclined people, for whom news about punishment is likely to be more salient. And if national media are the primary influences, this could explain why there is only a moderate amount of variation in perceptions of legal risk across different areas – everyone is influenced by the same national media messages, rather than varying local realities. Likewise, politicians might influence public perceptions of legal risks when they run on law-and-order themes. Incumbents might boast of the high punishment levels they have helped create, while challengers might reduce the public's perceived risk levels by criticizing incumbents for their supposedly lenient policies.

Unlike incapacitative effects, which occur regardless of how criminals perceive their punishment, deterrence can occur only to the extent that the risk of punishment is perceived by prospective offenders. The dimensions of punishment that influence its deterrent impact on criminal behavior have been summarized in the proposition that "the greater the certainty, severity, and swiftness (celerity) of punishment, the lower the crime rate will be" (e.g., Gibbs 1975, p. 5). However, a more precise restatement of this proposition, which stresses the essential role of perceptions, would be: "The greater the *perceived* certainty, severity, and swiftness of punishment, the lower the crime rate will be."

Of course, the abstract possibility of punishment for crime is perceived by virtually everyone. Some may believe that these risks are low, or that they are themselves unlikely to be caught and punished for any crimes they might commit, but almost everyone recognizes at least the theoretical possibility that they might suffer legal punishment if they violated the law. Nevertheless, this does not imply that increases in actual punishment levels (i.e., increases in actual certainty, severity, or swiftness of legal punishment) will increase deterrent effects and thereby reduce crime, since variations in actual punishment levels may not cause variations in the average

perceived level of punishment among prospective offenders. Indeed, critics of deterrence as crime control have long pointed to this reality-perception gap as a key weakness in deterrence-based crime control policies (Zimring and Hawkins 1973, p. 45).

The answer to the simple yes/no question "Can punishment deter crime?" is almost certainly "Yes," since at least a few people almost certainly refrain from some crimes due to a fear of legal punishment. But this is an irrelevant question from a policy standpoint since no policymakers are asking the simple yes/no policy question, "Shall we punish crime?" The more relevant policy question is: "Shall we have, as a means to crime reduction, *more* punishment than we have now?" It is not obvious that more punishment produces a stronger deterrent effect. because it would be uncertain whether increased punishment levels would cause increases in the perceived risk of punishment.

This way of framing the issue is pertinent because much of the debate over crime control in recent decades has been confined to variations on the theme of increasing legal punishments, and some is even more narrowly confined to strategies for increasing the severity and certainty of punishment. Thus, legislators debate bills that would mandate minimum sentences for certain crimes, increased penalties for repeat offenders, enhanced penalties for crimes committed with guns or in connection with drug trafficking, and many other strategies for increasing the severity of punishment. Law enforcement officials lobby for increased budgets and enforcement authority so that they can arrest more criminals, thereby increasing the certainty of arrest for crime. Prosecutors make similar appeals for resources that would enable them to increase the certainty of conviction. And many advocates argue for building more prisons so that both the probability and severity (length) of prison sentences can be increased.

These policy proposals are justified at least partly on the grounds that they will reduce crime through increased deterrence, by "sending a message," "getting the word out" that crime will not tolerated, that criminals will be "taught a lesson" that crime will not be "coddled," and that punishment will surely follow crime. Thus, it is asserted that policy changes producing increases in punishment levels will reduce crime by means of deterrence, that is, by means of an increased perception of legal risk among prospective offenders.

Deterrence theorists have routinely hypothesized about variations in the strength of deterrent effects across individuals, suggesting that various attributes of humans moderate the impact of punishment threats on criminal behavior. For example, legal threats are thought to have stronger effects on persons with greater stakes in conformity and stronger social bonds to conventional others who would ostracize them as a result of legal sanctions being imposed (Zimring and Hawkins 1973, pp. 96–128; Paternoster 1987; Nagin 1998). One might extend this reasoning to the link between actual and perceived punishment levels – perhaps the degree to which the former affects the latter is likewise moderated by some of the same variables. But even though the effect of actual punishment levels on perceived levels might vary across individuals, all theories of general deterrence nevertheless assume that the average effect is a significant positive one.

No matter how inclined and able people may be to rationally process and weigh information, and to consider potential costs and benefits of various courses of action, they cannot actually decide and act rationally unless there is at least some accuracy to their perceptions of those costs and benefits and thus some correspondence between reality and their perceptions of reality. In some realms of human activity, it is perfectly reasonable to assume a fairly close correspondence between perceptions and the realities of costs and benefits. In the sphere of economic behavior, narrowly construed, the assumption is particularly plausible, mainly because there is such a large volume of relevant information available to actors, and a relatively high degree of accuracy to that information. Consumers generally know exactly the price of different brands of goods, while investors know not only the price of a share of stock in any given firm, but also a great deal about the assets, liabilities, and past

profit performance of that company. Relevant information in the sphere of market behavior is comparatively voluminous, accurate, and easily obtained. Rational behavior, and predictable responses to changes in costs and benefits, are to be expected in such information-rich environments. Shaped by research experiences in this context, it is not surprising that some economists appear to consider it self-evident that there must be at minimum a significant positive, albeit imperfect, correlation between actual risks and perceived risks (Becker 1968; Ehrlich 1973).

Criminal behavior is quite different, especially with regard to one of the main risks associated with it, punishment. Information about legal risks is limited, often inaccurate, and hard to obtain, so the correlation of actual risks and perceptions of those risks is considerably weaker than in the realm of market behavior. If prospective offenders' perceptions of punishment risk bore no systematic relationship to punishment reality, variations in that reality would have no effect on deterrence of criminal behavior. People might well be deterred by the possibility of punishment, but they would be no more likely to be deterred in settings where risks were higher than in places where they were lower. Under such circumstances, investment in policies increasing punishment levels would be wasted, at least from a deterrence standpoint.

While a positive association between actual and perceived levels of punishment might seem patently obvious to some, there is good reason to question the linkage, and a fair amount of research on related topics that casts doubt on the assumption that the link is strong. Few people, whether criminals or noncriminals, are consumers of criminal justice statistics, and even criminals have only limited personal experience with crime and punishment. Further, depending on hearsay and gossip among their criminal associates may not be a reliable basis for forming even approximately accurate notions of levels or trends in CJS punishment activities.

Similarly, the news media provide neither criminals nor noncriminals with much reliable information on levels of either crime or punishment. At the macro-level, the amount of news coverage of legal *punishment* is unlikely to bear a very strong relation to the general level of actual punishment, since studies of the relationship between the volume of news coverage of *crime* and actual rates of crime find the relationship to be close to nonexistent (Garafalo 1981; Marsh 1989). If common punishment events such as court sentencings or admissions to prison receive even less publicity than the crimes that gave rise to them, it is unlikely that people could formulate even minimally accurate perceptions of punishment risks from news media coverage.

Indeed, various documented news media biases in coverage of crime and punishment could cause, in an irregular fashion, either overstated or understated perceptions of punishment risk. For example, some scholars have found that newspapers exaggerate the certainty of arrest by over reporting solved crimes (Roshier 1973, p. 37; Parker and Grasmick 1979, p. 371). Conversely, studies reviewed by Roberts (1992) indicate that news stories about suspects who "got off on a technicality" or who got a "slap on the wrist" sentence from a judge leads to the public perceiving the severity of sentences to be lower than it really is. Further, since estimates of the certainty of punishment necessarily reflect perceptions of the volume of criminal acts relative to punishments, the lack of correspondence between the volume of news media coverage of crime and actual crime rates may further weaken reality-perception correlations regarding the certainty of punishment.

Personal Experience of Punishment and Perceived Future Risk

On the other hand, already active criminals might draw on their own experiences and those of close associates to formulate their perceptions of punishment risk. To the extent that they accurately stored away these experiences when they occurred, and then accurately retrieved the information later, these experiences could improve both perceptions of past legal risks and forecasts of future risks. The research on personality traits common among known offenders, however,

does not encourage a view of criminals as disciplined and careful processors of information, likely to systematically recall and assess such past experiences. Indeed, within a population of persons who already evince tendencies towards risk-taking, past experience of punishment could even lead to a variant of the gambler's fallacy: "My string of past bad luck in getting caught is due to end; my chances of avoiding arrest are bound to improve because I've exhausted my share of bad luck." And of course the information situation is far worse for noncriminals, the people that deterrence-based policies are supposed to keep law-abiding. Noncriminals have no personal experience of criminal behavior leading to either punishment or no punishment, and thus no individual information base at all to use in formulating perceptions of legal risk.

Does the Experience of Punishment Increase Perceived Risk of Punishment? – The Resetting Effect (The Gambler's Fallacy)

The special deterrence hypothesis asserts that the experience of being personally caught and punished for crime should increase the offender's fear of future punishment, which in turn reduces criminal behavior. After all, it seems obvious that being caught doing crime should reinforce the idea that you are likely to be caught again in future if you continue committing crimes. At minimum, punishment experience should increase the punished person's perceived certainty of punishment. Perhaps if people were thoroughly rational, this would be true. If they are not, this proposition is not at all obvious, and a good deal of empirical evidence indicates it is wrong.

A series of careful studies indicates that the experience of being caught causes many offenders to conclude that it is unlikely they would be unlucky enough to get caught the next time they try crime, almost as if they had "used up" their bad luck. Like many gamblers, they irrationally believe that a larger number of unlucky experiences somehow improve one's odds of good outcomes in the future. Consequently, the experience of being caught doing a crime leads them to *reduce* their estimates of the probability of being caught in future (Paternoster and Piquero 1995; Piquero and Paternoster 1998; Piquero and Pogarsky 2002; Pogarsky and Piquero 2003; Pogarsky et al. 2005; Matsueda et al. 2006). More recently, Kleck and Barnes (2013), using survey data from a nationally representative sample of US adults, found that persons who had previously been arrested reported significantly *lower* levels of perceived arrest risk than those who had never been arrested (though these authors could not control for criminal behavior). Other researchers generally find no effect of arrest on perceived certainty of punishment (Piliavin et al. 1986; Pogarsky et al. 2004), or obtain very mixed findings (Wood (2007) found no evidence of an effect of punishment experience on perceived certainty, but did find a positive association of punishment experience with perceived severity). And a few other studies found that punishment experiences do appear to increase perceived certainty, as expected (Paternoster et al. 1985; Horney and Marshall 1992; Heckert and Gondorf 2000; Lochner 2007).

In sum, the full body of evidence does not indicate that experiencing a punishment will increase one's perceived risk of future punishment. The effects of punishment on perceived risk are highly variable, and do not, across the samples of subjects studied to date, produce the results predicted by the special deterrence thesis. Instead, many caught offenders seem to "reset" their estimated probability of arrest back down to some lower level that prevailed prior to capture (Pogarsky and Piquero 2003). This may partly explain why criminogenic effects of legal punishment seem to outweigh the special deterrent and other crime-reducing effects. While personally experiencing punishment may make its unpleasantness more vivid and concrete as Andenaes suggested (this effect has never been systematically studied), its special deterrent effect is nevertheless weak because the experience of punishment does not, on average, increase the punished person's perceived certainty of future punishment.

The Effect of Actual Punishment Rates on Individual Perceived Risk of Punishment

While some self-report studies have treated perceptions of punishment as a dependent variable, only a handful have assessed the impact on those perceptions of actual punishment levels prevailing in the person's area. One early study examined, in a limited way, the association between actual and perceived punishment levels. Erickson and Gibbs (1978) surveyed a random sample of Phoenix residents, asking them to estimate the probability of arrest for ten different offenses. Comparing their collective estimates with police statistics on arrest probabilities, they found a 0.55 Pearson correlation (rho =.39) between objective and perceived certainty of arrest, across ten offenses (p. 259). This study addressed variation in perceptions only across offense types rather than across individuals, and in arrest certainty across offense types, rather than across areas or time periods.

The most extensive investigation of the link between actual punishment levels and individual perceptions of punishment risk was conducted by Kleck and his colleagues (2005), and examined the perceived risks of punishment of persons interviewed in a national telephone survey of a representative sample of urban US adults. Their perceptions of the risk of arrest, conviction, sentencing to prison, average prison length, and swiftness of punishment for four different crime types were compared with the actual risks prevailing in the survey respondents' areas, as measured with official police and court data. The findings indicated that associations between perceived and actual legal risks were not significantly different from zero, for any of the types of risk or offense types.

Thus, there appears to be no significant association between perceptions of punishment levels and the actual levels of punishment that CJS agencies achieve, implying that increased actual punishment levels do not routinely reduce crime through general deterrence mechanisms.

Are Average Collective Perceptions of Legal Punishment Risk More Accurate?

Scholars from a diverse array of disciplines, addressing a wide variety of specific subjects, have argued that there is a "collective wisdom" in large populations that is not evident when one studies only individuals. For example, students of public opinion have theorized that "averaging many individuals' survey responses... tends to cancel out the distorting effects of random errors in the measurement of individuals' opinions" (Page and Shapiro 1992, p. 15), resulting in aggregated public opinions regarding policies and political leadership that are stable, meaningful, and responsive to changes in available information.

Does a sort of "collective wisdom" concerning punishment risks prevail at the aggregate level for local populations that is not evident at the level of individual persons? Perhaps the *average* perception prevailing within a population may correspond closely to punishment realities, even if any one individual substantially over- or underestimates risks. In his review of early deterrence research, economist Philip Cook (1980) acknowledged that many individuals may not react to changes in punishment risks in rational or predictable ways, but argued that some do, and therefore *average* population-wide perceptions may well respond to shifts in actual risks generated by the criminal justice system in a predictable and "rational" way. With respect to the deterrence of robbery, he hypothesized that "an increase in the true effectiveness of the system results in a corresponding increase in the *mean* of robbers' perceptions of effectiveness, and an increase in the number of robbers who are deterred," mainly because "the likelihood that a prospective offender will observe one or more friends apprehended is increased when the overall effectiveness of the system increases" (p. 225). If Cook is correct, increases in actual punishment levels should produce increases in *average* levels of perceived punishment risk, and thereby increase deterrent effects, even if there is little correspondence evident at the

individual level between perceptions of system effectiveness and actual effectiveness.

There are strong theoretical reasons to expect that this gap is a large one, and that perceptions of the risk of legal punishment are likely to have at best only a weak relationship with actual risks. First, people have limited capacities, or inclinations, to acquire, retain, and later make use of information, especially with respect to crime and punishment (Simon 1957; Kleck 2003). Defending the deterrence doctrine, Cook (1980) argued that deterrent effects could still exist in the face of limited information about the prospects of punishment, implying that incorporating these limits required only a modest revision of deterrence theory.

Second, people display certain consistent biases in acquiring, retaining, and using information, which tend to weaken the connection between actual contingencies such as legal risks, and peoples' perceptions of them. Tversky and Kahneman (1974; Kahneman and Tversky 1979) demonstrated in a long series of experiments that people faced with the need to decide under conditions of uncertainty make use of "a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (1974, p. 1124). These simplified decision rules may often produce reasonably successful outcomes, but they also result in consistent deviations from what utility maximization would predict, yielding decisions that, from the standpoint of orthodox rational choice theory, are "irrational." People deviated from simple rationality, in consistent ways, which the authors called "biases."

For example, people tend to ignore the prior probability of outcomes of a decision, and thus ignore the "base rate" or population-wide probability of it occurring. Regarding the deterrent effects of punishment, this implies that it may, for most people, be irrelevant what the population-wide probability of being caught when committing a crime is, perhaps because people give far more weight to their own experiences than to those of others, even if their own experiences may be very unrepresentative and unlikely to be repeated, and the experiences of a large population would provide a much more reliable basis for forecasting future risks for the individual. If the "base rate" information is irrelevant to decision-making, people are unlikely to acquire and retain such information in the first place.

The only direct test of the collective wisdom hypothesis found no support for it (Kleck and Barnes 2013). Regardless of whether one focuses on perceptions of individual persons, or the average of perceptions among large populations, there is generally no significant association between perceptions of punishment levels and actual levels of punishment. This in turn implies that increases in punishment levels do not routinely reduce crime through general deterrence mechanisms since the fundamental link between actual punishment levels and perceptions of punishment levels appears to be weak to nonexistent. Increases in punishment might reduce crime through incapacitative effects, through the effects of treatment programs linked with punishment, or through other mechanisms, but they are not likely to do so in any way that depends on producing changes in perceptions of risk.

These findings do not, on the other hand, imply that punishment does not exert any general deterrent effect. Rather, they support the view that any deterrent effect, however large or small it may be, does not covary with actual punishment levels to any substantial degree, since the perceptions of risk on which deterrent effects depend generally do not covary with punishment levels, within the range of levels prevailing in contemporary urban America. There may be some baseline level of deterrent effect generated by punishment-generating activities of the criminal justice system, but this level apparently does not consistently increase with increased punishment levels or diminish with decreased punishment levels. Thus, increased punishment levels are not likely to increase deterrent effects, while decreased punishment levels are not likely to decrease deterrent effects.

For those seeking ways to improve crime control, these findings suggest a need for either

(1) a shift in resources towards strategies whose success does not depend on general deterrence effects, or (2) different, nonroutine methods for generating effective deterrence messages. One approach in the latter category is to more narrowly target very specific deterrence messages at audiences who are at especially high risk of committing crimes in the near future. This was the main idea behind a program implemented in Boston and aimed at reducing youth gang violence. Rather than using apprehension, prosecution and punishment to send very broad "wholesale" deterrence messages aimed at the general population, the Ceasefire program delivered direct and explicit "retail deterrence" messages to a relatively small target audience of gang members and potential members. Unfortunately, evaluations of this sort of program are limited (Kennedy 1997).

It is also possible that unusually highly publicized punishment events may generate deterrent effects that the routine, largely unpublicized punitive activities of the criminal justice system do not. For example, it has been asserted that highly publicized executions exert an effect, albeit a possibly temporary one, on homicidal behavior (Phillips 1978; Stack 1987), and the same might be true of less extreme punishments, such as incarceration, if sufficiently publicized. There is, however, no persuasive evidence bearing on publicized punishment events other than executions or death sentencings. Further, there is a severe upper limit on how much publicity-dependent deterrent effects could be increased, since the very newsworthiness that is essential for gaining publicity would, in the absence of direct state control over news media, decline as soon as a given type of punishment event became more common. A few punishment events are highly publicized for a time, and may shift perceptions of legal risk upwards, but such effects are likely to be temporary, lasting only as long as the associated crime is the news media story *du jour*.

Criminals' awareness of legal risks may be largely confined to the most conspicuous features of their immediate environments at the time a crime is contemplated. They are aware of the presence of a police officer, patrol car, or bystander who might intervene or summon the police, but are not sensitive to the overall likelihood of arrest in their areas. These localized perceptions, however, may do little more than displace offenders to other places and times rather than deterring offenses altogether.

One can view prospective offenders' responses to punishment levels as characterized by a severely constricted rationality. While many people are capable of weighing perceived risks and rewards when deciding whether to do crime, they typically possess so little accurate information about key risks and rewards that this capacity for rational decision-making remains to a great extent inoperative.

Open Questions

The research linking perceptions of legal risks to actual risk levels has so far been cross-sectional in nature, and thus has not directly assessed the impact of changes over time in legal risk on perceived risks. Thus, while cross-area differences in objective risk appear not to influence differences in perceived risk across areas, further perceptual research, perhaps using a panel design, is needed to definitively establish whether changes in actual punishment levels produce changes in perceived risk.

It remains possible that prospective offenders are sensitive to changes in legal risks, but only for short periods of time. Thus, deterrent effects of legal threats might be elevated, but for so short a time that these effects are not evident in research using temporal units of analysis as long as years or months. For example, executions may deter homicides in the days just before or after the event, but analysts studying homicide counts for years would miss the effect because the number of homicides deterred would be tiny relative to annual murder counts. Thus, perceptual research examining crime levels in shorter time periods than years or months may prove useful.

Related Entries

- ▶ Certainty, Severity, and Their Deterrent Effects
- ▶ Deterrence of Tax Evasion
- ▶ Deterrent Effect of Imprisonment
- ▶ Focused Deterrence and "Pulling Levers"
- ▶ Offender Decision Making and Behavioral Economics
- ▶ Rational Choice and Prospect Theory
- ▶ Rational Choice Theory
- ▶ Rational Choice, Deterrence, and Crime: Sociological Contributions

Recommended Reading and References

Andenaes J (1952) General prevention: illusion or reality? J Crim Law Criminol Polit Sci 43:176–198

Becker G (1968) Crime and punishment: an economic approach. J Polit Econ 76:169–217

Cook P, Norval M (1980) Research in criminal deterrence: laying the groundwork for the second decade. In: Michael T (ed) Crime and Justice: an annual review of research, vol 2. University of Chicago Press, Chicago

Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. J Polit Econ 81:521–565

Erickson M, Gibbs J (1978) Objective and perceptual properties of legal punishment and the deterrence doctrine. Soc Probl 25:253–264

Garofalo J (1981) Crime and the mass media. J Res Crime Delinq 18:319–350

Gibbs J (1975) Crime, punishment and deterrence. Elsevier, New York

Heckert A, Gondolf E (2000) The effect of perceptions of sanctions on batterer program outcomes. J Res Crime Delinq 37:369–391

Horney J, Marshall IH (1992) Risk perceptions among serious offenders: the role of crime and punishment. Criminology 30:575–592

Jacob H (1979) Rationality and criminality. Soc Sci Q 59:584–585

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–291

Kennedy D (1997) Pulling levers: getting deterrence right. Natl Inst Justice J 236:2–8

Kleck G (2003) Constricted rationality and the limits of general deterrence. In: Blomberg T (ed) Punishment and social control: enlarged second edition. Aldine de Gruyter, New York

Kleck G, Barnes JC (2013) Deterrence and macro-level perceptions of punishment risks: is there a "collective wisdom?". Crime Delinq (in press)

Kleck G, Sever B, Li S, Gertz M (2005) The missing link in general deterrence research. Criminology 43: 623–660

Lochner L (2007) Individual perceptions of the criminal justice system. Am Econ Rev 97(1):444–460

Marsh HL (1989) Newspaper crime coverage in the U.S., 1983–1988. Crim Justice Abstr 21:506–514

Matsueda R, Kreager D, Huizinga D (2006) Deterring delinquents: a rational choice model of theft and violence. Am Soc Rev 71:95–122

Nagin D (1998) Criminal deterrence research at the outset of the twenty-first century. In: Michael T (ed) Crime and justice, vol 23. University of Chicago Press, Chicago

Page BI, Shapiro RY (1992) The rational public. University of Chicago Press, Chicago

Parker J, Grasmick H (1979) Linking actual and perceived certainty of punishment: an exploratory study of an untested proposition in deterrence theory. Criminology 17:366–379

Paternoster R (1987) The deterrent effect of the perceived certainty and severity of punishment: a review of the evidence and issues. Justice Q 4:173–215

Paternoster R, Piquero A (1995) Reconceptualizing deterrence: an empirical test of personal and vicarious experiences. J Res Crime Delinq 32:251–286

Paternoster R, Saltzman L, Chiricos T, Waldo G (1982) Perceived risk and deterrence: methodological artifacts in perceptual deterrence research. J Crim Law Criminol 73:1238–1259

Paternoster R, Saltzman L, Chiricos T, Waldo G (1985) Assessments of risk and behavioral experience: an exploratory study of change. Criminology 23:417–436

Phillips DP (1978) The deterrent effect of capital punishment: new evidence on an old controversy. Am J Soc 86:139–148

Piliavin I, Thorton C, Gartner R, Matsueda R (1986) Crime, deterrence, and rational choice. Am Soc Rev 51:101–119

Piquero A, Paternoster R (1998) An application of Stafford and Warr's reconceptualization of deterrence to drinking and driving. J Res Crime Delinq 35:3–39

Piquero A, Pogarsky G (2002) Beyond Stafford and Warr's reconceptualization of deterrence: personal and vicarious experiences, impulsivity, and offending behavior. J Res Crime Delinq 39:153–186

Pogarsky G, Piquero A (2003) Can punishment encourage offending? Investigating the resetting effect. J Res Crime Delinq 40:95–120

Pogarsky G, Piquero AR, Paternoster R (2004) Modeling change in perceptions about sanction threats: the neglected linkage in deterrence theory. J Quant Criminol 20:343–369

Pogarsky G, Kim KD, Paternoster R (2005) Perceptual change in the national youth survey. Justice Q 22:1–29

Roberts JV (1992) Public opinion, crime, and criminal justice. Crime Justice 16:99–180

Roshier B (1973) The selection of crime news by the press. In: Cohen S, Young J (eds) The manufacture of news. Constable, London
Simon H (1957) *Models of Man*. NY: Wiley
Stack S (1987) Publicized executions and homicide, 1950–1980. Am Soc Rev 52:532–540
Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131
Wood PB (2007) Exploring the positive punishment effect among incarcerated adult offenders. Am J Crim Justice 31:8–22
Zimring FE, Hawkins G (1973) Deterrence: the legal threat in crime control. University of Chicago Press, Chicago

Deterrent Effect of Imprisonment

Roberto Galbiati[1] and Francesco Drago[2]
[1]Department of Economics, CNRS-OSC and Sciences Po, Paris, France
[2]University Federico II Naples and CSEF, Napoli, Italy

Overview

During the last decades, societies have largely used incarceration as a central crime control tool. Between 1970 and 2008, the prison population per 1,000 inhabitants increased by a factor of more than 4.5 in the USA. Despite a dramatic difference in incarceration levels between the two sides of the Atlantic, also in European countries, prison population increased by a factor between two and three over the period 1970–2008 (Buonanno et al. 2011). This massive increase in incarceration had been coupled with a strong debate in social sciences over the magnitude of its impact on crime rates and over the reasons why crime rates might react to changes in prison population. Those favoring an increase in the severity of the criminal justice system have often argued that increasing prison sentences will lead prospective criminals to reduce their criminal activity. Criminals in fact would weight costs and benefits of crime, and an increase in the cost of criminal activities following longer expected prison sentences will induce a subset of them to refrain from engaging into crime. Those holding an opposite view maintained that an increase in incarceration might have a negative net effect. One of the main arguments against an increase in the severity of incarceration system maintains that while imprisonment may temporarily constraint inmates thus preventing them from criminal activity (*incapacitation*), the negative impact of longer or harsher incarceration for those who are incarcerated may enhance subsequent levels of criminal activity.

Which of the two views is more plausible crucially depends on the relative weight of the two effects: the behavioral response of prospective criminals with respect to an increase in expected sentences (general deterrence) and the behavioral consequences of longer or harsher prison sentences for those that experienced incarceration (specific deterrence).

Overall, estimating the relative weight of these two effects is fundamental to the understanding of whether the actual incarceration rates are at the optimal level. It is clear that prison sentences do not deter all the individuals from committing crime (otherwise prisons would be empty), but only a subset of all potential criminals. Therefore, an increase in the severity of prison sentences increases the incarceration rates. An incarceration rate is set at the optimal level if the social cost of incarcerating an additional individual (at that incarceration rate) is not below the social benefit deriving from the prevention of an eventual crime committed by that individual. While it is very difficult to compute the cost and the benefit from incarcerating an additional individual, it is very important to understand how potential criminals respond to prison sentence severity and the criminogenic effect of prisons. Both these effects enter the cost and the benefit from increasing the severity of prison sentences and incarceration rates.

Fundamentals of Deterrence Effects of Imprisonment

Definitions

In order to better understand what is the meaning of deterrence effects of imprisonment, it is useful

to refer to the pioneer economic model of crime Gary Becker (1968). Becker's contribution is crucial because it provides a simple theoretical framework presenting the commission of a crime as a rational choice reflecting a cost-benefit comparison. In this simple framework, an individual chooses to commit a crime if the expected payoffs (weighted by a utility function) from the commission of a crime exceeds the payoffs from not committing a crime. The expected payoff from committing a crime depends on the prison sentence from that crime, the probability to be caught and convicted to a prison sentence, and the gain deriving from that crime. The expected payoff from not committing a crime depends on the payoff from legal activities (e.g., expected wage from a legal job, disutility from working). An individual compares the expected costs and benefits from committing a crime and behaves accordingly. If the potential criminal opts for the criminal activity, we infer that he faces a net expected gain from the activity despite the expected punishment. This simple reasoning highlights that the probability of being punished and the costs associated with being punished are crucial elements for a criminal's choice. Rephrasing this point, the rational choice theory predicts that changes in the certainty and the severity of the punishment will induce changes in criminals' behavior.

The probability of punishment will depend upon many factors (number of police forces, efficiency of the justice system, and apprehension technology among others). For what concerns the severity of the punishment, two elements are crucial. One is length of the expected sanction and in particular the expected length of imprisonment. The other fundamental aspect is given by the conditions of detention. Both longer and harsher expected prison sentences represent a higher cost for prospective criminals. According to this framework, by varying these three elements, a policy maker can affect criminals' choice and thus can affect the level of crime rates.

The *general deterrent effect of imprisonment* is the response of potential criminals to the expected cost associated with the decision to commit a crime. This is an ex ante perspective. The choice of committing a crime is affected by the expected level of punishment that in turn reflects the probability of apprehension and conviction, the expected length of imprisonment, and the expected harshness of the imprisonment.

Since a share of those that are apprehended and convicted will actually spend some time in prison, it is natural asking if the time spent in prison and the condition of detention will affect criminals' propensity to commit a crime once they are released. If we assume that incarceration leads criminals to update their beliefs about the consequences of punishment, we might expect that having experienced harsher punishment should reduce the inmates' propensity to recommit a crime once released. Under this perspective, imprisonment is an experience good. Criminals ex ante do not perfectly know the cost of spending time in. Once they are convicted to a prison sentence and they make the experience of incarceration, ex post, they update their knowledge of prison conditions. Experiencing imprisonment and staying in prison longer will affect criminals' knowledge of the prison environment and in turn their evaluation of expected costs of committing a crime once they are released. For this reason, those having experienced harsher prison conditions should be less prone to recommit a criminal act. This is known as the *specific deterrence hypothesis*. On the other hand, however, experiencing harsher prison conditions could also imply higher propensity of committing crime once released. In fact, harsher prison conditions may lead to a higher human capital deployment and worse labor outcomes for former inmates (Waldfogel 1994). Since these factors affect the relative benefits of engaging into legal or criminal activities, they will affect former inmates' criminal choice. In this case, the theoretical predictions are not clear-cut. Hence, understanding the *specific deterrence effects of imprisonment* is ultimately an empirical question.

Current Issues and Controversies

Understanding how and whether potential criminals respond to changes in expected prison sentences (general deterrence) and if prison experience affects former inmates' decisions to recommit a criminal act after release

(specific deterrence) has been the object of a large and growing empirical literature. Most of the studies have focused on the American criminal justice system, but in recent years several studies have provided evidence on European ones. Many of these studies apply modern econometric techniques allowing to get rid of potential confounding factors that could affect the interpretation of the findings. These exercises thus provide important insights that could usefully inspire crime control policy interventions. Despite these relevant advances, international comparisons are still very difficult given the large diversity of the specific national contests.

Empirical studies aiming at understanding to what extent potential criminals respond to manipulations in expected prison sentences and/or to what extent the prison conditions experienced during incarceration by former inmates affect their post-release behavior need to face some major empirical challenge.

For what concerns *general deterrence*, longer prison sentences might deter potential criminals but might also reflect changes in the general attitude toward criminality or might anticipate expected increases in crime rates. For this reason, it is hard to disentangle if there is any causal effect of increasing expected prison sentences on criminal behavior. The same reasoning applies to the general deterrent effects of expected harshness of prison conditions. For example, overcrowded prisons might affect potential criminals' choices because more crowded prisons imply higher expected costs of detention but at the same time will mechanically reflect an increase in crime rates. Pinning down any *general deterrence* effect of harsher prison conditions is thus particularly challenging.

Understanding the *specific deterrent effect of prison* suffers from the same empirical challenges. Inmates suffering harsher prison conditions will probably react to them once they will be released, but being assigned to harsher prisons might reflect higher individuals' dangerousness. Thus, even in the presence of detailed individual level data, it is hard to understand if and to what extent prison conditions have any causal effect on former inmates' behavior without some clear empirical research design able to break the simultaneity between prison conditions and individuals' propensity to reoffend.

For these reasons, most of the recent research has focused on specific case studies allowing to quasi-experimentally disentangle the effect of the crucial policy manipulable elements on individuals' choices.

In particular researchers have investigated the general and specific deterrent effects of prison sentences' length and of prison conditions. Related to this last point, some interesting studies have focused on an important but largely neglected aspect of incarceration: the effect of peers' characteristics and behavior on inmates' post-release behavior. This aspect is crucial since there is a large and consistent evidence that peers' might affect people behavior in many domains, ranging from educational attainment to job search. If the same holds for crime, prisons might be a sort of criminal school where inmates learn or are influenced from each other. Thus, other inmates' characteristics or behavior might crucially affect ones' post-release behavior.

General Deterrence

Prison Sentences Length and Prison Conditions

Evidence on the effects of the severity of punishment on criminal activity is growing and consistently point out that, as predicted in the standard economic/rational choice approach to crime, potential criminals take into expected sentences' length when they decide whether to commit a criminal act. To understand this effect, most works in this field have studied the effect of incarceration rates on aggregate crime rates. In order to break the simultaneity between prison population and crime rates, researchers have resorted to case studies exploiting exogenous variation in prison population or expected prison sentences' length. An influential study based on US State level data, Levitt (1996), shows that releasing one prisoner is associated with an increase of 15 crimes per year. This estimate, however, includes deterrence and incapacitative effects. A major change in expected prison sentences in the USA have been introduced in some states through sentence enhancements'

laws such as California's "three strikes and your out" law or Proposition 8. These laws provide researchers with some variation on expected prison sentences' length useful to disentangle deterrence and incapacitation. Among these studies, Kessler and Levitt (1999) exploiting sentence enhancements targeting the most frequent and dangerous criminals find that some crime rates fell by 4 % after sentence enhancement, which, for example, increased the sentence for any "serious" felony offender by 5 years. Exploiting the increase in expected sentences induced by the California "three strikes law," Helland and Tabarrok (2007) compare the future offending of individuals convicted of two previous strikeable offenses with that of individuals who had been convicted of only one strikeable offense but who, in addition, had been tried for a second strikeable offense but were ultimately convicted of a nonstrikeable offense. The study finds that arrest rates were about 20 % lower for the group with convictions for two strikeable offenses providing evidence in line with previous studies.

Since the cost of imprisonment does not depend only on the expected prison sentence length but also on expected prison conditions, an important issue is to understand if and how potential criminals respond to expected harsher prison conditions. This is an interesting aspect since, setting aside ethical considerations, prison conditions are more easily manipulable than sentences' length. Such a deterrent effect of harsher prison conditions seems to exist in the USA where recent research (Katz et al. 2003), using death rates among prisoners as a proxy for prison conditions, shows that more punitive facilities have a small but statistically significant general deterrent effect.

These and several other studies taking seriously the econometric challenges behind the identification of the deterrent effect of prison sentences' length show that, at least for USA, potential criminals take into account expected sentence length and prison severity when they have to decide whether to engage into a criminal act.

Evidence outside the USA is less systematic, but a recent large criminal justice policy intervention in Italy has provided researchers the opportunity to study for a large sample of former inmates what is their response to a manipulation of expected prison sentences.

The Collective Clemency Bill passed by the Italian Parliament in July 2006 provided an ideal case study to understand how people respond to exogenous variations in prison sentences. This law provided for an immediate 3-year reduction in detention for all inmates who had committed a crime before May 2, 2006. Upon the approval of the bill, almost 22,000 inmates – about 40 % of the prison population of Italy – were released from Italian prisons on August 1, 2006. The bill states that if a former inmate recommits a crime within 5 years following his release from prison, he will be required to serve the remaining sentence suspended by the pardon (varying between 1 and 36 months) in addition to the sentence given for the new crime. This is equivalent to a policy manipulating incentives to commit a crime since it commutes 1 month of time of the original sentence to be served in 1 month more of expected sentence for future crimes. More importantly, this institutional framework manipulates prison sentences at the individual level in a random fashion. In particular, conditional on inmates' original sentences, the variation in the remaining sentence at the date of the pardon (and hence in the expected sentence for any crime) depends only on the date of an inmate's entry into prison, which is plausibly non correlated with individuals' characteristics potentially affecting their propensity to recidivate. The variation in the remaining sentence at the date of the pardon can thus be used to identify the causal impact of a manipulation of expected prison sentences on individuals' propensity to recommit a crime. Research based on data from the Italian experiment (Drago et al. 2009) shows that a marginal increase in the remaining sentence reduces the probability of recidivism by 0.16 percentage points (1.3 %). This means that for former inmates, 1 month less time served in prison commuted into 1 month more in expected sentence significantly reduces their propensity to recommit a crime. Interestingly, this deterrent effect of prison

sentences is quite homogeneous across inmates with different individual characteristics, but individuals convicted to relatively longer sentences do not seem to be deterred. This suggests that longer sentences ex ante deter more but experiencing longer time in prison might be deleterious in terms of former inmates' behavioral response to incentives.

Given this apparently consistent evidence suggesting that longer and harsher prison sentences deter crime, it could be tempting to draw immediate policy conclusions. However, it is extremely important to be cautious in making this further step. In fact it is still far to be clear for researchers what are the most cost-effective means to reduce criminal activity, and even just looking at deterrence, it is far from clear if criminals respond more to an increase in the severity rather than an increase in the certainty of punishment (Durlauf and Nagin, 2011). For example, in a pioneering study, DiTella and Shargrodsky (2009) show that electronic monitoring is more effective than incarceration in reducing ex post former prison inmates' criminal behavior. Moreover, existing evidence suggests that the effect of the aggregate impact of changing prison conditions on crime rates is relatively small (Katz et al. 2003). Finally, the evidence from the Italian case suggesting that those having spent long prison sentences do not respond to a manipulation in prison sentence after release suggests that the ex post effects of prison sentences might compromise the ex ante deterrent effect.

Since most criminals experiencing prison sentences experience multiple prison conditions during their life and constitute the core of the criminals' population, for the reasons explained above, understanding specific deterrence and the effect of incarceration on those that have experienced it is a crucial challenge for social scientists.

Specific Deterrence

Prison Conditions and Time Served

Opening the prison black box, we find very different punitive situations in terms of overcrowding, health services, social activities for inmates, and so on. All these elements might impact former inmates' propensity to reoffend. Those inmates having experienced harsher prison conditions might be more deterred by the threat of future sentences, but these same people might suffer from more human capital deployment and worse health conditions thus facing lower opportunity costs of criminal activity of engaging. While the issue of the deterrent effects of prison treatment appears particularly important for both researchers and policy makers, the empirical evidence is scarce. Only a few recent works analyze the effects of prison conditions on criminal behavior. The lack of evidence is mainly due both to the difficulty in obtaining access to reliable data on prison conditions and to the identification challenges faced by researchers interested in understanding how criminals respond to prison conditions. Nonetheless, the few existing studies focusing on the issue of prison conditions and former inmates behavior provide some crucial knowledge on the phenomenon. Focusing on women incarceration conditions, recent evidence exploiting the expansion of female penal system capacity in the United States (Bedard and Helland 2004) shows that, on average, increasing the distance of detention facilities from inmates' home tends to lower the female crime rate. Despite this results seem to conform to the deterrence hypothesis; from this case study, it is hard to understand if the deterrent effect is driven by the response of formers inmates or by the reaction of criminals who had never received a prison treatment (or both). Exploiting quasi-experimental designs and individual level data is crucial to separate general deterrence from specific deterrent effects of incarceration conditions. Studies using these two ingredients cast doubts on the capacity of harsher prison conditions to reduce former inmates' propensity to recidivate. By exploiting a discontinuity in the assignment of federal prisoners to security levels, American economists provide evidence that serving a sentence in a higher security prison implies a higher post-release propensity to commit a crime (Chen and Shapiro, 2007). In a similar vein, exploiting some randomness in the facility assignment rules and individual level data on inmates' post-release behavior, an Italian case study (Drago et al. 2011) shows that being assigned to a prison where mortality is

higher increases the propensity to recidivate for former inmates.

For the analysis of the effect of the amount of time spent in prison, recent quasi-experimental evidence from the USA and France (Kuziemko, 2007 and Maurin and Ouss 2009) suggests that spending more time in prison all else equal should reduce the propensity to recidivate for former inmates. Taken together the results of recent empirical literature focusing on the specific deterrent effects of prison in terms of prison sentences' length and detention conditions quality are mixed. Thus, deriving any strong policy conclusion based on the current evidence is not only hard but also not recommended.

Peer Effects Between Prison Inmates

Experiencing incarceration might affect post-release inmates' behavior not only through the effects of time spent in physical prison conditions or in prison activities but also through the effect of interactions with other inmates. Social scientists have documented that the exposure to peers' actions and characteristics might have a strong impact on individuals' behavior in various domains ranging from school performance to labor supply decisions. Peers might influence one's decision by providing information and by affecting social norms, constraints to action, and various other channels. Individual choices to engage into criminal behavior are also plausibly affected by peers' influence since peers affect the criminal markets' conditions, provide information about criminal opportunities, and act together. Estimating the impact of peer effects on criminal behavior and understanding the mechanisms through which peers affect individual criminal participation are of primary importance in the design of effective policies to prevent crime (Manski, 1993; Glaeser et al. 1996). Understanding if peer effects between prison inmates exist and estimating their magnitude is crucial to understand how prison experience affects former inmates' behavior. In this sense potential peer effects are a crucial component of prison experience, and understanding them is particularly important to assess overall general *and specific deterrent effects of prison* and thus to understand how the prison system works overall.

Peers effects between former prison inmates might take place because once in prison criminals respond to the characteristics of those they spend time with. For example, juvenile offenders serving time in the same correctional facility in Florida seem to have some influence on each other's subsequent criminal behavior (Bayer et al. 2009). The data on juvenile inmates spending some time in 1 of 169 Florida correctional facilities provide a complete record of past crimes, facility assignments, and arrests and adjudications in the year following release for each individual. These kind of data have allowed researcher to solve the simultaneity problem that usually hamper the identification of peer effects. From the Florida case study, we learned that peers exert their influence by influencing individuals who already have some experience in a particular crime category (Bayer et al. 2009). This evidence supports the idea that prisons are a sort of criminals' school.

Another plausible mechanism underlying peer effects is the maintenance of prison peer groups after release and the presence of complementarities in post-release behavior (e.g., joint crime production). Sociological and qualitative research on prison gangs (Skarbek 2010; Leeson and Skarbeck 2010; Fleisher and Decker 2001) and on former Italian inmates' post-release networks (Baccaro and Mosconi 2004; Santoro and Tucci 2006) supports this interpretation. Moreover, exploiting data on the Italian prison experiment described above, researchers have shown that former inmates from the same nationality that spent time together in the same prison facility tend to influence each others' post-release behavior (Drago and Galbiati, 2012). In particular, criminals respond to the behavior of those they spent time in prison with by increasing their propensity to commit crime if others' criminal activity increases. This suggests that peer groups formed in prison remain the same after release, and thus peers will continue to influence each other even after release from prison.

Thus, peer effects are a crucial aspect of inmates' interaction that have to be carefully taken into account in order to understand the overall effect exerted by prison sentencing on individuals' criminal choices.

Directions for Future Research

Despite a growing research effort, clean evidence about general and specific deterrent effects of prison is still scarce. Some more systematic research and a larger number of specific country case studies are most welcome to fill the knowledge gap.

Given that identifying causal relation is particularly difficult in this domain, a more systematic access and use of individual level data is needed. Some particularly fruitful lines of research should focus on understanding the relative weight of severity and certainty of punishment in determining the overall general deterrent effect of prison sentences. Continuing in the effort to open the specific deterrence black box is crucial in order to understand the dynamics of future crime rates since a large share of those that have been incarcerated in the USA and Europe during the last decades will be released in the following years.

Finally these research building blocks could and hopefully will be used to build a more comprehensive approach to determine the immediate and dynamic future effects of manipulating the length and harshness of prison sentences.

Related Entries

- ▶ Certainty, Severity, and Their Deterrent Effects
- ▶ Deterrence of Tax Evasion
- ▶ Econometrics of Crime
- ▶ Rational Choice, Deterrence, and Crime: Sociological Contributions

Recommended Reading and References

Baccaro L, Mosconi G (2004) Il girone dei dannati: ovvero il fenomeno della recidiva. Rass Penit e Criminol 2:212–237

Bayer P, Hjalmarsson R, Pozen D (2009) Building criminal capital behind bars: peer effects in juvenile corrections. Q J Econ 124(1):105–147

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76:169–217

Becker G (1978) The economic approach to human behavior. University of Chicago Press, Chicago, Chapter 1

Bedard K, Helland E (2004) The location of women's prisons and the deterrence effect of harder time. Int Rev Law Econ 24:147–167

Buonanno P, Drago F, Galbiati R and Zanella G (2011) Crime in europe and US: dissecting the reversal of misfortunes. Econ Policy

Chen MK, Shapiro JM (2007) Do harsher prison conditions reduce recidivism? a discontinuity-based approach. Am Law Econ Rev 9(1):1–29

Di Tella, Rafael, Ernesto Schargrodsky. (2009). Criminal recidivism after prison and electronic monitoring. National Bureau of Economic Research Working Paper 15602

Drago F, Galbiati R (2012) Indirect effects of a policy altering criminal behavior: evidence from the italian prison experiment. Am Econ J Appl Econo 2

Drago F, Galbiati R, Vertova P (2009) The deterrent effects of prison: evidence from a natural experiment. J Polit Econ 117(2):257–280

Drago F, Galbiati R, Vertova P (2011) Prison conditions and recidivism. Am Law Econ Rev 13(1):103–130

Durlauf S and Nagin D (2011). The deterrent effect of imprisonment. In Cook P.J, Ludwig J and McCrary J eds Controlling crime: strategies and tradeoffs. NBER

Fleisher MS, Decker SH (2001) An overview of the challenge of prison gangs. Correct Manag Q 5(1):1–9

Glaeser EL, Sacerdote B, Scheinkman JA (1996) Crime and social interactions. Q J Econ 111(2):507–548

Helland E, Tabarrok A (2007) Does three strikes deter? a nonparametric estimation. J Hum Resour 22:309–330

Katz L, Levitt SD, Shustorovich E (2003) Prison conditions, capital punishment, and deterrence. Am Law Econ Rev 5:318–343

Kuziemko Ilyana (2007) Going off parole: how the elimination of discretionary prison release affects the social cost of crime Working Paper no. 13380 (September), NBER, Cambridge, MA

Leeson PT, Skarbek DB (2010) Criminal constitutions. Glob Crime 11(3):279–297

Levitt SD (1996) The effect of prison population size on crime rates: evidence from prison overcrowding litigation. Q J Econ 111(2):319–351

Levitt S, Miles T (2007) Empirical study of criminal punishment. In: Mitchell Polinsky A, Shavell S (eds) Handbook of law and economics, vol 1. North-Holland, Amsterdam

Manski CF (1993) Identification of endogenous social effects: the reflection problem. Rev Econ Stud 60(3): 531–542

Maurin Eric Aurelie Ouss. (2009) Sentence reductions and recidivism: lessons from the bastille day quasi experiment IZA Discussion Paper no. 3990 (February), Inst. Study Labor, Bonn

Santoro E, Tucci R (2006) L'incidenza dell'affidamento sulla recidiva: prime indicazioni e problemi per una ricerca sistematica. Rass Penit e Criminol 10(1):79–158

Skarbek D (2010) Putting the 'con' into constitutions: the economics of prison gangs. J Law, Econ Organ 26(2):183–211
Waldfogel J (1994) The effect of criminal conviction on income and the trust reposed on workmen. J Hum Res 29:62–81

Development of International Criminal Law and Tribunals

Kai Ambos
Department of Foreign and International Criminal Law, Institute for Criminal Law and Justice, Georg-August-Universität Göttingen, Göttingen, Germany

Overview

This entry offers a brief reconstruction of the development of International Criminal Law (ICL) and an equally brief description of the existing International Criminal Tribunals and Courts. The former part will be divided into two sections: starting with the Versailles Peace Treaty and historical precedents, the Nuremberg and Tokyo trials are looked at as the first stage of the development, focusing on international conflicts. The establishment of the UN ad hoc Tribunals in the 1990s accompanied by the development of ICL for non-international conflicts constitutes the second stage. The third stage, i.e., the response of ICL to transnational terrorism after 11 September 2001, will not be covered by this outline. The description of the existing tribunals starts with the UN ad hoc Tribunals for the former Yugoslavia and Rwanda, continues with a more detailed characterization of the International Criminal Court, and finishes with the so-called mixed or hybrid tribunals.

The Versailles Peace Treaty and Historical Precedents

Crimes against the basic principles of humanity are nothing new to the history of mankind. The crusades of the eleventh century may be considered as early forms of genocide. Other examples of international crimes are the Spanish and Portuguese Conquista of the Americas accompanied by the extermination of great numbers of the native population, the massacre of thousands of the French Huguenots during St. Bartholomew's night 23 August 1572, and the massacre of Glen Coe in the year 1692. In all these cases, investigations never took place, and criminal sentences were never passed on the responsible persons except in the case of Peter von Hagenbach in the year 1474 (Cryer 2005, pp. 17 ff.).

At the beginning of the nineteenth century, the punishability of **piracy** was acknowledged under customary international law. Also, **slavery** was declared a crime of international concern due to numerous international treaties, which had been concluded since 1815 (Bassiouni 1999, pp. 305 ff.). Notwithstanding, the proposal of the president of the International Committee of the Red Cross (ICRC), Gustave Moynier, to set up an International Criminal Court (ICC) after the German-French War in 1870/1871 remained without any political resonance (König 2003, p. 60).

When the Allied and Associated Powers convened the 1919 Preliminary Peace Conference, the first international investigative commission was established, and the **Versailles Peace Treaty** was adopted. This treaty established a new policy of prosecuting war criminals of the vanquished aggressor state after the end of the hostilities. The legal basis of that policy was laid down in 1919 in the Paris Peace Treaties which created four groups of offences: crimes against the sanctity of the treaties, crimes against the international morals (which were not defined more precisely; Puttkamer 1949, p. 424), war crimes "in a narrow sense" (i.e., "violation of the laws and customs of war" according to Art. 228 of the Versailles treaty), and violations of the laws of humanity. The latter was not included in the Versailles treaty since the USA took the view that it could not be sufficiently precisely defined and thus was too vague as a basis for prosecutions (Bassiouni 1999, p. 65). At the same time, with the Versailles treaty, the individual criminal responsibility for

crimes against international law was for the first time recognized on a treaty basis. It was further recognized that such responsibility had no limits of rank or position.

In 1920, the Allied Powers decided to hand over the prosecution to Germany, which passed new legislation to be able to prosecute German suspects before its own Supreme Court (the *Reichsgericht*), sitting at **Leipzig**. However, only 12 Germans were prosecuted for war crimes (Ahlbrecht 1999, pp. 42 ff.). Thus, the so-called Leipzig Trials have been considered a failure and are widely cited as proof of the German unwillingness to seriously prosecute their own war criminals.

In international law, the term "**crimes against humanity**" was for the first time used in the context of the **genocide of the Armenians**, which led to a joint declaration of France, Great Britain, and Russia on 24 May 1915, asserting that all members of the Ottoman Government and those of its agents found to be involved in those massacres would be held personally responsible for the crimes. Prosecuted on the basis of the Turkish penal code, several ministers of the wartime cabinet and leaders of the Ittihad party were found guilty by a court martial of "the organization and execution of crime of massacre" (Schabas 2009, p. 25). At the international level, the Peace **Treaty of Sèvres**, signed on 10 August 1920 between the Ottoman Empire and the Allies (France, Italy, Japan, UK), in many aspects similar to the Treaty of Versailles, contained, as a major innovation, offences which were later qualified as crimes against humanity (Article 230). The treaty, however, never took effect. It was replaced by the Treaty of Lausanne of 24 July 1923, which included a "Declaration of Amnesty" for all offences committed between 1 August 1914 and 20 November 1922.

The First Ad Hoc Tribunals: Nuremberg and Tokyo

During World War II, the prosecution of "war crimes" became a primary objective. In 1942, the Allied Powers signed a declaration in St. James Palace in London, which established the **UN War Crimes Commission** (UNWCC). We will return to it below. The "Declaration of St. James" also laid down the foundation for the International Military Tribunal (IMT). This was followed by the Moscow Declaration of 30 October 1943, which confirmed the Allied quest for prosecution. Finally, the "Declaration of London" of 8 August 1945 – concluded by the governments of Great Britain, USA, France, and the Soviet Union – gave birth to the IMT.

The first series of World War II trials, the **Nuremberg trials**, took place under the terms of a charter drafted in London between June and August 1945 by representatives of the USA, Great Britain, Soviet Union, and France. It was therefore called the "London" or "Nuremberg Charter". The Nuremberg Charter contained three categories of offences: crimes against peace, war crimes, and crimes against humanity. As to the defenses, Article 7 rejected official position and Article 8 superior orders as grounds for excluding responsibility. The Allies set up the IMT to prosecute the "Major War Criminals". Twenty-three defendants were initially charged and 19 convicted (Engelhart 2004, p. 734 ff.).

The **Tokyo Trials** were based on the charter for the Far East, or *Tokyo Charter*, which was proclaimed on 19 January 1946. The charter was, unlike the London Charter, not part of a treaty or an agreement among the Allies. Representatives of the Allied nations, which had been involved in the struggle in Asia (the USA, Great Britain, France, Soviet Union, Australia, Canada, China, the Netherlands, New Zealand, India, and the Philippines), created the IMT for the Far East (IMTFE). It was composed of judges, prosecutors, and other staff from the Allied nations. The IMTFE recognized the same offences as the IMT: crimes against peace (as defined in the London Charter); "conventional war crimes, namely, violations of the laws or customs of war"; and crimes against humanity. The definition of crimes against humanity differed from that of the IMT Charter in two ways: first, the IMTFE Charter expanded the list of crimes to include imprisonment, torture, and rape. Second, it eliminated the requirement that "crimes against humanity" had

to be connected to war. As to possible defenses, the charter excluded – as did the IMT Charter – official position or superior orders. The prosecution selected 28 defendants; 25 were convicted (Osten 2003).

Post Nuremberg World War II Trials

The Nuremberg and Tokyo Trials were followed by a second series of prosecutions of Nazi leaders, pursuant to **Control Council Law** (CCL) No. 10. The most famous proceedings were the 12 trials before the US-American court in Nuremberg. Other important cases have been documented by the United Nations War Crimes Commission (UNWCC 1947–1949). It was formally established on 20 October 1943, and its task was basically twofold: on the one hand, to investigate war crimes, collect evidence, and identify those responsible and, on the other, to inform the Allied governments about the cases providing a sufficient basis for prosecution. In total, the UNWCC documented 89 war crimes trials on the basis of protocols of 2111 proceedings (Ambos 2002/2004, p. 140).

The proceedings instituted by the occupation powers ended a few years after the end of the war. However, the prosecutions of Nazi criminals have continued in and outside Germany until today. The most famous cases on the basis of universal jurisdiction were the trials against *Adolf Eichmann* (ILR 1968, pp. 5–14) and *Klaus Barbie* (Le Monde 5–6 July 1987, p. 1). The former was sentenced to death by the Jerusalem District Court on 15 December 1961, and the latter was first tried in absentia for war crimes and sentenced to death by the *Tribunal Permanent des Forces Armées de Lyon* in two judgements and later *in presence* sentenced to life imprisonment on 4 July 1987 for crimes against humanity. After having served 4 years of his sentence, Barbie died of leukemia in 1991. Other cases include that of *Paul Touvier* (ILR 1995, 338 ff., 357 ff.) in France, sentenced to life imprisonment before a *Cour d'Appel de Paris* in Versailles, France (20 April 1994), and that of *Imre Finta* (ILR 1995, 520 ff.) in Canada, finally acquitted by the Supreme Court (24 March 1994). Last but not least, *John Demjanjuk* was sentenced to 5 years imprisonment on 15 May 2011 but released pending an appeal; he died on 17 March 2012.

The principles resulting from the practical experience of the IMT were an important substructure for the upcoming development of ICL. The International Law Commission (ILC), founded in 1947 upon the recommendation of the Committee on the Progressive Development of International Law and its Codification (CPDIL), adopted seven principles on its second session in 1950 (ILC 1950, pp. 374 ff.). Those principles in conjunction with the Nuremberg Charter, the CCL 10, and the adjudication of the Nuremberg courts are called the "**Nuremberg Principles.**" They comprise rules on the general part (Principles I–IV, VI, and VII), on international crimes (VI), and on a procedural "fair-trial" norm (V).

The Development of International Criminal Law Prior to the Establishment of the UN Ad Hoc Tribunals

The Genocide Convention

Based on thoughts by Rafael Lemkin (1933, p. 117), Resolution 96 was adopted by the UN General Assembly on 11 December 1946. It declared genocide to be a crime of international concern and formed the basis for the drafting of a treaty by a group of experts. The Genocide Convention was adopted by the GA on 9 December 1948 and came into force on 12 January 1951. It is the most vital legal instrument on the crime of genocide (Schabas 2009, pp. 3 ff.).

The Hague and Geneva Law

First efforts to establish a "law of war" can be traced back to the middle of the nineteenth century focusing primarily on the humanization of war, first with regard to the admissible means and methods of warfare (so-called **Hague Law**) and then later increasingly with regard to the protection of the victims of armed conflict (so-called **Geneva Law**). While the Hague and Geneva laws regulate the situation of an armed conflict, i.e., the *ius in bello*, the law governing the resort

to force is called *ius ad bellum* (MacCoubrey and White 1992, p. 217).

The **Hague Law** was developed in two Hague Peace Conferences of 1899 and 1907. Since then, it has been amended and updated by various conventions, including by the 1977 First Protocol Additional (PA) to the 1949 Geneva Conventions (GC), and contains some "Hague elements." In substance, the Hague Conventions provide for three important principles still valid until today (Bailey 1972, p. 63): first, the *Martens Clause* according to which, notwithstanding the absence of specific regulations, "populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience" (cf. Preamble of the 1907 Hague Convention (IV)); second, the right to injure "the enemy is not unlimited" (Article 22 of the annexed Regulations of the 1899 and 1907 Hague Conventions II and IV); and third, the prohibition "[t]o employ arms, projectiles, or material of a nature to cause superfluous injury" (Art. 23 (e) 1899 Convention) or "calculated to cause unnecessary suffering" (Art. 23 (e) 1907 Convention).

The **Geneva Law** emerged from the Geneva Conventions of 1864, 1906, 1929, and 1949. It deals with the protection of noncombatants (civilians) and former combatants who are no longer willing or able to fight. The (modern) Geneva Law consists of the *four Geneva Conventions* (GC I–IV) of 12 August 1949 and the *three Additional Protocols* (AP I, II, III) of 18 June 1977 (AP I and II) and 8 December 2005 (AP III), respectively. This body of law constitutes the modern International Humanitarian Law (IHL) (Sassoli and Bouvier 2006, Part I, Ch. 3, pp. 121 ff.). Penal provisions can be found in the conventions that apply in the case of an international (armed) conflict (GC I–IV and AP I), while AP II contains no penal provisions whatsoever. This distinction reflects the *traditional "two-box approach"* differentiating between an international and a non-international armed conflict. It has been overcome with the seminal interlocutory decision of the *Tadic* Appeals Chamber (IT-94–1-AR 72) of the International Criminal Tribunal for the former Yugoslavia (ICTY) holding that IHL, in particular common Article 3 GC I–IV, provides for penal prohibitions in the case of a non-international conflict under certain conditions (para. 88 ff.).

As to the **grave breaches regime**, Articles 49 GC I, 50 GC II, 129 GC III, and 146 GC IV oblige the state parties to penalize conduct amounting to a grave breach of the GC, while Articles 50 GC I, 51 GC II, 130 GC III, and 147 GC IV contain the acts covered. The state parties have either to prosecute these acts before national courts or extradite those responsible to another state party (*aut dedere aut iudicare*). It is controversial whether the "grave breaches" norms provide for direct *individual criminal responsibility* given that they are only addressed to the state parties, obliging them to enact the respective penal prohibitions and ensure criminal prosecution. More detailed provisions concerning individual criminal responsibility can be found in AP I. Articles 86 and 87 AP I provide for *command or superior responsibility*. The superior may be liable for a failure to prevent crimes committed by his subordinates. In subjective terms at least a kind of negligence is necessary.

With regard to *grounds excluding responsibility* in general, the Geneva Law rejects the exclusion of criminal responsibility pursuant to a superior order implicitly. The recourse to *military necessity* is only possible in exceptional cases, namely, if the actions taken were necessary and proportional. This will rarely be the case if international crimes are committed. The *reprisal* defense was declared entirely unacceptable with a view to the protection of certain groups and objects (Prosecutor v. Martic, IT-95–11-R 61, 8.3.1996, para. 8 ff., 15 ff; Kupreškic et al. judgement, No. IT-95–16-T, 14.1.2000).

The Geneva Law also recognizes, at least partially, the *principles of legality and culpability*. Article 67 GC IV recognizes on the one hand the prohibition of retroactivity (*nullum crimen sine lege praevia*); on the other hand, it links the penalty to the offence thereby taking up the sentencing element of the principle of culpability (the punishment must conform to the actual culpability of the convicted person; see also Article 68 GC IV).

The Draft Codes of the International Law Commission and Private Initiatives

In 1947 the ILC was assigned to prepare a "Draft Code of Offences/Crimes against the Peace and Security of Mankind." The first Draft Code was adopted in the sixth session in 1954 (*Draft Code 1954*), which is composed of only four articles containing provisions on the general part and some criminal offences. After a definition of aggression was agreed upon by the GA in 1974, the ILC was again instructed to draft a code in 1981. The second Draft Code was adopted in 1991 (*Draft Code 1991*) and included provision concerning the general part and 12 offences, some of which show the political nature of the Draft Code 1991 (Articles 15–26). The third *Draft Code of 1996* basically rests upon the Draft Code of 1991. Great changes were undertaken in the special part, particularly by a sharp reduction of the former 12 offences to only 5: aggression, genocide, crimes against humanity, war crimes, and crimes against the UN and associated personnel. Apart from the last crime, this is exactly the catalogue of offences which was later included in the ICC Statute as the so-called core crimes. Prior to the 1996 Draft Code, in *1994*, the ILC had submitted a *Draft Statute* for an International Criminal Code (Ambos 2002/2004, pp. 444 ff.).

The various *unofficial proposals* for the development of ICL law may be subdivided into substantive law "Draft Codes" and procedural law "Draft Statutes." The most influential unofficial proposals have been the drafts of the Association Internationale de Droit Pénal (AIDP/ISISC/MPI 1995/1996) and of the International Law Association (Ambos 2002/2004, pp. 475 ff.) (ILA 1988).

The UN Ad Hoc Tribunals

The International Criminal Tribunal for the Former Yugoslavia

In reaction to massive violations of IHL and human rights in the former Yugoslavia beginning in 1991, the UN established the "United Nations Commission of Experts Pursuant to Security Council Resolution 780," which led to the establishment of the ICTY by the Security Council on 25 May 1993 (UN Doc. S/RES/827). The court was established as a subsidiary body of the Security Council according to Article 29 of the UN Charter. While the work of the court was initially supposed to last until the reestablishment of peace and security in the former Yugoslavia, the Security Council has in the meantime set up a so-called (first) **completion strategy** that fixed a time limit of 31 December 2004 for the end of investigations, 31 December 2008 for the end of trials of the first instance, and 31 December 2010 for the end of trials on appeal. On 22 December 2010, the Security Council adopted Resolution 1966 (2010) establishing the International Residual Mechanism for Criminal Tribunals with two branches, one for the ICTY and one for the ICTR; they commenced to operate on 1 July 2013 and 1 July 2012, respectively (UN SC-Res. 10141). According to the latest completion strategy, the tribunal continues its downsizing process (ICTY President 2012). All trials are expected to be completed by mid-2012, except for the case of *Radovan Karadžić*, which is expected to be completed in late 2013. Most appellate work is scheduled to be finished by the end of 2014 (ICTY President 2012). As of 19 November 2012, the ICTY has indicted 161 persons and concluded proceedings against 128 persons.

The **ICTY Statute** (ICTYS) contains 34 articles dealing with questions of substantive law (Articles 2–7, 24 ICTYS), procedural law (Articles 1, 8–10, 18–23, 25–30 ICTYS), as well as the organization of the tribunal (Articles 11–17, 31–34 ICTYS). Besides the Statute itself, several other legal instruments such as the Rules of Procedure and Evidence (RPE) have been adopted by the tribunal. It is composed of different chambers – the three Trial Chambers and the Appeals Chamber – the Office of the Prosecutor (OTP), and the Registry (Article 11 ICTYS). The chambers have 16 permanent judges and up to 12 so-called ad litem judges (Article 12(1) ICTYS) or ad hoc judges who can be appointed by the secretary general upon request of the President of the Tribunal (cf. Article 13*ter*, and *quarter* ICTYS). The OTP – though formally part of the tribunal – shall act independently as a separate

organ (Article 16 ICTYS). The Registry "serves" both the chambers and the Prosecutor (Article 11(c) ICTYS). One of its functions (Article 17 ICTYS) is also the setting up of an adequate defense for the accused, including the assigning and payment of council.

The tribunal's **jurisdiction** extends to all (natural) persons responsible for serious violations of IHL committed in the territory of the former Yugoslavia since 1991 (Articles 1, 6 ICTYS). According to Articles 2–5 of the ICTYS, the tribunal exercises jurisdiction *ratione materiae* over grave breaches of the four Geneva Conventions, violations of the laws, or customs of war, genocide, and crimes against humanity. The rule on **individual criminal responsibility** is provided for in Article 7 (1) ICTYS. It includes three groups of perpetrators: the politically responsible official person, the (military) superior, and the (committing) subordinate. Superior/command responsibility is laid down in Article 7 (3) ICTYS. Neither the official position of the accused nor the action pursuant to an order shall relieve a person from criminal responsibility (Article 7 (2), (4) ICTYS), but this fact may be considered in mitigation of punishment if the tribunal determines "that justice so requires" (Article 7 (4) ICTYS).

The International Criminal Tribunal for Rwanda

The International Criminal Tribunal for Rwanda (ICTR) was established by Security Council Resolution 955 of 8 November 1994. The **Statute** of the ICTR (ICTRS) resembles largely the one of the ICTY. Judges from both tribunals are assigned to the Appeals Chamber of both tribunals (Article 13 (3) ICTRS, Article 14 (4) ICTYS). The members of the Appeals Chamber of the ICTY shall also serve as the members of the Appeals Chamber of the ICTR (Article 13 (4) ICTYS).

The competence of the ICTR embraces the prosecution of serious IHL violations committed in the territory of Rwanda and by Rwandan citizens in the territory of neighboring states between 1 January 1994 and 31 December 1994 (Article 1 ICTRS). The ICTR exercises **jurisdiction**, similar to the ICTY, over genocide, crimes against humanity, and internal armed conflict crimes. Just like the ICTY, the ICTR introduced a first draft for a **completion strategy** in 2003, which has been continuously updated and developed since then. The latest report was submitted on 14 November 2012. The tribunal's president informed that "[a]s at 5 November 2012, the Tribunal has completed its work at the trial level with respect to 92 of the 93 accused. [. . .] The one remaining trial judgement will be delivered before the end of 2012, and appellate proceedings have been concluded in respect of 44 persons. The remaining appeals are projected to be completed by the end of 2014" (ICTR President 2012, Para. 3). As already mentioned above, the ICTR will then, like the ICTY, be transferred into the Residual Mechanism, which started operating on 1 July 2012.

The International Criminal Court

Negotiating History

In 1994 the UN General Assembly (GA) referred the ILC-Draft Statute to the "**Ad Hoc Committee** on the Establishment of an ICC." This committee presented a final report after two sessions in 1995. Then, the **Preparatory Committee** (PrepCom), established by the GA just after the 1995 report, took over. Its task was to prepare a draft for the Rome Conference to be held in 1998. The PrepCom held altogether six meetings from 25 March 1996 until 3 April 1998. Between the sessions, more or less formal meetings by states and delegations of states took place in order to smoothen out possible points of conflict before the actual sessions. In one of those "intersessionals," the important **Zutphen Report** was compiled (Sadat Wexler and Bassiouni 1998, pp. 7 ff., 129 ff.).

On 15 December 1997, the GA decided to arrange a **State Conference** for the establishment of the ICC in **Rome** (UN GA-Res 52/160). The conference was not only open to states but also for nongovernmental organizations (NGO). It commenced on 15 June 1998 and ended on 17 July 1998 with the adoption of the ICC Statute. One hundred fifty-nine governmental delegations

and 250 delegations of NGOs which had merged into the "Coalition for an ICC" attended the conference. Until the cessation of the conference, it was not entirely clear whether the ultimate goal – namely, the adoption of an ICC Statute – could be reached due to the opposition of important states. However, after intense negotiations, the conference adopted the court's Statute by a vote of 120 in favor to 7 against (USA, China, Libya, Israel, Iraq, Qatar, and Yemen), with 21 abstentions. A non-recorded vote was requested by the United States.

The Statute **entered into force** on **1 July 2002** after the deposit of the 60th instrument of ratification (cf. Article 126). Until the Statute was closed for signature on 31 December 2000, 139 states had signed, and as of 21 July 2012, 121 states had ratified it. Following the Rome Conference, a **Preparatory Commission** (PrepCommis) was first established in order to compile further legal instruments and to prepare the first meeting of the Assembly of State Parties (ASP). In addition, a working group on the crime of aggression was set up in order to reach a consensus on the definition and the conditions of jurisdiction pursuant to Article 5 (2) ICC Statute (on the final agreement see below).

The Rome Statute, the Structure of the Court, and Other Legal Instruments

General

The ICC Statute consists of 13 parts and 128 articles. The ICC was established as a permanent institution in The Hague (Articles 1, 3). While it is not an organ of the UN, it is linked to the latter by a "relationship agreement" (Article 2). The court is made up of a Presidency, a Pre-Trial Chamber (PTC), a Trial Chamber (TC) and an Appeals Chamber (AC), an Office of the Prosecutor (OTP), and a Registry (Article 34). Although the defense is not an organ of the court, an Office of Public Counsel was set up at the Registry (Regulation 77 of the Regulations of the Court).

The Judges

The judges are elected from two lists (Article 36 (5)): list A shall consist of candidates with established competence in criminal law and procedures and the necessary relevant experience. List B shall consist of candidates with established competence in relevant areas of international law, such as IHL and human rights law, and extensive experience in a professional legal capacity which is of relevance to the judicial work of the court. Especially the List B requirements are increasingly criticized for allowing diplomats without criminal trial experience to become ICC judges (Bohlander 2009, pp. 532 ff.; Ambos 2012a, pp. 224 ff). In addition, the candidates shall be of "high moral character, impartiality and integrity" and "possess the qualifications of their national law for appointment to the highest judicial offices" (Article 36 (3)(a) ICC Statute). They must be fluent in English or French (Article 36 (3)(c)). The judges shall be selected by lot to serve 3, 6, or 9 years (Article 36 (9)(b)). Only the judges elected for a term of 3 years are eligible for reelection (Article 36 (9)(c)). Judicial impartiality shall be secured by not engaging in any other occupation of professional nature. The judges shall represent the main legal systems of the world. On 10 September 2004, the ASP adopted a resolution on the "Procedure for the Nomination and Election of Judges of the International Criminal Court" (ICC-ASP/3/Res.6), providing for rules for the nomination and election of the judges. However, the respective "Advisory Committee on Nominations" of the ASP (Article 36 (4)(c) ICC Statute) has not yet been established. Updated information on the judges and chambers can be found at http://www.icc-cpi.int/.

The Office of the Prosecutor

The OTP shall act independently and as a separate organ of the court. Its first head was the Argentinean Luis Moreno Ocampo – he is "The Prosecutor" (cf. Article 42). He was elected on 21 April 2003 and took office on 16 June 2003. The ASP also elected two deputy prosecutors, Mr. Serge Brammertz (Investigations) and Mrs. Fatou Bensouda (Prosecutions). After Brammertz' leave to Lebanon in January 2006 (before assuming the Office of Chief Prosecutor of the ICTY in January 2008), only Fatou Bensouda from Gambia stayed

on as deputy prosecutor. On 1 February 2011, Phakiso Mochochoko from Lesotho was appointed as Head of the Jurisdiction, Complementarity, and Cooperation Division. The actual power now resides in an Executive Committee, composed of the prosecutor, the deputy prosecutor, and the heads of the different sections (JCCD, Investigation, and Prosecution) and supported by some external consultants. Moreno Ocampo's mandate expired in June 2012. Deputy Prosecutor Bensouda was elected unanimously as his successor on 12 December 2011 at the ASP's tenth session. She took office on 16 June 2012.

The OTP concluded several agreements with other organizations and persons, i.e., with the International Criminal Police Organization (Interpol) or so-called intermediaries (Ambos 2011a, p. 329 fn. 123). It cooperates with the UN by virtue of the UN-ICC agreement and the MONUC-Memorandum of the ICC.

Registry and Assembly of States Parties

The Registry is responsible for the administration and servicing of the court and is headed by the Registrar (Silvana Arbia, Italy, successor of Bruno Cathala, France). It consists of the Immediate Office of the Registrar, the Security and Safety Section, the Common Administrative Services Division, the Division of Court Services, Public Information, the Documentation Section, and the Division of Victims and Counsel (Lachowska 2009, p. 389).

The Assembly of States Parties (Article 112) is composed primarily of representatives of the states that have ratified and acceded to the Rome Statute. Other states, which have signed the Statute or the Final Act, may be observers in the Assembly (Article 112 (1)). The ASP is supposed to meet annually and can be seen as the decision-making organ of the court. It decides on various issues, such as the interpretation/application of the Statute, the adoption of legal texts and of the budget, and the election of the judges, the prosecutor, and the deputy prosecutor(s). Any dispute between two or more States Parties relating to the interpretation or application of the Statute shall be referred to the ASP. The Assembly may itself seek to settle the dispute or may make recommendations on further means of settlement of the dispute, including referral to the International Court of Justice in conformity with the Statute of that court (Article 119 (2)). The ASP shall also supervise the State Parties' compliance with their cooperation obligations under the Statute (Article 112 (2)(f) in relation to Article 87 (5)(b) and (7)). It is the only enforcement organ of the Rome system in this respect, unless the UN Security Council has referred a situation to the court (Article 87 (5)(b) *in fine*).

Legal Sources

Apart from the Statute itself, the two most important secondary legal sources of the ICC are the Elements of Crimes and the Rules of Procedure and Evidence, which were agreed on in the fifth session of the PrepCommis and were finally adopted in the first session of the ASP. In accordance with Article 9 ICC Statute, the **Elements of Crimes** shall assist the court in the interpretation and application of the core crimes (Articles 6, 7, 8, and 8*bis*). They are a subsidiary source, have to be consistent with the Statute (Article 9 (3)), and are subject to legal interpretation of the court. The **Rules of Procedure and Evidence** (RPE) complement the procedural regime of the Statute and have to be consistent with it (Article 51 (4)). In the event of conflict, the latter shall prevail (Article 51 (5)). They do not affect the procedural rules for any national court or legal system for the purpose of national proceedings.

On 26 May 2004, the judges adopted the **Regulations of the Court** (Article 52). The regulations were developed to fulfill the goal of speedy trials and to secure a fair trial for the accused. Regulations of the Registry and of the OTP were adopted on 3 March 2006 and 23 April 2009, respectively. Another legal source is the **Agreement on Privileges and Immunities** (Article 48) which grants certain immunities and privileges to the judges, the prosecutor and its staff, the Registrar and its staff, as well as to counsels, experts, witnesses, or any other person required to be present at the seat of the court. Other legal sources are inter alia: the Code of Professional Conduct for Counsel, the Code of Judicial Ethics, the agreement between the International

Criminal Court and the United Nations, and the agreement with the EU on cooperation and assistance.

The First Review Conference in Kampala

The first Review Conference took place in Kampala, Uganda, from 31 May to 11 June 2010, a year later than envisaged by Article 121. ICC States Parties, observer states, international organizations, NGOs, and other participants discussed proposed amendments to the Rome Statute and took stock of its impact to date. Debates focused on the impact on victims and affected communities, complementarity, cooperation, and peace and justice. Apart from these general discussions, the conference took three concrete decisions: most importantly, the crime of aggression was defined (Art. 8*bis* Statute), and the conditions for the exercise of jurisdiction were agreed upon (Arts. 15bis, 15ter; Ambos 2011b). Further, the deletion of Article 124 was rejected, and the war crimes of Article 8 (2)(b)(xvii) to (xix) were extended to a non-international armed conflict (Article 8 (2)(e)(xiii) to (xv) Ambos 2011a, p. 125).

Current Investigations

The OTP first initiated investigations mid-2004 with regard to two **situations**. One was the situation in **Uganda** closely connected with the activities of the so-called Lord's Resistance Army (LRA); the other one concerns crimes committed on the territory of the **Democratic Republic of Congo** (DRC) since 1 July 2002. In both situations, the respective governments made use of the possibility of a state referral in accordance with Article 13 (a), 14 ICC Statute. In June 2004 (DRC) and July 2004 (Uganda), the OTP determined that there is a reasonable basis to open a formal investigation into the situations (Article 53 (1)). On 21 December 2004, the **Central African Republic** (CAR) referred a situation to the ICC and requested an investigation by the OTP into the crimes committed on its territory since 1 July 2002. The OTP opened investigations on 22 May 2007. On 13 July 2012, Mali has self-referred a further situation to the ICC through the use of Article 14 ICC Statute. On 16 January 2013, the OTP formally opened an investigation into the alleged crimes, committed on the territory of Mali since January 2012.

While in all these situations, the court's jurisdiction was triggered by state self-referrals under Article 13 (a), 14 of the Statute, two other situations have been referred to the court by the Security Council pursuant to Article 13 (b), namely, **Darfur** (**Sudan**; Resolution 1593 of 31 March 2005) and **Libya** (Res. 1970 of 26 February 2011). The Darfur investigation was formally opened in June 2005. In the Libyan situation, the prosecutor announced very quickly, on 3 March 2011, that he will open a formal investigation. Last but not least, with regard to **Kenya**, the prosecutor acted for the first time proprio motu pursuant to Articles 13 (c) and 15. On 31 March 2010, Pre-Trial Chamber II authorized the prosecutor to open the investigation pursuant to Article 15 (4) with regard to crimes against humanity committed between 1 June 2005 and 26 November 2009.

Apart from these formal trigger mechanisms, Article 12 (3) offers non-States Parties the possibility to accept the jurisdiction of the court by a kind of **ad hoc declaration** "with respect to the crime in question." So far, this provision has been invoked in two cases. First, on 15 February 2005, **Ivory Coast** accepted the jurisdiction of the ICC with respect to alleged crimes committed from 19 September 2002, which was renewed by both the former President Laurent Gbagbo and current President Alassane Ouattara. On 3 October 2011, Pre-Trial Chamber III authorized the prosecutor to open an investigation into war crimes and crimes against humanity allegedly committed following the presidential election of 28 November 2010. On 22 February 2012, this authorization was expanded to include crimes allegedly committed between 19 September 2002 and 28 November 2010. Secondly, on 22 January 2009, the **Palestinian National Authority** lodged a declaration with regard to acts committed on the territory of Palestine since 1 July 2002, especially during the 2008/2009 Gaza war.

If an investigation is formally opened, several **cases** arise from the respective situation, persons are targeted, and, if they do not voluntarily

surrender to the court, **arrest warrants** are issued. So far, arrest warrants have been issued regarding the situations in Uganda (five arrest warrants issued, one case terminated because of the death of the suspect), Dafur/Sudan (five), DRC (seven, of which five have been executed), Libya (three, one case terminated), CAR (one issued and executed), and Côte d'Ivoire (two issued and one executed). Thus, only seven arrest warrants have been executed so far. Trials started against *Jean-Pierre Bemba Gombo*, *Thomas Lubanga Dyilo*, *Germain Katanga*, and *Mathieu Ngudjolo Chui*. The first judgement was delivered on 14 March 2012 in the Lubanga case (Ambos 2012b). On 10 July 2012, Lubanga was sentenced to a total period of 14 years of imprisonment. On 18 December 2012, Mathieu Ngudjolo Chui was found not guilty of the charges brought against him. In other cases summons to appear have been considered sufficient, and the suspects appeared voluntarily before the court, i.e., in the case of Abu Garda, a member of the Sudanese Janjawid militia (ICC-02/05-02/09-2), Banda Abakaer Nourain (ICC-02/05-03/09-3) and Mohammed Jerbo Jamus (ICC-02/05-03/09-2) and with regard to the so-called Ocampo six (Prosecutor v. Kirimimuthaura et al. ICC-01/09-02/11-01, Samoeiruto et al. ICC-01/09-01/11-01).

The "Mixed" Tribunals

The Legal Bases

As a result of the increasing internationalization of the prosecution of serious human rights violations, many so-called mixed or hybrid tribunals have been established in several states. These tribunals have a mixed national-international legal basis and recruit national and international (foreign) prosecutors and judges. The tribunals are either part of a **transitional UN administration** (Kosovo, East Timor), or based on a **bilateral agreement with the UN** (Sierra Leone, Cambodia, Lebanon), or on legislative provisions adopted by an **occupying power** (Iraq). A purely **national tribunal** for international crimes was created in Bangladesh.

Kosovo and East Timor

Following the armed conflict between Serb authorities and the Kosovo Liberation Army, **Kosovo** was placed under the interim administration of the UN on 10 June 1999. The competence of the transitional UN administration (UNMIK) in Kosovo for "maintaining civil law and order" and the representation by a Special Representative of the Secretary-General (SRSG) derive from SC-Res. 1244 (6) and (11 (i)). On this basis numerous "regulations" and "administrative directions" have been enacted in order to define the applicable law (Bohlander 2003, pp. 24 ff.). As a result, the Provisional Institutions of Self-Government (PISG) were established, including a government, a president, a parliament, and a court system. As to criminal justice, there exist three instances for the adjudication of crimes of international concern. The attempt to set up a special tribunal ("Kosovo War Crimes and Ethnic Crimes Court") failed; instead, international prosecutors and judges were assigned to all district courts in a 2:1 proportion (two international, one local judge). The court system includes a constitutional court, a supreme court, five district courts, a commercial court, 25 municipal courts, 25 minor offence courts, and an appellate court for minor offences. Shortly before the declaration of independence by Kosovo on 17 February 2008, the European Union Rule of Law Mission in Kosovo (EULEX) was established in February 2008 to "monitor, mentor, and advise Kosovo institutions in all areas related to the rule of law and to investigate, prosecute, adjudicate, and enforce certain categories of serious crimes" (Article 3(a) and (d) of the Council Joint Action 2008/124/CFSP). Through EULEX, 31 international judges and 15 international prosecutors support local judges and prosecutors. There is one state public prosecutor's office, five district prosecutors' offices, and seven municipal prosecutors' offices. EULEX exercises its executive authority over a special prosecutor's office, which includes eight international prosecutors, and focuses on serious crimes including human trafficking, money laundering, war crimes, and terrorism.

In **East Timor** – in accordance with SC-Res. 1272 (1999) of 25 October 1999 – the "United Nations Mission in East Timor"

(UNTAET) had overall responsibility for the administration of East Timor and was empowered to exercise all legislative and executive authority, including the administration of justice. The organization of the courts in East Timor was reorganized, and at the same time, panels with exclusive jurisdiction over serious criminal offences (genocide, war crimes, crimes against humanity, murder, sexual offences, and torture) were established within the District Court in Dili. The panels had exclusive jurisdiction only for offences committed in the period between 1 January 1999 and 25 October 1999 and operated until 20 May 2005. Moreover, within the Office of the General Prosecutor, there was a "Special Prosecutor" called the Deputy General Prosecutor for Serious Crimes, who headed the Serious Crimes Unit (Othman 2003, pp. 87 ff.). After the independence on 20 May 2002, the laws launched by UNTAET remained in effect, and the judges have been appointed by the Supreme Council of the Judiciary of East Timor (von Braun 2008, pp. 137 ff.). In 2005, the mandate of the Serious Crimes Unit expired, and its investigative functions were resumed by the Serious Crimes Investigation Team, assisting the Office of the Prosecutor-General of East Timor. Until the expiration of its mandate, the Serious Crimes Unit had indicted 391 people. While 84 defendants were convicted, three were acquitted in trials before the special panels, and more than 300 indictees remained at large, almost all of them in Indonesia.

Sierra Leone

In SC-Res. 1315 (2000), the Security Council asked the UN-Secretary General to negotiate an agreement with the Government of Sierra Leone to create an independent court to prosecute persons responsible for the commission of serious violations of IHL and crimes committed under Sierra Leonean law during the country's civil war. On 16 January 2002, such an agreement, accompanied by a Statute of the Special Court for Sierra Leone (SCSLS), was signed (Kelsall 2009, pp. 254 ff.). Thus, the court is based upon a bilateral, international law agreement between an international organization and a state and not solely upon a UN SC Resolution. It has the power to prosecute persons who bear the greatest responsibility for serious violations of IHL and Sierra Leonean law committed in the territory of Sierra Leone since 30 November 1996 (Article 1 SCSLS). This covers precisely crimes against humanity, war crimes, certain other international crimes, and certain crimes under Sierra Leonean law (Articles 2–5 SCSLS). The Special Court and the national courts shall have concurrent jurisdiction. The Special Court shall have primacy over the national courts of Sierra Leone and may at any stage of the procedure request a national court to defer to its competence in accordance with the SCSLS and the Rules of Procedure and Evidence (Article 8 SCSLS). In 2003, the prosecutor issued 13 indictments, of which two were withdrawn due to the deaths of the accused in December 2003. Thus far, the trials of three former leaders of the Armed Forces Revolutionary Council (AFRC), of two members of the Civil Defense Forces (CDF), and of three former leaders of the Revolutionary United Front (RUF) have been completed, including appeals. On 31 October 2009, the Special Court's eight convicted people were transferred to Mpanga Prison, Rwanda for sentence enforcement. The judgement against former Liberian President *Charles Taylor* was rendered on 26 April 2012. In a unanimous judgement (the "dissenting opinion" of the alternate Judge El Hadji Malick Sow from Senegal, expressed after the verdict had been delivered, does not count as a vote), Taylor was found guilty on all accounts of aiding and abetting the RUF and AFRC rebel groups and/or Liberian fighters operating in Sierra Leone and of having planned attacks on civilians. On 30 May 2012, he was sentenced to a term of 50 years imprisonment. Upon the delivery of the final judgment in the Charles Taylor case, the Special Court's mandate will be complete and the SCSL closes. However, many of its legal obligations will not terminate with the conclusion of all cases. In August 2010, the United Nations and the Government of Sierra Leone agreed to establish a Residual Special Court for Sierra Leone, which will be responsible for fulfilling the Special Court's obligations. The RSCSL Agreement and the RSCSL Statute were ratified by the Parliament of Sierra Leone in December 2011.

The RSCSL shall, pursuant to Article 1.1 of its Statute: "[. . .] maintain, preserve and manage its archives, including the archives of the Special Court; provide for witness and victim protection and support; respond to requests for access to evidence by national prosecution authorities; supervise enforcement of sentences; review convictions and acquittals; conduct contempt of court proceedings; provide defence counsel and legal aid for the conduct of proceedings before the Residual Special Court; respond to requests from national authorities with respect to claims for compensation; and prevent double jeopardy." At the beginning of 2012, the Registrar established the Residual Special Court for Sierra Leone Transition Working Group in order to coordinate work relating to the transition to RSCSL and closure of the Special Court (SCSL Ninth Annual Report 2011-2012, p. 7). The RSCSL will have its interim seat in The Hague and an office in Sierra Leone. It will be headed by a president (chosen by RSCSL judges), a prosecutor and a registrar.

Cambodia

In Cambodia, after long-lasting negotiations, a bilateral agreement with the UN was signed on 06 June 2003 "concerning the prosecution in Cambodian law of crimes committed during the period of Democratic Kampuchea" (during the period from 17 April 1975 to 6 January 1979). Parallel to the negotiations with the UN, a "Law on the Establishment of the Extraordinary Chambers" was prepared in order to prosecute these crimes by national institutions. The law was first adopted on 10 August 2001 and later amended (ECCC-Law). The extraordinary chambers have the power to bring trials against suspects who committed genocide, crimes against humanity, grave breaches of the Geneva Conventions, as well as certain other enumerated international and national crimes (Articles 2–8 ECCC-Law). The chambers are established within the existing court structure; they operate as a court of first instance, the Supreme Court being an appellate court and final instance (Kashyap 2003, pp. 192 ff.).

The extraordinary chambers are unique in structure and composition. The Trial Chamber is composed of five professional judges, of whom three are Cambodian (with one as president) and two foreign (Article 9 (1) ECCC-Law). The Appeals Chamber is composed of seven judges, of whom four are Cambodian (with one as president) and three foreign (Article 9 (2) ECCC-Law). While the courts in East Timor and Sierra Leone take their decisions with a simple majority, in the ECCC national judges dominate and a supermajority (a simple majority plus one) is necessary (Article 14 ECCC-Law). The judges agreed on rules of procedure on 12 June 2007.

