Cambridge Centre for Evidence-Based Policing

*Professor Lawrence W. Sherman, Chief Executive*

50 KEY CONCEPTS IN EVIDENCE-BASED POLICING

Part I: General Concepts in Evidence-Based Policy & Practice

# PART I A. : TESTING

## 1. A Program

A well-defined set of

a. activities (outputs) delivered
b. in a specific sequence, ideally with a
c. specific dosage, by people in
d. specific roles with
e. specific training and authority, to
f. other people who, or other units of delivery that, meet the
g. eligibility criteria for receiving the outputs.

**Negative Example:** Integrated offender management (IOM), said to be the same as preceding practices, yet with "more meetings." This label received hundreds of millions, but never met the definition of a program.

**Positive Example:** GPS-measured police patrol in Homicide hot spots of Trinidad.

## 2. Randomized Controlled Trial (RCT)

A research design intended to estimate the effects of a target program by

a. Identifying a substantial number of people or other units of analysis,
b. Selecting all units eligible for the target program and the RCT,
c. Recording the identifying details of each eligible unit,
d. Assigning some units (treatment group) to receive the program, while
e. Insuring that no other units receive the target program, then
f. Delivering the assigned treatments as randomized, and
g. Comparing outcomes for the two groups
h. During or after an appropriate followup period has elapsed.

## 3. Program Effects

The size of the difference in ***outcomes*** between the units in

a. The treatment group in any randomized experiment or quasi-experiment, and
b. The randomly assigned or appropriately matched comparison group.

## 4. Outcomes

In any analysis of a system designed to change the people or materials it processes, outcomes denote the end stage of the process, in which

a. Inputs of resources, including people and technologies, have been used to produce
b. Outputs of activities, the things that the system does with or to the people or units it processes, in order to produce
c. A more desirable state of those treated units, called "outcomes."

**Note:** the ambiguity of this term is found in its use for control groups of units that have not been processed by a particular system, but can be measured for the same "states" as units that were treated.

**Example:** Crime victims who meet their offenders in restorative justice conferences have less measured post-traumatic stress after the crime that those who do not, yet both groups have post-traumatic stress to some degree.

## 5. Long-term followup

Generally, in evidence-based policy, a follow-up period is considered long-term if it exceeds two to five years after treatment was delivered and outcome measures were collected. Two to four decades of followup have now been reported, raising the possibility of detecting many more program effects, including cost-effectiveness.

## 6. Small sample size

Generally, a small sample size is less than 100 units in each group under study, the treatment group and a control group.

**Note:** Unless the units are all highly similar in outcome measures to begin with, there is less statistical power with smaller samples than with larger ones of equal similarity. That means that true effects may not be found.

## 7. Departure from random assignment/treatment integrity

When units are not treated as randomly assigned, the research conclusion is threatened by the possibility of the remaining units that are treated as assigned being biased towards a certain direction in their outcomes. Note that:

a. In general, the more departure from random assignment in actual treatment, the more threat of a biased and internally invalid conclusion.
b. There is no clear boundary between "too much" and "not too much" departure from random assignment.
c. The best way to handle the problem is to analyse the units randomly assigned to a treatment group by that assignment, rather than by what actually happened to them. This practice is called "intention-to-treat" (ITT).
d. The effect of ITT may be to underestimate (or overestimate) the size of the effect (if any) of a randomly assigned treatment.

**Example:** Many hot spots policing experiments have had modest success in reducing crime in treatment locations compared to controls, but with some treatment hot spots receiving very little extra patrol. This tends to underestimate or over-estimate the size of the effects on outcomes had the full dosage of patrol been delivered as assigned.

## 8. Replication

Repeating an experiment, or any test, done in exactly the same way with a different sample, or in a different place, or with different units of analysis, as the original experiment.

Note: the more differences there are between the original experiment and the replication, the less value the replication has as a check on the original experiment. If the new results differ, the true reason could be that the different elements of the experiment were the cause of different outcomes.

## 9. Evidence-Based Testing

A true, evidence-based test of any hypothesis compares systematically observed outcomes found when treating exactly similar units in exactly different ways, with exactly the same measurement before, during and after the test.

Note: Testing is not just a matter of "trying something new," such as giving police BWV cameras to wear to "see what happens." Unless there is a similar group of officers facing similar work challenges without BWV cameras, there is no way to reliably identify the unique effects of the cameras—as distinct from some other reason why the outcomes may differ between officers given cameras and those not. A test, by definition, must provide an opportunity for a hypothesis to be proven wrong. The Cambridge Police Executive Programme motto is "Don't just *try* it—*test* it."

## 10. Inputs

See "outcomes" above. Inputs are

a. resources of money, people or materials that
b. may or may not be used to
c. produce activities or outputs that
d. may or may not affect outcomes

Examples: Inputs include police recruitment and training, cars, radios, cameras, computers, stations, lockup cells and uniforms.

## 11. Outputs

See outcomes and inputs above. Outputs are

a. Activities that systems undertake for the stated purpose of
b. Accomplishing some change in the people or units being processed that will affect the
c. Outcomes for those units at a future point in time.

Examples: In policing, outputs include patrolling the streets, taking and investigating crime reports, negotiating disputes, making arrests, and other duties.

## 12. Control Group

In any test of a hypothesis, a (pure) control group consists of people or other units that receive *nothing new* by way of outputs, while a treatment group of similar units receives a different program. A modified control group, sometimes called a comparison group, consists of people or other units of analysis that receives a different program from the program received by the treatment group.

Example: The control area in the Kansas City Gun Experiment received no new program in policing gun crime, while the treatment area did receive a

major new program of proactive encounters with individuals found in gun crime hot spots who were suspected of carrying or using weapons.

## 13. Hypothesis

This word can mean many things, all of which mean a claim that a systematic observation that has not yet been made will find a specific result. That observation can be with or without experimental manipulation.

a. An experimental hypothesis claims a difference in outcomes will be found between a treatment and a control group in a planned experiment that manipulates which outputs are administered to each unit.
b. A non-experimental hypothesis, also called a "prediction," simply states what a descriptive study will find, such as "Americans will report less respect for the police in the 2015 surveys than they did in the 2014 surveys."

## 14. Rival Hypotheses

Rival hypotheses consist of two or more claims about *why* a systematic observation has found something to be true, both (or all) of which are consistent with the facts.

Example: Displacement. When crime goes down in hot