Personal jurisdiction is limited since only the *senior leaders of Democratic Kampuchea and those who were most responsible* shall be brought to trial. Those leaders are former heads of the Khmer Rouge (Heder and Tittemore 2004). In Case 001, *Kaing Guek Eav alias Duch* was sentenced – after an appeal by the co-prosecutors – to life imprisonment. In Case 002, the trial against *Nuon Chea*, *Ieng Sary*, *Khieu Samphan*, and *Ieng Thirith* began on 27 June 2011. On 4 March 2013, Ieng Sary died in hospital from natural causes, as the Co-Prosecutors announced on 2 April 2013. On 7 September 2009, the international co-prosecutor filed two Introductory Submissions, requesting the co-investigating judges to initiate investigation of five additional suspected persons. These two submissions have been divided into what is known as Case files 003 and 004.

Iraq

The Iraqi Special Tribunal was established by the US Coalition Provisional Authority; its Statute was issued on 10 December 2003 by the Iraqi Governing Council and approved on 18 October 2005 by the first freely elected Parliament. The name of the court was changed to the "Iraqi Higher Criminal Court," which is now financed exclusively by the Iraqi government. The tribunal is not part of the regular Iraqi judicial system but an autonomous organ with its own rules and an own administrative capacity. The tribunal has jurisdiction over any Iraqi national or resident of Iraq accused of the core crimes (Articles 11–13,) committed since 17 July 1968 (takeover of the Ba'ath party) and up until and including

1 May 2003 (official ending of acts of war) in the territory of the Republic of Iraq, or elsewhere, including crimes committed in connection with Iraq's wars against the Islamic Republic of Iran and the state of Kuwait (Article 1(b)). Moreover, the tribunal has the power to prosecute certain violations of Iraqi laws (Article 14). The tribunal consists of the Tribunal Investigative Judges, one or more Trial Chambers, and an Appeals Chamber (Article 3). The Statute resembles the ICC Statute in its substantive law and procedural law provisions.

Besides former president *Saddam Hussein* (who was sentenced to death on 5 November 2006 and executed on 30 December 2006), several high-ranking Iraqi officials were tried, among others in the following cases: the *Al-Dujail* case against eight accused, which started on 19 October 2005 (all eight convicted, four sentenced to death – one after appeal – 3–15 years imprisonment, and one defendant was acquitted), and the *Al-Anfal* case, which started on 24 June 2007, against members of the former Ba'th regime (three sentenced to death, two to life imprisonment, and one acquitted). The third trial started in August 2007, relating to the brutal crushing of a Shiite rebellion in 1991 (three of the defendants were acquitted, four were sentenced to life imprisonment, six were given a long prison sentence, and two were sentenced to death). The fourth trial dealt with the execution of 42 merchants who were accused of raising their prices during the period when UN sanctions had been imposed against Iraq (two defendants were sentenced to death, one to life, 3–15 years and 1–6 years imprisonment, one defendant was acquitted). Further proceedings were initiated in the following cases: against 14 accused (inter alia against *Ali Hassan al-Majid*, *Hashim Hassan al-Majid*, *Tarik Aziz Issa*, members of the former Ba'th regime, and of the militia) for the deportation and forced movement of families in 1984. The indictment was issued on 4 November 2008, and the judgement was rendered on 2 August 2009. In another case, an indictment was issued against 14 persons (inter alia against the former minister for the interior *Sadun Shaker*) for the killing and forced displacement of Falili-Kurds. The judgement was rendered on 29 November 2010. Furthermore, in the *Al Jeboor* case, four accused were found guilty of crimes against humanity.

Lebanon and Bangladesh

The **Special Tribunal for Lebanon** (STL) was created by SC-Res. 1757 (2007) of 30 May 2007. The provisions of the document annexed to it, and the Statute of the Special Tribunal thereto attached, entered into force on 10 June 2007. The STL is based in The Hague and is neither a subsidiary organ of the UN nor is it a part of the Lebanese court system. Rather, it supersedes the national courts within its jurisdiction (Article 4 (1)). It is a hybrid court in the sense that it is composed of both national and international judges. The STL is unique in that the applicable law is national in character, while the ICTY and ICTR are limited to prosecuting crimes in violation of international law and the (other) hybrid tribunals prosecute crimes under both domestic and international law. In addition, the STL is the first UN-assisted tribunal to combine substantial elements of both a common law and a civil law legal system (U.N. Doc. S/2006/893).

According to Article 1 of the Statute of the Special Tribunal for Lebanon (STL), "[T]he Special Tribunal shall have *jurisdiction* over persons responsible for the attack of 14 February 2005 resulting in the death of former Lebanese Prime Minister Rafic Hariri and in the death or injury of other persons." However, the temporal jurisdiction was extended to include other attacks bearing the same, or similar, characteristics of the Hariri assassination (U.N. Doc. S/2006/893). As to the substantive law, the Statute stipulates that the tribunal shall apply provisions of the Lebanese Criminal Code relating to the prosecution and punishment of acts of terrorism (a crime that so far has not been within the province of an international tribunal) and crimes and offences against life and personal integrity, among others. Following an indictment by former prosecutor Daniel Bellemare on 17 January 2011, on 30 June 2011, the tribunal issued four arrest warrants that have not been executed so far. The STL's first case is *The Prosecutor v. Ayyash,*

Badreddine, Oneissi and Sabra. In this case, which awaits completion of the pre-trial phase, the Prosecutor accuses Ayyash, Badreddine, Oneissi and Sabra of criminal responsibility for the attack that killed former Lebanese Prime Minister Rafik Hariri and others on 14 February 2005 (STL President 2012-2013, p. 7). On 1 February 2012, the Trial Chamber had made an order for trial in absentia, which was confirmed by the Appeals Chamber on 1 November 2012 (STL President 2012-2013, p. 10). On 19 July 2012, the Pre-Trial Judge set 25 March 2013 as a tentative date for trial to start. However, many procedural problems, including incomplete disclosure and technical issues faced by the Defence in accessing certain disclosed material, caused the Pre-Trial Judge to postpone the commencement of trial (STL President 2012-2013, pp. 7, 9).

The jurisdiction of the **Bangladesh Tribunal** includes crimes against humanity, crimes against peace, genocide, war crimes, and "violation of any humanitarian rules applicable in armed conflicts laid down in the Geneva Conventions of 1949" and "any other crimes under international law" Article 3 (2) (a)-(f) International Crimes (Tribunals) Act 1973 (ICTA). Common law and customary international law are treated as primary sources of law, and the tribunal resembles the existing tribunals, albeit it conducts a purely domestic process (Linton 2001, pp. 221 ff.). On 20 November 2011, the first person was charged (Delwar Hossain Sayedee, a leader of Jamaat-e-Islami, an Islamist party opposed s Bangladesh's independence). Subsequently, the charges were "framed" (i.e., confirmed) by the tribunal. Further, the chief investigator Abdul Hannan Khan carries out investigations against 10 other suspects, including another six members of the Jamaat and two of the Bangladesh Nationalist Party. Apart from Sayedee, charges have been filed against Maulana Matiur Rahman Nizami, Ali Ahsan Mohammad Mujahid, Abdul Kader Mollah, Abdul Alim, Muhammad Kamaruzzaman, and Salahuddin Quader Chowdhury. Charges were framed against Kamaruzzaman on 4 June 2012 (the trial started on 2 July 2012), Nizami and Mollah on 28 May 2012, and Chowdhury on 4 April 2012.

Comparative Analysis

All these tribunals can be characterized as "mixed" not only because of their composition but also because of their organization, structure, and the applicable law. The tribunals apply **national and international law**. While there was first, in line with the ICTY and ICTR precedent, a certain preference for the common law system, in particular in terms of the applicable procedure, the ECCC introduced an inquisitorial-like French procedure, and the STL, as the first UN tribunal, combines elements of both legal systems.

It is common to all tribunals (with the exception of the STL) that they are **situated** in the state where the crimes of their subject matter jurisdiction took place. Thereby, a certain proximity to the local, crime-affected population is ensured. Either the tribunals are part of the local justice system (Kosovo, East Timor, Cambodia) or though special tribunals, somehow affiliated with the national system (Sierra Leone, Iraq, Lebanon). Contrary to the ICTY and ICTR, which have had no less than 15 years to terminate their proceedings, the mixed tribunals have a significantly **shorter time period** to conclude their work. For instance, the SCSL originally had only 3 years to fulfill its mandate. The problems become even more apparent if one looks at the **organizational problems** and the **tight resources** compared to the high operative cost they are bound to combat. Their budgets are remarkably lower than the ones of the ad hoc tribunals.

The crimes falling within the subject matter **jurisdiction** of the mixed tribunals are the **core crimes** genocide, crimes against humanity, and war crimes. The elements of the crime of these core crimes are related to the ICC Statute, but war crimes are in the majority of cases not as detailed codified as in Article 8 ICC Statute. In addition, all tribunals apply **specific violations of national law** depending on the situation: in East Timor, torture was added as an offence; in Kosovo incitement to national, racial, religious, or ethnic hatred, discord or intolerance ((1) of Regulation 4 (2000)), the illegal possession of weapons ((8) of Regulation 7 (2001)), and unauthorized border crossing ((3) of Regulation 10 (2001)) are criminal offences; in Cambodia, the destruction of cultural property during armed conflict can be

prosecuted; in Sierra Leone, offences relating to the abuse of girls and setting fire to dwelling houses were included; and in Iraq, the wastage of national resources is a violation, and the STL even applies exclusively national law. Furthermore, in Sierra Leone, adolescents between 15 and 18 years old can be brought to trial (Article 7 SCSLS).

The existence of **extrajudicial mechanisms** for dispute resolution, for instance, through **truth and reconciliation commissions** (Bassiouni 2003, pp. 711 ff.), leads to concurrent jurisdictions or at least entails difficulties of delimitation. In Sierra Leone, overlapping jurisdiction was meant to be avoided by prosecution of only the persons most responsible by the Special Court itself, basically leaving child and adolescents to the competence of the truth commission. The legitimacy of the Iraqi Court has been highly criticized from the outset due to its establishment by the US occupying power and its legal source (Megally and Zyl 2003). While the acceptance of the court by the Iraqi people may indeed be questioned, this fact is not an anomaly but lies at the heart of an international criminal justice system dominated by ad hoc tribunals. In fact, in this respect, all these tribunals face a **dilemma**: on the one hand, the national judiciary is generally not able and very often unwilling to carry out proceedings for internationalized core crimes; on the other hand, the "internationalization" of the courts and procedures gives rise to a deficit in its legitimacy in relation to the local population.

Related Entries

- ▶ Genocide
- ▶ International Responses to Victims in Criminal Justice
- ▶ Victims and the International Criminal Court

Recommended Reading and References

Ahlbrecht H (1999) Geschichte der völkerrechtlichen strafgerichtsbarkeit im 20. Jahrhundert. Nomos, Baden-Baden

Ambos K (2002/2004) Der Allgemeine Teil des Völkerstrafrechts – Ansätze einer Dogmatisierung. Duncker & Humblot, Berlin

Ambos K (2011a) Internationales Strafrecht. C.H. Beck, München

Ambos K (2011b) The crime of aggression after Kampala. Ger Yearb Int Law 53: 463–509

Ambos K (2012a) International courts and their judges. Book review. Crim Law Forum 23:223–228

Ambos K (2012b) The first judgment of the International Criminal Court (Prosecutor v. Lubanga): a comprehensive analysis of the legal issues. Int Crim Law Rev 12:115–153

Ambos (2013) Treatise on International Criminal Law, Vol. I: Foundations and General Part, Ch. I. Oxford University Press, Oxford

Association International de Droit Pénal, Instituto, Superiore Internazionale di Scienze Criminali, Max-Planck-Institut für ausländisches und internationales Strafrecht (1995/1996) Siracusa-Draft and updated Siracusa-Draft

Bailey SD (1972) Prohibitions and restraints in war. Oxford University Press, London

Bassiouni MC (2003) Introduction to international criminal law. Transnational Publishers, Ardsley

Bassiouni MC (1999) Crimes against humanity in international criminal law, 2nd ed. Kluwer Law International, The Hague

Bohlander M (2009) Pride and prejudice or sense and sensibility? A pragmatic proposal for the recruitment of judges at the ICC and other international criminal courts. New Crim Law Rev 12:529–542

Bohlander M (2003) The legal framework of the prosecution and the courts. In: Ambos K, Othman M (eds) New approaches in international criminal justice – Kosovo, East Timor, Sierra Leone and Cambodia. ed. iuscrim, Freiburg i. Brsg, pp 21–60

Cryer R (2005) Prosecuting international crimes. Cambridge University Press, Cambridge

Engelhart M (2004) Der Weg zum Völkerstrafgesetzbuch – Eine kurze Geschichte des Völkerstrafrechts. Jura, pp 734–743

Heder S, Tittemore BD (2004) Seven candidates for prosecution – accountability for the crimes of the Khmer Rouge, 2nd edn. Documentation Center of Cambodia, Phnom Penh

ICTR President (2012) Letter dated 22 May 2012 from the President of the International Criminal Tribunal for Rwanda addressed to the President of the Security Council. S/2012/836. 14 November

ICTY President (2012) Letter dated 16 November 2012 from the President of the ICTY, addressed to the President of the Security Council. S/2012/847. 19 November

International Law Commission (1950) Yearbook of the International Law Commission, 1950, Vol. II, Report of the International Law Commission covering its second session, 5 June – 29 July 1950, pp. 364 ff.,

Available at http://untreaty.un.org/ilc/publications/yearbooks/1950.htm.
International Law Reports (1968), pp. 5 ff.; (1995), pp. 338 ff., 357 ff., 520 ff.
Kashyap S (2003) The framework of the prosecution in Cambodia. In: Ambos K, Othman M (eds) New approaches in international criminal justice – Kosovo, East Timor, Sierra Leone and Cambodia. Ed. Iuscrim, Freiburg i. Brsg, pp 189–205
Kelsall T (2009) Culture under cross-examination. Cambridge University Press, Cambridge
König K-M (2003) Die völkerrechtliche Legitimation der Strafgewalt internationaler Strafjustiz. Nomos, Baden-Baden
Lachowska A (2009) The support work of the Court's registry. In: Doria J, Gasser H-P, Bassiouni MC (eds) The legal regime of the international criminal court – essays in honour of Professor Igor Blishchenko. Nijhoff, Leiden, pp 387–400
Lemkin R (1933) Akte der Barbarei und des Vandalismus als delicta juris gentium. Int Anwaltsblatt 18:117–119
Linton S (2001) Cambodia, East Timor and Sierra Leone. Crim Law Forum 12:185–246
MacCoubrey H, White ND (1992) International law and armed conflict. Dartmouth, Aldershot
Megally H, Zyl Pv (2003) U.S. justice with an Iraqi face? Int Her Tribune 4
Osten P (2003) Der Tokioter Kriegsverbrecherprozeß und die japanische Rechtswissenschaft. Berliner Wissenschaftsverlag, Berlin
Othman M (2003) The framework of prosecutions and the court system in East Timor. In: Ambos K, Othman M (eds) New approaches in international criminal justice – Kosovo, East Timor, Sierra Leone and Cambodia. ed. Iuscrim, Freiburg im Breisgau, pp 85–112
Puttkamer EV (1949) Die Haftung der politischen und militärischen Führung des 1. Weltkrieges für die Kriegsurheberschaft und Kriegsverbrechen. Arch Völkerrechts 1:424–449
Sadat Wexler LN, Bassiouni MC (1998) Model draft statute for the International Criminal Court based on the Preparatory Committee's text to the diplomatic conference. Nouvelles Études Pénales, Rome
Sassoli M, Bouvier A (2006) How does law protect in war? – cases, documents and teaching materials on contemporary practice in international humanitarian law. Vol. 1: outline of international humanitarian law. Vol. 2: cases and documents, 2nd edn. International Committee of the Red Cross, Geneva
Schabas W (2009) Genocide in international law, 2nd edn. Cambridge University Press, Cambridge
STL President (2012-2013) Special Tribunal for Lebanon, Fourth Annual Report 2012-2013
United Nations War Crimes Commission (1947–1949). Law Reports of Trials of War Criminals, Selected and prepared by the United Nations War Crimes Commission 1947–1949, Available at http://www.loc.gov/rr/frd/Military_Law/law-reports-trials-war-criminals.html. vol 1–15
von Braun L (2008) Internationalisierte Strafgerichte – eine Analyse der Strafverfolgung schwerer Menschenrechtsverletzungen in Osttimor, Sierra Leone und Bosnien-Herzogowina. Berliner Wissenschaftsverlag, Berlin

Development of Sex Offending

▶ Criminal Career of Sex Offenders

Development of the UCR and the NCVS

Lynn A. Addington
Department of Justice, Law, and Society, American University, Washington, DC, USA

Overview

The importance of the Uniform Crime Reporting Program (UCR) and the National Crime Victimization Survey (NCVS) cannot be emphasized enough as they are the only two ongoing national measures of crime in the United States that provide annual level and change estimates. While discussions of the UCR and NCVS often focus on the differences between the two, the two systems share many similarities in their development over time as well as their influence on how crime has been defined and studied. This entry examines the origins and development of these two data collection systems individually and describes the similarities between them. This entry concludes by examining how the UCR and NCVS have contributed to criminology by shaping views on what crimes "count," especially the focus on serious, violent street crimes. Alternative ways crime could be measured in the UCR and NCVS's existing frameworks are discussed.

Introduction

The Uniform Crime Reporting Program (UCR) and the National Crime Victimization Survey (NCVS) are the only two ongoing national measures of crime in the United States that provide annual level and change estimates. Data collected under the UCR use official records to measure crimes known to police. The NCVS data rely on survey responses to generate estimates of criminal victimization. This entry examines the origins and development of the UCR and NCVS individually and describes the similarities between the two data collection systems. This entry concludes by discussing how both systems have contributed to criminology by shaping what crime "counts," particularly the focus on violent, street crime. Finally, alternative ways are considered with regard to defining and measuring crime that could be used in the UCR and NCVS's existing frameworks.

Origins and Development of the UCR and NCS-NCVS

Uniform Crime Reporting Program

Before the UCR, very little information was available to study crime, and the existing data were primarily of prisoners collected by the Bureau of the Census. Information on individuals living in correctional facilities was included in the 1850 decennial Census (Rosen 1995). Before this national effort, a handful of states collected crime statistics using court and prison data. The first national collection of crime data using police records occurred in 1880 when the Census Bureau included homicides, arsons, and burglaries known to police as part of the 1880 decennial (Rosen 1995). Given the local level of data collection, a significant problem concerned the lack of uniformity in crime definitions across states. While this issue was acknowledged, it was not resolved with the early Census data collection efforts and these data were largely discounted (Rosen 1995).

By the 1920s, the need for national crime data had become apparent. Countries including France, Austria-Hungary, Sweden, Holland, Denmark, and Turkey were generating national crime statistics at this time (IACP 1929). Many criminologists and statisticians voiced their embarrassment that the United States lacked information on this important social indicator. In 1927, the National Crime Commission pronounced that "the United States had the worst criminal statistics of any civilized country" (as quoted by Rosen 1995, p. 220). Organizations such as the Social Science Research Council also supported the collection of crime data as a way "to provide a solid empirical base for 'scientific criminology.'" (Rosen 1995, p. 223). Police officials also demanded national statistics on the nature and extent of crime to "allay the public's undue fears" over media-created "crime waves" (Maltz 1977, p. 33).

While the need for national crime data was clear, the most appropriate way to generate these statistics in the United States was less so. Debates ensued over the appropriate data source to be used (see Rosen 1995, for a discussion). Initially interest focused on court records as police records were deemed to be unreliable (Maltz 1977). Ultimately the decision was made to utilize data closest to the criminal event, which supported the use of police records rather than court or correctional data (Rosen 1995; Maltz 1977). Along with the data source to be used, debates centered on who should collect and disseminate the data. The main two contenders were the Bureau of Investigation (the forerunner to the Federal Bureau of Investigation or FBI) and the Bureau of the Census (Maltz 1977). With the decision to collect police data, many favored the Census Bureau since it was a neutral organization (Maltz 1977). Law enforcement organization such as the International Association of Chiefs of Police (IACP) endorsed collection by the FBI (Maltz 1977). Eventually the FBI was selected due in part to lobbying by then FBI director J. Edgar Hoover (Maltz 1977).

While some scholars have questioned the emphasis on the IACP with regard to the original idea for the UCR (Rosen 1995), the IACP and its Committee on Uniform Crime Records undertook the immediate work for designing this data

collection system. After the decision to use police data had been made, many technical issues remained including "the structure of the data base, the source of the data, crime measures to be used, geographical uniformity of crime definitions, rules for scoring and counting offenses, [and] the mode of data collection" (Akiyama and Rosenthal 1990, p. 50). The work to address these issues culminated in the first UCR manual in 1929. This 464-page volume devoted more than half of its text to addressing the issues of which crimes would be reported to the UCR and uniformity in crime definition and counting procedures (IACP 1929). The remainder of the volume described other aspects of the UCR program, many of which are practices that still remain today (Akiyama and Rosenthal 1990). The handbook also explained the data submission procedure. Law enforcement agencies provided handwritten tallies of aggregate counts of crime, which was consistent with technological capabilities of the time. To promote participation in the voluntary UCR system, the FBI provided the necessary forms for free as well as and return, postage-paid envelopes (IACP 1929).

Initially the UCR collected seven crimes known to police, known as Index offenses. These crimes were murder and nonnegligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft. Later in 1982, arson would be added as the eighth Index offense by congressional mandate (FBI 2004). The basic criteria used to select these crimes were the seriousness of the crime, the similarity of rates of occurrence throughout all geographic regions of the country, the frequency of occurrence, and the likelihood of coming to the attention of police (Poggio et al. 1985). Other crimes had been considered for collection such as carrying concealed weapons, arson, and kidnapping. The IACP explained that such crimes were excluded because "[e]ither they are concealed when committed so that the police frequently do not know they occur, or else their statutory definitions vary to such a degree it would be impossible to obtain reasonably comparable results. Unless the returns are limited to those offenses which can be relied upon to give a trustworthy picture of crime, their purpose will be defeated at the start." (IACP 1929, p. 180). The IACP did emphasize that simply because a crime was not included in the UCR, it did not prevent a local law enforcement agency from collecting data for its own records.

In 1930, the FBI began collecting UCR data based on the program outlined by the IACP. That year, 400 law enforcement agencies provided data (FBI 2004). Since that time, many changes occurred (see Barnett-Ryan 2007, for a complete accounting), but the essence of the UCR Program remained basically the same for decades including the submission of handwritten aggregate tallies (Poggio et al. 1985, p. 21). In the late 1970s, pressure mounted for the FBI to modernize the UCR in order to capitalize on both the innovations in the capability of law enforcement agencies to collect more detailed crime data as well as the improvements in the scholarly study and understanding of criminality (FBI, n.d.; Poggio et al. 1985, p. 21).

In response to these calls for an update, the Bureau of Justice Statistics (BJS) and the FBI commissioned a team of experts to reevaluate the UCR (Poggio et al. 1985). The recommended plan that came out of this review was not fully implemented. Its call for more detailed crime information at the incident level, though, did provide the foundation for what would become the National Incident Based Reporting System (NIBRS) form of data collection for the UCR. South Carolina participated in the FBI's pilot incident-level reporting program and in 1991 became the first state to submit its UCR data in NIBRS format (Barnett-Ryan 2007). By 1992, North Dakota, Iowa, and Idaho also were submitting their crime data in NIBRS format.

The NIBRS format covers a wider variety of offenses than captured by the summary report system's Index, or as they are referred to today Part I, offenses. As the FBI has moved away from a "Crime Index" toward violent and property categories, the term "Index offenses" is no longer used in favor of "Part I" offenses (FBI, n.d.). NIBRS collects incident-level details for 46 Group A offenses, which include the 8 former Index offenses. Examples of these additional

crimes range from kidnapping and forcible sex offenses beyond rape (such as sodomy and sexual assault with an object) to vandalism, gambling offenses, and fraud offenses. In addition to expanding the number of crimes reported to the UCR, NIBRS captures incident-level characteristics not included in the summary reporting system. For Group A offenses, NIBRS can collect up to 53 distinct data elements to describe each criminal incident. These details include demographic information for victims, offenders, and arrestees; victim-offender relationship; information about weapons, injuries, stolen items; and clearance of the incident. NIBRS also includes information on up to 10 offenses that occur in a criminal incident rather than reporting on only the most serious offense as was previously the case under the summary system.

One of the biggest limitations with NIBRS has been the transition from summary-based to incident-based data collection. As with the summary reporting system, participation in the UCR remains voluntary under NIBRS and is not mandated by the FBI. Unlike the summary reporting system, law enforcement agencies must be certified before they are eligible to submit data in NIBRS format. In addition, states and agencies are under no deadline to convert to NIBRS. As a result of these factors, the conversion process has been gradual. As of 2012, 32 states are NIBRS certified and NIBRS agencies cover 27 % of the US population (JRSA, n.d.).

National Crime Survey-National Crime Victimization Survey

Just as the UCR played an essential role in shaping the national collection of US crime statistics, the NCVS and its predecessor the National Crime Survey (NCS) had a similar groundbreaking role. Before the NCS, no national crime statistics had been based on survey data. The idea for using victim surveys to measure crime originated out of a convergence of factors that included increased attention on crime as a social problem, decreased confidence in the reliability of UCR data, and overall advances in the field of survey methodology (see Cantor and Lynch 2000 for a discussion). In 1965, President Lyndon Johnson appointed two Presidential Commissions with charges that included "reduc[ing] the amount of crime that eluded the attention of police." (Cantor and Lynch 2000, p. 98). To accomplish these goals, the Commissions realized the need for more data to better understand and explain what appeared to be a growing crime problem (President's Commission 1967). The quality of US crime data at the time was strongly criticized by the Commission, which noted "the United States is today, in the era of the high speed computer, trying to keep track of crime and criminals with a system that was no less than adequate in the days of the horse and buggy" (President's Commission 1967, p. 123).

Victimization surveys offered a new source of data that addressed the Commissions' two pressing concerns. First, survey-based crime data could provide greater incident details than the existing aggregate-based UCR. Second, victimization data were separate from that collected by police. The unknown issue was whether victimization surveys would work, both in terms of respondents being willing to report victimization experiences and the ability to produce annual crime estimates. To test this method, the Presidential Commissions supported three sets of pilot victimization surveys: one household-based survey in Washington, DC; one group of city-based surveys; and one national sample. All of these pilots produced two important findings: (1) confirmation that victim surveys could provide reliable crime estimates and (2) identification that a significant amount of crime that occurred was never reported to the police (Cantor and Lynch 2000). These pilot surveys also helped to identify methodological issues, such as ways to facilitate recall, temporal placement, and collecting crimes against the household (Cantor and Lynch 2000).

Based on recommendations from the Commission and findings from methodological studies, the NCS was created and first implemented in 1972. Originally the NCS was a system of four victimization surveys – one national household sample (the Crime Panel), a city level household sample (the Central City surveys), and two Commercial Victimization Surveys (one at the national level and the other at the city level)

(Rennison and Rand 2007). Only the Crime Panel survived and is what is referred to here as the NCS. The original goals for the NCS concerned providing an indicator of the crime problem independent of police data and the UCR and an ongoing measure of victimization risk (Rennison and Rand 2007). The NCS also served other purposes that included shifting the focus of the criminal justice system to the victim and away from the offender and providing a measure to assess changes in reporting to police (Rennison and Rand 2007). As a result of this close tie to monitoring the UCR, the crimes covered by the NCS were closely linked with the UCR crimes. The NCS crimes included rape, robbery, aggravated and simple assault, personal larceny (pickpocketing and purse snatching), burglary, motor vehicle theft, and property theft.

The NCS was a nationally representative sample of households. Initially all household members aged 14 and older were personally interviewed to determine their victimization experiences over the past 6 months. Those household members aged 12 and 13 were interviewed by proxy. The Census field staff conducted the NCS interviews in person and over the telephone using a paper and pencil instrument. The NCS consisted of three parts: the control card, the screening questionnaire, and the incident report. The control card collected household information and respondent demographics. The screening questionnaire ascertained if the respondent experienced a criminal victimization in the past 6 months, and the incident report collected details of each victimization incident reported. Households remained in the sample for 3½ years and were interviewed at 6-month intervals. Since the sample was of households and not individuals, those who moved out of the household were not followed, but rather the new residents were included in the sample for the remainder of time the household was in sample.

As with the summary UCR system, the NCS also underwent a significant redesign effort (see Cantor and Lynch 2000, for an in-depth discussion). The redesign efforts began almost as soon as the survey went into the field (Cantor and Lynch 2000). The NCS was a significant event for both those interested in survey methodology as well as social statistics. One area of criticism came from academics who challenged the significant financial expenditure on a data system that provided limited analysis options and few variables that could be used for theory testing (Cantor and Lynch 2000). Critiques also came from the law enforcement community who questioned this new survey-based crime data collection method and the associated criticisms of police-based crime data (Cantor and Lynch 2000). At one point, the NCS had attracted enough criticism that its sponsor the US Department of Justice considered discontinuing the survey until it could be redesigned (Lynch 1990). While the survey was never suspended, it was the subject of a reevaluation by National Academy of Sciences and a 5-year program of research, instrument development, and redesign planning (Lynch 1990).

The redesign effort focused on developing a better screening instrument to promote respondent recall of victimization incidents and revising the incident report to collect additional details about the victimization event (Cantor and Lynch 2000). Other redesign work concerned research devoted to recall periods, respondent fatigue, and other methodological issues (Cantor and Lynch 2000). Research and development of these efforts began in the 1980s and the redesigned NCVS was introduced into the field in 1992. Unlike the UCR's NIBRS redesign effort which remains an ongoing project, the implementation of the redesigned NCVS instrument and methodology was completed in 18 months. The redesigned survey came with a new name: the National Crime Survey became the National Crime Victimization Survey.

The redesigned NCVS kept the same three part format (control card, screener, and incident report). Otherwise the survey had been completely overhauled, most noticeably in terms of the screener. The screener now included a series of short cues to prompt recall. In the NCS, each screener question corresponded to a particular crime, such as assault. In the NCVS, screener questions instead focused on aspects of the victimization such as where it occurred, what the respondent

was doing at the time, and who committed the offense. The premise was to facilitate recall and to use various memory triggers to do so. In addition, new crimes were added as part of the screening process (vandalism and sexual assault) and specific prompts were added to better collect data on certain crimes that had been included in the NCS (rape and domestic violence). Rape and domestic violence were more clearly screened by including questions related to sexually based offenses and offenses committed by people known to the respondent. Additional crimes were also collected in supplemental questionnaires. These supplements increased after the redesign and concerned such topics as workplace violence, school crime, stalking, and identity theft. Other changes that occurred with the redesign concerned how the survey was administered with the introduction of computer-assisted telephone interviewing (or CATI) and how repeated (or series) crimes were counted. Rennison and Rand (2007) provide a comprehensive discussion of these and other changes.

The NCVS also has experienced additional changes since its redesign. Until recently, many of these changes centered on efforts to reduce costs as the NCVS has been flat funded for most of its existence since the redesign. The most significant of these changes has been the dramatic cut in sample size (Rennison and Rand 2007). Concurrent with the flat funding and sample cuts, external pressures demanded that the NCVS do more in terms of data collection and speed in releasing data. Additional questions were added concerning hate crimes, crimes against the disabled, computer crimes, and identity theft. The NCVS was also under pressure to produce its annual estimates more quickly (Rennison and Rand 2007).

The deep cuts to the NCVS program generated an unsustainable situation and a request by BJS for the National Academies' Committee on National Statistics (CNSTAT) for a panel to review the NCVS and "consider alternative options for conducting the [NCVS]." (National Research Council 2008, p. 23). The CNSTAT Panel's review found the NCVS was in need of more resources and restoration. Specifically "[a]s currently configured and funded, the NCVS is not achieving and cannot achieve BJS's legislatively mandated goal to 'collect and analyze data that will serve as a continuous and comparable national social indication of the prevalence, incidence, rates, extent, distribution, and attributes of crime . . .'" (National Research Council 2008, p. 78). The CNSTAT Panel offered many recommendations to restore the NCVS as well as an overall call to encourage appropriate funding of the survey. In response to this report, BJS began conducting a redesign effort to address the recommendations. These efforts have included restoring the sample size as well as efforts to generate subnational victimization estimates (BJS n.d.).

Similarities Between the UCR and NCS-NCVS in Their Origins and Development

While most discussions that examine both the UCR and NCS-NCVS focus on their differences and points where they diverge (see Biderman and Lynch 1991; Lynch and Addington 2007a), the two systems share many similarities with regard to their origins and development as well as the current demands and tensions under which each still operate. With regard to their origins, both the UCR and NCVS started from a need for better measures to inform the crime problem. For the UCR, law enforcement officials wanted to combat misperceptions of crime based on media-generated crime waves, and social scientists wanted empirical measures for studying crime. For the NCS, government officials and policymakers wanted to address a perceived increase in crime that could not be addressed with aggregate police data. Both also were developed as a way to obtain data that were closer to the crime problem than previously were available. For the UCR, previous crime statistics used data from courts or prisons. Decades later these police data would come under fire for being unreliable and subject to manipulation. Victimization data from the NCS provided a means to get closer to the criminal event and to avoid the police reporting filter.

In their development, both the UCR and NCS-NCVS systems underwent massive redesign efforts. While the specific changes differed, both were based on changes in technology and in improved knowledge of the crime problem. Changes in technology permitted the UCR to have the capability to collect more detailed information on crime incidents as well as to accommodate the millions of cases generated by an incident-based system. The summary reporting system originated in the 1920s when handwritten tallies were considered state of the art. By the 1980s and 1990s, police agencies had the ability to provide more sophisticated crime data as their own collection and analysis programs were increasingly becoming computerized (Poggio et al. 1985). The technology also permitted scholars, policymakers, and others the statistical and computing capabilities to manipulate and analyze these data. The NCVS similarly benefited from advances in technology that enabled CATI interviewing and improved data collection. More sophisticated statistical software packages – particularly programs that could adjust for the NCS-NCVS's complex survey sample – extended the utility of the data for researchers.

Improvements in the knowledge of the crime problem also prompted changes for both the UCR and NCS. For the UCR, incident details provided by the NCS highlighted the summary reporting system's limitations. With the available details from the NCS, researchers and policymakers began to demand the production of police data that provided incident-level information needed to explain the crime problem (Poggio et al. 1985). Solely providing aggregate level data was no longer sufficient. The NCS similarly was critiqued by scholars who demanded more details and explanatory variables to permit modeling of victimization risk and development of theories of victimization. Here the critiques came from the NCS itself. Once it was demonstrated that a victimization survey could produce reliable national estimates of crime, researchers and policymakers demanded more information. As with the UCR, annual estimates – even annual estimates coupled by additional incident details – were not enough.

Both the NCVS and UCR continue to face external demands and pressures, particularly to provide more information and to provide it more quickly. As discussed above, the NCVS has suffered the double burden of flat funding and increased demand for specific crime data and for producing these data more quickly. The UCR also has faced demands for faster release of crime data (e.g., Rosenfeld 2011) and for the collection of new crime types such as cargo theft (FBI 2011) and new crime definitions such as rape (FBI 2012). The UCR and NCVS must balance these demands for expansion and change with ongoing concerns. Both systems are charged with the task of generating annual level and change crime estimates. Data collection changes must be balanced with the need for series continuity. Both systems operate within budget and funding constraints that limit what each can do and how much data can be collected. Both also are essentially volunteer systems and must be concerned with overburdening their respondents – whether these data providers are individual law enforcement agencies and state data analysis centers (for the UCR) or individual survey respondents (for the NCVS).

Effect of UCR and NCVS on Defining What Crime "Counts"

Unlike other social indicators, no natural metric exists for crime and as a result, the classification of crime and which crimes are included is a very important decision (Lynch and Addington 2007b; much of the following discussion draws upon this chapter). Since counting every crime would be an impossible task, both the UCR and NCS-NCVS needed to identify those crimes that would (and would not) be included in their data collection systems. Literally, these decisions determined what crimes "counted." By doing so, certain crimes and aspects of those crimes are emphasized and others virtually disappear. In its summary system, the UCR elevated legal aspects of offenses and virtually excluded all other attributes of crime events. This decision was largely due to limits of police data at the time the UCR

was created, but these choices continue to affect views of crime even though these same crimes might not be selected today. Given the NCS's initial focus as a check on the UCR, the crimes captured by the NCS largely mirrored those of the UCR. The redesign of both systems increased the number of crimes that were counted. NIBRS has expanded UCR crimes particularly in the areas of fraud, drug-related crimes, forcible sex crimes other than rape, and kidnapping as well as incident details that permit identification of crimes against children and those involving intimates. The NCVS expanded the crimes it collected through supplements as well as added questions to the screening questionnaire.

The decision of what crimes are counted in each system has had far-reaching effects on views of what crime is. The paradigm of crime problem is one dominated by street crime and more serious crime. This focus has largely ignored criminal activity that is less serious, but more prevalent such as bullying, vandalism, minor assaults, and threats. These crimes do not result in great loss or injury, but can be consequential to community, schools, and other social institutions as well as neighborhood cohesion and feelings of safety. In addition, such crimes may be leading indicators of more serious crime and social problems.

The UCR and NCVS, however, do not have to be limited to this paradigm of serious, street crime. As discussed by Lynch and Addington (2007b), an ignored potential of both NIBRS and NCVS is to go beyond recreating the UCR summary system's Index Crime Classification. By collecting incident-level details, the NCVS and NIBRS have the necessary information to experiment with alternative crime classifications. These alternative classifications can improve insights on crime and victimization for researchers and policymakers. Any crime classification needs to assess the risk that crime poses and does so in a way that can inform efforts to control this risk (Lynch and Addington 2007b). Rather than focusing on legal classifications, an alternative classification system could use categories that focus on relationship status, location, or activity at the time of the incident. Relationship status could compare strangers and intimates. Location could compare public and private places, and activities could focus on domains such as home, work, school, or leisure. These categories could be further defined using injury or property loss rather than legal definitions to separate violent and property crimes. Such a classification system emphasizes elements of the crime event rather than criminal behavior or legal definitions (Lynch and Addington 2007b). This scheme also directs attention away from a street crime paradigm of crime and permits more expansive consideration of the crime problem.

The NCVS and NIBRS already collect the necessary information to experiment with alternative crime classifications. Few researchers have capitalized on the opportunity to incorporate theses classifications in their work. This lack of work might be evidence of the pervasive effect of the original UCR crime classifications. Consideration of alternative classification models were mentioned in both redesign efforts. One basis for NCS redesign was to focus on victimization risk and ways for researchers to define crimes using different characteristics (Cantor and Lynch 2000). Designers of NIBRS identified several benefits of an incident-based system, one of which focused on the analytical flexibility that would allow "users to count and categorize crimes in ways they find meaningful" as well as "to explore a myriad of details about crime and law enforcement" (Poggio, et al. 1985, p. 4).

Conclusion

The importance of the UCR and NCS-NCVS cannot be emphasized enough, especially how they – separately and together – have shaped what is known about crime. These two measures of crime provide a rich source of information on trends over time, annual estimates of the crime problem, and details about the incident. While discussions of the UCR and NCS-NCVS often focus on the differences between the two and points where they diverge (Biderman and Lynch 1991; Lynch and Addington 2007a),

the two systems share many similarities in their development over time. Both likely will continue to evolve over time as technology advances, as more is learned about crime, and as the demand for more detailed data continues.

The UCR and NCS-NCVS also have affected views of crime by defining what crimes count (and which do not). The UCR originally was viewed as speaking primarily to law enforcement agencies. It was not until its move to NIBRS were the needs of outside researchers and policymakers specifically acknowledged (Poggio et al. 1985). The NCS originated out of a need, in part, to monitor the UCR. Since the NCS's redesign, calls have been made for the NCVS to turn away from this monitoring function (Cantor and Lynch 2000). With a change in focus for both systems, attention could be devoted to ways to address the crime problem through utilizing alternative crime classification systems as well as exploring techniques for measuring crimes that affect citizens on a more routine basis than serious violent street crimes.

Recommended Reading and References

Akiyama Y, Rosenthal HM (1990) The future of the uniform crime reporting program: its scope and promise. In: MacKenzie DL, Baunach PJ, Roberg RR (eds) Measuring crime: large-scale, long-range efforts. SUNY Press, Albany

Barnett-Ryan C (2007) Introduction to the uniform crime reporting program. In: Lynch JP, Addington LA (eds) Understanding crime statistics: revisiting the divergence of the NCVS and the UCR. Cambridge University Press, Cambridge, pp 55–89

Biderman AD, Lynch JP (1991) Understanding crime incidence statistics: why the UCR diverges from the NCS. Springer, New York

BJS (n.d.) NCVS redesign. http://bjs.ojp.usdoj.gov/index.cfm?ty=tp&tid=91. Retrieved 4 Feb 2012

Cantor D, Lynch JP (2000) Self-report surveys as measures of crime and criminal victimization. In: Duffee D (ed) Measurement and analysis of crime and justice. National Institute of Justice, Washington, DC, pp 85–138

Federal Bureau of Investigation (2004) Uniform crime reporting handbook. FBI, Washington, DC. http://www.fbi.gov/about-us/cjis/ucr/additional-ucr-publications/ucr_handbook.pdf. Retrieved 4 Feb 2012

Federal Bureau of Investigation (2011) Cargo theft electronic data submission specifications. http://www.fbi.gov/about-us/cjis/ucr/Cargo_Theft_Elec_Submission_Specs_2011.pdf. Retrieved 4 Feb 2012

Federal Bureau of Investigation (2012) Revised rape definition announced for uniform crime reporting program. http://www.fbi.gov/news/news_blog/revised-rape-definition-approved-for-uniform-crime-reporting-program. Retrieved 4 Feb 2012

Federal Bureau of Investigation (n.d.) UCR general FAQS. http://www.fbi.gov/about-us/cjis/ucr/frequently-asked-questions/ucr_faqs. Retrieved 4 Feb 2012

International Association of Chiefs of Police (1929) Uniform crime reporting: a complete manual for police, 2nd edn. J.J. Little & Ives, New York

Justice Research and Statistics Association (n.d.) Status of NIBRS in the states. IBR Resource Center, http://www.jrsa.org/ibrrc/background-status/nibrs_states.shtml. Retrieved 12 Feb 2013

Lynch JP (1990) The current and future national crime survey. In: MacKenzie DL, Baunach PJ, Roberg RR (eds) Measuring crime: large-scale, long-range efforts. SUNY Press, Albany

Lynch JP, Addington LA (2007a) Understanding crime statistics: revisiting the divergence of the NCVS and the UCR. Cambridge University Press, Cambridge

Lynch JP, Addington LA (2007b) Conclusion. In: Lynch JP, Addington LA (eds) Understanding crime statistics: revisiting the divergence of the NCVS and the UCR. Cambridge University Press, Cambridge, pp 297–334

Maltz MD (1977) Crime statistics: a historical perspective. Crime Delinq 23:32–40

National Research Council (2008) Surveying victims. National Academies Press, Washington, DC

Poggio EC, Kennedy SD, Chaiken JM, Carlson KE (1985) Blueprint for the future of the uniform crime reporting program: final report of the UCR study. U.S. Department of Justice, Washington, DC

President's Commission on Law Enforcement and Administration of Justice (1967) Task force report: crime and its impact – an assessment. U.S. Government Printing Office, Washington, DC

Rennison CM, Rand M (2007) Introduction to the national crime victimization survey. In: Lynch JP, Addington LA (eds) Understanding crime statistics: revisiting the divergence of the NCVS and the UCR. Cambridge University Press, Cambridge

Rosen L (1995) The creation of the uniform crime report. Soc Sci Hist 19:215–238

Rosenfeld R (2011) The big picture: 2010 presidential address to the American society of criminology. Criminology 49:1–26

Developmental Criminology

▶ Integrated Cognitive Antisocial Potential Theory

Deviance Theory

► Transitional Justice

Differences-in-Differences in Approaches

James H. Derzon
Battelle Memorial Institute, Centers for Public Health Research and Evaluation, Arlington, VA, USA

Synonyms

Randomized control trials; RCTs

Overview

Although almost universally acknowledged as the most powerful research design for causal attribution, the randomized control trial (RCT) is not without limitations in applied research settings. Foremost among those limitations is the generalizability of study findings due to vicissitudes in tested sample, setting, intervention, and result. If generalizability cannot be assumed, the value of the finding as evidence for future action may be tempered.

In addition, several practical and theoretical limitations may be imposed by the RCT. Probabilistic equivalence on all measured and unmeasured characteristics is difficult to assume in many applied settings. The ethical requirement of clinical equipoise for many interventions mathematically limits discovery of effective interventions to 25–50 % of trials. RCT research tends to produce evidence on which recipients respond to the intervention being tested and not produce evidence on the intervention characteristics responsible for that response.

The preference for RCTs may bias what interventions are studied and what evidence is considered. Many interventions are not amenable to randomization, and evidence of their effectiveness may be discounted if only evidence from RCTs is valued. If only RCTs are valued, the increasingly available archival and electronic records data is similarly discounted as a source of evidence of what works as randomization must occur prior to intervention. Finally, as a design for causal proof, RCTs tend to focus solely on program effectiveness and not on other considerations of implementation and maintenance; as a prospective design, with measures often established prior to the research being conducted, it may be a design ill-suited to observing and documenting unintended consequences.

Two additional limitations result from a hyperallegiance to RCTs for evidence. When RCTs are applied in real-world settings, they often assess the links between complex interventions and the individual mediators and moderators of distal outcomes using complex statistical models and multivariate methods. Including such findings in formal meta-analyses is difficult and undermines the accumulation of evidence. Also reducing the accumulation of evidence is the limited number of researchers qualified to conduct formal RCTs.

To supplement the evidence available for evidence-based practice, an effectiveness-surveillance system using a synthetic difference-in-differences design is proposed. Such a system would provide associational evidence of effectiveness and provide a valuable complement to the evidence generated by RCTs.

The RCT: A Powerful Design for Cause

In an ideal world, or the idealized world of the laboratory study, it is difficult to argue against the logic of the RCT for drawing causal inferences. In its simplest form, a sufficient number of units (typically individuals) are selected to be representative of a known population and are randomly assigned to either a treatment or a control condition. Those in the treatment condition are then exposed to a stimulus (also known as the manipuland, independent variable, treatment, or intervention) that is withheld from those

in the control condition. Under the logic of the design, the performance of the control group approximates what would have been the performance of the experimental group in the absence of the stimulus being tested. Since the experimental group cannot simultaneously both receive and not receive the stimulus, scientists rely on inference to attribute differences in performance to the presence and absence of the stimulus being tested. If, in a well-controlled trial, the outcome covaries with the presence and absence of the stimulus, then a causal relationship is generally inferred.

Since first proposed in its modern form by Sir Austin Bradford Hill (1952), the RCT as the premiere method of causal inquiry has grown in both importance and stature. RCTs reside below systematic review and meta-analysis at the peak of evidence-based medicine's evidence pyramids and have been granted priority by the Department of Education (Scientifically-based evaluation methods 2003). Well-implemented randomization creates two groups with "initial probabilistic equivalence," that is, "groups that are comparable to each other within the known limits of sampling error" (Cook and Campbell 1979, p. 341). This, in turn, allows the researcher to assume that differences in outcome are attributable entirely to the intervention and are not influenced (confounded) by preexisting subject characteristics. Moreover, if the sample prior to randomization is representative of a population, randomization allows the result of an RCT to be generalized to the population represented by the sample.

While the RCT is a powerful strategy for attributing cause in many research settings and has been effectively applied to answer many important research questions, it is merely one feature of a well-designed experiment. In and of itself, randomization does not ensure valid or reliable findings nor ensure inferences based on those findings are correct. Well-conducted randomization creates covariate-balanced groups equal in their probability of displaying the outcome(s) of interest and differing only in their receipt or nonreceipt of the treatment/intervention/stimulus being tested. What randomization does not control for are other vicissitudes of the research endeavor. These can include sample selection; research methods and procedures; the use of sensitive, reliable, and valid measures; application of appropriate statistics; appropriate interpretation of findings and their limitations; and, finally, complete and accurate reporting of the study and possible threats to the validity of the research findings, as discussed in detail elsewhere (e.g., Cook and Campbell 1979; Lipsey and Cordray 2000; Shadish et al. 2002). In other words, when applied to real-world research in real-world settings, unequivocal belief that in and of itself randomization is sufficient to produce credible evidence is not tenable.

Before accepting a knowledge claim resulting from an RCT (or any other evidence claim, for that matter), all aspects of the study design and execution must be examined and judged. As an exercise in applied logic, all parts of the research endeavor work together to form a conclusion. The result produced from a single well-conducted RCT has tremendous value and, arguably, no other single study design produces evidence of equal merit. However, the intrinsic value of the RCT design for generating knowledge should not be confused with the extrinsic value of the design for generating actionable knowledge (i.e., its capacity to document "worth"). Moreover, its elevation to the "the gold standard" for knowledge generation has had the unintended consequence of discounting other methods of inquiry (Gugiu and Gugiu 2010). The discussion that follows provides several potential, practical, and documented limitations of relying on RCTs for generating robust actionable knowledge in criminology and criminal justice, with some of these limitations also applying to quasi-experiments. This entry concludes by offering an alternative design for knowledge accretion that may produce evidence of sufficient merit and worth to meet the expanding need for evidence of effectiveness of what works, for whom, and under what conditions. In providing this perspective, we add to the already robust literature documenting the limitations of RCTs (e.g., Boruch 1975; Pawson 1994).

Key Issues/Controversies

In this section we discuss some of the key issues which may limit the extrinsic value of the RCT for creating actionable knowledge. After a brief introduction on the lack of standards for assessing external validity and generalizability, potential limitations of different aspects of the RCT for creating generalizable evidence are discussed. The section closes with brief discussions of a number of practical and theoretical limitations of relying on the RCT for knowledge generation.

Lack of Standards for Assessing Generalizability

Through its control of confounds, a sufficiently large and well-implemented RCT undoubtedly provides greater internal validity and better warrant for attributing results to the intervention tested than do other study designs. Moreover, there are clear standards for evaluating study quality, and these standards allow us to judge the internal validity of the knowledge claimed by the trial. However, the purpose of science is the identification of generalizable conclusions (Overton 1998), and "conclusions limited to a specific subset … are not scientifically informative" (Hunter and Schmidt 2000, p. 277). External validity estimates the extent to which the results obtained by a study generalize to an unstudied population; if subjects in an RCT are representative of the unstudied population, study results are likely to replicate well in these unstudied populations. Unfortunately, there are few established guidelines for estimating external validity (Rothwell 2005) and, as estimated in the social sciences, what guidelines do exist often seem reduced to simple sample demographics. In contrast, Rothwell postulates 39 issues that potentially affect external validity for clinicians to consider when evaluating clinical practice RCTs in medicine.

Potential Limitations of Generalizability

Below we discuss several potential limitations of generalizability in an RCT, particularly those that are designed to produce evidence of efficacy. These concerns, in particular, have been proposed as central to the failure of efficacy research to translate into effective behavioral and health promotion efforts (Glasgow et al. 2003). These limitations relate to generalizability of the sample, the setting, the intervention itself, and the result.

Generalizability of the Sample. How representative of unstudied populations are subjects who agree to participate in an RCT? For many practical reasons, study populations are likely selected on their proximity to research staff. If research-intensive settings (e.g., university towns) are fundamentally different from other settings, it may limit generalizability. Subjects in such settings, and those who participate in research generally, are likely more motivated to seek intervention than those not interested in participating and are willing to be randomized to a control condition in which they will receive no, limited, or delayed services. This willingness is ensured through the informed consent procedure, in which subjects voluntarily sign a plain-language form attesting to the fact that they understand the implications of participation. For many interventions this may not be problematic, but for efficacy research in which the sample is self-selected, homogeneous, highly motivated, and screened to control for comorbid conditions, it is reasonable to question the generalizability of the sample and if it is, in fact, representative of the population it claims to represent.

Generalizability of the Setting. In addition to much research being conducted in settings convenient to the researchers, it should be observed that all interventions occur within a political, economic, social, and cultural context. RCTs can be very effective at controlling these potential confounds within a study, but ignoring these external community-level characteristics when generalizing beyond the studied sample may ignore consequential causal factors. Simply put, even if demographic characteristics of the sample included in the RCT are similar to an unstudied population, generalizability will be compromised to the extent individual responsiveness to intervention is influenced by the political,

economic, social, and cultural context of the community in which the intervention is tested relative to the community to which the findings are hoped to generalize. Since variance is often controlled by limiting an RCT to one particular setting, such trials provide no way of estimating if or how the unique circumstances of that setting influenced results. These potential influences are rarely discussed in the limitations sections of articles, and their consequence is likely unappreciated in federal initiatives that require implementation of science-based programs whose efficacy may have been established in a limited setting.

Generalizability of the Tested Intervention. Randomization controls for confounding by preexisting subject characteristics, but other factors associated with implementing and delivering the intervention can bias and limit the generalizability of a result. RCTs are often used to test whether an intervention is efficacious (is it effective under optimal conditions?); whether the intervention is effective under real-world conditions is an entirely different question. Often the intervention tested by an RCT is not the intervention that gets disseminated or adopted; developers often refine their interventions based on the results of their research, and even simple interventions are amalgams of activities that are typically adapted to local conditions (Durlak and DuPre 2008). If the intervention being implemented lacks fidelity to the intervention that was tested, the implemented intervention cannot be considered evidence-based.

Generalizability of the Result. The effect of a given cause is captured by the findings of a well-conducted RCT. Since findings are based on a comparison of outcomes in each arm of the experiment (e.g., a *t*-test that compares the means and standard deviations of each condition), each arm contributes equally to the observed finding. A fundamental implication of this assertion is that both conditions of an experiment must be fully specified (Reichardt 2011). Only if both arms of the experiment are fully understood and explicated can the results of an RCT be interpreted. This counterfactual definition of effect is known as the Rubin model of causality (Rubin 2005) and is integral to experimental design. Unfortunately, one seldom reads even a cursory description of what services were received by the counterfactual sample.

Although reliable statistics are lacking, it is likely that many RCTs in applied settings are corrupted by compensatory services within the comparison group. For example, a multisite randomized control study for high-risk youth (HRY) found that 21 of the 46 randomly assigned "control" sites (45.7 %) received more prevention services than the sites randomized to receive the HRY intervention. An initial analysis concluded that the HRY intervention had no effect, whereas statistically controlling for compensatory exposure disclosed a strong and significant positive effect (Derzon et al. 2005). Depending on which counterfactual is assumed, both results are arguably correct (the intervention was ineffective relative to other services available and more effective than not receiving prevention services). Note that if compensatory services had not been measured, an incorrect conclusion could have been that the intervention was not effective. If the counterfactual is not defined and not documented as carefully as the intervention, it is hard to imagine how the result of an RCT can be understood.

Practical and Theoretical Limitations to the RCT

Practical and theoretical limitations to the RCT are discussed below. These include assuming initial probabilistic equivalence, clinical equipoise, decisions about what evidence is considered, and limitations to knowledge accretion.

Assuming Initial Probabilistic Equivalence. Recall that the purpose of randomization is to create initial probabilistic equivalence. This is an idea that is closely tied to the theory of the intervention, its assumed causal relationship with the outcome, and the potential for other factors to produce the outcome or that influence subjects' response to the intervention. As subject variability in response increases, confidence intervals tend to increase and commensurately larger

samples are necessary for statistical power and avoiding a type II error (not finding an effect for interventions that are, in fact, effective). Thus, for interventions in which measures are robust and the causal relationship is mechanistic and deterministic and the effect is produced only by the cause being tested, the assumption of initial probabilistic equivalence may be strongly supported. However, if there are multiple sufficient causes to produce the outcome (overdetermination), or if the theory and measures used account for only part of the variability in outcome (underspecification), the assumption of initial probabilistic equivalence – especially in small-sample studies – may be considerably weakened. In such cases, randomization may not create the covariance balance necessary to assume initial probabilistic equivalence, and biased results are possible. Although RCTs provide a strong defense against the violation of this assumption, in criminology, and in evaluation research generally, there is a long history of null findings that are, in part, attributable to a lack of statistical power (Lipsey 1990).

Clinical Equipoise. From a researcher's perspective, uncertainty and clinical equipoise are moral imperatives for the conduct of ethical research (Weijer et al. 2000; cf. Veatch 2007). Equipoise is defined as genuine uncertainty as to which arm in an experiment is likely to prove more effective. Once a treatment is known to be effective, it is generally considered unethical to withhold that treatment. Mathematically, if there is true uncertainty of the effectiveness of the practice being tested relative to its alternative (some will be superior, some inferior, and some equal), adherence to equipoise limits progress in advancing evidence-based practice to discovering 25–50 % of successful treatments when they are tested in RCTs (Djulbegovic 2009). To the extent equipoise is achieved, the ethical implementation of RCTs may sorely constrain the advancement of knowledge.

What Evidence Gets Considered? The RCT focuses attention on the units that can be randomized and on interventions that are amenable to randomization. Thus, much of the RCT research focuses on recipient characteristics and who responds to the intervention being tested (e.g., publications that present multivariate evidence accounting for individual differences in displaying the outcome). Without minimizing the value of this information – which is useful for intervention targeting – such results provide little insight for improving service delivery or identifying the active ingredients of multicomponent interventions. Since service providers control these aspects of interventions, it seems that equal attention to these features is warranted. Multisite randomized trials can provide such evidence, but these are expensive and rare. Meta-analysis and systematic reviews of RCTs may also provide that evidence, but building the evidence base necessary is dependent on sufficient reporting of these details from the original trial. In sum, while the typical RCT may tell us for whom the intervention is effective, it may provide little information to advance our understanding of how and why interventions achieve their results.

A second implication of a commitment to RCTs as "gold standard" evidence may be an a priori filtering of the kinds of interventions that can have "gold standard" evidence of effectiveness. For example, environmental interventions are often not amenable to randomization. Thus, if only RCTs are accepted as credible evidence, our understanding of the impact of these interventions will be highly constrained. For example, Cozens, Saville, and Hiller in 2005 published a review of the effectiveness of crime prevention through environmental design (CPTED) strategies. Interestingly, only one title cited in the review's bibliography contains the word "randomized." Yet, while many studies of environmental interventions cited in this review do not meet the gold standard, the results that are included are encouraging. Equally important, as a practical matter, relative to individual-based interventions, the fidelity of environmental interventions does not degrade once they are implemented, nor do such interventions require booster sessions and additional resources to maintain impact. Had Cozens et al. dismissed evidence for CPTED practices not obtained through RCTs, their review would likely have been much shorter and a good deal of knowledge would have been lost to the field.

A third implication is that, by definition, RCTs are prospective studies. Randomized assignment must occur prior to the intervention, and this requirement obviates the use of retrospective data. Particularly in medicine – although they are increasingly available in other domains – electronic records provide a potential wealth of data for conducting comparative effectiveness studies. Commensurate with this growth are statistical matching methods that are increasingly robust in producing initial probabilistic equivalence permitting causal inference in retrospective and observational studies (Stuart 2010). Prospective data collection is expensive, and even small RCTs often require considerable costs to build effective sampling frames, to create buy-in with both controls and subjects, and to implement with fidelity. Providing incentives to a control or comparison group that receives limited, no, or delayed services may also be a nontrivial expense. The cost of collecting these data may, in part, explain the propensity of RCTs to focus almost exclusively on documenting intervention effectiveness, while often ignoring or omitting many other issues of concern to program adopters (such as cost, training, local acceptance of intervention theory and activities, and other issues of implementation or maintenance).

Finally, unintended consequences are often not discussed in RCT research. This is not to suggest that RCTs cannot collect and report this information, but merely an observation that RCTs typically focus on establishing the efficacy or effectiveness of a program through a circumscribed set of measures. These measures are defined by a logic model developed in advance of the trial and this focuses attention on the measures adopted prior to the start of a trial. If not associated with the measures defined in advance, unintended consequences may be ignored and are almost certainly unmeasured, and a biased understanding of the impact of the intervention may be promulgated. Morell (2005) offers some reasons for unintended consequences and guidance on how evaluation methodologies can be improved to account for these unforeseen and/or unforeseeable consequences.

Limitations to Knowledge Accretion. One of the major limitations of relying solely on RCTs for generating knowledge surrounds the accumulation of knowledge about effective practices. RCTs are best suited for establishing causal linkages between discrete phenomena (a manipulation and an outcome) that are linked by simple, short causal chains (Victora et al. 2004). In addition to the probabilistic constraints of equipoise, when RCTs are applied in real-world settings, they often assess the links between complex interventions and the individual mediators and moderators of distal outcomes. One method to account for that complexity is to use multivariate methods to report conditioned findings. That is, propensity adjustment may be used to account for selection bias, and multiple regression or other forms of statistical modeling are used to analyze the effects of the intervention. There are often good theory-building reasons for doing this, but because the partial or semipartial estimate of impact is determined by the variables in the model, findings from these studies cannot easily be used to examine the stability of evidence across multiple trials unless each researcher modeled the same variables. Suffice it to say, they do not, and the simple main effects that could be included in evidence synthesis are often not reported (Alford and Derzon 2012).

Also limiting knowledge accretion may be the number of researchers available to conduct RCTs and other experiments in criminology and criminal justice. Although the number of PhDs granted in the USA in criminology and criminal justice more than doubled between 1997–1998 and 2007–2008, still only 104 PhD degrees were granted in 2007–2008 (ASA, 2010). The June 2012 newsletter of the American Society of Criminology's Division of Experimental Criminology boasts 167 active members (AEC/DEC, 2012). While these numbers certainly do not represent the total number of researchers conducting or capable of conducting RCTs in criminology and criminal justice, it does seem fair to question whether the tremendous need to produce evidence, and to conduct original research and replications across all the varied settings and populations that exist, can be met by the supply of researchers available to conduct RCTs, or the funding available to support researchers conducting RCTs.

In the end, it is fair to ask how successful RCTs are in providing the evidence necessary to support evidence-based practice. Several federal and privately funded groups now conduct evidence reviews in which they evaluate the methods, procedures, and internal validity of RCTs to make best-practice recommendations. The Center for the Study and Prevention of Violence (CSPV, 2012) summarizes the results of 12 such efforts. Programs supported by RCTs that meet commonly accepted criteria for providing credible evidence are recommended as model, exemplary, effective, or perhaps promising programs. Although the list of effective programs compiled by CSPV is long (over 450 programs), when broken out by outcome or contrasted with the number of programs studied, the list takes on a different character. For example, of the 491 programs included on CSPV's list in 2010, only 46 programs assessed a school-based ATOD or violence program, had implementation materials, and had been tested in two or more minimal quality trials (Alford and Derzon 2012). The Blueprints for Violence Prevention project at CSPV has examined over 900 programs to identify 11 effective and 22 promising violence prevention programs (CSPV, 2012). The problem is not limited to drug or crime studies. The US Department of Health and Human Services' Office of Adolescent Health screened 1,000 studies to identify 28 programs that likely reduce the probability of teen pregnancy or birth (DHHS/OAH, 2011). As another example, the US Department of Education's Institute of Education Sciences created the What Works Clearinghouse in 2002 and has screened at least 461 programs to improve academic achievement. Of these, 5 programs showed positive effects with a medium-to-large evidence base (DOE/IES, 2011).

The low ratio of programs of documented effectiveness relative to the number of programs examined suggests that relying on gold standard RCTs is not an efficient or a particularly effective standard on which to build evidence for science-based practice. It is possible that much of the evidence identified or submitted in support of these programs was not eligible because it was not obtained using an RCT, or that the RCTs submitted did not meet study quality inclusion criteria. Regardless, these numbers represent an astonishing paucity of evidence relative to need. As a corollary, how much skill, time, and money will be required to conduct high-quality RCTs that meet these study quality standards as valid, reliable, robust, and generalizable information for practice?

An Alternative: Seeing Beyond the Trees

The commitment to science-based intervention does not require a commitment to the RCT as the only path to identifying effective strategies for social betterment (Concato et al. 2000). It may be required for causal proof, but this is a standard that should follow and not lead the development of science-based practices. Alternative approaches are tenable and have proven useful. Key among these are natural experiments (e.g., difference-in-differences designs, regression discontinuity), adjustment for selection (e.g., matching, case-control, regression, fixed effects [sibling/person as own control], propensity scoring, and doubly robust estimation), and instrumental-variables analysis. Because of its applicability to criminology, the difference-in-differences approach is discussed in greater detail below.

Difference-in-differences (DID) approaches are natural experiments that contrast change in outcome for the intervention group (or groups) over time to change in one or more nonrandomly identified comparison groups. It is a popular design in criminology that relies, as perhaps all comparative research does, on the plausibility of the attribution that differences observed can be attributed to the intervention (Victora et al. 2004).

Given the many reasons outlined throughout this entry and the pressing need to develop evidence of what works in real-world settings, it seems an appropriate moment in the development of science to consider how advances in scientific methods, accountability in social programming, and information technologies can be harnessed for science to improve the human condition. Increasingly, social programs are tracking and monitoring performance data for program

improvement and accountability. The ubiquity of the personal computer, programs such as Microsoft Excel, and the development of the internet have made simple data management and the sharing of information over great distances possible. The past 20 years have seen a tremendous growth in the techniques and acceptance of meta-analytic methods for summarizing and analyzing findings generated by diverse instruments across uncoordinated settings. What is proposed is that these three phenomena be exploited to invert the phase model (e.g., Flay 1986; Greenwald and Cullen 1985) commonly adopted for advancing science to create a surveillance system that uses systematic data collection across multiple intervention settings to identify promising practices.

Using a synthetic DID model (S-DID) with distributed data collection and centralized analysis, local programs and evaluators would collect and submit summary performance data (i.e., pretest and posttest results) – and other relevant metadata (describing, e.g., the intervention, setting, and sample) – to a central repository that could then standardize those data and allow user-generated comparative effectiveness analysis. That is, findings from similar practices can be pooled to estimate average performance and the range of performance across multiple implementations, and these findings can be contrasted with performance scores from alternative practices. A user interface that graphically and quantitatively summarizes and displays the available evidence would give the local user tools to judge the breadth, depth, and consistency of evidence for each practice. Since findings are cumulative, the more settings that contribute, the more robust the estimates available. Homogeneous distributions of results are explained by sampling error, and metadata (user-provided information on setting, sample, and treatment characteristics) can be used by more sophisticated users to explore reasons for heterogeneity.

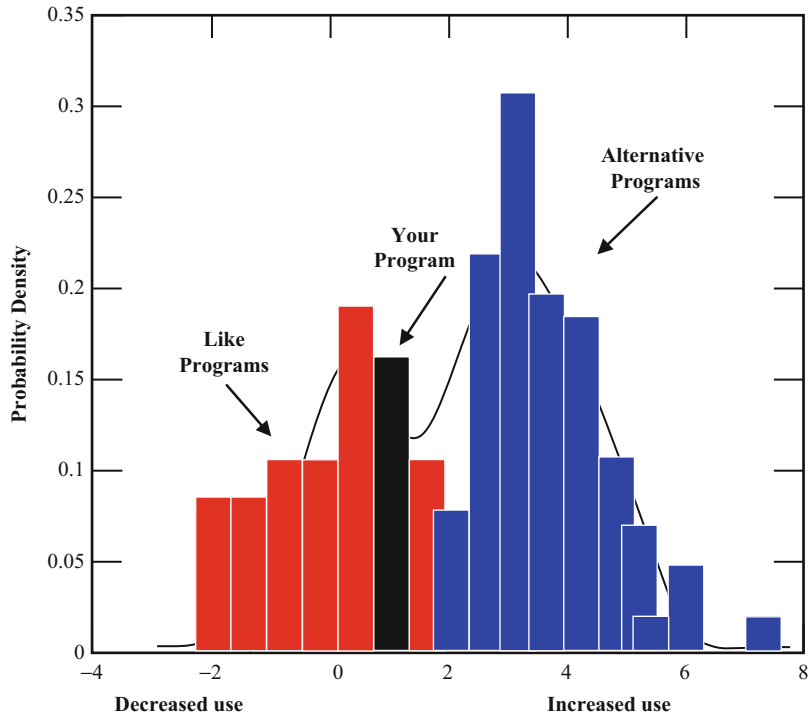
Persistent heterogeneity can be explored by adding categorical metadata variables to the system and entering results separately for each category. For example, if an intervention is hypothesized to be effective in urban but not rural settings, an urban/rural discriminator can be added to the metadata to distinguish subsequent submissions. As data in the two categories accrete, the homogeneity of evidence for each distribution can be tested and the results of the two distributions can be contrasted to confirm or refute the hypothesis. Hypothesized mediators that are continuous can likewise be entered as metadata and their influence on the outcome tested using correlational approaches.

Given the limitations of RCTs, the advantages of such a system are numerous. Interventions are tested in the settings in which they were adopted, and the diversity of adaptations, settings, and subjects provides an empirical basis for discerning what works where, when, and for whom. Because the diversity of interventions contributing to such a system is unlimited, opportunities for conducting ad hoc comparative effectiveness research – in near real time – are likewise unlimited. Since primary data are already being collected for performance-monitoring purposes, the cost of maintaining such a service, once engineered, would likely be modest. But the greatest benefit by far is that, by allowing local implementers and local evaluators to contribute evidence, the system democratizes evidence generation: all users who contribute create the evidence necessary for identifying evidence-based best practices. Interventions in real-world settings are tested, and the opportunity for innovation is distributed to all intervention practitioners.

What Is Required?

At the local level, data on outcomes are collected from service recipients. Summary statistics are generated for both pretest and posttest (this keeps data collection relatively unobtrusive and maintains subject confidentiality). Descriptive data that describe the sample, setting, activities, and context of the intervention are provided by program administrators, agents, or local evaluators.

Using a web interface, these locally derived data are then entered into a centralized data



Differences-in-Differences in Approaches, Fig. 1 An example of the possible output from S-DID

repository that permits the user to see how their change score compared with other change scores from similar interventions and then map that distribution against alternative interventions (see Fig. 1). Evidence for generalizability and testing of potential confounds could be accomplished using similarity scores (e.g., clustering based on proximity scores, with a commensurate loss of data from filtering) or meta-regression.

Limitations of S-DID

For all its potential, the S-DID approach is not without its own set of limitations. Foremost is the fact that it is a system dependent on distributed data collection. If the infrastructure necessary to support such a system were built, would local implementers and evaluators upload the data necessary for the system to work? Such a system is worthless unless populated with data from a variety of settings and samples. Moreover, would the data entered be reliable and valid? Although the summary statistics necessary to populate an S-DID database are fairly basic (e.g., means and standard deviations), whether local implementers and evaluators accurately collect data and convert their primary data into the necessary summary statistics is a nontrivial consideration. If the data entered are biased, flawed, or otherwise corrupt, they will be of little value to advancing knowledge.

Because evidence from the proposed system comes from multiple sites, it ameliorates, but does not obviate, the possibility of misattribution. Only the RCT, well-implemented, controls for third-factor influences and permits causal attribution. In this system, intervention selection is not random, and there may be unmeasured factors correlated with intervention selection that drive the result. Nonetheless, as interventions are tested across multiple settings and samples, the evidence generated in this system is self-correcting; if third factors determine the result, their impact will become apparent over time and through follow-on implementations. In exchange for causal attribution, the S-DID system allows plausible attribution and an early warning system to identify both likely effective and likely ineffective practices.

Differences-in-Differences in Approaches, Table 1 Comparison of RCTs with S-DID approaches to building evidence

Type of limitation	Randomized control trials	Synthetic difference-in-differences model
Participation	Only highly trained researchers produce effectiveness evidence	Practitioners and local programs produce effectiveness evidence
Politics	Proof of efficacy leads to approved solutions	Commitment to problem, not solutions
Power	Often underpowered and insensitive to impact	Pooling evidence increases the ability to detect effects that are present
Practicality	Distinguished by uniqueness and contribution to new knowledge	Only adopted interventions are tested
Price	Costs can be extraordinary, and much of the cost goes to not providing services	Costs likely minimal. Performance data are repurposed, remaining data are provided by program staff
Principles	Internal validity – did the intervention cause the result?	External validity – whether the intervention works, under what conditions, and for whom
Possibility	Burden of proof is "beyond a reasonable doubt" (e.g., $p < 0.05$), confounded by a publication bias toward successful trials	Burden of proof is the "preponderance of evidence" and the performance distribution of multiple implementations

Can It Work?

Two efforts currently under way are using the proposed evidence system to identify best practices. The CDC, through its Laboratory Science, Policy and Practice Program Office, is using both published and partner-submitted quality improvement data to identify best practices for laboratory medicine (https://www.futurelabmedicine.org/about/; Christenson et al. 2011), and Bill Hansen and the author are developing a web-based system for managing local substance use prevention trial data (Evaluation-based Effectiveness Testing; DHHS Grant # DA026219). The National Evaluation of Safe Schools/Healthy Students used an S-DID approach to estimate the correlates of effectiveness in a sample of 59 grantees (Derzon et al. 2012). Across a variety of applications, the approach is providing evidence of practices associated with effectiveness.

In sum, S-DID circumvents many of the limitations of traditional RCT research (see Table 1). It is not offered as a panacea for all the limitations of RCT research, but may be expected to supplement RCTs in our quest to understand what interventions are likely to be effective, under which conditions interventions tend to be effective, and for whom those interventions may be effective.

Related Entries

- ▶ Cambridge-Somerville Youth Experiment
- ▶ History of Randomized Controlled Experiments in Criminal Justice
- ▶ Place-Based Randomized Trials
- ▶ Propensity Score Matching
- ▶ Randomized Block Designs
- ▶ Randomized Experiments in Criminology and Criminal Justice
- ▶ Rational Choice Theory

Recommended Reading and References

Academy of Experimental Criminology, Division of Experimental Criminology (AEC/DEC) (2012) Newsletter. Academy of Experimental Criminology/The Division of Experimental Criminology Newsletter 7(1):1. Available at: http://gemini.gmu.edu/cebcp/DEC/AECDECJune12.pdf. Accessed 9/20/2012

Alford AA, Derzon J (2012) Meta-analysis and systematic review of the effectiveness of school-based programs to reduce multiple violent and antisocial behavioral outcomes. In: Jimerson S, Nickerson A, Mayer M, Furlong M (eds) Handbook of school violence and school safety, 2nd edn. Routledge, New York, pp 593–606

American Sociological Association (ASA) (2010) Report of the ASA task force on sociology and criminology programs. American Sociological Association,

Washington, DC. Available at: www.asanet.org. Accessed 8/7/2012
Boruch RF (1975) On common contentions about randomized field experiments. In: Boruch RF, Riecken HW (eds) Experimental testing of public policy: the proceedings of the 1974 Social Science Research Council Conference on Social Experiments. Westview, Boulder, CO
Center for the Study and Prevention of Violence (CSPV) (2012) Center for the study and prevention of violence: blueprints for violence prevention. Available at: http://www.colorado.edu/cspv/blueprints/index.html. Accessed 9/7/2012
Christenson RH, Snyder SR, Shaw CS, Derzon JH, Black RS, Mass D, Epner P, Favoretto AM, Liebow EB (2011) Developing laboratory medicine best practices: systematic evidence review and evaluation methods for quality improvement. Clin Chem 57:816–825
Concato J, Shah N, Horwitz RI (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342: 1887–1892
Cook TD, Campbell DT (1979) Quasi-experimentation: design and analysis issues for field settings. Houghton Mifflin, Boston
Cozens PM, Saville G, Hillier D (2005) Crime prevention through environmental design (CPTED): a review and modern bibliography. Property Manag 23:328–356
Department of Education, Institute of Education Sciences (DOE/IES) (2011) What works clearinghouse. Available at: http://ies.ed.gov/ncee/wwc/. Accessed 12/14/2011
Department of Health and Human Services, Office of Adolescent Health (DHHS/OAH) (2011) Evidence-based [teen pregnancy prevention] programs. Available at: http://www.hhs.gov/ash/oah/oah-initiatives/tpp/tpp-database.html. Accessed 12/14/2011
Derzon JH, Yu P, Ellis B, Xiong S, Arroyo C, Mannix D, Wells ME, Hill G, Rollison J (2012) A national evaluation of safe schools/healthy students: outcomes and influences. Eval Prog Plann 35(2):293–302, 10.1016/j.evalprogplan.2011.11.005
Derzon JH, Springer F, Sale L, Brounstein P (2005) Estimating intervention effectiveness: synthetic projection of field evaluation results. J Prim Prev 26: 321–343
Djulbegovic B (2009) The paradox of equipoise: the principle that drives and limits therapeutic discoveries in clinical research. Cancer Control 16:342–347
Durlak J, DuPre E (2008) Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. Am J Commun Psychol 41(3):327–350, 10.1007/s10464-008-9165-0
Flay BR (1986) Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. Prev Med 15:451–474
Glasgow RE, Lichtenstein E, Marcus AC (2003) Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. Am J Public Health 93(8): 1261–1267
Greenwald P, Cullen JW (1985) The new emphasis in cancer control. J Natl Cancer Inst 74:543–551
Gugiu PC, Gugiu MR (2010) A critical appraisal of standard guidelines for grading levels of evidence. Eval Health Prof 33(3):233–255. doi:10.1177/0163278710373980
Hill AB (1952) The clinical trial. N Engl J Med 247(4):113–119. doi:10.1056/NEJM195207242470401
Hunter JE, Schmidt FL (2000) Fixed vs. random effects meta-analysis models: implications for cumulative research knowledge. Int J Select Assess 8:275–292
Lipsey MW (1990) Design sensitivity: statistical power for experimental research. Sage, Newbury Park
Lipsey MW, Cordray DS (2000) Evaluation methods for social intervention. Annu Rev Psychol 51: 345–375
Morell JA (2005) Why are there unintended consequences of program action, and what are the implications for doing evaluation? Am J Eval 26:444–463
Overton RC (1998) A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychol Methods 3:354–379
Pawson R (1994) What works in evaluation research? Brit J Criminol 34(3):291–306
Reichardt CS (2011) Evaluating methods for estimating program effects. Am J Eval 32(2):246–272. doi:10.1177/1098214011398954
Rothwell PM (2005) External validity of randomised controlled trials: "To whom do the results of this trial apply?". Lancet 365:82–93
Rubin DB (2005) Causal inference using potential outcomes. J Am Stat Assoc 100(469):322–331. doi:10.1198/016214504000001880
Scientifically-based evaluation methods. Priority proposal by the US Department of Education. 68 Fed Reg 62445 (2003)
Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, New York
Stuart EA (2010) Matching methods for causal inference: a review and a look forward. Statist Sci 25(1):1–21. doi:10.1214/09-STS313
Victora CG, Habicht J-P, Bryce J (2004) Evidence-based public health: moving beyond randomized trials. Am J Public Health 94:400–405
Weijer C, Shapiro SH, Cranley Glass K (2000) For and against: clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. Br Med J 23:756–758
Veatch RM (2007) The irrelevance of equipoise. J Med Philos Forum Bioethics Philos Med 32(2):167–183. doi:10.1080/03605310701255776

Differential Association Theory

Gerben Bruinsma
Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), Amsterdam, The Netherlands
UV University Amsterdam, Amsterdam, The Netherlands

Synonyms

Differential social organization; Early Chicago school theory; History of criminological theories: causes of crime; Social learning theory

Overview

The differential association theory (DAT) of Edwin H. Sutherland is one of the key theories in criminology. The theory and its empirical support, however, are not undisputed. There is much confusion about DAT in the criminological literature, caused partly by Sutherland who changed his theory several times. Early in his career, Sutherland embraced the multiple-factor approach and the interactionist theory of William Thomas (1863–1947), then leaned to social disorganization and culture conflict, and finally settled his own DAT in the late 1930s and 1940s. In this entry, changes in DAT are discussed and the complexity of DAT is illuminated by exploring the intellectual roots of the theory. Several propositions of DAT can be understood better when the theory is situated within broader sociological and psychological traditions. The entry continues by presenting the underlying assumptions of DAT and the arguments of its critics.

Introduction

The differential association theory (DAT) has a history that goes back to the 1920s when a scholar in sociology with a minor in economics was invited to write a textbook on criminology with less focus on European data and research (Bruinsma 1985; Gaylord and Galliher 1988; Goff and Geis 2011). Edwin Hardin Sutherland (1883–1950) published then his "Criminology" that would be influential in the field of criminology for decades (Sutherland 1924). In this first edition, hardly any elaborated perspective on crime causation was presented but in subsequent revised editions, the DAT has been posed, elaborated, and changed in just a few pages each. Besides that, Sutherland, who was by nature an intellectual doubter, tried to improve his DAT continuously. As a result, DAT has been published over the years in scattered works in sometimes slightly different versions, but also in radically changed statements. The concept of *differential association* was introduced in Sutherland's study on the professional thief (Sutherland 1937). Despite the confusion, DAT is still one of the most influential but complex theories in criminology (Geis 1976; Kubrin et al. 2009; Vold et al. 2002). Its complexity is also likely responsible for its abominable history of empirical testing and for the fact that criminologists simplified DAT in the last 60 years into "*the bad influence of delinquent peers on young people's crime involvement*" (Bruinsma 1992; Cressey 1952). This entry will go beyond that simplicity.

Development of the Differential Association Theory

DAT has a remarkable history because of the fact that Sutherland has published different versions of his explanation of crime in four successive editions of his textbook (*Principles of*) *Criminology* over a period of 20 years and in some other publications as well. DAT has been presented by Sutherland in different forms, using varieties in wording, in central concepts, and in causal mechanisms. The history of DAT reflects more or less the intellectual personality and life history of a great scholar, always doubting whether his theory can address the explanation of crime: "*It is a story of confusion, inconsistencies, delayed recognition of implicit meanings, and of much*

borrowing from and stimulation by colleagues and students" (Sutherland 1956[1942], p. 13). Sutherland was sensitive to the theoretical influences of other criminologists and for critical remarks on his thoughts by his colleagues and friends. The editor of Lippincott Sociological Series invited Sutherland to publish his (*Principles of*) *Criminology*, a book that would be most influential in criminology for decades.

In the first edition of this textbook of 1924, Sutherland was, as he acknowledged later (1956), a follower of the multiple-factor approach. In his critical discussion of the state of art in criminology of that period, Sutherland was looking for information in the empirical research "*that enable us to state that such a person with such and such attitude in such and such situation will always become delinquent*" (Sutherland 1924, p. 82). According to Sutherland each individual is capable of committing crimes but "*it requires contacts and direction of tendencies to make either a criminal or a law-abiding person*" (Sutherland 1924, p. 118).

Furthermore, he adapted the perspective of Chicago sociologists about changing societies in which the people are constantly being influenced by different values and norms and that the family has becoming less influential in socializing young people: "*Social disorganization is a condition of progress as well as of delinquency*" (Sutherland 1924, p. 133). In summary, in the first edition some general assumptions of the DAT can be found:

(a) The search for a universal explanation of crime
(b) Attention to the interaction of the individual and his/her social environment
(c) Interest in cultural and macro social conflicts and their consequences for the individual
(d) The idea that crime – like all other behavior – is learned, and not the result of heritable defects

In the second edition, published ten years later with the changed title *Principles of Criminology*, Sutherland's own vision is more clearly present. Consistent with his thesis about crime as learned behavior, Sutherland added: "*Failure to follow a prescribed pattern of behavior is due to the inconsistency and lack of harmony in the influences which direct the individual*" (Sutherland 1934, p. 52). More strongly than ever before, he joined the so-called "Chicago school" of sociology and criminology, in which macro social (cultural) conflicts are assumed to be central in the explanation of human behavior.

In his classical study *The Professional Thief*, Sutherland (1937, pp. 206–207) for the first time put forward the concept of differential association: "*Differential association is characteristic of the professional thieves, as of all other groups. . .The differential element in the association of thieves is primarily functional rather than geographical. Their personal association is limited by barriers which are maintained principally by the thieves themselves*." Interactions with other people are a necessary condition to enter the complex world of professional thieves on the border line of conventional society. Differential association is here defined in a narrow sense: *the underworld of criminals*.

In 1939, Sutherland opened his third, and largely revised, edition of *Principles of Criminology* with a chapter in which he tentatively presented his own theory in the form of seven propositions. With these propositions, he tried to connect three levels of explanations: (a) the *macro* level (that of culture conflicts), (b) the *meso* level (that of social disorganization), and (c) the *individual* level (that of having contacts with criminals, i.e., in 1939, differential association). He presented the following seven propositions as a general theory without naming it DAT (Sutherland 1939, pp. 4–9):

(1) The processes which result in systematic criminal behavior are fundamentally the same in form as the processes which result in systematic lawful behavior.
(2) Systematic criminal behavior is determined in a process of association with those who commit crimes, just as systematic lawful behavior is determined in a process of association with those who are law-abiding.
(3) Differential association is the specific causal process in the development of systematic criminal behavior.

(4) The chance that a person will participate in systematic criminal behavior is determined roughly by the frequency and consistency of his contacts with the patterns of criminal behavior.
(5) Individual differences among people in respect to personal characteristics or social institutions cause crime only as they affect differential association or frequency and consistency of contacts with criminal patterns.
(6) Cultural conflict is the underlying cause of differential association and therefore of systematic criminal behavior.
(7) Social disorganization is the basic cause of systematic criminal behavior.

He (1939, p. 9) summarized his theory on the same page as follows:

> Systematic criminal behaviour is immediately to differential association in a situation in which cultural conflicts exist, and ultimately to the social disorganization in that situation. A specific or incidental crime of particular person is due generally to the same process, but it is not possible to include all cases because of the adventitious character of delinquency when regarded as specific or incidental acts.

The crux of this version is that systematic criminal behavior is determined by the process of association with criminals just as systematic law-conforming behavior is developed in a process of association with law-conforming people. Cultural conflict is the underlying cause of differential association and social disorganization is the basic cause. In this version, Sutherland ascribed a limited sense to associations with criminal behavior patterns, namely, only associations with a criminal subculture. Belonging to such a subculture implies that the ratio of contacts turns in favor to one side.

The 1947 Version of Differential Association Theory

Edwin Hardin Sutherland published the final version of his theory in 1947. The years after 1947 Sutherland kept on searching for ways to improve his general theory of crime but his sudden and unexpected death by a stroke in 1950 prohibited that.

Sutherland introduced his nine propositions of DAT with the statement "*The following paragraphs state such a genetic theory of criminal behavior on the assumption that a criminal act occurs when a situation is appropriate for it, as defined by the person who is present*" (Sutherland 1947, p. 6). The added clarifications are from Sutherland (1947, pp. 6–7).

(1) *Criminal behavior is learned.*

According to Sutherland, this proposition implies that criminal behavior cannot be inherited. Besides that, he stated that a person who did not learn how to commit criminal behavior could not invent that behavior.

(2) *Criminal behavior is learned in interaction with other persons in a process of communication.*

In many respects, Sutherland added, this communication can be verbal as well as nonverbal.

(3) *The principal part of the learning of criminal behavior occurs within intimate personal groups.*

Negatively, this statement implies that impersonal communication like films or newspapers plays an unimportant part in the genesis of crime.

(4) *When criminal behavior is learned, the learning includes*:
(4a) *Techniques of committing the crime, which are sometimes very complicated and sometimes very simple.*
(4b) *The specific direction of motives, drives, rationalizations, and attitudes.*

(5) *The specific direction of motives and drives is learned from definitions of the legal codes as favorable or unfavorable.*

In some societies an individual is surrounded by persons who invariably define the legal codes as rules to be observed, while in others he is surrounded by persons whose definitions are favorable to the violation of the legal codes.

(6) *A person becomes delinquent because of an excess of definitions favorable to violation of the law over definitions unfavorable to violation of law*.

This is the principle of differential association. It has to do with criminal and anti-criminal associations. When somebody becomes criminal, it is because he has contacts with criminal behavior patterns and because of a relative isolation with anti-criminal behavior patterns. Sutherland added that everybody assimilates with the surrounding culture, unless other patterns are in conflict with this culture. This sixth statement also implies that neutral associations have none or little impact on the development of crime.

(7) *Differential associations may vary in frequency, duration, priority, and intensity*.

Associations can vary according to these modalities.

(8) *The process of learning criminal behavior by association with criminal and anti-criminal patterns involves all of the mechanisms that are involved in any other learning*.

This statement implies that learning criminal behavior is not limited to the process of imitation.

(9) *While criminal behavior is an expression of general needs and values, it is not explained by those general needs and values because noncriminal behavior is an expression of the same needs and values*.

With this statement, Sutherland opposed himself against criminologists who tried to explain the occurrence of criminal behavior by general needs and values, like searching for happiness, social status, or wealth.

Changes in the Differential Association Theory by Sutherland

Some remarkable changes compared to its predecessor can be observed in the final version of Sutherland's DAT (Bruinsma 1985; Chiricos 1967; Cressey 1952, 1964, 1969):

(a) All references to explanations of processes at higher levels of aggregation were removed from earlier statements of DAT.

(b) More than ever before, Sutherland emphasized the learning process of individuals. In statement one, he claims *that* criminal behavior is learned; in statement two, *by which* criminal behavior is learned; the third statement indicated the *context in which* the learning process take place; and in statements four and five, *what* is learned. In addition, statement eight finally supplies information on *the manner in which* criminal behavior is learned.

(c) Much confusion remained about the central concept of differential association (see § 7.2). The explanation of crime in the final version shifted from an excess of contacts with criminals to an excess of positive definitions. This is a remarkable shift in DAT. In 1939, Sutherland (Sutherland 1939, p. 6) wrote "*The ratio of criminal acts to lawful acts by a person a roughly the same as the ratio of the contacts with the criminal and with the lawful behaviour of others*." Cressey (Sutherland and Cressey 1966) added that it is all about the presentation of positive and negative definitions and that the person who presents them not necessarily needs to be a criminal.

(d) Compared to the 1939 edition, Sutherland abandoned the idea that associations need to be criminal themselves ("crime is the cause of crime"). In the 1947 version, it is about the presentation of behavior patterns, despite the character of the person who presents these.

(e) In the final version of DAT, the modalities of associations (duration, priority, frequency, and intimacy) were introduced. Sutherland replaced the 1939 concept of consistency of associations by intimacy and added two new modalities in addition to the frequency of associations: priority and duration. "*In a precise description of the criminal behaviour of a person these modalities would be stated in*

quantitative form and a mathematical ratio be reached. A formula in this sense has not been developed and the development of such a formula would be extremely difficult" (Sutherland 1947, p. 7).

Roots of Sutherland's Differential Association Theory

DAT cannot fully be understood without discussing its intellectual roots. The fundaments of DAT can be found in (symbolic) interactionism (Fisher and Strauss 1978). It is a sociological and social psychological perspective on society and individuals that originated in Chicago in the late nineteenth century and early twentieth century. Sutherland tried to translate the basic elements of this sociological perspective into the field of criminology. Some of these elements emerged from earlier works of sociologists and psychologists who developed (symbolic) interactionism, others from the publications of Sutherland's colleagues and students during his professional life. In the studies of Charles Cooley (1854–1929), George Mead (1863–1931), and William Thomas (1863–1947), the focus is on social processes and on the interaction of the individual with his environment. They all focus on primary intimate groups as the most important social environment of people (Warr 2001). By interacting and communicating with others, the self (that regulates one's social behavior) can develop into a generalized other: "*The organized community or social group which gives to the individual his unity of self may be called 'the generalized other'. The attitude of the generalized other is the attitude of the whole community*" (Mead 1974 [1934], p. 154). As in DAT, the face-to-face interactions and responses of others are emphasized in studying the social behavior of people, whether deviant, criminal, or conforming behavior. From others individuals learn how to define the social situations that they encounter in daily life. In general, symbolic interactionism emphasized the changeability of societies and the heterogeneity of values within them, exercising their impact on people's motives, attitudes, and behaviors. Mead emphasized the interactions with other people in which definitions (= normative constraints) are constantly reformulated and reworked in specific contexts (the process of learning). Being close to this principle, Sutherland views the development of a criminal self as a result of the individual's communications and interactions with the social environment.

According to Vold and his coauthors (2002, pp. 160–162), DAT adopted also another basic element from Mead's symbolic interactionism, that is, that the contents of what is learned (techniques, motives, drives, attitudes, and definitions) are cognitive elements. It was Mead's idea that human beings act toward objects based on the meanings that the objects have for them. People construct relatively permanent "definitions of their situation" out of their experiences they derive from particular situations, and form a relatively set of ways of looking at reality. Cognitive factors determine behavior or, as DAT stated that principle, an excess of positive definitions toward law violations. Mead argued that the self, as he called it, is socially constituted. The organized society is prior to the individual, and the self has a social structure in which the consciousness is socially organized from the social organization of society in general. Gongaware and Dotter (2005) demonstrated clearly that Sutherland owed much of the symbolic interactionist theory of George Mead and that Sutherland's principles can be perceived as a translation and elaboration of that theory in the field of crime causes.

Although Sutherland skipped the word "conflict" in his final version of DAT, he was convinced that culture conflict is the underlying cause of crime in society. His discussions with Thorsten Sellin (1896–1994) strengthened his view that culture conflicts are a consequence of the transition of society from homogeneous to heterogeneous and of the increasing social differentiation (Sellin 1938; Sutherland 1939). Migration processes accelerate changes in society and consequently culture conflicts will emerge. Sutherland nevertheless posits culture conflict as an important causal factor in his differential organization theory, explaining variations in crime rates of social and ethnic groups.

Assumptions of the Differential Association Theory

DAT is based on various assumptions. These assumptions are partly testable, partly not. The assumptions are that:

(i) Crimes are not inherited.
(ii) Crimes can be learned in social interactions.
(iii) Individuals are not just a tabula rasa in which the environment put definitions favorable to commit crimes and where crimes are the output of an unobservable process.
(iv) Individuals play not only a passive role in the learning process but also an active role in acquiring positive definitions, close to symbolic interactionism rejecting the idea of an oversocialized conception of individuals.
(v) Individuals do not invent criminal behavior.
(vi) DAT implies a recursive process in which committing crimes will continue as long as individuals have contacts with criminal behavioral patterns.
(vii) The mechanisms underlying the learning process are active continuously.
(viii) Personal traits or propensities like self-control or susceptibility play no role in the emergence of crime.
(ix) Processes at a higher societal level are the fundamental causes behind the causes at the individual level.

A number of these assumptions have been the subject of several debates after the death of Sutherland in 1950 (Cressey 1952, 1964, 1969). Most of these assumptions, however, can be translated in propositions and subsequently be tested empirically. An important assumption of Sutherland that personality characteristics play no role in the occurrence of crime needs to be clarified more fully in order to understand Sutherland's thoughts. Sutherland took the criticism very seriously that DAT cannot ignore personality characteristics, or psychological variables as he called them: "*I believe it is the most important and crucial question in criminological theory*" (Sutherland 1956[1942], p. 25). Originally he underlined that these variables should be incorporated in his DAT, later he rejected that idea (Sutherland 1956; 1956 [1942]).

Criticism on the Differential Association Theory

Criticism on the Nine Propositions

Sutherland theorized in his first proposition that crime is learned. This statement, however, offers no information on how criminal behavior is learned. Statement *one* contains no conditions that can lead to criminal behavior. Others have pointed at the ideological content of this proposition (Warr 2001, p. 184). However, from an explanatory point of view, the proposition is lacking to supply any information how the learning process is taking place and what is learned.

A similar critique is valid for the *second* statement of DAT. Every individual interacts with other people without generating crime. There is no specification of conditions by what kinds of interaction crime will occur. The statement informs us only that learning is not a solitary or isolated process, but a *social* process in which other people are involved. Who are involved is pointed at in the next statement: intimate groups. But Sutherland did not define in statement *three* in what intimate groups crime is learned. In practice, criminologists limit themselves almost always to "delinquent friends" or "peers" in general as empirical indicator for intimate groups. Just in statement *four*, a first condition of the occurrence of crime is mentioned. The variable "techniques to commit crimes" (whether complicated or simple) is assumed to be a necessary condition for committing a crime. Besides that, it is stated that the specific direction of motives, drives, rationalizations, and attitudes is relevant. However, which motives, drives, rationalizations, and attitudes remain unclear as well as how these have an impact on the occurrence of crime.

In statement *five*, some conditions about the origins of motives and drives are indicated. Statement five can be formulated as "If laws are defined positive, then motives and drives will emerge conform the law, and if laws are defined as negative then criminal motives and drives will emerge" (Opp 1974). Conditions for the occurrence of crime are stated in proposition *six*.

Sutherland claimed that when an excess of definitions favorable to violate laws over definitions unfavorable is present, crime will occur. In statement *seven* it is indicated that associations can be attributed different qualitative characteristics like frequency, duration, priority, and intensity, but contains no explanandum.

The nine statements of DAT vary considerably in scope and range. In modern epistemological perspectives, statements 4, 5, 6, and 8 can be classified as testable hypotheses. Statements 1, 2, 3, and 9 are general orientations or general theoretical outlooks in which clusters of undefined variables are related to each other and cannot be tested properly. The seventh statement is not only a general theoretical outlook but it also contains a metascientific sentence of methodological nature that quantifying is a heuristic useful activity in theory formation.

Criticism on the Concepts

The Excess of Positive Definitions to Violate the Law

Several complex and abstract concepts play a role in the nine propositions of DAT. Some are very problematic for the theory. The most worrisome is the concept of an excess of definitions favorable to violate laws that will cause crime. What is meant by the concept "an excess of positive definitions to violate laws over definitions unfavorable to violate laws"? To answer this question one has to clarify four issues:

(a) What are definitions favorable to law violations?
(b) What is an excess of definitions favorable to law violations?
(c) Is it the excess of definitions *of individuals whose criminal behavior has to be explained* or is it the excess of *presented* position definitions?
(d) How do individuals recognize and adopt these definitions from other people?

DAT is embedded in symbolic interactionism in which the learning of definitions is part of the development of the social self. People construct permanent definitions out of their experiences from particular situations and form a relatively enduring set of ways of looking at things. In communication with others these definitions are transferred (or, to do justice to the symbolic interactionist roots of DAT, the word of exchanging definitions positive to violating the law between people should be preferred, expressing the active role of individuals in building a set of attitudes) to the individual and consequently adapted to regulate his behavior in a given context. The definitions are cognitive elements like behavioral techniques, motives, drives, and attitudes. Sutherland only used the attitudes part of symbolic interactionism. These can be interpreted as a *kind of moral rules that guide a person's behavior*. If people would communicate to each other that a certain kind of behavior is not appropriate in a particular context, you may apply that moral rule as a consequence of that rule. If the social environment is proclaiming the attitude that a certain kind of behavior is appropriate in that context, you may apply that rule. The underlying concept of definitions resembles closely the concepts of *beliefs* of Hirschi's social control theory (2006[1969]) or of *morality* in situational action theory of Wikström (2010). It is common practice to include neutralization techniques being part of the concept of positive definitions, implying that part of the positive definitions to violate the law also consist of rationalizations that take away barriers to commit crimes. As such, they are a category of positive definitions.

The second complexity is the measurement of an "excess." Can an excess of positive definitions to violate laws be counted empirically? A number of critical comments (Bruinsma 1985; Glueck 1956; Matsueda 1982, 1988; Short 1960) encountered problems in the measurement of the excess of definitions.

At the theoretical level some additional critical questions can be raised. (i) Let us assume that Sutherland is right and that there is a kind of crossover in the formation of definitions; why is the crossover then put exactly on .5? Why not on .6 or .4? (ii) If we insist that a kind of crossover exists, then the chance that DAT will be refuted in empirical research will be 100 % because the number of conforming behaviors exceeds the number of criminal behaviors.

As a consequence, the definitions about these kinds of behaviors will also be unbalanced. (iii) Why does the presence of positive definitions, that are less than .5 of all definitions of a person, have no significance at all for the commission of crimes? (iv) We have to consider whether positive and negative definitions are about *all* violations of the law. Do people have a definition about all possible crimes? If yes, that would imply that someone must have an infinitive number of definitions. That would be very unrealistic. Lastly, (v) another objection can be put forward on this central concept of DAT. One can question whether an excess of positive definitions to violate the law is about general attitudes toward committing crimes or *specific* attitudes toward *specific* crimes. Is it possible that people have a positive definition to commit fraud and at the same time have negative definitions to other forms of crime? In DAT there is no sign whether these definitions are about specific crimes or crimes and rule breaking behavior in general.

However, the most important weakness of DAT is the indistinctness whether an excess of positive definitions is an attribute of the *person whose behavior has to be explained* or has to do with *his environment*. Almost all criminologists assume that the excess of positive definitions is of the person whose criminal conduct has to be explained, but in statement six of DAT, *no object* is mentioned to which the attribute of an excess of positive definitions favorable to violating the law is related. The debate has always been about the question whether not everybody who has contacts with criminal associations will become criminal. Cressey defended that important question not very clearly: "*Thus, he* (Sutherland) *does not say that persons become criminals because of associations with criminal behavior patterns; he says that they become criminals because an* overabundance *of such associations, in comparison with associations with anti-criminal behavior patterns*" and thus strengthened the confusion (Cressey 1964, pp. 25–26). This theoretical issue needs to be sorted out by solid empirical research.

The Modalities of Associations

Sutherland identified four modalities of associations in the seventh proposition of the 1947 version of DAT: It is stated that associations can vary in frequency, duration, priority, and intensity. The meanings of priority and duration are "*obvious and need no explanation*" (Sutherland 1947, p. 7). About priority he wrote "'*priority' is assumed to be important in the sense that lawful behaviour developed in early childhood may persist throughout life. This tendency, however, has not been adequately demonstrated, and priority seems to be important principally through its selective influence*" (Sutherland 1947, p. 7). Warr (Warr 2001, p. 185) suggested that Sutherland referred with this modality to the Freudian point of view that was accepted as a common fact at his time. It remains unclear whether the theory is about the frequency of the contacts a person has *now* or is it about the contacts he had f*or the first time in his life*. In the latter interpretation it is assumed that the younger an individual is when he first comes in contact with criminal behavior patterns, the more impact they have for the rest of his life. For instance, it is assumed that the parents have more impact on the development and acceptance of positive definitions to law violation than other persons later in life.

The modality *frequency of contacts* got no further specification by Sutherland. It is assumed that the more often you see people and talk to them, the more influential that association will be. Frequency can thus be either the number of times you meet other people or the amount of time people spend together in a defined period of time (day, week, month, year) or the number of times multiplied by the hours spent together. That last option is never used in criminological research. The third modality *duration* is left open by Sutherland with no further specification. If incorporated in the frequency of contacts, then duration will be redundant as a separate modality of contacts with criminal or conforming behavior patterns. Opp (Opp 1974) suggested in his modification of DAT to let out this modality for that reason.

The last modality, the *intensity* of criminal associations, also named *identification or attachment*, is the most complex of the four, because a (developmental) psychological notion is

introduced in DAT. By intensity Sutherland meant the prestige of the source of association. With this modality he tried to overcome the critics of his time that police officers and prison guards have frequent contact with criminals but show no criminal behavior. The suggestion was that police officers seldom have intimate relationships with criminals and hardly ever attribute prestige to them. That makes them less vulnerable for the impact of them. Identification is generally regarded as the necessary condition for an individual to obtain preferences, attitudes, values, and norms from a group someone belongs to or wants to belong to. Acquiring of these happens in an interaction process called socialization. Crucial in this socialization process is the existence of real or symbolic models with which the individual identifies himself.

It can be concluded that in DAT it is left open what the function of this proposition is in the theory and what the effects of these modalities are on the excess of positive definitions favorable to criminal behavior. Probably it is implied that the more frequent, the longer, and the higher the priority and the more intense the contact with associations, the greater the chance that an excess of positive definitions over definitions negative to law violations will occur (the sentence is deliberately formulated without referring to any object, considering the discussion of the sixth proposition of DAT in the previous paragraph). There are no indications presented that one of the "modalities" is more important than one of the others, nor if, and how, these modalities are related to each other. Is priority more important than duration? And if so, how much more important? In the history of DAT, there has never been an attempt made to qualify the relative weights of the four modalities into mathematical formulas to assess the impact of each of them on the formation of positive definitions favorable to violating the law.

Differential Associations with Whom?

In DAT it is stated that the interactions and communications are in intimate social groups of significant others. That allows a great number of people with whom a person can have associations. In many research, DAT is perceived as the theory of the bad peers or delinquent friends (Bruinsma 1992). Sutherland never limited DAT to have contacts with just peers or delinquent friends. People interact with parents, siblings, grandparents, members of the extended family, neighbors and other neighborhood residents, schoolmates and teachers, colleagues, shop owners, sport mates and trainers, acquaintances, and strangers. All can potentially influence an individual by exchanging positive and negative definitions toward law violations through interactions and communications. It is surprising to observe that siblings have not attracted much more attention in empirical research. It can be suggested that siblings share many interactions and communications about definitions and behavioral techniques lasting for many years.

Conclusions

The differential association theory of Edwin Sutherland has a long history in criminology. It has been presented by Sutherland in different wordings and scope in scattered publications over the years, many times in a very brief statement. The differential association theory can be placed within the tradition of symbolic interactionism of early Chicago sociology. There is much criticism on the nine statements of the theory as well as on the concepts Sutherland has used. The theory is not undisputed in criminological theory. Hirschi and Gottfredson (1980, p. 9) stated that DAT neither predicts nor explains criminal behavior.

During the years, a number of modifications of DAT were published, of which some became very well known in the field of criminology. Gresham Sykes and David Matza modified proposition 4 of DAT into their Neutralization Theory (Matza and Sykes 1957; Sykes and Matza 1957), Dan Glaser into the Differential Identification Theory based on proposition seven and into the Differential Anticipation Theory (Glaser 1978), and Richard Cloward and Lloyd Ohlin into the Differential Opportunity Theory based on a combination of DAT and anomie theory of Robert Merton (Cloward and Ohlin 1960); De Fleur and Richard Quinney

made a formalized model of DAT according to the principles of formal logic and set theory (Fleur and Quinney 1966); the German sociologist Karl-Dieter Opp reformulated the theory guided by the methodological rules of critical rationalism (Opp 1974, 1976) and one that is based on behavioral theory that is still present in today's criminology: the social learning theory (SLT) of Akers (Akers 1973, 1998). Akers have several times claimed that his social learning theory is superior to Sutherland's one: "*... social learning is not meant to be an alternative, competitive, or rival theory to Sutherland's position. It is, instead, a broader theory that integrates processes of differential association and definitions from Sutherland's theory, modified and clarified, with differential reinforcement and other principles of acquisition, continuation, and cessation from behavioral learning theory*" (Akers 1998, p. 47). This entry, however, demonstrated that this claim is not valid and that the roots of both are too distinct to compare. DAT is based on symbolic interactionism, social learning on behaviorism. They differ in scope, vision on human nature, and how people learn and decide to commit a crime in a certain setting.

Acknowledgement I thank professors Sally Simpson and Lieven Pauwels for their comments and suggestions for the improvement on a previous version.

Recommended Reading and References

Akers RL (1973) Deviant behavior. A social learning approach. Wadsworth Publishing, Belmont

Akers RL (1998) Social learning and social structure: a general theory of crime and deviance. Northeastern University Press, Boston

Bruinsma GJN (1985) Criminaliteit als sociaal leerproces: een toetsing van de differentiële associatietheorie in de versie van K.-D. Opp. Gouda Quint, Arnhem

Bruinsma GJN (1992) Differential association theory reconsidered: an extension and its empirical test. J Quant Criminol 8:29–50

Chiricos TG (1967) The concept of cause. A developmental analysis of the theory of differential association. Issues Criminol 3:91–99

Cloward RA, Ohlin LE (1960) Delinquency and opportunity: a theory of delinquent gangs. Free Press, New York

Cressey DR (1952) Application and verification of the differential association theory. J Crim Law Criminol Police Sci 43:43–52

Cressey DR (1964) Delinquency, crime and differential association. Martinus Nijhoff, Den Haag

Cressey DR (1969) Epidemiology and individual conduct. In: Cressey DR (ed) Delinquency, crime and social process. Harper & Row, New York, pp 557–577

Fisher BM, Strauss AL (1978) Interactionism. In: Bottomore T, Nisbet R (eds) A history of sociological analysis. Heinemann, London, pp 457–498

Fleur MJD, Quinney R (1966) A reformulation of sutherland's differential association theory and a strategy for empirical verification. J Res Crime Delinq 3:1–22

Gaylord MS, Galliher JF (1988) The criminology of Edwin Sutherland. Transaction Books, New Brunswick

Geis G (1976) Editorial: revisiting Sutherland's criminology (1924). Criminology 14(3):303–306

Glaser D (1978) Crime in our changing society. Rinehart & Winston, New York

Glueck S (1956) Theory and fact in criminology: a criticism of differential association. Brit J Criminol 7:92–98

Goff C, Geis G (2011) Edwin Hardin Sutherland: the development of differential association theory. In: Cullen FT, Johnson CL, Myer AJ, Adler F (eds) The origens of American criminology. Transaction, New Brunswick/London, pp 37–62

Gongaware TB, Dotter D (2005) Developing the criminal self: Mead's social psychology and Sutherland's differential association. Sociol Spectr 25:379–402

Hirschi T (2006[1969]) Causes of delinquency, 5th edn. Transaction, New Brunswick/London

Hirschi T, Gottfredson M (1980) The Sutherland tradition in criminology. In: Gottfredson M, Hirschi T (eds) Understanding crime. Current theory and research. Sage, Beverly Hills, pp 7–19

Kubrin CE, Stucky TD, Krohn MD (2009) Researching theories of crime and deviance. Oxford University Press, New York/Oxford

Matsueda RL (1982) Testing control theory and differential association: a causal modeling approach. Am Sociol Rev 47(4):489–504

Matsueda RL (1988) The current state of the differential association theory. Crime Delinq 34(3):277–306

Matza D, Sykes GM (1957) Techniques of neutralization: a theory of delinquency. Am J Sociol 22:664–670

Mead GH (1974[1934]) Mind, self & society from a standpoint of a social behaviourist, vol 1 (ed. and with an introduction by Ch. W. Morris, 19 edn. The University of Chicago Press, Chicago/London

Opp K-D (1974) Abweichendes Verhalten und Gesellschaftsstruktur. Luchterhand, Neuwied

Opp K-D (1976) Methodologie der sozialwissenschaften. Einführung in Probleme ihrer Theorienbildung. Rowohlt, Reinbeck bei Hamburg

Sellin T (1938) Culture conflict and crime. Social Science Research Council, New York

Short JF Jr (1960) Differential association as a hypotheses: problems of empirical testing. Soc Probl 8:14–24
Sutherland EH (1924) Criminology. Lippincott, Philadelphia
Sutherland EH (1934) Principles of criminology. Lippincott, Philadelphia
Sutherland EH (1937) The professional thief. The University of Chicago Press, Chicago
Sutherland EH (1939) Principles of criminology. Lippincott, Philadelphia
Sutherland EH (1947) Principles of criminology. Lippincott, Philadelphia
Sutherland EH (1956a) Critique of the theory. In: Cohen AK, Lindesmith AL, Schuessler K (eds) The Sutherland papers. Indiana University Press, Bloomington, pp 30–41
Sutherland EH (1956[1942]) The development of the theory. In: Cohen AK, Lindesmith LA, Schuessler K (eds) The Sutherland papers. Indiana University Press, Bloomington, pp 13–29
Sutherland EH, Cressey DR (1966) Principles of criminology, 7th edn. Lippincott, Chicago
Sykes GM, Matza D (1957) juvenile delinquency and subterranean values. Am Sociol Rev 26(6):712–719
Vold GB, Bernard TJ, Snipes JB (2002) Theoretical criminology, 5th edn. Oxford University Press, New York
Warr M (2001) The social origins of crime: Edwin Sutherland and the theory of differential association. In: Paternoster R, Bachman R (eds) Explaining criminals and crime: essays in contemporary criminological theory. Roxbury, Los Angeles, pp 182–191
Wikström P-OH (2010) Situational action theory. In: Fisher B, Lab S (eds) Encyclopedia of victimology and crime prevention. Sage, Thousands Oaks/London, pp 875–879

Differential Association-Reinforcement Theory

► Social Learning Theory

Differential Enforcement

► Biased Policing

Differential Group Organization

► Differential Social Organization

Differential Offending

► Juvenile Justice System Responses to Minority Youth

Differential Social Organization

Ruth Triplett
Department of Sociology and Criminal Justice, Old Dominion University, Norfolk, VA, USA

Synonyms

Differential group organization

Overview

In this entry, the concept of differential social organization is traced from the publication of the fourth edition of *Principles of Criminology* by Edwin Sutherland in 1947. Sutherland viewed differential social organization as the explanation for group crime rates that corresponded to differential association, his individual-level theory of involvement in crime. Sutherland left the factors that comprise differential social organization largely unexplored, however, along with exactly how it should be linked to differential association. Along with some powerful criticisms of his macro theory as cultural deviance, this lack of attention by Sutherland may explain why differential social organization never received the attention that researchers and theorists alike have given to differential association. Recent work by Akers (1998), Matsueda (2006), and Sampson and Graif (2009), however, offer insights into the concept of differential social organization and address some of the questions that Sutherland left unanswered. These works, along with those of others such as Harding (2009); Browning (2009); Haynie et al. (2006); and De Coster et al. (2006), offer new avenues for exploration for those interested in explaining neighborhood

variation in rates of crime, in particular, and encourage new questions about organization both for and against crime.

Understanding Differential Social Organization

Differential social organization, sometimes referred to as differential group organization, is a central concept in the fourth edition of Edwin Sutherland's *Principles of Criminology*. Though used to explain aggregate crime rates in a way that corresponded to his individual-level theory of crime, differential association, two key questions about differential social organization were left largely unexplored by Sutherland. First, Sutherland gives little guidance for understanding the key factors important in understanding organization both for and against crime. Second, he says little about how differential social organization and differential association should be linked. Perhaps as a result of this lack of attention, the concept has received little attention from other criminologists compared to differential association. Recent work though shows that it offers interesting possibilities for shedding new light on well-researched areas such as neighborhoods and crime.

In what follows, Sutherland's development of the concept of differential social organization is outlined along with its influence on the work of Cloward and Ohlin (1960). Their expansion of his ideas gives early insight into factors important in shaping organization for crime. In addition, criticisms of Sutherland's work that might well have hampered the willingness of criminologists to give it much attention are reviewed. Attention then turns to an explication of recent work that attempts to answer questions left unexplored by Sutherland. Understanding what Sutherland means by differential social organization, however, requires a short discussion of two other concepts, differential association and culture conflict. A review of the three concepts follows.

Background Description

In 1921, Edward Hayes, then chair of the Department of Sociology at the University of Illinois, asked Edwin Sutherland to write a textbook on criminology (Gaylord and Galliher 1988). Hayes was the editor of a book series for Lippincott, and it was this request that led Sutherland to devote himself to criminology. The first edition of *Criminology*, as his text was originally titled, was published in 1924. By 1939, he included in it a statement of his theory of differential association which he revised until the publication of the final version in 1947 when the fourth edition of *Principles of Criminology* was published (Gaylord and Galliher 1988). With the publication of this edition, three theories are offered to explain crime at three different but corresponding levels – differential association (individual), differential social organization (group), and culture conflict (societal).

Differential association is the most well known, developed, and tested aspect of the three ideas. In 1947, Sutherland laid out his theory of differential association in nine propositions. The theory explains individual involvement in crime as the result of a process of learning, through association. The learning includes both techniques and "...the specific direction of motives, drives, rationalizations and attitudes" (1947, p. 6) favorable to the commission of crimes. It predicts that an individual "...becomes delinquent because of an excess of definitions favorable to law violation over definitions unfavorable to violation of the law" (6).

After the statement of the ninth proposition, Sutherland writes that what was just stated at the individual level can also be stated at the group level. Differential social organization is founded on his belief that "...crime is rooted in social organization and is an expression of that organization. A group may be organized for criminal behavior or organized against criminal behavior. Most communities are organized both for criminal and anti-criminal behavior and in that sense the crime rate is an expression of the differential group organization" (1947, pp. 8–9). He writes of the link between differential social organization and differential association that – "The person's associations are determined in a general context of social organization. A child is ordinarily reared in a family; the place of residence of the family is

determined largely by family income; and the delinquency rate is in many respects related to the rental value of the houses. Many other factors enter into this social organization, including many of the small personal group relationships" (1947, p. 8).

Sutherland did not develop his ideas about differential social organization to the extent that he did those surrounding differential association. In his discussion of differential social organization, he does not explain, for example, what the ". . .many other factors. . ." of social organization are, either for or against crime, that he considered important. However, there are two sources beyond *Principles of Criminology* that are important to examine for further insight into what Sutherland meant by differential social organization. The first is a chapter he wrote on theft during wartime published in 1943, 4 years before the fourth edition of *Principles of Criminology*. The second is what can be drawn from those individuals and works that influenced his ideas about social organization.

In a chapter entitled "Crime" in an edited volume entitled *America in Wartime*, Sutherland outlined a theory of theft during wartime that was based on the concept of differential social organization. Unhappy with the explanations offered by others, Sutherland (1943) offers differential social organization as an alternative which, he argues, can fit the known facts about theft during wartime. In doing so, he describes a number of aspects of organization both against theft and for theft that add to the meaning of differential social organization. Sutherland (1943) lists external opportunities relating to guardianship of homes and goods; socialization of children by parents, schools, and churches; supervised activities for children; attitudes towards property ownership by those who do not own much property; and clarity in the meaning of property ownership and ownership rights as factors important in the organization *against* theft. Factors important in organization *for* theft that he listed included various changes in contacts with criminal patterns. Sutherland argued that frequency of theft was increasing which increased the chance that any one individual would have contact with someone who stole and collusion of individuals with thieves, and that there had been an increase in the number of people stealing to sell to the black market. In addition, Sutherland included a cultural element in the social organization for crime. Sutherland proposed that during wartime, there were more people for whom theft was the only alternative to not surviving. Since culture allowed for this exception to the rule that people should not steal, he argued, there could be predicted an increase in theft. "The two aspects of organization together are called differential group organization. The balance between the opposed organizations determines whether the crimes committed, the reactions against crimes, and the convictions rates increase or decrease" (203). As Matsueda (2006) writes, it is clear from this entry that Sutherland saw differential social organization as consisting of a variety of different structural and cultural factors.

A final important part of Sutherland's ideas on differential social organization is found in the conclusion of this entry. Here he writes that "organization has two principal constituent elements, namely consensus in regard to objectives and implementation for the realization of objectives. . .Each of these is found in the organization for crime and the organization against crime" (203). This statement suggests that organization involves not simply structures and culture, which might be considered fairly stable, but the more dynamic process of interaction that must occur for a group to develop agreement on a common goal and take action towards achieving that goal (see Matsueda 2006 and the discussion of his work below). It is notable that these two components of social organization correspond with the definition of social disorganization used by contemporary theorists interested in neighborhood variation in crime rates that involves the ability to agree upon and work towards a common goal.

In addition to Sutherland's own writing, Matsueda (2006) argues that an examination of some of the figures whose work influenced him can give insight into the meaning of differential social organization. Most notable perhaps is the

influence of Shaw and McKay (1942) who are the founders of the theory of social disorganization. In an essay on the development of his theory, Sutherland (1942[1973]) reports having originally "borrowed" the term social disorganization from Shaw and McKay. Later, he changes this for differential social organization under the encouragement of Albert Cohen because "...the organization of the delinquent group...is social disorganization only from an ethical or some other particularistic point of view" (1942[1973], p. 21). Gaylord and Galliher (1988) later argue that he changed the terminology in response to two factors. First critiques come out of social disorganization, some of which drew attention to the fact that it seemed to be evaluating the organization of certain areas, "theirs," as deficient to "ours." They report that Sutherland wanted to get away from the negative connotation associated with the concept. Second, they argue that he found the concept differential social organization better at conveying his idea that society can be organized both for and against crime. The change in terminology does not mean that he did not see value in Shaw and McKay's ideas or their research findings, both of which he continued to cite in his text. Matsueda (2006) argues, though, that he is reconceptualizing Shaw and McKay's social disorganization to be weak organization against crime.

A second important influence on Sutherland's ideas of differential group organization is found in Tannenbaum (1938). Most broadly, Sutherland saw value in Tannenbaum's work on identity and the role of social reactions against crime in the shaping of criminal identify (Gaylord and Galliher 1988). This is connected with the influence of symbolic interactionism on Sutherland and the role it plays in differential association. Specifically in terms of differential social organization, however, his influence can be seen in Sutherland's emphasis on reactions to crime as key part of differential social organization. On the one hand, the community response to crime is part of the organization against crime. On the other, the development of a criminal identity is part of organization for crime (Matsueda 2006).

Overall then a review of Sutherland's own writings and those who influenced his work suggest that differential social organization involves social structural and cultural elements. Also involved are the processes that lead to the development of a common goal and action towards the achievement of that goal. With safety from theft and violence as a common goal, control of the behavior of others, including reactions to misbehavior, is key element of the action towards achieving that goal.

Along with differential association and differential social organization, a third concept is also important in understanding Sutherland's work. This is culture conflict or, what Cressey (1960, 1968) later came to call, normative conflict. This refers to "...conflict between legal and other norms that arise through the societal growth process..." (Cressey 1968, p. 50). For Sutherland, culture conflict was the basic explanation for crime. Culture conflict then explains why there is crime in society; differential social organization explains why some groups have higher crime rates than others in that society; and differential association explains why a particular individual is involved in crime. How all three link up is left largely undeveloped by Sutherland.

Even after reviewing Sutherland's writings and influences, and placing the concept of differential social organization in relation to differential association and culture conflict, it remains clear that Sutherland never developed his ideas on differential social organization in the systematic way that he did his ideas of differential association. This lack of development left it open to interpretation by others, both supporters and critics, in ways that influenced the interest in and thus development of his ideas by other criminologists. One particularly important development of his ideas of social organization for crime is found in the work of Cloward and Ohlin (1960). One particularly important critique is found in Kornhauser's (1978) *Social Sources of Delinquency*. A discussion of each of these works follows.

Two early scholars whose work explores the ideas of differential social organization are

Cloward and Ohlin (1960). In *Delinquency and Opportunity*, they link Merton's theory of anomie with Sutherland's differential association and differential social organization to explain the nature, development, and stability of delinquent subcultures among males in lower class areas. Their theory centers on the idea that youths want to achieve success but have differential access to opportunities to do so. Cloward and Ohlin argue that Merton's theory of anomie recognized differences in conventional opportunities for achieving success but ignored differences in opportunities for crime. While Sutherland's differential association recognized differences in opportunity to learn about criminal opportunities, he ignored differences in legitimate opportunities. They felt that inclusion of both in one theory would enhance the ability of theory to explanation of the nature and existence of various types of subcultures.

Cloward and Ohlin (1960) posit that youths in the lower class, having little in the way of legitimate opportunities for success, struggle to find options for achieving higher status. Just what options they have were argued by Cloward and Ohlin (1960) to depend in large part of the social organization of the lower class neighborhood in which they found themselves. They noted two factors in particular. The first was integration of different age levels which helped with the transmission of delinquent skills and values. The second was integration of conventional and deviant values. They argued integrating illegitimate opportunities structures and values with those of the legitimate world enhanced the stability of criminal roles. Differences in the organization of a given area in terms of opportunity structures are what lead to differences in subcultures. They describe three ideal subculture types – delinquent, conflict, and retreatist. Delinquent subcultures involve theft and organized criminal activities as alternative ways of achieving status. These emerge in neighborhoods where there exist illegitimate opportunities for achieving status because of both an integration of age levels and conventional and deviant values. Conflict subcultures, center on the use of violence, and emerge where few illegitimate opportunities exist. Finally, it is the retreatist, revolving around drugs, which consists largely of those individuals who are not able to successfully compete in either the legitimate or illegitimate opportunity structure.

By positing the existence of different subcultures and relating them to access to opportunity for learning and integration into legitimate or illegitimate life, Cloward and Ohlin were starting on a path of exploration of what factors are key to understandings differential social organization. Though there work was highly influential both in criminology and in the arena of public policy, continued criticisms of social disorganization and cultural deviance theories alike were to turn criminologists' attention away from exploring the role of social organization in understanding crime. In particular, the influential work of Kornhauser (1978) did much to turn attention of criminology, in general, and those interested in social disorganization theory, in particular, away from differential social organization.

There are two parts to her critique that are important to the discussion here. The first is her critique of the work of Shaw and McKay (1942) and their "mixed" model of neighborhood crime. She points out that they present us with a model that combines two different theories – control theory and cultural deviance. The control part of the theory predicts that in high-crime-rate neighborhoods, neighborhood controls are weak, freeing youths to be delinquent. These youth then develop delinquent traditions in the neighborhood that are transmitted to others. The cultural deviance part, then, argues that youths in neighborhoods with delinquent subcultures commit delinquent acts because of their socialization into a delinquent way of life. She argues that this mixed model is not logically possible because control and cultural deviance theories are based on differing assumptions about human nature. On the one hand, control theories assume that youths are capable of delinquency when free from control and that no other motivation is necessary. On the other, cultural deviance theories assume youths are not capable of deviance. They are merely following the norms and values of a subculture into which they have been successfully and completely

socialized. Kornhauser suggests that the cultural deviance part of the model is not necessary and, for reasons discussed below, not even logically tenable. She then presents us with a pure control model of neighborhood crime.

Kornhauser (1978) combines her argument for a social disorganization theory that is based on a pure control model with a powerful critique of cultural deviance theories. She uses Sutherland's differential association and his discussion of culture conflict as the prime example of a cultural deviance theory. Her argument starts with the assumptions she believes are behind cultural deviance theories. These are that "...man has no nature, socialization is perfectly successful, and cultural variability is unlimited" (p. 34). She then points out that the first is not actually consistent with what cultural deviance theorists argue. If socialization is perfect, she argues, it must be because people *by their nature* are perfectly adaptable. Though others have argued that she misinterprets Sutherland (see, e.g., Akers 1998), the critique resonated with others. Discussion and research on social disorganization largely followed the pure control model. Cultural deviance theories continued to fall into disfavor, including Sutherland's ideas about cultural conflict. Differential social organization received relatively little attention, and what it did receive did not differentiate between culture conflict and differential social organization. The lack of attention Sutherland gave to differential social organization relative to differential association is mirrored by the level of attention the field has given it. There are, however, several recent exceptions which show various ways of exploring this idea, and it is to these that the discussion now turns.

State of the Art

Recent work on differential social organization has offered new opportunities for development and research on the factors that may shape organization both for and against crime. Here, the focus is on three different works that contribute to the understanding of differential social organization. The first work, by Akers (1998), explores linking differential social organization with the individual-level differential association. Two others, Matsueda (2006) and Sampson and Graif (2009), critically examine various aspects of differential social organization.

Akers, who has spent his career developing and testing social learning theory as an elaboration of differential association, has now taken his work a step further by articulating the link between differential social organization and differential association. Though his initial work focused exclusively on rewriting the individual-level differential association, Akers (1998) has maintained a career long interest in integrating the individual-level explanation with the structural component. It is in the 1998 publication of *Social Learning and Social Structure* that Akers makes his most complete statement of that link as well as providing preliminary evidence regarding his proposed model. The model includes four different groups of social structural factors whose effects on crime he argues are mediated by social learning variables. The four social structural factors are differential social organization, differential location in the social structure, social disorganization and conflict, and differential location in primary, secondary, and reference groups.

Akers uses differential social organization to refer to "...ecological, community or geographical differences across systems..." (332) such as neighborhoods and the division between urban and rural. Importantly, Akers (1998) argues that differences in crime across these areas are related to structural and/or cultural differences, some of which are known and some of which are not. Differential location in the social structure refers to factors such as age, gender, race, and class that are indicative of a group's position in the social structure. Next, Akers draws from three major social structural theories to discuss the effects of social disorganization, anomie, and culture conflict. He argues that when social cohesion and integration are disrupted, disorganization and anomie result in an area or group, with a resulting increase in the crime rate. Further, social, political, and/or economic injustices may result in the development of different belief systems that lead to conflicts and, through conflicts, higher area or group rates of crime. Finally,

Akers posits that individuals are differentially located in primary, secondary, and reference groups such as family, school, and peers, in such a way as to lead to differences in social control and socialization.

As is seen above, Akers leaves unspecified much of what comprises the essential aspects of differential social organization. A recent test of the theory by Lee et al. (2004) leaves questions about the important social structural and cultural factors that make up differential social organization largely unanswered. The one measure of differential social organization is size of the community. Akers explication of the link between differential social organization and differential association, though, and his empirical tests of the link are important steps in understanding how differential social organization contributes to our understanding of individual criminal involvement.

Another theorist whose work has advanced the understanding of differential social organization is Matsueda (2006), a longtime advocate of Sutherland's work. In the current work, his focus is on adding to the theory a dynamic component that is implied, but never explicitly recognized. He draws his thinking about the dynamic aspect of organization from Sutherland's (1943) ideas that social organization involves the development of consensus over goals to be achieved and the means by which to achieve that goal. He then draws from Mead's (1934) theory of symbolic interaction and ideas of social control to discuss just how it is that consensus over goals and means is built up and turned into action. He argues Mead's ideas are consistent with Sutherland's differential social organization for both involve the development of meaning regarding goals and means as well as the development of a self that is built up over time and in interaction with others as solutions to common problems are sought.

Central to this statement is how collective action as a solution to a common problem, such as crime control or safety, is built up. Collective action, which he views as a types of social organization, is defined as "those acts commonly defined as occurring outside of institutional contexts in informal groups or gatherings, tending to be more spontaneous and creative, and requiring the building of coalitions and consensus in the absence of a strong normative system" (19). Understanding how a problem results in collective action depends on several factors including the frame and individual thresholds. Framing refers to how the problem is defined. Individual thresholds deal with how individuals vary on the proportion of people in a collective action that it takes for them to decide to join in. One example he uses to illustrate his points deals with collective efficacy. To view this in terms of process and action, he argues, involves consideration of how residents frame the issue of crime in their neighborhood, how much value individuals in the neighborhood place on safety, and how capable individuals in the neighborhood are in winning others to their view, as well as the mix of individuals with various levels of individual thresholds. Collective efficacy does not depend then just on social networks in the neighborhood then on "efficacious individuals" embedded in the network and their work in framing a problem and engaging others.

While Matsueda radically expands our understanding of the role of processes in social organization, recent work by Sampson and Graif (2009) on types of social capital illustrates important ways of thinking about key components of differential social organization across neighborhoods. Echoing somewhat Matsueda's emphasis on "efficacious individuals," Sampson and Graif distinguish between types of social capital at the resident and leadership levels of neighborhoods. They argue that the social capital that may reside in the abilities of the residents is not the same as that which is involved with neighborhood leaders and the two types of social capital do not necessarily correlate. Analyses of social capital need to distinguish then between resource possibilities of social networks, leadership in the neighborhood, and organizational capacity. Along with these structural dimensions are cultural aspects of neighborhood life including the normative climate and expectations about safety. Using these different dimensions of social capital, their analysis uncovers four different types – cosmopolitan, conduct norms, urban village, and institutional alienation – that vary on

the levels of social capital existing among residents and leaders.

Beyond these three highlighted works are others who have recently explored various aspects of social organization at the neighborhood level and its effects on crime. Browning's (2009) work on "negotiated coexistence" illustrates how the inclusion of offenders in neighborhood social networks which depend upon reciprocity and exchange can lower the ability to control the behavior of individuals in the network. Haynie et al. (2006) present a rare empirical assessment of neighborhood disadvantage and its effects on exposure to violent peers. They rightly note that while many predict neighborhood structural characteristics will affect whom youths will interact with and what they will be exposed to, empirical research testing these ideas is limited. Similarly, Harding (2009), drawing on work such as Cloward and Ohlin (1960) on integration of age groups, is interested in expanding knowledge on the transmission of culture in disadvantaged neighborhoods. He explores how neighborhood disadvantage increases time spent by adolescents with older youths. Finally, De Coster et al. (2006) are also interested in culture as found in the "street milieu." In their study, they examine that, for example, the effects of deviant peer relationships, easy access to firearms, witnessing violence, being the victim of violence, and expectations of death all affect involvement in violence by shaping the ability of families to control the children.

Controversies and Open Questions

How different types of social organization are related to crime rates across groups and areas is a question that has fascinated criminologists since the field's inception. An increasing emphasis on social process theories emerged in the 1970s, turning attention away from social structural theories as criticisms of these theories grew. Interest never died completely out though, and work on social organization continued, slowly but surely. Neighborhood research on social disorganization, collective efficacy, and social capital is an example of the continued vitality in this area. Though there is a great deal of interesting work being done, questions remain for exploration.

Researchers in interested in social disorganization theory and neighborhood crime have been exploring some aspects of neighborhood organization for and against crime since Shaw and McKay. Initially research focused on social structural factors such as poverty, mobility, and heterogeneity and connected them to social networks and organizational participation as factors important in understanding the ability of residents in a neighborhood to agree upon and work towards common goals. More recent work has radically expanded our ideas on key factors in understanding neighborhood organization – collective efficacy, bridging and bonding social capital, institutional capacity, neighborhood leadership – many of which are now being connected to neighborhood types. The focus here is on the ability to organize *against* crime found in resources and structures which aid in the development of a common goal and action towards achieving that goal. Many of the ideas need further explication and testing, but a solid ground has been laid for future work.

If Sutherland is right though, there is a need to explore the factors that shape organization *for* crime. Cloward and Ohlin (1960) saw integration of different age levels along with integration of the conventional and criminal worlds as important for opening opportunities for success in the criminal world. Harding's (2009) research on integration of age levels and its effects on exposure to neighborhood culture shows the importance of this. What other factors, structural or cultural, are key to understanding organization for crime? It seems that much of what has been done in this area is related to culture – alienation from the larger culture and the police, for example, as well as understanding of cultural adaptations to living in high-crime areas. Related to this is the need to examine how organization for and against crime are interrelated.

Perhaps the most interesting and difficult aspect of differential social organization is

grappling with the role of culture. Most seem to have moved away from notions of a deviant subcultural where residents are predicted to adhere to beliefs supportive of violence or crime to more subtle notions of rationalization and adjustments to poverty and feelings of alienation from the larger society and the police. Key to continued development is distinguishing culture and its effects from those of social structure. Work thus far, though often identifying cultural and social structure factors, often fails to distinguish them in ways that aids in our understanding of their independent effects much less their relationship. Without a careful delineation, the ability to untangle the relationship between social structural factors and their effects from those of culture will continue to limit our understanding of both.

Tackling culture leads to another question. What about the relationship between social disorganization and social control theories on the one hand and social disorganization, differential social organization, and social learning on the other? Kornhauser's (1978) powerful criticism that Shaw and McKay originally presented an untenable mixed model because of the contradictory assumptions behind control and cultural deviance seemed to have the effect of banishing consideration of culture from most analyses of neighborhood crime rates under the social disorganization framework. In fact, social disorganization is most commonly argued to be the macro level equivalent of social control theories focusing, while cultural deviance and social learning are similarly linked. More and more it seems that social control and social learning are linked processes but are these different versions of each?

Related Entries

- ▶ Differential Association Theory
- ▶ Disadvantage, Disorganization and Crime
- ▶ Early Chicago School Theory
- ▶ Neighborhood Effects and Social Networks
- ▶ Social Capital and Collective Efficacy

Recommended Reading and References

Akers R (1998) Social learning and social structure: a general theory of crime and deviance. Northeastern University Press, Boston

Browning C (2009) Illuminating the downside of social capital: negotiated coexistence, property crime, and disorder in urban neighborhoods. Am Behav Sci 52:1556–1578.

Cloward R, Ohlin L (1960) Delinquency and opportunity. Free Press, New York

Cressey D (1960) Epidemiology and individual conduct: a case from criminology. Pacific Sociol Rev 3:47–58

Cressey D (1968) Culture conflict, differential association and normative conflict. In: Marvin W (ed) Crime and culture: essays in honor of Thorsten sellin. Wiley, New York, pp 43–54

De Coster S, Heimer K, Wittrock S (2006) Neighborhood disadvantage, social capital, street context, and youth violence. Sociol Q 47:723–753

Gaylord M, Galliher J (1988) The criminology of Edwin Sutherland. Transaction Publishers, New Brunswick

Harding D (2009) Violence, older peers and the socialization of adolescent boys in disadvantaged neighborhoods. Am Sociol Rev 74:445–464

Haynie D, Silver E, Teasdale B (2006) Neighborhood characteristics, peer networks, and adolescent violence. J Quant Criminol 22:147–169

Kornhauser R (1978) Social sources of delinquency: an appraisal of analytic models. University of Chicago Press, Chicago

Lee G, Akers R, Borg M (2004) Social learning and structural factors in adolescent substance use. West Criminol Rev 5:17–34

Matsueda R (2006) Differential social organization, collective action and crime. Crime Law Soc Ch 46:3–33

Mead GH (1934) Mind, self and society. University of Chicago Press, Chicago

Sampson R, Graif C (2009) Neighborhood social capital as differential social organization: resident and leadership dimensions. Am Behav Sci 52:1579–1605

Shaw C, McKay H (1942) Juvenile delinquency and urban areas. University of Chicago Press, Chicago

Sutherland E (1942[1973]) Development of the theory. In: Karl S (ed) Edwin Sutherland: on analyzing crime. University of Chicago Press, Chicago, pp 13–29

Sutherland E (1943) Crime. In: William O (ed) American society in wartime. University of Chicago Press, Chicago, pp 185–206

Sutherland E (1957) Principles of criminology, 4th edn. J.B. Lippincott Company, Chicago, revised

Sweetser F (1942) A new emphasis in neighborhood research. Am Sociol Rev 7:525–533

Tannenbaum F (1938) Crime and the community. Ginn and Co, Boston

Disadvantage, Disorganization and Crime

Karen F. Parker and Richard Stansfield
Department of Sociology and Criminal Justice, University of Delaware, Newark, DE, USA

Overview

A fundamental community-level theory, social disorganization theory posits that crime and delinquency are more pronounced in areas characterized by persistent poverty, population heterogeneity, and residential mobility, which combine to disturb the capacity of neighborhoods to maintain informal social control. These ideas have been well investigated and empirically supported, leading social disorganization theory to become the most well-known theory of neighborhood crime in the field today. Yet as time passes, scholars have turned their attention to the ways that central cities have changed. This includes the influx of service-based economies to city centers, the spatial concentration of poverty, and how joblessness has overcome the economic prospects and hopes of so many urban residents. And while the removal of manufacturing industries affected jobless rates, preexisting racial residential segregation meant the ill effects of deindustrialization were predominantly felt among minorities and the poor. It is these arguments, directly associated with Disadvantage theory, that have led researchers to incorporate both disorganization and disadvantage into their community-level research on crime.

The interconnections between disadvantage and disorganization are the central focus of this entry. After providing a brief review of the fundamental ideas driving social disorganization and disadvantage theories, some of the advantages to incorporating both approaches jointly when attempting to capture community conditions, neighborhood processes, and crime rates will be discussed. At the same, it is important to acknowledge that neighborhoods are not influenced just by structures and process within (Kubrin and Weitzer 2003; Sampson et al. 2002). Community studies of disadvantage and disorganization must also account for changes in the larger society, such as political changes in crime policies, growing diversity in the US population, and severe downturns in the economy and/or times of economic recession. To illustrate, the 1990s saw unprecedented levels of immigration from Central and Southern America. The majority of new arrivals settled in urban neighborhoods within a few port cities. Yet despite facing high levels of community disadvantage in these neighborhoods, the arrival of new immigrant groups has not led to higher rates of crime, raising questions about additional factors overlooked by community-level research. These critical issues are offered as aspects for future research or open questions, as they have significant implications for disorganization and disadvantage.

The Fundamental of Social Disorganization and Urban Disadvantage Theories

Social Disorganization Theory

One of the most fundamental approaches to the study of violence emanates from the Chicago school research of Shaw and McKay. In essence, Shaw and McKay (1942) argued that neighborhood dynamics lead to social disorganization in communities, which account for the variations in crime and delinquency. Under this line of inquiry, Shaw and Mckay (1931, 1942) set out to explain variation in offending informed largely by the work of Park et al. (1925). Park was a proponent of the concentric zone modeling of cities, observing that industries tend to be located within inner-city zones and cities grow naturally from the inside out. Shaw and McKay believed that characteristics of the inner-city zone could explain delinquency and crime due to three elements – low economic status, ethnic heterogeneity, and residential mobility. First, communities with low economic status lack adequate money and resources to generate the needed formal and informal controls necessary to reduce crime and delinquency. Second, racial and ethnic

heterogeneity, which is often accompanied by fear and mistrust, impedes communication and patterns of interaction in a community. Third, residential mobility was said to disrupt a community's network of social relations by acting as a barrier to the development of extensive friendship networks, kinship bonds, and local social ties.

Believing the conditions of disorganization to be structurally conducive to crime rates, Shaw and McKay (1931) examined robbery and mental illness and found that rates were higher in the inner-city zones, where residents lived adjacent to industry in conditions of population growth and environmental hazard. Rates progressively declined, moving away from the inner city into transitional zones however, and lowest of all on the outskirts of the city in the commuter zone. An equally important aspect of their work was that rates of delinquency remained high in the inner-city zones regardless of the ethnic group residing in those areas. That is, they suggested that structural conditions of disorganization were the key to understanding urban crime, not the characteristics or motivations of individuals. Over time, "social disorganization" became known as referring to the inability of a community structure to realize the common values of its residents and maintain effective social control (Bursik 1988). That is, social disorganization addressed the structural barriers that impede the development of the formal and informal ties that promote the ability of the community to solve common problems.

The social disorganization model however struggled to hold popularity within criminology throughout much of the 1970s and 1980s, not least because as Bursik and Grasmick (1993) argued, the authors failed to supply a refined concept of social disorganization and how its constituent parts were each related to crime and delinquency. In an effort to revitalize interest in the approach, Bursik (1988) noted 5 key reasons for the failure of social disorganization to be applicable to contemporary society and criminology. These included normative assumptions made about organization in neighborhoods and the dependency on official sources of data for testing social disorganization, noting that many neighborhoods have their own informal mechanisms of dealing with delinquency; the study was cross-sectional, yet cities are dynamic and not static; and simply at the time, criminology was more interested in theories of crime focused on explaining individual tendencies to offend. Perhaps the key argument, however, was that social disorganization had not reached its full potential as a community theory yet could do so if these flaws were addressed in future work.

A renewed interest in the theory has been witnessed and research flourishes today largely because of the ability of criminologists such as Robert Sampson to show the relevance of Shaw and McKay for understanding crime in contemporary society. This renewed interest has led to some important developments. First, researchers have extended the types of structural conditions in urban areas that impede social control, including population size, poverty, and family disruption (see Sampson and Groves 1989). As another example, Bursik and Grasmick (1993) expand Shaw and McKay's claims of the inability of communities to maintain social control, offering a "systemic model" linking community control to crime rates at three levels: private (friendships or households), parochial (organizational ties), and public (citywide efforts to regulate neighborhoods via police presence).

Third, the mediating linkages between structural conditions, social control, and crime rates have been further articulated. While Shaw and McKay identified the capacity of the community to control group-level dynamics as a key mechanism linking community characteristics with crime rates, Sampson and Groves (1989) clarified the types of social networks in a community at both the informal (e.g., friendship ties) and formal (e.g., participation in community organizations) level. They suggested that communities unable to (1) control teenage groups through collective social control, (2) form informal local friendship networks, and (3) offer local participation in formal and voluntary organizations will experience high rates of crime and delinquency (Sampson and Groves 1989). The authors found that structural characteristics did predict levels of

disorganization, and in turn, the weakly organized communities had higher rates of crime. These findings were replicated 10 years later using updated responses from the British Crime Survey (Lowenkamp et al. 2003) highlighting the strength and stability of the findings.

More recently, Sampson et al. (1997) conceptualize community ties a bit differently to include the inability of community residents to collectively display trust and deal with problems (such as disorderly teens or drugs) via "collective efficacy." Collective efficacy has been found to largely mediate the relationship between structural disadvantage and crime (Sampson et al. 1997). Not only does this new line of inquiry show why problems differentially exist between neighborhoods, but it also shows that the neighborhoods experiencing problems are less able to do anything about it. For example, the focus of trust and cohesion suggests that when thinking about intervening and exercising informal social control – neighbors are able to know the likelihood of a neighbor collaborating or backing them up based on trust.

The dynamic element of collective efficacy thus makes an enormous leap forward in explaining the disorganization, social control, and crime relationship. This is important given that the direct evidence on traditional disorganization factors such as residential instability and ethnic heterogeneity has been more mixed in recent research, particularly in the case of ethnic heterogeneity (Morenoff et al. 2001; Sampson et al. 2002). Yet there are limitations, particularly when identifying the key factors behind precisely how or why ties are activated and mobilized for the purpose of social control. Recent research also questions the causal pathway of disorganization and crime, in which crime may cause feedback effects on disorganization, disorder, and a community's capacity to build social ties (Kubrin and Weitzer 2003). Thus, while these efforts expand upon and extend social disorganization theory, the field continues to call for further conceptual refinement, inclusion of other factors rarely examined by disorganization researchers, and rigorous methodology to unpack the dynamic process between disorganization and crime.

Disadvantage Theory

Numerous scholars have documented the growing disadvantages in urban areas and how these disadvantages are spatially concentrated in poor black areas particularly. Building on this foundation, criminology has attended to the structural aspects of concentrated disadvantage in producing urban crime. Wilson (1987) claims that the increase in economic marginalization of blacks in the inner city was due, in a large part, to a set of spatial and industrial changes in the political economy. In *The Truly Disadvantaged*, he points to deindustrialization, coupled with suburbanization of middle-class African Americans, as central to the rise in poverty and social isolation inner-city residents. Central to his arguments, Wilson illustrates the dramatic shift away from local manufacturing-based economies during the 1970s and 1980s, particularly in the Midwest and Northeast regions of the United States (Kasarda 1995). Advances made in technological industries contributed to this shift, as did the motivation to secure cheaper labor and the growing demand for service industries. That is, the process of deindustrialization marks a change in the industrial mix of urban areas or, more specifically, when the share of jobs shift from manufacturing to administrative and information services in many American cities (Kasarda 1995). The process of deindustrialization had a number of adverse outcomes on city residents.

First, the movement of manufacturing jobs into the suburbs or overseas and shift in industrial mix of local economies resulted in a spatial mismatch, which emphasizes the geographic changes in the number of low-skilled jobs in the inner city and limited mobility of African Americans due to residential segregation. Essentially a "mismatch" occurs as residential segregation prevents blacks from following employers to the suburbs. The low wages and earnings associated with manufacturing work served as a disincentive for inner-city residents to seek longer commutes for work outside their neighborhoods and city boundaries. Yet inside city boundaries, as central cities are transformed into service-based economies and centers for administration and information processing

(Kasarda 1995), these areas see a reduction in the number of potential employment opportunities available to the low-skilled worker. That is, the requirements of the new economy also created a "skills mismatch" between workers and the service jobs available, further disadvantaging blacks living in the inner city.

Second, while the removal of manufacturing jobs affected unemployment or jobless rates, preexisting segregation in labor markets through discrimination in hiring and denying equal opportunity meant that the ill effects of deindustrialization were predominantly felt among minorities and the poor. The shift in labor demand had real consequences for minority groups and women. Scholars have linked the high rates of joblessness among black males to the transformation of the US economy or industrial restructuring (Wilson 1987). And while the literature examining the influence of restructuring on women is small and inconclusive, some research suggests women benefited from industrial restructuring, as the shift increased the number of jobs in female-dominated service occupations, while other scholars argue that black women lost rather than gained with the expansion of service-based industries. For example, Smith and Tienda (1987) find that Latino and black women faced greater declines in jobs than Asian and white women, suggesting that the already high levels of economic marginalization among women was further concentrated among minorities. Adding to the crippling effects of industrial restructuring on women, the disabling of welfare or the elimination of federal welfare entitlements such as AFDC to black women with children whom were disproportionately represented, only further guaranteed that African American women would be vulnerable to shifts in the local economies. To capture the social transformation of urban areas, Wilson's term "concentration effects" reflects the continued deterioration of employment opportunities and job networks, decline in schools, and diminishing number of marriageable partners, thus contributing to higher levels of family dissolution, which only further aggravated the weak labor attachments and conventional role models for inner-city residents (Wilson 1987, p. 58).

Taken together, these scholars focus on how blacks are residentially concentrated in poor urban areas with limited access to job networks and opportunities to participate in formal local labor markets. Racial residential segregation and black economic disadvantage (joblessness, poverty concentration, and deindustrialization) are thus considered the macrolevel sources of racial discrimination in the urban environment, leaving blacks to face a more disadvantaged residential environment than the poor of other racial groups. A wealth of recent empirical work has documented that these precise structural conditions are uniquely related to urban violence (Krivo and Peterson 1996; Parker 2008).

In sum, there is strong evidence that factors of concentrated disadvantage and racial isolation remain direct predictors of many criminal outcomes. But while these measures provide a telling story in their own unique relationships to crime, they are not insulated from measures of disorganization. Dynamic neighborhood processes, particularly those related to collective efficacy, are not produced in a vacuum; rather, they appear to emerge mainly in environments with sufficient socioeconomic resources and residential stability (Sampson et al. 2002). As Sampson and Wilson (1995, p. 53) conclude, "the intersection of race, place, and poverty goes to the heart of our theoretical concerns with society and community organization."

Disadvantage and Disorganization: Why They Should Be Examined Together?

This section discusses the benefits and advantages to the understanding of community conditions and crime by assessing disadvantage and disorganization jointly. While neglecting other important issues, arguments here focus on three major advantages: (1) by placing the focus on local communities rather than individual circumstances or motives, we more fully understand the structural processes that lead to crime; (2) together, the role that race plays in the USA becomes more lucid, as race is largely infused with economic status and place throughout the

United States; and (3) the ability to move beyond cross-sectional examination of crime rates to incorporate how changes occur in communities and crime over time.

Community Focus

Sampson and Wilson (1995) largely accepted the tenets of Shaw and McKay's disorganization theory; however, they disagreed that the disorganization and growth out of the city was part of a "natural process" that cities went through. Instead they argued that racial discrimination, and the consignment of African Americans into inner-city neighborhoods based on economic and political decisions, was the key to understanding disorganization and crime in contemporary American society. Through making this connection, Sampson and Wilson have helped illustrate the connection and overlap between disadvantage and disorganization. A wealth of structural research has also suggested that variations in disorganization are intimately linked to racial inequality and racial segregation (Peterson et al. 2006). Accordingly, recent research has utilized disadvantage and disorganization together to provide a greater understanding of community processes. Grattet (2009), for example, uses Wilson's notion of concentrated disadvantage to understand social disorganization and defended neighborhoods, noting that neighborhoods without adequate resources lack the legitimate means (and informal control mechanisms) of defending its area from outsiders.

This is a positive trend in research, moving away from an exclusive focus on structural factors toward placing the community at the heart of the discussion of disadvantage and disorganization. As Sampson and Wilson (1995) describe, disadvantaged communities are not only more likely to be deprived of resources necessary to mobilize crime control, but the resulting isolation of communities affects the capacity to form cohesive networks based on mutual trust and common goals. Due to political and economic forces, the disappearance of jobs (Wilson 1987, 1996), and the increasing social isolation of the urban poor, cultural values in these communities have had the potential to shift away from middle-class norms of achieving success to subcultural norms where the pursuit of status and respect replaces conventional means (Anderson 1999). Indeed as Anderson describes, the structural conditions of the neighborhood are linked to public cynicism and trust in formal social control and serve to reshape residents' values. Studies from this perspective, including the work of Elijah Anderson (1999), highlights that structural conditions and cultural responses to these conditions in which people find themselves jointly shape neighborhood organization and neighborhood crime. As those with the misfortune of living in segregated, isolated, and disorganized communities are also then exposed to conditions and values that may permit their criminal involvement, a community focus is critical.

Race, Ethnicity, and Place

Racial groups continue to face different social and economic realities (see, e.g., Sampson and Wilson 1995; Wilson 1987). Discrimination and racial competition in opportunity structures are core issues that stratify racial groups in the urban context. That is, structural forms of racial discrimination, such as residential isolation, poverty concentration, and segregated labor markets, not only stratify groups spatially but can serve as a source of conflict between them. Disorganization and disadvantage arguments allow race to be examined as a structural yet local, as opposed to, individual or biological feature of communities. In this way, this literature advances the study of race, urban inequality, and violence by highlighting the core conditions that lead to racial differences in life circumstances and crime within neighborhoods and US cities.

Second, by combining theories of disorganization and disadvantage, scholars acknowledge the persistent disadvantaged position of African Americans as a structural feature of urban areas. For example, Massey and colleagues (see Massey and Denton 1993) provide evidence of whites' exclusion of black participation in the local labor

market and whites' resistance to blacks residing in integrated communities. Research also reveals the inequalities across race and ethnic groups as they compete in the labor force and face differential treatment as a result of economic restructuring. A striking feature of contemporary United States is that black employment remains lower than white employment despite recent narrowing of educational and occupational inequalities. Another important finding is that blacks face greater barriers to residential mobility than other race and ethnic groups. According to Massey and Denton (1993), blacks remain spatially isolated and residentially segregated from whites at all levels of economic status, while Hispanics and Asians have experienced clear improvements as their economic status increases. Similarly, Wilson (1987) argues that blacks reside in areas of extreme poverty concentration, a reality not known to poor whites.

By examining disorganization and disadvantage together, one gains insights into the ways that race becomes embedded in the structural characteristics of areas, resulting in the racial disparities in crime rates so often found in empirical studies. Yet as Sampson and Wilson (1995) note, the ability of researchers to make meaningful comparisons about the conditions that produce racial violence is significantly hindered by the fact that so few whites reside in areas of concentrated disadvantage when compared to blacks. Thus, as research continues to explore the interconnections between disorganization and disadvantage, it is important to consider not only the level of disparities between racial groups in structural characteristics, opportunity structures like access to jobs, and rates of crime and delinquency but also what larger historical and society forces are allowing those disparities to persist.

Change

Fundamental to Chicago-style criminology is its attention to dynamic changes in urban neighborhoods and crime. In fact, the changing nature of, and level of spatial interdependency between, neighborhoods is embedded in the social disorganization tradition (Morenoff et al. 2001). Yet the element of change, once core to the theory, has been largely neglected in the existing research. In an effort to sway researchers, Byrne and Sampson (1986, p. 17) called for longitudinal analysis when they stated: "By far the vast majority of ecological studies have examined the relative effects of structural characteristics on crime in cross-sectional analysis at one point in time. This is somewhat ironic given that classic ecological theory is concerned with the processes of change in urban areas." Some 20 years later, Sampson (2002) reiterates this troubling trend toward cross-section research, further arguing that the lack of research has caused a series of fundamental questions concerning our understanding of cities and communities.

Cross-sectional designs largely dominate the macrolevel research on crime. In fact, one can find few examples of empirical research that examines changes in macrolevel characteristics and crime rates over time (see examples in Messner et al. 2005). By far the greatest impediment to longitudinal analysis is data, although some researchers have overcome this issue. Conducting national time-series analyses of white and black arrests rates, LaFree and colleagues find differences in the influence of economic indicators on white and black arrest rates over time. For example, based on their longitudinal analysis between 1957 and 1990, LaFree and Drass (1996) found that income inequality (as a measure of resource deprivation) has a positive impact on the changes in white and black arrest rates, but absolute economic well-being (their composite measure that combines median income and unemployment) did not significantly influence homicide rates for either racial group. While studies have begun to examine changes at the city level (or higher levels of aggregation), research focusing on community-level changes is still largely missing from the literature. Thus, examining changes in both communities and crime over time is central to disorganization and disadvantage arguments. The full set of conditions which may lead to

disorganization can only be captured when examining long-term processes of urban disadvantage and development (Kubrin and Weitzer 2003).

Future Directions and Open Questions for Disadvantage and Disorganization

While much has been gained by combining arguments of disadvantage and disorganization in the study of urban crime, more work is warranted. So much of the empirical work on communities and crime has focused on how to delineate the connections between community characteristics and crime within the community itself. However, larger societal processes heavily influence communities as well. Three areas that should be accounted for in future work on disadvantage and disorganization include (1) trends in Hispanic immigration and growing diversity, (2) the move toward more conservation crime policy and the impact of mass incarceration on urban communities, and (3) the need to expand the ways we measure the local economy, particularly as the economic recession looms on.

Hispanic Immigration and Growing Diversity

As discussed, structural factors such as racial segregation act to concentrate and isolate disadvantaged minorities into communities with high rates of population turnover, systematic flight of the most qualified residents, few professional employment opportunities, and high rates of family disruption (Bursik 1988; Massey and Denton 1993; Sampson and Wilson 1995; Wilson 1996). As new immigrants have typically settled in these communities, and because immigration may be associated with conflicting norms and the inability of people to agree to common values and reach common goals (Bursik 1988; Sampson et al. 1997), recent immigration has been argued to predict higher rates of crime. New immigrants may also face additional economic barriers and blocked opportunities which make it difficult to abide by conventional societal norms. As an example, new arrivals may initially be faced with limited opportunities entering into the labor market due to language barriers, a lack of well-paying jobs within close proximity, and discrimination (Martinez 2002).

Despite the fact that new immigrants often reside in communities that are characterized by disadvantage and disorganization, research tends to show that the presence of new Latino immigrants results in either a negligible or negative effect on crime (see, e.g., Martinez 2002; Ousey and Kubrin 2009). Many explain the "Latino paradox" as the result of the importance placed on family and religious conservatism in Latino communities, which may counteract the criminogenic effect of the economic disadvantage that often accompanies the immigration experience. The emphasis placed on family and religion may serve to lower crime by increasing informal social control on family members through a greater sense of obligation, reference group identity, and community support. This may be particularly important for young Hispanics, a group that is disproportionately large as compared to the white population and at risk for criminal offending due to the link between age and crime.

Contemporary immigration is also vastly different from immigration of Shaw and McKay's day, not least because of the human capital and financial resources many immigrants bring. Even among the less-skilled and less-educated recent arrivals, research has documented high rates of employment among recent Hispanic immigrants, even if the employment happens to be at a low-wage level (Martinez 2002). Jobs that are considered "undesirable" and/or those offering wages that are low by American standards are often acceptable to recently arrived immigrant laborers as they are offer much better wages than similar jobs in their country of origin. In addition to basic economic sustenance, the higher rates of Hispanic labor market involvement are important for an understanding of offending patterns because employment provides stability and regulation of daily behavior by requiring that people follow consistent schedules and spend time engaging in conventional activities (Wilson 1996).

Information emerging from this body of immigration-crime literature thus provides

interesting caveats to traditional theories of disadvantage and disorganization. Specifically, whether there is a positive or negative relationship of neighborhood conditions and crime for Hispanics may depend on a number of factors such as local labor market structures, the presence of racial discrimination facing different groups, and the presence of a dual frame of reference in a community, through which residents evaluate their socioeconomic conditions relative to past realities in their country of origin (Martinez 2002). The presence of vibrant co-ethnic communities awaiting recent arrivals, as in the case of the Cuban enclave in Miami, for example, offers another example of how some communities may be able to provide a protection against outside forces of discrimination while offering employment opportunities through informal social networks within. A challenge for criminologists lies in how to use social disorganization theory and theories of disadvantage together in order to explain how one ethnic group may have overcome conditions of disadvantage while many low-income blacks have yet to do so.

Crime Policy and Mass Incarceration

Another important open question for scholars of disorganization and disadvantage is the extent to which structural characteristics influence arrest and incarceration and the resulting implications back on the community itself. Surprisingly, criminologists have paid little attention to the determinants of arrest rates across time and space, despite the strong tradition of neighborhood studies examining the relationship between social and economic inequality and crime arrests (Mosher 2001). Yet it is hard to deny the impact of an era of mass imprisonment and the war on drugs on neighborhood organization and disadvantage. Mass incarceration may have one of two key implications for ecological analyses studying social disorganization and disadvantage. Firstly, in communities of high disadvantage, a significant proportion of the population may no longer be present in the neighborhood, as they were three decades ago. Recent research has also revealed that important distinct relationships exist between concentrated disadvantage and the use of social control against blacks, when compared to whites (Parker et al. 2005). These authors argue that the concentration of black disadvantage and racial isolation may influence black arrests because these factors serve to amplify perceived group differences, thus increasing antiblack dispositions, which in turn may increase the pressure on police to exert control. Given the racial and economic disparities in the experience of arrests and incarceration, some communities have experienced more concentrated levels of incarceration than others. And there have been well-documented neighborhood problems associated with the removal of significant numbers in a community, such as the removal of male role models for children and employable residents. If estimates of Clear et al. (2001) are any indication, among the worse-off communities as much as 30 % of the population may be gone.

In addition to the role crime policy and mass incarceration has played in removing family and community role models from neighborhoods, it is as important to consider that most of those incarcerated during the mass incarceration era are eventually released back into the community. This issue of reentry into communities, and its implication for the organization of neighborhoods, is also critical to consider. In interviews in Tallahassee, Florida, Clear and colleagues (2001) target understanding the implications of this process of removing and returning offenders, for social networks, social capital, and subsequently informal social control. The impacts go far beyond financial and fundamentally alter the shared sense of identity, interpersonal networks, self-esteem levels, and a community's reputation as a good place to live. The importance of both of these arguments is brought to light if we consider the more recent work on collective efficacy (Sampson et al. 1997). As a theory focusing on trust, community ties, and relationships with neighbors, both the removal and reentry of convicted community members are likely to have significant implications for family disadvantage, informal neighborhood networks, mutual trust, and shared goals in the neighborhood. Based on recent data from the National Institute of Justice that suggests that

the era of mass incarceration may be stabilizing, and possibly even ending, considerations of mass incarceration and crime policy must be a critical element to future studies of social disorganization and disadvantage.

Economic Downturn and Times of Recession

Given the historical importance of residential patterns, economic conditions, and other community-level measures of disadvantage or disorganization, it is not surprising that the looming economic recession presents researchers with an important call to expand the ways we measure the local economy. To illustrate, the massive trend in housing foreclosures nationally could easily contribute to neighborhood conditions such as residential mobility and disadvantage. Research has begun to examine the possible effects of foreclosure on neighborhood levels of crime. As Katz et al. (2011) discuss, the concern is less about the act of foreclosure itself and more about the neighborhood conditions that a foreclosure is believed to create, compromising the social environment much in the same joblessness has been thought to do (e.g., Wilson 1996). In this vein, one may expect that neighborhoods with high rates of foreclosure will similarly begin to undergo a process in which residents withdraw from a community, both physically and psychologically, and in turn, informal social control mechanisms break down. A conclusion on the importance of such measures may be, at best, premature, especially as the preliminary evidence suggests that foreclosed homes may not have the long-term negative impact on crime that many scholars have feared (Wallace et al. 2012).

Nevertheless, within the current economic climate, new measures of disadvantage and disorganization can be helpful to expanding our knowledge of what structural factors are important to consider within particular time and spatial contexts. Traditional measures of disadvantage, social ties, and collective efficacy have proven effective, but may not fully account for variation in neighborhood crime in times of recession or severe economic downturn. Other factors are thus also important to consider, like housing foreclosure and lacking consumer sentiment. By studying the interconnections between disorganization and disadvantage, particularly as influenced by macroeconomic conditions, wider social patterns in migration, and in the context of decisions by political elites, the utility of disorganization and disadvantage remains more pertinent than ever.

Related Entries

- ▶ Crime and the Racial Composition of Communities
- ▶ Early Chicago School Theory
- ▶ History of Criminological Theories: Causes of Crime
- ▶ Neighborhood Effects and Social Networks
- ▶ Poverty, Inequality, and Area Differences in Crime
- ▶ Race and Ethnicity in Social Disorganization Theory
- ▶ School Structural Characteristics and Crime
- ▶ Social Disorganization and Terrorism
- ▶ Understanding Victimization Frequency

Recommended Reading and References

Anderson E (1999) Code of the street: decency, violence, and the moral life of the inner-city. WW Norton, New York

Burgess EW (1967) The growth of the city: an introduction to a research project. In: Park R, Burgess E, McKenzie RD (eds) The city. University of Chicago Press, Chicago

Bursik RJ Jr (1988) Social disorganization and theories of crime and delinquency: problems and prospects. Criminology 26:519–551

Bursik RJ, Grasmick HG (1993) Neighborhoods and crime: the dimensions of effective community control. Lexington, New York

Byrne JM, Sampson RJ (1986) The sociology ecology of crime. Springer, New York

Clear TR, Rose DR, Ryder JA (2001) Incarceration and the community: the problem of removing and returning offenders. Criminology 47(3):335–351

Grattet R (2009) The urban ecology of bias crime: a study of disorganized and defended neighborhoods. Soc Probl 56(1):132–150

Kasarda JD (1995) Industrial restructuring and the changing location of jobs. In: Harrell AV, Peterson GE (eds) Drugs, crime, and social isolation. The Urban Institute, Washington, DC

Krivo L, Peterson RD (1996) Extremely disadvantaged neighborhoods and urban crime. Soc Forces 75:619–648
Kubrin CE, Weitzer R (2003) New directions in social disorganization theory. J Res Crime Delinq 40(4):374–402
Lafree G, Drass K (1996) The effect of changes in intraracial income inequality and educational attainment on changes in arrest rates for African Americans and whites, 1957–1990. Am Sociol Rev 61(4):614–640
Lowenkamp CT, Cullen FT, Pratt TC (2003) Replicating Sampson and Groves' test of social disorganization theory: revisiting a criminological classic. J Res Crime Delinq 40:351–373
Martinez R Jr (2002) Latino homicide: immigration, violence and community. Routledge, New York
Massey D, Denton N (1993) American apartheid: segregation and the making of the underclass. Harvard University Press, Cambridge, MA
Messner SF, Deane GD, Anselin L, Nelson BP (2005) Locating the Vanguard in Rising and Falling Homicide Rates Across U.S. Cities. Criminology 43:661–696
Messner SF, Raffalovich L, McMillan R (2001) Economic deprivation and changes in homicide arrest rates for white and black youths, 1967–1998. Criminology 39(3):591–614
Morenoff JD, Sampson RJ, Raudenbush SW (2001) Neighborhood inequality, collective efficacy and the spatial dynamics of urban violence. Criminology 39:517–558
Mosher C (2001) Predicting drug arrest rates: conflict and social disorganization perspectives. Crime Delinq 47(1):84–104
Ousey GC, Kubrin CE (2009) Exploring the connection between immigration and violent crime rates in U.S. cities, 1980–2000. Soc Probl 56:447–473
Park R, Burgess EW, McKenzie RD (1925) The City. University of Chicago Press, Chicago
Parker KF (2008) Unequal crime decline: theorizing race, urban inequality and criminal violence. New York University Press, New York
Parker KF, Stults BJ, Rice SK (2005) Racial threat, concentrated disadvantage and social control: considering the macro-level sources of variation in arrests. Criminology 43(4):1111–1134
Peterson RD, Krivo LJ, Browning CR (2006) Segregation and Race/Ethnic Inequality in Crime: New Directions. In: Cullen FT, Wright JP, Blevins KR (eds) Taking Stock: The Status of Criminological Theory. Transaction Publishers, New Brunswick
Sampson RJ, Groves B (1989) Community structure and crime: testing social disorganization. Am J Sociol 94:774–802
Sampson RJ, Wilson WJ (1995) Toward a theory of race, crime, and urban inequality. In: Hagan J, Peterson RD (eds) Crime and inequality. Stanford University Press, Stanford
Sampson RJ, Raudenbush SW, Earls F (1997) Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277:918–924
Sampson RJ, Morenoff JD, Gannon-Rowley T (2002) Assessing "neighborhood effects": social processes and new directions in research. Annu Rev Sociol 28:443–478
Shaw CR, McKay HD (1931) Social factors in juvenile delinquency. Report on the causes of crime, national commission of law observance and enforcement. GPO, Washington, DC
Shaw CR, McKay HD (1942) Juvenile delinquency and urban areas. University of Chicago Press, Chicago
Smith SA, Tienda M (1987) The doubly disadvantaged: women of color in the US labor force. In: Stromberg A, Harkess S (eds) Working women, 2nd edn. Mayfield, Palo Alto
Wallace D, Hedberg EC, Katz CM (2012) The impact of foreclosures on neighborhood disorder before and during the housing crisis: Testing the spiral of decay. Social Science Quarterly 93(3):625–647
Wilson WJ (1987) The truly disadvantaged: the inner city, the underclass, and public policy. University of Chicago Press, Chicago
Wilson WJ (1996) When work disappears: the world of the new urban poor. Alfred A Knopf, New York

Disclosure

▶ Naming and Shaming of Corporate Offenders

Discrete Spatial Choice

▶ Crime Location Choice

Discriminant Validity of Disorder and Crime

Jacinta M. Gau[1] and Travis C. Pratt[2]
[1]Department of Criminal Justice, University of Central Florida, Orlando, FL, USA
[2]School of Criminology and Criminal Justice, Arizona State University, Phoenix, AZ, USA

Synonyms

Order maintenance; Quality of life; Zero tolerance

Overview

The broken windows thesis posits a causal relationship between disorder and crime whereby disorder occurs initially and, if left unchecked, ultimately produces serious crime. This causal sequence requires that disorder and crime be empirically separable phenomena; that is, they must possess discriminant validity. If discriminant validity is absent, then the broken windows thesis is fatally flawed because disorder cannot be said to cause crime if disorder and crime are two facets of the same underlying problem. This entry reviews three studies that have spoken to the issue of discriminant validity and have concluded that disorder and crime are not, in fact, unambiguously distinct factors. The future of the broken windows thesis is thus uncertain, and if order maintenance policing does reduce some types of crime, this might be accomplished through a net-widening effect rather than through the informal social control mechanisms that are keys to the thesis.

Introduction

For policy analysts, the question "Where does crime come from?" is subordinate to "What institution of control can be utilized to reduce crime?" In their 1982 *Atlantic Monthly* article that laid out the central tenets of what would become known as the "broken windows thesis," policy analysts James Q. Wilson and George Kelling selected the police as the institution that would get the limelight. Wilson and Kelling argued that the police were the keys to crime reduction and prevention at the community level and that the best way for them to accomplish this would be to proactively target disorderly people and conditions.

Today, the idea of proactive, order maintenance policing has become part of the common vernacular in discussions of police goals and activities. In 1982, however, Wilson and Kelling faced a serious obstacle that impeded their proposed line of reasoning: The professional model of policing and several constitutional rulings issued by the US Supreme Court had made it unacceptable and legally questionable for police officers to interfere in matters not pertaining directly to serious crime. Wilson and Kelling, then, could not merely assert that police should start concentrating their efforts on reducing annoying yet nondangerous, noncriminal behaviors like loitering and panhandling – police leaders and policymakers would have pointed to the various industry standards and court rulings and thrown their hands up at the suggestion. Wilson and Kelling therefore had to be creative.

And they were. They laid out a formula wherein (a) disorder was uncoupled from serious crime and (b) disorder led to serious crime in a causal fashion. The reasonableness of the claim that disorder is well within the legitimate purview of police enforcement activities was thus made obvious. It boiled down to a simple syllogism: Police are responsible for controlling crime, and disorder causes crime if nobody does anything about it, so police are responsible for controlling disorder.

The problem is that Wilson and Kelling offered no empirical evidence for the unlinking of disorder and crime; in fact, there really was no reason to think that Wilson and Kelling's formulation of the disorder-crime relationship was any more valid than, say, Garofalo and Laub (1979) contention that people's fear of "crime" is actually a very general feeling of anxiety and apprehension about neighborhood-level conditions of sociostructural malaise wherein crime and disorder are part of overarching judgments about quality of life. Wilson and Kelling, then, did not *prove* that disorder and crime are distinct from one another and that disorder causes crime – they *speculated* that such was the case.

Wilson and Kelling's oversight may not have been problematic had others stepped up to fill in the gaps, but this did not happen. Instead, the disorder-crime distinction and supposed causal sequencing was accepted as fact, and efforts were launched to put broken windows policing into action in the streets. Doing so violated the basic principle of scientific inquiry that requires theories and their constituent predictions (i.e., hypotheses) to be empirically tested repeatedly before being taken

as valid. Scientific testing is hardly a minor detail. The broken windows thesis has assumed a position of preeminence in policing without any strong evidence that Wilson and Kelling were right. Worrall (2006a, p. 379) put it well when he argued that "just because theorists have pushed for the separation of crime and incivility indicators does not mean both concepts are distinct... the incivilities [broken windows] thesis has evolved with little attention to the validity of its measures."

This entry assesses the evidence pertaining to the *discriminant validity* (i.e., the empirical uniqueness) of disorder and crime. Discriminant validity is said to be present when two factors that are theorized to be separate and unique from one another actually do display these qualities in empirical tests. The question of discriminant validity is vital to the broken windows thesis (BWT) because this framework posits that disorder is an independent variable and crime is a dependent variable. This means that these two phenomena must be related but distinct and separable from one another. If disorder and crime possess discriminant validity, then this portion of the broken windows thesis holds up; if they do not, then the BWT argument is seriously weakened. Disorder can only cause crime if disorder is *not* crime, and crime, likewise, can only be caused by disorder if crime is *not* disorder.

Disorder and Crime Under the Broken Windows Thesis

There is no shortage of practitioners and researchers who claim that the BWT has been firmly empirically supported (e.g., Kelling and Coles 1996; Kelling and Sousa 2001; Skogan 1990). Most of the studies claiming to find support for the thesis, however, have not actually tested the thesis itself but, rather, have assessed the crime-reduction impact of police efforts to crack down on disorder. As will be discussed later, some evidence does suggest that order maintenance policing may be promising under certain conditions; however, it is not at all clear that the mechanism by which order maintenance policing may effectively reduce disorder and crime is actually that specified by the broken windows thesis. Researchers who have studied the thesis itself and the hypotheses derived from it have found little empirical support for it and have offered several reasons why the disorder-crime causal link posited by the BWT may be flawed (Harcourt 2001; Harcourt and Ludwig 2006; Sampson and Raudenbush 1999, 2004; Taylor 2001).

Perhaps the biggest challenge to most of the existing broken windows studies is researchers' tendency to rely on cross-sectional data. Cross-sectional data are those that are gathered at a single time point, such as when a survey asks respondents to report their perceptions about various neighborhood and community characteristics. Cross-sectional research designs make the establishment of causation difficult because there is no time lapse between the measurement of the independent variable and that of the dependent variable, and thus, it is very easy to mistake correlation (i.e., simple, noncausal coincidence) for causation. To the extent that areas plagued by serious crime also suffer high levels of social and physical disorder (Sampson and Raudenbush 2004), cross-sectional research may merely capture the convergence of these two phenomena in space and time and erroneously portray this contemporaneousness as evidence of causality. Alternatively, Taylor's (2001) longitudinal study showed that disorder measured at one time point did not consistently or strongly impact crime measured at a later date. This severely undermined the broken windows thesis and called its core tenet into serious question.

In the search for reasons why disorder may not, in fact, cause crime in the manner advocated by Wilson and Kelling (1982), the issue of discriminant validity stands prominent. As explained in the preceding paragraphs, disorder can only be a predictor of crime if disorder and crime are empirically distinct phenomena. Content overlap between these two factors would violate the principles of scientific inquiry because the independent and dependent variables would essentially be the same thing.

Some proponents of the broken windows thesis might argue that discriminant validity is not a prerequisite for the veracity of the claim that

disorder causes crime. Two lines of argument might be advanced in this regard. First, it might be proposed that an absence of discriminant validity would merely signal that the broken windows process has occurred – disorder went unchecked, disorder caused crime, and now crime and disorder are both prevalent in the area. This line of reasoning, however, tautologically begs the question of whether disorder causes crime, as merely pointing to the co-occurrence of the two problems proves neither that disorder temporally and causally precedes crime nor that the two problem types are empirically distinguishable.

Second, it could be contended that disorder and crime are more like a continuum of problems than two separate issues, with crime being on the more serious side of the continuum and disorder on the less serious side. The idea of such a continuum is, in and of itself, quite plausible; however, it contradicts the logic of the BWT. The continuum model would render broken windows unfalsifiable by making it impossible to figure out exactly what the independent and dependent variables are. An accurate empirical test could never be conducted because there would always be doubt as to whether something that was measured as crime should have actually been considered disorder or vice versa. The crucial assumption made by Wilson and Kelling is that disorder causes crime. This premise requires that disorder and crime be two different problems; that is, it necessitates discriminant validity.

It is immediately apparent that this requirement is violated to at least some extent in the BWT, because many of the behaviors the thesis identifies as disorder (e.g., prostitution, tagging, vandalism, minor drug dealing) are crimes by statute or ordinance. There is, therefore, overlap between disorder and crime from a legal or definitional standpoint. This fact provides preliminary evidence against the scientific validity of the broken windows thesis.

The only way that the broken windows thesis can survive this dilemma is if disorder and crime are *perceived* as being unambiguously different in the public's view. Broken windows is grounded in social psychology and hinges on subjective perceptions and on the emotions that people feel and the inferences they draw when they see either the presence or absence of disorder in their neighborhoods of residence. In other words, disorder does not exist in some kind of objective "reality" or "truth" (Piquero 1999; Sampson and Raudenbush 2004), but, rather, disorder takes on whatever meaning is ascribed to it by local residents. If people *see* a difference between disorder and crime, then broken windows might be a viable theory of crime production despite the aforementioned statutory overlap.

One of the most significant demonstrations of the similarity between disorder and crime comes from Sampson and Raudenbush (1999), who relied on systematic social observations of disorder rather than on survey-based perceptual measures. Their results clearly and consistently showed a weak relationship between neighborhood-level disorder and crime once structural covariates were controlled. The only exception to this pattern was with respect to the crime of robbery, which was, interestingly, the only crime type linked to disorder in Skogan's study published in 1990 (Harcourt 2001). Moreover, it appeared that disorder and crime both sprung from the common origin of weak collective efficacy – one of the most consistently robust community-level predictors of such problems (Pratt and Cullen 2005). Far from the causal sequence predicted by the BWT, then, it appeared that disorder and crime were both "symptoms" of the same "disease": the breakdown of a community's ability to regulate its group members and enforce standards of civility and legal conformity.

Testing for discriminant validity between disorder and crime requires the use of a statistical modeling procedure that was created specifically for the purpose of testing proposed factor structures (e.g., Gau 2010). This analytical technique is called confirmatory factor analysis (CFA). Confirmatory factor analysis has allowed tests of discriminant validity within the BWT framework; without this modeling tool, tests of this sort would be much more difficult and the results would likely be harder to interpret. Because CFA is so central to the discriminant validity debate, the next section will briefly

outline this analytical strategy and will be followed by summaries of the studies that have employed CFA to test for discriminant validity.

Confirmatory Factor Analysis

Confirmatory factor analysis is a statistical procedure that permits extensive examination of factor structure. A *factor* is something that is not observed directly but accounts for the similarity between a set of objects. For instance, a fire truck, a stop sign, and the pen a professor uses to grade student exams may all be objects representative of the factor "red in color." Certain variables or objects can be indicators of different factors depending on the full item set. If the fire truck were put into a group with a police car and an ambulance, then the factor would be "emergency vehicles." Confirmatory factor analysis permits researchers to propose hypothesized models by saying, "I think *this* variable (e.g., vandalism) is part of *this* factor (e.g., disorder) and I think that *that* variable (e.g., robbery) is part of *that* factor (e.g., crime)" and to then test these predictions.

The quality of a hypothesized model is assessed using fit indices and factor loadings. Fit indices are established criteria that have certain maximum or minimum values – depending on the index – that are indicative of good model fit. Factor loadings are similar to regression coefficients and measure the strength of the relationship between a variable and the factor it is predicted to represent. In the context of tests for discriminant validity in broken windows research, this type of validity would be present when all of the fit indices met their respective thresholds for good fit and the items thought to measure disorder all loaded highly on the disorder factor *and not* on the crime factor, while the crime items loaded well on crime *and not* on disorder. If the fit indices or factor loadings do not meet these requirements, then there is a problem with the model and discriminant validity might be absent.

Confirmatory factor analysis is a popular statistical procedure in psychological research and has gained recognition and use in criminal justice and criminology, as well. The software required to run these analyses is widely available, reasonably priced (as far as software goes), and sports user-friendly, point-and-click interfaces (Gau 2010). There is, in short, no good reason for CFA to not be a staple of broken windows research, which makes it surprising that only a handful of studies have used this method to assess disorder and crime factors for measurement validity (Armstrong and Katz 2010; Gau and Pratt 2008; Piquero 1999; Worrall 2006a). The next section details the studies that have used CFA to test for discriminant validity in perceptual measures of disorder and crime.

Empirical Inquiries into the Discriminant Validity of Disorder and Crime

Worrall (2006a) and Armstrong and Katz (2010) addressed the issue of discriminant validity by entering variables tapping survey respondents' perceptions of disorder, crime, and personal victimization experiences into a series of CFA models. Worrall employed data from 12 cities in 12 different states, and Armstrong and Katz used a sample of survey respondents in a single city in Arizona. The results of both studies painted a murky, inconclusive picture. Survey respondents did not make clear, consistent distinctions between the three phenomena, but neither did they unambiguously see them as one and the same. One-factor models positing disorder as being a single factor along with crime or victimization did not display good fit, yet neither did two-factor models allowing disorder to function as a factor apart from the other two. Overall, then, the authors of both studies concluded that discriminant validity had not been established and that this key premise of the broken windows thesis was not empirically supported.

Gau and Pratt (2008) added to the debate by examining survey data from respondents in neighborhoods in 21 municipalities in Washington State. The variables of interest came from 17 crime and disorder items that asked respondents to indicate whether each issue was *1 = no problem*, *2 = uncertain*, *3 = a problem*, or *4 = a serious problem* in their neighborhoods.

While some of the items were obvious indicators of crime (e.g., gun violence) and others were clearly measures of disorder (e.g., noise), many others (e.g., vandalism, youth gangs) were ambiguous and did not lend themselves clearly to either category. The authors therefore relied on prior literature (Giacopassi and Forde 2000; Reisig and Parks 2004; Sampson and Raudenbush 1999; Taylor 2001) to construct the disorder and crime factors. The disorder factor contained the following items: dogs running at large, drunk drivers on the road, people drinking to excess in public, groups of teens or others hanging out and harassing people, youth gangs, people using illegal drugs, vandalism, noise, traffic problems, and garbage/litter. The crime indicators were people's homes being broken into and things being stolen, people being robbed or having their purses/wallets taken, domestic/intimate partner violence, rape/sex crimes, child abuse, violent crime, and gun violence.

Gau and Pratt (2008) constructed two different models so that each one could be tested against the other to determine which one seemed superior to the other. This strategy is necessary for valid hypothesis testing; it is not a good idea to run only one model because then it is impossible to determine whether the tested model truly is the best one or whether some other model that was not tested might fit better. The first model that Gau and Pratt specified predicted that all of the disorder and crime variables were part of a single factor. This model represented the prediction that people see no difference between disorder and crime and that they instead tend to lump them together as part of the general condition of the neighborhood. Support for this model would undermine the broken windows thesis. The second model predicted that disorder and crime were two separate, distinct factors, each measured by the variables listed above. This model represented the original broken windows framework and predicted that disorder and crime possess discriminant validity. Support for this model would strengthen the broken windows thesis.

Both models displayed similar fit indices that offered lukewarm validation of each of them. Some indices lent the impression that the models were good, while others fell short of their ideal values. On the basis of these alone, then, it was not clear whether the one-factor model or the two-factor model was the better representation of crime and disorder.

There is, however, another way to evaluate model fit above and beyond fit indices: the between-factor correlation. A between-factor correlation of 0.85 or greater indicates that the factors cannot be justifiably separated (see Gomez et al. 2005) because discriminant validity is unacceptably low. It becomes impossible to tell whether items that are purported to be indicators of one factor are not, in fact, equally plausible indicators of the other factor. Gau and Pratt (2008) discovered that the correlation between the crime and disorder factors was 0.92. The two-factor model, therefore, had to be rejected because the correlation between the crime and disorder factors made factor separation indefensible. Perceptions of crime and perceptions of disorder appeared to constitute a single factor; that is, discriminant validity was not present.

It is possible that discriminant validity in individual-level perceptions of crime and disorder is only part of the story; that is, the separability of disorder and crime might be influenced by neighborhood – or community-level conditions. Wilson and Kelling (1982) hinted that there might be a tipping point whereupon a neighborhood becomes so overrun by disorder and crime that police cannot do much other than try to keep pace with calls for service. The tipping point concept may apply to resident perceptions, too. People living in areas where disorder has been present for some time may hardly notice these incivilities because they are simply part of the neighborhood landscape, whereas people residing in areas that have been disorder-free in the past may react very strongly when even the slightest sign of incivility springs up. In other words, the broken windows thesis might be more applicable to some neighborhoods or communities than to others, depending upon extant levels of disorder in those places.

Gau and Pratt (2010) investigated the possibility of spatially patterned discriminant validity by creating a difference score that

captured the extent to which people saw disorder levels as being higher, lower, or similar to crime levels in their neighborhoods. They regressed the difference score on several variables, including respondents' perceptions of disorder and crime. The results indicated that the difference between crime and disorder increased as perceived neighborhood-level disorder increased. The authors then split the sample into low-disorder and high-disorder subsamples to test for interaction effects. Among the low-disorder group, increases in disorder caused a reduction in the difference score, while among the high-disorder group, increases in disorder were associated with a closing of the gap. Evidence for a perceptual tipping point was thus uncovered: People living in areas where disorder is mild or absent are more sensitive, and when they see visible signs of incivility, they are likely to interpret these signs as ominous harbingers portending the arrival of serious criminals. People accustomed to seeing disorder in their neighborhoods, on the other hand, were better able to make a distinction between disorder and crime and to more clearly see them as two separate phenomena.

Implications for the Broken Windows Thesis and Order Maintenance Policing

The absence of clear, consistent discriminant validity in the three CFA-based studies described above renders indefensible the claim that disorder is an independent variable and crime is a dependent variable. These two phenomena are no doubt interrelated; however, this relationship is likely correlational rather than causational (see also Sampson and Raudenbush 1999). The possibility that people's ability to discriminate between disorder and crime is dependent upon the amount of disorder present in their neighborhoods of residence (Gau and Pratt 2010) adds another layer of complexity to the matter. These findings have important implications for criminological theory, the measurement of concepts in broken windows research, and crime control policy.

Theoretical and Methodological Implications

The broken windows thesis insists that people see disorder as a visible indicator that the community is spiraling out of control. If, however, people do not differentiate between crime and disorder, then crime itself could serve as the visible indicator of the lack of informal social control in a community. If this is so, then broken windows theory is untenable because it is tautological to claim that crime causes itself.

The only way for the BWT to survive is for there to be a careful refinement of the measures of disorder and crime (Kubrin 2008). Wilson and Kelling (1982) outlined an interesting theoretical framework but they failed to take the quintessential step of defining their terms. It is perhaps not surprising, then, that there has been wide variation in the items used as indicators of disorder and of crime (see Armstrong and Katz 2010; Gau and Pratt 2008, 2010; Giacopassi and Forde 2000; Reisig and Parks 2004; Sampson and Raudenbush 1999; Taylor 2001; Worrall 2006a). It is difficult to test for a relationship between an independent variable and a dependent variable when one is not sure what the independent variable actually is or how it should be measured.

It also might simply be time to retire the BWT from active duty on the criminological stage. There is at this point no solid reason to believe – and plenty of reason *not* to believe – that disorder causes crime in the manner stated by Wilson and Kelling (1982). In addition to the apparent absence of discriminant validity between the BWT's core theoretical constructs, there are identifiable conceptual problems with the thesis. The broken windows idea was based on the notion that if a rock is thrown through a window and the damage done to the window is not repaired immediately, then pretty soon all the windows will be shattered. This seems logical on its surface, but there is a troubling question embedded in this scenario: Who threw the rock? If people are prowling about a neighborhood and maliciously throwing rocks at things, then it seems that a problem is already in place and that the newly smashed window is incidental to deeper, more pressing issues. Misplacing attention on the window diverts effort and resources away from

identifying and addressing the genuine, underlying problems in a community that allow both disorder *and* crime to take root and flourish (Sampson and Raudenbush 1999).

Policy Implications

As stated at the outset of this entry, the broken windows thesis is a policy-based proposal (hence the limited attention paid to theoretical and measurement issues). Wilson and Kelling (1982, p. 36) made police the centerpiece of disorder and crime reduction by arguing that "Though citizens can do a great deal, the police are plainly key to order-maintenance." Broken windows theory has had a profound effect on the field of policing and has revolutionized some departments' missions and functions.

There is evidence that broken windows policing (often also called order maintenance policing, quality-of-life policing, proactive policing, and zero tolerance policing, though some broken windows advocates take issue with this last label due to its negative connotation) can reduce some types of crime, some of the time, in some jurisdictions. Sampson and Cohen (1988) discovered a relationship between order maintenance policing and reductions in robbery rates, though they were not entirely sure that this effect was attributable to deterrence or to broken windows with respect to the underlying theoretical mechanism. Braga et al. (1999) found that a disorder crackdown seemed to reduce crime, yet paradoxically did not reduce disorder or nuisance offending. Worrall's (2006b) analysis indicated that arrest rates for public order offenses had an immediate suppressive impact on assault, a delayed impact on burglary, and no effect on larceny. Kubrin et al. (2010) discovered a relationship between proactive policing and robbery rates. On the other hand, Novak et al. (1999) studied a disorder crackdown and found no evidence that the crackdown reduced any type of serious crime.

An important aspect of existing evaluations of disorder-based policing is the absence of attention paid to the mediating variables that were central to Wilson and Kelling's (1982) original specification of the BWT. True order maintenance policing that reduced crime in a way consistent with Wilson and Kelling's predictions would do so by enhancing social cohesion among area residents and stimulating social interaction and the frequenting of public spaces by law-abiding persons whose collective informal surveillance provided guardianship for all those present. Absent evidence that cohesion and informal social control emerged as a result of a concerted disorder-reduction effort by police, then order maintenance policing is empirically indistinguishable from crackdowns and hot-spots policing (e.g., see Sherman and Weisburd (1995), where a randomized crackdown-backoff tactic in crime hot spots reduced crime and disorder simply through sporadic, unpredictable increases in police presence).

Another alternative theoretical mechanism by which order maintenance policing might reduce crime in a manner inconsistent with the BWT was hinted at by Kelling and Coles (1996). They noted that nearly one quarter of the people arrested by police for disorderly behaviors were either carrying illegal weapons or had outstanding arrest warrants in connection to serious crimes. Order maintenance policing, then, might simply be a form of net widening that legitimizes police interference in people's freedom of movement for myriad reasons and thus permits them greater power to investigate people they feel as somehow suspicious or worthy of scrutiny. Net widening might accomplish the goals laid out by the BWT, but not via the same means, and therefore, it is not proper to call this "broken windows policing," strictly speaking.

The evidence to date that disorder and crime lack discriminant validity results also bears on the issue of citizen satisfaction with the police. If citizens do not see disorder as a problem apart from crime, then they are unlikely to understand the logic behind order maintenance policing. Victims of crimes like robbery and residential burglary may be unlikely to fully endorse police crackdowns on jaywalking and littering. Indeed, though citizens rate order maintenance as an important endeavor (Skogan and Hartnett 1997), they still see traditional law enforcement against serious crimes as the primary role of the police (Skogan 1990; Webb and Katz 1997).

A tunnel-vision focus on disorder could discredit police in the eyes of community residents.

Finally, the possibility that the discriminant validity of disorder and crime is dependent upon neighborhood context (specifically, the amount of disorder already present in the area) implies that order maintenance policing may be appropriate for better-off areas and inappropriate for more disadvantaged neighborhoods. Bucolic neighborhoods luxuriously free of dire problems might be an optimal place for police to adopt a quality-of-life approach in order to maintain the status quo and prevent disorder from seeping in. These are the places, too, where residents are most likely to hold positive attitudes toward police and to be willing to help police keep the social order in check.

By contrast, areas wherein disorder has already left its mark may, perhaps ironically, be the worst candidates for order maintenance policing simply because of the volume and severity of incivilities present in the area. Order maintenance policing places vast discretion in the hands of police and charges them with a mandate to intervene in the lives of people who do not pose an immediate or obvious threat to public safety. Exacerbating the danger inherent in the combination of discretion and low visibility is that order maintenance has been interpreted by many police departments as being best accomplished via the widespread use of *Terry*-type stop and frisks for suspicious behavior.

The convergence of these conditions – discretion, a disorder mandate, and reliance on *Terry* stops – has implications for police treatment of citizens in disadvantaged areas. Neighborhood-level disadvantage may encourage police misconduct (Kane 2002) and the use of more severe forms of coercive force (Terrill and Reisig 2003) by giving police the impression that the entire area has descended into lawlessness (Klinger 1997) and by offering the promise of impunity for bad deeds, given the social and political powerlessness of the people who reside there. In some disadvantaged, high-crime areas, the use of *Terry* stops has been racially discriminatory (Fagan and Davies 2000). Already-rocky relationships between the police and community can be further strained when officers stop people for what the people themselves see as "nothing" and, especially, when officers are rude and disrespectful during these encounters. This can worsen existing tensions and lead to even more crime and disorder as police systematically undercut their own legitimacy and authority (Gau and Brunson 2010).

Conclusion

When Wilson and Kelling wrote their *Atlantic Monthly* magazine article in 1982, they wanted the police to do something that the police could not do at that time: focus their efforts on the eradication of disorder. The professional model of policing and several US Supreme Court rulings had limited police duties to reactive response to calls for service and to the identification and arrest of people suspected of having committed serious crimes. In order to legitimize a broader conceptualization of the police role and to credibly argue that police can and should intervene in disorderly activities, Wilson and Kelling posited a restructuring of the disorder-crime relationship such that these two problems were linked sequentially and causally with disorder being a precursor to crime. The broken windows thesis thus casts disorder as an independent variable and crime as a dependent variable.

In order for the BWT's formulation of disorder and crime to be accurate, though, disorder and crime must be empirically separate phenomena; that is, they must possess discriminant validity. Preliminary evidence that discriminant validity is not present arises from the fact that many activities technically classified as disorder are, in fact, criminal by statute or ordinance. The broken windows thesis, however, is grounded in subjective, individual perceptions of disorder, so the thesis could survive the legal criticism if people see a difference irrespective of legal overlap.

Studies employing confirmatory factor analysis – a statistical modeling procedure that permits tests of factor structure and examinations of discriminant validity – have found either mixed support or no support for the central

BWT proposition that people discriminate between disorder and crime. Gau and Pratt (2008) discovered that the disorder and crime factors were too highly correlated to justify separation and that disorder and crime were, therefore, both part of a single factor. Worrall (2006a); Armstrong and Katz (2010) reported that one – and two-factor models alike fell short of the criteria necessary to establish good fit. Discriminant validity was either weak or was absent entirely across the different sets of models the authors ran. Finally, Gau and Pratt (2010) constructed a score measuring the difference between people's perceptions of disorder and those of crime and found that neighborhood disorder was a significant predictor of that difference score such that people living in low-disorder areas had trouble distinguishing between the two problem types, while those in high-disorder areas did seem to draw a line between them.

The apparent absence of discriminant validity between disorder and crime, and the evidence suggesting that perceptual discrimination may vary by neighborhood type, has several implications for the broken windows thesis and the policing strategy arising therefrom. The thesis itself needs to be refined and its measures more thoroughly defined. More radically but nonetheless on the table as an option, the thesis needs to be abandoned in favor of alternative explanations for the covariance between disorder and crime that are better conceptualized and, therefore, hold more promise for theoretical validity and useful policy. Order maintenance policing might, moreover, hold more promise in areas wherein disorder is not currently present and can be kept out of the neighborhood in a preventive fashion. Once disorder has taken hold, however, a new strategy might be required.

In sum, the broken windows thesis has focused national attention on disorder and crime. It is clear from a long line of research that both matter greatly to people's fearfulness and their satisfaction with their neighborhoods of residence; however, it is time to question the causal, disorder-to-crime sequence posited by Wilson and Kelling's (1982) broken windows thesis. It simply might not be true. Researchers and practitioners should consider alternative theoretical models that hold promise for reducing disorder and crime and improving people's quality of life.

Related Entries

- ▶ Broken Windows Thesis
- ▶ Defining Disorder
- ▶ Disorder: Observational and Perceptual Measures
- ▶ Social Disorder and Physical Disorder at Places

Recommended Reading and References

Armstrong T, Katz C (2010) Further evidence on the discriminant validity of perceptual incivilities measures. Justice Q 27:280–304

Braga AA, Weisburd DL, Waring EL, Mazerolle LG, Spelman W, Gajewski F (1999) Problem-oriented policing in violent crime places: a randomized controlled experiment. Criminology 37:541–580

Fagan J, Davies G (2000) Street cops and broken windows: *Terry*, race, and disorder in New York City. Fordham Urban Law J 28:457–504

Garofalo J, Laub J (1979) The fear of crime: broadening our perspective. Victimology 3:242–253

Gau JM (2010) Basic principles and practices of structural equation modeling in criminal justice and criminology research. J Crim Justice Educ 21:136–151

Gau JM, Brunson RK (2010) Procedural justice and order maintenance policing: a study of inner-city young men's perceptions of police legitimacy. Justice Q 27:255–279

Gau JM, Pratt TC (2008) Broken windows or window dressing? Citizens' (in)ability to tell the difference between disorder and crime. Criminol Public Policy 7:163–194

Gau JM, Pratt TC (2010) Revisiting broken windows theory: examining the sources of the discriminant validity of perceived disorder and crime. J Crim Justice 38:758–766

Giacopassi D, Forde DR (2000) Broken windows, crumpled fenders, and crime. J Crim Justice 28:397–405

Gomez R, Burns LG, Walsh JA, Hafetz N (2005) A multitrait-multisource confirmatory factor analytic approach to the construct validity of ADHD and ODD rating scales with Malaysian children. J Abnorm Child Psychol 33:241–254

Harcourt BE (2001) Illusion of order: the false promise of broken windows policing. Harvard University Press, Cambridge

Harcourt BE, Ludwig J (2006) Broken windows: new evidence from New York City and a five-city social experiment. Univ Chicago Law Rev 73:271–320
Kane RJ (2002) The social ecology of police misconduct. Criminology 40:867–896
Kelling GL, Coles CM (1996) Fixing broken windows. Simon & Schuster, New York
Kelling GL, Sousa WH Jr (2001) Do police matter? An analysis of the impact of New York City's police reforms. The Manhattan Institute, New York
Klinger DA (1997) Negotiating order in patrol work: an ecological theory of police response to deviance. Criminology 35:277–306
Kubrin CE (2008) Making order of disorder: a call for conceptual clarity. Criminol Public Policy 7: 203–214
Kubrin CE, Messner SF, Deane G, McGeever K, Stucky TD (2010) Proactive policing and robbery rates across U.S. cities. Criminology 48:57–97
Novak KJ, Hartman JL, Holsinger AM, Turner MG (1999) The effects of aggressive policing of disorder on serious crime. Polic An Int J Police Strateg Manag 22:171–189
Piquero A (1999) The validity of incivility measures in public housing. Justice Q 16:793–818
Pratt TC, Cullen FT (2005) Assessing macro-level predictors and theories of crime: a meta-analysis. In: Tonry M (ed) Crime and justice, vol 32. University of Chicago Press, Chicago, pp 373–450
Reisig MD, Parks RB (2004) Can community policing help the truly disadvantaged? Crime Delinq 50:139–167
Sampson RJ, Cohen J (1988) Deterrent effects of the police on crime: a replication and theoretical extension. Law Soc Rev 22:163–189
Sampson RJ, Raudenbush SW (1999) Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. Am J Sociol 105:603–651
Sampson RJ, Raudenbush SW (2004) Seeing disorder: neighborhood stigma and the social construction of broken windows. Soc Psychol Q 67:319–342
Sherman LW, Weisburd D (1995) General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial. Justice Q 12:625–648
Skogan WG (1990) Disorder and decline: crime and the spiral of decay in American neighborhoods. University of California Press, Berkeley
Skogan WG, Hartnett SM (1997) Community policing, Chicago style. Oxford University Press, New York
Taylor RB (2001) Breaking away from broken windows. Westview Press, Boulder
Terrill W, Reisig MD (2003) Neighborhood context and police use of force. J Res Crime Delinq 40:291–321
Webb VJ, Katz CM (1997) Citizen ratings of the importance of community policing activities. Polic An Int J Police Strateg Manag 20:7–23
Wilson JQ, Kelling GL (1982) Broken windows: the police and neighborhood safety. The Atlantic Monthly, March 29-38
Worrall JL (2006a) The discriminant validity of perceptual incivility measures. Justice Q 23:360–383
Worrall JL (2006b) Does targeting minor offenses reduce serious crime? A provisional, affirmative answer based on an analysis of county-level data. Police Q 9:47–72
Xu Y, Fiedler ML, Flaming KH (2005) Discovering the impact of community policing: the broken windows thesis, collective efficacy, and citizens' judgment. J Res Crime Delinq 42:147–186

Disorder: Observational and Perceptual Measures

Donald L. Anthony and Douglas D. Perkins
Program in Community Research and Action, Peabody College, Vanderbilt University, Nashville, TN, USA

Overview

This entry provides a synopsis of disorder-related research methods, measures, and analytic frameworks. The entry is intended to give readers a better understanding of trends in disorder-related research and the many facets of disorder analysis; aspects covered in the literature range from the various ways in which individuals and communities conceptualize disorder to the roles of homeownership, organizational participation, and land use in predicting crime in neighborhoods. Methods employed in extensive studies in Chicago, Baltimore, and Salt Lake City are cited here for illustrative purposes. The entry concludes with a summary of major findings from those studies as well as suggestions for future disorder-related research. Literature on disorder frequently refers to social or physical *incivilities* (Taylor and Hale 1986), and this entry will use the terms *disorder* and *incivilities* interchangeably.

The Evolution of Disorder Research

Conceptualizations of neighborhood-level disorder and its association with crime have evolved

over the last half century; likewise, methodologies employed in disorder-related research have evolved as well. Early research into disorder and crime was heavily influenced by Chicago-school sociologists such as Gerald D. Suttles, whose 1968 work *The Social Order of the Slums* examined the social and economic conditions of a poor Chicago neighborhood. Suttles employed a mixed-methods approach, analyzing quantitative demographic data and presenting ethnographic accounts of everyday life in the Chicago slums.

With Wilson and Kelling's advancement of incivilities theory in a 1982 issue of *The Atlantic*, interest in disorder-related research was renewed. As incivilities theory became a mainstay of the social sciences, disorder-related studies sought to test and refine incivilities theory largely by employing sharpened research methods and analytical tools. Methodological advances have allowed researchers and audiences to better understand connections between disorder and crime. Over the last half century, research on disorder and crime has employed diverse data collection methods and analytic frameworks. While gathering residents' perceptions and conducting passive surveys are still important in disorder-related research, recent studies tend to rely more on objective measures and active means of data gathering.

Analytic methods have been greatly improved by newer technologies. Geographic information systems (GIS) and hierarchical linear modeling (HLM) software are two examples of such technological advances; GIS gives researchers the ability to analyze spatial patterns and produce informative and interactive graphic representations of their findings, while HLM allows researchers to explore multiple levels of data and effects in an efficient manner. The following sections summarize trends in disorder-related research. These trends include clarification of the units of analysis, evolution of measurement tools, and advances in methods for analyzing data. Extensive studies undertaken by researchers in Chicago, Baltimore, and Salt Lake City serve as primary settings; other exemplary studies are cited when appropriate.

Units of Analysis

Neighborhoods have been conceptualized in many ways ranging from strictly geographic definitions to delineations that account for social and cultural contexts. Disorder-related studies tend toward the former, though questions about the appropriate scope for geographic boundaries persist. Early disorder-related studies typically grouped neighborhoods based on Census Bureau-defined boundaries. The use of such delineations has been problematic for researchers because census-defined neighborhoods – particularly census tracts – often include artificial or imposed boundaries. Overly broad census-defined neighborhoods may cross major roads, railroad tracks, and water bodies. Neighborhoods delineated in such a manner often extend far beyond finer, resident-recognized neighborhood boundaries such as street blocks.

Recognizing the need to define neighborhoods in ways that match residents' own definitions and are amenable to research, recent studies have focused on census-defined block groups (a collection of contiguous census blocks) and an even more precise unit of analysis, individual street block faces. When the basic unit of analysis is an individual street block face, a resident's neighbors are those on either side of one's street bounded by cross streets. (By "block" we generally mean street block as opposed to polygonal census block, which does not include neighbors across the street.) By examining lower geographic levels such as blocks and block groups, researchers have been able to better account for demographic characteristics that tend to be consistent within compact areas but vary – sometimes wildly – within larger areas such as census tracts. The studies highlighted in this entry utilized blocks as the basic units for their analyses. In the Baltimore study, for example, the authors focused their attention on 50 randomly selected street blocks (Perkins et al. 1992; Perkins and Taylor 1996). In the Salt Lake City study, researchers focused on 58 blocks in a transitional suburban setting (Brown et al. 2004b). Sampson and Raudenbush (1999, 2004) used the block group as the basic neighborhood delineator in a series of studies on disorder

and crime in Chicago; however, like the studies in Baltimore and Salt Lake City, the Chicago-based study relied on individual street blocks for comparative purposes.

Using the street block as the primary unit of analysis gives researchers the ability to compare data at both the block level and the individual or household level. Such a multilevel approach allows researchers to determine the extent to which disorder as well as demographic and other potential mediating factors, such as collective efficacy, operate at the individual or block level. This is critically important because if virtually all the variance is at the individual level, it suggests that the impact of disorder and its mediators are mainly personal and subjective. If substantial block-level effects exist on top of any individual variation, it essentially shows that those effects are more than simply a matter of individual perception or psychology. They are "real" and "objective" and have an even greater impact insofar as being shared by one's neighbors. To obtain block-level data, researchers often aggregate individual or household-level data. Although less common, it is helpful to also identify and obtain measures that are not just aggregated but commensurate to the level being studied (Shinn 1990).

Data Gathering and Measurement Tools

Just as conceptualizations of neighborhood delineation have evolved, data gathering and measurement tools have changed as well. Disorder-related research has become increasingly focused on employing multiple methods of data gathering and measurement, often relying on advanced quantitative methods not available to previous generations of researchers. Technological and theoretical advances have offered researchers new ways of collecting and examining data. Video recordings, for example, allow researchers to verify their observations of street life and other neighborhood conditions. Thus, the *objectivity* of data-gathering methods has greatly improved.

Aggregation is not the only way to obtain more objective measures. Recent research has also compared residents' perceptions and independent measures of disorder and crime. While residents' perceptions continue to be important indicators, such subjective data do not form the sole basis on which disorder is measured. Disorder can be more objectively measured with trained independent observers rating settings using predefined scales or checklists. Crime can be measured by analyzing police reports, although such reports have often been charged with political bias and there is often an information gap since crimes are not always reported. Thus, even with the emphasis on objective measures, researchers tend to see a continuing need for resident surveys and interviews.

As disorder-related research evolved, researchers sought to increase objectivity by employing independent observers to conduct "windshield surveys" (Taylor et al. 1985). These surveys were typically conducted by independent raters who drove through neighborhoods and either made open comments or graded neighborhood characteristics on a number of predetermined criteria. While windshield surveys can be useful in a cursory sense, they may lead to inadvertent oversights on the part of the raters. Walking observations may be more useful as they allow the rater to take a slower pace and make closer observations, often while having impromptu interactions with residents and others in the neighborhood; such interactions can provide valuable insights not typically gleaned through passive windshield surveys.

Researchers have created various measures and scales for conducting independent visual surveys. In their study of physical decay in low, moderate, and high-income Baltimore neighborhoods, Taylor et al. (1985) conducted assessments of the physical conditions of 808 street blocks. Independent raters evaluated a wide-ranging set of neighborhood factors including building setbacks, traffic volume, types of land uses, graffiti, and building vacancies. The researchers created three neighborhood-level subscales using aggregated data collected in the block assessments. A street accessibility subscale aggregated measures such as street width, traffic volume, and street lighting. A land use subscale aggregated measures related to type of land

use, that is, residential or nonresidential. A final subscale, density, included measures such as block length and building height. By aggregating data into three subscales, the authors were able to draw comparisons across blocks.

Other observational tools have been used to measure both physical and social disorder. One such tool is the Block Environmental Inventory (BEI). Researchers studying Baltimore neighborhoods in 1987 employed the BEI as a means of examining the social and physical conditions of 50 residential blocks, most of which were assessed by two trained raters to allow for inter-rater reliability tests (Perkins et al. 1992). The BEI included observations at both the house and block levels. At the house level, the BEI consisted of three main categories: incivilities, defensible space, and territorial markers. The incivilities subscale included items such as litter, vandalism, and dilapidation. The defensible space subscale included visibility, barriers, lighting, and security bars. The territoriality subscale consisted of signs warning of dogs, sitting places, plantings, and home decorations. Block-level items included people outside (by gender, activity, and general age group), open lot frontage (e.g., playgrounds, public gardens, institutional yards), abandoned cars, street trees, and vacant lots. The raters themselves were also part of the BEI; they were to observe whether people noticed them as they assessed the neighborhood (Perkins et al. 1992).

Another means of objective data gathering is through the use of Systematic Social Observation (SSO). Sampson and Raudenbush (1999, 2004) employed this method in a study of Chicago neighborhoods. The SSO method included the videotaping of social and physical disorder in 23,816 block faces across 80 neighborhoods. The authors asserted that SSO was an important methodological advance because it could be replicated. Of the 23,816 block faces that were videotaped by independent raters, 15,141 were coded and subsequently utilized in the study. Notably, because of the extensive coding required, SSO may be cost and/or time prohibitive for researchers on unfunded or underfunded projects.

Resident surveys continue to be popular in disorder-related research. Sampson and colleagues (1989, 1999, 2004) employed resident surveys in both Chicago and the United Kingdom to determine residents' perceptions of disorder and crime. Likewise, Perkins and colleagues (1990, 1993) in Brooklyn and Queens, New York; Perkins and Taylor (1992, 1996) in Baltimore; and Brown, Perkins, and Brown (2004a, b) in Salt Lake City employed resident surveys primarily as a means of comparing resident perceptions to independent observations. Surveys, particularly face-to-face interviews, often contextualize objective data, giving researchers unique insights into the subject neighborhoods.

Researchers have employed several analytical tools for determining actual occurrences of crime. Sampson and Raudenbush (2004) relied on publicly available data to help them objectively measure both crime and demographics. Particularly, they relied on Census Bureau data and police records of violent crimes. Likewise, Taylor et al. (1985) relied on police reports from the Baltimore Police Department to measure serious crime in their subject neighborhoods. In addition to reported crime and surveyed victimization, Perkins and Taylor (1996) also analyzed newspaper articles related to crimes in or near neighborhoods included in their study as a measure of media influence on localized fear.

Obtaining reliable measures of crime often proves more difficult than accurately measuring perceptions or media accounts. Police reports are perhaps the most common sources of crime data employed by researchers. However, because crime records may be under- or overcounted for budgetary or political ends and because reporting to police varies by type (e.g., higher-value property crimes and homicides more likely to be reported than lesser property and personal crimes), neighborhood, and city, resident surveys and interviews often provide very different information and insights.

Means of Analysis

Collecting data at both the individual and block levels allows researchers to employ complex statistical methodologies to find associations

between disorder and crime. In some cases, however, simpler analytic techniques may yield useful data about the neighborhoods being studied. Using the Chicago, Baltimore, and Salt Lake City studies, this section provides a brief overview of the various statistical techniques employed in disorder-related research.

Simple correlations may reveal important relationships among variables. Taylor et al. (1985), for example, found that actual occurrences of crime measured by analyzing Baltimore police reports were positively correlated with physical decay, nonresidential land uses, and vacancy rates. Perkins et al. (1992) utilized correlations to examine relationships among actual disorder, perceived disorder, and perceived crime in Baltimore neighborhoods. After controlling for race, education level, tenure, and block size, four statistically significant correlations emerged at the block level: (1) the number of street lights was negatively correlated to perceived robberies, (2) the presence of yard decorations was negatively correlated to perceived drug dealing and fighting in the street, (3) yard decorations were negatively correlated to perceived burglaries, and (4) the presence of block watch signs was negatively correlated to perceived burglaries. Clearly, in both Baltimore studies, the use of simple correlations yielded helpful information: certain environmental conditions and features were associated with neighborhood crime reports and perceptions.

Multiple regression allows researchers to determine the degree to which a combination of independent variables are predictive of a dependent variable. Perkins et al. (1992) employed multiple regression to determine whether territorial functioning and defensible space features could supplement observed disorder in predicting perceived disorder and crime. The authors found that adding territorial markers and defensible space features to their model explained more of the variance in perceived disorder and crime.

Hierarchical linear modeling (HLM) and structural equation modeling (SEM) are examples of advanced statistical techniques employed by researchers. With HLM, the effects of predictors at multiple levels and/or over multiple points in time on outcomes can be determined. For example, Perkins and Taylor (1996) used HLM to compare the power of three different methods of measuring aggregated community disorder (surveyed resident perceptions, on-site observations by trained raters, and content analysis of crime- and disorder-related newspaper articles) to predict subsequent fear of crime, controlling for individual-level perceptions of social and physical disorder and independent ratings of physical disorder on respondents' properties. Sampson and Raudenbush (2004) used HLM to examine the relationship between independently observed social and physical disorder and perceived disorder both within and between neighborhoods.

SEM has been employed in disorder-related studies as a means of exploring complex models including latent variables. For example, Sampson and Groves (1989) used SEM in finding that socioeconomic status, ethnic heterogeneity, residential mobility, family disruption, and urbanization predicted crime and delinquency when mediated by local friendship networks, unsupervised teenage peer groups, and low organizational participation.

Other Common Disorder Research Methods

Other methods have proven useful in disorder-related research; among these are qualitative analyses and longitudinal studies. Qualitative data can give researchers insights that may be impossible to glean, or less richly developed, through surveys, systematic environmental assessments, crime reports, or other quantitative sources. In-depth open-ended or semi-structured interviews, ethnographic accounts, and oral histories are examples of such methods. Earlier disorder-related research conducted by Suttles and other Chicago-school sociologists relied heavily on ethnography; many researchers now utilize qualitative research as one component in a toolbox of data collection methods. Longitudinal studies allow the researcher to discern trends over a period of time. By following up snapshot data with data collected at a later time or at subsequent intervals, the researcher may gain a better understanding of phenomena that are not static in nature. Because neighborhoods are

ever-changing, examining the neighborhood over a period of time can help the researcher contextualize and better understand phenomena and findings.

Several recent disorder-related studies have used many of the research and analytic methods discussed here. The following section highlights findings from those studies, focusing on the methodologies employed in both the data-gathering and data analysis stages. These studies are cited for illustrative purposes; references to additional studies of interest are provided following the entry's conclusion.

Major Findings Using Mixed Research Methods

This section provides a brief overview of recent major findings in the disorder-related literature. As a comprehensive review of the literature is beyond the scope of this entry, this section focuses on findings based on the above methods and means of analysis. Findings are categorized based on the following themes: objective versus subjective measures of disorder and crime, advanced statistical analyses, and challenges to accepted theories.

Objective and Subjective Measures

Early disorder-related research tended to focus on residents' perceptions of neighborhood disorder, occasionally comparing those perceptions to observations by trained raters. Recent research has in some ways attempted to standardize disorder-related research by balancing residents' perceptions with objective measures. Data in recent studies include results of independent neighborhood assessments, analyses of media accounts of crime, and analyses of police reports.

Taylor et al. (1985) examined the relationship between observed and perceived physical disorder in their study of Baltimore neighborhoods. Residents and independent raters evaluated four specific indicators of physical disorder: housing vacancies, empty lots, litter, and unkempt properties. After combining the indicators to form a single scale, the authors found a high correlation between resident perceptions and independent observations of physical disorder.

The Block Environmental Inventory (BEI) was used by Perkins et al. (1992) as a means of objectively measuring disorder, territorial markers, and defensible space features in Baltimore neighborhoods. Researchers obtained data on residents' perceptions by surveying 412 residents across 50 street blocks either by telephone or in person. Interviewers asked how big a problem (on a three-point scale) each of a list of 12 specific items related to physical disorder, social disorder, and neighborhood crime are on respondents' blocks. Perkins et al. used both correlation and regression analyses to determine whether the objective measures of disorder gleaned through the BEI were associated with subjective measures. The authors correlated objective and subjective measures of five specific forms of physical disorder: litter, vandalism, dilapidated exterior, vacant housing, and trashed lots. Correlations for all five forms of disorder were statistically significant. Litter was the strongest correlation, while dilapidated exterior was the weakest. After controlling for race, education, homeownership, and block size, all of the correlations were weakened but still statistically significant. The authors concluded that residents' perceptions and independent observations of physical disorder are strongly related but that the relationship is affected in part by block demographics.

In a 2004 study of Salt Lake City neighborhoods, Brown et al. considered associations among several predictors of crime: home attachment, observed physical disorder, perceived physical disorder, social ties and collective efficacy, and home ownership and demographics. Besides the demographic variables, all other predictors were scales or composites of multiple measures. Using multilevel modeling (HLM), the authors found that observed disorder and perceived disorder were not related at the individual level (i.e., people tend not to perceive their own properties as disordered). However, as expected, observed and perceived disorder were significantly related at the block level (Brown et al. 2004b).

Sampson and Raudenbush (1999, 2004) found little difference between observed and perceived

disorder. For their study, social disorder included adults congregating, public intoxication, selling drugs, and prostitution. Physical disorder ranged from minor offenses (e.g., cigarettes on the street) to more prominent offenses (e.g., gang graffiti on buildings). The researchers employed SSO by having independent raters videotape block faces; the tapes were then coded for both social and physical disorder. The objective data were compared to survey data gathered from 3,864 residents. The authors found that residents' perceptions of both physical and social disorder were significantly associated with independent measures of disorder.

Specific findings regarding the relationship between objective and subjective measures of disorder have varied by study. However, all of the studies cited in this section have found at least some degree of correlation between independent observations and resident perceptions of disorder. Variation among the findings appears to be attributable at least partially to the particular research setting, research questions, analytic methods, and the variables or characteristics being studied.

Advanced Statistical Analyses

Quantitative modeling techniques such as multiple regression, hierarchical linear modeling (HLM), and structural equation modeling (SEM) have been utilized to analyze disorder and crime data. Each method brings its own set of strengths to data analysis. Multiple regression allows researchers to determine the predictive values of independent variables on a particular outcome. Hierarchical linear modeling allows researchers to compare data on multiple levels; change in individuals or neighborhoods can be traced over time, individuals can be compared with their neighbors, and blocks or neighborhoods can be compared to each other. Structural equation modeling allows researchers to determine the effects of latent variables. Both HLM and SEM are relatively new tools. The following subsections explore findings attributable to the use of these tools; specifically, the subsections address findings related to personal characteristics, place characteristics, home and income characteristics, and other important factors.

Personal Characteristics

Quantitative analyses reveal that certain personal characteristics are associated with individuals' perceptions of disorder and crime. Studies from Chicago, Baltimore, and Salt Lake City offer diverse and wide-ranging findings on the associations among crime, fear, disorder, and personal characteristics. With regard to sex, females are likely to perceive more disorder than males (Sampson and Raudenbush 2004). Likewise, females are likely to express more fear of crime than males at both the within-block and between-blocks levels (Perkins and Taylor 1996).

Age plays a role in perception and fear, though the findings have been somewhat inconsistent. At the block level, age is a significant predictor of fear of crime. In one study, however, age was not a significant predictor of fear at the individual level (Perkins and Taylor 1996). In another study, within a given block, older individuals perceived less disorder (Sampson and Raudenbush 2004), an unexpected finding that may be due to older residents being less aware of specific changing disorder cues.

Race and ethnicity have significant effects at multiple levels. African Americans tend to perceive less disorder than do other residents of their block; however, controlling for social context, blocks with more African Americans tend to have more disorder as perceived by residents of the block. Latinos are also positively associated with perceived disorder at the between-blocks level and are likely to perceive more disorder at the within-block level (Sampson and Raudenbush 2004). Race has been shown to be a significant predictor of fear of crime as well (Taylor et al. 1985; Perkins and Taylor 1996).

Marital status plays some role in perception of disorder. Separated and divorced individuals are likely to perceive more disorder at the within-block level than their married and single counterparts. Between blocks, marital status is not significantly associated with perceptions of disorder (Sampson and Raudenbush 2004). Clearly, sex, age, race and ethnicity, and marital status are all important personal characteristics that shape residents' perceptions of crime and/or disorder,

although their effects tend to vary by location and model.

Place Characteristics

Taylor et al. (1985) found two important correlations between crime and place-based characteristics. First, the authors found a strong positive correlation between crime and vacant housing, indicating that the higher the percentage of vacant housing units in a neighborhood, the higher the amount of crime in that neighborhood. A second statistically significant positive correlation was found between crime and nonresidential land uses. Both vacancy and land use were significant predictors of fear. Perkins and Taylor (1996) found that the physical conditions of nonresidential lots were better predictors of fear of crime than the conditions of residential areas. Land use appears to be a consistent indicator of fear as shown by two separate Baltimore-based studies.

Home and Income Characteristics

Researchers have examined the associations between crime and home-based characteristics such as tenure and home attachment, which are both inversely related to observed disorder (Brown et al. 2004b). At the within-block level, homeownership is negatively associated with police reports of crime. At the between-blocks level, home attachment is negatively associated with police reports of crime (Brown et al. 2004b).

Socioeconomic status (SES) has been shown to be a significant predictor of perceived disorder, fear of crime, and neighborhood confidence. Fear and confidence are particularly affected by perception of disorder in moderate-income neighborhoods (Taylor et al. 1985). In a United Kingdom-based study, Sampson and Groves (1989) considered the association between SES and crime. The authors found that SES was a significant predictor of crime. However, a closer look revealed that the presence of unsupervised peer groups mediated the effects of SES on crime. In fact, unsupervised peer groups mediated 80 % of the effect of SES on mugging and street robbery, 34 % of the effect of SES on stranger violence, and 68 % of the effect of SES on total victimization. The presence of unsupervised peer groups had a worsening effect on both household and property victimization, while organizational participation had an inverse effect on both forms of victimization. These findings led Sampson and Groves to conclude that the presence of unsupervised peer groups and the absence of organizational participation were more important to remediate than low SES.

Other Outcomes

Brown, Perkins, and Brown (2004a) employed a longitudinal approach to examine the spillover effects of cleaning up a brownfield in Salt Lake City and replacing it with new housing. The researchers were interested in incumbent upgrading – the idea that residents will invest in their own properties if their neighbors do likewise. Independent disorder assessments and resident surveys were taken from 1993 to 1996 (T1) and again from 1998 to 1999 (T2). Police reports of crime were measured from 1995 to 1996 (during T1) and again from 1999 to 2000 (T3). The researchers examined physical disorder using objective raters; home attachment, homeownership, perceived disorder, and crime were measured using resident surveys.

Using correlation analysis, the researchers found that from T1 to T2, observed disorder in the areas surrounding the former brownfield decreased, especially in the blocks closest to the brownfield. HLM analysis indicated that crime reports decreased at T2 in the neighborhoods closest to the brownfield, while crime reports tended to increase in the neighborhoods farther away. The authors found that the presence of independently observed physical disorder on specific properties increased the likelihood of crime on those properties; further, physical disorder on a particular block increased the likelihood of crime on properties within that block. Conversely, the presence of new houses built on the former brownfield was associated with decreased crime and disorder on nearby blocks. The blocks in which disorder increased between T1 and T2 were more vulnerable to crime at T3. The authors argued that litter and unkempt lawns were two types of physical disorder that could lead to residents investing less in their homes.

In summary, recent research has not revealed any single, dominant predictor of crime. Rather, the research has created a patchwork of findings which – when taken as a whole – provide a clearer picture of how individuals perceive and experience disorder, fear, and crime in their local settings. While some personal characteristics appear to influence both degree of fear and perceptions of crime and disorder, territorial markers, defensible space features, homeownership, home and place attachment, organizational participation, and lack of activities for young people appear to work in some combination to explain the occurrence of crime within neighborhoods. Simple answers have not emerged in disorder-related research, regardless of the methods or analytic means employed.

Challenges to Accepted Theory

The studies included in this section have resulted in findings that challenge long-held assumptions and theories. Taylor et al. (1985) found that nonresidential land use was significantly correlated with crime and was a significant predictor of perceived physical decay. In a later study, Taylor et al. (1995) concluded that nonresidential land use in a neighborhood has a statistically significant effect on physical deterioration. Although parks well used by families may tend to deter crime, in general most nonresidential land uses act as magnets for neighborhood deterioration and subsequent loss of social control and ultimately (either directly through routine travel patterns of offenders or indirectly through increased disorder) for crime. This conclusion directly refutes Jane Jacobs' (1961) argument that businesses have positive impacts on the neighborhoods in which they are located. The authors did note that Jacobs wrote in a context in which neighborhoods and neighborhood businesses were typically more closely aligned ethnically. Further, in that context, longer hours kept by business owners may have discouraged loitering in the evening hours. Taylor et al. suggested that policy focus on preventing or slowing physical deterioration and loss of social control. All of the recommended policies begin with acknowledgement of the relationship between nonresidential land uses and disorder.

Sampson and Raudenbush (2004) directly challenged the broken windows and incivilities theories. In their study of Chicago neighborhoods, the authors employed SEM to examine the effects of structural characteristics – such as race, ethnicity, and class – on physical and social disorder. Observed disorder was a moderate correlate of predatory crime (e.g., robbery, aggravated assault, rape, and homicide) and varied with neighborhood demographic composition. When the demographic characteristics were taken into account, the connections between observed disorder and all but one predatory crime (robbery) were rendered statistically insignificant. The authors argued that concentrated poverty and lack of collective efficacy at the neighborhood level explained disorder and most types of predatory crimes. Rather than using disorder as a predictor of crime, Sampson and Raudenbush argued that disorder was – like crime – an outcome of broader structural factors. The authors concluded by arguing that the mere mitigation of disorder may be inadequate means of combating crime; addressing structural factors and collective efficacy would be a more appropriate response.

Troublesome Findings and Directions for Future Research

The relationship between disorder and crime is still a relatively new field of study for social scientists. Thus, many of the findings warrant closer examination. This section considers potential areas for future research. Three specific issues raised in the literature reviewed in this entry are presented for consideration: the spatial dimensions of disorder and crime, measurements of social disorder, and the future of *broken windows*.

While units of analyses have been discussed and debated among researchers, few studies have explicitly examined the extended spatial implications of neighborhood disorder. Brown et al. (2004a) did consider extended neighborhoods in their study of a former brownfield site in Salt Lake City. They found that blocks farther away from the cleaned up area had more crime

and disorder. Their study was situated in a first-ring suburban community; it is uncertain whether studies set in an inner city or outer-ring suburb would result in the same or similar findings. Future research might focus on the spatial dimensions of disorder and crime in various settings, comparing perceptions, observations, and occurrences of crime across different types of neighborhoods in different types of communities.

Researchers have found it relatively simple to define physical disorder. After all, such incivilities are visible, sometimes highly so. Broken windows, graffiti, and even cigarette butts in gutters are observable; there would likely be agreement among researchers and residents that most of these are undesirable. Defining social disorder has proven more difficult. A clear way of characterizing social disorder has eluded researchers for the most part. Researchers who describe social interactions and relationships as disorders or incivilities risk pathologizing individuals, neighborhoods, and communities. Studies examined in this entry characterize noticeable offenses such as prostitution and drug dealing as social disorder. Yet, social disorder may extend beyond such egregious examples. Future research could delve further into social disorder, seeking to better understand and define it while making sure not to present marginalized and oppressed individuals, neighborhoods, and communities as abnormal or at fault. Interestingly, early disorder-related research was largely centered on the social aspects of disorder; the focus on physical disorder is more recent. In this way, clarifying and better defining social disorder represents a return to the roots of disorder-related research.

Perkins et al. (1990) join earlier researchers in pointing out that measuring social climate is a controversial issue in disorder-related studies. As this entry has shown, researchers often rely on measures of social climate to draw conclusions about groups and blocks. Yet such measures are typically collected at the individual level and aggregated to some higher level (e.g., street block or census tract). The validity and reliability of social climate in disorder-related research warrants scrutiny; future research may focus on issues of data aggregation (Shinn 1990).

Finally, the work of Sampson and colleagues has called key components of the broken windows theory into question. Their findings indicate that disorder and crime are not necessarily involved in a cause-and-effect relationship. Rather, disorder and crime are both outcomes of greater structural forces. In spite of these findings, broken windows theory continues to be a dominant framework in community policing. Future studies may wish to replicate Sampson's methods in other settings. By exploring disorder, demographic characteristics, and crime reports in multiple settings, researchers can corroborate or challenge Sampson and colleagues' findings. Additionally, researchers may choose to employ other data collection and analytic methods. For example, rather than replicating the SSO used by Sampson and Raudenbush (2004), researchers seeking to replicate the UK study may opt to use the BEI or some other data collection mechanism. Challenging a dominant paradigm and effecting change in policy and practice are not easy tasks; replicable studies in various settings may help shift the focus from broken windows to a more comprehensive discussion of social context and its manifestations in both disorder and crime.

Conclusion

This entry has provided an overview of recent disorder-related research, emphasizing the ways in which data have been collected and analyzed. Over the last few decades, researchers have supplemented resident perceptions of disorder with independent observations made by trained raters. Further, media accounts and police reports of crime have added points of comparison in some disorder-related research. Means of analysis have also evolved; the development of efficient analytic techniques such as hierarchical linear modeling and structural equation modeling has allowed researchers to explore layers of data and the effects of latent variables, respectively. Using these data-gathering methods and analytic techniques, researchers have uncovered personal and environmental characteristics that increase

the likelihood of both perceiving and experiencing crime and disorder. Finally, implications recent studies have for future research include topics as diverse as geographic scope, clarification of social disorder, and the future of the broken windows theory.

Related Entries

- Broken Windows Thesis
- Crime Prevention through Environmental Design
- Defining Disorder
- Designing Products Against Crime
- Differential Social Organization
- Disadvantage, Disorganization and Crime
- Discriminant Validity of Disorder and Crime
- Drug Trafficking
- Informal Social Control
- Law of Community Policing and Public Order Policing
- Order Maintenance Policing
- Poverty, Inequality, and Area Differences in Crime
- Problem-oriented Policing
- Prostitution
- Public Housing and Crime Patterns
- Race and Ethnicity in Social Disorganization Theory
- Race, Ethnicity, and Youth Gangs
- Social Disorder and Physical Disorder at Places

Recommended Reading and References

Brown BB, Perkins DD, Brown G (2004a) Crime, new housing, and housing incivilities in a first ring suburb: multilevel relationships across time. Hous Policy Debate 15(2):301–345

Brown BB, Perkins DD, Brown G (2004b) Incivilities, place attachment and crime: block and individual effects. J Environ Psychol 24:359–371

Franzini L, Caughy MOB, Nettles SM, O'Campo P (2008) Perceptions of disorder: contributions of neighborhood characteristics to subjective perceptions of disorder. J Environ Psychol 28(1):83–93

Gau JM, Pratt TC (2010) Revisiting broken windows theory: examining the sources of the discriminant validity of perceived disorder and crime. J Crim Justice 38(4):758–766

Haney TJ (2007) 'Broken windows' and self-esteem: subjective understandings of neighborhood poverty and disorder. Soc Sci Res 36(3):968–994

Jacobs J (1961) The death and life of great American cities. Random House, New York

Kurtz EM, Koons BA, Taylor RB (1998) Land use, physical deterioration, resident – based control and calls for service on urban street blocks. Justice Q 15(1):121–149

Perkins DD, Florin P, Rich RC, Wandersman A, Chavis DM (1990) Participation and the social and physical environment of residential blocks: crime and community context. Am J Commun Psychol 18:83–115

Perkins DD, Meeks JW, Taylor RB (1992) The physical environment of street blocks and resident perceptions of crime and disorder: implications for theory and measurement. J Environ Psychol 12:21–34

Perkins DD, Taylor RB (1996) Ecological assessments of community disorder: their relationship to fear of crime and theoretical implications. Am J Commun Psychol 24:63–107

Perkins DD, Wandersman A, Rich RC, Taylor RB (1993) The physical environment of street crime: defensible space, territoriality and incivilities. J Environ Psychol 13:29–49

Robinson JB, Lawton BA, Taylor RB, Perkins DD (2003) Multilevel longitudinal impacts of incivilities: fear of crime, expected safety, and block satisfaction. J Quant Criminol 19(3):237–274

Ross CE, Mirowsky J (2001) Neighborhood disadvantage, disorder, and health. J Heal Soc Behav 42(3):258–276

Ross CE, Mirowsky J (2009) Neighborhood disorder, subjective alienation, and distress. J Heal Soc Behav 50(1):49–64

Sampson RJ, Groves WB (1989) Community structure and crime: testing social – disorganization theory. Am J Sociol 94(4):774–802

Sampson RJ, Raudenbush SW (1999) Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. Am J Sociol 105(3):603–651

Sampson RJ, Raudenbush SW (2004) Seeing disorder: neighborhood stigma and the social construction of "broken windows". Soc Psychol Q 67(4):319–342

Schweitzer JH, Kim JW, Mackin JR (1999) The impact of the built environment on crime and fear of crime in urban neighborhoods. J Urban Technol 6(3):59–73

Shinn M (1990) Mixing and matching: levels of conceptualization, measurement, and statistical analysis in community research. In: Tolan P, Keys C, Chertok F, Jason LA (eds) Researching community psychology: issues of theory and methods. American Psychological Association, Washington, DC, pp 111–126

Suttles GD (1968) The social order of the slum: ethnicity and territory in the inner city. University of Chicago Press, Chicago

Taylor RB (1987) Toward an environmental psychology of disorder: delinquency, crime, and fear of crime. In: Stokols D, Altman I (eds) Handbook of environmental psychology, vol 2. Wiley-Interscience, New York, pp 951–986

Taylor RB (1996) Neighborhood responses to disorder and local attachments: the systemic model of attachment, social disorganization, and neighborhood use value. Sociol Forum 11(1):41–74
Taylor RB (1997) Social order and disorder of street blocks and neighborhoods: ecology, microecology, and the systemic model of social disorganization. J Res Crime Delinq 34(1):113–155
Taylor RB, Hale M (1986) Testing alternative models of fear of crime. J Crim Law Criminol 77(1):151–189
Taylor RB, Shumaker SA (1990) Local crime as a natural hazard: Implications for understanding the relationship between disorder and fear of crime. Am J Commun Psychol 18(5):619–641
Taylor RB, Shumaker SA, Gottfredson SD (1985) Neighborhood-level links between physical features and local sentiments: deterioration, fear of crime, and confidence. J Archit Plan Res 2(4):261–275
Taylor RB, Koons BA, Kurtz EM, Greene JR, Perkins DD (1995) Street blocks with more nonresidential land use have more physical deterioration: evidence from Baltimore and Philadelphia. Urban Aff Rev 31:120–136
Wilcox P, Quisenberry N, Cabrera DT, Jones S (2004) Busy places and broken windows? Toward defining the role of physical structure and process in community crime models. Sociol Q 45(2):185–207
Wilson JQ, Kelling GL (1982) Broken windows: the police and neighborhood safety. Atlantic 249(3):29–38

Disproportionate Minority Contact

► Juvenile Justice System Responses to Minority Youth

Dissocial Personality Disorder

► Psychopathy and Antisocial Personality Disorder

Distributional Analysis

► Quantile Regression Models to Analyze Experimental Data

DNA

► Identification Technologies in Policing and Proof

DNA Databasing

► DNA Technology and Police Investigations

DNA Fingerprinting

► DNA Technology and Police Investigations

DNA Identification

► DNA Profiling

DNA Profiling

Tracy Alexander
Cold Case Investigation, LGC Forensics, Abingdon, UK

Synonyms

DNA identification; Forensic DNA matching; Genetic profiling

Overview

The use of DNA for identification purposes has often been the subject of controversy. This is particularly true in the UK at present with the imminent implementation of legislation to destroy a significant number of physical samples and to remove the related profiles from the national DNA database. This chapter covers the basic biology behind the use of DNA profiling and its use in the criminal justice system in the UK, particularly in terms of the relevant legislation and the provision of services by private laboratories. These issues are fundamental to the understanding of the problems that have always existed in terms of DNA interpretation,

particularly in terms of mixed and partial profiles, plus emerging issues that have arisen with advances in research and technology.

Fundamentals of DNA Profiling in Forensic Science

DNA is a complex chemical compound comprised of relatively simple building blocks generally found within the nucleus of cells. The genome is the entirety of this cellular DNA, and encodes all the genetic information that governs an organism's structure and function, and is unique to the organism, except in genetically identical individuals. During sexual reproduction, the information determining the physical characteristics is inherited, half from each parent. Small sections of the molecule contain specific variations in the genetic code that can be statistically evaluated to assist in the process of individualization and therefore identification, most notably in human subjects. Analysis of the variable sections of DNA is frequently employed to determine paternity and to resolve immigration disputes, but public perception is focussed on its application as an aid to police investigative processes, to identify victims of crime and, by the transfer, exchange, and persistence of traces of DNA, the offenders responsible for crime.

The use of DNA identification methodologies has revolutionized crime scene investigation and the presence of forensic science in the courtroom. Historically, evidence of identity was limited to direct eye witness or, since the late 1800s, photographic recognition. Fingerprint identification has been in common use in the UK since the establishment of the Fingerprint Bureau in 1901, with the first conviction employing this technology being for a murder in 1905. Consequently, public confidence in the veracity of "dactyl" fingerprints and the legislation governing its utility has had considerable time to develop, whereas the revolution of DNA identification is still a relatively recent phenomenon and still presents legal and ethical issues that have yet to be fully resolved.

DNA

Deoxyribonucleic acid is a macromolecule containing the genetic instructions for the replication and function of all known living organisms (excepting a small number of viruses). Specific segments known as genes carry this genetic information and encode for proteins, for example, are known as genes. The presence of DNA in living material was originally recognized in 1869 by a Swiss physician, Friedrich Miescher. By 1937, the chemical composition of DNA had been identified and x-ray diffraction had demonstrated that DNA had a regular structure and the link between DNA and hereditary characteristics had been postulated. In 1953 Francis Crick and James Watson proposed the double-helix model of DNA structure. Two strands joined together by four distinct nitrogenous bases form a compact right-handed helix, thus allowing for a large amount of information to be encoded in a relatively small space. Crick and Watson's DNA breakthrough helix discovery was enabled through the work of two other scientists: Maurice Wilkins, who pioneered X-ray crystallography, and Rosalind Franklin who refined the technique for work with DNA. Together they identified the four bases which link the two strands of the helix; adenine, cytosine, thymine, and guanine (A, C, T, and G). These nitrogenous bases are attached to a ribose sugar on the phosphate-sugar backbone of each DNA strand.

Each molecule of DNA includes a pattern of these bases bonded together. Adenine forms hydrogen bonds with thymine while cytosine similarly engages with guanine. This base pairing holds the two strands of the double helix in close proximity to each other. It is the sequence of these base pairs within genes that is the genetic "blueprint" for the organism, encoding sequential instructions for amino acids, which are the building blocks of proteins. The information encoded in DNA is held in compact structures called chromosomes, of which humans have a normal complement of 46: 23 of maternal and 23 of paternal origin. Humans share almost all of their DNA sequence with other humans and much of the sequence information with other organisms. Only about 1 % is specific to humans when compared to a near relative such as

a chimpanzee, and only about one tenth of that 1 % of DNA differs from one human to the next (with the exception of identical twins). However, this small fraction of the DNA still comprises some three million base pairs.

When analyzing person-specific DNA variations for forensic purposes, the sections of DNA examined are called short tandem repeats (STRs). As the name suggests, these are short sequences of DNA normally just four bases long repeated as adjacent blocks. It is predominantly the number of repeats of the core sequence that varies within individuals (e.g., GATA, GATA, GATA). STRs are, fortuitously, found in the noncoding regions of the genome (i.e., not within the genes) but the position (locus) of these on the chromosomes does not usually change from one person to the next. For crime scene investigation in the UK, a DNA profile is produced by measuring the physical length of the DNA at ten STR loci simultaneously, and the DNA analyzer displays the results as a series of peaks on a graph known as an electropherogram (EPG). The information carried therein is statistically very powerful, and the probability of a match being from someone other than the suspect and unrelated to them is given a numerical value. Assessments of match probabilities are made with reference to STR profiles separated into different racial groups in order that the most conservative figure is given.

DNA in the form of linear chromosomes exists within the nucleus of every nucleated cell. Several highly specialized cell types lose their nuclei and their DNA, for example, red blood cells are packed with hemoglobin while white blood cells retain their DNA, but in general DNA recovered from a white blood cell will be identical to DNA from any other cell type or tissue from the same individual. The nucleus is not the only source of DNA; cells contain tiny sausage-shaped organelles known as mitochondrion. These are the powerhouses of the cell specifically designed to make energy. Each mitochondrion has between one and ten copies of circular DNA that code for proteins required for the energy production and each cell may have hundreds of mitochondrion depending on its function (e.g., mitochondrion are found in large numbers in muscle cells). Hence in adverse conditions when the genomic DNA is too severely degraded to be useful, it is sometimes still possible to get some useful genetic information from the mitochondrion. The amount of genetic variation is far lower than genomic DNA but mtDNA can be very useful in identifying family members. The mitochondrion are inherited through the maternal line as at fertilization it is only the sperm's nucleus that is injected into the egg, not the mitochondrion.

While all nucleated cells contain DNA, not all nuclear DNA is contained within cells. Like many of the elements that go to make up an organism, the DNA is recycled. Dead cells are harvested and the DNA broken down. In this process, long stretches of DNA can be found in the plasma and recently also discovered in sweat. This discovery has far-reaching implications for the evidential value of "touch" DNA, particularly in terms of secondary and tertiary transfer between individuals and touched items (Quinones and Daniel 2012).

Development of Technology

The first paper reporting the use of DNA in a criminal context was published in 1985 by Alec Jeffreys, Professor of Genetics at Leicester University (Jeffreys 1985a).

Jeffreys was researching inherited variation in human DNA and he demonstrated how a DNA profile could be used to resolve issues of identity and kinship. Its initial use in legislative practice was to demonstrate that a child was the legal offspring of two individuals already granted asylum in the UK (Jeffreys 1985b).

DNA technology was the subject of research and development for the purposes of criminal investigation throughout the late 1980s and early 1990s primarily by the Forensic Science Service in the UK (FSS), previously known as the Home Office Forensic Service Laboratory (HOFSL). The FSS was the major provider and innovator in forensic science in the UK until its closure in March 2012. The multi-locus probe was introduced to routine casework in 1987 but required a large amount of sample material to produce a profile, although a full profile could give a likelihood ration of one in a million. This was followed by the single-locus probe in 1989 which allowed smaller

samples to be tested and could produce statistics of one in 20–30 million.

However, both techniques were limited by the quantity of starting material required and regularly failed due to the poor quality and small amounts of DNA recovered from crime scenes. A major breakthrough came about with the development of the polymerase chain reaction (PCR) by American biochemist and Nobel Prize winner Kary Mullis. This enabled the biochemical copying of small amounts of DNA. This meant that sufficient target sequences could be amplified so as to be easily detectable by DNA analysis. During the PCR process the sample of DNA is treated with cycles of heating and cooling that denature the DNA and divide it into two separate strands, and a DNA primer is then used to anneal to these strands. Primers are short pieces of DNA containing sequences complementary to the target regions. The annealed primers together with a TAQ polymerase (an enzyme which enables a chain reaction to happen) allows two new copies of the sequence under investigation to be produced, one from each strand. This process of denaturing and rejoining was repeated again and again in a cyclic manner allowing amplifications of DNA even from samples with a relatively limited amount of starting material.

The flexibility of this method meant that only the targeted sections are copied and simultaneously those copies are marked by the addition of fluorescent tags. The fluorescent tags mean that the copied sections of DNA are visible when illuminated by laser light. In the UK, most DNA samples are subject to 28 cycles of PCR, doubling the amount of DNA available for analysis at each cycle.

The first use of PCR was in the detection of Human Leucocyte Antigen DQ alpha (HLA), introduced by the FSS in 1991, thereby allowing the analysis of much smaller stains. However, PCR has now become a fundamental part of all current profiling methodologies. The STR Quad system was introduced in 1994 which looked at four regions of DNA followed by, in 1995, Second Generation Multiplex (SGM) which gave a profile of six areas of an individual's DNA plus the sex indicator area, giving an average discrimination potential of 1:50 million. The SGM technique facilitated the introduction of a computerized database in 1995. SMGPlus®, used since 1999, is the main technique in use in the UK and looks at ten areas plus the sex indicator area and increases the discrimination potential to 1:1,000 million. However, many countries are moving to 16 or 17 loci multiplexes which the UK plans to do by 2014.

Use in the Criminal Justice System

In the UK, a Royal Commission was set up in 1993 to look at the opportunities that new DNA technologies might be able to offer the criminal justice system. The development of rapid, automated testing and the use of digital tools enabled interrogable databases to be quickly compiled. However, in order to deploy the technology, the police required the power to take DNA samples from those involved in criminal offences. The ability to take samples was restrictions in the Police and Criminal Evidence Act. PACE 1984 had been specific about the consent and authority required before a sample could be taken. Samples were categorized as "intimate" or "non-intimate" and regulations covered who could and could not take samples. The Criminal Justice and Public Order Act 1994 redefined intimate and non-intimate samples: mouth swabs were redefined as non-intimate and could be obtained without the person's consent. Though a suspect could refuse to open his mouth, it was then permissible to pluck head hairs with roots from which DNA could be obtained. Consent was still required to obtain blood and still required a qualified practitioner to take it. The Police Reform Act (2002) changed the regulations concerning the taking of samples: a police constable could now take non-intimate samples or could delegate this power to a "designated person" such as a civilian forensic officer. It also created the requirement for all new police officers to supply DNA samples to the Police Elimination Database.

In 1995, the evolution of DNA STR profiling technology and the subsequent change in legislation meant that the Home Office was in a position to create a database. The technology was simplified and automated, and the world's first criminal intelligence database. This was launched in April of that

year: the UK National Criminal Intelligence DNA Database (NDNAD) (www.genome.wellcome.ac.uk/doc_wtd020879.html). Scotland and Northern Ireland have databases separate to that in England and Wales but all three are intersearchable.

First Use in a Criminal Investigation

In 1983 Lynda Mann was found raped and murdered in a deserted footpath in Leicestershire. Conventional grouping tests on semen samples from the body suggested that her killer was a person with blood type A and an enzyme type shared by approximately 10 % of males in the general population, but with no further evidence, the case remained unsolved. In 1986 the murder of another girl, also in Leicestershire, was linked by police through modus operandi. Police held a prime suspect, Richard Buckland, who confessed to the second murder but not the first. Jeffreys, in conjunction with the FSS, using extraction methods which enabled DNA from semen to be separated from DNA from vaginal cells, demonstrated that the murders were committed by the same person and that that person was not Buckland. Leicestershire Constabulary and the FSS began an investigation in which 5,000 local men were asked to volunteer blood or saliva samples, but after 6 months, no matches had been found. Later one of those men was heard bragging that he had been paid £200 to give a sample on behalf of Colin Pitchfork. Pitchfork was arrested in September 1987 and samples taken from him matched those of the double killer. Pitchfork admitted the murders and was convicted in 1988, becoming the first man to be convicted on DNA evidence, with Buckland being the first person to be proved innocent by DNA profiling. It was also the first time that the mass DNA screening of a population had been undertaken, a process that has been carried out on numerous occasions since. Even in cases where no suspect has been identified through this process, it has been beneficial in quickly eliminating a large number of individuals as being the donor of a profile believed to be crime related.

Persistence of DNA and Use in Historic Cases

Trace amounts of DNA can be recovered from bones as much as 5,500 years old, with opportunities within the forensic world for the identification of the victims and perpetrators of crime. In 1992, DNA testing gave compelling evidence linking remains recovered in Brazil in 1985 with Dr Joseph Mengele who died in 1979 by comparing a sample taken from the femur of the skeleton with samples taken from his widow and his son, which indicated full parental inclusion. In 1991, skeletal remains found in a shallow grave in Yekaterinburg were identified by Russian authorities as those of Tsar Nicholas II, the Tsarina Alexandra with three of their children. Remains discovered in a nearby smaller grave in 2007 were identified for the remaining two children using mtDNA, in part using samples from the Duke of Edinburgh who shares the same maternal link. Improvements in technology have meant that DNA is a frequently employed technique in resolving "cold cases," such as in the recent conviction in the UK of David Burgess for the murder of Yolande Waddington in 1966.

Databases

In April 2007, responsibility for the NDNAD was transferred from the Home Office to the National Policing Improvement Agency (NPIA). The NPIA published the following statistics for the databases of England and Wales, Scotland, and Northern Ireland combined as at January 2012 (www.npia.police.uk/en/13338.htm).

Estimated total number of individuals retained on NDNAD	5,882,724
Total number of subject sample profiles retained on NDNAD	6,889,385
Total number of subject sample profiles retained on NDNAD from volunteers	43,915
Total number of crime scene sample profiles retained on NDNAD	409,695

In the USA, the Combined DNA Index System (CODIS) is the central database for

DNA profiles created by federal, state, and local crime laboratories and funded by the Federal Bureau of Investigation (FBI). Originally holding only the profiles of sex offenders, it has been extended to include a much wider range of offenses and encompasses:

- The Convicted Offender Index
- The Arrestee Index
- The Forensic Index (crime scene profiles)
- The Missing/Unidentified Persons Index
- The Missing Persons Reference Index

All 50 states have legislation governing the collection, storage, and retention of DNA and it is state law rather than federal law which governs which crimes qualify for CODIS. CODIS databases exist at local, state, and federal level and separate laboratories can retain or share information as they choose. However, CODIS is the largest database in the world.

Interpol also operates DNA Gateway, inaugurated in 2002, containing more than 117,000 profiles submitted by 61 member countries. Police from any of the 190 countries belonging to Interpol can access information (http://www.interpol.int/INTERPOL-expertise/Forensics/DNA).

Legislation

In the UK, legislation regarding the collection, storage, and use of data held on the NDNAD has developed over time, and no single piece of legislation covers every aspect legislative amendments have been made to old laws, and case law originating from judges' rulings has redefined the application of the legislation.

The Doheny and Adams ruling (1997) addressed the way in which DNA evidence should be presented in court. An expert could no longer give an opinion on whether a crime stain came from a suspect but had to explain its probability. In 2000, the Lashley judgement in the Appeal Court ruled that DNA evidence alone was insufficient to bring a conviction and supporting evidence was also required. However, this can be as limited as geographical proximity to the offence; living in or having visited the region where a crime scene stain is matched can be enough. Furthermore, in 2000, challenges to convictions in two cases, *R v. Wier* (murder) and *R v. "D"* (rape), sparked further reform. PACE (1984) required samples to be destroyed after acquittal or discontinuance; Wier and "D" were identified using unlawfully held DNA samples. The convictions were appealed, but the Lords found that it would have been against the cause of justice for the convictions to be set aside, and the Criminal Justice and Police Act 2001 amended PACE so that now all DNA data collected from persons arrested for an offence could be kept, whether found guilty or not guilty. The Criminal Justice Act (CJA) 2003 further extended this so that data could be logged from anyone arrested for an offence, irrespective of whether they were eventually charged. Within a short space of time, the database doubled in size, but the problem of holding records of innocent people was created. Samples and profiles could only be destroyed by application to the Chief Constable of the arresting force. In addition, the holding of samples for the prevention or detection of crime is exempt from the Human Tissue Act (2004), brought about in part as a response to the discovery of the retention of the organs of children without consent by the Alder Hey Children's Hospital.

Recent legislative changes will have a tangible impact on DNA identifications. The Nuffield Council on Bioethics published a report in September 2007 on "The forensic use of bioinformation: ethical issues" which recommended that proposals to extend police powers even further to include the taking of DNA for minor offences such as littering should not be implemented. In addition there has been a growing perception among civil liberties groups that the retention of samples on the NDNAD of individuals who were never convicted of an offence infringed the civil liberties of those whose DNA profiles were stored. The Appeals of "S" and Marper particularly apply: both were arrested in 2001 in separate incidents, but both cases were dropped. On application to the Chief Constable of South Yorkshire, both were refused the right to have their samples destroyed. Between 2002 and 2004, they were refused a judicial review of the decision, and their appeal was rejected first by the Court of Appeal and then by the House of Lords. However, in 2008,

the European Court of Human Rights found in their favor, stating that the indefinite retention of profiles on a database interferes with a right to a private life and is particularly important for minors. Following consultation, this led to further amendments to PACE (1984) being promoted in the Crime and Security Act 2010 which passed into law but which has not been enacted to date. The coalition government elected in 2010 has further revised provisions for the rights of the individual in the Protection of Freedoms Act 2012, with far-reaching consequences for the NDNAD. The terms of the act require that approximately one million records of people on the database in England and Wales, and the copies held elsewhere, must be removed. The law does not require the removal of records of adults who have been convicted or have accepted a caution from the police, and people arrested for (but not convicted of) a serious offence can have their records retained for 3 years in the first instance, or a further two if there is the approval of a court (www.genewatch.org/sub-539488).

To date it has been possible to carry out a familial search on the UK NDNAD. On arrest, two buccal swabs are routinely taken, one to be kept as a backup in case the first sample fails to yield a profile. The "A" sample is processed; the remaining "B" sample is stored. When a crime scene sample has not given an immediate match on the database, it has been possible to look for previously loaded profiles that show similarities. Geographical factors and known information about the suspect, such as age, are taken into account to narrow the number of near matches. Once a manageable number of matches are obtained, it is possible to profile the "B" sample looking only at Y-STRs (DNA information obtained only from the Y chromosome and so only paternally inherited). This information allows the formulation of family trees indicating the presence of a male relative who fits the criteria for the offender but who has never been arrested for a recordable offence. The first successful prosecution relying on this procedure was in 2004 when Craig Harman was convicted of manslaughter for throwing a brick from a bridge which killed a lorry driver. Harman had left his blood on the brick (having injured his hand before taking it) but did not at that point have a police record. Forensic experts at the FSS found a profile with similar characteristics using the new techniques of familial searching through a relative whose profile was on the database, and as a result, the police traced Harman. Familial searching has always been limited to the most serious of cases and requires approval from the DNA ACPO lead. However, the application of familial searching in this way will no longer be available to police forces as following the Protection of Freedoms Act 2012, the UK has decided to destroy all "B" samples (over six million samples), even if taken from convicted offenders.

For further details, please refer to:

- Police and Criminal Evidence Act (PACE) 1984
- Criminal Justice and Public Order Act 1994
- Criminal Evidence Act 1997
- Criminal Justice and Police Act (CJPA) 2001
- Police Reform Act 2002
- Criminal Justice Act (CJA) 2003
- Serious Organised Crime and Police Act 2005
- Policing and Crime Act 2009
- Crime and Security Act 2010
- Protection of Freedoms Act 2012

Some rights of the suspect have been set aside when arrested or detained under certain sections of the Terrorism Act 2000, and specific rights apply to suspects under the age of majority, i.e., 18 (www.legislation.gov.uk).

It is interesting to note that the Association of Chief Police Officers (ACPO) recently announced, before the actual implementation of the Protection of Freedoms Act 2012, a new operation to capture the DNA of individuals whose profiles are not currently held on the database. Using powers under the Crime and Security Act 2010, which became law last year, the aim of Operation "Nutmeg" is to gather DNA profiles from criminals who were convicted before 1995 (when the database was launched). Initially the operation will target 11,993 criminals convicted of serious offences such as murder, manslaughter, and rape

over the past 40 years. The success of the initiative is impossible to guess but there is obviously scope for further sampling, plus potential implications for the removal or retention of the samples currently targeted for destruction.

Provision of Services

The Forensic Science Service pioneered the use of the DNA database. Originally the Home Office Forensic Science Laboratory, it became an executive agency of the Home Office in 1991 but in 2005 changed its status from executive agency to a government-owned company. Demand for DNA services exceeded all expectations particularly after the launch of the NDNADB and large backlogs of samples gave private companies an opportunity to join the market by providing assistance. Police forces paid the FSS for their services but the increasing use of competitive tendering resulted in a loss of market share and the FSS ceased to be financially viable. It closed in March 2012 and services are now provided by a number of private companies such as LGC Forensics and Cellmark, who contract their services to the various constabularies in England and Wales. Private companies are able to invest in additional resources if it will have a beneficial effect on their profit margins and this, together with the competition in the forensic market, has led to some advantages such as reduced costs and turn round times, with a standard sample result being delivered in about 3 days. Forensic services in Scotland are provided by the Scottish Police Services Authority and in Northern Ireland by Forensic Science Northern Ireland. Each of these service providers will have separate provisions for the testing of different types of sample, and each must demonstrate the highest standards for preventing cross-contamination. Environmental testing is routinely carried out to monitor background levels and to inform processes to address any potential issues. It is also a requirement of the UK Accreditation Service (UKAS) and essential in order to maintain ISO 17025 accreditation, an assessment of internal standards, issued by the International Standards Organization and applied by UKAS in the UK.

Sampling

The NDNAD holds samples from three sources: personal samples from those arrested or charged, crime scene samples, and voluntary personal samples. Prior to 2004, these were non-evidential Criminal Justice (CJ) samples and required a confirmatory sample on rearrest. Since 2004, samples are taken under the provision of the Police and Criminal Evidence Act (PACE) and can be used for evidential purposes. Previously the taking of a confirmatory sample could entail a delay of up to 2 weeks during which time the suspect had the opportunity to abscond or to undertake multiple crimes, knowing their arrest was imminent.

There are three possible samples that can be taken from arrested persons for search against and inclusion on the database: blood, buccal scrapes (from the epithelial lining on the inside of the cheek against the buccal muscles), and pulled head hairs. Blood samples used to be taken from major crime suspects due to the likelihood of obtaining a usable profile, but improvements in technology have meant that this is no longer necessary and buccal scrapes are by far the most commonly taken. Pulled head hairs are an option if a suspect refuses to give a buccal sample voluntarily.

From a crime scene, there are many sources from which to obtain a DNA profile, with varying degrees of evidential value depending on the circumstances.

Blood is commonly found particularly at major crime scenes. The red blood cells contain the protein hemoglobin which bonds with oxygen to carry it around the body and there are about 4.5–5 million red blood cells per microliter (one thousandth of a milliliter) of blood, but they have no nuclear or mitochondrion DNA. It is therefore only the white blood cells, of which there are only 5–10,000 per microliter of blood, which can be analyzed for a DNA profile. In addition to the obvious opportunities such as blood left by suspects or victim's blood on weapons, the analysis of blood patterns (BPA) and subsequent profiling can be a vital tool in reconstructing a sequence of events.

Saliva does not contain DNA at point of production but epithelial cells from the inside of the cheek regularly slough off and are deposited in the saliva in sputum and on items coming into contact with the mouth, such as cigarette butts, drinking vessels, masks, gags, and on licked stamps and envelopes. Food is also an option but as saliva contains digestive enzymes, the material is often broken down, making odontology a potentially favorable evidential option.

Hairs with roots are a good source of DNA but dead hairs that have fallen out (telegen hairs) usually only contain mitochondrion DNA. The source of a hair may be identified by examining it in section. Determining the source may add value to an investigation, for example, finding head hairs on the boot of a suspect from a victim who has been kicked in the head may indicate the individual responsible for a fatal injury. Chest hairs on a particular knife in a multiple-stabbing incident with more than one suspect may indicate which weapon caused wounds to the chest and thereby which suspect is responsible for the chest injuries.

Sexual assaults are often resolved by finding an exchange of material between victim and suspect. Therefore a large number of samples are taken from victims in order to find traces from the suspect, either at a police station, hospital, or rape suite. These include internal and external vaginal and anal swabs plus samples from the mouth and any other pertinent areas, such as where the victim has been licked or bitten. Urine samples are also taken for toxicology but urine may be examined for DNA as some semen may be washed away with the urine during collection. On arrest, a suspect will also have numerous samples taken in order to find traces from the victim, including swabs from the glans and shaft of the penis. Exhibits are also collected from a crime scene to identify DNA from all parties involved. The ability to split DNA between semen and vaginal material is vital in the investigation of sexual assault using DNA evidence (the cellular and seminal fractions). In addition, a victim of rape is asked for permission to take a DNA sample from a child produced as a result in order to identify the offender, which is achieved by removing the mother's profile from that of the child, leaving the profile of the child's father. This can also be achieved after a sufficient period of development with the products of conception after a miscarriage or abortion.

Semen is only produced by postpubescent men with approximately one million spermatozoa per ejaculate. Vasectomized and naturally azoospermic males have a vastly reduced number of spermatozoa per ejaculate but it is often still possible to obtain a profile.

Footwear and clothing are often analyzed for "wearer" DNA as well as trace materials. This may come from cells sloughed off by close contact with the item or from sweat, as with "touch" DNA. Other items that have come into contact with biological material can also yield a DNA profile. One of the suspects in the Canary Wharf IRA bombings in November 1992 was identified from nasal debris found on a piece of green tissue paper recovered from a van containing a bomb that had failed to explode.

Trace amounts of DNA can be found in urine. They are hard to identify as they are contained only in discarded cellular material washed from the walls of the urinary tract and are diluted by other waste products. Similarly, very fresh fecal material may also yield trace amounts of DNA in mucal secretions on the outside of the stool.

DNA is also used for identification of the deceased. Venous blood is preferable but not always possible, particularly if the victim has bled profusely. Where a corpse is badly burned or heavily decomposed, samples may be sought in deep muscle tissue where nuclear DNA has not been exposed to degradation, or if there is none remaining, DNA may be obtained from bones or teeth.

In all circumstances, a protocol for the collection of samples must be strictly followed to ensure that there is no cross-contamination. Officers collecting samples will wear protective clothing, and once obtained, samples are sealed in sterile containers before processing. Crime scene samples and arrestee samples are processed separately, often at a completely different laboratory site.

Success Rates

It is assumed that subject samples should be almost 100 % successful in yielding a DNA profile.

Statistics for crime scene samples show approximate success rates for DNA profiles as follows:

Blood stains	93 %
Semen[a]	81 %
Saliva	76 %
Hair with roots	53 %
Other tissue	50 %
Other material[b]	26–30 %

[a]Not vasectomized or azoospermic individuals
[b]From unidentifiable source material

Processing and Analysis of Samples

Arrestee samples are usually buccal scrapes, one from the inside of each cheek. The two samples, "A" and "B," are kept separately with the "B" sample being stored in the event that initial testing fails. Crime scene samples then go to a casework laboratory. Since the samples come in many forms, each sample will be treated differently to extract the DNA, and some will require presumptive testing, to locate the DNA for profiling and to establish whether a crime scene stain is from saliva, blood, or semen.

Stains and samples are then chemically treated to extract the DNA either manually or robotically and a process of quantification follows. A fresh sample will almost always provide an adequate amount of DNA. The smallest standard starting template permissible is a 1 ng (one millionth of a gram per milliliter) sample. This is subjected to the standard 28 cycles of PCR which provides sufficient amplified DNA to produce a profile which can be converted into a numerical code. If sufficient information from the profile is obtained, then this is sent to the NDNAD where it can be digitally compared to the millions of subject and crime stain profiles held there. Should there be no match, the profile from the crime scene sample will be held on the database indefinitely in case a match should appear in the future, either person to person, person to crime stain, or crime stain to crime stain. However, should the database indicate a match with a profile held on the database, then the information is passed on to the investigating force and a warrant for the arrest of that person is issued. Provided there is other supporting evidence, this means that the police can proceed with a prosecution.

High-Sensitivity Profiling

In some cases, the possibility of extracting DNA from crime scene samples will be reduced if the sample is old, degraded, or otherwise small in quantity. These can be subjected to more specialized techniques depending on the severity of the crime. All high-sensitivity work is performed in ultra-sterile conditions.

The FSS developed low copy number (LCN) to deal with samples containing insufficient good-quality DNA for standard profiling. The sample is subjected to 34 cycles of PCR to give more copies from which to draw a profile, but there are inherent difficulties. During 2007, a review was set up to examine the standards of science used in the analysis of LCN DNA, but before the review was complete, a challenge was made in the Northern Ireland case of *R v. Hoey*, a defendant charged with various offences in relation to the Omagh bombing in 1998. The case against him derived chiefly from DNA evidence using the LCN procedure, but the judge was not satisfied as to the integrity of the process and the prosecution case failed. This case was cited in the appeals of Reed, Reed, and Garmson in 2009 where it was found that in a very high proportion of profiles obtained using the LCN procedure, the profiles were not capable of robust and reliable interpretation.

DNASenCE (sensitive capillary electrophoresis) devised by LGC Forensics removes the impurities which interfere with the PCR process. The resulting profile is enhanced by a factor of 13. If necessary, the extracted material is loaded onto the capillary tube in greater concentration. This can enhance the resulting profile by a factor of 62. Other laboratories have devised similar

enhancement procedures. The major advantage of post-PCR cleanup and enhancement methodologies is that the original sample remains for resampling using alternative techniques, whereas 34-cycle LCN uses up all the original material. The profile can then be searched against the NDNAD as with a standard SGM + sample (Gross et al. 2009).

In addition to DNA SenCE profiling, other methodologies have been developed, for example, Minifiler STR analysis which has been particularly useful in "cold-case" investigations. The eight mini-STRs examined are smaller versions of eight of the regions looked at in standard SGM + profiling. The smaller size means they are more robust and less prone to degradation but resulting profiles are still compatible with database searching. Y-STR profiling is also of benefit in sexual assault cases where there is semen present but no sperm or where the victim's profile is likely to mask the profile of the offender, for example, scrapings from a victim's fingernails.

Promega's Powerplex 16 and Applied Biosystems' Identifiler both examine the same ten SGM + sites plus the sex determination site, as well as an additional five "smaller" sites in common and one more which differs in each case. These processes are useful where a sample is degraded and where the new, smaller sites are likely still to be extant.

Next Generation Multiplexes

Next Generation Multiplexes (NGMs) examine an increased number of sites comprising those presently used for SGM + profiling plus additional sites. There is a directive from the European Network of Forensic Science Institutes (ENFSI) urging member states using DNA profiling for forensic science purposes to standardize their methodologies and profile storage across the European Union by applying a European standard set of alleles to improve cross-border compatibility. The aspiration is that this will encourage more cross-border searching. Several commercial companies have developed NGM kits that meet the ENFSI recommendation and in the UK the decision on which Next Generation Multiplex (NGM) kit will be adopted rests with the National Policing Improvements Agency. Once this has been decided, all subject and crime scene samples will be processed using an NGM and the NDNADB will be gradually "upgraded." The potential advantages of this will be a harmonization of approach across Europe as well as greater discrimination in individual matches. The major disadvantage is that the profiles will of necessity be more complex and therefore much more difficult to interpret, particularly in the case of mixtures.

Contrary to the impression given in "CSI"-type media, even a standard crime scene sample can still take 48 h to process, with further time taken for interpretation, depending on the complexity of the result. Under exceptional circumstances, this time can be reduced, but only with greatly increased demands on equipment time and personnel therefore incurring greatly increased costs. Obviously, the more complex the profiling method used, the longer it will take to interpret the findings.

Mixtures and Partial Profiles

Partial profiles and mixtures of profiles are particularly difficult to interpret. A partial profile occurs when there is insufficient good-quality DNA to produce a full profile. The match probability of such a profile can be significantly reduced to the point where it is not possible to reach any conclusion as to evidential or even intelligence value. In order to calculate the probability of a partial match, reference databases are used to estimate the proportion of the STR profile in the corresponding populations. The rarity of certain characteristics is also taken into account.

A notable case using LCN profiling involving a mixture of profiles was that of *R. v. Broughton* (2010), jailed for 2 years and 8 months in 1999 in connection with violent animal rights activism when police found a firebomb in his car. He was arrested again in 2007 and remanded in custody after incendiary devices were found in Oxford University colleges. A jury cleared him of possessing explosives but failed to reach a verdict on other charges and so there was a retrial in 2009.

If an individual has inherited the same genetic information from each parent only one peak is shown. A mixed profile is apparent when more than two peaks are shown, as below.

From just a two person mixture it is possible to see four peaks at one locus.

From four peaks at one locus, there are these possible variations;

15,16 : 17,18
17,18 : 15,16
15,17 : 16,18
16,18 : 15,17
15,18 : 16, 17
16,17 : 15,18

DNA Profiling, Fig. 1 DNA electropherograms to demonstrate mixture interpretation

He was subsequently convicted and sentenced to 10 years in prison for conspiracy to commit arson. However, in 2010, this conviction was overturned on the basis that the DNA evidence had been unreliable. In the retrial, the uncertainty regarding the reliability of the DNA evidence hinged on differing interpretations of the results of processing which yielded the DNA profiles of more than one person on a matchstick which formed part of the incendiary device. The defense maintained that Broughton's DNA could not be present, and both prosecution and defense teams assembled impressive teams of DNA experts and statisticians using different software packages to calculate likelihood ratios. There was however fundamental disagreement on the question of how likely Broughton was to be included in the mixture or even how many profiles the mixture collected from the matchstick contained. Despite these differences of opinion, the jury returned a guilty verdict.

Mixed Profile Resolution

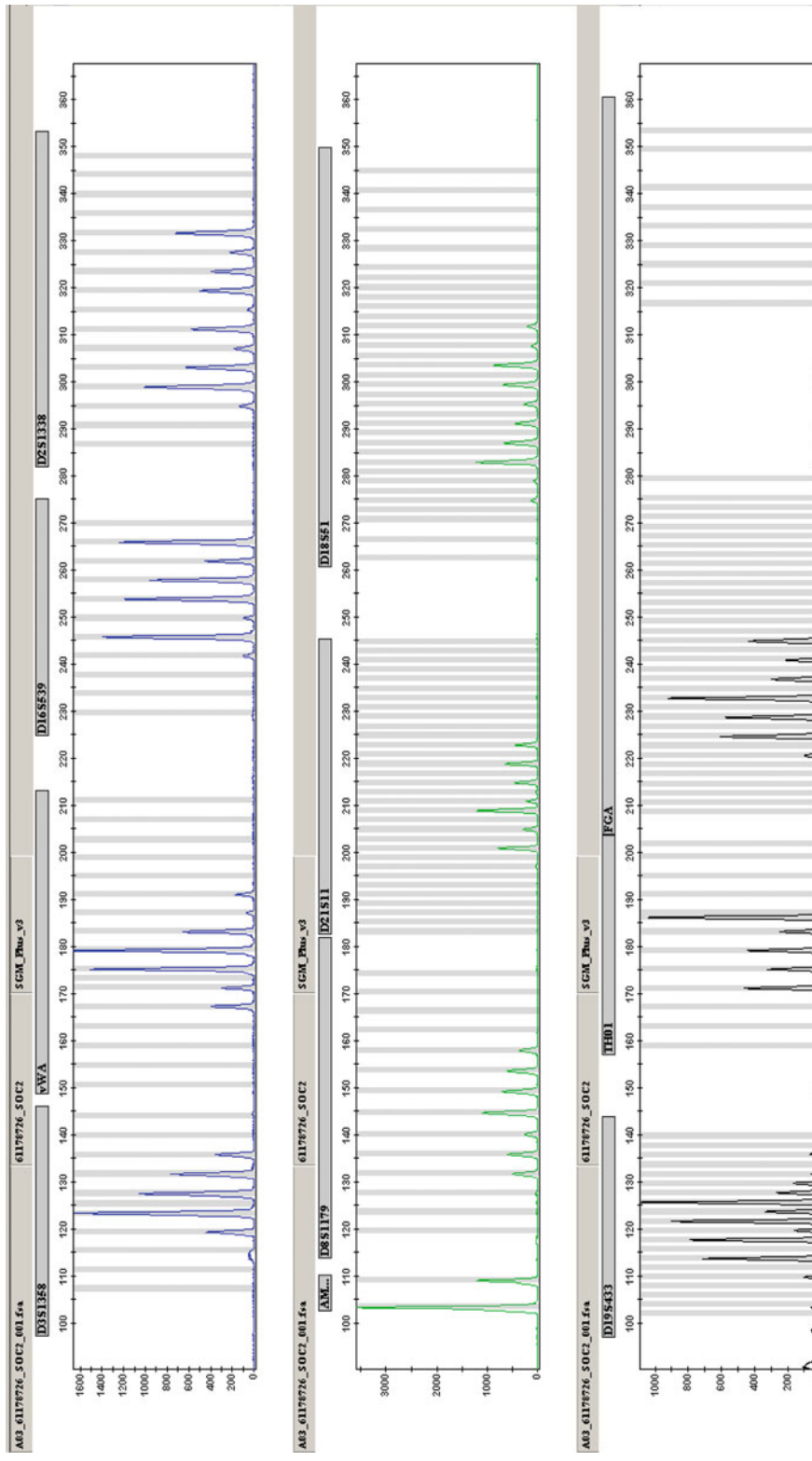
The components of a DNA profile are represented by a series of peaks that are measured and given a numerical value. A single-source profile will have two peaks at each locus, one from each parent (Figs. 1 and 2).

The more contributors to the mix, the more difficult interpretation becomes.

Without being able to subtract known profiles, such as from a victim or others known to have had the potential to contribute, the mixture has too many variables to interpret with certainty.

Interpretation

Each laboratory will set its own interpretation guidelines, giving a peak height at which a numerical value can be "read," but these are



DNA Profiling, Fig. 2 Example of a complex mixture

subject to variables including the nature of the offence and the source material. In addition, STR profiles can show "stutter," a term applied to peaks in a profile caused by the stochastic effects of PCR (Butler and Hill 2006). It is not certain why stutters occur, but they happen during the PCR process and show as a small peak usually one repeat smaller than the main band. Although a mostly clean, single-source profile may show stutter, this is easy to absorb in the overall interpretation, but the small bands often align with common alleles and so can be impossible to discriminate from true small peaks genuinely present from another profile as part of a mixture. Two peaks from the same source should be approximately the same size, but sometimes unidentifiable problems in profiling can lead to peak imbalance dramatically changing the appearance of a profile even though there is no background interference. In addition, stochastic effects can also occur. Referred to as "drop-in" and "dropout," drop-in is probably caused by contamination from an individual, the laboratory, or consumables. Dropout may be caused by a number of factors. There may be a problem with the primer leading to failure in amplification or the allele may be much larger than the usual size at a particular locus and is not seen with the others. DNA testing laboratories are continually attempting to adjust in order to cope with these difficulties but too little is known to be able to accommodate all eventualities successfully (Gill et al. 2000).

The Limitations of the System

Even with the incredible advances in technology, crime scene DNA profiling can be used to conclusively exclude a person from an inquiry, but even though it may provide compelling evidence of association, it is not proof of identity. Interpretation, particularly of complex mixtures, can be affected by many variables from interpretation guidelines differing between forensic providers to personal bias. In addition, it is estimated that the addition of partial profiles to the NDNADB may mean that approximately 0.1 % of matches are adventitious, that is, occurring by chance (Werrett 1997). The current sensitivity of profiling means that approximately 75 % of crime scene profiles generated are mixtures. Dr. Itiel Dror has recently expanded his studies on bias in fingerprint experts being dependent on external emotive factors to DNA-reporting officers (Dror 2012). In addition any process involving human beings is subject to human error, accidental or deliberate, and the pressure to perform efficiently in a commercial market place places additional pressures on forensic operators. Finally, despite extensive research in many different areas from collection to extraction to profiling, not enough is known about the propensity of individuals to leave their DNA and how that DNA is subsequently transferred between objects and individuals. While the Next Generation Multiplexes will increase the discriminatory power of profiling, the inherent increase in sensitivity will lead to more background DNA becoming part of a profile, so mixtures will be even more common, and the increase in the number of sites will make these mixtures even more difficult to interpret accurately and will take longer, with inevitable repercussions. Despite public perception, DNA profiling is never the complete answer to solving crime but continues to assist police forces as an aid to the criminal justice process.

Related Entries

▶ DNA Technology and Police Investigations

Recommended Reading and References

Butler JM. Fundamentals of forensic DNA typing

Butler JM, Hill CR (2006) Scientific issues with analysis of low amounts of DNA, 10th edn. American Medical Association

Dror IE (2012) Cognitive forensics and experimental research about bias in forensic casework. Sci Justice 52(2):128–130

Gill P et al (2000) An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. Forensic Sci Int 112:17–40

Gross T, Thomson J, Kutranov S (2009) A review of low template STR analysis in casework using the DNA SenCE post-PCR purification technique. Forensic Sci Int: Genet Suppl Ser 2(1):5–7

Jeffreys AJJ (1985a) Hypervariable "minisatellite" regions in human DNA. Nature 314(6006):67–73
Jeffreys AJA (1985b) Positive identification of an immigration test-case using human DNA fingerprints. Nature 317(6040):818–819
Quinones, Daniel (2012) Cell free DNA as a component of forensic evidence recovered from touched surfaces. Forensic Sci Int-Genet 6(1):26–30
Werrett DJ (1997) The national DNA database. Forensic Sci Int 88:33–42

DNA Technology and Police Investigations

Robin Williams and Matthias Wienroth
Centre for Forensic Science, Northumbria University, Newcastle Upon Tyne, UK

Synonyms

DNA databasing; DNA fingerprinting; DNA profiling; Forensic genetics

Overview

A forensic DNA profile is constructed by measuring highly polymorphic sequences of DNA in order to compare biological samples (especially blood, semen, skin cells, saliva, vaginal and nasal secretions, sweat, and other human tissue) found at a crime scene with samples taken from known individuals and those found at other crime scenes. First introduced in the 1980s, forensic DNA profiling is increasingly important to police investigations and criminal prosecutions in a large number of cases. However, despite its acknowledged successes in many criminal jurisdictions, debate continues about the general utility of forensic DNA technology to criminal investigations, the significance of due process and human rights challenges to the increasingly routine uses of DNA databasing, and the socio-ethical acceptability of some recent innovations in forensic DNA analysis.

History

The comparison of biological material recovered from scenes of crime with that taken from known individuals has a long pedigree in forensic science; the practice has been used widely to assist investigations and support the prosecution of offenders. Several different technologies, each seeking to capture varying degrees of distinctiveness exhibited by particular biological attributes, have been used to make these comparisons. Historically, the most widely accepted of these has been serological analysis in which blood samples are assigned to one of a small number of "ABO" blood types in order to include or exclude individuals as the possible sources of such material. However, in 1985, Alec Jeffreys and colleagues at the University of Leicester published two papers which demonstrated a new method for capturing individual differences at the genetic level (Jeffreys et al. 1985a, b). Described as "providing a level of individual specificity that was light-years beyond anything that had been seen before" (Newton 2004), the potential forensic science implications of what was first called "DNA fingerprinting" were quickly realized by Jeffreys and others.

The first, and highly prominent, deployment of this new technology in criminal investigations occurred only 2 years after Jeffreys' initial – and largely adventitious – laboratory discovery. Dawn Ashworth, a 15-year-old girl, went missing from her home in Northamptonshire on 31 July 1986. Her body was discovered 2 days later, and blood typing of semen recovered from her revealed identical features with semen obtained from the body of Lynda Mann who had been raped and murdered by an unidentified individual 3 years earlier (in both cases the semen donor was a Blood Group A secretor). The prime suspect for the murder of Dawn Ashworth was 17-year-old Richard Buckland, and following his arrest on 5 August 1986, he confessed to Ashworth's murder. However, Buckland was not Blood Group A and could not be linked to the semen recovered from Ashworth's body. In addition, he denied involvement in Mann's death. Faced with the contradiction between the biological evidence and Buckland's confession, investigators requested

Jeffreys to extract DNA from both recovered semen stains and to compare them to a blood sample taken from Buckland. Jeffreys' analysis concluded that Buckland's sample did not match the crime scene semen, but the semen taken from both crime scenes matched each other. Following Buckland's exoneration, the first ever mass DNA screening eventually resulted in the identification of Colin Pitchfork as the source of the semen, and Pitchfork was convicted of the two murders on 22 January 1988.

Despite this and other investigative successes in the UK and the USA, several technical limitations to this early form of forensic genetic analysis meant that it could be used only in a relatively small number of criminal investigations. These limitations included: the need to obtain relatively large quantities of DNA to undertake analysis; the time taken to complete the analytical process (several days or, in some cases, weeks); its unsuitability for use with degraded samples; the limited number of genetic markers that could be analyzed simultaneously; and the small number of samples that could be processed at one time. In addition, the criminal justice reception of forensic DNA fingerprinting was marked by an initial period of legal skepticism and methodological difficulty, especially concerning the estimation of "random match probability" and the communication of DNA profiling results by expert witnesses. The history of these "DNA Wars," especially the way in which they were conducted in several high-profile US criminal prosecutions, has been well documented by legal and human science scholars (see, for example, Kaye 2010; Lynch et al. 2008; Lynch and Jasanoff 1998).

Following these early controversies, and the stabilization of confidence in DNA profiling that resulted from their resolution, a series of subsequent changes in the methods for producing profiles also gradually overcame the technical shortcomings described above. The development of polymerase chain reaction (PCR), first reported by Kary Mullis in 1985/6, and its subsequent automation played a central role in this development by enabling the profiling of small and even degraded DNA samples through amplification. In the early 1990s, the shift to short tandem repeat (STR) multiplexes also meant that genetic information garnered through PCR became more easily comparable and digitizable, which made possible the automated computerized comparison of databased profiles.

DNA Evidence Recovery

In most jurisdictions, only trained crime scene examiners and forensic scientists recover biological material at scenes of crime. Some, however, also permit police officers to do so, at least in cases other than those of the most serious crimes. Regardless of these differences, investigators are increasingly conscious that aspects of the collection, packaging, recording, transporting, and laboratory handling of forensic genetic material can be subject to challenge in the course of judicial hearings, so all these practices are subject to strict protocols. What might otherwise be compelling DNA evidence can become "valueless if the authenticity of the samples used in the investigation cannot be confirmed" (Lincoln 1997). This means that very high levels of care are required to avoid prejudicing the weight otherwise given to the presence of DNA evidence recovered from crime scenes by lapses in continuity or by the use of inappropriate collection or preservations methods.

One clear feature of the trajectory of DNA profiling since the late 1980s has been the increasing success of forensic laboratories at being able to obtain analyzable quantities and qualities of DNA from an increasing variety of sources, including "trace" or "touch" DNA. In the early 1990s, Wiegand and associates were able to derive and profile DNA from debris obtained from fingernail scrapings and later from epithelial cells left on a victim's body following strangulation as well as cells left on strangulation tools. DNA analysis using an increased number of amplification cycles has been used successfully in UK forensic science casework since 1999 and a large number of studies have reported on the actual and potential uses of such "low copy number" (LCN) or low

template (LT-DNA) analysis. While this technique has made possible the production of profiles from very small and also degraded samples of DNA, these profiles are subject to a series of technical effects which can make their interpretation very difficult. There has also been one significant UK judicial ruling that questioned the robustness of this particular practice (Weir, J. [2007] "The Queen v Sean Hoey," The Crown Court Sitting in Northern Ireland). For these reasons, and despite recent work conducted on behalf of the UK Forensic Regulator, the use of LT-DNA remains contested in many jurisdictions.

DNA Databasing

Throughout the late 1980s and into the early 1990s, all applications of forensic DNA profiling technology to criminal casework required the temporary storage of DNA profiles from nominated suspects, or other persons of interest, so that each individual's profile could be compared with those obtained from crime scene samples. The limited accounts of these practices suggest that such profiles (and usually the physical samples from which they were derived) were held for varying periods of time in local police or forensic laboratory collections. These collections were not usually regulated by formal mechanisms or by external bodies, but their existence caused little unease, perhaps because the practice was not widely known outside of restricted police and forensic science circles. However, it was not long before the use of DNA profiling expanded beyond reactive forensic casework in some jurisdictions. Forensic scientists in the UK, the USA, Austria, and New Zealand argued that searching DNA profiles recovered from crime scenes against profiles held in larger and centralized databases could, by the provision of "cold hits," facilitate the early identification of many more potential suspects (as well as the exclusion of some others as persons of interest). Once further technological advances made it possible to construct easily transportable digital representations of profiles and store them in continuously searchable computerized databases, the stage was set for a vastly expanded role for DNA profiling in many criminal investigations. In turn, this meant that the limitations of existing and varied local collections required them to be replaced by more extensive forensic DNA databases capable of contributing to the successful detection and prosecution of more offenders – at least in jurisdictions where there was the political will and administrative resolve to support such innovations (descriptions of the early days of DNA databasing in these and other jurisdictions can be found in Hindmarsh and Prainsack 2010).

The last 20 years have witnessed an increasing number of criminal jurisdictions in which such forensic DNA databases have been established – usually, but not always, at a national level. The first national forensic DNA database was created in England & Wales in 1995. Three years later, it was followed by the official launch of the US Federal Bureau of Investigation's Combined DNA Index System (CODIS), although all 50 US state databases were not fully connected through CODIS until 2004. Many other nation states established their own national forensic DNA databases during the last decade of the twentieth century, and each year sees the addition of more states authorizing or creating such collections. These hybrid scientific-legal innovations have usually required legislative changes and the provision of additional funds to create and populate such databases with profiles obtained from known subjects and those obtained from biological material recovered from crime scenes. Particular commercial actors – largely but not exclusively biotech companies – have also been prominent advocates and supporters of these state-driven criminal justice ambitions, and their interests have complemented the enthusiasm of many prominent policing stakeholders.

There is also widespread public support for the use of forensic DNA profiling in most contemporary democratic societies. Allowing for some simplification, the global trajectory of forensic DNA database expansion has followed a distinctive shape in which the types of

people from whom DNA samples can be taken without consent, profiled, and retained have become greater. This has most often begun by sampling only those involved in the most serious crimes against the person, then moving on to include those involved (or suspected of being involved) in a range of property crimes. Police have also been authorized to take DNA samples at earlier points in investigative and judicial inquiries (for example, at the point of arrest rather than the point of charge, or even conviction). The period during which DNA profiles and samples can be retained has often been extended, and more DNA has been recovered from crime scenes. In addition to these developments, which together have resulted in the existence of much larger databases, the investigative applications of databased profiles have also expanded; there are increasing efforts to make possible the sharing of DNA profile information between criminal jurisdictions, for example, within the European Union via the Prüm Treaty, and beyond the EU through INTERPOL.

Current Databasing Practice

Taking DNA from suspects, and using these and crime scene DNA profiles as intelligence to support criminal investigations and as evidence to support prosecutions, are now central aspects of policing in a large number of countries. In those jurisdictions that have established a "national" DNA database, the vast majority of profiles entered are obtained from suspects during the investigation of crime. For those eager to promote and extend the powers of the police to sample criminal suspects, emphasis is placed on the immediacy of DNA profiling to exonerate individuals from, as well as implicate individuals in, further criminal investigations. The point is often made that enabling the police to obtain DNA samples from suspects introduces a reliable and objective method of evaluating their presence at a scene of crime. Yet, in order to obtain such a sample, the suspect must undergo a procedure to extract bodily material, and beliefs about the significance of this bodily intrusion color legislative decisions about forensic DNA sampling. Most states have enacted legislation that carefully specifies the situations in which the police may legally "interfere" with bodily integrity. Affording the police the authority to take DNA samples without consent involves making a number of legislative decisions. The first involves deciding at what stage in criminal procedure an individual should be subject to compulsory DNA sampling. A second, and related, issue is whether the police themselves should possess the authority to administer the collection, or if they should be required to obtain judicial approval. A third decision pertains to the types of offenses that should allow the compulsory sampling of suspects. And, following that, a fourth issue is whether such sampling should be relevant to the specific offense in question. A subsequent question arising is that of DNA retention: from whom, for how long, in what form (biological sample and/or digitized profile), and in what ways may the retained DNA material be used in the future.

Some states have reduced the significance of these issues and thus maximized the possible sampling of suspects. Equally, other states have minimized sampling because they have assigned more significance to them. The obvious example in the former category is the UK (England & Wales) which permits compulsory DNA sampling at the earliest point of investigation (upon arrest of a suspect) by the police for any "recordable" offense regardless of whether such a sample is relevant to the investigation. An example of a country in the latter category is France where nonconsensual sampling of suspects is completely prohibited.

In some states of the European Union (for example, The Netherlands, Luxembourg, and Malta), police are able to take compulsory DNA samples from individual suspects but such sampling requires judicial authority. Requiring the police to obtain this authority is a significant element in the distribution of powers across the criminal justice system. It prohibits the automatic sampling of criminal suspects by the police and transfers authority elsewhere, requiring the police to make a strong argument for compromising a person's bodily integrity. Nevertheless, it is increasingly common in many

jurisdictions for the police to take DNA during the investigation of certain types of offenses. Most often, these are serious offenses which involve violence against persons. And some countries possess legislation which limits the collection of DNA from suspects in relation to specific – usually more serious – offenses. The situation in the USA is made more complex by differences between the states of the Union, but there seems a general trend there to establish "arrestee databases" in which DNA is taken, stored, and speculatively searched on arrest, even if subsequently removed (or at least sequestered) when individuals are not prosecuted, or if prosecutions do not result in findings of guilt. The many and constant changes in legislative frameworks governing forensic DNA databasing across the world are monitored by several police and civil society groups; from time to time, publications by INTERPOL and by a consortium of UK and US civil society groups provide serially updated accounts of these developments (see www.interpol.int and http://dnapolicyinitiative.org/).

Key Issues and Controversies: Claims-Making and the Measurement of Utility

Criminal justice researchers in the UK were among the first attempting to establish the value of DNA profiling and databasing to criminal investigations other than by citing conspicuous case successes. A series of studies, all funded by the Home Office, reported mixed success in this regard, although some UK government publications confidently asserted significant gains in the proportion of crimes detected when DNA evidence was available to investigators (McCartney 2006a, b). More recent studies have been carried out in the USA, mostly funded by the National Institute of Justice. Two of these have been especially ambitious. Roman and colleagues (Roman et al. 2008) carried out the first randomized control trial of the use of forensic DNA profiling across several US police districts in which the results of burglary investigations during which DNA evidence was collected but not analyzed were compared with those in which such evidence was made available to investigators. A second major study, of the role of forensic science in securing prosecutions, but also providing separate data specifically on DNA profiling, has recently been completed by Peterson and others (Peterson et al. 2010). This work followed the investigative process in a number of different kinds of crime (including homicide, rape, aggravated assault, robbery, and burglary) through the criminal justice system in order to assess the contribution of forensic science to the outcome of investigations and prosecutions. While a range of forensic evidence types was considered, particular attention was given to DNA analysis because of its ability to provide individualizing evidence capable of associating particular suspects to crime scenes. Despite these and other studies, there remains disagreement over the extent to which research has yet provided a robust account of the utility of DNA profiling and databasing to criminal investigations. The UK Human Genetics Commission, along with academic researchers in many jurisdictions, has commented on the need for better data to be provided by the custodians responsible for the operation of national DNA databases as well as by the police who are responsible for the use of DNA match information in support of individual investigations (Human Genetics Commission 2009).

Key Issues and Controversies: DNA Databases and Human Rights

All sampling and databasing of the genetic profiles of individual suspects by the police – especially those taken without consent – involves consideration of a number of legal, social, and ethical issues. As reflected in the comment of the Council of Europe Committee of Ministers from 10 February 1992, such sampling must "[t]ake full account of and not contravene such fundamental principles as the inherent dignity of the individual and the respect for the human body, the rights of the defence and the principle of proportionality in the carrying out of criminal justice." For some critics, forensic DNA sampling and databasing threaten the bodily integrity of citizens who are subject to the

forced and nonconsensual sampling of their genetic material based on decisions of police and judicial actors prior to findings of guilt by relevant authorities. In addition, there is concern that DNA databasing also violates privacy rights by allowing the use of profiles and the storage of biological samples, storage which itself creates the potential for the future misuse of such samples held in state and privately owned laboratories.

Accordingly, these and other legal, social, and ethical issues have been explored in several academic and policy studies. In the UK, two major agencies – the Nuffield Council on Bioethics (2007) and the Human Genetics Commission (2001, 2002, 2009) – have both published substantial critical reports on the distinctively forceful legislative and operational developments in DNA databasing that occurred in England & Wales between the establishment of the National DNA Database in 1995 and the decision of the Council of Europe's European Court of Human Rights in the case of S & Marper v the UK Government in 2008.

In addition, the monitoring group Genewatch UK have also been actively interrogating official statements and statistics on the National DNA Database for a number of years as well as appearing before several House of Commons Select Committees that have inquired into aspects of forensic DNA profiling and databasing in England & Wales (see http:www.genewatch.org/). In the USA, the American Civil Liberties Union has frequently criticized the state and federal expansion of DNA collection and retention, and the American Society of Law, Medicine and Ethics sponsored a series of national workshops and conferences on US developments in DNA profiling, attended by many of the major academic, scientific, and operational experts in the field between 2003 and 2006.

Key Issues and Controversies: Recent Innovations in Forensic DNA Analysis

There are many investigations in which genetic material has been recovered from a crime scene but no matches with databased profiles have been made. In such circumstances, investigators may seek other ways to infer personal features of an unknown individual from the DNA that they have left at the scene. New forms of genetic knowledge, technological improvements in sample processing, and the premium on the investigative ingenuity necessary for the detection of "hard-to-solve" serious crime contribute the means and desirability for constant innovations in methods for interrogating the informational content of biological samples obtained from scenes of crime. At present, analysis can be undertaken to gain information about phenotypical attributes, "biogeographic ancestry," and "familial relationships." Some interrogations involve the direct examination of coding regions of the human genome – genes themselves – while others rely on new ways of using information from the noncoding areas already examined by conventional forensic profiling. The most significant of these approaches are briefly described in the following sections.

"*Genetic Ancestry*," "*Population Groups*," *and Forensic Investigations*. One particularly complex area of genetic information of interest to criminal investigators has been that of patterned human genetic diversity. Knowledge of the differential variability of genotypes according to population groups has informed the calculation of random match probabilities since the early days of DNA profiling. While population genetics are largely of interest to the specialized forensic community, the ability to infer the "biogeographic" origin of an individual who left otherwise unidentified biological material at a crime scene may be of significant interest to investigators. Reliable inferences of the "racial origins," "ethnic origin," "ethnic affiliation," or even "ethnic appearance" of such an individual can be used to focus subsequent inquiries, to determine an interview strategy, to compare with witness statements, or to design an intelligence-led mass DNA screen. However, there are significant conceptual and operational uncertainties surrounding such categorizations of individuals. Moreover, there is a danger, well

articulated, for example, by Duster and colleagues, that "race" will be reified in the attempt to define distinctive human population groups and subgroups. These critics also point to the ways in which questions of genetic "ancestral attribution" for these limited and pragmatic purposes can easily become confused with more ambitious theoretical assertions concerning the biology of "race" as well as "some old and dangerously regressive ideas about how to explain criminal conduct" (Duster 2003: 151).

Despite these problems, a number of forensic laboratories and agencies have added to their analysis of autosomal DNA STRs and SNPs, Y-chromosome STR and Y-chromosome SNP multiplexes for the analysis of loci whose polymorphic range is already databased by a variety of international consortia. The Y-chromosome is an especially suitable site for such investigations because of the low rate of recombination on this chromosome. This means that particular male-specific haplotypes are preserved across generations and vary systematically across different population groups. The research and reference databases used to inform developments of this kind are the Y-STR haplotype reference database, the US population database, and a European population database (see www.yhrd.org).

Autosomal Single Nucleotide Polymorphisms (SNPs). There have been several surveys of such polymorphisms, many of which are collected together in a global haplotype collection known as the "HapMap" (www.hapmap.org). SNPs have a much more limited polymorphic range than STRs, so that about four times as many SNPs are needed to produce profiles capable of discriminating individuality as those used by STR typing. Nevertheless, the establishment and expansion of SNP forensic databases alongside current STR collections is not out of the question, and the analytical scope of SNPs means that they can serve valuable forensic identification functions, especially in situations where samples are too degraded to make STR typing possible. The Y Chromosome Consortium has provided a "phylogenetic tree" which describes the history of 18 major lineages of diverse SNP haplotypes across human population, and other scholars recently discussed the development and use of Y-SNP multiplexes to support inferences of population origin.

It is difficult to determine the significance of these efforts to provide information about genetic ancestry. A clear preference for SNP markers over STRs seems to have emerged over the last few years, and large numbers of SNPs are now combined to form "ancestry informative markers," some of which have been used in forensic casework. However, until recently, there have been problems in standardizing such markers and in the standardization of haplotype nomenclature. Furthermore, the increase in populations of mixed origin is a feature of complex urban societies so that "indirect deductions about individuals are often unreliable" (Jobling 2001: 161). Even when the pattern of differential SNP distributions is used to "improve" the accuracy of such inferences, as in the case with "proportional ancestry" studies, significant uncertainties remain.

Inferring Specific Phenotypical Features. In addition to efforts at identifying the genetic ancestry of unmatched crime scene stains, forensic scientists and police investigators remain interested in whether interrogations of such stains may yield information about a wide repertoire of visible characteristics of their donors. For the purposes of police investigations, the ability directly to determine individuals' physical characteristics may be more appealing than inferring those characteristics from assumptions of biogeographic ancestry. The most frequently used method of direct interrogation – of the amelogenin locus – determines the biological sex of the DNA source and is already incorporated into the majority of multiplex systems. Aside from this test, however, the research literature reveals limited success in attributing phenotype from genotype in ways that are practically useful to investigators.

An initial review of forensic work in this field suggests that positive results remain scarce, and are focused on probabilistic inferences about pigmentation (see for example Kayser and de Knijff 2011). While analysis of the human melanocortin-1 receptor gene can be used to

indicate "red hair" in the relevant subject, hair loss and hair coloring can make this test problematic when applied in investigative contexts. Both STR and SNP profiling are of interest to forensic scientists keen to develop predictive tests for a range of other observable physical characteristics including eye color, skin type, and height. It seems likely that SNP analysis may prove more successful than STR markers as the basis for such tests. This is not simply because most genomic mutations are single base changes but also because there is considerable research being carried out beyond the forensic community to identify SNP polymorphisms and their effects on a variety of human attributes.

Familial Searching. The term "familial searching," as used by forensic scientists and police officers, refers to a form of database searching reliant on knowledge about the probability of matches between the STR markers of two members of the same family (as opposed to the probability of matches between these markers when the individuals compared are unrelated). This practice makes use of understandings of inheritance that prefigured the discovery of the structure of DNA and which had been largely applied to understanding variation in human, animal, and plant phenotypical characteristics (for a summary account of these assumptions as applied to the forensic context, see Bieber et al. 2006). Because familial searching relies on identifying a pool of possible genetic relatives of a suspect, who are then subject to more direct investigation (typically by being interviewed by the police), forensic science policy makers in the UK and elsewhere have also acknowledged that a number of ethical issues need to be addressed when this strategy is being considered (see Greeley et al. 2006; Haimes 2006). Issues arise in both the searching of profiles on a database and in the subsequent investigative trajectories that follow the provision of a list of individuals derived from such a search. A genetic link between individuals might be previously unknown by one or both parties and police investigations may make such information known to them for the first time. Equally an investigation may reveal – to investigators, if not to informants – the absence of genetic links which participants assumed to have existed. There is also the question of whether this kind of use of an individual's databased DNA violates promises of privacy and confidentiality made when genetic material was originally donated voluntarily, for example, in the course of a mass DNA screen. Furthermore, assertions about criminality, geography, and familial relatedness that are central to the use of this forensic methodology are especially problematic – even if they do accord with the rhetorical endoxa of many detectives – and they reveal pervasive problems associated with the confusion between "genetic" and "social" relatedness ("families" are not only constituted through genetic lines but through clusters of non-genetically related individuals) as well as the implicit assumption that criminality is fostered because of such relatedness (either for genetic or social reasons).

Conclusion

This entry has outlined the history of the main DNA technologies that are currently used in police investigations along with their reception by criminal justice actors, especially in the UK and the USA. The uncertainties reflected in the "DNA Wars" of the 1980s and 1990s have diminished as scientific and legal agreements about the strength and limitations of many of these technologies have stabilized. Police enthusiasm for DNA technology has grown, and in many jurisdictions, legislators have increased police powers in order to maximize its potential uses to support criminal investigations and prosecutions. A variety of international bodies (especially the European Network of Forensic Science Institutes, The American Society of Crime Laboratory Directors, and the International Society for Forensic Genetics) support efforts to standardize forensic genetic practice and shape the education of relevant scientific and legal personnel. At the same time, new normative and empirical uncertainties arise as legislative, operational, and technical innovations are critically appraised by scientific, legal, and human science scholars.

Questions of scientific adequacy, investigative utility, and ethical acceptability, as well as the relationship between each of these questions, are sure to engage criminological interest for the foreseeable future.

Related Entries

- ▶ Cognitive Forensics: Human Cognition, Contextual Information, and Bias
- ▶ DNA Profiling
- ▶ Forensic Science and Criminal Inquiry
- ▶ Forensic Science Culture
- ▶ History of Forensic Science in Policing
- ▶ Identification and the Development of Forensic Science

Recommended Reading and References

Aronson JD (2007) Genetic witness: science, law and controversy in the making of DNA profiling. Rutgers University Press, New Brunswick

Bieber FR, Brenner C, Lazer D (2006) Finding criminals through DNA of their relatives. Science 312:1315–1316

Duster T (2003) Backdoor to eugenics, 2nd edn. Routledge, New York

Fraser J, Williams R (2009) Handbook of forensic science. Willan, Cullompton

Gerlach N (2004) The genetic imaginary: DNA in the Canadian criminal justice system. University of Toronto Press, Toronto

Greeley HT, Riordan DP, Garrison NA, Mountain JL (2006) Family ties: the use of DNA offender databases to catch offenders. J Law Med Ethics 34:248–262

Haimes E (2006) Social and ethical issues in the use of familial searching in forensic investigations: insights from family and kinship studies. J Law Med Ethics 34(2):263–276

Hindmarsh R, Prainsack B (2010) Genetic suspects: global governance of forensic DNA profiling and databasing. Cambridge University Press, Cambridge

Human Genetics Commission (2001) Whose hands on your genes? A discussion document on the storage protection and use of genetic information. HGC, London

Human Genetics Commission (2002) Inside information: balancing interests in the use of personal genetic data. Department of Health, London

Human Genetics Commission (2009) Nothing to hide, Nothing to fear? Balancing individual rights and the public interest in the governance and use of the national DNA database. HGC, London

Jeffreys AJ, Wilson V, Thein SL (1985a) Individual-specific 'fingerprints' of human DNA. Nature 316: 76–79

Jeffreys A, Wilson V, Thein SL (1985b) Hypervariable 'minisatellite' regions in human DNA. Nature 314: 67–73

Jobling MA (2001) Y-chromosomal SNP haplotype diversity in forensic analysis. Forensic Sci Int 118:158–162

Kaye DH (2010) The double helix and the law of evidence. Harvard University Press, Cambridge, MA

Kayser M, de Knijff P (2011) Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet 12(3):179–192

Lincoln PJ (1997) Criticisms and concerns regarding DNA profiling. Forensic Sci Int 1997:23–31

Lynch M, Jasanoff S (1998) Contested identities: science, law and forensic practice. Soc Stud Sci 28(5–6):675–686

Lynch M, Cole S, McNally R, Jordan K (2008) Truth machine: the contentious history of DNA fingerprinting. Chicago University Press, Chicago

McCartney C (2006a) The DNA expansion programme and criminal investigation. Brit J Criminol 46:175–192

McCartney C (2006b) Forensic identification and criminal justice: forensic science, justice and risk. Willan, Cullompton

Newton G (2004) Discovering DNA fingerprinting 2004. http://genome.wellcome.ac.uk/doc_wtd020877.html Accesed 21 December 2012

Nuffield Council on Bioethics (2007) The forensic uses of bioinformation: ethical issues. The Nuffield Council, London

Peterson J, Sommers I, Baskin D, Johnson D (2010) The role and impact of forensic evidence in the criminal justice process. National Institute of Justice, Washington, DC

Roman JK, Reid S, Reid J, Chalfin A, Adams W, Knight C (2008) The DNA field experiment: cost-effectiveness analysis of the use of DNA in the investigation of high-volume crimes. Urban Institute, Washington, DC

Williams R, Johnson P (2008) Genetic policing: the use of DNA in criminal investigations. Willan, Cullompton

Domestic Burglary

- ▶ Residential Burglary

Domestic Terrorism

- ▶ Homegrown Terrorism in the United States

Domestic Violence

Russell P. Dobash[1,2,3] and
Rebecca Emerson Dobash[1,2,3]
[1]School of Law, University of Manchester, Manchester, UK
[2]School of Social Justice and Inquiry, Arizona State University, Glendale, AZ, USA
[3]Centre for Law, Crime and Justice, School of Law, Strathclyde University, Glasgow, Scotland

Violence has long been a subject of interest for criminology, but it was not until the last quarter of the twentieth century that gender became a significant explanatory factor and domestic violence became a focus of research. Although an early, classic work by Wolfgang, Patterns of Criminal Homicide (1958), included a groundbreaking investigation of domestic murders, this was mostly ignored until the 1980s. In the USA, where most criminological research was conducted, the focus was on gang delinquency. This focus drew upon the pioneering work of Merton (1938) and produced a number of empirically informed accounts of gang delinquency and gang violence (Short and Strodtbeck 1965). The prevailing characterization of violent events was of a "face game" with disputes involving "honor" in which both victim and offender "agreed" that violence was an appropriate means of settling the conflict as verbal responses escalated to physical violence (Luckenbill 1977). The concept of "victim precipitation" defined the mutual and equally culpable participation of both parties as each "agreed" to the escalation from verbal encounter to physical violence.

When domestic violence (DV), now termed intimate partner violence (IPV), was "discovered" in the mid-1970s, the sparse accounts of physical and sexual violence against women within criminology relied heavily upon the concept of "victim precipitation" with its implicit conception of male-male encounters, provocation, and mutual agreement to violence, and this was applied to accounts of conflicts between men and women culminating in men's violence against women (e.g., Amir 1971). Similarly, within the psychiatric literature, women victims of domestic violence were deemed to be "provocative," "aggressive," and/or "masculine," and thus responsible for the violence men used against them (for a review see Dobash and Dobash 1979, 1992). Thus, as social scientists began to focus on domestic violence in the 1970s and 1980s, often in conjunction with activist community groups, many looked outside criminology to the emerging feminist literature that stressed gender, male domination, aggression, and violence as more relevant to this area of research. Such approaches have generally dominated the study of IPV, and the body of literature is now voluminous and includes criminology, socio-legal studies, sociology, anthropology, psychology, evolutionary psychology, health, and medicine. Here, the focus is upon the extent of the violence, the nature of violence and controlling behavior, the context in which it occurs, the characteristics of male perpetrators, and the response of the criminal justice system.

Since the "discovery" of this form of violence in the 1970s, there have been debates about definitions, terminology, research methods, and the resulting findings. Over time, several terms have been used: "wife beating," "wife abuse," "wife battering" which were followed by "woman beating, abuse, or battering," "domestic violence," and "intimate partner violence." There have been debates about *how much of this violence exists* in a given place at a particular time; *who commits the violence* (men, women, or both, i.e., symmetry or asymmetry in perpetration by men and/or women); a *narrow vs. broad definition of what counts as violence*, narrowly restricted to physical acts of violence or broadly expanded to include other acts that are emotional, financial, and the like; *what should be done about it and by whom*; as well as other debates not discussed here (see Dobash et al. 1992; Dobash and Dobash 2004).

Nature and Extent of Intimate Partner Violence

Here, the focus is on serious, physical "violence" against women rather than upon aggression,

controlling behavior, emotional abuse, or financial deprivation. This is not to say that such nonviolent aggressive acts are not problematic or worthy of concern, intervention, or research. However, it is physical acts of violence that are most likely to be defined as illegal or criminal and to warrant intervention by the justice system including police, courts, probation, and prison. Around the world, several national and international studies have attempted to establish the nature and extent of intimate partner violence. When the focus is upon serious physical violence, the overwhelming evidence indicates that it is women, not men, who are significantly more likely to be the victims of violence from an intimate partner, to suffer serious consequences and injuries, and to require emergency attention and hospitalization (Krahe et al. 2005; Tjaden and Thoennes 1998; Watson and Parsons 2005).

National population-based surveys in several industrialized countries suggest that about one-quarter of adult women will at sometime in their life experience at least one act of violence from a male intimate partner (Backman and Saltzman 1995; Greenfield et al. 1998; Mirrlees-Black 1999; Tjaden and Thoennes 1998; Wilson et al. 1995; Walby and Allen 2004). Evidence collected by the World Health Organization reveals that intimate partner violence is a "common experience" for women throughout the world (Krug et al. 2002). A WHO meta-analysis of 50 population-based studies from 35 countries found lifetime estimates of between 10 and 69 %. In most countries this varied from 10 % to 50 % (Krug et al. 2002). A subsequent carefully conducted WHO study involving probability samples and standardized face-to-face interviews with 24,097 women at 15 sites in ten countries revealed a lifetime prevalence of 13–61 % for moderate or severe violence (acts capable of inflicting injury and hospitalization) from partners of ever cohabiting or married women with most countries reporting prevalence of 25–45 % (Garcia-Moreno et al. 2006).

Sexual violence against women in intimate relationships also occurs with the WHO research revealing that between 6 % and 59 % of women had experienced "forced sexual intercourse" by an intimate male partner, with most areas varying from 10 % to 50 % (Garcia-Moreno et al. 2006; Watson and Parsons 2005). Research conducted primarily in the USA indicates that 10–15 % of ever married or cohabiting women have been raped by an intimate partner, about one-quarter of all rapes involve intimate partners, and a considerable proportion of physically abused women also suffer sexual assault (Campbell 1999; Randall and Haskall 1995; Russell 1990; Ullman and Siegel 1993). Results of the comprehensive WHO study indicate that the violence is often severe and results in serious injuries, it occurs on a frequent basis, there is a strong relationship between frequency and severity, and it is intrinsically related to sexual violence (Garcia-Moreno et al. 2006). This definitive study corroborates and confirms the results of scores of other national and local surveys conducted in countries throughout the world.

Crime victimization surveys and evidence gathered from other sources have persistently shown that women are much more likely than men to be victimized by an intimate partner, to suffer injuries, and to require medical treatment. In the USA, National Crime Surveys have been conducted annually since 1972, and these and victimization surveys conducted in other countries have persistently shown that women constitute 70–90 % of all victims of assaults between intimate partners and that women are much more likely than men to report serious injuries (Archer 2000; Gaguin 1977–1978; Schwartz 1987; Johnson and Sacco 1995; Tjaden and Thoennes 1998; Worrall and Pease 1986). Evidence from other sources, such as police, court, and accident and emergency records, gathered in numerous countries over a number of years suggests that in an overwhelming proportion of cases, women are the victims of intimate partner violence (Abbott et al. 1995; Archer 2000; Dobash et al. 1992; Dobash and Dobash 2004; Nazroo 1995).

Extensive population-based surveys and intensive studies provide information about the specific nature of the physical acts of violence perpetrated by men in intimate relationships and repeatedly reveal a range from the more frequent

use of pushing, shoving, slapping, punching, and kicking to the less frequent use of weapons and strangulation (Dobash and Dobash 1979; Dobash et al. 1998; Johnson 1996; Tjaden and Thoennes 1998).

Injuries

Research suggests that the most common injury from violent attacks involves bruising and lacerations of the face and body. Fractures, concussions, miscarriages, and internal injuries also occur although less frequently (Campbell 1998; Dobash and Dobash 1998; Coker et al. 2000; Kyriacou et al. 1999; Stark and Flitcraft 1992). Health and medical research indicates that permanent disfigurement, physical disability, and damage to hearing and vision sometimes occur, that abused women are six to eight times more likely to use health services than non-abused women, and that violence during pregnancy threatens the health of the woman and the fetus (Campbell 1998; Walton-Moss et al. 2005). Women are also likely to experience an ongoing sense of fear, helplessness, entrapment, and loss of self-respect (Dobash and Dobash 2004; Watson and Parsons 2005), as well as other long-term negative consequences to their health and well-being.

Constellation of Abuse

Intimate partner violence is frequently linked to other acts that do not involve physical violence but are controlling, intimidating, and coercive. Such acts do not break bones or cause bruising or bleeding, but may result in fear, intimidation, and damage to self-worth. The impact of coercive and controlling behaviors usually rests upon the fact that physical violence has previously been used and could be used again. This provides a firm foundation upon which to use intimidation in order to maintain authority and control. We have defined this combination of physical violence, injuries, and controlling behavior as the "constellation of abuse" (Dobash and Dobash 1984; Dobash et al. 2000). Our historical and intensive studies confirm the link between violence and other forms of intimation and coercion (Dobash and Dobash 1979; Dobash et al. 1998, 2000), and this link has been corroborated in several population surveys (Tjaden and Thoennes 1998; Walby and Allen 2004; Wilson et al. 1995). A survey by the World Health Organisation (WHO) revealed a direct relationship between violence against intimate partners and various forms of controlling and intimidating acts by men (e.g., restricting her mobility and access to friends and family). While levels of controlling behaviors vary from country to country (from 21 % to 90 %), systematic research from ten countries indicates a strong statistical relationship between "severe" restrictive controls and physical and/or sexual violence (Coker et al. 2000). Violence and other sustained forms of abuse not only result in physical injury and psychological distress but may also result in chronic mental health problems (Coker et al. 2000; Kyriacou et al. 1999).

This constellation of abuse may even extend beyond the "end" of a relationship as some men continue try to control and/or punish the woman for leaving or for beginning a new relationship. The woman may be stalked and/or subjected to further violence, and others (relatives, friends/neighbors, and new partners) may also be subjected to violence or intimidation in an ongoing effort to punish and control. A representative sample survey of 8,000 women in the USA revealed that 81 % of the respondents who reported having been "stalked" by a former partner indicated that they had previously been assaulted by the person who continued to harass them. Additionally, 31 % of these women also reported a previous incident of sexual assault (U.S. Dept. of Justice 1998a).

Type of Relationship: Marriage, Cohabitation, and Dating

The type of intimate relationship has been found to be important in several respects. Research on lethal and nonlethal violence suggests that cohabiting relationships are more at risk of lethal

and nonlethal violence than marital relationships (Miethe and Regoeczi 2004; Shackleford and Mouzos 2005; Wilson et al. 1995). Additional research expanded to include dating, nonresidential relationships revealed that both cohabiting and dating relationships were at a greater risk of lethal violence than marital relationships (Dawson and Gartner 1998; Dobash et al. 2007; Johnson and Hotton 2003). The preponderance of evidence suggests that cohabitation and serious dating/engaged relationships have a greater risk of lethal and nonlethal IPV than state-sanctioned marriage. It may be that such relationships are more tenuous, involve less commitment, have fewer outside supports in the form of relatives or the state, and these circumstances leave the couple with fewer resources to deal with the inevitable conflicts of domestic life as well as reduce the likelihood of external intervention. Alternatively, any observed difference in the risk of IPV relating to the type of relationship may be associated with the distinct characteristics of those in these different types of relationship. Men and women who cohabit or are in a boy/girlfriend relationship are likely to be younger, poorer, and in other ways categorically different than those who are married (Brownridge and Halli 2002). It may be, however, that the observed differences are related both to the characteristics of the individuals involved and to the sociocultural factors associated with various types of relationships.

Conflict, Nonlethal, and Lethal Violence

Conflict usually precedes violent events, and conflicts (often chronic) in intimate relations characterized by violence include a number of recurring issues associated with daily life including: money, children, housekeeping, sex, fidelity, jealousy, possessiveness, authority, and the continuation of the relationship (Dobash and Dobash 1979; Dobash et al. 2000). Men's sense of entitlement, jealousy, and possessiveness are major issues and may be even more apparent in cases that end in murder (Block and Christakos 1995; Campbell et al. 2007; Dobash et al. 2007; Dobash and Dobash 2011; Polk and Ranson 1991; Serran and Firestone 2004; Wilson and Daly 1998). At the point when women attempt to leave, or terminate a relationship, issues of possessiveness and "ownership" become very apparent, and the combination of these factors appears to contribute to an elevated risk of lethal and nonlethal violence against women. Evidences from several studies of intimate partner murder suggest that at the time of the murder, one-third to one-half of women killed by a partner were either separated from their partner or had indicated their intention to leave the relationship (Browne et al. 1999; Dawson and Gartner 1998; Dobash et al. 2007 Johnson and Hotton 2003; Wilson and Daly 1993). The early stages of separation appear to be the most risky, although some men stalk and kill an ex-partner several years after separation.

Lethal violence in intimate relationships primarily involves men as perpetrators and women as victims, although abused women do sometimes kill their male abuser. In a number of countries, the ratio of male-to-female victims is around one to five, and in some societies there are no reports of women killing male intimate partners (Campbell et al. 2007; Daly and Wilson 1988; Dobash et al. 2007). The homicides of women intimate partners constitute 40–60 % of all murders of women, whereas no more than 10 % of men who are murdered are killed by an intimate partner (Campbell et al. 2007; Dahlberg and Krug 2002). In the USA such murders are one of the leading causes of premature deaths of women and the seventh leading cause of death for African-American women aged 15–45 (U.S. Office of Justice 1998b). When men murder an intimate partner, it usually occurs in the context of sustained, long-term violence and abuse. Research suggests that at least 60 % of all murders of women by intimate partners are associated with the chronic violent abuse by the male perpetrator (Campbell et al. 2007; Dobash et al. 2007; Moracco et al. 1998). A considerable proportion of these cases also involve a history of sexual violence which occurs in about 15 % of the murders of women (Campbell et al. 2007; Dobash et al. 2007; Mathews et al. 2004).

When women murder a male partner, it usually occurs in the context of the man's physical and sexual violence against her and frequently involves self-defense and/or retaliation (Browne 1987).

Collateral Murder Involving Intimate Partner Conflict/Violence

Men not only murder their women partners and ex-partners in the context of intimate partner violence and conflict, they also kill others, such as children or new partners, who might be described as collateral victims (Dobash and Dobash 2012; Langford et al. 1998). Children appear to be the most likely collateral victims, but research suggests that those who attempt to shelter or protect women, usually family and friends, as well as new male partners, are also at risk (Dobash et al. 2007; Dobash and Dobash 2012). The widespread availability of firearms, particularly in the USA, means that police officers attempting to intervene in "domestic disputes" are at considerable risk of injury or death. Men sometimes commit familicide, the murder of the entire family including their woman partner and children, often followed by suicide (Wilson and Daly 1998; Websdale 2010). Such actions are extremely rare among women. Most mass murders (more than two victims) are committed by men in the context of intimate partner conflict/violence, as are the majority of murders followed by suicide (Liem 2010; Websdale 2010).

Male Perpetrators of Intimate Partner Violence and Murder

Childhood and Adulthood

Clinical survey and longitudinal studies suggest that men who have committed violent acts against an intimate woman partner are significantly more likely than those who have not to have experienced adversity in childhood and various problems in their adult life (Browne et al. 1999; Moffit et al. 1998; Schumacher et al. 2001). Witnessing domestic violence, along with physical abuse and disadvantage in childhood, has been linked to the subsequent perpetration of lethal and nonlethal violence against a woman intimate partner (Ehrensaft et al. 2003; Gondolf 2002; Holtzworth-Munroe and Stuart 1994). Nonlethal and lethal violence against an intimate partner has also been associated with problems in adulthood including poor educational achievement, chronic unemployment, a history of arrest, and convictions for violence or other offenses, as well as problems in relationships with others, particularly intimates (Campbell et al. 2003; Dobash et al. 2000; Dutton and Hart 1992). Chronic substance abuse, particularly of alcohol, often features in the adult lives of both abusers and IPMurderers. It is not merely the consumption of alcohol but binge drinking and/or persistent alcohol abuse that are risk factors for violence, severe violence, and lethality (Fals-Stewart 2003; Finney 2004; Moracco et al. 1998; Walton-Moss et al. 2005). With respect to previous offending among those who commit IPMurder, studies have shown that a prior criminal record for any type of offense is a correlate of intimate partner murder (Dawson and Gartner 1998; Dobash and Dobash 2009; Moracco et al. 1998; Grann and Wedin 2002).

Personality

Focusing on personality, some studies have identified distinct characteristics of abusers, while others have found little difference in the characteristics of abusers and the wider population (Dutton and Kerry 1999; Gondolf 2002; Holtzworth-Munroe and Stuart 1994). Focusing only on men who have murdered an intimate partner, the results of several studies suggest that a considerable proportion of men who kill their partners are less likely to exhibit distinct personality problems than abusers and that a considerable proportion differ very little from the wider population (Dobash and Dobash 2009; Echeburua et al. 2003; Grann and Wedin 2002; Weizmann-Henelius et al. nd). While adverse experiences in childhood and problems in adulthood such as alcohol abuse may increase the risk of serious and more injurious forms of violence,

clearly they cannot be considered necessary or sufficient conditions either for lethal or nonlethal violence.

While evidence suggests the importance of adversity in the backgrounds of the perpetrator's of IPV and IPMurder, demonstrating that a considerable proportion of these men are similar to those who, for example, murder other males, there is a body of evidence indicating that some men who assault and murder their partners may not be characterized by such backgrounds. Several studies indicate that a reasonable proportion of men who murder their intimate partner have grown up in a stable family, experienced few problems in childhood, were steadily employed, and appeared to be reliable partners and parents (Dobash and Dobash 2009; Dobash et al. 2004; Echeburua et al. 2003; Grann and Wedin 2002; Weizmann-Henelius et al. nd). This evidence is in sharp contrast to the men who murder other men where difficult backgrounds and criminal records are the norm (Dobash et al. 2004). In a UK study of men who murdered their partner, 40 % appeared to have relatively "conventional" backgrounds (Dobash and Dobash 2009). Additional evidence collected after the murder, however, suggests that around one in five of these men had assaulted their partner at least once at some point in the relationship and their orientations to relationships with women, their violence, and the victims of the murder parallel those of men who have experienced adversity in their backgrounds. These men reacted angrily and violently to their partner's attempts to terminate the relationship and were reluctant to express remorse or regret regarding the murder (Dobash and Dobash 2011, nd).

Criminal Justice and Intimate Partner Violence

The CJ system has long been the focus of concern for community and pressure groups attempting to assist women victims of IPV (Dobash and Dobash 1992). Activists and advocates learned that the CJ system was at best reluctant and at worst opposed to intervening in "domestic disputes." This is confirmed by systematic evidence from several countries which revealed that the norm was under enforcement of the laws of assault and that assaults between strangers were more likely than those between family members to result in arrest (Dobash and Dobash 1979). Violence against women in the home was not considered to be a "real crime" and was, instead, treated as a social and personal problem best dealt with by social services and others. Training guidelines, such as those of the International Association of Police Chiefs, suggested that this was a "personal matter" and that arrest should be a "last resort." If the legal system was to be involved, it was through civil remedies such as injunctions and protection orders. This legacy of legal ideas, principles, and practices constituted the family as a private domain where the man was "in charge" and usually immune from prosecution. Even near the end of the twentieth century, this legacy was boldly reflected in homicide "special immunity" statutes in some US states whereby a man who murdered his wife when she was engaged in a sexual act of infidelity was immune from prosecution (Daly and Wilson 1988).

Beginning in the 1960s, the orientations within criminal justice in the US were merged with "psychiatric ideals" with attempts to provide interventions that would "facilitate human helping" in "family fights" and "domestic squabbles" (Dobash and Dobash 1992). Short-term, immediate interventions, such as "mediation," were endorsed. Model programs were created around the ideals of "crisis intervention," and by the 1980s "crisis intervention" and mediation were standard policy and procedure in large police departments throughout the USA (Oppenlander 1982; Lerman 1984). Arrest was a "last resort" to be invoked only in cases of "wanton" or "brutal" assault, and even into the 1980s half of US police departments prohibited arrest in all but the most injurious types of domestic violence (Sherman 1992). Such "model" programs exemplified the approach to domestic violence that was neither seen as a serious criminal problem nor as the responsibility of the criminal justice system. However, subsequent lawsuits and class actions against the police for failure to protect women

victims of violence (almost exclusively in the USA) and legislative and administrative mandates at national and local levels brought about changes in such policies.

Police officers in many countries are now given specialized training, often provided by the same pressure groups that lobbied for change. In some cases, protocols mandating enhanced police action are now in place with prescriptive instructions limiting police discretion and requiring actions to protect victims of violence and their children. In several jurisdictions, presumptive arrest of the perpetrator based on "probable cause" is now seen as an appropriate action, and arrest is mandated with such practices sometimes enacted by dedicated domestic violence police units (Iovanni and Miller 2001).

Criminal justice developments have occurred throughout the world. In the USA, the UK, and many other countries, pressure groups have contributed to changes in thinking and practice, particularly concerning the treatment of victims who had previously been deemed outside of the remit of police work. A comprehensive study of European Union countries found significant changes in criminal justice responses (European Commission 2010), and the authors concluded that the concept of "due diligence" to "protect, prosecute, and prevent" is widely accepted as is criminalization of IPV with mediation and conciliation no longer acceptable in most countries. In most, but not all, EU countries, policies and laws associated with domestic violence are usually gender specific, and it is generally acknowledged that women are the usual victims of intimate partner violence.

In some countries there have also been significant developments in the prosecution services, courts, and sentencing. Most importantly, intimate partner violence is now on the agenda and considered an important problem worthy of criminal justice consideration and intervention through prosecution and criminal courts. In some countries dedicated domestic violence courts have been created to facilitate the processing of such cases. New sentencing options have also been developed that mandate offender attendance to innovative feminist-oriented cognitive behavioral interventions that require offenders to deal with and acknowledge their violent acts and take responsibility for them (Dobash et al. 2000; Gondolf 2002). Such programs are an important part of the criminal justice response to intimate partner violence.

It is prudent to ask if such changes have made any difference. Has intimate partner violence been reduced? Are women and their children safer? Are offenders taking responsibility for their actions and changing their behavior? Systematic evaluations of innovations are rare, most have been conducted in the USA and most have focused on the effectiveness of arrest. In the 1980s, a number of studies in the USA attempted to compare the impact of arrest with doing nothing or responding in other ways such as mediation. Initially, arrest appeared to reduce repeat incidents of violence over a 12-month period, and subsequently a pro-arrest strategy was adopted in some jurisdictions (Sherman and Berk 1984). However, subsequent studies conducted in other cities reached conflicting conclusions ranging from studies finding that arrest had "no effect on repeat violence" to those showing a "reduced effect" and to those showing an "increase" in violence. Despite these contradictory findings, policies supporting arrest are now in place in the USA and elsewhere and are deemed to be a useful intervention, although practice has not always followed on from such policies (Ferraro 1989; Jaffe et al. 1993; Zorza and Woods 1994; Logan et al. 2006).

Summary and Conclusions

Violence against women is now recognized as a worldwide problem. While the prevalence of this violence may vary from country to country, it is clear that the problem exists in all those societies where it has been investigated. Violence against women is now recognized as an issue of human rights by the United Nations, UNESCO, the European Union, and other national and international bodies. Where once there was indifference to, or even outright support of, violence by men against women partners, this has now been drastically eroded although not eliminated.

Societal values and institutional responses have changed. It is clear that the efforts of women's groups throughout the world have made a difference in raising the issue and working toward more effective responses to women who have been the victims of intimate partner violence. In the USA, research indicates that there has been a significant reduction in intimate partner violence in those cities that have introduced meaningful interventions for IPV. In addition, national figures show a significant reduction in intimate partner murder.

Recovering from a shaky start, researchers within criminology have, along with those from other disciplines, played an important part in placing this problem on the academic and public agendas, in helping to transform ways of thinking about this problem, and in replacing uninformed speculation with solid evidence. Evidence reveals that violence against women by intimate partners is widespread and serious, but also that it is amenable to positive interventions from the justice system and others that work toward reductions in its frequency and severity and thus toward the well-being of women, their children, and, ultimately, the society at large.

Related Entries

► Surveys on Violence Against Women

Recommended Reading and References

Abbott J, Johnson R, Koziol-McLain J, Lowenstein SR (1995) Domestic violence against women: incidence and prevalence in an emergency department population. J Am Med Assoc 273:1763–1767

Amir M (1971) Patterns in forcible rape. Chicago University Press, Chicago

Archer J (2000) Sex differences in aggression between heterosexual partners: a meta-analytic review. Psychol Bull 126:651–680

Backman R, Saltzman LE (1995) Violence against women: estimates from the redesigned survey. Special report of the bureau of justice statistics. U.S. Dept. Justice, Washington D.C

Block CR, Christakos A (1995) Intimate partner homicide in Chicago over 29 years. Crime Delinq 41:496–526

Browne A (1987) When battered women kill. The Free Press, New York, Collier Macmillan

Browne A, Williams KR, Dutton DG (1999) Homicide between intimate partners: a 20-year review. In: Smith MD, Zahn MA (eds) Homicide: a sourcebook of social research. Sage, Thousand Oaks, pp 149–164

Brownridge DA, Halli SS (2002) Understanding male partner violence against cohabiting and married women: an empirical investigation with a synthesized model. J Fam Violence 17:341–361

Campbell JC (1998) Abuse during pregnancy: progress, policy, and potential. Am J Public Health 88:185–187

Campbell JC (1999) Forced sex and intimate partner violence. Violence Against Women 5:1017–1035

Campbell JC, Webster D, Koziol-McLain J, Block CR, Campbell D, Curry MA et al (2003) Risk factors for femicide in abusive relationships: results from a multisite case control study. Am J Public Health 93: 139–152

Campbell JC, Glass N, Sharps PW, Laughon K, Bloom T (2007) Intimate partner homicide: review and implications of research and policy. Trauma Violence Abuse 8:246–269

Cavanagh K, Dobash RE, Dobash RP (2007) The murder of children by fathers in the context of child abuse. Child Abuse Negl 31:731–746

Coker AL, Smith PH, McKeown RE, King MJ (2000) Frequency and correlates of intimate partner violence by type: physical, sexual, and psychological battering. Am J Public Health 90:553–559

Dahlberg LL, Krug EG (2002) Violence – a global health problem. In: Krug EG, Dahlberg LL, Mercy JA, Zwi AB, Lozano R (eds) World report on violence and health. WHO, World Health Organisation, Geneva, pp 1–21

Daly M, Wilson M (1988) Homicide. Aldine De Gruyter, New York

Dawson R, Gartner R (1998) Differences in the characteristics of intimate femicides: the role of relationship state and relationship status. Homicide Stud 2:378–399

Dobash RE, Dobash RP (1977/78) Wives: the 'appropriate' victims of marital violence. Victimology 2:426–442

Dobash RE, Dobash RP (1979) Violence against wives: a case against the patriarchy. Free Press & Open Books, New York

Dobash RE, Dobash RP (1984) The nature and antecedents of violent events. Br J Criminol 24:269–288

Dobash RE, Dobash RP (1992) Women, violence and social change. Routledge, London

Dobash RE, Dobash RP (eds) (1998) Rethinking violence against women. Sage, Thousand Oaks

Dobash RP, Dobash RE (2004) Women's violence to men in intimate relationships: working on a puzzle. Br J Criminol 44:324–349

Dobash RE, Dobash RP (2009) Out of the blue: men who murder an intimate partner. Fem Criminol 4:194–225

Dobash RE, Dobash RP (2011) What were they thinking? Men who murder an intimate partner. Violence Against Women 17:111–134
Dobash RE, Dobash RP (2012) Who died? The murder of collaterals related to intimate conflict. Violence Against Women 18:662–671, Jan
Dobash RP, Dobash RE, Wilson M, Daly M (1992) The myth of sexual symmetry in marital violence. Soc Probl 39(1):71–91
Dobash RP, Dobash RE, Cavanagh K, Lewis R (1998) Separate and intersecting realities: a comparison of men's and women's accounts of violence against women. Violence Against Women 4:383–414
Dobash RP, Dobash RE, Cavanagh K, Lewis R (2000) Changing violent men. Sage, Thousand Oaks
Dobash RP, Dobash RE, Cavanagh K, Lewis R (2004) Not an ordinary killer-just an ordinary guy: when men murder an intimate woman partner. Violence Against Women 10:577–605
Dobash RE, Dobash RP, Cavanagh K, Medina-Ariza JJ (2007) Lethal and non-lethal violence against an intimate partner: comparing male murderers with non-lethal abusers. Violence Against Women Int Interdisc J 13(4):1–27
Dobash RE, Dobash RP (nd) When men murder women. Oxford University Press, New York (forthcoming)
Dutton D, Hart S (1992) Risk markers for family violence in a federally incarcerated population. Int J Law Psychiatry 15:101–112
Dutton D, Kerry G (1999) Modus operandi and personality disorder in incarcerated spousal killers. Int J Law Psychiatry 22:287–299
Echeburua E, Fernandez-Montalvo J, Amor PJ (2003) Psychopathological profile of men convicted of gender violence. A study in the prisons of Spain. J Interpers Violence 18:798–812
Ehrensaft MK, Cohen P, Brown J, Smailes E, Chen H, Johnson JG (2003) Intergenerational transmission of partner violence: a 20 year prospective study. J Consult Clin Psychol 71:741–753
European Commission (2010) Standardise national legislation on violence against women, violence against children & sexual orientation violence: European Commission Directorate for Justice. Daphne, Brussels
Fals-Stewart W (2003) The occurrence of partner physical aggression on days of alcohol consumption: a longitudinal study. J Consult Clin Psychol 71:41–52
Ferraro K (1989) Policing women battering. Soc Probl 36:61–74
Finney A (2004) Alcohol and intimate partner violence: key findings from the research. Home office bulletin 216. Research, Development and Statistics Directorate, London
Gaquin DA (1977/78) Spouse abuse: data from the national crime survey. Victimology 2:632–643
Garcia-Moreno C, Jensen H, Ellsberg M, Heise L, Watt C (2006) Prevalence of intimate partner violence. Findings from the World Health Organisation multi-country study on women's health and domestic violence. Lancet 368:1260–1269
Gondolf EW (2002) Batterer intervention systems: issues, outcomes and recommendations. Sage, Thousand Oaks
Grann M, Wedin M (2002) Risk factors for recidivism among spousal assault and spousal homicide offenders. Psychol Crime Law 8:5–23
Greenfield L, Rand MR, Craven D, Klaus PA, Perkins CA, Ringel C, Warchol G, Maston C, Fox JA (1998) Violence by intimates: analysis of data on crimes by current or former spouses, boyfriends and girlfriends. US Dept. of Justice, Washington, DC
Holtzworth-Munroe A, Stuart GL (1994) Typologies of male batterers: three subtypes and the differences among them. Psychol Bull 11:476–497
Iovanni L, Miller SL (2001) Criminal justice system responses to domestic violence: law enforcement and the courts. In: Renzetti CM, Edelson JL, Bergen RK (eds) Sourcebook on violence against women. Sage, Thousand Oaks, pp 303–327
Jaffe PG, Hastings E, Reitzel D, Austin GW (1993) The impact of police laying charges. In: Hilton NZ (ed) Legal responses to wife assault: current trends and evaluation. Sage, Newbury Park
Johnson H (1996) Dangerous domains: violence against women in Canada. Nelson, Scarborough
Johnson H, Hotton T (2003) Losing control: homicide risk in estranged and intact intimate relationships. Homicide Stud 7:58–84
Johnson H, Sacco VF (1995) Researching violence against women: statistics Canada's national survey. Can J Criminol 37:281–304
Krahe B, Bieneck S, Moller I (2005) Understanding gender and intimate partner violence from an international perspective. Sex Roles 52:807–827
Krug EG, Dahlberg LL, Mercy JA, Zwi AB, Lozano R (2002) World report on violence and health. World Health Organisation, Geneva
Kyriacou D, Anglin D, Taliaferro E, Stone S, Tubb T, Linden J, Muelleman R, Barton E, Kraus JF (1999) Risk factors for injury to women from domestic violence. N Engl J Med 341:1892–1898
Langford L, Isaac N, Kabat S (1998) Homicides related to intimate partner violence in Massachusetts: examining case ascertainment and validity of the shr. Homicide Stud 2:353–377
Lerman LG (1984) Mediation of wife abuse cases: the adverse impact of informal dispute resolution on women. Harv Women's Law J 7:57–113
Liem MCA (2010) Homicide followed by suicide: a review. Aggress Violent Behav 15:247–261
Logan TK, Shannon L, Walker R (2006) Police attitudes toward domestic violence. J Interpers Violence 21:1365–1374
Luckenbill DF (1977) Criminal homicide as a situated transaction. Soc Probl 25:176–186
Merton RK (1938) Social structure and anomie. Am Sociol Rev 3:672–682

Miethe TD, Regoeczi WC (2004) Rethinking homicide. Exploring the structure and process underlying deadly situations. Cambridge University Press, New York
Mirrlees-Black C (1999) Domestic violence: findings from a new British crime survey self completion questionnaire. Home Office, London
Moffit TE, Caspi A, Silva PA (1998) Developmental antecedents of partner abuse: a prospective study. J Abnorm Psychol 107:375–389
Moracco KE, Runyan CW, Butts JD (1998) Femicide in North Carolina, 1991–1993: a statewide study of patterns and precursors. Homicide Stud 2:422–446
Nazroo J (1995) Uncovering gender differences in the use of marital violence: the effect of methodology. Sociology 29(3):475–494
Oppenlander N (1982) Coping or copping out? Criminology 20:449–465
Polk K, Ranson D (1991) The role of gender in intimate homicide. Aust N Z J Criminol 24:15–24
Randall M, Haskall L (1995) Sexual violence in women's lives. Violence Against Women 1:6–31
Russell DEH (1990) Rape in marriage. Indiana University Press, Indianapolis
Schumacher JA, Feldbau-Kohn S, Smith Slep AM, Heyman RE (2001) Risk factors for male-to-female partner physical abuse. Aggress Violent Behav 6:281–352
Schwartz MD (1987) Gender and injury in marital assault. Sociol Focus 20:61–75
Serran G, Firestone P (2004) Intimate partner homicide: a review of male proprietariness and the self-defense theories. Aggress Violent Behav 9:1–15
Shackleford TK, Mouzos J (2005) Partner killing by men in cohabiting and marital relationships: a comparative, cross-national analysis of data from Australian and the United States. J Interpers Violence 20:1310–1324
Sherman LW (1992) Policing domestic violence: experiments and dilemmas. Free Press, New York
Sherman LW, Berk RA (1984) The specific deterrent effects of arrest for domestic violence. Am Sociol Rev 49:261–272
Short JF, Strodtbeck FL (1965) Group process and gang delinquency. University of Chicago Press, Chicago
Stark E, Flitcraft A (1992) Spouse abuse. In: Last RJ, Wallace R (eds) Public health and preventive medicine, 13th edn. Appleton & Lange, Norwalk
Tjaden P, Thoennes N (1998) Prevalence, incidence, and consequences of violence against women: findings from the national violence against women survey. Research in brief (Nat. Instit. Just. & Centre for Disease Control/Prevention). US Dept. of Justice, Washington, DC
Ullman SE, Siegel JM (1993) Victim-offender relationship and sexual assault. Violence Vict 8:121–134
US Dept. Justice (1998) Stalking and domestic violence. The third annual report to congress under the violence against women act. US Dept. Justice, Washington, DC
US Dept. Justice-Office of Justice Programs (1998) Bureau of justice statistics factbook: violence by intimates-analysis of data on crimes by current or former spouses, boyfriends, and girlfriends. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC
Walby S, Allen J (2004) Domestic violence, sexual assault and stalking: findings form the British crime survey. Home office research study 276. Home Office, Research, Development and Statistics Directorate, London
Walton-Moss BJ, Manganello J, Frye V, Campbell JC (2005) Risk factors for intimate partner violence and associated injury among urban women. J Commun Health 30:377–389
Watson D, Parsons S (2005) Domestic abuse of women and men in Ireland: report of the national study of domestic abuse. Nat Crime Council and Economic and Social Research Institute, Dublin, pp 1–224
Websdale N (2010) Familicidal hearts: the emotional styles of 211 killers. Oxford University Press, New York
Weizmann-Henelius G, Gronroos M, Putkonen H, Eronen M, Lindberg N et al (nd) Gender specific risk factors for intimate partner homicide—a nationwide register-based study. J Interpers Violence (forthcoming)
Wilson M, Daly M (1993) Spousal homicide risk and estrangement. Violence Vict 8:3–16
Wilson M, Daly M (1998) Lethal and nonlethal violence against wives and the evolutionary psychology of male sexual proprietariness. In: Dobash RE, Dobash RP (eds) Rethinking violence against women. Sage, Thousand Oaks, pp 199–230
Wilson M, Johnson H, Daly M (1995) Lethal and nonlethal violence against wives. Can J Criminol 37:331–362
Wolfgang M (1958) Patterns in criminal homicide. University of Pennsylvania Press, Philadelphia
Worrall A, Pease K (1986) Personal crime against women: evidence from the 1982 British crime survey. Howard J 25:118–124
Zorza J, Woods L (1994) Mandatory arrest: problems and possibilities. National Center on Women and Family Law, New York

Double-Blind Administration

▶ Eyewitness Research

Dowry Deaths

▶ Lawful Killings

Drug Abuse and Alcohol Dependence Among Inmates

Javier A. Cepeda[1] and Sandra A. Springer[2]
[1]Center for Interdisciplinary Research on AIDS and Department of Epidemiology and Public Health, Yale School of Public Health, Yale University, New Haven, CT, USA
[2]Center for Interdisciplinary Research on AIDS and Yale School of Medicine, Yale University, New Haven, CT, USA

Synonyms

AIDS; Alcohol dependence; Buprenorphine; CJS; Criminal justice system; Cocaine; Extended-release Naltrexone; Hepatitis C; HIV; Human Immunodeficiency Syndrome; Incarceration; Jail; MAT; Medication assisted therapy; Mental illness; Methadone; Methamphetamine; Naltrexone; Opioid dependence; Prison; Reentry; Release; STDs; TB; Tuberculosis; Vivitrol

Overview

The massive increase in the number of individuals within the United States (USA) criminal justice system (CJS) in the past 30 years can be largely attributed to the "War on Drugs" campaign with drug-related arrests increasing more than fivefold from 1970 to 2005. Many of the incarcerated individuals have a history of substance use disorders (SUDs), and have reported a use of drugs within the past month prior to arrest. According to the US Bureau of Justice Statistics (BJS), in 2002, the prevalence of offenders who used drugs more than once a week for at least 1 month prior to incarceration exceeded 60 % (Karberg and James 2005) and approximately 50 % of those incarcerated at the federal, state, and local levels met DSM-IV criteria for drug abuse or dependence.

The large burden of drug and alcohol dependence within the US CJS complicates treatment programs since drug and alcohol dependency is highly associated with comorbid psychiatric disorders and chronic infectious diseases. Effective pharmacotherapy for opioid and alcohol dependence exists in the community; however, access to treatment in the US CJS remains inadequate. Less than 10 % of local jail inmates received treatment (Karberg and James 2005).

The public health implications of SUDs treatment for offenders are numerous. Incarceration provides an opportunity to initiate or continue treatment for drug and alcohol dependency among inmates, whereupon both the individual and the community can be greatly benefited upon release. In addition to treatment, strong linkages should be established to ensure continuity of care. Failure to link services for offenders upon release to the community may increase recidivism, relapse to substance abuse, and increase transmission of sexually transmitted diseases (STDs) including human immunodeficiency virus (HIV), and viral hepatitis B and C to the uninfected.

This entry will concentrate mostly on the epidemiology, treatment, and public health implications of opioid and alcohol dependence among inmates in the US CJS. While abuse of others drugs, such as cocaine and methamphetamine, is highly prevalent among offenders, much of the clinical and epidemiological literature on incarcerated persons has been limited to research involving dependency on alcohol and opioids. Effective medically assisted therapy (MAT) for cocaine, crack, and methamphetamine dependence is still lacking, and to date, only behavioral interventions have been implemented to treat dependence for these substances. Conversely, there are FDA-approved pharmacotherapies and behavioral interventions for opioid- and alcohol-dependent persons within the community, and some of these have been successfully implemented in some incarcerated settings across the USA. This entry will discuss the strengths and limitations of various evidence-based practices of SUDs treatment administered in incarcerated settings.

Definition of Substance Dependency

The Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) classifies substance

Drug Abuse and Alcohol Dependence Among Inmates, Table 1 DSM-IV criteria for substance dependence and substance abuse. In using the DSM-IV criteria, one should specify whether substance dependence is with physiologic dependence (i.e., there is evidence of tolerance or withdrawal) or without physiologic dependence (i.e., no evidence of tolerance or withdrawal)

Dependence (3 or more in a 12-month period)	Abuse (1 or more in a 12-month period) Symptoms must never have met criteria for substance dependence for this class of substance
Tolerance (marked increase in amount; marked decrease in effect)	Recurrent use resulting in failure to fulfill major role obligation at work, home, or school
Characteristic withdrawal symptoms; substance taken to relieve withdrawal	Recurrent use in physically hazardous situations
Substance taken in larger amount and for longer period than intended	Recurrent substance related legal problems
Persistent desire or repeated unsuccessful attempt to quit	Continued use despite persistent or recurrent social or interpersonal problems caused or exacerbated by substance
Much time/activity to obtain, use, recover	
Important social, occupational, or recreational activities given up or reduced	
Use continues despite knowledge of adverse consequences (e.g., failure to fulfill role obligation, use when physically hazardous)	

dependency as displaying three out of the seven of the listed criteria in Table 1 in a span of 12 months. Substance abuse, as defined by DSM-IV, is demonstration of at least one of the criteria in Table 1 in a span of 12 months.

Epidemiology and Socio-demographic Profile of Drug and Alcohol Abuse and Dependence in the Criminal Justice System

Prevalence of drug abuse among offenders in the US CJS remains high at the county, state, and federal level. In 2005 alone, there were 1.65 million drug arrests. Among all US county jails, nearly 60 % of inmates reported a history of drug dependency at some point in their life and about half of all US inmates met DSM-IV criteria for drug dependency (Mumola and Karberg 2006).

The majority of those incarcerated are men; however, the incarceration rate for women has been increasing since the early 1990s. From 1990 to 1998, the number of women behind bars jumped 71 %. This escalation of incarcerated women was mostly attributed to a surge in the arrest of female drug users. In 2002, about 61 % of incarcerated women satisfied DSM-IV criteria for drug dependency, versus 54 % of incarcerated men in local jails (Karberg and James 2005).

Minorities, especially blacks and Latinos, are disproportionately represented in the criminal justice system. While only making up about a quarter of the US population, more than 60 % of the US criminal justice population is black or Latino. Conversely, white inmates have been reported as having a relatively higher prevalence of substance (alcohol and/or drug) abuse or dependence (78 %) compared to blacks (64 %) and Hispanics (59 %) (Karberg and James 2005).

Opioid Abuse and Dependence

Abuse and dependence to opioids continues to plague the USA. There are approximately 900,000 opioid-dependent persons within the USA. In 2004, there were approximately 1.2 million state prisoners, of whom 23.4 % had ever used heroin or opiates in their lifetimes, and 8.2 % of convicted inmates reported using heroin or other opiates 1 month prior to arrest (Mumola and Karberg 2006). Approximately 11 % of all male inmates, and 20 % of all female inmates reported using opioids daily in the 6 months prior to arrest. Since the 2000s, nonmedical abuse of prescription opioids has led to an upsurge in number of arrests. A nationwide survey from years 2002 to 2004 reported that 30 % of arrestees had used prescription opioids for nonmedical purposes.

Stimulant Abuse and Dependence

The prevalence of abuse of stimulants (cocaine/crack and methamphetamines) among inmates in state and federal prisons exceeds that of marijuana, opioids, hallucinogens, and other illicit depressants combined. In 2004, approximately one half of all federal and state prisoners with a history of drug use reported using cocaine and/or crack (Mumola and Karberg 2006). Additionally, methamphetamine (METH) abuse is a growing problem in the USA. From 1997 to 2004, use of METH was the only drug that increased across all measures collected in a nationwide survey (Mumola and Karberg 2006). There are stark differences in METH use by race/ethnicity. At the federal level, the prevalence of METH abuse among white, Hispanic, and blacks is 29 %, 5 %, and 1 %, respectively (Mumola and Karberg 2006). METH use also differs by gender. Among state inmates, 17 % of women used METH in the month prior to arrest, as opposed to 10 % of men (Mumola and Karberg 2006). Furthermore, a nationwide survey of a sample of local police officers revealed that METH was overwhelmingly their drug of concern compared to cocaine, marijuana, and heroin (48 % vs. 22 %, 22 %, and 3 % respectively). Unlike treatment for opioids and alcohol, there is no effective FDA-approved pharmacotherapeutic treatment yet for cocaine and METH dependency. This is particularly problematic, given the increasing abuse of methamphetamines and high prevalence of cocaine abuse among inmates in state and federal prisons.

Alcohol Abuse and Dependence

Alcohol use disorders (AUDs) encompass a wide variety of conditions including abuse, hazardous drinking, binge drinking, and dependence. Approximately 4 % of the US population meets criteria for alcohol dependence (Grant et al. 2004), while the prevalence of alcohol dependence in state jails, however, is roughly tenfold higher than in the general population (Mumola and Karberg 2006). In 2002, nearly 50 % of all inmates met criteria for alcohol abuse or dependence (Mumola and Karberg 2006). The prevalence of alcohol abuse or dependence among female inmates was 39 %, while half of all men met the criteria for dependence or abuse (Mumola and Karberg 2006). Whites were relatively more likely to abuse alcohol or be dependent (58.6 %), compared to blacks (42.7 %) and Hispanic inmates (41.8 %).

Management and Treatment of Opioid Dependence Within the Criminal Justice System

Despite the proven effectiveness of MAT to treat substance dependence in the community, long-term implementation is still rare across US incarcerated settings. The two most commonly administered pharmacotherapeutic options to manage and treat opioid dependence are opioid substitution therapy (OST) in the form of methadone and buprenorphine. The other options are antagonists at the mu-receptor in two formulations: once daily oral naltrexone (ReVia®), or the once monthly extended-release depot formulation of naltrexone (Vivitrol). See Table 2 for a summary of available MATs for the treatment of opioid dependence.

Methadone

Methadone is a full opioid mu-receptor agonist that produces similar effects to morphine and heroin. In terms of public health benefits, methadone has been associated with a reduction in risky injection practices (thereby reducing transmission of blood-borne viruses), improvement in overall health, and reduction in mortality. Methadone maintenance therapy (MMT) is also regarded as extremely cost-effective, with more than 50 % of the benefits affecting individuals who do not use drugs. In spite of the evidence demonstrating the advantages of MMT, it still remains difficult to obtain for the majority of heroin users. Access to methadone still remains a challenge for the majority of opioid-dependent persons in the USA for several reasons including long waiting lists and because some state Medicaid programs do not cover the cost of MMT.

Drug Abuse and Alcohol Dependence Among Inmates, Table 2 Medically assisted therapies for treatment of opioid or alcohol dependence

MAT	Mechanism	Form and frequency of administration	Advantages	Disadvantages
Methadone	Full opioid agonist	Liquid or tablet, daily	Inexpensive, no need to experience opioid withdrawal symptoms prior to administration, prevents relapse	Strict federal regulations, accessibility, diversion to illicit drug markets, tolerance, side effects
Buprenorphine	Partial opioid agonist, partial opioid antagonist	Sublingual tablet, film, daily or alternate day	Lower risk of diversion, administered in office setting (less stigma), prevents relapse	More expensive than methadone, patient must experience minimal withdrawal before administration, requires 8-h training to obtain ability to prescribe
Extended-release naltrexone	Full opioid antagonist	Once monthly intramuscular injection	Adherence advantage, no risk of dependence, no risk of overdose, FDA-approved for opioid and alcohol dependence treatment, reduces cravings, no special licensing or special training	Injection, expensive, 5–7 days of opioid abstinence needed prior to treatment for opioid dependence
Naltrexone	Full opioid antagonist	Tablet, daily or alternate day	No risk of dependence or overdose, reduces cravings, very effective for treatment of alcohol dependence	5–7 days of opioid abstinence needed, not found to be effective for preventing relapse for opioid dependence due to low adherence to oral form
Disulfiram	Inhibits acetaldehyde dehydrogenases involved in metabolizing alcohol	Tablet, daily	No risk of tolerance, possible use in treatment of cocaine dependence	Questionable efficacy, low compliance
Acamprosate	Blocks N-methyl D-aspartate receptors	Tablet, 2 tablets three times daily	Reduces relapse, can be used with opioid agonists for opioid dependence	Poor adherence, may only be effective with support groups

DSM-IV criteria for sub- stance dependence and substance abuse. In using the DSM-IV criteria, one should specify whether substance dependence is with physiologic dependence (i.e., there is evidence of tolerance or withdrawal) or without physio- logic dependence (i.e., no evidence of tolerance or withdrawal)

Authorities in the CJS still have strong reservations about offering methadone in prisons or jails. The predominating perception among criminal justice officials was that methadone itself "is replacing one addiction with another" (McMillan and Lapham 2005). Other reasons include concerns regarding cost, diversion, and overdose. Only a select few of incarcerated settings have set up the infrastructure necessary to implement a methadone maintenance treatment (MMT) program. Several research studies were conducted to investigate the effectiveness of the MMT program at Rikers Island (New York City) and found a decrease in risky injection practices 6 months post-release. Moreover, the implementation of a directly observed therapy approach of administration of methadone in the jail greatly alleviated correctional officials' concern about possible diversion of methadone for illicit purposes (Tomasino et al. 2001). Furthermore, an RCT comparing (1) methadone *after* release, (2) methadone started *in* prison and transferred after release, and (3) counseling alone among opioid-dependent prisoners in Baltimore found

that individuals who received only counseling were approximately seven times more likely to relapse to opioids and cocaine 12 months post-release to the community, when compared to individuals who received both counseling and had initiated methadone in prison (Kinlock et al. 2009). Overall, both of these studies lend evidence that initiation of MMT is not only feasible within correctional institutions in the USA, but also that MMT has broad implications in reducing risky and illicit drug-using behavior when prisoners return to the community.

Buprenorphine

In 2002, the FDA approved ***buprenorphine (BPN)***, *a partial mu-receptor agonist and antagonist* for the treatment of opioid dependence. Providers need to take an 8-h course on BPN induction and maintenance practices to receive the necessary additional DEA license authorization to prescribe to patients. Since BPN can be administered in an office setting, there is less stigma associated with it compared to receiving daily doses of methadone at a federally sanctioned clinic. The dose response profile of BPN plateaus at a 32-mg dose (no additional effect for doses higher than 32 mg). Depending on the patient, BPN can be administered every 2–3 days, allowing for a more flexible maintenance regimen.

Due to its partial opioid agonist activity at the mu-receptor, BPN generally stimulates milder euphoric effects compared to methadone. BPN is most often administered sublingually, and the variant that includes naloxone (commercially known as Suboxone®) is thought to prevent misuse via injection of BPN. Recently, BPN has been introduced as a sublingual film that is more efficient in absorption than the tablet form. Furthermore, the chances of diversion are thought to be reduced due to a 10-digit bar code on the packaging that can be traced back to the patient should it be found in the illegal drug market and as a means to avoid accidental ingestion among minors.

Due to its relatively new arrival, research investigating the effectiveness of BPN in prison settings is still limited. Investigators conducted a RCT comparing BPN maintenance to MMT. The authors found that patients on BPN were more likely to attend community treatment centers compared to those randomized to receive methadone after release. Ninety-three (93) percent of the patients on BPN said that they intended to continue treatment after being released, while only 44 % of those on methadone expressed the same sentiment. While in jail, patients on BPN were significantly less likely to voluntarily drop out of receiving treatment than those on methadone. Despite these findings demonstrating more acceptability of BPN compared to methadone while the patients were incarcerated, there were no significant differences in terms of self-reported relapse to opioid use, rearrests, or reincarceration between the two groups. BPN was shown to reduce recidivism among inmates who initiated BPN while incarcerated compared to those who received no form of OST (Levasseur et al. 2002).

Implementation of a BPN maintenance program in a Baltimore prison highlighted some challenges distinct from MMT (Kinlock et al. 2010). The authors highlight diversion of BPN as the most significant challenge compared to methadone. Nurses had to wait 10 min per patient receiving a sublingual dose of BPN to ensure that the tablet had been completely dissolved. Methadone, on the other hand, is taken in liquid form, making diversion more difficult if the inmate is being directly observed. These findings are consistent with those from the Rikers Island study, with the authors reporting that approximately 10 % of the patients on BPN attempted to divert the drug, while only 1 % of the patients on methadone attempted diversion. Fears of diversion of BPN could be potentially mitigated if the sublingual film is used rather than the tablet, due to the former dissolving more rapidly than the latter. The FDA in 2012 ordered that tablet formulation of BPN would be replaced with the sublingual film in order to mitigate concerns for diversion and decrease deaths associated with inadvertent use of the sublingual tables by children.

Oral and Extended-Release Naltrexone

Buprenorphine and methadone are classified as types of opioid substitution therapy due to their affinity for the mu-receptor; however,

naltrexone, a full opioid antagonist, completely blocks the mu-receptor. Naltrexone can be administered orally daily, or once a month by intramuscular injection (extended-release naltrexone). The oral form, ReVia®, approved in 1984 for the treatment of opioid dependence, has NOT been shown to be effective in relapse prevention or in recidivism due to compliance issues mainly. In October of 2010, the FDA approved the use of extended-release naltrexone (XR-NTX) for the treatment of opioid dependence after results from a double-blind multicenter RCT performed in Russia. In this study, 126 opioid-dependent individuals were randomized to receive the XR-NTX compared to 124 subjects who received placebo. The XR-NTX group was 1.58 times more likely to be abstinent from opioids at end of the 24-week study period compared to the control group. Of note, 40 % of the subjects who received XR-NTX had HIV disease, and approximately 90 % had Hepatitis C infection without any serious liver function abnormalities. The extended-release version of naltrexone does provide a possibility of less adherence concerns, given its once monthly dosing as well as possible fewer side effects compared to the oral form of naltrexone as well as methadone and BP-N. Furthermore, because XR-NTX is an antagonist at the mu opioid receptor, there are no concerns for adverse events such as respiratory depression or overdose from the treatment as can be seen with methadone. Also due to its antagonist properties, there is not a concern for diversion as has been seen with buprenorphine. Therefore, XR-NTX may be a viable option for the CJS where such concerns predominate.

Before initiating ***naltrexone*** maintenance therapy, the patient must be free of opioids within the last 5–7 days. This is perhaps the most challenging aspect to naltrexone initiation, since many patients are fearful of experiencing any withdrawal symptoms and can often not resist the urge to continue using opioids in this period. While sudden abruption of not taking naltrexone does not produce withdrawal symptoms, the patient must be on oral naltrexone for at least 30 days to demonstrate efficacy in treatment outcomes with this form of naltrexone; however, it is unclear how long one should be maintained on XR-NTX at this time to demonstrate efficacy.

Thus far, only a few studies have examined the feasibility and effectiveness of oral or XR-NTX maintenance therapy among incarcerated populations. In one double-blind RCT, investigators compared oral naltrexone to another opioid antagonist called cyclazocine among 40 inmates with prior history of opioid addiction (Brahen et al. 1977). At the end of the study, inmates who were on naltrexone reported fewer side effects after induction with a placebo and very low toxicity. In fact, three patients on cyclazocine dropped out of the study due to complications arising from the treatment itself. None of the patients randomized to naltrexone dropped out of the study for these reasons (Brahen et al. 1977). Indeed, a recent meta-analysis reported that the evidence claiming oral naltrexone as a superior treatment to other therapies for preventing relapse is conflicting (Minozzi et al. 2011).

There has been only one publication, thus far, on the effectiveness of XR-NTX among incarcerated populations. In this Norwegian study, 46 heroin-addicted inmates were randomized to either XR-NTX or methadone (Lobmaier et al. 2010). At 6 months post-release, the investigators found similar reductions in frequency of heroin use and criminal behavior between the two groups, suggesting that XR-NTX may be as equally effective as methadone at treatment of opioid relapse and recidivism. RCTs evaluating the effect of XR-NTX among opioid-dependent incarcerated populations are ongoing at the present time.

Behavioral Interventions for Offenders with Drug Dependence

Oftentimes pharmacotherapy is integrated with psychosocial support, either in the form of SUDs 12-step counseling, cognitive behavioral therapy, case management, or enrollment in therapeutic communities. Prison-based SUDs treatment programs have been shown to reduce recidivism and drug use. Specifically,

cognitive behavioral therapy (CBT) interventions for offenders focusing on reducing criminal "thinking" can lead to criminal behavior. The objectives of CBT are to prevent relapse, to recognize situations that can often lead to drug use or criminal activities, to assist patients on how to effectively deal with these situations, to enhance social support networks, and to advance feelings of self-efficacy.

Contingency management (CM) is another type of behavioral intervention employed among drug abusing offenders that aims to promote abstinence through positive reinforcement. The theory supporting CM as an effective strategy is that rewarding positive behavior will eventually replace punishable behavior. Findings from two meta-analyses lend evidence to this theory (Griffith et al. 2000; Lussier et al. 2006). Modeled after community-based therapeutic communities, *in-prison therapeutic communities (TCs)* are a type of intervention that move drug-using offenders into separate facilities from the rest of the incarcerated population. This kind of residential treatment allows for intensive interventional and psychosocial support. The two TC modalities focus on the 12-step self-help model and the relapse prevention model. An evaluation of the Amity prison TC revealed statistically significant better outcomes in terms of reincarceration among inmates enrolled in TC versus inmates who dropped out of the program versus inmates randomized to a control group. Unlike OST, TCs have gained wide acceptance in the prison system. Since 1994, many states have adopted the TC model and received federal funding to expand residential substance abuse treatment (RSAT) programs in prisons and jails.

Management and Treatment of Alcohol Dependence Within the Criminal Justice System

The prevalence of alcohol use disorders (AUDs) among incarcerated individuals is staggering, as previously discussed. Behavioral interventions have typically been viewed as the bastion for alcohol abuse treatment; however, these interventions have only been marginally effective in treating alcohol-related disorders (Springer et al. 2011). Specifically, behavioral treatment can include motivational enhancement therapy (MET), cognitive behavioral therapy, and self-help (e.g., 12-step) programs. MET, although less structured than CBT, has been shown to reduce cravings of alcohol with integrated pharmacotherapy using naltrexone (Monterosso et al. 2001). Currently, there are four FDA-approved pharmacotherapies available to treat alcohol dependence: naltrexone, extended-release naltrexone, acamprosate, and disulfiram. (These are summarized in Table 1).

Naltrexone

After being approved for opioid treatment by the FDA in 1984, oral naltrexone was approved 10 years later for the treatment of alcohol dependence after demonstrating a reduction in the pleasure produced from consumption of alcohol. Extended-release naltrexone (Vivitrol®) was FDA-approved for the treatment of alcohol dependence in 2006. Ethanol is believed to activate receptors in the opioid response system, which in turn activates numerous neurotransmitters, including dopamine. Since the efficacy of oral naltrexone is strongly contingent upon adherence, its usefulness as a treatment option for heavy drinkers has been questioned extensively. In many cases, compliance is the limiting factor in efficacy of oral naltrexone, while monthly injectable XR-NTX may overcome these challenges, due to fewer side effects and a better adherence profile (Mannelli et al. 2007). Although there has not yet been a head-to-head comparison of oral NTX and XR-NTX, evidence suggests that XR-NTX is highly efficacious for treatment of alcohol dependence, as well as compared to counseling or combined with counseling (O'Malley et al. 2007).

Disulfiram

Physiologically, disulfiram functions completely differently than naltrexone. Disulfiram blocks the oxidation of alcohol that results in increased levels of acetaldehyde. When a patient is administered disulfiram, yet continues to drink alcohol,

the buildup of acetaldehyde can produce several negative effects such as nausea, vomiting, chest pain, and increased heart rate shortly after consumption. Similar to oral naltrexone, adherence to disulfiram is problematic unless the patients are highly motivated. Most noncontrolled trials among offenders demonstrated a beneficial effect of disulfiram, while one study did not (Bourne et al. 1966). One study among those on probation in Atlanta found 50 % of the subjects were abstinent from alcohol for over 3 months after daily observation of disulfiram administration (Bourne et al. 1966). Additionally, another study in New York Later studies, however, found that disulfiram was of no more benefit than placebo. One of these studies compared disulfiram to group therapy among alcoholic offenders in New Orleans and reported only a modest improvement in abstinence between the two groups.

Acamprosate

While both oral naltrexone and disulfiram have been available for decades, ***acamprosate*** has been available only since 2004. The biological mechanism of acamprosate's activity is still not fully understood, yet it has been shown in numerous randomized clinical trials to maintain abstinence among alcohol dependents (who have already been detoxified from alcohol) modestly better than those who were administered placebo. In summary, as evidenced by findings from a large multicenter clinical trial known as the COMBINE study, oral naltrexone coupled with psychosocial support is still considered the most effective pharmacotherapeutic treatment for alcohol dependence (Anton et al. 2006; Garbutt et al. 2005).

Contraindications and Severe Adverse Events of Naltrexone, Disulfiram, and Acamprosate

According to the manufacturer's instructions, naltrexone should only be administered at the recommended doses. Liver injury has been documented when doses are excessive. *Naltrexone* should NOT be administered to patients with acute hepatitis or liver injury and is contraindicated in persons with childs-pugh class C cirrhosis. Naltrexone is also contraindicated in patients undergoing acute opioid withdrawal.

Disulfiram is contraindicated in patients receiving metronidazole, paraldehyde, alcohol, or alcohol-containing preparations (such as cough syrups). Patients with severe heart disease, coronary occlusion, and psychoses should avoid disulfiram. Disulfiram should never be administered to a patient currently under the influence of alcohol or without his/her knowledge. Several cases of hepatitis and liver failure have been documented in patients administered disulfiram.

Acamprosate is contraindicated in patients who have previously demonstrated hypersensitivity to acamprosate and in patients with severe kidney impairment. Similar to naltrexone, there has not been a rigorous evaluation on the safety of acamprosate in pregnant women.

The safety of naltrexone, disulfiram, and acamprosate has not been rigorously evaluated in pregnant women and is generally considered contraindicated in pregnancy. If the decision is made to use these drugs during pregnancy, careful consideration of the benefits and risks of administering these medications during pregnancy is necessary.

Medical and Psychiatric Comorbidities Associated with Dependency to Alcohol and Drug Abuse

Incarceration provides not only an opportunity to curb drug and alcohol use disorders, but also to address other underlying comorbidities. Infectious diseases, such as human immunodeficiency virus (HIV), tuberculosis (TB), viral hepatitis B (HBV) & C (HCV), and sexually transmitted infections (STIs) such as syphilis, gonorrhea, and chlamydia, are all strongly associated with drug and alcohol abuse. The burden of these diseases among individuals who have passed through the criminal justice system is disproportionately high, compared to those who have not. In 2006, it was estimated that one out of seven individuals

living with HIV in the USA had been through the CJS or 16 % of the incarcerated population (Maruschak 2008; Spaulding et al. 2009). For HCV and TB, more than a third of all those infected in the USA were either incarcerated or just released from a correctional facility.

Human Immunodeficiency Virus (HIV)

The rate of HIV and AIDS among those incarcerated is approximately three times and four times higher respectively than the US general population (Maruschak 2008, 2009; Spaulding et al. 2009). The high rates of HIV among incarcerated persons are in part due to the overlap of increased frequency of alcohol and drug use among persons who are incarcerated and the high-risk injection-drug-using behaviors and unprotected sexual behaviors that occur under the influence of these substances. Concurrent opioid and alcohol use is significantly associated with higher morbidity and mortality, decreased adherence to highly active antiretroviral therapy (HAART), and increased transmission of HIV among HIV-positive CJS populations. Among those with HIV disease who have comorbid alcohol or opioid dependence, adherence to HAART is low (Springer and Altice 2005). Thus, treatment of their chemical dependence should be prioritized among HIV-positive individuals in order to maintain adherence to HAART. In a small study among HIV-positive opioid-dependent incarcerated offenders in Connecticut, administration of both BPN and HAART resulted in low opioid positivity in urinalysis and high HIV viral load suppression at 3 months post-release (Springer et al. 2010; Springer et al. 2012). Currently, evaluation of administration of HIV treatment and MAT among incarcerated offenders before release is an active area of investigation.

Hepatitis C Virus (HCV)

With prevalence rates ranging from 16 % to 49 %, HCV is the most common blood-borne disease in the CJS. Injection-drug use is a strong risk factor for HCV acquisition. Most individuals with chronic HCV do not develop any symptoms; however, alcohol abuse can greatly accelerate end-stage liver disease (ESLD) and death (Springer et al. 2011). HCV treatment typically has been expensive and logistically difficult to implement in correctional settings due to its long duration. Newer and more effective treatments such as boceprevir and telaprevir that have shown more sustained virologic response when compared to standard treatment of pegylated interferon-ribavirin offer potential improved treatment opportunities for the incarcerated population as well. The CJS should consider evaluating the cost-effectiveness of HCV treatment, given the immense disease burden among inmates and potential risk to the community upon release.

HIV/HCV Coinfection

The prevalence of HIV/HCV coinfection is estimated to be between 50 % and 90 % among injection-drug users (IDUs). Estimates of prevalence of IDU among incarcerated adults range from 3 % to 28 %, and high-risk behaviors, such as sharing syringes, persist behind bars (Hammett 2006). Despite the potential for HIV and HCV acquisition in incarcerated settings, needle exchange programs are prohibited in American prisons and jails. Given the high burden of comorbid HIV and HCV disease in prisoners and the notion that ESLD is now the number one cause of death in HIV-positive persons, it behooves the CJS to consider treatment of HCV as well as treatment of drug and alcohol dependency to improve adherence to HCV treatment as a measure to decrease morbidity and mortality.

Tuberculosis (TB)

Transmission of TB is associated with unstable living conditions, poverty, and history of injection-drug use and is facilitated in crowded environments such as prisons and jails (Altice et al. 2010). The resurgence of TB in the last three decades can be largely attributed to the emergence of HIV (Altice et al. 2010). Susceptibility to TB is particularly high among those with high HIV viral loads due to suppression of the immune system from the virus (Friedland 2010). It has been estimated that in any given year, 40 % of all TB-positive individuals in the USA have circulated through the CJS (Hammett et al. 2002).

At intake, all inmates are screened for TB. Given the high prevalence of drug and alcohol users who are TB positive in correctional settings, treatment of TB is imperative. First-line treatments for latent TB infection, such as rifampin and isoniazid, are highly effective (95 % cure rate) with successful adherence (Friedland 2010). Incarceration provides an opportunity where patients can be treated by directly observed therapy (DOT) after patients have detoxed from alcohol or drug use. Chronic alcohol use in conjunction with TB treatment can cause severe hepatotoxicity. Moreover, once released into the community without proper linkage to health care services, resumption of alcohol or drug can disrupt compliance to treatment regimens and promote drug-resistant strains of TB (Friedland 2010). Therefore, it is imperative that alcohol treatment be instituted in concert with effective TB treatment to prevent adherence interruptions and liver toxicity.

Sexually Transmitted Diseases (STDs)

Previous research has shown that drugs and alcohol are associated with risky sexual behaviors that increase the likelihood of transmission of HIV and other STDs. HIV and other STDs disproportionately affect women within the CJS. Women are more likely to be incarcerated for crimes related to drugs or sex work, two behaviors that are strong correlates of STD transmission. Additionally, incarcerated adult women have high rates of chlamydia (6.3 %), syphilis (7.5 %), while juvenile women in detention facilities have high rates of gonorrhea (5.7 %). Screening and treatment of STDs in incarcerated settings should be administered in addition to treatment for opioid and alcohol dependency. Such interventions could prevent jails and prison from becoming incubators of STDs, given the strong association between drug/alcohol dependency, risky sexual behavior, and incarceration.

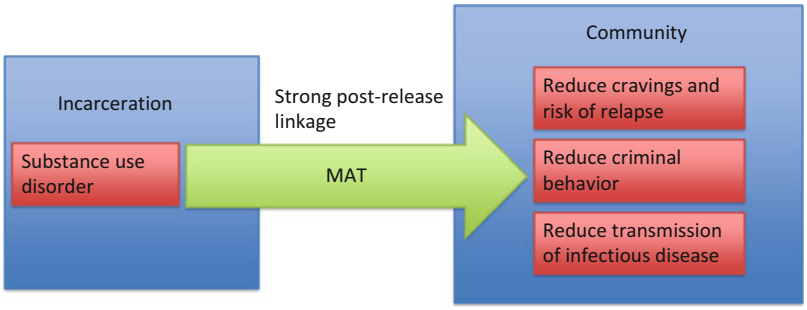
Mental Illness

In addition to the large burden of infectious disease comorbidities among drug offenders, substantial proportions also suffer from severe mental illness. Schizophrenia, major depression, and bipolar disorder are all highly associated with substance abuse. In 1998, it was reported that nearly a quarter of federal inmates, and over a third of jail inmates with psychiatric disorders, had a history of alcohol dependence (Ditton 1999). Upon release, individuals with comorbid mental illnesses often fare poorly in both mental illness treatment and substance abuse treatment programs due to the lack of integrated care in addressing both these ailments in one health center. Thus, untreated comorbid psychiatric disorders place released drug offenders at high risk for criminal behavior, parole violation, and consequent reincarceration.

Linkage to the Community Post-Release

The majority of CJS populations will eventually be released to the community. One of the greatest challenges facing recently released offenders is relapse to drug or alcohol use. Nearly 90 % of individuals with opioid dependence not on drug treatment will relapse within 12 months after being released (Kinlock et al. 2002). Moreover, overdose is the number one cause of death among all released incarcerated persons in Washington state (Binswanger et al. 2007). Additionally, relapse is correlated with criminal behavior and reincarceration.

During this critical transition period, a robust framework must be in place to coordinate access to drug treatment, mental health services, HIV care, or other medical services once an inmate returns to the community. Case management services are seen as an effective linkage between the criminal justice system and the community. Various case management models exist; however, several share core principles such as client advocacy, monitoring, referral, and planning. Several case management interventions have been able to successfully link substance users to treatment upon release (Lincoln et al. 2006), but participation in case management services offered in correctional settings still remains a significant barrier to its effectiveness. While in some instances inmates may be mandated to utilize these services as conditions of their parole,



Drug Abuse and Alcohol Dependence Among Inmates, Fig. 1 Schematic of reasons to initiate MAT to incarcerated persons with comorbid drug and alcohol dependence prior to release to the community

those who volunteer to enroll in case management services often build stronger rapport with their case managers. Case managers are not viewed as representatives of the CJS; thus, clients are more willing to "open up" to them about illicit behavior. By establishing trust with their clients, case managers can advocate more efficiently on behalf of them and ensure greater retention with their services. Overall, it is not conclusive whether case management alone is sufficient to improve the health or curb drug use of released prisoners.

Future Directions

The "revolving door," in reference to repeat offenders with problematic drug and alcohol abuse, continues to plague the US CJS and imposes a heavy burden on taxpayer dollars. Correctional care in the form of drug and alcohol treatment among offenders should be prioritized in order to disrupt this cycle. The very high influx of drug and alcohol dependents cycling through the CJS should be seen as an opportunity to initiate treatment and rehabilitation. In conjunction with behavioral interventions, effective pharmacotherapy has been shown to sustain abstinence and reduce cravings for opioids and alcohol for recently released inmates. Despite this, it cannot be emphasized enough that several major hurdles must be overcome in order to efficiently deliver correctional substance abuse care for incarcerated persons. First, improving education of correctional staff would likely increase acceptance of opioid and alcohol pharmacotherapies among correctional staff. The percentage of CJS settings offering medication-assisted therapy in the USA is dismally low, despite evidence indicating positive outcomes for offenders. Furthermore with the approval of Vivitrol® for treatment of opioid dependence as well as alcohol dependence, treatment options are expanded for offenders and the opioid antagonist may be more appealing to the CJS, given its once monthly adherence benefit and no overdose or diversion concerns.

Second, participation in psychosocial support programs in jails or prisons should be increased. Psychosocial groups, especially 12-step programs such as AA, have been very successful in reducing relapse to alcohol and drug use as has therapeutic communities and CBI.

Third, before release, correctional health care providers and case managers should work in concert to ensure that adherence to medically assisted therapy and medications for other comorbidities is not jeopardized, particularly for HIV disease. Relapse to drug and alcohol use is a significant risk factor for noncompliance or treatment failure, yet rigorous follow-up and treatment of substance use disorders can decrease relapse and recidivism rates. See Fig. 1.

The evidence-based practices presented in this entry are mostly limited to opioid and alcohol dependence. Polysubstance use is common among offenders, yet currently there is no federally approved medically assisted therapy for other highly abused addictive substances such as cocaine or methamphetamine. CBT has shown promising results in treating cocaine dependence; however, an effective pharmacologic treatment for long-term abstinence does not exist as of yet. Disulfiram, a pharmacotherapy for alcohol dependency, is

being investigated as a possible treatment for reducing cocaine use. Additionally, clinical trials are currently being conducted to evaluate the efficacy of an anti-cocaine vaccine. In a trial with 115 individuals however, less than 40 % of subjects reached the requisite antibody level needed to block cocaine's access to the brain. Treatment for methamphetamine dependency is of paramount importance now, as it is becoming a more emergent problem in many parts of the country. Potential immunotherapies for methamphetamine treatment have also been proposed; however, they still remain in the preclinical phase. Other pharmacotherapies show promise as well, such as XR-NTX and methylphenidate, but are still being evaluated.

Similar to management of most chronic diseases, treatment of substance use disorders among offenders requires a multidisciplinary approach inside and outside correctional settings. Moreover, a sizeable fraction of drug-using offenders suffer from numerous psychiatric and infectious comorbidities, which have broad societal and public health implications. Attention to inmates' medical needs must be addressed in prisons and jails as a measure of safety, not just for the individual, but also from a public health standpoint.

Recommended Reading and References

Altice FL, Kamarulzaman A, Soriano VV, Schechter M, Friedland GH (2010) Treatment of medical, psychiatric, and substance-use comorbidities in people infected with HIV who use drugs. Lancet 376(9738):367–387. doi:10.1016/s0140-6736(10)60829-x

Anton RF, O'Malley SS, Ciraulo DA, Cisler RA, Couper D, Donovan DM, Gastfriend DR, Hosking JD, Johnson BA, LoCastro JS, Longabaugh R, Mason BJ, Mattson ME, Miller WR, Pettinati HM, Randall CL, Swift R, Weiss RD, Williams LD, Zweben A (2006) Combined pharmacotherapies and behavioral interventions for alcohol dependence: the COMBINE study: a randomized controlled trial. JAMA 295(17):2003–2017. doi:10.1001/jama.295.17.2003

Binswanger IA, Stern MF, Deyo RA, Heagerty PJ, Cheadle A, Elmore JG, Koepsell TD (2007) Release from prison – a high risk of death for former inmates. N Engl J Med 356(2):157–165. doi:10.1056/NEJMsa064115

Bonczar TP, Glaze LE (2009) Publication probation and parole in the United States, 2008. US Department of Justice/Bureau of Justice Statistics, Washington, DC, NCJ 228230

Bourne PG, Alford JA, Bowcock JZ (1966) Treatment of skid-row alcoholics with disulfiram. Q J Stud Alcohol 27(1):42–48

Brahen LS, Capone T, Wiechert V, Desiderio D (1977) Naltrexone and cyclazocine. A controlled treatment study. Arch Gen Psychiatry 34(10):1181–1184

Coviello DM, Cornish JW, Lynch KG, Alterman AI, O'Brien CP (2010) A randomized trial of oral naltrexone for treating opioid-dependent offenders. Am J Addict 19(5):422–432. doi:10.1111/j.1521-0391.2010.00070.x

Ditton PM (1999) Mental health and treatment of inmates and probationers. US Department of Justice/Bureau of Justice Statistics, Washington, DC, NCJ 174463

Friedland G (2010) Infectious disease comorbidities adversely affecting substance users with HIV: hepatitis C and tuberculosis. J Acquir Immune Defic Syndr 55(Suppl 1):S37–S42. doi:10.1097/QAI.0b013e3181f9c0b6

Garbutt JC, Kranzler HR, O'Malley SS, Gastfriend DR, Pettinati HM, Silverman BL, Loewy JW, Ehrich EW (2005) Efficacy and tolerability of long-acting injectable naltrexone for alcohol dependence: a randomized controlled trial. JAMA 293(13):1617–1625. doi:10.1001/jama.293.13.1617

Grant BF, Dawson DA, Stinson FS, Chou SP, Dufour MC, Pickering RP (2004) The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. Drug Alcohol Depend 74(3):223–234. doi:10.1016/j.drugalcdep.2004.02.004

Griffith JD, Rowan-Szal GA, Roark RR, Simpson DD (2000) Contingency management in outpatient methadone treatment: a meta-analysis. Drug Alcohol Depend 58(1–2):55–66

Hammett TM (2006) HIV/AIDS and other infectious diseases among correctional inmates: transmission, burden, and an appropriate response. Am J Public Health 96(6):974–978. doi:10.2105/ajph.2005.066993

Hammett TM, Harmon MP, Rhodes W (2002) The burden of infectious disease among inmates of and releasees from US correctional facilities, 1997. Am J Public Health 92(11):1789–1794

Karberg JC, James DJ (2005) Substance dependence, abuse, and treatment of jail inmates, 2002. US Department of Justice/Bureau of Justice Statistics, Washington, DC, NCJ 209588

Kinlock TW, Battjes RJ, Schwartz RP (2002) A novel opioid maintenance program for prisoners: preliminary findings. J Subst Abuse Treat 22(3):141–147

Kinlock TW, Gordon MS, Schwartz RP, Fitzgerald TT, O'Grady KE (2009) A randomized clinical trial of methadone maintenance for prisoners: results at 12 months postrelease. J Subst Abuse Treat 37(3):277–285. doi:10.1016/j.jsat.2009.03.002

Kinlock TW, Gordon MS, Schwartz RP, Fitzgerald TT (2010) Developing and implementing

a new prison-based buprenorphine treatment program. J Offender Rehabil 49(2):91–109. doi:10.1080/10509670903534951
Levasseur L, Marzo JN, Ross N, Blatier C (2002) Frequency of re-incarcerations in the same detention center: role of substitution therapy. A preliminary retrospective analysis. Ann Med Interne (Paris) 153(3 Suppl):1S14–1S19
Lincoln T, Kennedy S, Tuthill R, Roberts C, Conklin TJ, Hammett TM (2006) Facilitators and barriers to continuing healthcare after jail: a community-integrated program. J Ambul Care Manage 29(1):2–16
Lobmaier PP, Kunoe N, Gossop M, Katevoll T, Waal H (2010) Naltrexone implants compared to methadone: outcomes six months after prison release. Eur Addict Res 16(3):139–145. doi:10.1159/000313336
Lussier JP, Heil SH, Mongeon JA, Badger GJ, Higgins ST (2006) A meta-analysis of voucher-based reinforcement therapy for substance use disorders. Addiction 101(2):192–203. doi:10.1111/j.1360-0443.2006.01311.x
Magura S, Lee JD, Hershberger J, Joseph H, Marsch L, Shropshire C, Rosenblum A (2009) Buprenorphine and methadone maintenance in jail and post-release: a randomized clinical trial. Drug Alcohol Depend 99(1–3):222–230. doi:10.1016/j.drugalcdep.2008.08.006
Mannelli P, Peindl K, Masand PS, Patkar AA (2007) Long-acting injectable naltrexone for the treatment of alcohol dependence. Expert Rev Neurother 7(10):1265–1277. doi:10.1586/14737175.7.10.1265
Maruschak L (2008) HIV in Prisons, 2006 (NCJ 222179). US Department of Justice, Bureau of Justice Statistics, Washington, DC
Maruschak L (2009) HIV in prisons, 2007–08. US Department of Justice/Bureau of Justice Statistics, Washington, DC
McMillan GP, Lapham SC (2005) Staff perspectives on methadone maintenance therapy (MMT) in a large southwestern jail. Addict Res Theory 13(1):53–63. doi:10.1080/16066350512331328159
Minozzi S, Amato L, Vecchi S, Davoli M, Kirchmayer U, Verster A (2011) Oral naltrexone maintenance treatment for opioid dependence. Cochrane Database Syst Rev (4), CD001333. doi:10.1002/14651858.CD001333.pub4. http://www.thecochranelibrary.com/view/0/index.html
Monterosso JR, Flannery BA, Pettinati HM, Oslin DW, Rukstalis M, O'Brien CP, Volpicelli JR (2001) Predicting treatment response to naltrexone: the influence of craving and family history. Am J Addict 10(3):258–268
Mumola C (1999) Substance abuse and treatment of state and federal prisoners, 1997. US Department of Justice/Bureau of Justice Statistics, Washington, DC, NCJ 172871
Mumola C, Karberg JC (2006) Drug use and dependence, state and federal prisoners, 2004. US Department of Justice/Bureau of Justice Statistics, Washington, DC, NCJ 213530
O'Malley SS, Garbutt JC, Gastfriend DR, Dong Q, Kranzler HR (2007) Efficacy of extended-release naltrexone in alcohol-dependent patients who are abstinent before treatment. J Clin Psychopharmacol 27(5):507–512. doi:10.1097/jcp.0b013e31814ce50d
Rinaldo D (2008) Medicaid financing of medication-assisted treatment for opiate addiction. Accessed 28 Jan 2013
Spaulding AC, Seals RM, Page MJ, Brzozowski AK, Rhodes W, Hammett TM (2009) HIV/AIDS among inmates of and releasees from US correctional facilities, 2006: declining share of epidemic but persistent public health opportunity. PLoS One 4(11):e7558. doi:10.1371/journal.pone.0007558
Springer SA, Altice FL (2005) Managing HIV/AIDS in correctional settings. Curr HIV/AIDS Rep 2(4):165–170
Springer SA, Chen S, Altice FL (2010) Improved HIV and substance abuse treatment outcomes for released HIV-infected prisoners: the impact of buprenorphine treatment. J Urban Health 87(4):592–602. doi:10.1007/s11524-010-9438-4
Springer SA, Azar MM, Altice FL (2011) HIV, alcohol dependence, and the criminal justice system: a review and call for evidence-based treatment for released prisoners. Am J Drug Alcohol Abuse 37(1):12–21. doi:10.3109/00952990.2010.540280
Springer SA, Qiu J, Saber-Tehrani AS, Altice FL (2012) Retention on buprenorphine is associated with high levels of maximal viral suppression among HIV-infected opioid dependent released prisoners. PLoS One 7(5):e38335. doi:10.1371/journal.pone.0038335
Tomasino V, Swanson AJ, Nolan J, Shuman HI (2001) The Key Extended Entry Program (KEEP): a methadone treatment program for opiate-dependent inmates. Mt Sinai J Med 68(1):14–20

Drug Corners

► Spatial Perspectives on Illegal Drug Markets

Drug Courts

Michael Rempel
Drug Courts, Criminal Justice Intervention, Center for Court Innovation, New York, NY, USA

Overview

Drug courts are the oldest, most prolific, and most studied of the major alternative court models,

which also include domestic violence, mental health, community, and reentry courts. Similar to these other models, drug courts organize their cases on a separate court calendar, presided over by a specially trained judge. What distinguishes drug courts is their focus on cases involving an underlying drug addiction. To treat the addiction, drug courts employ a combination of treatment and judicial oversight, generally for 1 year or longer. Judicial oversight generally involves regular drug testing, meetings with court-affiliated case managers, and status hearings before the judge. At these hearings, the judge and participant directly converse, while the attorneys often remain silent. The judge typically responds to progress with verbal praise or tangible incentives (e.g., certificates, journals, or gift cards) and to noncompliance with interim sanctions (e.g., more frequent status hearings, community service, or short jail stays). Drug courts are voluntary programs; those who do not wish to participate can have their cases handled in a conventional fashion instead.

The first drug court was an adult program that opened in Miami-Dade County in 1989. In 2009, two decades later, 2,459 drug courts had been established. These courts include 1,317 adult, 476 juvenile, 322 family dependency, 172 DWI, 89 tribal, and 83 other drug courts (Huddleston and Marlowe 2011). The original adult model enrolls an estimated 55,000 defendants per year in the USA (Bhati et al. 2008). Comparable estimates are unavailable for any other model. Internationally, drug courts have spread to countries as varied as Australia, Canada, Chile, Jamaica, Mexico, New Zealand, Norway, and the United Kingdom.

Numerous studies have shown that the original adult drug court model reduces re-offending, as compared with conventional prosecution (Gutierrez and Bourgon 2009; Mitchell et al. 2012; Shaffer 2011). Research suggests that adult drug courts reduce drug use as well (Government Accountability Office 2011; Rossman et al. 2011).

Over the past decade, scientific study surrounding drug courts has begun to evolve from evaluations of program impact (do they work) to studies of which target populations and program features enhance their success. Research has begun to coalesce, for instance, around the idea that the judge – through judicial status hearings and conversational interactions with participants – plays a particularly influential role (Marlowe et al. 2003; Rossman et al. 2011). This research bolsters the initial premise of the model that ongoing judicial oversight can substantially bolster the effects obtainable through community-based treatment alone.

This entry reviews why drug courts arose, describes their core policies and practices, surveys the research literature, and summarizes several current issues and controversies.

Origins

Several overlapping trends created conditions ripe for drug courts. Foremost among them was the strain on criminal justice systems nationwide, resulting from skyrocketing court and correctional caseloads. For instance, from the early 1980s to the late 1990s, criminal court caseloads increased by more than 50 % nationally (Ostrom and Kauder 1999). Over this same period, the increase in correctional caseloads was even greater. As compared with 500,000 jail and prison inmates in 1980, there were two million inmates at all points throughout the 2000s (Bureau of Justice Statistics 2012).

To explain these trends, key factors include the crack epidemic of the 1980s, combined with state and federal policies targeting drug-related crimes for aggressive enforcement. Beginning in 1998, data collected as part of the federal Arrestee Drug Abuse Monitoring (ADAM) Program confirmed the preexisting beliefs of many criminal justice officials that a high percentage of criminal arrestees were drug-involved. The 2003 survey found that 70 % of arrestees in approximately three dozen jurisdictions tested positive for at least one of nine illegal drugs, and 37 % of the arrestees were "heavy users," defined by self-reported use of marijuana, cocaine, heroin, methamphetamine, or PCP in at least 13 of the past 30 days (Zhang 2003).

In this context, drug courts arose as a solution that could garner support from many sides of the

ideological spectrum. For liberals, drug courts held the potential to expand access to treatment while counteracting a growing trend towards the mass incarceration of drug offenders. For conservatives, drug courts held the potential to increase public safety (through recidivism reductions) and to limit the costs of incarceration to state and local taxpayers. Not surprisingly, federal funding for drug courts has persisted across three presidential administrations.

The creators of the first drug courts, however, began with a relatively nonideological and practical objective: By routing drug cases to a single court calendar, the drug court model offered important efficiency advantages for overburdened court systems. Indeed, research demonstrates that efficiency was a critical objective motivating the earliest drug courts, but by the mid-1990s, efficiency was supplanted by the goals of rehabilitation and recidivism reduction (McCoy 2003).

Buttressing the case for rehabilitation as a legitimate court focus was the theory of *therapeutic jurisprudence*, which first gained currency in the early 1990s, just as drug courts began to spread (Wexler 1993). The theory argues that it is legitimate for the law to incorporate therapeutic considerations in addition to strictly legal ones that sentencing does not have to reflect "just desserts" alone (punishment proportional to offense) but can also embody a desire to address the underlying problems of the litigant. To facilitate the rehabilitation process, therapeutic jurisprudence embraces the idea that judges should adopt a nontraditional judicial role, including a willingness to engage in the classic drug court strategy of direct conversational interaction with defendants or other litigants.

Core Policy Components

Specific policies and practices vary from jurisdiction to jurisdiction, but in broad outline, the drug court model is relatively similar everywhere. To a large extent, this similarity reflects the influence of the National Association of Drug Court Professionals (NADCP), a trade association in existence since 1994 that produces drug court-related publications, provides technical assistance, and runs a popular annual training conference for drug court practitioners.

In 1997, NADCP convened a national group of experts and issued a document known as *Defining Drug Courts: The Key Components* (Bureau of Justice Assistance 1997). The document identifies ten policy components that all drug courts should adopt. Since the document was issued, federal and state agencies have routinely required drug courts to demonstrate that they follow all ten components in order to receive funding. The *Key Components* document has led drug courts to be defined by the following policies. (The list below does not verbatim reproduce the original "ten key components" but summarizes the resulting policies that now inform the field.)

- *Early Identification*: Potential participants are assumed to be especially receptive to the drug court intervention at the "crisis moment" comprised by the initial arrest or case filing. Therefore, drug courts screen and assess potential participants for drug dependence and other psychosocial problems as soon as possible after case initiation.
- *Community-Based Treatment*: Treatment is considered most effective when tailored to the individual. Therefore, drug courts link participants to community-based treatment, drawing on a continuum of possible outpatient and residential treatment modalities.
- *Legal Leverage*: Participants are considered most likely to remain engaged in treatment when faced with undesirable legal consequences for dropping out. Therefore, drug courts establish clear jail or prison alternatives for unsuccessful participation while using the positive incentive of a charge dismissal or reduction for graduates. (In family dependency drug courts, the legal incentives revolve around child reunification but are less cut and dry, since reunification must ultimately follow from the "best interests of the child.")
- *Judicial Status Hearings*: Participants are assumed to perform better when under close surveillance by the court and when routinely

reminded of their responsibilities. Therefore, drug courts require participants to report back to court regularly, often weekly or biweekly at the outset of participation, for judicial status hearings on their progress.

- *Direct Judicial Interaction*: As an authority figure, the judge is assumed to be in a unique position to motivate compliance. Therefore, the judge directly converses with participants during scheduled judicial status hearings, asks participants about their needs, acknowledges and praises progress, and admonishes participants for noncompliance.
- *Drug Testing*: Frequent and random drug testing is assumed to deter drug use and also to assist drug court staff in understanding whether participants are currently receiving a sufficient intensity of treatment. Therefore, drug courts administer random drug tests.
- *Interim Sanctions and Incentives*: Classic behavioral modification theory recommends the consistent use of sanctions and incentives. Therefore, the judge imposes sanctions for noncompliance (including short jail stays) and distributives incentives for progress (including regular symbolic incentives, such as praise and courtroom applause).
- *Multiple Chances*: The physiological effects of withdrawal, as well as psychological and other barriers to recovery, may make relapse-free drug court participation unlikely for many. Therefore, drug courts use interim sanctions to respond to initial relapses or other noncompliance while terminating participants only for repeated or severe misbehavior.
- *Case Management*: Court-affiliated case managers are used as a critical liaison between community-based programs and the court (i.e., judge and attorneys). Therefore, most drug courts employ staff members who devise the treatment plan, suggest modifications based on progress, and communicate with treatment providers. Case managers may also meet regularly with participants to motivate progress or address newly apparent needs.
- *Ancillary Services*: It is assumed that participants may have multitude of needs, including co-occurring mental health disorders, lack of education, vocational or employment deficits, and family dysfunction. Therefore, drug courts are prepared to link participants to additional services other than substance abuse treatment.
- *Collaboration*: Drug courts are predicated on the notion that the court, attorneys, and treatment providers should seek the same goals: the rehabilitation of the individual and the consequent reduction in that individual's threat to public safety. Therefore, once a participant is enrolled, all parties curtail the adversarial process and are supposed to work together to promote the recovery of the individual. To facilitate collaboration, most drug courts hold regular "staffing" meetings attended by an interdisciplinary "drug court team." In these meetings, the team discusses specific cases and makes decisions, usually by consensus, on how to respond to their compliance and service needs.

Research on Adult Drug Courts

More than 150 evaluations have been conducted to date, of which by far the greatest number has focused on the adult drug court model.

Do Adult Drug Courts Work?

Research has amply demonstrated the positive impact of adult drug courts on re-offending. Across more than 90 studies, including statewide evaluations in California, Indiana, Maryland, New York, Ohio, and Washington, adult drug courts have consistently produced recidivism reductions – although the precise magnitude of the impact has varied from site to site. Three *meta-analyses* – which synthesize the findings of other studies – variously estimated that adult drug courts produce an average reduction in the rearrest or reconviction rate of 8–13 percentage points (Gutierrez and Bourgon 2009; Mitchell et al. 2012; Shaffer 2011). (e.g., a 13-point reduction might involve a rearrest rate of 50 % in the comparison group declining to 37 % among those in drug court.) Most evaluations tracked defendants over 2 years or less, but several extended

the follow-up period to 3 years or longer and still reported positive effects (Gottfredson et al. 2006; Rempel et al. 2003).

Only a handful of studies have directly examined effects on future drug use, but their results also were mostly positive (e.g., see Gottfredson et al. 2005). NIJ's *Multi-Site Adult Drug Court Evaluation*, a 5-year study of 23 drug courts and six comparison jurisdictions (Rossman et al. 2011), found that in the year prior to 18-month follow-up, drug court participants were significantly less likely to report any drug use (56 % vs. 76 %) or any "serious" use (41 % vs. 58 %). (Serious use omits marijuana and "light" alcohol use, with the latter defined as less than four drinks per day for women and less than five drinks per day for men.) The same study also detected positive effects when examining the results of oral swab tests that were conducted at the 18-month follow-up.

Among the plethora of single-site studies in the literature, three involve the random assignment of defendants to drug court and control conditions. The first of these randomized controlled trials took place in Maricopa County (AZ) and detected mixed results after 1 year (Deschenes et al. 1995) but more positive results after a 3-year timeframe (Turner et al. 1999). A randomized trial of a New South Wales (Australia) drug court reported a significant reduction in rearrests over 18 months (Shanahan et al. 2004). Finally, a randomized trial of the Baltimore drug court found that over 3 years, the court significantly reduced rearrests (Gottfredson et al. 2006) as well as drug and alcohol use (Gottfredson et al. 2005).

Do Adult Drug Courts Save Money?

Part of the original rationale for adult drug courts was to generate cost savings for increasingly overburdened criminal justice systems. Moreover, an array of cost-benefit studies almost universally confirms that drug courts save money, at least in the long term. An evaluation of nine adult drug courts in California found that the median drug court saved $5,139 per participant (Carey et al. 2005). This study also found that across multiple public agencies, including the court, prosecutor, public defender, law enforcement, probation, corrections, and treatment, the largest savings were accrued by corrections – through reductions in incarceration – and the only agency that incurred a net cost was treatment (e.g., through Medicaid/Medicare payments). Other cost studies have reported similar findings. For instance, across six sites in Washington State, five produced cost savings at an average of $3,892 per participant (Barnoski and Aos 2003). *NIJ's Multi-Site Drug Court Evaluation* detected average savings of $5,680–$6,208 per participant across its 23-site drug court sample. However, this last study came with an important asterisk: Drug courts entailed higher up-front costs (for treatment and other services) than comparison jurisdictions. Drug courts ultimately produced savings by reducing recidivism: that is, by reducing the costs that would otherwise have been produced by future crimes. These crime-related savings largely materialized, because drug courts achieved a significant reduction in the *most serious* future crimes that have otherwise produced substantial healthcare and property-related costs to victims. In sum, drug courts achieved a significant return on investment with high-risk offenders who might otherwise have committed serious crimes, but they produced far smaller savings, if any, with low-risk offenders.

Why Do Adult Drug Courts Work?

Much of the early research on adult drug courts focused on the bottom-line question of *whether* they work, but recent studies have yielded a rich array of findings concerning *why* they work – which theories of change explain their capacity to alter participant behavior for the better.

Treatment. A long-standing prior literature finds that when drug-addicted individuals are retained in treatment for significant periods – at least 90 days and ideally up to 1 year – those individuals tend to engage in less posttreatment drug use and criminal behavior. Research also shows that cognitive-behavioral therapy (CBT) is particularly effective in creating the pro-social thought, attitudinal, and decision-making changes that can, in turn, elicit pro-social behaviors (see Lipsey et al. 2007). Finally, research

suggests that treatment is most effective when it does not adopt a "one-size-fits-all" approach but is tailored to the individual characteristics, learning style, and needs of each participant (Andrews and Bonta 2006). Despite the sizable preexisting literature on this subject, drug court research has yet to uncover a clear treatment effect. Instead, some research suggests that many drug courts refer participants to treatment providers that have failed to adopt evidence-based practices (Lutze and van Wormer 2007). A recent meta-analysis concurs that evidence-based practices are underutilized and finds that when they *are* used, drug courts produce larger recidivism reductions than otherwise (Gutierrez and Bourgon 2009). This analysis indicates that treatment *can* contribute to positive impacts, but because treatment quality varies substantially, it does not always produce its desired effects.

Deterrence. Based on research with other offender populations, it is possible to *deter* future misbehavior with legal sanctions that involve *certainty* (each infraction elicits a sanction), *celerity* (sanction imposed soon after the infraction), and *severity* (sanction is sufficiently undesirable to deter noncompliance but not so severe as to preclude upgrading to a more serious sanction after subsequent infractions) (e.g., Marlowe and Kirby 1999). In a drug court context, deterrence entails routine surveillance through judicial status hearings, drug tests, and case manager meetings; threat of interim sanctions; and threat of incarceration for final termination. Of these practices, several studies have found that drug testing and judicial status hearings are effective in reducing crime and drug use (e.g., Gottfredson et al. 2007; Marlowe et al. 2003; Rossman et al. 2011). Yet, several studies involving in-depth participant focus groups suggest that it is less the deterrent effect of surveillance and more the positive engagement effect of motivational interactions with a judge that leads judicial status hearings in particular to be effective (e.g., Goldkamp et al. 2001).

Providing a clearer test of deterrence theory, research also suggests that the threat of jail or prison for failing drug court altogether is a key factor motivating compliance. NIJ's *Multi-Site Adult Drug Court Evaluation* found that participants who perceived themselves to face more severe consequences in the event of program failure engaged in less noncompliance than others while enrolled in the program and less crime and drug use at follow-up (Rossman et al. 2011). This research is consistent with previous studies, which generally link greater legal leverage to improved treatment outcomes (e.g., Rempel and DeStefano 2001; Young and Belenko 2002). Qualifying these findings, research also makes clear that it is not only the factual jail or prison sentence that participants face in the event of failure that influences their performance. Drug court participants who face similar legal consequences may have differing *perceptions* of those consequences due to what the participants were told, by whom, how often they were reminded of their responsibilities, and how well they actually understood those consequences. Research has shown that eliciting participant *perceptions* of legal pressure comprises the critical link to improved behavioral outcomes (Young and Belenko 2002).

As opposed to the threat of incarceration for final program failure, research has been less clear concerning the deterrent effect of *interim sanctions*. In determining when and how to use sanctions, drug courts often employ a great deal of individualized discretion – that is, using different sanctions with different participants and not necessarily imposing a sanction in response to each and every infraction (Rempel et al. 2003; Rossman et al. 2011). Yet, individualized discretion may vitiate the behavior modification principle of *certainty*, which holds that it is best to impose a sanction each time and to employ similar sanctions in response to similar infractions. Moreover, it is unclear at this time whether interim sanctions might yield clearer and more positive effects if drug courts adhered more consistently to best sanctioning practices.

Procedural Justice. Procedural justice concerns the fairness of court procedures and interpersonal treatment while a case is processed (Tyler and Huo 1990). Key dimensions include *voice* (litigants have their side heard), *respect* (litigants are treated with dignity and respect), *neutrality* (decision-makers are seen as trustworthy and unbiased), and *understanding*

(court language and decisions are readily understood). Some research finds that when litigants believe the court process was fair, they become more likely to comply with court orders and to follow the law in the future (e.g., Tyler and Huo 2002). Procedural justice acts as the positive flip side to deterrence, eliciting compliance through fair procedures rather than through threat of consequences. The randomized controlled trial of the Baltimore drug court found that greater participant perceptions of procedural justice were associated with less crime and drug use at 3-year follow-up (Gottfredson et al. 2007). Similarly, in *NIJ's Multi-Site Adult Drug Court Evaluation*, drug court participants held significantly greater perceptions than the comparison group that the judge treated them fairly – and in turn, these perceptions strongly predicted reduced crime and drug use at follow-up. This same study found that specific drug courts in which the researchers independently rated the judge as having a more engaging demeanor – more respectful, fair, attentive, enthusiastic, consistent, predictable, caring, and knowledgeable – produced better outcomes than other drug courts. In light of this research, the federal Office of Justice Programs recently identified both *procedural justice* and *judicial interaction* as design features that all drug courts should fully embrace in order to be evidence-based (OJP 2012).

Research on Other Drug Court Models

This section reviews what has been learned to date about juvenile, family dependency, and DWI drug courts. In addition, reentry courts receive attention in a separate alternative courts entry (Lindquist et al., this volume).

As compared with the adult model, some have argued that juvenile drug courts may not be as successful, since juveniles do not tend to be addicted to drugs. Instead, they are more often casual drug users, especially of marijuana, who face a series of other social and psychological problems, including ties to deviant peer groups, low family functioning, poor educational performance, poor impulse control, and developmental disabilities. Still other juveniles may not have any severe problems of this nature but may instead be engaging in common teenage deviance that is likely to desist on its over time, in the absence of intervention. Indeed, results to date have been mixed. A recent meta-analysis of 34 juvenile drug court evaluations detected an average reduction in re-offending of eight percentage points; however, the average effect was smaller in the most methodologically rigorous studies, and when isolating drug-related re-offending in particular, there was not any significant impact (Mitchell et al. 2012).

Although these results are ostensibly disappointing, some researchers have found that when juvenile drug courts employ certain evidence-based practices, they can be more effective. Specifically, several studies have found that juvenile drug courts produce positive outcomes when they facilitate pro-social peer activities, limit contact with antisocial peers, involve family members in judicial status hearings, and employ evidence-based treatments such as multisystemic therapy, which entails comprehensive engagement with the youth, parents, teachers, and other systems in which the youth are involved (Salvatore et al. 2010; Schaeffer et al. 2010).

By comparison with juvenile drug courts, research findings have tended to be more positive for the remaining drug court models. Family dependency drug courts are somewhat unique in that they seek both to reduce substance abuse by the respondent parent and to achieve a positive permanency outcome for the child. Across several family dependency drug courts that have been evaluated, five of seven produced increased treatment completion rates for the respondent, and six of eight produced increased rates of parent–child reunification than equivalent comparison groups (e.g., see Fritsche et al. 2011; Green et al. 2009).

DWI courts have also yielded positive impacts. A recent meta-analysis isolated the impact of 28 DWI courts on re-offending and found that all except four reduced re-offending to at least some degree, with an average reduction of 12 percentage points (Mitchell et al. 2012). It is perhaps unsurprising that the results for DWI

courts mirror those for the original adult model. These two models are highly similar, with both focusing on drug-involved adult criminal defendants – except DWI courts serve those whose court case involved DWI or DUI charges.

Issues and Controversies

Drug courts have spawned a large number of issues and controversies. Several prominent examples follow, of which some concern the legal and constitutional implications of the drug court model, whereas other issues concern specific findings in the evaluation literature.

The Therapeutic Judicial Role

Some have expressed concern that it may not be legally appropriate for judges to serve in any other capacity than as neutral arbiters of facts and legal questions. In a well-publicized commentary, the Honorable Morris B. Hoffman succinctly summarized this position as follows: "Judges have the right to exercise only those powers necessary to dispose of the cases before us. When we succumb to the very human temptation to do more – to fill the void that is so achingly apparent in so many of the dysfunctional people we see every day – we not only risk being wrong, but we risk being imperial (2000: 1478)." Hoffman further posits that when the judicial branch involves itself in social policymaking, or when individual judges apply therapeutic methods, it violates the constitutional separation of powers, which leaves it exclusively to the legislative and executive branches to develop and implement social policy (2000: 1479).

Others judges have sought to articulate how legal due process can be maintained even as judges extend their focus beyond legal process alone to the *outcomes* their decisions produce, such as recidivism reductions (see Hora et al. 1999). Moreover, a survey of more than 1,000 trial court judges nationwide found, on balance, support for judging methods that are common to drug courts and other alternative courts. Eighty percent of the responding judges believed it was very or somewhat important for judges to consider "the individual needs or underlying problems of the litigant," whereas only 25 % believed that that "problem-solving compromises the neutrality of the court" (Farole et al. 2008). Nonetheless, some would question whether majority opinion matters on these kinds of questions. The therapeutic judicial role that drug courts embrace – and especially the use of direct interaction between judge and participant coupled with a less adversarial process between the attorneys – remains highly controversial as legal practices. As a legal protection for drug court participants, while not opposing the therapeutic judicial role per se, the National Association of Criminal Defense Lawyers (NADCL) (2009) recommends that defense attorneys be present at all judicial status hearings, in the event that issues arise (e.g., related to sanctions or possible program termination) that require zealous defense advocacy.

Risk of Net Widening

A second concern, also rooted in legal due process principles, is that drug courts may inappropriately deepen and lengthen the criminal justice involvement of many defendants. Known as "net widening," this concern takes two essential forms. First, some have linked the establishment of drug courts to the increased criminalization of nonviolent drug behavior: that is, to rampant drug arrests and prosecutions (Drug Policy Alliance 2011). However, there is relatively little evidence to support this concern beyond a reported massive increase in drug arrests after the founding of the Denver (CO) Adult Drug Court (Hoffman 2000).

The second variant of net widening appears in the argument that, despite their widely promulgated status as an alternative to incarceration, the average drug court *increases* the criminal justice involvement of their participants – and increases rather than reduces time spent incarcerated (Drug Policy Alliance 2011). Here, critics can point to hard evidence supporting their position. A number of multisite studies found that adult drug court participation lasts about 15 months on average, representing a longer period of time than the jail or prison sentence that most participants would have otherwise faced under conventional prosecution (Rempel et al. 2003; Rossman et al. 2011).

These same studies also found that those who fail drug court receive a significantly longer jail or prison sentence than they would have received had they not enrolled in the first place. After considering the lengthier sentences imposed on those who fail with the complete avoidance of jail or prison sentences for those who graduate from drug court, the *average* effect on incarceration appears to be a wash. Neither *NIJ's Multi-Site Adult Drug Court Evaluation* nor a randomized controlled trial of the Baltimore drug court detected a net difference in incarceration sentences between drug court and comparison defendants (e.g., see Rossman et al. 2011). Further, in a study of six New York State drug courts, when combining those who graduated and failed the program, participants averaged significantly shorter incarceration sentences than the comparison group in three sites, a significantly longer sentence in one site, and no difference at all two sites (Rempel et al. 2003). Drug court proponents might counter that, even if drug courts do not produce an average reduction in incarceration on the initial criminal case, they may still reduce incarceration in the long term through recidivism reductions. This is exactly why the cost-benefit literature has found that adult drug courts save money in the long term. Nonetheless, those who are concerned with net widening tend to predicate their position on the basic legal fairness of the initial court outcomes, regardless of whether outcomes are positive in the long term.

In consideration of these issues, NACDL (2009) recently advocated a number of reforms. First, contrary to the practice in many drug courts, NACDL proposed that prosecutors should not be allowed to reject cases for drug court when formal charge and criminal history criteria indicate that the case is eligible to participate. NACDL noted that some prosecutors may tend to reject the highest-risk cases, even though such cases are the least vulnerable to net-widening concerns. More generally, NACDL recommended limiting drug court eligibility to those defendants who would otherwise face lengthy sentences under conventional prosecution while developing less intensive programming for defendants who face less legal exposure.

Adult Drug Court Volume

The net-widening critique typically translates into a policy recommendation to limit participation to those individuals who would otherwise face substantial legal exposure (Drug Policy Alliance 2011). Some social scientists, however, have suggested that drug courts should *increase* rather than reduce their numbers. One analysis estimated that in 2005, whereas 1,471,338 adult arrestees in the USA either abused or were dependent on drugs, only 55,365 (3.8 %) were enrolled in an adult drug court (Bhati et al. 2008). Even allowing that these estimates are now several years dated and were based on extrapolations from multiple data sources, this analysis makes clear that adult drug courts reach a small fraction of the eligible pool. This same study also projected that whereas adult drug courts in 2005 produced financial benefits to society worth about $624 million, if these courts had served all substance-abusing and substance-dependent defendants, the benefit might have reached $46 billion – resulting from an investment of $13.7 billion in treatment. Importantly, the implication that drug courts should serve more participants does not automatically contradict the aforementioned concerns about net widening. Some might reconcile these considerations in proposing that adult drug courts should reach a higher percentage of all *serious* cases – those that face substantial legal exposure – while carefully restricting eligibility where net-widening concerns might apply.

Uneven Treatment Quality

A long-standing research literature supports the general benefits of substance abuse treatment. Yet, drug court research has yet to provide clear confirmation of the positive effects of *treatment* – as opposed to judicial status hearings or other drug court practices. This does not mean that treatment is unimportant, for less than a handful of drug court studies to date have attempted a tease out its specific impact. Still, available research indicates that many drug court programs do not provide what national experts would term *evidence-based treatment*. Many treatment programs suffer from inadequate staff training, high turnover, lack of written curricula specifying what each treatment session

should cover, and insufficient use of proven cognitive-behavioral therapy techniques (Lutze and van Wormer 2007). In addition, a serious drug addiction is a brain disease that can often be addressed in part through medication-assisted treatment (MAT). However, many drug courts do not have access to local treatment programs whose staff is trained to provide and monitor such treatment (Lutze and van Wormer 2007). Of final note, whereas drug court participants often possess multiple problems other than substance abuse, including pro-criminal thought patterns, ties to antisocial peers, antisocial personality patterns, and employment deficits, many drug courts neither conduct a rigorous assessment for these other needs nor are able to make appropriate treatments available.

Some might conclude that drug court staff should monitor and seek to improve the quality of treatment. Supporting such a conclusion, a recent review found that drug courts that do adhere to one or more evidence-based assessment and treatment practices produced significantly larger reductions in re-offending than other drug courts (Gutierrez and Bourgon 2009). However, in many jurisdictions, particularly small ones with a limited number of available providers, it may simply not be possible for a court to order or obtain best treatment practices. Accordingly, the question of how to improve the quality of treatment received by drug court participants is one that depends not merely on clear guidance from research but also on overcoming practical obstacles to the dissemination of evidence-based practices within local provider communities.

Appropriate Target Population

General research on offender interventions recommends varying program intensity based on risk of re-offense. Specifically, research indicates that high-risk defendants – those who are especially likely to re-offend in the absence of any intervention – require a particularly intensive form of treatment. By contrast, low-risk defendants may not be well served by an intensive and lengthy intervention such as drug court; since they are unlikely to re-offend in any case, low-risk defendants may be better served if they are left alone (Andrews and Bonta 2006).

Little research has put the risk principle to the test in drug courts. However, it is conceivable that the literature on juvenile drug courts is somewhat mixed, in part because these courts may order some teenagers to a year or more of intensive program participation, who might otherwise have desisted from crime on their own – that is, their risks and needs may have been too low to merit subjection to the intensive drug court model.

Regarding adult drug courts, one study of the Los Angeles program confirmed that it was more effective with high- than low-risk defendants (Fielding et al. 2002). Similarly, *NIJ's Multi-Site Adult Drug Court Evaluation* found some evidence that those who were particularly likely to commit serious crimes in the future and those who presented with a more severe addiction at baseline were especially likely to benefit from their participation. This same study uncovered few other differences in the magnitude of the drug court impact, based on participant demographics (age, race, or sex) or self-reported motivation at baseline (Rossman et al. 2011).

Interestingly, the social science-based concern that drug courts are best suited to high-risk defendants dovetails with the aforementioned defense bar position that drug courts should limit eligibility to high-risk/high-leverage defendants, who would otherwise face a substantial jail or prison sentence in the absence of drug court participation. In considering these issues, the National Association of Drug Court Professionals recently issued a two-part series of policy papers that essentially advised drug courts to abandon a "one-size-fits-all" program model in favor of multiple *tracks* (Marlowe 2012). In this system, programs would apply the full drug court model to defendants who pose a high risk of re-offending and have serious treatment needs related to a substance-dependence problem (track 1). Those who combine a high risk of re-offending in general but with less serious drug treatment needs in particular would receive intensive judicial oversight but less intensive community-based treatment (track 2). Conversely, those who combine a low risk of re-offending with serious treatment needs would receive less intensive oversight – that is, status hearings only when noncompliant – coupled

with regular treatment (track 3). Finally, those who are both low risk and low need might receive general prevention services but would not be required to attend any form of intensive judicial oversight or treatment (track 4). Although the state of Missouri intends to implement a track system along these lines, it is unclear whether a new national movement will be launched in this direction. Moreover, research has not been conducted to determine whether a track-based system in fact yields improved drug court outcomes. In short, the development of tailored responses to different defendant populations is a potential cutting-edge policy direction that may or may not ultimately be the subject of broad experimentation nationwide.

Conclusion

As the first alternative court model, drug courts have existed since 1989, spawned the development of other alternative courts in the USA and internationally, and been the subject of a growing body of research and evaluation studies. At the same time, important unanswered questions persist. These include the effectiveness of models other than adult and DWI drug courts, the role of treatment in explaining drug court success and how to improve its quality, the appropriate target population for the full-length drug court model, and the solution to a number of competing legal and social science considerations that variously mitigate in favor of and against expanding the number of individuals who are enrolled in drug courts. Regardless of the effectiveness of any one model, the diffusion of drug courts nationwide has transformed the role that many courts are now prepared to embrace in devising solutions to vexing social problems.

Related Entries

- ▶ Community Courts
- ▶ Mental Health Courts
- ▶ Reentry Courts
- ▶ Therapeutic Jurisprudence

Recommended Reading and References

Drug Policy Alliance (2011) Drug courts are not the answer: toward a health-centered approach to drug use. Drug Policy Alliance, New York

Andrews DA, Bonta J (2006) The psychology of criminal conduct, 4th edn. LexisNexis, Newark

Barnoski R, Aos S (2003) Washington state's drug courts for adult defendants: outcome evaluation and cost-benefit analysis. Washington State Institute for Public Policy, Olympia

Bhati A, Roman J, Chalfin A (2008) To treat or not to treat: evidence on the prospects of expanding treatment to drug-involved offenders. The Urban Institute, Washington, DC

Bureau of Justice Assistance (1997) Defining drug courts: the key components. Bureau of Justice Assistance, Washington, DC

Bureau of Justice Statistics (2012) Correctional populations: key facts at a glance. Washington, DC: Bureau of Justice Statistics. http://bjs.ojp.usdoj.gov/content/glance/tables/corr2tab.cfm

Carey SM, Crumpton D, Finigan MW, Waller MS (2005) California drug courts: a methodology for determining costs and benefits, phase II: testing the methodology: final report. NPC Research, Portland

Deschenes E, Turner S, Greenwood P (1995) Drug court or probation? An experimental evaluation of Maricopa County's drug court. Justice Syst J 18:55–73

Farole DJ, Rempel M, Byrne F, Chang Y (2008) Problem-solving and the American bench: a national survey of trial court judges. Center for Court Innovation, New York

Fielding JE, Tye G, Ogawa P, Imam IJ, Long AM (2002) Los Angeles county drug court programs: initial results. J Subst Abuse Treat 23:217–224

Fritsche S, Bryan J, Kralstein D, Farley E (2011) The Bronx family treatment court 2005–2010: impact on family court outcomes and participant experiences and perceptions. Center for Court Innovation, New York

Goldkamp J, White MD, Robinson JB (2002) An honest chance perspective on drug courts: findings from drug court participant focus groups in Brooklyn, Las Vegas, Miami, Portland, San bernadino, and Seattle. Crime and Justice Research Institute, Philadelphia

Gottfredson DC, Kearley B, Najaka SS, Rocha CM (2005) The Baltimore city drug treatment court: 3-year outcome study. Eval Rev 29(1):42–64

Gottfredson DC, Kearley B, Najaka SS, Rocha CM (2006) Long-term effects of participation in the Baltimore city drug treatment court: results from an experimental study. J Exp Criminol 2(1):67–98

Gottfredson DC, Kearley B, Najaka SS, Rocha CM (2007) How drug treatment courts work: an analysis of mediators. J Res Crime Delinq 4(3):3–35

Green BL, Furrer CJ, Worcel SD, Burrus SWM, Finigan MW (2009) Building the evidence base for family drug treatment courts: results from recent outcome studies. Drug Court Review VI(2):53–82

Gutierrez L, Bourgon G (2009) Drug treatment courts: a quantitative review of study and treatment quality. Retrieved 30 Sept 2011 from the Public Safety Canada Web site: http://www.publicsafety.gc.ca/res/cor/rep/2009-04-dtc-eng.aspx
Hoffman MB (2000) The drug court scandal. North Carolina Law Rev 78:1439–1533
Hora PF, Schma WD, Rosenthal JTH (1999) Therapeutic jurisprudence and the drug treatment court movement: revolutionalizing the criminal justice system's response to drug abuse and crime in America. Notre Dame Law Review 74:439–537
Huddleston W, Marlowe DB (2011) Painting the current picture: a national report on drug courts and other problem-solving court programs in the United States. National Drug Court Institute, Washington, DC
Lipsey MW, Landenberger NA, Wilson SJ (2007) Effects of cognitive-behavioral programs for criminal offenders. The Campbell Collaboration, Oslow
Lutze FE, van Wormer JG (2007) The nexus between drug and alcohol treatment program integrity and drug court effectiveness: policy recommendations for pursuing success. Crim Just Pol Rev 18(3):226–245
Marlowe D (2012) Alternative tracks in adult drug courts: Matching your program to the needs of your clients: Part two of a two-part series. Alexandria: National Drug Court Institute
Marlowe DB, Kirby KC (1999) Effective use of sanctions in drug courts: lessons from behavior research. Natl Drug Court Inst Rev 2(1):1–31
Marlowe DB, Festinger DS, Lee PA, Schepise MM, Hazzard JER, Merrill JC, Mulvaney FD, McLellan AT (2003) Are judicial status hearings a key component of drug court? During-treatment data from a randomized trial. Crim Justice Behav 30:141–162
McCoy C (2003) The politics of problem-solving: an overview of the origins and development of therapeutic courts. Am Crim Law Rev 40:1513–1544
Mitchell O, Wilson DB, Eggers E, MacKenzie DL (2012) Assessing the effectiveness of drug courts on recidivism: a meta-analytic review of traditional and non-traditional drug courts. J Crim Just 40:60–71
National Association of Criminal Defense Lawyers (NADCL) (2009) America's problem-solving courts: the criminal costs of treatment and the case for reform. National Association of Defense Lawyers, Washington, DC
Office of Justice Programs (2012) Seven program design features: adult drug court principles, research, and practice. Washington, DC: Bureau of Justice Assistance, National Institute of Justice, and National Center for State Courts. http://research2practice.org/resources/
Ostrom B, Kauder N (1999) Examining the work of state courts, 1998: a national perspective from the court statistics project. National Center for State Courts, Alexandria
Rempel M, DeStefano CD (2001) Predictors of engagement in court-mandated treatment: findings at the Brooklyn treatment court, 1996–2000. J Offender Rehabil 33(4):87–124
Rempel M, Fox-Kralstein D, Cissner A, Cohen R, Labriola M, Farole D, Bader A, Magnani M (2003) The New York state adult drug court evaluation: policies, participants, and impacts. Center for Court Innovation, New York
Rossman SB, Roman JK, Zweig JM, Rempel M, Lindquist CH (eds) (2011) The multi-site adult drug court evaluation. The Urban Institute, Washington, DC
Salvatore C, Henderson MA, Hiller ML, White E, Samuelson B (2010) An observational study of team meetings and status hearings in a juvenile drug court. Drug Court Review 7(1):95–124
Schaeffer CM, Henggeler SW, Chapman JE, Halliday-Boykins CA, Cunningham PB, Randall J, Shapiro SB (2010) Mechanisms of effectiveness in juvenile drug court: altering risk processes associated with delinquency and substance abuse. Drug Court Review 7(1):57–94
Shaffer DK (2011) Looking inside the black box of drug courts: a meta-analytic review. Justice Quart 28(3):493–521
Tyler TR, Huo YJ (2002) Trust in the law. Russell Sage, New York
Wexler DB (1993) Therapeutic jurisprudence and the criminal courts. Wm Mary L Rev 35:279–299
Young D, Belenko S (2002) Program retention and perceived coercion in three models of mandatory drug treatment. J Drug Issues 22(2):297–328
Zhang A (2003) Drug and alcohol use and related matters among arrestees 2003. National Opinion Research Center, Chicago

Drug Courts' Effects on Criminal Offending

David B. Wilson[1], Ojmarrh Mitchell[2] and Doris Layton MacKenzie[3]
[1]Department of Criminology, Law and Society, George Mason University, Fairfax, VA, USA
[2]Department of Criminology, University of South Florida, Tampa, FL, USA
[3]Justice Center for Research, The Pennsylvania State University, University Park, PA, USA

Overview

Drug courts are problem-solving courts that divert generally low-level drug-involved offenders from conventional prosecution. These courts use their

legal and moral authority to monitor offenders' abstinence from drugs and compliance with individualized treatment programs. This is done through regular status hearings before the court and routine drug testing. All drug court participants are expected to be engaged in some form of drug treatment. Successful complete of a drug treatment program is celebrated with a graduation ceremony and generally reward with dismissal of charges or expunging the initiating conviction.

This entry summarizes the findings from a recent systematic review and meta-analysis conducted by Mitchell et al. (2011) of the numerous studies that have examined the effectiveness of this criminal justice innovation. The systematic search identified 154 independent, eligible evaluations, 92 evaluations of adult drug courts, 34 of juvenile drug courts, and 28 of drunk-driving (DWI) drug courts. The findings most strongly support the effectiveness of adult drug courts, as rigorous evaluations find reductions in recidivism and these effects generally persist for at least 3 years. The evidence also cautiously suggests that DWI drug courts are effective in reducing recidivism, although few rigorous experimental evaluations have been conducted of DWI courts and the findings of these more rigorous evaluations are inconsistent. For juvenile drug courts, they found considerably smaller effects on recidivism than for adult drug courts but the evidence is suggestive of positive effects.

Background Description

Since they emerged in the late 1980s, drug courts have proliferated across the United States and more recently are being adopted elsewhere. Recent data indicate that there are over 2,400 drug courts in operation in the United States (Huddleson and Marlowe 2011). Other countries that have adopted drug courts include Canada, the United Kingdom, New Zealand, Australia, South Africa, Bermuda, and Jamaica (Berman and Feinblatt 2005).

Initially started entirely for nonviolent drug-involved adult offenders, the drug court model has grown to include specialized juvenile drug courts and DWI (driving while intoxicated) courts. The drug court model helped spawn a broader movement of therapeutic jurisprudence that includes a range of problem-solving courts, including mental health courts, prostitution courts, domestic violence, and others.

According to the National Association of Drug Court Professionals (1997, see also Hora 2002), there are several key components of a drug court: (1) early identification of eligible offenders; (2) a collaborative, non-adversarial, court processing; (3) regular judicial monitoring in the form of status hearings for drug court clients before the judge; (4) required drug treatment, (5) urine testing; (6) and the use of graduated sanctions for failure to comply with the court requirement and rewards for successful progress toward program completion. The specific treatment requirements are individualized to the needs of each client, and there is variability in how different drug courts implement these key components.

Eligible drug court clients are identified by prosecutors and diverted from traditional prosecution. The drug court client must, however, agree to the diversion to the drug court and the terms of drug court participation. The typical drug court client is a nonviolent drug individual charged with a drug or property offense, although some drug courts do accept individuals charged with or previously convicted of a violent offense. Most drug courts do not accept individuals with drug trafficking offenses. Thus, the ideal is that drug court clients are those individuals for whom drug use is presumed to play an important role in their criminal behavior.

As a non-adversarial judicial process, drug court clients must agree in some form to their guilt. One model is a pre-plea arrangement where no formal plea or other findings of guilt are formally entered but the client does stipulate that the facts of the case are correct, ensuring a successful prosecution if they are unsuccessful in the program. The case is dismissed if the participant completes the drug court. A post-plea/predisposition model requires a formal plea from the participant but delays sentencing. The case is dismissed for program graduates but

sentence is imposed for those that fail. For more serious offenders, a post-plea/post-disposition model may be used. Under this approach, drug court participation is essentially a condition of probation and successful program completion is rewarded with a suspected sentence of incarceration, that is, he or she avoids a jail or prison sentence.

All drug court clients are required to participate in some form of drug treatment. A drug court may refer clients to a broad range of treatment services available throughout their community or may have a working relationship with a single treatment provider. The intensity of treatment varies and, ideally, is tailored to the needs of the client.

Drug court clients appear regularly before the judge for status hearings. These status hearings are crucial as it is here that the drug court judge and clients converse directly, and it is in these hearings where judges in collaboration with other court actors most clearly use the authority of the court to motivate treatment compliance. The court uses various rewards (e.g., praise, tokens of achievement, movement to the next phase of the program) and sanctions (e.g., increased treatment attendance or urine testing, short jail stays) to compel compliance to program requirements. Clients advance through three or more progressively less intense stages before completing the drug court, which typically takes about 12- to 18-months. Successful completion of the drug court is acknowledged with a formal graduation ceremony.

An early review of the evidence on drug courts conducted by the US General Accounting Office (GAO 1997) was unable to arrive at a firm conclusion, either for or against, regarding their effectiveness in reducing crime and drug use. In contrast, a 2001 review by Belenko drew a cautious but positive conclusion regarding the effectiveness of drug courts based on 37 evaluations. Both of these reviews, however, comment on the lack of evidence on outcomes measured post program. That is, most of the studies examined drug use and criminal behavior during the period of the drug court. Thus, it was not clear whether the reductions in drug use and criminal behavior would continue after graduation from a drug court.

Recent reviews have been more favorable. The GAO updated its review and concluded that drug courts were effective, at last during the program period, at reducing criminal behavior (U.S. General Accountability Office 2005). Furthermore, cost-benefit evaluations examined in this report also showed that drug courts yield a net benefit to society. The meta-analysis by Wilson et al. (2006) of 55 drug court effectiveness studies tentatively concluded that drug court participants have lower rates of recidivism than similar offenders who did not participate in drug courts. These findings held for evaluations that measured recidivism during and after program participation. Like the earlier reviews, these findings were tempered by the generally weak methodological rigor of the evaluations.

Evidence of Effectiveness of Drug Courts

This entry summarizes the findings of a recent meta-analysis and systematic review on the effectiveness of drug courts by Mitchell et al. (2011) published by the Campbell Collaboration (www.campbellcollaboration.org). This meta-analysis is an update of an earlier meta-analysis by the same authors (Wilson et al. 2006). This review examined the evidence of the traditional adult drug court and two newer adaptations of this model: juvenile drug courts and DWI drug courts.

Summary of Methods

The focus of the meta-analysis by Mitchell and colleagues (2011) was to synthesize all evaluations of the effectiveness of drug courts that used an experimental or quasi-experimental design and examined drug use or criminal behavior as an outcome. The evaluations must have compared a drug court (adult, juvenile, or DWI) to a comparison condition, such as probation. Studies that simply compared program graduates with program failures were excluded.

Mitchell and colleagues implemented a search strategy designed to find all published and unpublished evaluations that satisfied the

Drug Courts' Effects on Criminal Offending, Table 1 Mean odds ratio by type of recidivism measure

Outcome	Mean	95 % CI		Q	k^a	Tau^2
		Lower	Upper			
Adult drug courts						
General recidivism	1.66	1.50	1.84	442.19*	92	0.178
Drug recidivism	1.70	1.39	2.08	323.98*	42	0.368
Drug use	1.45	0.92	2.28	15.78*	4	0.165
Juvenile drug court						
General recidivism	1.37	1.15	1.63	66.31*	34	0.105
Drug recidivism	1.06	0.69	1.63	29.65*	14	0.357
Drug use	1.50	0.67	3.34	2.05	3	0.359
DWI drug court						
General recidivism	1.65	1.35	2.02	78.40*	28	0.159
Drug recidivism	1.59	1.22	2.09	16.84	14	0.054
Drug use	1.87	0.34	10.23	5.02*	2	1.227

*Distribution heterogeneous ($p \leq 0.05$)
[a]Number of odds ratios

above criteria. Their approach included keyword searching on numerous bibliographic databases and culling the reference list of prior reviews and eligible studies. In addition, websites of several organizations involved in promoting drug courts or organizations that conduct evaluations of drug courts were searched. The search was updated through August of 2011. This process yielded 181 studies that reported the results of 154 independent evaluations of drug court programs. See Mitchell and colleagues (2011) for a complete list of these references.

Results across studies were converted to odds ratios and these served as the basis for the meta-analysis. The odds ratio reflects the odds of success (or failure) in one group compared to another. In this case, the odds of success in the drug court versus the comparison. These were coded such that odds ratios greater than one reflected less criminal behavior or drug use among the drug court participants relative to the comparison condition and odds ratios less than one more criminal behavior. That is, values greater than one favor the drug court whereas values less than one favor the comparison.

Odds ratios were averaged across studies using the inverse variance weight method assuming a random effects model. This model gives greater weight to larger studies and assumes that the odds ratios vary both as a result of study level sampling error and true study level differences.

Summary of Results

Of the 154 evaluations included in the Mitchell and colleagues meta-analysis, 92 were evaluations of adult drug courts, 34 were of juvenile drug courts, and 28 evaluated DWI courts. Only 8 of these 154 evaluations were of drug courts outside of the United States (the other countries represented in the review included Australia, Canada, Guam, and New Zealand). The vast majority of these evaluations were relatively recent and have been conducted since 1999.

Mitchell and colleagues examined three different outcomes across the 154 studies: (1) general recidivism (typically measured as rearrest for any offense), (2) drug-related recidivism (typically measured as rearrest for a drug offense), and (3) drug use (usually measured as self-reported drug use or via urinalysis), for each type of court. Table 1 reports the mean odds ratio for each of these outcome types by each court type. All of these mean odds ratios favor the drug court over the treatment group, although the effect for the outcome of drug recidivism for juvenile drug courts is near a no different value. The pattern of evidence clearly favors the drug courts with most studies observing

positive effects. For the general measure of recidivism, the percentage of studies observing positive effects was 88 % for adult drug courts, 70 % for juvenile drug courts, and 85 % for DWI courts.

These effects are easier to interpret if they are translating into simple percentages. For example, the mean odds ratio of 1.66 for general recidivism for the adult drug courts translates into a reduction in recidivism from 50 % (a common rate in the comparison conditions) to 38 % for the drug court participants. This is a small to moderate effect by most standards but arguably one of practical and policy significance. The comparable percent success rates for general recidivism for juvenile courts and DWI courts are 42 % and 38 %, respectively. All three of the mean odds ratios for the most general recidivism measure reported by each study are statistically significant.

An effect of drug courts on future criminal behavior is presumed to be mediated by reductions in drug use and drug dependence. That is, the core logic of drug courts is a presumed connection between drug use and more general criminal behavior. Thus, it is surprising that roughly half of the evaluations of drug courts did not report results specific to drug use. For both adult drug and DWI courts, the results for drug-specific measures of recidivism (rearrest for drug-related offenses) are quite similar to those for general measures of recidivism. The more direct measures of drug use, such as self-report or urinalysis, also produce roughly similar results for these two court types. The latter effects were not statistically significant; however, given the small number of studies contributing to the mean (and correspondingly large confidence interval), this result is not surprising.

In contrast, the pattern of mean odds ratios for juvenile drug courts across the three outcome types is counter-intuitive. The mean odds ratios are positive and statistically significant for general recidivism and more direct measures of drug use but essentially a value of no difference for drug-related measures of recidivism. There is not clear explanation for this pattern other than the smaller number of studies evaluating drug use outcomes.

Prior reviews of the drug court literature were critical of the high percentage of evaluations that only examined effects during the period of program enrollment. To address this concern, Mitchell and colleagues examined effects separately for different follow-up periods and clearly differentiated effects that overlapped either partially or completely with the period of program enrollment.

More specifically, they first compared results measured during the first, second, third, and fourth years post entry into the program using all available studies at each time period. Next they examined only those studies reporting results at 12- and 24-months post program entry and also those studies reporting results at 12-, 24-, and 36-months post program. This was only done for adult drug courts as there was insufficient data for these analyses for the juvenile and DWI courts. The analyses showed that the effects of adult drug courts on recidivism are not limited to the short term. Rather, the research suggests that drug court participants have reduced recidivism during and after drug court treatment, and these effects appear to last at least 3 years post-drug court entry.

Although the overall findings are encouraging and suggest that drug courts are generally effective, the findings were highly heterogeneous across studies. That is, there was great variability in the size of the effects across the different studies. This variability may be due to the numerous methodological differences across the studies or due to actual difference in the structure and design of the drug court.

Mitchell and colleagues examined the influence of numerous methodological features on study outcomes. Of greatest concern was whether selection bias might account for the overall positive findings. Because the vast majority of the studies included in this meta-analyses were quasi-experimental (i.e., did not randomly assign participants to the drug court and comparison conditions), it is plausible that naturally occurring differences between the groups might upwardly (or downwardly) bias the effectiveness

estimates. For example, a selection mechanism resulting in the drug court program having more highly motivated individuals would result in a positive odds ratio even if the treatment was not effective. The evaluations included in this meta-analysis included studies with a low likelihood of selection bias, that is, experimental studies with random assignment to conditions, and studies with a high likelihood of selection bias, that is, studies with included drug court eligible offenders who elected not to participate.

Studies were grouped into four categories. The most rigorous category was randomized experiments. The second was high quality quasi-experimental designs. These rigorous quasi-experiment studies typically used subject-level matching on key variables or propensity score matching to ensure comparability of clients in each condition. The third category was standard quasi-experimental designs. These typically used either a historical comparison group that met drug court eligibility criteria constructed from archival data or a group of offenders who were eligible but not referred to the drug court program. The critical feature here is that the participants did not self-select into the drug court or comparison condition. The fourth and weakest category of studies was quasi-experimental designs that involved comparing drug court clients to drug offenders who were eligible for the drug court but declined participation (drug court refusers).

The evidence suggests that the overall mean results are at least somewhat positively biased. The weakest quality studies did, on average, produce the largest results across the three types of courts. Thus, it is important to focus more closely on the findings from the higher quality studies.

For adult drug courts, Mitchell and colleagues concluded that the more rigorous studies still provided evidence of general effectiveness. This finding was complicated by the inconsistent results across the three randomized experimental studies of adult drug courts. They concluded, however, that all three experimental evaluations of adult drug courts provide evidence of these courts' effectiveness in reducing recidivism. Two of the three evaluations find recidivism reductions in the first year after program entry, and the remaining evaluation finds a recidivism reduction at 3 years post-program entry. Furthermore, the vast majority of the rigorous quasi-experimental evaluations of drug courts find moderate reductions in general recidivism.

For juvenile drug courts, the picture is more complicated. The average effects for the 11 rigorous quasi-experimental studies and for the single experimental study were positive. Unfortunately, these effects were small overall (ranging from a mean odds ratio of 1.22 to 1.39 for general and drug-specific measures of recidivism) but not statistically significant. Thus, the most rigorous evidence suggests that juvenile drug courts have small effects but clearly additional research is needed to firm-up this conclusion.

For DWI drug courts, the rigorous quasi-experimental studies had a large and statistically significant mean odds ratio for both general recidivism and drug-specific recidivism (1.99 and 1.78, respectively). Unfortunately, the mean odds ratio for the four randomized studies was very small and statistically nonsignificant. However, Mitchell and colleagues noted that three of these four evaluations reported positive effects but that a single influential negative finding seriously attenuated the overall mean odds ratio. Thus, Mitchell and colleagues concluded the evidence of effectiveness for DWI courts was encouraging but that clearly additional high quality studies are needed.

As already discussed, drug courts vary substantially in how they implement the various key elements. Mitchell and colleagues explored numerous differences in the structure of the drug courts across studies and whether these differences were related to the observed effects. These analyses were guided by a framework proposed by Longshore et al. (2001). This framework has five dimensions and predicts that the most effective drug courts (1) will use the court's leverage (rewards and sanctions) to motivate offender change, (2) will serve populations with less severe problems, (3) will have high program intensity, (4) will apply rewards and

sanctions predictably, and (5) will emphasize offender rehabilitation as opposed to other court goals like speedy case processing and punitive sanctioning.

Unfortunately, it was not possible to categorize the studies on all of these dimensions. Mitchell and colleagues were able to code variables that related to the first three dimensions but not the last two. Furthermore, the categorizations were often rather crude and were unable to cleanly differentiate the studies on each dimension.

The first dimension is the use of sanctions and rewards or what Longshore and colleagues refer to as leverage. To examine leverage, Mitchell and colleagues coded studies by the method of case disposition (pre-plea, post-plea, or mixed) and what happens to the charges/sentences upon graduation (dismissed/expunged or not dismissed/expunged). Mitchell and colleagues found mixed support for the leverage hypothesis. Consistent with the hypothesis, drug courts that dismiss/expunge charges upon graduation had higher mean odds ratios on both measures of recidivism. This difference is statistically significant for the drug recidivism measure, but is nonsignificant for the general recidivism measure. Counter to their hypothesis, however, drug courts that predominantly used post-plea case processing did not have greater reductions in recidivism than other courts. Post-plea courts are presumed to have greater leverage because offenders have been convicted and face immediate sentencing upon failure in drug court whereas the in the pre-plea model, the case must still be processed.

In practice, what may matter most are not these structural differences but the certainty and swiftness of sanctions and rewards. Mitchell and colleagues were not able to categorize courts on this dimension based on the information provided in the evaluation reports. Evidence supporting the importance of the certainty and swiftness of sanctions is emerging from studies of Project HOPE [see related encyclopedia entry], a variant on the drug court model. The key feature of Project HOPE is the certainty and swiftness of sanctions, typically a night in jail, for failure to remain drug free (i.e., a dirty urinalysis) or adhere to a simplified set of terms and conditions of probation.

The second dimension of Longshore and colleagues framework is the intensity of the program. Mitchell and colleagues categorized the studies on several variables designed to address this dimension. These included the length of the program, frequency of drug testing, frequency of status hearings, and number of phases through which the clients progressed.

The analyses did not generally support the hypothesis that more intense programs are associated with greater reductions in recidivism. The one exception was for analyses involving the number of status hearings. For the adult drug courts, courts with more than two status hearings in first treatment phase had larger effects of drug-related recidivism than other courts. The number of status hearings was also positively related to improved outcomes for juvenile drug courts. Other than the number of status hearings, these findings suggest that drug courts with greater program intensity are no more effective in reducing recidivism than other courts. Once again, however, these analyses do not fully capture the differences in the intensity of the drug treatment provided by the various drug courts, nor do they capture the differences in the quality of drug treatment services.

The third dimension of Longshore and colleagues' framework posits that drug court programs that serve less severe populations will be most effective. This hypothesis is in conflict with a core element of Andrews and colleagues' (1990) principle that more effective programs serve more severe populations. This "risk principle" has found empirical support (see Andrews and Bonta 1992). Mitchell and colleagues examined this issue by categorizing studies by whether individuals with violent criminal histories or with extensive prior convictions were eligible for the program.

Analyses of the general recidivism effect sizes support Longshore and colleagues' prediction. Specifically, samples that included only nonviolent offenders had statistically larger mean odds ratios on the general recidivism

measure than samples that included violent offenders. Similarly, samples with minor criminal history had larger mean odds ratios than evaluations based on samples with more criminal history; however, this difference was not statistically significant. These findings were not replicated in the analysis of the drug-related recidivism outcome measure. Here there were no differences in the mean effects odds ratios on either of the measures of population severity. For the juvenile and DWI courts, there were no meaningful patterns across these categorizations. Thus, these findings are more supportive of Longshore and colleagues' hypothesis than Andrews and colleagues' risk principle although the finding is clearly weak and inconsistent across the different types of drug courts.

Key Issues and Controversies

The rapid expansion of drug courts has been based on a presumption that they are effective and on the local problems that they solve, such as jail overcrowding. The approach also resonated with the get-tough approach to crime that has dominated crime policy during that past quarter century. Drug courts also help address the problem of increasing prison and jail populations by diverting a portion of drug-involved offenders toward community sanctions and treatment. Thus, the drug court model satisfied both get-touch and rehabilitation ideologies, allowing for a broad base of support that facilitated adoption of the approach.

This general belief in the value of drug courts is not without detractors. For example, a recent report by the Justice Policy Institute (Walsh 2011) questioned our "addiction" to drug courts. The report does not question the basic effectiveness of drug courts but argues that there are unintended consequences to our use of them as a solution to what is essentially a public health problem: drug addiction. First, the report argues that drug courts have expanded the involvement of the criminal justice system into the lives of low-level offenders, or what is often called, net widening. Many of the individuals who enter a drug court program would have had their cases dismissed or been diverted to a community treatment program prior to the development of drug courts. Second, the report argues that criminal sanctions for individuals who fail in the drug court are often greater than what they would have received through conventional processing. Third, the report argues that a criminal conviction should not be a criterion for low-income and other disadvantages groups to gain access to needed drug treatment. And finally, the report argues that these negative consequences often affect minority groups more severely, both because they are less likely to be eligible given higher rates of offense histories that preclude eligibility and through higher failures rates for those admitted. The report argues for strengthening community-based treatment services for drug-involved individuals that do not require a criminal conviction and diverting many of the existing drug court clients to these community programs.

These criticisms of drug courts are specific to drug courts within the United States and reflect the broader ecosystem within which they operate. Addressing these criticisms would require larger changes within this ecosystem, primarily a major policy shift from a criminal justice approach to drug addiction to a public health approach. In the absence of such a shift, these courts are likely to continue to serve as an important means of facilitating access to drug treatment and improving treatment compliance through judicial sanctions.

Conclusions

The meta-analysis by Mitchell and colleagues draws generally positive conclusions regarding the effectiveness of drug courts for reducing criminal behavior and drug use. The vast majority of the programs that evaluate the effectiveness of these programs find that participants have lower rates of recidivism than nonparticipants. The analysis suggests that the typical effect is roughly a reduction from 50 % to approximately 38 % for adult drug courts.

For juvenile drug courts, the effects appear to be smaller, roughly a reduction from 50 % to 43.5 %. The findings for DWI courts are promising but not unambiguous, given the mixed findings and limited number of studies.

The findings are not without caveats. The number of randomized controlled trials (true experiments) is limited. These high quality studies do indicate, however, that drug courts can work. The high variability in effects across studies suggests differential effectiveness across variations in drug courts. That is, some courts may be highly effective whereas others may be ineffective. The meta-analysis by Mitchell and colleagues explored several hypotheses regarding differential effectiveness and found limited support for differential effects based on the leverage of the court. Additionally, there was weak support that drug courts will be more effective with less severe populations.

Despite the rather large number of studies examining the effectiveness of drug courts, this body of literature provides limited insight into the critical components of effectiveness. It is likely that the quality of the drug treatment services provided to drug court clients and the degree to which these treatments are evidence based are related to program's success. Furthermore, the findings from deterrence research in criminology and from behavioral psychology would imply the importance of certainty and swiftness in the application of judicial sanctions and rewards. The collection of studies reviewed by Mitchell and colleagues provides limited information on the role of these important features of the drug court model. Clearly, future research on the relationship between drug court elements and effectiveness is needed.

Related Entries

- ▶ History of Substance Abuse Treatment
- ▶ Problem-Solving Courts
- ▶ Reentry Courts
- ▶ Rehabilitation

Recommended Reading and References

Andrews DA, Bonta J (1992) The psychology of criminal conduct. Anderson, Cincinnati

Andrews DA, Zinger I, Hoge RD, Bonta J, Gendreau P, Cullen FT (1990) Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. Criminology 28(3):369–397

Berman G, Feinblatt J (2005) Good courts: the case for problem-solving courts. New Press, New York

Hora PF (2002) A dozen years of drug treatment courts: uncovering our theoretical foundation and the construction of a mainstream paradigm. Subst Use Misuse 37:1469–1488

Huddleston CW III, Marlowe DB, Casebolt R (2008) Painting the current picture: a national report card on drug courts and other problem-solving court programs in the United States. National Drug Court Institute, Washington, DC

Longshore D, Turner SW, Morral A, Harrell A, McBride D, Deschenes E et al (2001) Drug courts: a conceptual framework. J Drug Issues 31(1):7–26

Mitchell O, Wilson DB, Eggers A, MacKenzie DL (2011) Drug courts' effects on criminal offending for juveniles and adults. Campbell Syst Rev. http://www.campbellcollaboration.org

National Association of Drug Court Professional (1997) Defining drug courts: the key components. Bureau of Justice Assistance, Washington, DC

U.S. General Accountability Office (2005) Adult drug courts: evidence indicates recidivism reductions and mixed results for other outcomes. United States General Accounting Office, Washington, DC

U.S. General Accounting Office (1997) Drug courts: overview of growth, characteristics, and results. United States General Accounting Office, Washington, DC

Walsh Natasssia (2011) Addicted to courts: how a growing dependence on drug courts impacts people and communities. Justice Policy Institute, Washington, DC. http://www.justicepolicy.org

Wilson DB, Mitchell O, MacKenzie DL (2006) A systematic review of drug court effects on recidivism. J Exp Criminol 2(4):459–487

Drug Dealing Places

- ▶ Spatial Perspectives on Illegal Drug Markets

Drug Distribution

- ▶ Drug Trafficking

Drug Enforcement

Nicholas Corsaro
School of Criminal Justice, University of Cincinnati, Cincinnati, OH, USA

Synonyms

Drug markets; Law enforcement; Policing

Overview

The US Office of National Drug Control Policy (ONDCP) has consistently allocated between US $12 billion to U2003S$17 billion per year to fund the Federal Drug Control Budget since the turn of the century (ONDCP 2011). In addition, the cumulative costs of drug law enforcement for federal, state, and local governments has been estimated to exceed US$35 billion annually (Boyum and Reuter 2005). These budgetary figures illustrate a persistent commitment for the continued "war on drugs" waged by different agencies within the law enforcement community to investigate, build cases, arrest and ultimately incarcerate individuals involved in illicit drug distribution.

Given this clear enterprise, it is important to understand the different practices and procedures utilized by drug law enforcement officials, including: (a) investigation techniques, (b) street-level enforcement, and (c) problem-oriented policing strategies designed to disrupt drug markets. Within this chapter, a variety of different drug enforcement approaches will be discussed in detail. In addition, potential drug law enforcement effectiveness and crime reduction impact will be examined within each sub-section, where appropriate and applicable.

Drug Investigations

In the past 20 years plus, there has been a shift away from reactive (i.e., unfocused and reactionary) policing toward more proactive police tactics. Proactive policing typically refers to self-initiated activities during uncommitted patrol time. A cornerstone role of drug law enforcement is to identify and eventually build cases against illicit drug offenders. A common practice of drug investigators is to engage in covert operations through undercover policing, and to rely on the use of confidential informants. In this section, these tactics and different policing approaches will be discussed in extensive detail.

Girodo (1985) notes that undercover investigations are typically either strategic or tactical in nature; strategic investigations are typically referred to as undercover probes that are used to target a group or geographic area in addition to standard surveillance activity. Tactical investigations target specific individuals, seek evidence regarding their intentions to commit a specific crime, and are geared toward eventual prosecution. In this case, the main purpose of conducting an extensive undercover drug investigation is to gather convincing evidence in order to meet the required standard of evidence (i.e., proof beyond reasonable doubt) needed for a conviction at trial (Sherman 1987, p. 88). In terms of a time-line, Girodo (1985) notes that undercover drug investigations can occasionally last a day or two, but most often occur over several weeks to a few months; and, in some cases can last years.

However, the use of undercover surveillance is not always clear-cut and certainly not without its controversies. Gary Marx (1988) illustrated that undercover policing has both intended and unintended consequences. Among other points of concern, Marx (1988) laments that individuals targeted by an investigation can be deceptively provoked to commit a crime. The use of coercive undercover techniques (i.e., coercion or temptation) by law enforcement has the potential to erode public confidence in police, particularly when they are used to target the politically affluent or those from marginalized communities. Thus, the short-term benefits of a single arrest will most likely not outweigh lingering doubts about procedural unfairness. Many have argued the lack of control over undercover policing should promote more stringent regulations.

Ultimately, legitimacy through the rigorous case-building and procedurally fair (i.e., uncorrupt) investigations that promote public safety and foster crime prevention is the ideal balance, which serves as the primary goal of drug investigations.

Marx (1988) also illustrated that effectiveness in undercover policing can be viewed as a paradox. Success or failure can be defined both as an arrest (because in this case a crime has been detected and justice pursued) as well as a non-arrest (i.e., the absence of an arrest may indicate a deterrent effect). Thus, evaluating the overall efficiency and utility of undercover investigations is an arduous task.

Another important aspect of investigative work by drug law enforcement is the use of confidential informants (CIs), which are usually individuals that are networked within the local drug trade. CIs typically assist police in their investigations, often in exchange for leniency. James Q. Wilson (1968) noted that the use of CIs often centers on the following primary functions: providing leads, casework facilitation, and occasionally testifying in court. Certainly, a great deal of research exists regarding the tactical components of covert drug investigations (see Miller 1987). Related specifically to illegal drug investigations, CIs are considered the 'sine qua non' of narcotics work.

In terms of outlining a framework of drug investigations, the primary function of CIs is to assist investigations by providing detailed information about the nature of drug market networks, which usually requires at least some intimate knowledge of the relationship between users and dealers. In a study that relied upon interviews with 24 undercover officers, Jacobs (1997) outlined a "contingent tie" perspective (i.e., perceptual bonds that may be either weak or strong depending on context and history between individuals) that can be used to explain the social-link between users and dealers, as well as police with CIs. In essence, the reliance on CIs as an investigative tool requires multiple steps and social interactions between drug investigators and potential CIs.

Jacobs (1997) also illustrated that police often first attempt to establish these contingent ties by purchasing drugs directly from potential lower-level distributers (e.g., pretending their 'usual' dealer was unable to provide their normal service and a single purchase would serve as a favor to the undercover officer). Then, investigators use threats of a potential arrest and likely conviction to encourage cooperation among these primary actors in order leverage them to serve as an intermediary informant with others involved in drug market networks. In addition, fictitious relationships between officers and CIs usually need to be established, which usually fall under one of three forms: jailhouse buddies (e.g., police and CIs pretend to know one-another from jail or prison), co-workers (e.g., acting as though they were introduced to one-another through work), or subcultural associates (e.g., conveying they have been long-term offending "colleagues").

A failure to protect CIs (i.e., "burning informants") as part of their investigation makes it virtually impossible to replenish this vital fact-finding resource. Jacobs (1997) noted that drug investigators will likely "cut out" (i.e., remove) CIs from their investigation by complaining about unfair (and fictitious) taxation directly to dealers. For example, undercover officers often ask dealers to engage in future transactions directly with them in order to "cut out" the mid-level person – who happens to also be the CI. This type of deception is used to capitalize on the CIs relationship with others involved in drug networks without putting them in the middle of the investigation, past the onset point. In addition, Manning (1977) illustrated that the reliance on CIs is usually temporary and requires a continued commitment given the high-turnover, and replacement found in illicit drug networks. Thus, law enforcement officials must be invested perhaps more in protecting rather than exploiting their relationship with CIs, particularly over the long-term.

Ultimately, investigations are a preliminary step in terms of initiating law enforcement driven approaches to disrupt illicit drug networks and open-air drug markets. While some have argued there should be additional constraints placed on the process of investigations, most researchers regard undercover investigations and the use of confidential informants a vital component in

drug investigations. After a thorough inquiry is conducted, the next traditional step for police officials is to engage in street-level enforcement, which can take multiple approaches and has a variety of sensitive issues that are typically considered.

Street-Level Drug Enforcement

In a systematic review of over 130 street-level drug enforcement strategies, Mazerolle et al. (2007) synthesized their findings into a finite number of approaches that do not necessarily require interagency cooperation between police and other criminal justice organizations (i.e., those that tended to be "police-only"), some of which are discussed in detail here, including: drug seizures; crop eradication; crackdowns; raids; and search and seizures.

Drug seizures are often referred to as "supply-reduction" strategies that generally center on reducing trafficking, dealing, and the amount of drugs available in a specific geographic area. The seizure of drugs can take place during a routine investigation (described above) or can occur as part of a larger intervention or suppression effort. While evaluative research on seizures of general drugs is rather scant, a majority of evaluations have focused on the impact of heroin-based seizures. Wood et al. (2003) reported a non-significant association between police-led heroin seizures and price, availability, drug-related deaths (including overdoses), as well as overall rates of crime and arrest. Little empirical research exists demonstrating a tangible relationship between drug supply reduction strategies and crime associated with illegal drug markets. This is not surprising given that five-sixths of the crime and violence associated with illegal drug distribution is motivated primarily by drug money while only one-sixth is motivated by drugs themselves (see Caulkins and Reuter 1998). Thus, supply reduction strategies should also be evaluated, in part, on their periphery influence regarding financial increases within drug markets brought forth by street-level enforcement.

Another supply-reduction approach is crop eradication. Research on crop eradication has primarily focused on the removal, confiscation, and destruction of cannabis. Overall effectiveness is difficult to measure, primarily given the findings by Potter, Gaines, and Holbrook (1990) which illustrates how growers and distributers often adapt to the increased attention by police (e.g., moving cannabis plants from outdoors to indoors). Cost-effectiveness is also difficult to determine given the extensive amount of time devoted to most crop reduction strategies.

To this point, the drug law enforcement approaches discussed here have generally been restricted to investigations, undercover work, the use of CIs, and supply-reduction strategies. However, street-level drug enforcement can also involve both undercover and uniformed officers, and thus are usually very highly visible to the surrounding community. This is particularly true concerning the use of police crackdowns and raids. Sherman (1990, p. 1) defined a police crackdown as "a sudden increase in officer presence, sanctions, and the threat of apprehension either for specific offenses or for all offenses in specific places." In offense-specific crackdowns, police focus on a particular type of activity, such as public drunkenness, prostitution, and drug dealing. Location-based crackdowns are aimed at reducing crime in a targeted area. Again, Mazerolle et al.'s (2007) review of the evaluation literature shows that the vast majority of drug enforcement crackdowns generally target street-level drug markets (particularly focusing on "harder" drugs such as crack, cocaine, and heroin); however, some crackdowns also focused on drug distribution within indoor locations – including both residential and commercial sites (e.g., motels/hotels, clubs, etc.).

A synthesis of evaluation-based research suggests that crackdowns produce widely varying results in terms of altering drug and crime problems. Roughly 1/3 of the evaluations indicated a positive effect (i.e., a reduction in drug problems), 1/3 demonstrated no discernable effect, and 1/3 actually showed evidence of a negative effect (i.e., an unintended increase in drug problems). Of those studies that illustrated a reduction

in drug problems, the following were consistent themes observed across the evaluations: (a) the reductions in crime were more likely observed with property and violent offenses rather than drug-specific incidents; (b) the impact appeared to be short-term; and, (c) potential effects were more likely observed in geographically contained areas near bridges, rivers, and borders where displacement was less likely to occur due to natural barriers (Mazerolle et al. 2007).

While crackdowns are described as a sudden increase in police presence, raids are equally intensive but are usually directed within specific indoor locations (e.g., a single residence within an apartment complex or a single bar). Like crackdowns, police raids appear to have a short-term impact on reducing the supply of drugs as well as crime and disorder problems. Cohen, Gorr, and Singh (2003) showed that police enforcement activities that target drug dealing in nuisance bars (i.e., drug hotspots) have the capacity to have a substantive impact on levels of drug dealing, at least during those periods of active raid-based enforcement. In the Cohen et al. (2003) study, there were distinct and active periods of police raids in nuisance bars conducted by the narcotics squad, followed by back-off periods. The authors' note that drug dealing in street markets that are associated with nuisance bars show considerable resiliency after raid and suppression strategies are withdrawn by law enforcement. However, over reliance on raids and crackdowns also has the potential to alienate even the most cooperative residents within high-crime areas, which can have long-term repercussions from a community crime perspective.

Problems with Intensive Street-Level Drug Enforcement

The continual use of police crackdowns, raids, and enhanced searches and seizures in order to control drug related problems in high-crime areas often comes at a tangible cost. Personal contact with police is perhaps the most important factor that shapes individuals' perceptions of law enforcement. Cao et al. (1996) also illustrated that community context plays a major role in terms of residents' appraisals of law enforcement. Citizens in disadvantaged and high-crime areas are often skeptical, concerned, and in most cases resistant to potentially intrusive police practices.

In terms of the intersection between race and community context, Weitzer (1999) found that Blacks and Whites in middle-class neighborhoods were less likely to perceive police to be abusive or to engage in enforcement-based misconduct than were Black residents in lower-class neighborhoods. In many cases, police seem to behave differently depending on the type of neighborhood they patrol. Terrill and Reisig (2003) found police were more likely to use force in disadvantaged neighborhoods on suspects. Similarly, Kane (2005) found police were more likely to engage in misconduct in distressed communities. Instances of over-policing (i.e., aggressive enforcement) and under-policing (i.e., lack of concern) can erode public confidence in formal mechanisms of social control.

However, research also consistently indicates a desire for integrative law enforcement responses to crime and drug problems in high-risk geographic contexts. In a study of the perceptions of Black residents in high-crime areas, Brooks (2000) found a paradox in that many African-Americans residing in poor neighborhoods and who were disproportionately affected by high-crime problems often associated with illicit street drug markets were frustrated with levels of crime, desired tough legal enforcement, but were also fearful of police. Carr et al. (2007) similarly found that youths from high-crime communities who were negatively disposed to police also were overwhelmingly in support of increased and tougher law enforcement approaches. Finally, in an interesting twist, Cooper et al. (2005) conducted a study that relied upon interviews with 40 drug-injecting residents and found that "study participants lauded the reductions in public drug activity that constant [police] monitoring produced" (Cooper et al. 2005, p. 681). Thus, there appears to be a constant tension in terms of maintaining a balance between levels of enforcement that are perceived appropriate by local residents contrasted with efforts that are viewed as excessive and unwarranted.

To illustrate, rapid response time by police are largely unrelated to on-scene arrests (some contend fewer than 29 for every 1,000 cases). However, expedient citizen reporting time significantly influenced the likelihood of on-scene arrests. Thus, outcome-driven processes that substantively improve police effectiveness are more important than simply focusing on traditional organizational police processes.

Integrative policing approaches that rely upon a diverse set of strategies as well as neighborhood residents to assist in the performance of their duties are also more likely to stimulate cooperation when citizens' view the police favorably and with legitimacy. Strategies that promote positive pro-social aspects of neighborhood integration appear to have a mediating effect on targeted neighborhood crime outcomes. In the next section, more extensive detail is provided for problem-oriented policing applications that have shown considerable promise in reducing drug related incidents.

Problem-Oriented Policing (POP) Strategies

Beginning in the 1980s, there was a groundbreaking shift in policing strategies that saw the advent of new approaches to law enforcement, including hotspots policing (i.e., strategically focusing on small geographic places to reduce high levels of crime), community policing (i.e., law enforcement approaches that promote community engagement) and problem-oriented policing (i.e., problem solving through the use of a variety of analytical approaches). Concerns with effectiveness and evaluation are consistent with the problem-oriented policing (POP) framework introduced by Herman Goldstein (1990).

Morris and Heal (1981) also advocated the use of "situational policing" as a potential approach to ensure the assimilation between various communities interests with police-led crime prevention activities. The POP model calls for police to use scanning, analysis, response, and assessment (SARA) in order to determine and define crime problems (i.e., root-causes) rather than reacting to crime incidents (i.e., symptoms). Goldstein (1990) provided an "inventory" of actions that he argued should be employed by law enforcement interested in adopting the problem-oriented model. Some key components of Goldstein's (1990) inventory included: concentrating on high-call locations; integration with additional government and private agencies; use of mediation (rather than simply using arrest-only); mobilization and collaborating with the local community; using civil and criminal law to control public nuisances; and, opportunity reduction (see also Clarke, 1997).

Eck (2006) more recently framed problem-oriented policing around three major principles: empirical, normative, and scientific. The empirical principle states that the public demands police handle a diverse range of crime problems, and that they are not invested in specifically how police handle such problems (i.e., arrest need not be the only "tool" available to police). The normative principle illustrates that police should address and ultimately reduce crime problems rather than simply respond to incidents. Finally, the scientific principle asserts that police should use scientific evidence to identify and address core crime problems. In essence, the tools officers utilize to address crime problems such as illegal drug distribution should be analytical, unique, and crafted specifically to each problem.

Mazerolle et al. (2006) conducted a rigorous meta-analytic review of street-level drug law enforcement strategies. The study assessed the relative effects of different approaches by comparing problem-oriented policing (outlined above) with community-wide policing, hotspots policing, standard (i.e., unfocused) law enforcement efforts. Effect size comparisons (i.e., changes in targeted outcomes) indicated that problem-oriented policing programs and geographically-based interventions that involved cooperative partnerships between police and third parties (e.g., community members, social service providers, additional government agencies, etc.) tended to be far more effective at controlling drug problems than traditional and law-enforcement only interventions. Mazerolle et al. (2006) concluded that policing initiatives

that forge cross-agency and cross-community partnerships had the most pronounced and sustained effect on drug crime problems.

Recent development in POP drug enforcement strategies have focused on the use of "pulling levers" focused deterrence initiatives to disrupt the crime and violence associated with street-based drug markets. Pulling levers policing relies on actors across multiple agencies (i.e., police, prosecution, probation and parole, social service providers, and community leaders) to use data-driven approaches in order to target repeat and chronic offenders within high-crime locations (Goldstein 1990). Pulling levers strategies are designed to utilize specific deterrence in order to inform chronic and persistent offenders of the sanctions that are specifically available to criminal justice officials that can be used to obtain leverage and ultimately reduce recidivism (Kennedy 1997).

The initial pulling levers policing strategy that used this type of focused deterrence framework occurred in Boston where citywide levels of youth homicide significantly declined by roughly 66 % after implementation (Braga et al. 2001). A number of sites have replicated the approach, and a series of evaluations have provided further evidence of a significant violent crime impact. While the majority of these strategies and subsequent evaluations have focused predominantly on reducing youth, gang, and gun violence, officials have recently begun adapting the approach to disrupt levels of crime and violence associated with street level drug markets (Kennedy 2009).

Officials in High Point, North Carolina were the first to use the problem-oriented policing model while also incorporating a pulling levers intervention in order to disrupt neighborhood crime problems facilitated by local drug markets. More specifically, they combined the threat of enhanced sanctions to targeted individuals that were identified through a detailed drug investigation in order to emphasize the seriousness of their message through the use of a public neighborhood call-in session. In addition, the public communication strategy was designed to demonstrate interagency and community cooperation and encourage improved social service provisions for desisting offenders in an effort to permanently disrupt several open-air drug markets across the city (Kennedy 2009). Initial results from High Point indicated significant declines in drug and violent offending as well as perceptional changes among residents in target areas (Kennedy 2009). Replications of the High Point drug market program evaluated in Nashville, TN (Corsaro et al. 2010) and Rockford, IL (Corsaro et al. forthcoming) also demonstrated similar promise. Thus, the use of specific deterrence combined with detailed drug investigations appears to warrant additional attention as a potential mechanism to disrupt open-air drug markets. However, we also note that POP strategies are not always universally successful.

Problems with Problem-Oriented Policing

While a variety of promising POP strategies were presented here in terms of impacting drug related crime, it is also important to note that in practice POP has a number of limitations that should be carefully considered. Eck (2006) contends that problem-oriented policing was originally designed to be held within police headquarters, though in reality decentralization occurred where problem-identification and subsequent problem-solving tasks were shifted to police officers within operational units.

In addition, Cordner and Biebel (2005) conducted an organizational study of POP in practice within an organization that appeared to be devoted to the POP model. Study results indicated that patrol officers often engage in activity that could be defined as problem-solving; however, these same patrol officers tended to concentrate their efforts on symptoms of greater neighborhood problems (e.g., drug deals) and indeed very rarely formally used the SARA model in order to disrupt central community crime problems (e.g., personal crime, property crime, traffic violations). Cordner and Biebel (2005) found that while patrol officers engaging in problem-solving were thoughtful and analytical in their various approaches to specific neighborhood crime problems, they rarely engaged in a full range of POP activities since these require extensive data analysis, integrative and dynamic

approaches, and evidence-based evaluations. In essence, the true POP model requires more time, effort, and energy than many officials are willing to invest.

Conclusion

Boyum and Reuter (2005) estimated that the imprisonment risk per drug transaction in the USA was about 1 in 4,500 for drug sellers that average roughly 1,000 transactions annually. In addition, they calculated the risk of imprisonment for distributing 0.2 g units of cocaine to be less than 1 in 15,000 per sale for distributers. When coupled with the combined state and federal annual budget averages that range between an estimated $30 and $35 billion per year, objective readers will likely question the efficiency, practicality, and sustainability of current criminal justice approaches to street level drug law enforcement.

In this chapter, a detailed review of drug enforcement strategies, history, processes, and evaluations of effectiveness were presented. Given that illicit drug networks and open-air drug markets appear to be unique across different locations, there does not seem to be a single catch-all drug law enforcement approach that solves the laundry list of drug crime problems. However, proper steps and continual commitment geared toward drug investigations, street level enforcement, community cooperation, and long-term commitment to solving drug crime problems has shown promising results.

In essence, effective POP strategies are built upon the total range of drug enforcement strategies discussed here. Investigations are a single step in the process; the use of CIs helps build stronger cases; suppression policing directed at specific individuals or within specified locations can have a short-term impact; inter-agency cooperation helps drive reductions in drug related crime problems; community notification sessions are designed to improve police-community relationships; and community informal social control helps sustain potential positive drug crime reduction gains. When the cumulative efforts of these drug enforcement strategies are present when policing illegal drug markets, the evidence suggests that crime reduction benefits are more likely to occur.

Related Entries

- ▶ Problem-Oriented Policing
- ▶ Spatial Perspectives on Illegal Drug Markets

Recommended Reading and References

Boyum D, Reuter P (2005) An analytic assessment of U.S. drug policy. AEI Press, Washington, DC

Braga AA, Kennedy DM, Waring EJ, Piehl AM (2001) Problem-oriented policing, deterrence, and youth violence: an evaluation of Boston's operation ceasefire. J Res Crime Delinq 38:195–226

Brooks RW (2000) Fear and fairness in the city: criminal enforcement and perceptions of fairness in minority communities. South Calif Law Rev 73:1219–1275

Cao L, Frank J, Cullen FT (1996) Race, community context and confidence in the police. Am J Police 15:3–22

Carr PJ, Napolitano L, Keating J (2007) We never call the cops and here is why: a qualitative examination of legal cynicism in three Philadelphia neighborhoods. Criminology 45:445–480

Caulkins JP, Reuter P (1998) What price data tell us about drug markets. J Drug Issues 28:593–612

Clarke RV (1997) Situational crime prevention: successful case studies. Harrow and Heston, Albany

Cohen J, Gorr WL, Singh P (2003) Estimating intervention effects in various settings: do police raids reduce illegal drug dealing at nuisance bars? Criminology 41:257–292

Cooper H, Moore L, Gruskin S, Krieger N (2005) The impact of a police drug crackdown on drug injectors' ability to practice harm reduction: a qualitative study. Soc Sci Med 61:673–684

Cordner GW, Biebel EP (2005) Problem-oriented policing in practice. Criminol Public Policy 4:155–180

Corsaro N, Brunson RK, McGarrell EF (2010) Evaluating a policing strategy intended to disrupt an illicit street-level drug market. Eval Rev 34:513–548

Corsaro N, Brunson RK, McGarrell EF (forthcoming) Problem-oriented policing and open-air drug markets: examining the rockford pulling levers deterrence strategy. Crime Delinq

Eck JE (2006) Science, values, and problem-oriented policing: why problem-oriented policing? In: Weisburd D, Braga A (eds) Police innovation: contrasting perspectives. Cambridge University Press, New York, pp 117–132

Girodo M (1985) Health and legal issues in undercover narcotics investigations: misrepresented evidence. Behav Sci Law 3:299–308
Goldstein H (1990) Problem-oriented policing. McGraw-Hill, New York
Jacobs BA (1997) Contingent ties: undercover drug officers' use of informants. Br J Sociol 48:35–53
Kane RJ (2005) Compromised police legitimacy as a predictor of violent crime in structurally disadvantaged communities. Criminology 43:469–498
Kennedy D (1997) Pulling levers: chronic offenders, high-crime settings, and a theory of prevention. Valparaiso Univ Law Rev 31:449–484
Kennedy D (2009) Deterrence and crime prevention: reconsidering the prospect of sanction. Routledge, New York
Manning PK (1977) Police work: essays on the social organization of policing. MIT Press, Cambridge, MA
Marx GT (1988) Undercover: police surveillance in America. University of California Press, Berkeley
Mazerolle L, Soole DW, Rombouts S (2006) Street-level drug law enforcement: a meta-analytical review. J Exp Criminol 2:409–435
Mazerolle L, Soole D, Rombouts S (2007) Drug law enforcement: a review of the evaluation literature. Police Q 10:115–153
Miller GI (1987) Observations on police undercover work. Criminology 25:27–46
Morris P, Heal K (1981) Crime control and the police, vol 67, Home office research study. Home Office, London
Office of National Drug Control Policy (2011) ONDCP available at http://whitehousedrugpolicy.gov/publications/policy/11budget/exec_summary.pdf. Accessed on 25 July 2011
Potter G, Gaines L, Holbrook B (1990) Blowing smoke: an evaluation of marijuana eradication in Kentucky. Am J Police 97:97–116
Sherman LW (1987) Reinventing probable cause: target selection in proactive investigations. J Soc Issues 43:87–94
Sherman LW (1990) Police crackdowns: initial and residual deterrence. Crime Justice Rev Res 12:1–48
Terrill W, Reisig MD (2003) Neighborhood context and police use of force. J Res Crime Delinq 40:291–321
Weitzer R (1999) Citizen perceptions of police misconduct: race and neighborhood context. Justice Quart 16:819–846
Wellford C, Pepper JV, Petrie C (2005) Firearms and violence: a critical review. The National Academies Press, Washington, DC
Wilson JQ (1968) Varieties of police behavior. Harvard University Press, Cambridge, MA
Wood E, Tyndall M, Spittal P, Li K, Anis A, Hogg R et al (2003) Impact of supply-side policies for control of illicit drugs in the face of the AIDS and overdose epidemics: investigation of a massive heroin seizure. Can Med Assoc J 168:165–169

Drug Marketing

▶ Drug Trafficking

Drug Markets

▶ Drug Enforcement

Drug Substitution Programs and Offending

Ambros Uchtenhagen[1] and Nicole Egli Anthonioz[2]
[1]Research Institute for Public Health and Addiction, University of Zurich Medical School, Zurich, Switzerland
[2]School of Criminal Sciences, University of Lausanne, Lausanne, Switzerland

Synonyms

Drug-related crime; Heroin-assisted treatment; Opioid replacement treatment; Opioid substitution treatment

Overview

The prescription of opioids as a therapy for opiate dependence has a long history, and opioid substitution treatment is one of the best documented and researched therapeutic approaches in medicine. It is considered a cornerstone in the worldwide efforts to help opiate addicts and to relieve the burden for afflicted societies, mainly through a significant reduction of HIV infections in heroin injectors and the population at large and through a significant reduction of acquisitive crime by heroin users.

However, "fighting fire with fire" has always met strong opposition. Many only accept abstinence as an outcome of treatment. The research

evidence of a substantial reduction of opiate-related crime through substitution treatment can facilitate widespread political acceptance by demonstrating the benefits of this approach for the societies as well as for addicts and their families.

This entry has two parts: the first describes the historical development of this treatment approach, and the second provides the international research evidence on the reduction in opiate-related crime associated with substitution treatment based on a systematic review of eligible studies comparing substitution treatment with diverse control groups. This entry examines the benefits of methadone, buprenorphine, and heroin substitution maintenance.

A Short History of Opioid Maintenance Treatment

Introduction

Maintaining opiate addicts on opiates has a long history. Opium was one of the most effective medications in ancient medicine, and its widespread use frequently led to chronic use and dependence. Medical and nonmedical rationing practice was introduced to satisfy the needs of those who were unable to discontinue its use. The discoveries of morphine and heroin and of intravenous injections were followed by even more widespread medical use and consequently concerns about chronic dependence. The idea of prescribing injectable opiates as a substitute for street heroin started in USA and was abolished on the basis of prohibitionist legislation, while it continued as regular practice in the UK. A new approach to maintaining opiate addicts on substitution therapy was initiated in the USA in 1963, with the prescription of oral methadone in the framework of a comprehensive treatment program. This approach slowly found increasing acceptance and is nowadays considered a cornerstone in the management of opiate dependence and for the prevention of HIV/AIDS in opiate injectors. The concept of heroin maintenance treatment was reactivated in order to reach out to treatment resistant heroin addicts. Based on the unanimously positive outcomes, heroin maintenance has become routine treatment for otherwise untreatable heroin addicts in Switzerland, the Netherlands, Denmark, and Germany.

The Origins: The Exceptional Power of Opium

Prescribing and using opiates as an effective medication for a range of ailments goes back to the origins of Chinese, Sumeric, and Egyptian medicine, and probably beyond (the first traces are known from Neolithic burial sites). In Homeric times, opium in wine was offered against all evils, and in ancient Egypt, sponges soaked in opium were used during surgery. Authoritative Roman and Arabic texts have praised the uses of opium throughout the Middle Ages, and the opium-made "theriak" was used as a panacea for many centuries, in spite of prohibitive warnings by the Church. Opium tincture (laudanum) was still considered to be the most essential drug at the dawn of modern medicine by Paracelsus. It is inevitable that its addictive potential became manifest, but we do not know when and where opiate addiction was first observed and when the use of opiates as a maintenance regime started. However, opiate dependence must soon have been recognized to be difficult to overcome (Seefelder 1996).

Medical and Nonmedical Maintenance by Rationing Systems

Maintenance was the logical answer to the failure in overcoming dependence. We do know that the Roman emperor Marc Aurel was maintained on opium by the eminent physician Galenus and that the Mugal emperor Jahangir received opium maintenance from his chief physician. And far into the twentieth century, opium dispensing to registered opium addicts was practiced in some Asian countries like Pakistan, providing obviously dependent people with their daily dose. This system was abolished after the adoption of the 1961 UN Single Convention.

A comparable system of an alcohol rationing system (Bratt system) was practiced in Sweden for many years. It became unpopular as a "paternalistic" interference of state and was abolished after the Second World War.

Morphine Maintenance

Opium is rich in alkaloids with different characteristics. They were identified successively and developed further synthetically. Morphine, distilled from opium in 1804, was a revolution in pain management and found widespread use especially after the invention of syringes for injection. Once heroin (diacetylmorphine) was developed in 1874, it was again used as an effective medication for many conditions, frequently prescribed as "patent medicines." In Europe, it was also considered to be a cure for morphine, cocaine, and opium dependence, until its own addictive potential became known. In the USA, 44 narcotic clinics were set up, most of them after the Harrison Act, which left the dependent persons without supply, and introduced tapering off morphine for detoxification purposes. By then the maintenance concept was reinvented, based on the frequency of relapses ("Tennessee system," 1914). While the southern clinics, treating mainly iatrogenic morphine addicts, were quite successful, the New York clinics failed in maintaining young disintegrated heroin addicts (Musto 1999; Mino 1990).

Prohibition Against Maintenance

All narcotic clinics, successful or not, were closed by law in 1923, while a few doctors continued to prescribe heroin to addicts until the Single Convention of 1961, when heroin became a controlled substance and its use restricted to scientific purpose. The concept of maintaining patients on an otherwise illegal or unwelcomed consumption was incompatible with the Puritan idea of prohibition.

The one country resisting the temptation of a strict prohibition in Europe was the United Kingdom where the Rolleston Committee in 1926 recommended heroin prescribing for chronic addicts (which at the time were mainly socially integrated patients who became dependent on morphine after treatment for pain management).

The Revival of the Maintenance Model: "Fighting Fire with Fire"

It was again in the USA where the maintenance treatment of opiate addicts was reintroduced, using methadone, a synthetic opioid, instead of morphine or diamorphine, on the basis of its advantages (oral application, longer half-life). According to the needs of the new target population – mainly socially disintegrated urban heroin injectors – a comprehensive health and social support program went along with methadone prescribing in contrast to just handing out prescriptions for unsupervised use (Dole and Nyswander 1965).

The positive effects of this new approach, especially on the health and criminal behavior of patients, confirmed by an increasing number of evaluation studies, led to a growing acceptance of the maintenance concept, in spite of "abstinence-only" arguments and opposition. But the decisive factor to speed up methadone (and buprenorphine) maintenance was the advent of the HIV epidemic. Drug injecting was recognized as a major factor for transmitting the viral infection (for hepatitis C as well). Throughout Europe, the idea of maintaining heroin addicts on opioids became increasingly acceptable, instead of restricting treatment to detoxification and abstinence-oriented approaches. The number of countries providing methadone maintenance increased from 7 in 1980 to 28 by 2005, and the number of countries providing buprenorphine maintenance rose within a few years to 21 (EMCDDA 2006). In 2008, there were ca. 670,000 patients in substitution treatment (EMCDDA 2010). The former main objective to reduce criminality and to curb the illegal heroin market was replaced gradually by a major public health concern (WHO/UNODC/UNAIDS 2004). Finally, in 2006 the World Health Organization succeeded in putting methadone and buprenorphine on the list of essential medicines, based on evidence about the safety and effectiveness of these substances.

Opioid maintenance treatment is the predominant treatment option for opioid users in Europe. It is generally provided in outpatient settings, though in some countries it is also available in inpatient settings and is increasingly provided in prisons. Opioid maintenance is available in all EU Member States. In most countries, specialized public outpatient services are the main providers of substitution treatment. However, office-based general practitioners, often in shared care arrangements with specialized centers, play an increasing role

in the provision of this type of treatment (EMCDDA 2010). The only other psychotropic substance where maintenance is practiced is tobacco, using nicotine replacement patches.

Outside of Europe, we see a different picture. By 2009, opioid maintenance treatment was available in 70 countries, but only an estimated 8 % of drug injectors received it (Mathers et al. 2010). The reasons are manyfold: most of the research evidence comes from developed countries in the West, and injecting addicts are discriminated against as morally or legally deviant; therefore only abstinence is considered to be the legitimate goal of interventions, even at the cost of infringing on human rights and medical ethics. Compulsory nonmedical reeducation camps are still frequent in some Asian countries (WHO 2009).

The Quest for Prescribing the "Original Drug": New Research, New Practice

The increasing number of methadone patients inevitably has led to an increasing but less important number of "methadone-resistant" patients who continued to inject heroin in spite of adequate methadone dosages and care. As the systematic review of evaluations will show, methadone treatment by itself did not provide entirely satisfactory results. At the same time, the HIV epidemic made it a priority to increase coverage, i.e., to reach out to as many injectors as possible. Thus, prescribing heroin as the original and preferred substance of addicts was proposed and has been tested.

Even if the AIDS epidemic is at the origin of all initiatives to prescribe heroin in Switzerland, in the Netherlands, in Germany, in Canada, and in Spain, the idea of heroin maintenance was ready for a revival. In the UK, where the number of heroin addicts had increased and their characteristics changed, drug dependence clinics for maintenance treatment were opened in London in the 1960s, and notification of any addicts receiving a heroin prescription had to be provided to the Home Office. The much debated objectives were medical care and, at the same time, social control of addicts. For many reasons, heroin prescribing was more commonly replaced by methadone prescribing, until the AIDS epidemic led to a reconsideration and to new experimentation with injectable and smokable heroin, now in a perspective of public health interests (Strang and Gossop 1996).

Another debate started in Canada in the 1950s and again in the early 1970s, in the Netherlands and in the USA during the 1970s. The arguments were mainly focusing on heroin addiction as a chronic condition and the need to restrict its negative health and social consequences. Nonmedical options for providing heroin to addicts in the Netherlands (tolerated "home dealers" and "heroin bars") were started but then considered to be failures (Mino 1990).

Feasibility studies on heroin prescribing were initiated in Australia. Large-scale experimental studies were finally set up in Switzerland (1994–1996), the Netherlands (1995–1997), Germany (2002–2005), Spain (2003–2004), Canada, and the UK. The Swiss cohort study provided the first positive outcomes, acknowledged by an international expert committee set up by WHO (Ali et al. 1998). As they could not separate the effects of heroin prescribing from the effects of concomitant care, the other countries set up randomized controlled trials, thereby adding essential new findings to the already established positive outcomes (see below).

A common characteristic of the new experimental studies was the aim to cover heroin addicts unable to profit from other treatments including methadone maintenance, and the provision of pharmaceutical heroin in the framework of a comprehensive assessment and treatment program. In order to avoid overdose and diversion, the intake of injectables must be made under visual supervision of staff in the clinics. In Switzerland, the Netherlands, and Germany, heroin maintenance has reached the status of a routine treatment and is paid by health insurance. Belgium decided to start another trial, and Denmark has set up heroin maintenance without further research.

Benefits and Risks of Opioid Maintenance: A Summary Statement

A summary of findings confirms the feasibility of maintenance treatment in terms of patient

satisfaction and widespread acceptance, the safety for patients and staff, the significant improvements in somatic and mental health, and a reduction in risk behavior and illicit drug use, in drug-related crime and in public nuisance offenses (Uchtenhagen 2003, WHO 2009). Maintenance therapies are increasingly integrated into treatment and care systems without negative effects for other approaches. Main risks are overdose mortality in the introductory phase of methadone maintenance and the diversion of prescribed opioids to the illegal market; they can be avoided by appropriate regimes. Economic studies have evidenced superior benefits compared to costs.

How the Treatment's Effect on Crime Was Evaluated Across the World

Methods

Under the umbrella of the Campbell Collaboration Crime and Justice Group, a systematic review was carried out (Egli et al. 2009) in order to investigate the effect on crime of substitution programs, be it substitution by methadone or heroin, as opposed to any other type of treatment (in particular, abstinence, detoxification, psychotherapy) or no treatment at all, as well as comparative effects between different substitution therapies. All primary references used in the systematic review can be found in the Campbell Collaboration publication (Egli et al. 2009); they are identified in the current text by the first authors' name and the year, without a space between.

A search strategy for the identification of studies was defined and followed. Relevant studies were identified through abstracts, bibliographies, and databases such as Medline, the National Criminal Justice Reference Service (NCJRS), Harm Reduction Journal, Journal of Substance Abuse Treatment, National Treatment Agency for Substance Misuse (NHS), and National Treatment Outcome Research Study (NTORS), as well as the bibliographies of relevant reviews. In order to be eligible, a study had to fulfill the following criteria:

- At least one of the study groups had to undergo a substitution program (using, e.g., methadone and/or opiates as substitution drugs).
- A measure of re-offending had to be given, since only such effects were included in the meta-analysis, as opposed to medical or social outcomes.
- Level 4 or higher on the Sherman scale had to be met.

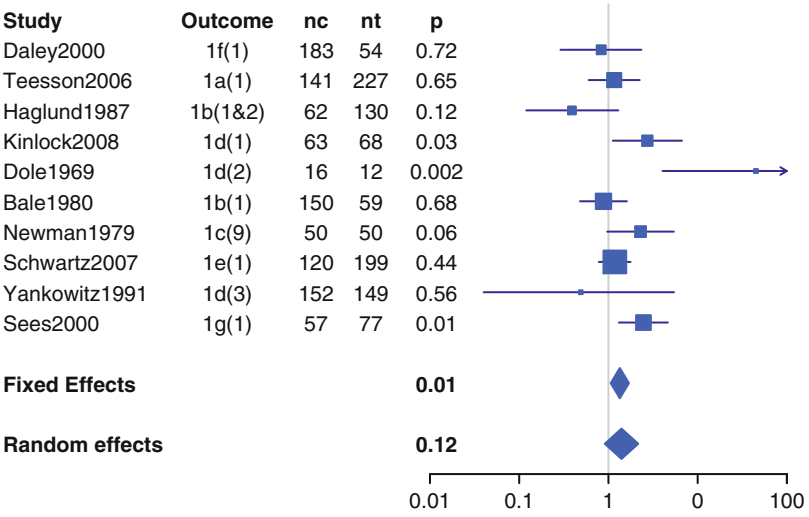
The methods used, be it for review or for the analysis of the data, are as specified in Practical Meta-Analysis by Lipsey and Wilson (2001); odds ratios have been used for measuring the effects observed. An odds ratio greater than 1 stands for a reduction in the outcome measure (e.g., a positive effect with respect to the response variable). For the computation of mean effect sizes, the inverse variance method of meta-analysis (Lipsey and Wilson 2001) was used.

Illustrations (forest plots) have been created using R (www.r-project.org). Also, in these illustrations, the outcomes have been coded as follows: the criminal behavior (1: any offense; 2: Property crime; 3: theft) followed by the measure used (a: commission; b: arrest; c: conviction; d: incarceration; e: illegal income; f: cost of crime (estimated); g: ASI legal; h: charged) and finally the way it has been measured, in parentheses (1: self-report; 2: official sources; 3: known status; 9: unknown).

Results

Methadone Maintenance Treatment (MMT)

Seven RCTs (randomized controlled trials) and three high-quality quasi-experimental studies (i.e., level 4 on the Sherman Scale) were found for MMT. The control groups differed widely: three were wait-list control designs (Dole 1969; Schwartz 2007; Yancovitz 1991), one used a placebo control (Newman 1979), in one the control group received counselling (Kinlock 2008), in three others the control condition received detoxification (Daley 2000; Haglund 1978; Sees 2000), in one the control received treatment in the community (Bale 1980), and one received residential treatment (Teesson 2006). The results for these studies are shown in Fig. 1. Not surprisingly, the distribution is



Drug Substitution Programs and Offending, Fig. 1 Descriptors and forest plot for studies fulfilling criteria 4 or 5 of the Sherman Scale where treatment group receives MMT and control groups receive no substitution treatment. The size of the boxes in the forest plot is proportional to the weight of the study in the summary measures. The confidence interval of Dole (1969) is shortened for representation (*arrow*). *nt* is the number of subjects in the treatment group, while *nc* is the number of subjects in the control group. Outcomes have been coded

heterogeneous ($p < 0.05$), indicating meaningful variation in effect across these studies. The random effects model does not indicate a significant effect of methadone maintenance with respect to these control groups; the mean effect measure is, however, in favor of MMT. Overall effects for the four different types of control groups have also been computed; none of these show a significantly greater effect of the maintenance treatment over the treatment in the control groups. All of these effects grouped by control treatment are, however, in favor of MMT.

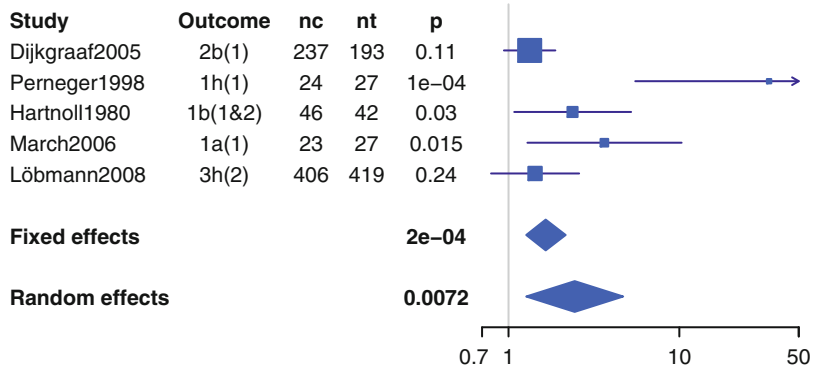
Effects of Heroin Substitution Treatment

Six studies of heroin substitution programs were found, five of which were RCTs. In four of the RCTs, the control group underwent methadone maintenance (Dijkgraaf 2005; Hartnoll 1980; Löbmann 2008; March 2006), while in one (Perneger 1998), the control group underwent a conventional treatment. The results are shown in Fig. 2.

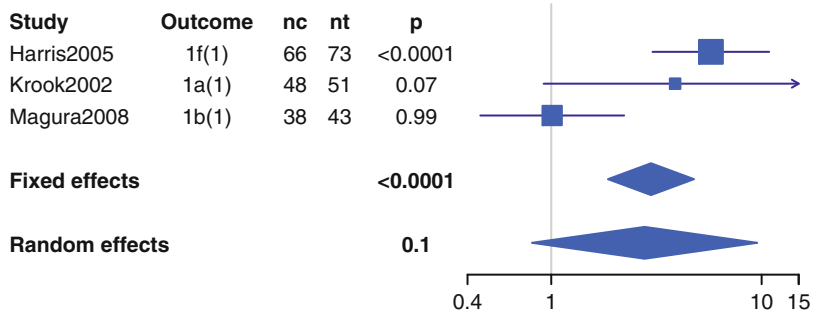
A very large effect was obtained for the study by Perneger (1998), along with a very large confidence interval. Both the small sample size and the variety of control treatments explain these two observations. Overall, for these RCTs, homogeneity is rejected ($p = 0.004$). If Perneger (1998) is not included in the analysis, due to the different treatment of the control group with respect to all other included studies, homogeneity is accepted ($p = 0.21$). The fixed effects mean effect size is then 1.55 [1.18; 2.02] ($p = 0.0015$). Here, a significant decrease in the criminality measures is, therefore, present for heroin over methadone maintenance treatment. However, in all of these trials except Dijkgraaf (2005), a selection effect is present that is more or less pronounced. In order to be admitted into these studies, subjects had to have followed (and failed) a program before (this selection effect is strongest for the study March (2006), where subjects were required to have had MMT at least twice before admission). Therefore, these results are overall more properly interpreted as showing a positive effect of HMT over MMT in subjects having already failed at a treatment, including MMT.

Effects of Buprenorphine Substitution Treatment

Three randomized controlled trials concerning the effect of buprenorphine maintenance on



Drug Substitution Programs and Offending, Fig. 2 Descriptors and forest plot for studies where treatment groups receive heroin maintenance and the control group either MMT or another standard treatment. The size of the boxes is proportional to the weight of the study in the summary measures. The confidence interval for the study by Perneger (2008) has been cut for representation (*arrow*). The outcomes have been coded. *nt* and *nc* are defined as above



Drug Substitution Programs and Offending, Fig. 3 Descriptors and forest plot for studies where treatment group receives buprenorphine maintenance. The confidence interval for Krook (2002) has been cut on the right (*arrow*) for representation purposes. The outcomes have been coded and *nt* and *nc* are defined as above

criminal activities were found. In two of these studies, the control group was MMT (Harris 2005; Magura 2008), while in the third (Krook 2002), the control group received a placebo. The individual and overall effects are shown in Fig. 3. The homogeneity analysis rejects homogeneity ($p < 0.05$). The overall effect is positive, but not significant ($p = 0.10$). Only one of the three studies has a significant odds ratio, the study by Harris (2005), which shows the largest effect size, and is also the largest of the three studies.

When the study with a differing control group (Krook 2002) is excluded, homogeneity is still refuted (Q test $p = 0.0008$). Overall, there is no significant reduction in criminality when buprenorphine instead of methadone is used, although the findings suggest a slight advantage for buprenorphine relative to methadone (or placebo).

Discussion

Results show that MMT has a greater effect on criminal behavior than non-maintenance-based treatments, but not significantly so. Furthermore, the comparison between this maintenance treatment and heroin maintenance showed a significant advantage for heroin maintenance; this is true particularly for persistent users who have participated in a program before and failed. Buprenorphine maintenance also shows an advantage over alternative treatments, but one of these alternative treatments was placebo; over MMT, one study shows

a significant advantage for buprenorphine, while a second study shows almost equivalent effects.

Two systematic reviews of substitution programs have been published by the Cochrane Collaboration: Ferri et al. (2006) and Mattick et al. (2009). While these reviews do not focus on crime as an outcome measure, a comparison of results with the findings from the Egli et al. (2009) review summarized above is relevant. In Mattick et al. (2009), three studies comparing methadone maintenance to no opioid replacement therapy with respect to their effect on criminal behavior are included. The results obtained are similar to the results obtained here in two respects: firstly, the effect of MMT seems to reduce criminal behavior more than the alternatives that do not include maintenance, and secondly, this effect is not significant.

In Ferri et al. (2006), four trials comparing methadone maintenance to heroin maintenance are included. One study showed a reduction in the risk of being charged when on heroin maintenance; this is in line with the results obtained here. Also, two studies considered criminal offending and social functioning in a multi-domain outcome measure, and again, heroin plus MMT yields better results than MMT alone. Again, this is in line with the results obtained here, in that heroin maintenance reduced criminality more than MMT.

Related Entries

- ▶ Drug Abuse and Alcohol Dependence Among Inmates
- ▶ Drug Courts
- ▶ Drug Courts' Effects on Criminal Offending
- ▶ History of Substance Abuse Treatment
- ▶ Randomized Experiments in Criminology and Criminal Justice

Recommended Reading and References

Ali R, Auriacombe M, Casas M, Cottler L, Fasrrell M, Kleiber D (1998) Report of the external panel on the evaluation of the Swiss scientific studies of medically prescribed narcotics to heroin addicts. World Health Organisation, Geneva

Dole VP, Nyswander M (1965) A medical treatment for diacetylmorphine (heroin) addiction. A clinical trial with methadone hydrochloride. J Am Med Assoc 193:80–84

Egli N, Pina M, Skovbo Christensen P, Aebi MF, Killias M (2009) Effects of drug substitution programs on offending among drug-addicts. Campbell Syst Rev 2009:3

EMCDDA (2006) European Monitoring Centre on Drugs and Drug Addiction EMCDDA: annual report 2006. Office for Official Publications of the European Union, Luxemburg

EMCDDA (2010) 2010 Annual Report.http://www.emcdda.europa.eu/online/annual-report/2010

Ferri M, Davoli M, Perucci CA (2006) Heroin maintenance treatment for chronic heroin-dependent individuals: a cochrane systematic review of effectiveness. J Subst Abuse Treat 30(1):63–72

Lipsey MW, Wilson DB (2001) Practical meta-analysis. Sage, Thousand Oaks

Mathers BM, Degenhardt L, Ali H, Wiessing L, Hickmann M, Mattick RP et al (2010) HIV prevention, treatment and care services for people who inject drugs: a systematic review of global regional and national coverage. Lancet 375: 1014–1028

Mattick R, Breen C, Kimber J, Davoli M (2009) Methadone maintenance therapy versus no opioid replacement therapy for opioid dependence (Cochrane Review). Cochrane Database Syst Rev 3

Mino A (1990) Analyse scientifique de la litérature sur la remise contrôlée d'héroïne ou de morphine (Scientific analysis of studies on the controlled prescription of heroin or morphine). Federal Office of Public Health, Bern

Musto DF (1999) The American disease. Origins of narcotic control, 3rd edn. Oxford University Press, New York

Seefelder M (1996) Opium. Eine Kulturgeschichte (Opium. A cultural history), 3rd edn. Ecomed, Landsberg

Strang J, Gossop M (1996) Heroin prescribing in the British system: historical review. Eur Addict Res 2:185–193

Uchtenhagen A (2003) Substitution management in opioid dependence. In: Fleischhacker WW, Brooks DJ (eds) Addiction. Mechanisms, phenomenology and treatment. J Neural Transmission suppl 66:33–60

Uchtenhagen A (2008) Heroin-assisted treatment in Europe: a safe and effective approach. In: Stevens A (ed) Crossing frontiers. International developments in the treatment of drug dependence. Pavilion, Brighton, pp 53–82

WHO (2009a) Assessment of compulsory treatment of people who use drugs in Cambodia, China, Malaysia and Viet Nam. An application of selected human rights principles. World Health Organisation, Geneva

WHO (2009b) Guidelines for the psychosocially assisted pharmacological treatment of opioid dependence. World Health Organisation, Geneva
WHO/UNODC/UNAIDS (2004) Position paper: substitution maintenance therapy in the management of opioid dependence and HIV/Aids prevention. World Health Organisation, United Nations Office on Crime and Drugs, Joint United Nations Programme on HIV/Aids

Drug Trade

► Drug Trafficking

Drug Trafficking

Henry H. Brownstein
NORC at The University of Chicago, Bethesda, MD, USA

Synonyms

Drug distribution; Drug marketing; Drug trade

Overview

In an organic sense, drugs are chemicals. When people ingest drugs, they pass through their body to their brain where they interfere with the neurotransmitters that transfer signals across synapses changing the messages that the brain sends to the body and thereby affecting the way the body works. In strictly economic terms, however, drugs are a commodity. That is, they are a product that is exchanged between people. Some drugs in some nations are socially and legally considered to have value for one reason or another and are deemed acceptable for use by some or all people under some or all circumstances. These are produced, distributed, and consumed under authority of law and sanctioned by less formal social norms for commercial purposes and personal use by approved individuals in approved circumstances. Others drugs are not acceptable for any purpose or any person under any circumstance. Today and for some time now, the production, distribution, and consumption of those drugs that are not authorized by law or sanctioned by social norms have been considered a serious problem for individuals, communities, and nations. There are very real public health concerns about drugs, all drugs but in particular illicit drugs or licit drugs used in illicit ways, and what they do to individuals and their minds and bodies. And there are very real public safety concerns about the impact of drugs on people and their communities and nations, especially with regard to the consequences of the commercial transactions involving those drugs that are not recognized by law. So it is not surprising that for the last century or so, there has been considerable attention among social scientists and public policymakers to questions and concerns about the trafficking and marketing of illicit drugs and their relationship to crime and violence.

Understanding Drug Trafficking and Illicit Drug Markets

As a commodity drugs need to be produced and distributed in order to reach consumers. Among those drugs that are illegal, there are differences in how and where they are produced, and therefore, how and at what scale they are distributed. For example, heroin may be sold to consumers in nations such as the United States as powder but it starts as an agricultural product cultivated in other parts of the world and requires a manufacturing process to convert the plant product into the form preferred by consumers and a distribution process that crosses national boundaries. Methamphetamine, on the other hand, can be produced by cooking the required but easily available chemicals at a fixed location with enough equipment to be called a laboratory, or simply by shaking the chemicals together in a 2-L bottle while sitting in the back of a moving car. Cocaine starts as a plant grown in places like South America, but by the time it reaches consumers in the United States, it is in the form of

a powder manufactured in a laboratory setting or as little rocks known as crack that are made almost anywhere from a small amount of the powder for small quantity sales. So a broad commercial perspective is helpful for a general understanding of the trafficking and trade in illicit drugs.

Industrial economists distinguish *industry* from *markets*. Whereas industry refers to groupings of individual *businesses* that share common techniques and processes for the production of a certain commodity, markets refer to groupings of the *consumers* of the product of that industry (Andrews 1949). When the product is illicit drugs, whatever specific drug it is, the markets of consumers for those drugs are inextricably linked to the production and distribution techniques and processes and the organization and operation of the illicit drug industry. For example, people who purchase heroin from a local dealer in their own community are nonetheless connected to the individuals and organizations that are involved in the international, national, and regional agricultural, manufacturing, and trafficking operations that are necessary to produce heroin and bring it to market.

The illicit drug industry and its related markets have been studied from a number of perspectives. Looking at drugs as a commercial enterprise, economists have studied the illicit drug industry for one or more drugs and their various markets in terms of things like price and purity (e.g., Caulkins 2005; Caulkins and Reuter 1998; Rhodes et al. 2000). Public health researchers have focused on the impact of drug trafficking and markets on communities in terms of issues related to things like drug-related morbidity and mortality and drug treatment and prevention (e.g., Curtis et al. 1995; Sommers et al. 2006). Criminologists have studied the illicit drug industry and markets in terms of issues related to public safety, criminal activity, and violence (e.g., Goldstein et al. 1992; Weisburd and Mazerolle 2000). Overall the questions raised by this body of research ask how is the industry and how are the markets organized and how do they operate? How large are they in terms of production and in terms of economic value? Specifically in terms of trafficking, given that the product is illegal, how does it get from producers to consumers? And what is the impact of the trafficking and marketing of illicit drugs on people, communities, and nations?

Drug Trafficking

From the founding of the United States as a nation through to the late 1800s, it was common for home remedies to be freely available to citizens for dealing with personal problems of pain and illness. David Musto, a historian of drugs in America, called the late nineteenth century a period of "wide availability and unrestrained advertising" (1991, p. 42). During this period, the production and distribution of home remedies was not regulated by government, and many of these remedies contained palliative substances such as opium or stimulating substances such as cocaine (Inciardi 2007). Then in the early twentieth century, concerns for health and safety problems believed to be associated with their use resulted in the passage in 1906 by the United States Congress of the Pure Food and Drug Act, which imposed quality and packaging standards on all food and drug products (Inciardi 2007). Following that in 1914, the Harrison Narcotics Act was passed and in 1937 the Marijuana Tax Act, and together they gave the federal government some measure of regulatory control over the production and distribution of drugs (Musto 1999). Unfortunately while these government actions through taxation regulated the production and distribution of drugs that were earlier found in popular home remedies, they did not address the consumer demand for those drugs. So the illicit drug industry and markets filled the void to meet the demand.

As the illicit drug industry grew, the government of the United States increasingly became concerned about the impact of these drugs on public health and safety. In the middle of the twentieth century, there was not sufficient scientific evidence to know what that impact really was, but there were strong beliefs and opinions and those were used to support the direction of

public policy toward illicit drug use and users and trafficking and markets. In 1971, President Richard Nixon formally declared war on drugs (Inciardi 2007). With his declaration, attention shifted from drug users to drug trafficking, and in 1973, Nixon created the Drug Enforcement Administration (DEA) as a component of the United States Department of Justice specifically charged with fighting the war on drug trafficking. Later President Ronald Reagan reaffirmed the war on drug trafficking, and then President George H.W. Bush with Congress passed the Anti-Drug Abuse Act of 1988 thereby forming the Office of National Drug Control Policy (ONDCP) as a policy agency to lead the national drug strategy. With ONDCP, the United States might finally have a Drug Czar, but the focus on drug trafficking as the problem and drugs as the enemy started with Nixon's declaration of war and his formation of the DEA to fight that war.

While the United States was intensifying its war effort against drugs and drug traffickers, not all nations followed suit. Many others favored a policy of harm reduction emphasizing efforts to manage the harm to drug consumers and their communities from drug use rather than trying to control the supply of drugs through trafficking (Inciardi and Harrison 2000). For example, in the Netherlands, drug policy historically has favored protecting the health of users and reducing their health risks by focusing on programs of prevention and treatment (van Laar et al. 2011). Given a tradition known as "gedoogbeleid" that favors discretion when dealing with drugs and drug users, since the late twentieth century, the Dutch have formally and systematically not enforced laws involving small quantities of cannabis and have even established guidelines allowing certain retail establishments to sell cannabis to consumers without fear of prosecution as long as they adhere to established guidelines, such as not advertising and not selling to minors (MacCoun and Reuter 1997). Under international pressure in recent years, the Dutch government has placed greater limits on how these establishments can operate and who they can serve, but the tradition and policy allow the trade to continue. This does not mean that the government of the Netherlands favors unregulated trafficking or trade in illicit drugs. A recent report issued by the Netherlands National Drug Monitor to the European Monitoring Centre for Drugs and Drug Addiction (van Laar et al. 2011) notes that along with addressing the need to prevent drug use, the needs of drug users for treatment and rehabilitation, the reduction of harms faced by drug users, and lessening any disturbances to public safety they might cause in their communities, combatting drug production and trafficking is one of its main objectives. But the report continues, the "primary aim of Dutch drug policy is focused on health protection and health risk reduction" (van Laar et al. 2011, p. 15). Similarly, in the United Kingdom, a harm reduction approach to drugs has been in place going back to a report issued in 1926, *The Report of the Departmental Committee on Morphine and Heroin* (Bennett 1988).

While sovereign nations appropriately and naturally each have their own policy toward drugs, drug users, and drug trafficking, there has been international cooperation going back to 1909 when 13 nations gathered in Shanghai for the International Opium Commission (Musto 1991). At that meeting, which focused on opium and opiates, no binding decisions were made and no treaties signed. But other meetings followed. In 1911 there was a meeting of 12 nations in Hague resulting in an agreement for each nation to enact legislation to control narcotics trade and to fund educational programs. Then starting in 1961, there was a series of three international Conventions held under the auspices of the United Nations. Through the first two, all previous multinational treaties that had been negotiated from 1912 to 1953 were consolidated, and controls over 100 different substances considered to be narcotic drugs were tightened, and through the third in 1988, a focus was set on international drug trafficking (United Nations 1988).

Despite national and international efforts to control or manage it, today the illicit drug industry is well established on a global scale with active local markets and large and profitable national and international corporate-type

manufacturers and distributors. Systems and procedures are in place to move drugs that are produced in one part of the world to other parts of the world where consumers are waiting to purchase and use them. With the focus on drug trafficking, the question then is what is the nature and scope of international trafficking? What is the impact of the traffic in illicit drugs on both producer and consumer nations and on their citizens?

Since the industry and the markets operate outside of the law, there are no official records of how much is produced or how much is sold or purchased. There are no official records of costs related to production or distribution of any illicit drug product, and no official records of how many consumers there are or what consumers have paid or are willing to pay for their drugs. So statisticians, economists, and other social scientists use available data sources to try to calculate the scope and scale of the illicit drug industry as a whole, and for particular drug industries and markets, such as heroin or cocaine. They use these data to derive estimates of drug consumption and estimates of the supply of drugs. For consumption estimates, they sometimes use data from surveys of samples of people who tell an interviewer about their experience using drugs, or not. For example, in the United States, estimates are derived using data from the National Survey of Drug Use and Health (NSDUH), an annual nationwide household survey produced for the Federal government by the Substance Abuse and Mental Health Services Administration (SAMHSA) and administered to a random sample of 70,000 respondents age 12 and older asking them about their use of and experience with various illicit drugs. Estimates of consumer demand in the United States also come from the Treatment Episode Data Set (TEDS) similarly maintained by SAMHSA, though in this case, counting the drug treatment and using experiences of about 1.5 million people annually admitted to drug treatment.

In addition to survey data, estimates of drug consumption and supply are also derived from available official record data from Federal, state, and local government agencies that have operational involvement in some way or other with acknowledged users of various illicit drugs. These agencies include, for example, law enforcement agencies that arrest drug users and dealers or traffickers, and treatment service providers that work with drug users to try to help them stop using drugs. Law enforcement records, for example, have been used to derive estimates of the supply of drugs based on a variety of law enforcement activities including crimes known to the police and arrests, but also things like the amount of particular drugs seized by various local, regional, and national law enforcement agencies. In the United States, the agency record data for supply estimates comes from national aggregate counts of crime and agency counts of arrests from sources like the Uniform Crime Reports (UCR), an annual national count of crimes known to the police and arrests made by the police. More directly focused on the actual supply of drugs known to be in the country, another estimate comes from data on drugs seized and analyzed in laboratories by law enforcement for the DEA System to Retrieve Information from Drug Evidence (STRIDE) program (NDIC 2011), or the tactical intelligence collected by the El Paso Intelligence Center (EPIC), established by the DEA in 1974 for the collection and dissemination of information related to drug trafficking particularly along the United States border with Mexico.

In summary, what is known about the scope and scale of drug trafficking or even drug using is limited. No official records of activity in the illicit drug industry or drug markets are maintained. All estimates are indirect, derived from proxy measures. They are based on an accounting of the activity of individuals and agencies that in some way work with people who are involved in or with the drug industry. So there is some uncertainty about what is known or believed to be known about the illicit drug industry and drug trafficking.

Illicit Retail Drug Markets

Similarly there are no official records to support what is known or believed to be known about

local drug markets and the retail trade in illicit drugs. However, there has been considerable research conducted over the last few decades on the organization and operation of illicit retail markets at the local level (National Institute of Justice 2003; National Research Council 2010, 2001). There are several ethnographic studies of markets in particular cities, and there are a number of studies that provide sociological, geographical, and economic analyses of market organization and operation. Important findings of these studies include the identification of differentiated roles among buyers and sellers of illicit drugs, the characteristics of social relationships and of structural forms in different local markets, and the variation among patterns of distribution and consumption across places like neighborhoods and by the different types of drugs being transacted.

An illicit retail drug market can be defined as the set of people, facilities, and procedures through which illicit drugs are transferred from sellers or dealers to buyers or users (National Research Council 2001, p. 160). As such they are economic enterprises and therefore operate in response to the forces of supply and demand. But unlike legitimate commercial enterprises, illicit drug markets participants are regularly faced with the an odd mix of rapid turnover and overlapping roles among sellers and buyers, a broad range of product price and quality in a limited geographic area, and an absence of any legitimate authority to settle disputes over things like market share or product quality. Consequently, the natural economic forces of supply and demand in such markets are tempered by the need of buyers and sellers to worry about things like the trustworthiness of the people they are dealing with and the fairness and security of those dealings. By definition in an illicit retail drug market, all transactions are criminal transactions. As a result every individual who participates as a buyer or seller in such a market risks a hostile encounter with law enforcement officers and criminal justice agencies. In addition they risk intimidation, coercion, or victimization by other buyers and sellers over market-related disputes or for just being there.

Drug Markets, Crime, and Violence

Estimates from official statistical data from surveys on drug using and the supply of illicit drugs known to law enforcement along with the evidence of a large body of social scientific research strongly suggest that there is a relationship between illicit drug trafficking and drug markets and crime and violence. For the most part, the research on the connection between drugs and crime and violence has focused on drug use, but there are also studies that directly link crime and violence with drug trafficking and drug markets. In an early study of homicide rates in the United States during the twentieth century, for example, criminologist Margaret Zahn found that the rate of homicide varied over time in relation to the establishment of markets for illicit products with rates of homicide being at their highest in the 1920s and 1930s, during the control of alcohol under Prohibition, and in the 1960s through 1970s, during periods of disruption in the heroin and cocaine markets (Zahn 1980). More recently, studies have found higher rates of crime and violence during periods of disruption in organization and operation of particular illicit retail drug markets. For example, in the late 1980s when crack cocaine markets were newly emerging in urban areas, researchers found that most of the crime and violence associated with those markets involved things like disputes between dealers over territory or disputes between sellers and buyers over the quality of the product sold (Brownstein 1996).

The focus of drug policy in the United States on drug trafficking has been linked over the years to a stated relationship between international drug trafficking and crime and violence. The National Drug Intelligence Center (NDIC) of the United States Department of Justice in its annual threat assessment has regularly reported on violence and crime related to transnational crime organizations in producer nations, notably in Latin America and Asia, and its impact in and on the United States (NDIC 2011). In recent years, concern has focused on the Transnational Criminal Organizations in Mexico with the NDIC reporting that seven such organizations in Mexico "control much of the production,

transportation, and wholesale distribution of illicit drugs destined for and in the United States" and that among them "a dynamic struggle for control of the lucrative smuggling corridors leading into the United States [has resulted] in unprecedented levels of violence in Mexico" (NDIC 2011, p. 7).

Similarly crime and violence have been associated with domestic illicit retail drug markets. In the 1970s in Chicago, Patrick Hughes conducted a groundbreaking epidemiological study of heroin addicts and concluded, "... heroin dealers must have a reputation for violence, otherwise addicts and other deviants would simply take their drugs and their money" (1977, p. 31). Not only did he observe that heroin dealers would use violence in the conduct of their business but also that other participants in the market were at risk of being the perpetrators or victims of crime and violence themselves through confrontations with others including drug users, drunks, and gang members (Hughes 1977, p. 31). A body of research conducted since that time provides additional though inconclusive support for the notion that while drug markets more often than not are peaceful, there is an observable relationship between the organization and operation of illicit retail drug markets and criminal violence (see Reuter 2009).

In 1985, sociologist Paul Goldstein conceptualized the ways that drugs and violence might be related (Goldstein 1985). He suggested that violence might be a consequence of drug ingestion, a response by an addicted drug user needing drugs and having to use force to get them, or the product of the normal organization and operation of the illicit drug trade. The latter he called systemic and suggested it could include things like disputes between drug dealers over claims to territory or among dealers and buyers over the quality of the product being sold. Just about the time that Goldstein was writing and talking about his tripartite framework, the level of violence and crime began to rise dramatically in United States cities concurrent with the introduction of crack cocaine. Public policymakers, criminal justice and law enforcement practitioners, and the news media around the country became alarmed and responded in the belief that the increasing crime and violence could be linked to the use of crack cocaine in city neighborhoods (Brownstein 1996). But studies using the tripartite framework provided preliminary evidence that in fact the growing crime and violence had more to do with market dynamics than with user behavior (Goldstein et al. 1992).

Conclusions and Future Research

In summary, there are limits to what is known and can be known about the illicit drug industry, drug trafficking, and drug markets. To some extent those limits result from the lack of official records available to be used for scientific or policy analyses. This problem limitation is not likely to be overcome in that the industry exists outside of the law. In addition, the research conducted in this area to date largely has followed the personal agenda of individual researchers and research teams rather than being guided by an organized and systematic program specifically designed to guide researchers and their studies in the direction of questions that need to be answered to adequately and effectively inform policy and practice. The unfortunate consequence has been that much of the policy that is made and many of the practices and programs that are initiated are guided by unfounded assumptions or political principles rather than by scientific evidence.

Nonetheless, there are some things that researchers have found that are convincing. The drug industry, with its consumer markets and the trafficking operations that comprise it, is highly organized and does operate not only on the local level but also on a global scale. Depending on the drug (and among illicit drugs and drugs used in illicit ways, there are many distinctions in what they do and how they are obtained and used), in some way or ways, there is a relationship between local retail market outlets for the product, large- and small-scale producers or manufacturers of the product, and the organized enterprises and well-established procedures and practices put into place

to move the product from producers to consumers. And as all this happens outside of the law and legitimate authority, things do not always go well and sometimes the outcome is harmful and a threat to personal and public health and safety.

While there may be one or more studies that have addressed a particular question, it does not necessarily mean that the question has been adequately answered. So questions important for understanding, explaining, and addressing the institution of drug trafficking and any social and personal problems related to it remain open. As noted earlier, among the questions probably worth asking but certainly not all of the questions that need to be asked are the following: How is the industry and how are the markets organized, and how do they operate? How does drug trafficking work? How do different drugs go from producers to consumers? How much of all and different drugs are being produced, and how great is the demand for them? What is the value of illicit drug production worldwide, and what is the value of all drugs being consumed? What is the impact of the illicit drug industry on people, communities, and nations? Are there any positive outcomes? What are the harms? Besides scientific evidence, what if any ethical or moral principles need to be considered when making public decisions about policies and practices related to drugs and drug trafficking? What if any practical limitations are there that may restrict what can or cannot be done about drugs and drug trafficking?

What needs to be done now is not just a continuation of the current practice of individual researchers raising individual questions based on personal interest or public attention and designing studies. What needs to be done is the work by teams of researchers, policymakers, and practitioners to design and develop a broader research agenda, probably international even more than national, to identify and study relevant questions about the drug industry and markets and drug trafficking to then allow policy to be made and practices implemented that will lead to outcomes that benefit and do not harm people, communities, and nations.

Related Entries

- ▶ Drug Abuse and Alcohol Dependence Among Inmates
- ▶ Drug Enforcement
- ▶ History of Organized Crime in Mexico
- ▶ History of Organized Crime in the United States
- ▶ History of Substance Abuse Treatment
- ▶ Italian Mafia
- ▶ Sex Offenders and Criminal Policy
- ▶ Trafficking in Human Beings

Recommended Reading and References

Andrews PWS (1949) Manufacturing business. Macmillan, London

Bennett T (1988) The British experience with heroin regulation. Law Contemp Probl 51:299–314

Brownstein HH (1996) The rise and fall of a violent crime wave – crack cocaine and the social construction of a crime problem. Harrow & Heston, Guilderland

Caulkins JP (2005) Price and purity analysis for illicit drug: data and conceptual issues. Heinz Research. Paper 25. http://repository.cmu.edu/heinzworks/25. Accessed 25 Oct 2011

Caulkins JP, Reuter P (1998) What price data tell us about drug markets. J Drug Issues 28:593–612

Curtis R, Friedman SR, Neaigus A, Jose B, Goldstein M, Ildefonso G (1995) Street-level drug markets: network structure and HIV risk. Soc Netw 17: 229–249

Goldstein PJ (1985) The drugs/violence nexus: a tripartite conceptual framework. J Drug Issues 14: 493–506

Goldstein PJ, Brownstein HH, Ryan PJ (1992) Drug-related homicide in New York: 1984 and 1988. Crime Delinq 38:459–476

Hughes PH (1977) Behind the wall of respect: community experiments in heroin addiction control. University of Chicago Press, Chicago

Inciardi JA (2007) War on drugs IV: the continuing saga of the mysteries and miseries of intoxication, addiction, crime and public policy, 4th edn. Allyn & Bacon, Boston

Inciardi JA, Harrison LD (eds) (2000) Harm reduction – national and international perspectives. Sage, Thousand Oaks

MacCoun R, Reuter P (1997) Interpreting Dutch cannabis policy: reasoning by analogy in the legalization debate. Science 278:47–52

Musto D (1991) Opium, cocaine, and marijuana in American history. Sci Am 265:40–47

Musto D (1999) The American disease, 3rd edn. Oxford University Press, New York
National Drug Intelligence Center (2011) National drug threat assessment 2011. No. 2011-Q0317-001. USDOJ, Washington, DC
National Institute of Justice (2003) Toward a drugs and crime research agenda for the 21st century. Special report. USDOJ, Washington, DC
National Research Council (2001) Informing America's policy on illegal drugs: what we don't know keeps hurting us. National Academy Press, Washington, DC, Commission on Behavioral and Social Sciences and Education
National Research Council (2010) Understanding the demand for illegal drugs. National Academy Press, Washington, DC, Committee on Law and Justice
Reuter P (2009) Systemic violence in drug markets. Crime Law Soc Chang 52:275–284
Reuter P, MacCoun R, Murphy P (1990) Money from crime: a study of the economics of drug dealing in Washington, DC. Drug Policy Research Center, RAND, Santa Monica
Rhodes W, Layne M, Johnston P, Hozik L (2000) What America's users spend on illegal drugs 1988–1998. Office of National Drug Control Policy, Washington, DC
Sommers I, Baskin D, Baskin-Sommers A (2006) Methamphetamine use among young adults: health and social consequences. Addict Behav 31: 1469–1476
United Nations (UN) (1988) United Nations conference for the adoption of a convention against illicit traffic in narcotic drugs and psychotropic substances. http://www.unodc.org/pdf/convention_(1988_en.pdf. Accessed 5 June 2012
van Laar M, Cruts G, van Gageldonk A, van Ooyen-Houben M, Croes E, Meijer R, Ketelaars T (2011) Report to the EMCDDA by the Reitox National Focal Point. Final version as approved by the Scientific Committee of the Netherlands National Drug Monitor (NDM) on the 22nd of December 2010. European Monitoring Centre for Drugs and Drug Addiction
Weisburd D, Mazerolle LG (2000) Crime and disorder in drug hot spots: implications for theory and practice in policing. Police Q 3:331–349
Zahn M (1980) Homicide in the twentieth century United States. In: Inciardi JA, Faupel CE (eds) History and crime. Sage, Beverly Hills

Drug-Related Crime

► Drug Substitution Programs and Offending

Dutch Colonial Police

Marieke Bloembergen and Ellen Klinkers
KITLV, Leiden, The Netherlands

Overview

This entry discusses the development of the Dutch colonial police forces in the late nineteenth and twentieth century in the context of colonial state formation. The differences in the organization of the police forces and the practice of policing in the Dutch colonies (the Dutch East Indies, Suriname, and the Dutch Antilles/Dutch Caribbean) were significant, but political policing and a difficult relationship between police and community were overall the result of a weak and inefficient colonial state. Decolonization did not imply a breach of Dutch involvement with the police in the (former) Dutch West Indies. In Indonesia, the independent republic that replaced the Dutch East Indies after a period of war, Japanese Occupation, military actions, and diplomacy (1942–1950), the breach was more definite. However, the history of the decolonization of the Indonesian police still needs to be investigated in order to understand the continuities and discontinuities in policing and police methods in colonial and postcolonial Indonesia.

International Perspectives and Key Issues

The end of the twentieth century marked a turning point in the research on colonial police and state formation. Failing humanitarian interventions and unsuccessful international peace missions in those years may have provoked this change. The image of a repressive police, effectively maintaining colonial power, dominated academic historical writings until well into the 1980s (Anderson and Killingray, 1991, 1992; Arnold, 1986), but since the late 1990s and early 2000s, historians are inclined to emphasize

the fundamental weakness of the colonial state and thus the ineffectiveness of the police (Bickers 2004; Campion 2002; Chandavarkar 1998; Sinclair 2006). The research on Dutch colonial policing on which this present entry is mainly built (Bloembergen 2009, 2011a, 2012; Broek 2011; Klinkers 2011) links up with this approach and is in particular inspired by the insights of Rajnarayan Chandavarkar (1998) and David Campion (2002) that police violence was the result of a weak and inefficient state rather than a strong, decisive state. The failure of the colonial police was caused, as they argue, not only by inadequate management but also by a difficult relationship between police and the community.

The tensions within the twofold task of the police – to maintain order and to provide in the social need of security – were more manifest in colonial societies with their racial biased organization structures and fragmented power relations than elsewhere: the cooperation of the local population was required for providing safety and protection, but at the same time, the police had to maintain order on behalf of a foreign regime which could provoke hostility and popular opposition. The police was therefore faced with unsolvable dilemmas that incited repression, in such a way that the challenge became not how the colonial state maintained control through the police, but how they could achieve this in spite of the police.

It is a question whether the histories of the modern colonial police forces in the Dutch East Indies in Asia and the Dutch West Indies – consisting of Suriname and the Antilles (in the Caribbean world) – can be analyzed and understood within one integrated Dutch police system. On the one hand, the three organizations of police forces (for the Dutch East Indies, Suriname, and the Antilles) all emerged at the end of the nineteenth century in the context of modern colonial states-in-development; they shared as their main objectives the prevention and detection of crime, surveillance, and the maintenance of colonial power, while, in the context of a modern colonial state-in-development, colonial authorities and entrepreneurs waged similar debates about security and the organization and mission of the colonial police. The fear of a loss of control over the indigenous or colonized people in a changing colonial world and the colonial state's responsibility for the safety and protection of the inhabitants were key reasons for maintaining the colonial police. The state's concern for security legitimized colonial power, while security also seemed indispensable for economic development (Bloembergen 2009, pp. 71–106; 109–136).

On the other hand, however, the organization of the police forces in each of the Dutch colonies and their practice of policing differed importantly. Obviously, there were differences in scale and population. The colonial regime of the Dutch East Indies, after a period of military expansion until 1910, had to oversee a million population in an immense archipelago consisting of thousands of islands, stretching from the west of Sumatra to the east of New Guinea. However, the Dutch East Indies' modern colonial police forces, at the peak of their development in the early 1930s, consisted of only 54,000 men, at a population of 60 million souls (including less than 200,000 Europeans). Police control in the Dutch East Indies was therefore fragmented by definition and dependent on the effectiveness of local administration and security control. Meaningful, the semi-traditional village police – a service provided by male inhabitants of the villages – formalized by the colonial administration, was maintained throughout the colonial period.

Suriname is about 12 times smaller than the Dutch East Indies, and is about four times the size of the Netherlands, and thinly populated. The majority of the people live in the capital of Paramaribo and along the northern coast. Over three-fourths of Suriname is covered by tropical rainforest. The islands of the Antilles (Aruba, Curacao, St. Maarten, Bonaire, St. Eustatius, and Saba) are small, scattered widely over the Caribbean Sea, and densely populated. Until the 1950s, the colonial police in Suriname and the Antilles did not exceed a few hundred men each, with minimal specialization. These West Indian colonies were populated by Europeans and the descendents of African slaves.

In addition, Asian contract laborers and their descendents lived also in Suriname, while the interior was inhabited with small groups of Amerindians and Maroons. Thus, there were no indigenous systems of authority and control in Suriname and the Antilles.

The police in the Dutch East Indies was the largest and most professionalized of the three colonial police organizations. Nevertheless, the police in the Dutch East Indies was fragmented and revealed the weaknesses of the colonial state. Political consciousness and tensions in the colonies in the 1930s placed ideals of professionalization and modernization under pressure. Political policing became more prominent. The anxiety of the colonial state in the Dutch East Indies resulted in violent police actions, not diminishing but overshadowing policing out of care and protection. Policing in Suriname became in those years also subject to repression and gathering intelligence. In this period, the colonies shared politics of policing. Nevertheless, the outcome was different. While, after a violent decolonization war between 1945 and 1949, the independence of Indonesia was recognized by the Netherlands in 1949, a process of gradual decolonization of the Dutch West Indies was started.

As, in the end, the police forces in the Dutch East Indies and West Indies developed in distinct and mostly separate contexts, their histories will be dealt with in the following in more detail, separately.

Policing the Dutch East Indies

The lack of security was not an uncommon phenomenon in nineteenth century colonial Java. And policing had always been a tragic business. That is at least the idea that rises from colonialists' complaints about the dreadful performance of the Javanese *politieoppasser* (literal meaning: police caretaker) or the formidable sleep of the indigenous village wards – the two main tools of civil colonial policing during the nineteenth century. But with the army and a diverse flock of armed security guards and auxiliary forces at hand, the unsafe circumstances, especially in rural Java, never seemed a problem to the colonial authorities – at least until the last quarter of the nineteenth century. Only then, in the context of a "liberal" colonial state-in-progress and in the context of military, administrative, and economic expansion, worries were explicitly addressed – in public and governmental discussions – about rampant lack of security in Java. At the same time, a new ideal seems to emerge, the ideal of effective and good policing. The wracked result of this all was the police reform of 1897, the first of a series of three large police reforms – in 1897, in 1911–1914, and in 1918–1920 – by which the Dutch colonial state endeavored to get more control over the organization of security surveillance.

Fundamentals

The modern police of the Dutch East Indies, thus developed in the heyday of the "ethical policy" that held sway between 1900 and 1920, was the answer to a typical colonial problem: the struggle of a colonial state that wanted to be civilized but witnessed its legitimacy crumbling. Compelled to use force to enforce its authority, it nonetheless also sought to govern by consent. The modern police embodied these two contrary forces of the ethical policy, or the Dutch version of the *mission civilicatrice*, namely, simultaneous efforts to achieve development and control (Locher-Scholten 1981). Since the colonial police was the instrument used to pursue these diverse goals, it became a two-headed beast: in trying to safeguard the state's authority, it provoked resistance, while, in reaching out to fulfil society's need for security, it needed the cooperation of the local population (Campion 2002, pp. 1–2).

Three meaningful police reforms between 1897 and 1920 contributed to the installation of a *would be* modern police force, or the general police, that operated in the 1920s and 1930s. These reforms were all important steps in the colonial state's effort to gain control over security surveillance. After the reform of 1918–1920, the Netherlands Indies officially (although not effectively) had a centralized and uniform modern police force. It was in itself pluralistic: it consisted of (a) the *gewapende politie*

(an armed militarized police force, in hands of European administration) and (b) the general police, with its many divisions – the administrative police (dating from the nineteenth century), a modern city police (created in 1912–1914, large towns), the field police (*veldpolitie*; created in 1918–1920 for the security surveillance of the rural areas), and, mainly meant for political control, the *gewestelijke recherche* (the regional investigation departments). The Attorney General was in charge of central control, the director of Interior Administration of the management of the police.

The modern Dutch East Indies' police force never had, nor could it have, a monopoly of security surveillance. Therefore, the colonial budget was too small, and the territory and the population number too large. Manned for more than 96 % by Indonesians and directed by a small minority of European staff, the size of the force peaked in the 1930s at 54,000, and this on a population of almost 60 million souls, which counted less than 200,000 Europeans. Only a very small minority of Indonesian policemen could reach the higher ranks of the police force. Through its organization and hierarchy, the police force therefore reflected an essential characteristic of the colonial state: namely, how this state worked both on principles of cooperation and inclusion, and on difference and exclusion (compare Cooper 2005). Due to the dualistic principle of colonial administration in the Dutch East Indies, the system of policing and surveillance remained moreover very much fragmented. The effectiveness of policing and surveillance depended to a large extent on the functioning of the local administration – a European administrative hierarchy or the *Binnenlands Bestuur* (Interior Administration), paralleled by an indigenous administrative hierarchy, or the *pangreh praja*, which was responsible for administrative police and village police. This village police, partly a colonial construction as well, was a forced service of guarding and patrolling, which male adult inhabitants of a village or of indigenous quarters (kampong) in the towns provided in turn. While ideally the modern colonial police tools were to encapsulate these local semi-traditional tools, the village police, whether effectively functioning or not, remained for economic reasons the core of colonial security surveillance at the very local level (Bloembergen 2012). Next to that, the modern colonial police force also depended on other local or private forms of security surveillance and on how the modern police tools interacted with these.

Trustworthy to her civilizing mission, the Indies' government, however, kept up the ideal of the police as a *good* police force, that is, "a force of men with high mental attitudes towards life, a good morale and a strong character" – this according to an official governmental enunciation in the Volksraad in 1931. The Police School in Soekaboemi, set up in 1914 (10 years before a comparable school existed in the Netherlands), was the symbol and promoter of this ideal. The curriculum of the police academy reflected the same ambivalent notion of fear and concern that shaped the modern colonial police, and thus the aims to empower and gain loyalty, and to civilize. Thus, police recruits, from the lowest to the highest rank, were trained to perform as a tool for (political) control and as a *civil* security tool – at least as a more civilized tool than the army. On the one hand, they were trained in semi-military discipline, the use of arms and physical exercises, and, after communists revolts erupted in West Java and on the west coast of Sumatra at the end of 1926 and early 1927, also on how to recognize a "communist conspiracy" – all reflecting the police needs of a police state; on the other hand, they were educated on the principles of a constitutional state, criminal law and justice, methods and rules of *civilized* modern policing, and the idea that postponement of violence should be a *leitmotif* (Bloembergen 2009:203–247; 299–332; Bloembergen 2011a, b). Also, they were meant to perform as a tool of "civilization": to bring order, safety, neatness, cleanliness – in short "light" in colonial society. This idea became part of the self-image of the policemen who were trained at the police academy.

The image of good, civilized police and the idea of civilization and neatness through policing seemed to have become more important even in the 1930s. With further development of Indonesian

nationalism and fiercer political policing, it became harder for the colonial authorities to ignore opposition against colonial rule. After the violent repression of communist revolts in 1926–1927, mass arrests, and the internment of around 1,300 presumed communists without trial, the colonial government subsequently refined and extended the organization of political policing and enlarged the police force in general. The international economic crisis, which forced the government to cut down policing expenses, did not hamper the artificial image of *rust en orde* (law and order) or *zaman normal* (times of normalcy) by which the 1930s have been characterized. Meanwhile, the police, more visible because of the extension of political policing and being watched while watching, had become the standard for the quality of colonial government. For those groups who felt oppressed by the police, the police embodied "the dirty work of empire" (Orwell 2001 [1936]).

Consequences: Political Policing

Colonial anxiety was the main condition under which a centralized Dutch East Indies Intelligence Service was set up in 1916. In the context of modernizing colonial society, this anxiety was fed, in short, by peaks of panic about religious revolts and rural unrest in the countryside, a growing awareness of an indigenous movement towards association and progress (the *pergerakan*), the visible Japanese drive for expansion in Asia, and, therefore, double awareness of the insecure position of the Netherlands in Asia – a position which leaned on Dutch neutrality in international relations. After the installation of a Central Intelligence Service, set up during the Great War, and slightly reformed in 1919, the colonial government had – on paper – strong means to assess and control indigenous political activities: (1) a Central Intelligence Service (Politieke Inlichtingendienst, PID), in 1916, renamed Algemeene Recherche Dienst (General Intelligence Service, ARD), in the hands of the Attorney General; (2) a weekly survey of Javanese, Malay, and Malay Chinese newspapers (*Overzicht van de Inlandsche en Maleische Pers*, IPO); (3) the new Regulation for association and gathering, or article 8a, a tool in the hands of the police to act (=to attend/watch meetings and, officially when directed to do so, to intervene whenever "disturbances were expected") (Sunario 1926, p. 19); and (4) steady background intelligence, provided by the advisor for Native Affairs (Laffan 2002).

From ARD to the local administrators, from the modern police in all its diverse units to local spies, and back to the center, the will "to know" and means to arrive at knowledge were there. But did this result in perspicacity? Practice was rough.

In the night of 12 November 1926, the colonial police and administrators on Java were completely surprised by communist revolts that took place at several locations at the same time – in the main capital Batavia, in Bandung, and, most thoroughly, in Banten (West Java). Completely surprised they were again in January 1927, when another communist uprising occurred in the west coast of Sumatra. These uprisings, both organized by the Partai Kommunis Indonesia (Indonesian Communist Party, PKI), could occur despite the existence of a Central Intelligence Service, despite intense police surveillance, despite repression of strikes in the first half of the 1920s (organized by the Union of Tram and Railway workers), and despite forced exile of most of these organizations leaders. The latter measures could be implemented on the base of the so-called exorbitant rights of the governor-general, a regulation dating from 1854 which gave the governor-general the right to expel persons living in the Dutch East Indies when it was in the interest of colonial law and order. These violent interventions and repressive measures did have the effect that the preparations of the intended revolt were obstructed by loss of strong leadership, internal conflict, and miscommunications within the PKI divisions. Moreover, since the end of 1924, the colonial authorities notably feared organized unrest. Therefore, this was a double failure of the political police and the ARD: they were surprised by badly organized uprisings, which they had foreseen (Bloembergen 2009:247–259).

This clear failure of Governor-General Johan Paul van Limburg Stirum's expressed wish, in 1916, to be "informed [. . .] on what's going on in the Native's mind," was above all a failure of central control of the political police. The main

reasons are as follows: due to the fragmentation of colonial authority, political policing remained to a large extent a local business. Miscommunication, due to language problems, unbridgeable distances, misunderstandings, and conflicting interests of police, administrators, and local spies did the rest. And this miscommunication was enhanced by the mechanisms of secrecy and misleading that accompanied spying and the PKI's preparations of a revolt. As one police expert would point out later on, despite ARD, central police control, and regional intelligence forces, there was no effective interlocal (nor supra-local) collaboration between police forces and administrations.

The way the colonial state subsequently repressed the communist uprisings was a typical solution of a colonial state that went on the defensive: what followed was mass repression, mass custodies, and mass internments to a special camp in Boven-Digoel (in short "Digoel") in New Guinea – and this without any form of trial. The PKI was no longer tolerated.

The communist revolts moreover stimulated a general enlargement of the police force: the set up of a refined system of intelligence gathering and reporting (the *Politiek-Politioneele Verslagen*) and the set up of stronger guidelines for recognizing "radicalism" and keeping firm while watching. The memory of the fierce repression of the communist uprisings and the spectre of "Digoel" moreover made clear to nationalist/anti-colonial opposition, where the government drew the line. In that sense, 1927 was a turning point: a more pronounced policy of political policing made political police and the police state-in-development more effective – or so it seemed (Bloembergen 2009, pp. 247–298, 2011a).

The system of political policing still fell prey to bureaucratic retarding and the inability to process more information, to miscommunication, and to misunderstandings as well. Moreover, because the political police were, in their task of political control, guided by the search for signs of communism, they further blinded the state. The system of political surveillance therefore stimulated the inability and disinclination of the majority of the colonial government to try to understand the aims, aspirations, reach and depth of the national movement, and other drives for progress among the indigenous population. The system was also the product of this inability and disinclination.

Moreover, precisely because of the scale, violence, and visibility of their repressive actions, the police also showed the fragile position and hampering legitimacy of the colonial state. While the police looked out, the public watched, criticized, and ridiculed the police. In the end, the colonial police therefore was not so much a vehicle for discipline, but a vehicle for the loss of prestige and corrosion of colonial authority. This is what made colonial policing a dirty task, and in the end, a tragedy.

Future Directions: Decolonization

This entry does not cover the history of the police during the Pacific War and the Japanese occupation of the Dutch East Indies (1942–1945) nor the issue of policing and decolonization in the Dutch East Indies/the Indonesian Republic, during the Indonesian revolution and Dutch military aggressions in the period 1945–1949. These histories still need to be investigated, although Ambar Wulan (2009) made a beginning with her study on the Indonesian Republican police and intelligence gathering during the Indonesian Revolution, 1945–1949. To be short, for these periods there are, to a certain extent, continuities with prewar colonial policing: the problem of the legitimacy of the colonial state, and thus of the police, only had become more acute. The practice of policing and the nature of the security problems, however, also changed, becoming even more complicated and more pluriform. The Dutch colonial police, for several reasons, could no longer play the role it had played before the Pacific War. They resumed their tasks in a society transformed and also unsettled by the Japanese occupation and war economy, a society involved in a decolonization war, and a society in a permanent state of civil wars. Moreover, the colonial police had to concur with the new police force, developed by the Indonesian Republic in the same period. These circumstances complicated the tasks of policing and the issue of police loyalties. A study of the colonial police in this period would have to address the relationships

and interactions between this force and the colonial army and Dutch troops during the two military aggressions in 1947 and in 1948–1949 (notably referred to at the time as police actions) and those between this force, the Republican police force, and the various Dutch and Indonesian security forces developed in this period. The question concerning the functioning of the colonial police during this period is, therefore, not easy to answer and deserves new research.

Policing the Dutch West Indies: Suriname and the Netherlands Antilles/ Dutch Caribbean

The marechaussee in Suriname became operative on the first of July 1863, the same day that 33,000 slaves became free citizens in Suriname. This newly established police force symbolized the historical transformation of colonial society. The developments in the police force of Suriname in the late nineteenth century were closely intertwined with the changes in a society adapting to the abolition of slavery. Society was on the move. Social boundaries were no longer determined by slavery and freedom. Disciplining and law enforcement, until then largely based on military and plantation discipline, had to be reshaped. In addition, society became more complex because of the arrival of indentured laborers from Asia and the West Indies. Chinese, more than 34,000 British Indian laborers, and almost 33,000 Javanese immigrants from the Dutch East Indies moved to Suriname to work the plantations, while increasingly ex-slaves found a living outside the plantation economy. The unfolding developments of the Surinamese police force show the struggle of the colonial authorities to accept, mold, and structure the socioeconomic changes in colonial society in the aftermath of slavery. The search for a suitable colonial police force resulted in a series of reorganizations which came to a conclusion with the founding of the *korps gewapende politie* (armed police force) in 1895.

Maintaining colonial power was a major concern for the planters and local authorities. The police had to be efficient, leaning towards military discipline, but also decent and generous towards the people to legitimate state control and to realize its civilization mission. That is, to create a society, conform western European culture with people willing to work for the benefits of colonial prosperity (Klinkers 2011).

Fundamentals

In November 1862, more than 6 months before the abolition of slavery, the minister of Colonial Affairs gave permission to the governor of Suriname to introduce a new colonial police force, *het korps marechaussee* (marechaussee). On the Dutch island of Curaçao, a brigade of marechaussee had been operating since 1838. This brigade was not an institutional part of the ministry of war like the Royal Marechaussee in the Netherlands but stood under the command of the Attorney General. As a result, the colonial authorities in Suriname chose unanimously for the installation of a civilian police force with a strong military character. Military of European descent and encamped in Suriname were recruited to join the marechaussee. The policemen were supposed to live in barracks to keep a safe distance from the local community.

The Attorney General had a difficult time to maintain the marechaussee up strength. The solution for this shortage of personnel was sought in the foundation in 1868 of a second police force, the *korps inlandse politie* (inland police force). The inland police force would be manned with creoles, that is to say, descendents of former slaves in Suriname. The entrance of Creoles into the police force seemed inevitable, but their incorporation into the marechaussee was considered problematic. Racial prejudices certainly abounded. The obsession with the separation of local and European policemen in different police forces was remarkable, since Suriname had a long history of cooperation between black and white men in maintaining law and order. Black overseers were considered as mediators between slaves and whites at the plantations. Besides, patrols of black (both enslaved and free) and white men pursued runaway slaves in the interior of Suriname in the seventeenth and eighteenth

century. The racial separation in the post-abolition police force can be understood as a redefinition of social boundaries in the aftermath of slavery (Compare Cooper and Stoler 1997, p. 7). The marechaussee was supposed to represent the colonial state and symbolized the unbroken supremacy of the white colonial population after emancipation.

However, there was no fundamental difference in the police practices of both forces. The removal of the institutional distinction between inland police and marechaussee took decades with endless discussions. Eventually Governor T.A.J. van Asch van Wijck decided to turn daily practice into policy and merged the two systems into one armed police force. The institutional changes did not fundamentally change the system. What remained was one police force with a military character, with a growing shift from European to local personnel through time. The colonial authorities failed to establish a strong police force during the twentieth century, because of the lack of financial means. Besides, the authorities seem to distrust the local policemen and clang to the army as a loyal ally of colonial power. In addition, the rise of (semiprivate) policemen fractured the system of law enforcement even more.

The initial idea to organize the police systems in the colonies of the Dutch West Indies in the same way faded to the background. The Antillean police force passed through several stages during colonial times, beginning with the *brigade koloniale marechaussee* (colonial marechaussee) (1839–1918), followed by *korps veldwacht* (rural police) (1873–1918, 1932–1949), *detachement marechaussee* (marechaussee) (1911–1918), *korps burgerpolitie* (civilian police) (1918–1949), and *korps militaire politietroepen* (military police troops) (1928–1949). The police was highly militarized and had, in contrast to the Surinamese police, a large influx of Dutch military and civil policemen throughout all its stages, except for the rural policemen. Personnel was distributed over six islands, but the majority was located on Curaçao. The expansion of the oil industry in the 1920s, requiring security and causing a large recruitment of people on the island, had been a reason for reorganization, expansion, and further militarization of the police system with the formation of a *korps militaire politietroepen* besides the already existing *korps burgerpolitie*. The bifurcated police system did not live up to the expectations of a strong and efficient police force. The weakness of the system of maintenance order became apparent when a small group of Venezuelans overpowered the police station in a fortress on Curaçao in 1929 (Broek 2011). The Venezuelans under command of R.S. Urbina not only succeeded in overpowering the fortress but also captured Governor L.A. Fruytier for almost 1 day. The incident was considered as an embarrassment for colonial power.

Consequences: Political Policing

Ideals of an effective and a respectable police force, of civilizing society and its people, were frequently at odds with maintaining colonial power and control. This happened in uncertain times, as in the years following the abolition of slavery, but became most apparent in the 1930s. The global economic crisis not only caused poverty and unemployment in Suriname and Curaçao but raised also political consciousness and opposition against the colonial administration. Surinamese laborers who had been employed in the oil industry on Curaçao in the 1920s lost their jobs and returned to their home country, disillusioned but inspired by the island's labor unions and left-wing press (Ramsoedh 1990, p. 33). The Surinamese colonial authorities feared that communism and nationalism would conquer influence, a fear that was stimulated by the communist revolts on Java and Sumatra in the Dutch East Indies in 1926 and 1927 (see above).

In Suriname, repression and political policing became more manifest than ever before. The press, labor unions, and other signs of political consciousness were not tolerated and repressed. An anti-revolution law was announced in 1933 to defy the feared threat of communism and to maintain colonial order. Most illustrious was the arrest of the Surinamese writer Anton de Kom who had been involved with communist and nationalist groups during his sojourn in the Netherlands. His decision to travel to Suriname to talk about the history of slavery caught the

attention of the Dutch intelligence service that was surveying de Kom for some time already. They warned Governor A.A.L. Rutgers that he was suspected to poison the Surinamese people with anti-colonial thoughts. As a result, de Kom stood under police surveillance continuously, was forbidden to speak in public, and detained several times. A group of people gathered awaiting his release after he had been arrested again. The police ended this protest with force; 22 people were wounded and two were killed (Woortman and Boots 2009, pp. 63–135).

In this period of anxiety, the colonies shared politics of policing. The authorities in Suriname and the Antilles informed each other about possible security threats and measures. After the appointment of J.C. Kielstra as a governor of Suriname in 1933, the events in the Dutch East Indies gained importance. Kielstra had been a colonial administrator in the Dutch East Indies. Kielstra increased the repressive regime further, even though there were no indications that communism and nationalism would manifest themselves in Suriname as strongly as in the Dutch East Indies (Klinkers 2011, pp. 105–156).

Future Directions: Decolonization

Suriname and the Dutch Antilles became autonomous parts of the Dutch Kingdom in 1954 as stipulated in the 'Charter for the Kingdom of the Netherlands.' The Netherlands remained responsible for foreign affairs, defense, and good governance (Oostindie and Klinkers 2012, pp. 21–27). The police were internal Surinamese and Antillean affairs from then on. The Dutch gave up leading positions in the police organization in Suriname, which was already mainly staffed by local men reflecting the multicultural society. The Dutch input in the Antillean police forces was and would maintain relatively strong.

However, self-rule did not imply a breach of Dutch involvement in Surinamese police organization. Schooling and training, technical assistance, and cooperation between Dutch police and those in the West Indies intensified, while joint international crime investigation became more important. The organizations in Suriname and the Antilles started to expand and professionalize from the 1950s onwards. In Suriname, special branches, such as an intelligence service, children's police, and traffic police were founded. The armed police force changed its name into *korps politie Suriname* (Suriname police force) in 1973, emphasizing the civilian mission of the police, even though military aspects in its presentation and protocol would be uphold. The ineffectiveness of the fragmented Antillean police system had been acknowledged before. The three forces (civilian police, military police troops, and the village police) were transformed into the *korps politie Nederlandse Antillen* (Dutch Antilles police force) in 1949 (Broek 2011, pp. 147–149).

The fear that the independence of Indonesia would become a source of inspiration for the people in Suriname was a reason for gathering intelligence in the 1950s. But more than the struggle for independence, the Dutch government – encouraged by the United States – feared that communism would spread as an ink spot in the Caribbean area after the Cuban revolution of 1959. The Dutch government's secret intelligence service, *binnenlandse veiligheidsdienst* (BVD), assisted in the founding and professionalization of the Antillean police forces and Surinamese intelligence services (*veiligheidsdienst Nederlandse Antillen, centrale inlichtingendienst Suriname*), both subdivisions of the local police organizations. The cooperation was not entirely successful because of mutual distrust, reason for the Dutch BVD to work together with the marines on the Antilles and the Dutch troops in Suriname.

The Dutch government had to bring foreign policy into practice in a changing colonial reality. The work of army and police, intertwined during colonial times as both had been occupied with the internal security, had to be unraveled. The army or marines were still allowed to assist the police in times of crisis, but when and how this was supposed to happen was hard to determine. A conflict between Suriname and Guyana police force about territorial claims in the southwestern border area in 1968–1969 and a revolt on Curaçao in 1969 demonstrate how difficult it was to act in a mutually acceptable way and according to 1954 Charter.

The border dispute between Suriname and (British) Guyana harks back to the nineteenth

century but came to a crisis after Guyana became independent in 1966. After the Guyana police force dispatched Surinamese workers out of the area, the Surinamese government demanded a military response to the Guyanese action. The Dutch government sympathized with the Surinamese territorial claims but refused military assistance which would intensify the conflict. The Surinamese government founded in response a special police unit, the *defensie politie* (defense police), to guard and protect the territory. Eventually, the conflict would never come to an armed clash, but simply lost its urgency. Without coming to a conclusion the territorial claims still lingers on (Klinkers 2011, pp. 157–244).

The labor conflict in Curaçao which ended in a revolt on May 30, 1969, challenged the Charter even more. Thousands of demonstrators marched through the streets of Willemstad, looting and burning. The police killed two people. Eventually, the local authorities called the Dutch Marines for assistance to end the uproar. This military action was successful in restoring order but was criticized at home and abroad for this neocolonial act (Broek 2011, pp.153–190; Oostindie 1999). The revolt changed the attitude towards the Charter. It made the Dutch government realize that their possibilities for intervention to guarantee good governance overseas were limited. Besides, the continuing dependency of Dutch financial aid and the unlimited stream of migrants were other consequences of the Charter. The government in The Hague headed for independence of the West Indies, which was accepted in Suriname in 1975. The Antilles, on the other hand, refused sovereignty and are connected with the Netherlands until today, even though the ties loosened and the islands are no longer united in one country since October 2010 (Oostindie 2011). Aruba took the lead and received a separate status in 1983 already; the *korps politie Aruba* was founded in 1985 as a result of new responsibilities and legal arrangements. As international crime, like drugs and human trafficking, increased, the cooperation with and investments in the (former) Dutch Caribbean police systems remained of major importance for both areas.

Related Entries

- ▶ French Colonial Police
- ▶ High Policing

Recommended Reading and References

Ambar Wulan G (2009) Polisi dan politik: Intelijen kepolisian pada masa revolusi tahun 1945–1949. Rajawali pers, Jakarta

Anderson DM, & Killingray D (eds) (1991) Policing the empire: Government, authority and control, 1830–1940. Manchester University Press, Manchester

Arnold D (1986) Police power and colonial rule: Madras 1859–1947. Oxford University Press, Delhi

Bickers R (2004) Empire made me: an Englishman adrift in Shanghai. Penguin, London

Bloembergen M (2007) The Dirty work of empire: modern policing and public order in Surabaya. Indonesia (Cornell University) 83:119–150

Bloembergen M (2009) De geschiedenis van de politie in Nederlands-Indië: Uit zorg en angst. Boom/KITLV, Amsterdam/Leiden

Bloembergen M (2011a) The perfect policeman. Colonial policing, modernity and conscience on Sumatra's West Coast in the early 1930s. Indonesia 2011:165–193

Bloembergen M (2011b) Being clean is being able: policing colonial *netheid* and gay vices in the Netherlands-Indies in the 1930s. In: van Kees D, Taylor J (eds) Cleanliness and culture: Indonesian histories. KITLV Press, Leiden, pp 117–146

Bloembergen M (2012) Vol, meurtre et action policières dans les villages Javanaises: Les dynamiques locales de la sécurité aux Indes néerlandaises dans les annees 1930. Genèses 82:8–36

Bowling B (2010) Policing the Caribbean. Oxford University Press, Oxford

Broek Aart G (2011) De geschiedenis van de politie op de Nederlandse Caribische eilanden, 1839–2010: Geboeid door macht en onmacht. Boom/KITLV, Amsterdam/Leiden

Campion DA (2002) Watchmen of the raj: The United Provinces police, 1870–1931 and the dilemmas of colonial policing in British India. Dissertation, University of Virginia

Chandavarkar R (1998) Police and public order in Bombay, 1880–1947. In: Rajnarayan Chandavarkar, imperial power and public order in Bombay. Cambridge University Press, Cambridge, pp 181–233

Cooper F (2005) Colonialism in question: theory, knowledge, history. University of California Press, Berkeley

Cooper F & Stoler A (eds) (1997) Tension of empire: colonial cultures in a bourgeois world. University of California Press, Berkeley

Cyrille F (1998) Politiemodellen beïnvloeden elkaar wederzijds: Het beleid aangaande het politiewezen in

Nederlandsch-Indië. Het Tijdschr voor de Polit 60:4–21
Gouvernementsblad Suriname
Klinkers E (2011) De geschiedenis van de politie in Suriname, 1863–1975: Van koloniale tot nationale ordehandhaving. Boom/KITLV, Amsterdam/Leiden
Laffan MF (2002) "A watchful eye"; The Meccan plot and changing Dutch perceptions in Indonesia. Archipel 63:80–108
Lamur HE (1973) The demographic evolution of Surinam 1920–1970: a socio-demographic analysis. Martinus Nijhoff, The Hague
Locher-Scholten E (1981) Ethiek in fragmenten: Vijf studies over koloniaal denken en doen van Nederlanders in de Indonesische archipel, 1877–1942. HES, Utrecht. [Ph.D.-thesis, Leiden University]
McCoy Alfred W (2009) Policing America's empire: The United States, the Philippines, and the rise of the surveillance state. University of Wisconsin Press, Madison
Oostindie G (1999) Curaçao 30 mei 1969: verhalen over de revolte. Amsterdam University Press, Amsterdam
Oostindie G (2011) In: Spapens T, Groenhuijsen M, Kooijmans T (eds) Een koloniaal instrument: De politie in de Antillen en Suriname. Liber Amicorum Cyrille Fijnaut. Antwerpen/Cambridge, Intersentia, Universalis
Oostindie G, Klinkers I (2012) Gedeeld Koninkrijk: De ontmanteling van de Nederlandse Antillen en de vernieuwing van de trans-Atlantische relaties. Amsterdam University Press, Amsterdam
Orwell G (2001) Shooting an elephant. In: Davison P (ed) Orwell and politics: animal farm in the context of essays, reviews and letters selected from the complete works of George Orwell. Penguin, London, pp 17–24. [Originally published in 1936]
Ramsoedh H (1990) Suriname 1933–1944: Koloniale politiek en beleid onder gouverneur Kielstra. Delft, Eburon
Sinclair G (2006) At the end of the line: colonial policing and the imperial endgame 1945–80. Manchester University Press, Manchester
Sunario (1926) Het recht van vereeniging en vergadering der Indonesiërs. Perhimpoenan Hakim Indonesia, Leiden
Woortman R, Boots A (2009) Anton de Kom: Biografie. Uitgeverij Contact, Amsterdam/Antwerpen

Dutch Crime Networks

Toine Spapens
Tilburg University, Tilburg, The Netherlands

Introduction

This entry offers a brief history of Dutch crime networks. There have always been gangs in the Netherlands that were involved in serious crimes, such as violent robberies, extortion, large-scale theft, and the supply of illegal goods and services, and that is still true today. In that respect, Dutch networks do not differ from crime groups in most other countries. The reason to address these networks separately is the fact that, particularly since the 1970s, Dutch crime groups have managed to establish a far-flung international network for the import and reexport of different types of narcotic drugs. Furthermore, since the 1990s, the Netherlands has also developed into a major producer and exporter of synthetic drugs and cannabis. A key feature of the Dutch networks is that they organize these activities primarily in loose-knit cooperatives revolving around criminal entrepreneurs who transact business with others in shifting coalitions. To some extent, it is even more appropriate to view them as a single network (a *meso network*) from which temporary cooperatives of varying composition (*micro networks*) spring up (Spapens 2006, 2010, 2011, 2012).

Here, Dutch networks are defined as groups composed of lawbreakers who live in the Netherlands. This does not necessarily mean that they are Dutch nationals. What is important, however, is that they must have criminal connections to other network members. Indeed, without such "criminally exploitable ties," they would be unable to take part in the coalitions. This definition therefore excludes itinerate gangs that also operate on Dutch territory and sometimes stay there for longer periods, but without visible connections to indigenous criminals.

Section "Bandits, Smugglers, and Local Providers of Illegal Goods and Services" starts with a brief description of the historical roots of Dutch networks. Section "The Emergence of 'Dutch-Style' Organized Crime" then addresses the emergence of Dutch-style organized crime from the 1970s onwards, resulting mainly from a shift towards drug trafficking. Section "Large-Scale Drug Production" focuses on the production of synthetic drugs and cannabis. Section "Schengen and Its Effects" goes on to discuss the most recent trends and developments in Dutch crime networks. Finally, Section "Conclusion" offers some concluding remarks.

Bandits, Smugglers, and Local Providers of Illegal Goods and Services

Criminal networks do not spring up out of the blue, and it is therefore important to take a brief look at the historical roots of organized crime in the Netherlands. Large gangs of bandits plagued the Low Countries as far back as the seventeenth and eighteenth centuries, as they did other parts of Northwest Europe (cf. Fijnaut 2013). In the Netherlands, such groups were particularly active in the present-day province of North Brabant, at the time a very rural and densely forested area. Although it became part of Holland after the end of the 80 Years' War, Brabant (as it was then known) was not a member of the States General and treated almost as a colony. During the Dutch Revolt, it had been a principal battleground, with both sides committing atrocities against the local population (Adriaensen 2007). As a result, the central government had little legitimacy with the people of Brabant. They supported the gangs, and the local authorities were susceptible to corruption. Moreover, the gangs usually set up camp on the borders of jurisdictions because this often led to disagreements between the respective magistrates about who should take action, resulting in inactivity. Finally, the brigands also made sure not to commit any crimes within the jurisdictions where they resided.

One of these bands, the White Feather gang based in Kaatsheuvel, totaled almost a hundred members, including women and children (Grootswagers 1983). Heavily armed and militarily organized – some of their ranks consisted of mercenaries and deserters who had stayed behind after the end of the 80 Years' War – the gangs committed burglaries and violent robberies in the countryside and extorted farmers by threatening to set fire to their farmhouses unless they paid them off (Egmond 1994). The gangs were highly mobile and operated in parts of the Netherlands outside Brabant, in what is now Belgium, and in the German border areas. Their demise finally came after 1798 when, under French influence, the State was unified and gendarmerie units were assembled to operate in the countryside. In only a few short years, the gangs were history. Some of their members' children, however, would carry on the legacy.

New opportunities arose when Belgium gained its independence in 1830, and it became lucrative to smuggle all sorts of goods back and forth across the new border. Sint Willebrord ('t Heike), a hamlet that had once been the lair of brigands, was located close by, and it rapidly turned into a notorious smugglers' den. Large-scale smuggling continued until 1960, when the Benelux Customs Union came into effect and border checks were abolished. After that, some smugglers reverted to large-scale theft, for instance, cars and armed robberies, or continued trafficking in illegal goods, notably amphetamines. Soon, however, they found other more rewarding illegal activities.

Other parts of the Netherlands proved less fertile ground for committing large-scale serious crimes. Government legitimacy was much stronger there, and because most people lived in towns, law enforcement was easier. Things changed when the Industrial Revolution – which only got going in the Netherlands late in the nineteenth century – led to the rapid growth of the cities in the western part of the country. Newcomers there faced poor living conditions, and poverty and the loss of existing social structures led to problems such as crime and alcoholism. Such problems were harder to control in the growing cities, especially because the police force had only just begun to refashion itself into a professional organization. Like the United Kingdom, which faced similar problems on a much larger scale, there was a revival of moral values in the Netherlands. From the second half of the nineteenth century onwards, the upper class did much to improve housing and hygiene as well as organize better welfare schemes for those who were unable to support themselves (De Swaan 1989). Furthermore, the government passed strict laws on vices such as gambling and prostitution. The government even considered a ban on stock market speculation, because it viewed it as a form of gambling (!), but parliament failed to pass the proposal. This helped lower the general crime figure on the one hand,

but on the other, people still wanted to have "fun," and this inevitably created bigger market opportunities for those who could provide it. A *penoze* (urban underworld) emerged, and "local heroes" would occasionally crop up who were involved in all sorts of illegal activities simultaneously – illegal gambling, prostitution, trafficking in illegal firearms, fencing stolen goods, small-scale drug dealing, and sometimes extortion – and who also had a stake in bars and restaurants.

These activities were usually concentrated in specific neighborhoods, the Red Light District of Amsterdam being one example. Although the famous red-tinted windows did not appear until 1930, the district had always been a rough place where sailors and city folk alike came for prostitutes, gambling, and drinking. Moreover, the Red Light District was always an international meeting point. Opium use, for example, was already a problem in the Chinese community before the Second World War, and inevitably, friendly relations with other members of the districts' underworld allowed for the supply to Dutch customers as well, albeit on a small scale (Wubben 1986).

Still, these developments are not exceptional. Bands of brigands were also common in other parts of Northwestern Europe. Hamburg, for example, has its Reeperbahn, which is comparable to the Red Light District. And while in 1960s Amsterdam had "local heroes" like "Zwarte" (Black) Joop de Vries, London had the Kray Twins. Dutch networks as these are now known only emerged in the mid-1970s, and the key to this was the wholesale trade in hashish and marihuana in particular.

The Emergence of "Dutch-Style" Organized Crime

Youth culture developed rapidly in the Netherlands in the 1960s, as it did in other western countries. One result was the increasing scale of narcotic drugs use, particularly hashish and marihuana. These drugs had to be imported from abroad and the main source countries then were Pakistan, the Lebanon, and Morocco. At first, the drugs were smuggled in small quantities by young people, often users themselves, who had traveled to the east. As demand increased, traditional criminals also started to discover the market, and some of the "hippies" managed to expand their businesses as well. In 1974, for example, the coast guard intercepted a fishing trawler, the "Lammie," that carried two tons of hashish. Investigations revealed that some well-known members of the Red Light District underworld had organized the transport (Middelburg 2001).

Although the Netherlands was a party to the UN Single Convention on Narcotic Drugs, its government has always been relatively lenient towards drugs use and the possession of small quantities for personal use. It was agreed that drug addicts should be given help or treatment, and not serve as law enforcement targets (Leuw 1994, p. 34). With regard to cannabis, the government decided in 1976 to make a distinction between soft drugs (i.e., hashish and marihuana) and hard drugs (all other narcotic drugs). The Board of Procurators General subsequently issued a directive stating that the public prosecution service would no longer prosecute soft drugs possession as long as the amount did not exceed 30 g. Blanket legalization was impossible because that would violate the Single Convention. Logically, this also meant deciding how to act towards dealers selling soft drugs to customers. The 1970s, for example, saw the emergence of coffee shops selling hashish and marihuana. In 1979, another directive ordered prosecutors to refrain from active investigation of dealers and coffee shops unless they put public safety or health at risk, or openly tried to promote and expand their business (Spapens and Van de Bunt forthcoming).

Although historical research is lacking, it is clear that these changes also expanded the market for crime groups importing narcotic drugs. To begin with, coffee shops, particularly in Amsterdam and the border regions, started to attract large numbers of foreign youth who wanted to try out soft drugs without the risk of apprehension. As there was no system of

licensing in place yet, the same crime groups that imported the drugs could also open coffee shops. Furthermore, criminals from the south of the country, who had accumulated capital from smuggling and armed robberies, teamed up with their counterparts in the western cities and started to invest in large shipments. The Dutch importers quickly succeeded in establishing business relations with wholesale producers in source countries.

In the mid-1980s, concern grew within the Dutch police force about such groups developing into organized crime syndicates of mafia-type proportions (Sietsma 1985). Moreover, when cocaine became fashionable in the 1980s, the importers began to serve that market too. In particular, Klaas Bruinsma's network seemed to be bringing increasing quantities of drugs into the country, and it tried to take over other businesses as well, such as the installation of gambling machines in bars and restaurants. Law enforcement response was slow, however, and it took until 1991 before the first special investigation squad was set up. That same year, Bruinsma was murdered, so the investigation instead focused on his "heirs," supposedly a trio that had taken over the "management of the organization," which was code-named Delta.

The competent authorities abruptly dismantled the investigation team in 1993 when an undercover operation spiraled out of control. It involved a criminal informer whom agents allowed to bring large quantities of drugs into the country in order to build trust with the alleged top leaders of the Delta group. That way, the police would be able to obtain evidence against them. After 2 years, however, the operation was not any closer to achieving its goal, whereas the informer had imported thousands of kilos of soft drugs, and possibly cocaine as well, with the consent of the police, who also allowed him to sell it on the illegal market. This resulted in a major scandal that led to a parliamentary enquiry.

Although one of the causes of the problem was the inadequate regulation of special investigative powers – the parliamentary enquiry revealed that other investigation teams had used similar methods – another was the fact that the police wrongly viewed these crime groups as well organized and hierarchically structured. Instead, they more closely resembled a network organization. "Entrepreneurs" within these networks continuously formed different coalitions in order to import and sell wholesale quantities of drugs. As far back as 1990, Dutch criminologists also came to the conclusion that "Dutch-style organized crime" was an amalgam of loosely knit networks instead of well-structured "firms," and this was confirmed in subsequent studies (Van Duyne et al. 1990; Van Duyne 1995; Fijnaut et al. 1996; Kleemans et al. 1998, 2002; Klerks 2000).

According to the literature, Dutch networks generally operate in line with the following basic model, which can be applied to drug trafficking and to other types of organized crime involving illegal goods.

A typical criminal cooperative consists of about 10–30 people. To begin with, most coalitions consist of a stable core made up of a criminal entrepreneur and a two or three well-trusted associates, usually family members or longtime friends.

Around this core, there is a second shell consisting of personnel carrying out the more crucial tasks, such as arranging cover-up loads. Unskilled workers, who are the most numerous, also stem from relatively close social circles, usually from the "extended family." Specialists, however, such as people who can set up complex money laundering schemes, often come from outside the direct criminal *milieu*. A good example of this is Willem Endstra, a real-estate dealer who also invested money on behalf of criminals. Specialists may get involved in criminal activities because of financial problems, or because they become friendly with a criminal entrepreneur and are attracted by the money or the excitement. Criminals may also deliberately maneuver people with specific skills into a situation of financial or other type of dependency and then more or less force them to cooperate (Spapens 2006).

The third shell of accomplices takes care of the high-risk tasks, such as the actual smuggling.

Although these persons know at least one of the other members of the group, they are kept in the dark about the extent of the activities.

The entrepreneur, aided by his core of long-term associates, strikes the business deals with other organizers at home and abroad, whereas the operational personnel varies with the requirements of the specific "projects." These may differ depending on the type of crime – entrepreneurs are usually active in different fields simultaneously – but also on the agreements made in the business deals, for instance, who will be responsible for smuggling. If the other party, for example, has access to a proven smuggling route, it makes sense to let them take care of that part of the activities. Although the actual involvement of the workers varies, for practical reasons, they are often recruited from more or less the same pool of individuals.

The key persons of the crime groups are usually 30–40 years old, but some may continue their careers well past the age of retirement, interrupted of course by stints in prison. Being a leader of a criminal group involved in drug trafficking requires different types of human capital, such as a reputation in criminal circles, organizing skills, and contacts with suppliers and buyers abroad. In extended families, such capital is often transferred from one generation to another. Sons or sons-in-law – women still seem to play a modest role – can take advantage of the reputation and the network of their fathers, which enables them to engage in large illegal business deals quickly (Spapens 2006).

Immigration is another important factor. Nowadays, tens of different nationalities live in the Netherlands, and members of these communities may be able to bridge the gaps between supply and demand for specific illegal goods and services between their countries of origin and the Netherlands. For example, in the 1990s, Surinamese nationals were involved in the trafficking of cocaine to Europe; Chinese groups in the trafficking of migrants from China to the United Kingdom, but also in ecstasy production in the Netherlands; Turkish families in the trafficking of heroin from Turkey, but they also smuggled ecstasy back to the Turkish Riviera; Moroccan groups imported hashish; and former Yugoslavians smuggled illegal firearms to the Netherlands and narcotic drugs back to their countries of origin. Groups of different nationalities are most certainly not islands within the Dutch *meso* network, and all sorts of contacts and business relations exist. When it comes to the trading of illegal goods, criminal groups usually do not specialize in just one commodity.

In the Netherlands, levels of violence and corruption are relatively low. Most of the narcotic drugs produced in (see Section "Large-Scale Drug Production") or imported into the Netherlands are exported abroad. There is no indication of turf wars over contacts with suppliers and buyers. Violence is usually the result of business deals going wrong, cheating, or the theft of a shipment of illegal goods. Corruption is also relatively rare. It usually involves relatively low-level customs officers who are able to ensure that containers in which illegal goods are concealed can be safely brought into the country. Criminals may also bribe police officers to provide information on ongoing criminal investigations.

Although it took some time for law enforcement agencies to gain a clear understanding of the composition and working methods of the Dutch networks in the early 1990s, and to adjust their investigative techniques accordingly, they gained experience over time and managed to bring several major cases before the courts, after which soft drug imports did appear to decrease. As it turned out, however, that was partly because the networks discovered drug production as a lucrative activity in the mid-1990s.

Large-Scale Drug Production

Drug production had already started back in the 1970s, with the manufacture of amphetamine, and expanded from the early 1990s onwards, when ecstasy production and cannabis cultivation took off. Production and sale, however, cannot be treated separately. After all, there is no point in producing large quantities of drugs (or importing them, for that matter) if you are unable to

find wholesale customers. Export is the key to this, because the Dutch domestic market is relatively small.

In the mid-1970s, crime groups, particularly those in the south of the country, started to produce amphetamines. Amphetamines were only included in the Dutch Opium Act in 1975, much later than elsewhere in Europe. Consequently, Dutch providers started to smuggle "speed" to dealers abroad. With the sales network in place, they also succeeded in taking up illegal production when amphetamines were finally criminalized.

Large-scale use of MDMA, or ecstasy as it is popularly known, started in the second half of the 1980s. Here, the story is comparable to that of amphetamine. The United Nations had already added MDMA to the list of controlled psychotropic substances in 1985, but it was not included in the Opium Act until 3 years later. Other countries were quicker to respond to the UN requirement. Logically, this once again resulted in the Netherlands developing into a source country. After 1988, illegal production quickly took off and by the next year, the police had discovered the first clandestine laboratory (Weijenburg 1996). Consumer demand for ecstasy skyrocketed in the early 1990s. Foreigners preferred Dutch ecstasy because of its good quality, and this resulted in a thriving export trade. The Dutch-Belgian border area in the south of the Netherlands became a heartland of ecstasy production. Here, criminals had already built up experience with synthetic drug production, and soon, most networks in this part of the country switched to manufacturing ecstasy or started to combine it with existing illegal activities. Apart from the south, production also appeared to concentrate in the west of the country, particularly around Amsterdam (KLPD 2005).

In the first half of the 1990s, dealers in other Western European countries, particularly in the United Kingdom, were the most important buyers of Dutch-produced ecstasy. In the second half of the decade, the United States also emerged as a very attractive market. Members of the Jewish Diaspora played a crucial role in establishing connections between Dutch suppliers and American wholesale buyers. They included extended families whose members lived in the United States, the Netherlands, and Israel. Groups of Chinese living in the Netherlands got involved in delivering PMK, an essential controlled chemical for the manufacturing of MDMA, which they imported from China, and after the turn of the century, they also began to set up ecstasy laboratories and smuggling large quantities of MDMA to Canada (Spapens 2006).

Under increasing pressure from the United States – in 2000, President Clinton personally expressed his anger about Dutch ecstasy flooding into his country to Dutch Prime Minister Kok – the Dutch decided to crack down hard on the crime groups producing MDMA. This resulted in a notable reduction in production. To begin with, the police successfully investigated a number of key players, who were sent to prison for lengthy periods (at least by Dutch standards). Next, diplomacy resulted in the Chinese government taking more effective measures against the illegal production of PMK in their country. More prosaically, the Dutch lost the still lucrative US market to Chinese crime groups, who, having learned the production process in the Netherlands, set up laboratories in Canada. Only Australia then remained as a major overseas destination, although small shipments carried by couriers traveling by plane may still go anywhere in the world. It is also important to note that ecstasy is now far less popular than in the halcyon days of the 1990s. Finally, Dutch criminals also had an excellent and far less risky alternative that generated major profits: cannabis cultivation.

The Dutch climate is generally not very favorable to growing cannabis, and marihuana users used to prefer the "pot" imported from abroad. This changed in the early 1990s with the introduction of indoor cultivation methods. Americans who came to the Netherlands allegedly introduced the technique (Boekhout van Solinge 2008). The Dutch then improved the product by raising the percentage of THC, the working component of marihuana. Hobby growers started to set up nurseries and began to supply the coffee shops. At first, even the government welcomed this development because it reduced the need for imports, which were largely in the hands of the Dutch criminal

networks. In hindsight, this was quite naïve, because the same groups of course quickly recognized the new market (Weijenburg 1996).

At first, Dutch networks started to set up relatively small-scale cannabis nurseries in the homes of members of their extended network, usually in economically weak neighborhoods (Bovenkerk and Hogewind 2003). Because the police put little effort into investigating cannabis cultivation and the penalties for those caught were lenient – often limited to a fine or community service – volunteers who wanted to earn extra money were lining up. However, the Dutch networks found managing a large number of nurseries for people who themselves had little growing experience more labor-intensive than they liked. They also attracted attention because the criminals sometimes resorted to violence, particularly if they suspected the growers of embezzling some or all of the harvest and selling it themselves to make a bigger profit.

Around the turn of the century, the Dutch networks found better solutions. On the one hand, they opened "grow shops" (Spapens et al. 2007). Here, growers could not only buy all the necessary equipment and get advice, but they could also purchase cuttings and sell the harvest. On the other hand, the professional criminals started to set up very large plantations – 5000 to 50,000 plants – themselves or, even better, gave someone with the necessary skills the equipment and had him run the plantation for them at his own risk. The production process was outsourced, and entrepreneurs could go back to focusing on trading the product. Today, they usually sell the best quality homegrown cannabis to coffee shops, whereas medium and low-grade products find their way to foreign dealers (Spapens et al. 2007).

Schengen and Its Effects

In 1995, the agreement between the Governments of the States of the Benelux Economic Union, the Federal Republic of Germany, and the French Republic on the gradual abolition of checks at their common borders, better known as the Schengen Agreement, came into effect. In 1998, the Treaty of Amsterdam further expanded the Schengen zone. The "open borders" had both visible and less-visible consequences for drug trafficking from the Netherlands to other parts of continental Europe and particularly Germany and France, because fixed customs controls at the Dutch-Belgian border had already been abolished in 1960. Things did not change with regard to the United Kingdom, but the country remained one of the important destinations for narcotic drugs from the Netherlands.

The most visible effect was a sharp increase in the number of Belgian, German, and French drug tourists who visited the border towns in particular. Foreign drugs users were naturally not a new phenomenon in the Netherlands. As far back as the 1970s, heroin addicts, especially from Germany, came to the country for good quality and low prices. To avoid the risk of being checked at the border, they generally did not take home major supplies for personal use. The same applied to foreigners who came to the Netherlands to buy hashish or marihuana in a coffee shop. Because of these limitations, most drug tourists came from the border areas. "Schengen" meant that the risk of taking quantities of drugs across the border virtually disappeared, and it thus became attractive for users living further inland to drive to the Netherlands regularly. Furthermore, general economic prosperity led to many French and German youth owning their own cars and having enough money to drive several hundreds of kilometers round trip to buy drugs in a Dutch coffee shop. A few years ago, thousands of drug tourists a day came to border towns such as Maastricht, Roosendaal, and Terneuzen (Spapens 2008; Fijnaut and De Ruyver 2008).

Another effect was the emergence of drug-dealing houses in the border towns. Of course, not every drug tourist was interested only in hashish or marihuana, and the amount they were permitted to buy in a coffee shop was limited to 5 g per person. Drug-dealing houses offered larger quantities and other types of narcotic drugs as well. In 2008, the number was estimated at 150 in the district of South Limburg alone (Fijnaut and De Ruyver 2008, p. 149).

Finally, suppliers operating in border areas periodically ship quantities of drugs, known as "kilo shipments," directly to dealers in Germany and France. A foreign dealer may have a weekly order of two kilos of marihuana, 1,000 XTC pills, 100 g of cocaine, and 50 g of heroin, depending on his customer base. The Dutch dealer will contact his suppliers of different types of illicit drugs to fill the order. Next, a courier will drive to the Netherlands to pick up the drugs. Because the risk of apprehension is virtually nonexistent, such deliveries have developed into an almost on-demand service (Spapens 2008).

Although it seems that the major entrepreneurs within the Dutch networks still control the import and export of wholesale quantities of narcotic drugs, there are indications that members from the "second shells" are now also organizing kilo shipments on their own. We do not have a clear picture of how this works, because tackling small-scale shipments – by Dutch standards – is not at all a priority with law enforcement officials, given the fact that their in-tray already overflows with cases involving hundreds or even thousands of kilos of drugs.

Generally, the authorities have also been slow to respond to these developments, mainly because "soft drugs" were long associated with "love and peace." Law enforcement agencies also found it difficult to crack down on cannabis cultivation because the sale in coffee shops was not prosecuted. Furthermore, painstaking criminal investigations resulted in relatively short prison sentences that hardly affected the broader network. Finally, there is less pressure from abroad than in the case of ecstasy production in the 1990s. Starting in 2004, however, some changes began to take place. To begin with, if a small-scale nursery was found in a private home, the dweller would be prosecuted, but also taxed for the extra income, and if he rented the house, he would also run the risk of eviction. Second, the police and the public prosecution service set up task forces to combat cannabis cultivation and the networks involved, aimed specifically at making it more difficult for people to grow cannabis. Third, "Joint Hit Teams," consisting of Dutch, Belgian, and French police officers, were established to patrol the main motorways used by drug tourists and kilo couriers and to take action against drug-dealing houses. Finally, in 2012, foreign customers were banned from coffee shops in the south of the Netherlands and plans are to extend this ban to the rest of the country as of January 1, 2013. However, there has also been fierce criticism of this measure, particularly because Dutch customers also need to register, which most of them refuse, and whether the government will go through with it is an open question.

Conclusion

According to the routine activities theory, crime requires motivated offenders, suitable targets or opportunities, and ineffective responses by the authorities (Cohen and Felson 1979). To begin with, we can conclude that in the Netherlands, there has been no lack of the first. Dutch networks nowadays consist of individuals from a wide range of backgrounds and nationalities. Second, Dutch networks have been able to exploit opportunities in the narcotic drug market, most notably those arising from the Dutch government's unusual approach towards soft drugs in the 1970s and the open borders of the Schengen Agreement in the 1990s. The extensive trade network that has since developed is the key to the specific characteristics of the Dutch networks. Finally, the authorities did indeed respond slowly to international requirements for criminalizing amphetamine and MDMA, and we can also conclude that it took the police and the public prosecution service a relatively long time to react to the development of the Dutch networks in the 1970s and 1980s, and to subsequent drug production in the 1990s. When law enforcement agencies did finally step up measures against ecstasy production, they proved beyond a doubt that Dutch networks are not invincible. In the past decade, however, these networks have broadened and diversified, making it more difficult to tackle them solely by means of law enforcement action. Measures that would dramatically reduce market opportunities, such

as regulation of the sale and production of hashish and marihuana in other EU Member States, however, are politically unfeasible in the short and medium term.

Related Entries

- ▶ Careers in Organized Crime
- ▶ Drug Enforcement
- ▶ Drug Trafficking
- ▶ Gangs and Social Networks
- ▶ Network Analysis in Criminology
- ▶ Organized Crime, Types of
- ▶ Routine Activities Approach
- ▶ Social Network Analysis of Organized Criminal Groups
- ▶ Triads and Tongs

Recommended Reading and References

Adriaensen L (2007) Staatsvormend geweld. Overleven aan de Frontlinies in de Meierij van Den Bosch. Stichting Zuidelijk Historisch Contact, Tilburg

Boekhout van Solinge T (2008) De bestrijding van de cannabisteelt in Nederland in historisch perspectief.' In: Decorte T (ed) Cannabisteelt in de lage landen. Perspectieven op de cannabismarkt in België en Nederland. Acco, Leuven/Voorburg, pp 141–159

Bovenkerk F, Hogewind W (2003) Hennepteelt in Nederland: het probleem van de criminaliteit en haar bestrijding. Willem Pompe Instituut voor Strafrechtswetenschappen, Utrecht

Cohen L, Felson M (1979) Social change and crime rate trends: a routine activity approach. Am Sociol Rev 44:588–605

de Swaan A (1989) Zorg en de staat. Bert Bakker, Amsterdam

Egmond F (1994) Op het verkeerde pad. Georganiseerde misdaad in de Noordelijke Nederlanden 1650–1800. Uitgeverij Bert Bakker, Amsterdam

Fijnaut C (2013) The history of organized crime: a contribution to a more balanced approach. In: Paoli L (ed) The Oxford handbook of organized crime. Oxford University Press, Oxford,. (forthcoming)

Fijnaut C, de Ruyver B (2008) Voor een gezamenlijke beheersing van de drugsgerelateerde criminaliteit in de Euregio Maas-Rijn. Euregio Maas-Rijn, Maastricht

Fijnaut C, Bovenkerk F, Bruinsma G, van de Bunt H (1996) Eindrapport georganiseerde criminaliteit in Nederland. Sdu Uitgevers, The Hague

Grootswagers K (1983) De historie van het Loonse kerkdorp Kaatsheuvel. Uitgeverij Brabantia Nostra, Breda

Kleemans E, van den Berg E, van de Bunt H (1998) Georganiseerde criminaliteit in Nederland. Wetenschappelijk Onderzoek- en Documentatiecentrum, Ministerie van Justitie, The Hague

Kleemans E, Brienen M, van de Bunt H (2002) Georganiseerde criminaliteit in Nederland, Tweede rapportage op basis van de WODC-monitor. Wetenschappelijk Onderzoek- en Documentatiecentrum, Ministerie van Justitie, The Hague

Klerks P (2000) Groot in de hasj, Theorie en praktijk van de georganiseerde criminaliteit. Samson/Kluwer Rechtswetenschappen, Antwerpen

Korps Landelijke Politiediensten (2005) Criminaliteitsbeeldanalyse 2002–2004. Zoetermeer

Leuw E (1994) Initial construction and development of the official Dutch drug policy. In: Leuw E, Haen Marshall I (eds) Between prohibition and legalization. The Dutch experiment in drug policy. Kugler Publications, Amsterdam/New York, pp 23–40

Middelburg B (2001) De Godmother. LJ Veen, Amsterdam

Sietsma K (1985) Rechercheren naar de recherche. In: Fijnaut C, Kuijvenhoven A (eds) De recherche onder de loupe, inleidingen voor het recherchecongres op 14 en 15 oktober 1985 te Zutphen. J.B. van den Brink, Lochem

Spapens T (2006) Interactie tussen criminaliteit en opsporing. Intersentia, Antwerpen/Oxford

Spapens T (2008) Georganiseerde misdaad en strafrechtelijke samenwerking in de Nederlandse grensgebieden. Antwerpen/Oxford, Intersentia

Spapens T (2010) Macro networks, collectives and business processes: an integrated approach to organized crime. Eur J Crime Crim Law Crim Justice 18:185–215

Spapens T (2011) Interaction between criminal groups and law enforcement: the case of ecstasy in The Netherlands. Glob Crime 12:19–40

Spapens T (2012) Netwerken op niveau. Inaugural address Tilburg University, Tilburg

Spapens T, Bunt van de H (forthcoming) Soft drugs regulation and market dynamics in the Netherlands. European Journal on Criminal Policy and Research.

Spapens T, van de Bunt H, Rastovac L (2007) De wereld achter de wietteelt. Boom Juridische uitgevers, The Hague

van Duyne P (1995) Het spook en de dreiging van de georganiseerde misdaad. Sdu Uitgeverij, The Hague

van Duyne P, Kouwenberg R, Romeijn G (1990) Misdaadondernemingen: ondernemende misdadigers in Nederland. Gouda Quint, Arnhem

Weijenburg R (1996) Drugs en drugsbestrijding in Nederland. Vuga Uitgeverij BV, The Hague

Wubben H (1986) Chineezen en ander Aziatisch ongedierte, Lotgevallen van Chinese immigranten in Nederland, 1911–1940. De Walburg Pers, Zutphen

Dynamical Simulation in Criminology

Charlotte Gerritsen[1] and Michel C. A. Klein[2]
[1]Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), Amsterdam, The Netherlands
[2]Department of Computer Science, VU University Amsterdam, Amsterdam, The Netherlands

Overview

Simulation techniques from the area of Artificial Intelligence can be beneficial for the field of Criminology because they can be used to gain more insights in criminological phenomena (that were not clear based on just the informal theory), without actually having to experiment with these phenomena in the real world. The main goal of this article is to explain the process of modelling and simulation step by step and to illustrate the benefits of this approach for the field of Criminology.

Introduction

Within the field of Criminology a number of standard research methods exist to study criminal behavior, e.g., victim surveys, offender surveys, social experiments, and the analysis of police data. Based on these methods multiple theories have been developed that provide insight into delinquent behavior. However, these theories are usually "informal," meaning that they are written in natural language or described graphically and, thus, in principle ambiguous. In contrast, researchers from the areas of Computer Science and Artificial Intelligence have recently started investigating whether theories from Criminology can be translated into a formal, unambiguous, machine-readable notation, so that they can be used for *simulation* (e.g., Gerritsen 2010; Liu and Eck 2008; Malleson and Brantingham 2009). The main assumption behind that work is that simulation techniques can be beneficial, because they can be used to gain more insights in criminological phenomena (that were not clear based on just the informal theory), without actually having to experiment with these phenomena in the real world. The main goal of this article is to explain the process of modelling and simulation step by step and to illustrate the benefits of this approach for the field of Criminology.

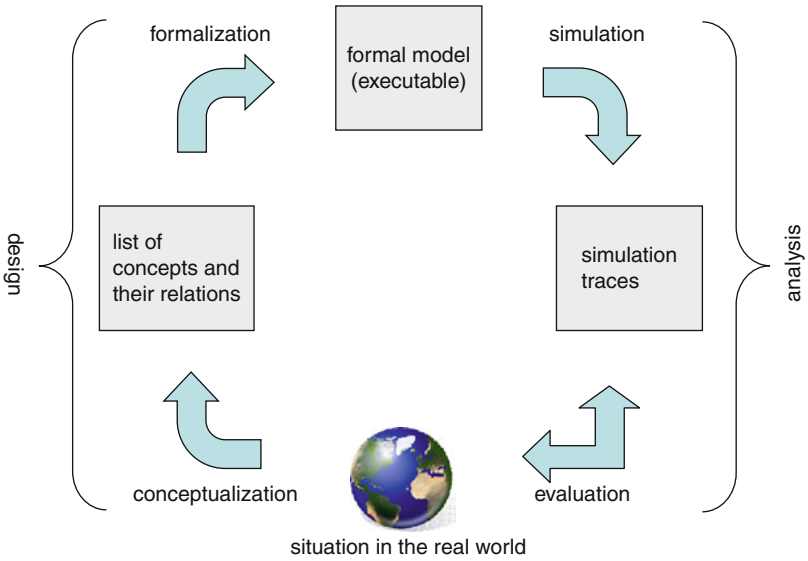
Modelling and Simulation

The main concept involved in the area of modelling and simulation is the concept of a *model*. A model is a representation of an object, system, or idea in some form other than that of the entity itself (Shannon 1975). Such a representation describes the most important concepts and their relations. However, it is almost never possible to describe all these aspects completely and unambiguously, e.g., because the exact relation between concepts in reality is not known or because not all factors that influence a concept are known. When building such a model it is important to realize that this has some important consequences:

1. A model is a simplification, not a complete representation of the reality.
2. A model is based on several choices and assumptions of the creator of the model.
3. Characteristics of the model are not necessarily characteristics of reality.

These consequences should be taken into account when working with models, as they imply that the conclusions that are drawn based on the model are based on the assumptions that are made.

In general, the main goal of a model is usually that it enables the user to study a certain process that exists in reality in a convenient manner. There may be several possible reasons why studying a model is more convenient than studying the process itself. One possible reason is that the process of interest does not yet exist. An example of such a situation is studying the impact of the introduction of surveillance cameras in a mall on people's behavior. Another reason can



Dynamical Simulation in Criminology, Fig. 1 General methodology for modelling

be that the process itself cannot be easily studied directly. For example, think about cognitive processes: it is not easy to measure things that go on in a human mind, while one can analyze relatively easily how a model of it behaves. In such a situation, a model may provide the researcher more insight in the process under investigation, even if it is not completely identical to the real world. One other common reason for studying models instead of the process in reality is prediction: one would like to see what happens in the future. Economical and weather models are examples of models that are often used for prediction, but also the prediction of the development of crime rates in a city is a possibility. Finally, experimenting using a model can be cheaper and more feasible than experimenting with real processes. For example, experimenting with different surveillance strategies is much more convenient in a simulated world than in the real world (Bosse and Gerritsen 2010).

In order to study a model, it is often useful if the model can imitate the dynamics of the process over time. This process is called *simulation*, and models that can be used to simulate behavior over time are called *dynamical models*. This type of model is the focus of this article. A simulation model imitates the dynamics of a process over time and thus helps to clarify the interaction between different aspects. The result of a simulation is a sequence of states of the model at subsequent time points, which is called a *trace*, *simulation trace*, or *simulation run*. The outcomes of a simulation are usually represented in a graphical form.

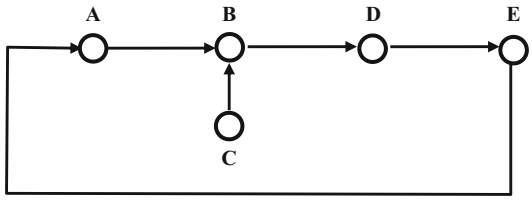
Modelling and Simulation Cycle

A general methodology for designing and analyzing models, the *modelling and simulation cycle*, is depicted in Fig. 1.

The modelling cycle consists of two main phases that comprise four steps. The first phase is the design phase in which the model is built. This phase consists of the conceptualization and the formalization of a process. The second phase is the phase in which the model is analyzed. This phase consists of the simulation of the process and the evaluation of the model. Below these four steps of the process are explained in more detail.

Conceptualization

Before starting the conceptualization, it is important to decide what exactly should be modelled. This means that it should be decided which process should be simulated and which questions need to be answered by using the model.



Dynamical Simulation in Criminology, Fig. 2 Graphical representation of the relationship between concepts

Then, there are two relevant aspects that need to be distinguished in the conceptualization step.

First, the relevant *concepts* need to be identified. What are the factors that play a role in a process? A concept can be an object but it can also be an event. Which concepts are important depends on the process that is modelled.

The second task is the specification of the *relationships* between the concepts that have been identified. At this moment it is specified *whether* different concepts influence each other, but not yet *how* they influence each other. This is specified in detail in the formalization step. The results of specifying the relationships between concepts can be a list of statements in the form of "A affects B," where it is possible that concepts are influenced by multiple concepts. It is also an option to represent these relationships in a graph. Each concept should be depicted as a node (circle) and the relationships as edges (arrows) between the nodes (see Fig. 2).

Formalization

In the formalization step, the concepts and relations between concepts specified in the conceptualization step are defined in more detail. The formalization depends on the representation mechanism that is used for the model. When considering existing simulation approaches, the following two classes can be identified (Bosse et al. 2007): *logic-oriented* approaches and *mathematical* approaches, usually based on difference or differential equations. Logic-oriented approaches are good for expressing qualitative relations but less suitable for working with quantitative relationships. Mathematical approaches are good for the quantitative relations, but expressing conceptual, qualitative relationships with them is hard to impossible. These representations will also be used in the case study presented in the next section. The exact form of the formalization (i.e., the formula's chosen) is a choice made by the modeller, based on the assumed form of relations between concepts. A number of standard ways of specifying relations are available to the modeller (e.g., logistic growth, weighted average).

Simulation Experiments

Depending on the representation of the model and the software tools used, it is possible to simulate the model and generate a simulation trace. To perform a simulation experiment a number of steps have to be taken. First the question or the pattern to be addressed by the experiment is formulated, for example, to investigate the amount of crime that is performed in a certain period. As a next step, choices have to be made with respect to the initial values of variables and the values of the parameters involved in the simulation model. Setting up a simulation experiment requires a structured plan on which values are chosen for these variables and parameters, thereby establishing a particular *scenario*. The settings of these values are determined by the process under investigation. If one is simulating the dynamics of crime over a city and the precise locations of the different houses are known, then this information can be used to set the parameters that represent the locations of the houses in the simulation. Alternatively, it is also possible to set parameters by using fictive information. Moreover, different scenarios can be used in order to compare the outcomes for different circumstances: changing the values in a number of different simulations. For a thorough evaluation, multiple scenarios need to be simulated.

Some models contain a stochastic element, i.e., parameters for which the values are determined by a probability. With this kind of models, running the same model several times results in possibly different traces, which all reflect the outcome of a different simulation for the same model.

Evaluation

Evaluation consists of verifying whether the model is a correct representation of the system that it represents. This can be done by formulating certain *properties* that hold in the actual situation and check whether they also hold in the model. These properties often express higher-level characteristics of the behavior of the process rather than the direct influences, for example, stable end situations (e.g., if all environmental factors are stable eventually also the amount of crimes will stabilize), or the effect of the occurrence of several events at the same time.

A distinction can be made between two types of properties:

1. Quantitative properties: statements about numerical characteristics of a model. Examples of quantitative properties are "the crime numbers never decline with more than 10 percent" or "the amount of burglaries increases with 5 percent each year."
2. Qualitative properties: statements about nonnumerical relations or characteristics of the model. Examples of qualitative properties are "eventually state property X will be true" (e.g., an offender will be arrested), "event A always occurs before event B," "parameter X (e.g., the crime rate) is always decreasing," "the simulation always reaches an equilibrium," etc.

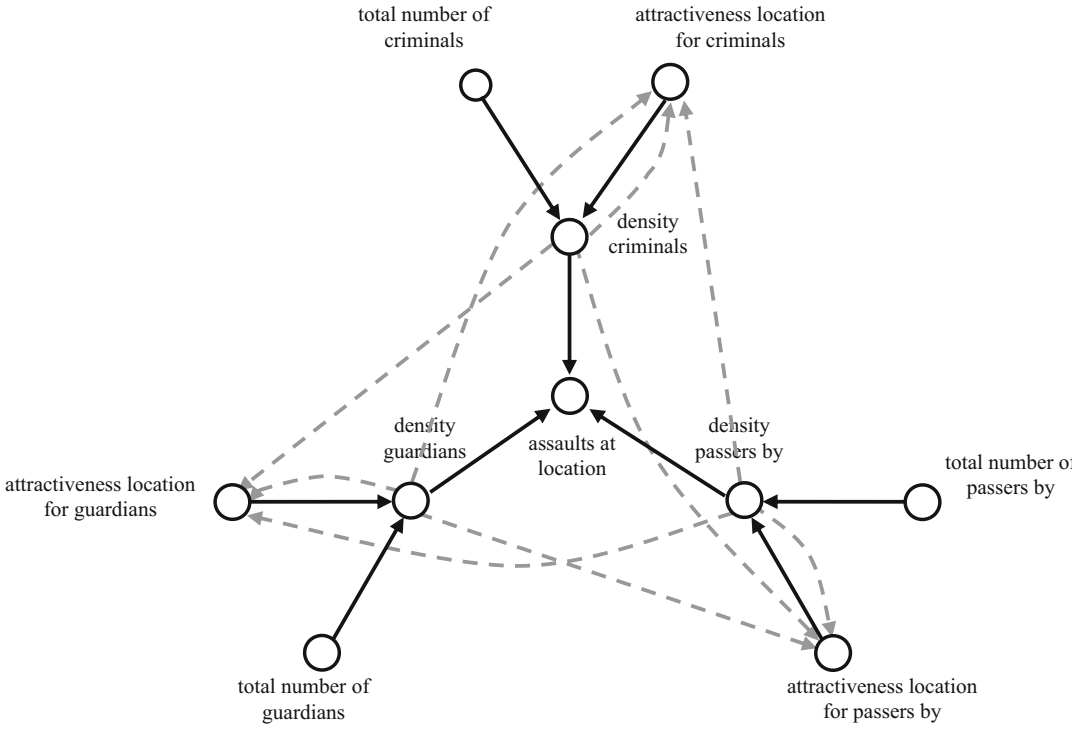
Various computer programs exist that automatically check properties of a simulation model.

When the evaluation shows that the model does not resemble the real process well enough, there may be three possible causes, namely, (1) a modelling flaw, (2) wrongly chosen values for the parameters that describe the influence of the concepts on each other, or (3) an invalid theory that was used as a basis. In the first case, the modelling cycle has to be continued with a reconceptualization phase. In the second case, better values for parameters have to be chosen. This can be done heuristically, by trying different values that result in a more realistic simulation or by mathematical approaches that find optimal parameter values. This latter mechanism is called *parameter tuning*. To apply parameter tuning, realistic data about the simulated process are required. The tuning process then finds parameter values that, when used in the model, achieve results that are as close as possible to the empirical data. In the third case (i.e., there is a discrepancy between the theory used as a basis for the model and the phenomena observed in the real world), the model can be used as a tool to find out in more detail where the theory is incorrect (e.g., by comparing simulation results with empirical data). The result of this step is usually that the theory itself is altered or refined.

When the evaluation gives a satisfactory result, the validated model can be used as tool for studying the process that has been described. Different hypothetical scenarios can be simulated and compared to improve the understanding of the process or evaluate the effect of changes in the scenario.

Case Study: Simulation and Analysis of the Dynamics of Hot Spots

In this section the modelling and simulation cycle is applied to a concrete example: the spatiotemporal dynamics of crime, more particular the emergence of the so-called criminal *hot spots*. This is one of the main research interests within Criminology. Hot spots are places where many crimes occur (Sherman et al. 1989). After a while the criminal activities shift to another location, for example, because the police has changed its policy and increased the number of officers at the hot spot. Another reason may be that the passersby move away when a certain location gets a bad reputation. The reputation of specific locations in a city is an important factor in the spatiotemporal distribution and dynamics of crime (Herbert 1982). For example, it may be expected that the amount of assaults that take place at a certain location affects the reputation of this location. Similarly, the reputation of a location affects the attractiveness of that location for certain types of individuals. For instance, a location that is known for its high crime rates will attract police officers, whereas most citizens will be more likely to avoid it. As a result, the amount of criminal activity at such a location will



Dynamical Simulation in Criminology, Fig. 3 Overview of simulation model for spatiotemporal dynamics of crime

decrease, which will affect its reputation again. Below, it is explained how the spatiotemporal dynamics of crime can be modelled by referring to the interaction of three groups, namely, *offenders*, *guardians*, and *passersby* (i.e., potential victims). This case study is inspired by the research presented in Bosse et al. (2011).

Conceptualization

In order to make a simulation model of the spatiotemporal dynamics of crime, several aspects are important. First, it is important to know the total number of the different groups involved, i.e., the *number of offenders*, *number of guardians*, and *number of passersby*. Next, it is assumed that the world (or city) that is addressed can be represented in terms of a number of different *locations*. It is important to know how many individuals (or *agents*) of each type are present at each location: the *density of offenders*, *density of guardians*, and *density of passersby*. Furthermore, to describe the movement of the different agents from one location to another, information about the *reputation* (or *attractiveness*) of the locations is used, so also this information is needed. This attractiveness is different for each type of agent. For example, passersby like locations where it is safe, e.g., locations where many guardians are present and no offenders. On the other hand, guardians are attracted by places where a lot of offenders are present, and offenders like locations where there are many passersby and no guardians. Finally, to be able to represent the idea of hot spots (locations where many crime takes place), the number of *assaults* per location is modelled. The idea is that more assaults take place at locations where there are many offenders and passersby and few guardians, cf. the *Routine Activity Theory* by (Cohen and Felson 1979).

The dynamic relationships are visualized in Fig. 3. To enhance clarity, the relationships that affect attractiveness of locations are depicted as gray dashed lines.

Dynamical Simulation in Criminology, Table 1 Mathematical formalization of concepts used for crime simulation model

Concept	Formalization
Total number of offenders	c
Total number of guardians	g
Total number of passersby	p
Density of offenders at location L at time t	c(L,t)
Density of guardians at location L at time t	g(L,t)
Density of passersby at location L at time t	p(L,t)
Attractiveness of a location L for a group x	β(L,x,t)
Attractiveness of a group y for a group x	B(x, y)
Assault rate at location L	assault_rate(L,t)

Dynamical Simulation in Criminology, Table 2 Logical formalization of concepts used for crime simulation model

Concept	Formalization
Density of a group at a location	density_of_at(G:GROUP, L:LOCATION, D:REAL)
Total number of a group	total_number_of(G:GROUP, N:REAL)
Attractiveness of a location for a group	has_attractiveness_for(L:LOCATION, G:GROUP, A:REAL)
Weight factor of attractiveness of group G2 for G1	are_attracted_by(G1:GROUP, G2:GROUP, B:REAL)
Assault rate at location L	assaults_at(L:LOCATION, A:REAL)

Formalization

In this section, as a next step, for each of the concepts introduced in the previous section, both a mathematical and a logical formalization (a logical atom) are introduced; see Tables 1 and 2. Note that these two forms of representation are presented purely as illustrations of the difference of logical and mathematical modelling, but both are formalizations of the same model.

Note that the first three concepts of Table 1 do not contain an argument t for time, since the values of these concepts remain constant over time.

Dynamical Simulation in Criminology, Table 3 Sorts used

Sort	Description	Elements
GROUP	Different groups in the society	{criminals, guardians, passers_by}
REAL	The set of real numbers	the set of real numbers
LOCATION	Different locations in the environment	{loc1, loc2, . . .}

Here, the variables A and B in the two attractiveness predicates are assumed to range over the interval [0, 1]. Note that some logical atoms make use of the so-called *sorts* (i.e., limited sets of elements that belong to a certain group, similar to types in programming languages). The specific sorts that are used in the presented model, and the elements that they contain, are shown in Table 3. Moreover, note that the attractiveness of certain groups for other groups is not shown in Fig. 3.

The dynamic relationships between these concepts are formalized below, both as mathematical formulae and as logical relationships. The first relationship determines how attractive each location is for the different groups (where it is assumed that $B11 + B12 + B13 = 1$):

SR1 (Simulation Rule 1)

at any point in time,	
if	the density of group G1 at location L is D1,
and	the density of group G2 at location L is D2,
and	the density of group G3 at location L is D3,
and	the total number of agents within group G1 is N1,
and	the total number of agents within group G2 is N2,
and	the total number of agents within group G3 is N3,
and	agents of group G1 are attracted by agents of group G1 with strength B11,
and	agents of group G1 are attracted by agents of group G2 with strength B12,
and	agents of group G1 are attracted by agents of group G3 with strength B13,
and	G1 and G2 and G3 are three distinct groups,
then	the current attractiveness of location L for group G1 is
	$B11^*D1/N1 + B12^*D2/N2 + B13^*D3/N3$

Mathematical formalization:

$$\beta(L,x,t) = B(x,c)\cdot c(L,t)/c + B(x,g)\cdot g(L,t)/g + B(x,p)\cdot p(L,t)/p$$

Logical formalization (note that the first line indicates the sorts to which the variables belong):

$$\forall G1,G2,G3:\mathrm{GROUP}\ \forall L:\mathrm{LOCATION}\ \forall D1,D2,D3,N1,N2,N3,B11,B12,B13:\mathrm{REAL}$$
$$\mathrm{density\_of\_at}(G1,L,D1)\wedge \mathrm{density\_of\_at}(G2,L,D2)\wedge \mathrm{density\_of\_at}(G3,L,D3)\wedge$$
$$\mathrm{total\_number\_of}(G1,N1)\wedge \mathrm{total\_number\_of}(G2,N2)\ \mathrm{total\_number\_of}(G3,N3)\wedge$$
$$\mathrm{are\_attracted\_by}(G1,G1,B11)\wedge \mathrm{are\_attracted\_by}(G1,G2,B12)\wedge \mathrm{are\_attracted\_by}(G1,G3,B13)\wedge$$
$$G1\neq G2\wedge G1\neq G3\wedge G2\neq G3$$
$$\twoheadrightarrow$$
$$\mathrm{has\_attractiveness\_for}(L,G1,B11^{*}D1/N1 + B12^{*}D2/N2 + B13^{*}D3/N3)$$

This formula expresses that the attractiveness of a location is based on some kind of reputation of the location for the respective types of agents. This reputation is calculated as a linear combination of the densities of all agents. Several variants of a reputation concept can be used. The only constraint is that they are assumed to be normalized such that the total over the locations equals 1. For example, a natural parameter setting for offenders would be to have B(c,p) high since offenders need victims to assault, and to have B(c,g) low because offenders try to avoid guardians. For the guardians, B(g,c) is likely to be high since offenders attract guardians, whereas B(g,p) is rather high as well. Finally, for the passersby the B(p,c) can be taken low as passersby prefer not to meet offenders, and B(p,g) (and possibly also B(p,p)) high because guardians (and other passersby) protect the passersby.

Based on this attractiveness function, the movement of the different types of agents between the locations can be determined:

SR2 (Simulation Rule 2)

at any point in time,	
if	the density of group G1 at location L is D1,
and	the total number of agents within group G1 is N1,
and	the current attractiveness of location L for group G1 is A1
then	the density of group G1 at location L is
	$D1 + \eta^{*}(A1 - D1/N1)^{*}N1^{*}\mathrm{delta\_t}$

Mathematical formalization:

$$c(L,t+\Delta t) = c(L,t) + \eta 1^{*}(\beta(L,c,t) - c(L,t)/c)^{*}c^{*}\Delta t$$
$$g(L,t+\Delta t) = g(L,t) + \eta 2^{*}(\beta(L,g,t) - g(L,t)/g)^{*}g^{*}\Delta t$$
$$p(L,t+\Delta t) = p(L,t) + \eta 3^{*}(\beta(L,p,t) - p(L,t)/p)^{*}p^{*}\Delta t$$

Logical formalization:

$$\forall G1:\mathrm{GROUP}\ \forall L:\mathrm{LOCATION}\ \forall A1,D1,N1:\mathrm{REAL}\ \mathrm{density\_of\_at}(G1,L,D1)\wedge \mathrm{total\_number\_of}(G1,N1)\wedge \mathrm{has\_attractiveness\_for}(L,G1,A1)$$
$$\twoheadrightarrow$$
$$\mathrm{density\_of\_at}(G1,L,D1 + \eta^{*}(A1 - D1/N1)^{*}N1^{*}\mathrm{delta\_t})$$

Here, the formula for offenders expresses that the density $c(L, t + \Delta t)$ of offenders at location L on $t + \Delta t$ is equal to the density of offenders at the location at time t plus a constant η (expressing the rate at which offenders move per time unit) times the movement of offenders from t to $t + \Delta t$ from and to location L multiplied by Δt. Here, the movement of offenders is calculated by determining the relative attractiveness $\beta(L, c, t)$ of the location (compared to the other locations) for offenders. From this, the density of offenders at the location at time t divided by the total number c of offenders (which is constant) is subtracted, and the result is multiplied with c, resulting in the change of the number of offenders for this location. For the guardians and the passersby similar formulae are used.

In order to measure the assaults that take place per time unit, also different variants of formulae can be used, for example:

SR3 (Simulation Rule 3)

at any point in time,	
if	the density of offenders at location L is C,
and	the density of guardians at location L is G,
and	the density of passersby at location L is PB,
then	the amount of assaults that take place at L is the minimum of γ1*C−γ2*G and PB

Mathematical formalization:

$$assault\_rate(L,t) = min(\gamma_1 \cdot c(L,t) - \gamma_2 \cdot g(L,t), p(L,t))$$

Logical formalization:

$$\forall L : \text{LOCATION } \forall C, G, PB : \text{REAL}$$
$$\text{density\_of\_at}(\text{offenders}, L, C) \wedge$$
$$\text{density\_of\_at}(\text{guardians}, L, G) \wedge$$
$$\text{density\_of\_at}(\text{passers\_by}, L, PB)$$
$$\twoheadrightarrow$$
$$\text{assaults\_at}(L, \min(\text{gamma1}^{*}C - \text{gamma2}^{*}G, PB))$$

Here, the assault rate at a location at time t is calculated as the minimum of the possible assaults that can take place and the number of passersby. Here the possible number of assaults is the capacity per time step of offenders (γ1) multiplied by the number of offenders at the location minus the capacity of guardians to avoid an assault (γ2) times the number of guardians. In theory this can become less than 0 (the guardians can have a higher capacity to stop assaults than the offenders have to commit them), therefore the maximum can be taken of 0 and the outcome described above.

Simulation

This section presents an example simulation trace that was generated on the basis of the simulation model for the spatiotemporal dynamics of crime. The parameter settings used for the simulation are presented in Table 4. As shown there, the simulation comprises four locations (L1, L2, L3, and L4), 800 offenders, 400 guardians, and 4,000 passersby. The initial distribution of these agents over the four locations is shown in Table 5.

Dynamical Simulation in Criminology, Table 4 Parameter settings

Simulation length	100
Locations	4
Number of offenders	800
Number of guardians	400
Number of passersby	4,000
Speed factor η (for all groups)	0.5
Δt	0.1

Dynamical Simulation in Criminology, Table 5 Initial distribution of agents

	L1	L2	L3	L4
Offenders	300	100	250	150
Guardians	50	150	125	75
Passersby	1,500	500	750	1,250

For the attractiveness function (SR1), the following formulae have been used:

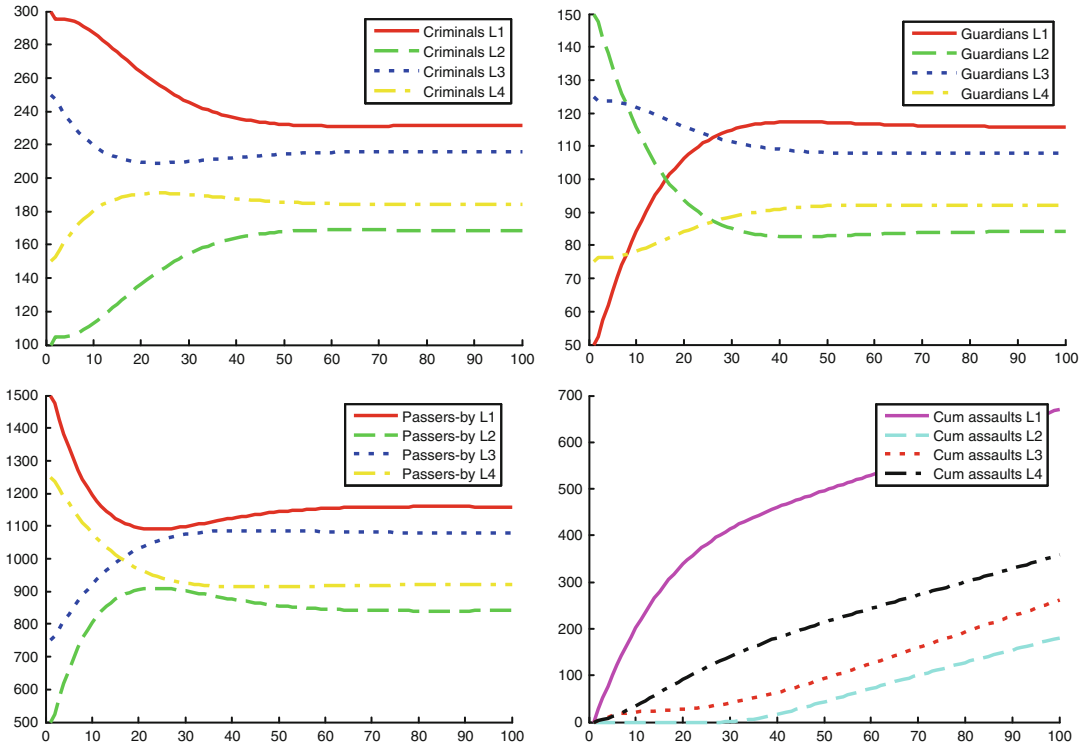
$$\beta(L,c,t) = p(L,t)/p \text{ for offenders}$$
$$\beta(L,g,t) = c(L,t)/c \text{ for guardians}$$
$$\beta(L,p,t) = g(L,t)/g \text{ for passers} - \text{by}$$

Note that this corresponds to the following settings for the different B(x,y):

$$B(c,c) = 0, B(c,g) = 0, B(c,p) = 1$$
$$B(g,c) = 1, B(g,g) = 0, B(g,p) = 0$$
$$B(p,c) = 0, B(p,g) = 1, B(p,p) = 0$$

Obviously, more complex attractiveness functions can be used as well.

A simulation trace for this situation is depicted in Fig. 4. Note that this simulation has been performed in the MATLAB simulation environment. The first three graphs depict the movement of, respectively, offenders, guardians, and passersby over the different locations. The last graph depicts the amount of assaults performed.



Dynamical Simulation in Criminology, Fig. 4 Example simulation trace for crime simulation model

As shown in Fig. 4, from the beginning of the simulation many passersby move away from location 1 (where there are many offenders and few guardians) and towards location 2 (where there are many guardians and few offenders). The guardians follow the opposite pattern: they move away from location 2 and towards location 1. As soon as the number of guardians at location 1 has increased, this location becomes more attractive for the passersby. The offenders first move away from location 1, towards location 2, but as soon as the passersby come back to location 1, a significant part of the offenders stays there. Eventually, all populations stabilize.

Although this is just one example simulation trace, it illustrates the power of simulation to study criminological processes. The way the simulation model has been set up enables the researcher to full in any desired combination of parameter settings, thereby reproducing (an artificial instance of) a particular real-world situation of interest. For example, one can manipulate the number of locations, the ratio between the different types of agents, or the attractiveness functions. By running simulations for these different settings, one can gain more insight in how the corresponding scenarios might develop in reality.

Evaluation

As shown, the model presented in the previous sections is capable of producing very interesting patterns. For example, eventually the number of offenders, guardians, and passersby on each location stabilizes. Nevertheless, as yet no guarantee has been given that these patterns are realistic. In other words, the model has not been validated (with empirical data). In order to fully validate a model, an obvious approach is to compare the output of the model with empirical data. For the presented case study, ideally this would mean that one would compare for each time

point of the simulation whether the numbers of offenders, assaults, and so on would match the numbers if the process would take place in reality. However, for obvious reasons, such a detailed comparison is not always feasible. Moreover, it is even not always necessary, since the results of a simulation model that is not fully validated can still provide interesting insights to the researcher (as long as one is confident that the basic mechanisms of the model are plausible). For this reason, sometimes a more modest form of validation can be performed. For example, it is possible to specify higher-level requirements about the behavior of the model. These requirements might result from knowledge by domain experts. For example, if it is known from empirical data that hot spots never move if the characteristics of the locations stay the same, the following property might be formulated. This property expresses, in a semiformal notation, that after a certain time point t, the densities of all three groups stay around a certain equilibrium:

GP1 (Global Property 1): Stable Number of Groups at Locations

For all traces, there is a time point t such that	
for all time points t1 and t2 after t, for all groups g and locations l,	
if	the total amount of agents of group g is n,
and	at t1 the density of group g at location l is x,
and	at t2 the density of group g at location l is $x2$,
then	the difference between x and $x2$ is smaller than α % of n.

Using an automated tool for checking properties of simulation runs (Bosse et al. 2009b), property GP1 has been checked against the trace shown in Fig. 4. It was found, among others, that for an α of 1.0 (i.e., 1 %), stabilization of offenders occurs at time point 35, stabilization of guardians occurs at time point 28, and stabilization of passersby occurs at time point 38. Assuming that one would have similar information about a corresponding real-world process (e.g., one would know for a given scenario how long it takes until each group stabilizes), then a comparison of these numbers can be used to confirm that the model shows plausible behavior.

Also other types of properties can be specified. Some examples are the following (only in informal notation):

- "for each group, the attractiveness of each location is always between 0 and 1" (checking this property can be useful for the modeller to check whether the model is free of errors that make the variables go out of their ranges).
- "between time points t1 and t2, the density of group g at location l is monotonically increasing/decreasing."
- "for all traces $\gamma 1$ and $\gamma 2$, if there are more locations in $\gamma 1$ than in $\gamma 2$, then at the end of the simulation there will be less crime in $\gamma 1$."

Thus, checking certain properties allows the modeller to investigate whether the model behaves according to the expectations. Note that it is part of an iterative process: when the model does not behave according to the expectations, you return to the first step of the modelling and simulation cycle (conceptualization), in order to improve and refine its basic mechanisms.

Discussion

Within the area of Criminology, analysis of complex social and environmental processes, such as the spatiotemporal dynamics of crime, is an important challenge. For instance, criminologists are interested in the question where criminal hot spot may emerge and when they will emerge. Answering such questions is not easy since the large number of factors involved can make the process quite complicated. Moreover, it is often not feasible to perform experiments with these processes in the real world. The use of computer simulation can be a solution to this problem. Computer simulation may help to investigate complex (spatiotemporal) processes within criminology in a relatively fast and cheap manner.

In this article, the different methodological steps involved in modelling and simulation have been explained, and the methodology has been illustrated for a particular case study. Note that the main goal of the case study is to help researchers in their theory building, to shed

more light on the process under investigation. As such, the presented model can be used as an analytical tool to see how certain aspects influence the spatiotemporal dynamics of crime. This tool can for example be used to investigate a fictive process, e.g., how do passersby, guardians, and offenders react to each other when their ratio is 4:1:2? And what happens when the number of guardians is increased?

As soon as the model can be shown to be sufficiently realistic (i.e., the global patterns produced by the model match reality), it may also become of interest for policy makers, e.g., to study certain "what-if" scenarios. For example, one may investigate how the crime level of a certain city will change if the policy makers invest in more surveillance in a certain area (Bosse et al. 2010b). Other interesting insights from simulation models in the criminological domain are the finding that anticipating strategies for police investments seem to be more effective than reactive strategies (Bosse et al. 2010b); the finding that school and parents have a relatively large influence on delinquent behavior among juveniles (Bosse et al. 2009a); and that for a reasonable amount of guardians, hot spot policing seems to be more effective on the crime rates than area hot spot policing (Bosse et al. 2010).

When one considers the level of detail of similar tools that have been proposed in the literature (such as Baal 2004; Brantingham et al. 2005; Groff 2005; Liu et al. 2005; Melo et al. 2005; Reis et al. 2006), it turns out that different perspectives are taken. For example, some authors have attempted to develop simulation models of crime displacement in existing cities, which can be directly related to real-world data (e.g., Liu et al. 2005), whereas others deliberately abstract from empirical information (e.g., Bosse et al. 2010). The point of view taken in the current article can be situated in between these two extremes. Initially, the simulation model was developed to study the spatiotemporal dynamics of crime per se. However, its basic concepts have been defined in such a way that they can be directly connected to empirical information, if this becomes available. As a result, an interesting challenge for future work will be to explore the possibilities to connect the basic concepts of the model to real-world data.

Related Entries

- ▶ Agent-Based Modeling for Understanding Patterns of Crime
- ▶ Agent-Based Models to Predict Crime at Places
- ▶ Simulation as a Tool for Police Planning

Recommended Reading and References

Baal PHM van (2004) Computer simulations of criminal deterrence: from public policy to local interaction to individual behaviour. Ph.D. Thesis, Erasmus University Rotterdam, Boom Juridische Uitgevers

Bosse T, Gerritsen C (2010) An agent-based framework to support crime prevention. In: Hoek W van der, Kaminka GA, Lespérance Y, Luck M, Sen S (eds) Proceedings of the ninth international conference on autonomous agents and multi-agent systems, AAMAS'10. ACM Press, pp 525–532

Bosse T, Jonker CM, van der Meij L, Treur J (2007) A language and environment for analysis of dynamics by simulation. Int J Artif Intell Tools 16(3):435–464

Bosse T, Gerritsen C, Klein MCA (2009a) Agent-based simulation of social learning in criminology. In: Sharp B (ed) Proceedings of the international conference on agents and artificial intelligence, ICAART'09. INSTICC Press, pp 5–13

Bosse T, Jonker CM, van der Meij L, Sharpanskykh A, Treur J (2009b) Specification and verification of dynamics in agent models. Int J Coop Inf Syst 18:167–193

Bosse T, Elffers H, Gerritsen C (2010) Simulating the dynamical interaction of offenders, targets and guardians. Crime Patterns Anal 3(1):51–66

Bosse T, Gerritsen C, Hoogendoorn M, Jaffry SW, Treur J (2011) Agent-based versus population-based simulation of displacement of crime: a comparative study. Web intelligence and agent systems: an international journal, IOS Press, to appear

Brantingham PL, Glässer U, Singh K, Vajihollahi M (2005) Mastermind: modeling and simulation of criminal activity in urban environments. Technical report SFU-CMPTTR-2005–01. Simon Fraser University

Cohen LE, Felson M (1979) Social change and crime rate trends; a routine activity approach. Am Sociol Rev 44:588–608

Gerritsen C (2010) Caught in the act: investigating crime by agent-based simulation. Ph.D. Thesis. VU University Amsterdam

Groff ER (2005) The geography of juvenile crime place trajectories. Ph.D. Thesis. University of Maryland
Herbert DT (1982) The geography of urban crime. Longman, Harlow
Liu L, Eck J (eds) (2008) Artificial crime analysis systems: using computer simulations and geographic information systems. Hershey, PA
Liu L, Wang X, Eck J, Liang J (2005) Simulating crime events and crime patterns in RA/CA model. In: Wang F (ed) Geographic information systems and crime analysis. Idea Group, Singapore, pp 197–213
Malleson N, Brantingham P (2009) Prototype burglary simulations for crime reduction and forecasting. J Int Symp Environ Criminol Crime Anal 2(1):47–66
Melo A, Belchior M, Furtado V (2005) Analyzing police patrol routes by simulating the physical reorganisation of agents. In: Sichman JS, Antunes L (eds) Multi-agent-based simulation VI, proceedings of the sixth international workshop on multi-agent-based simulation, MABS'05. Lecture notes in artificial intelligence, vol 3891. Springer Verlag, 2006, pp 99–114
Reis D, Melo A, Coelho ALV, Furtado V (2006) Towards optimal police patrol routes with genetic algorithms. In: Mehrotra S, et al. (eds) ISI 2006. LNCS 3975, pp 485–491
Shannon RF (1975) Systems simulation. Prentice-Hall
Sherman LW, Gartin PR, Buerger ME (1989) Hot spots of predatory crime: routine activities and the criminology of place. Criminology 27:27–55

Dyssocial Personality Disorder

► Psychopathy and Offending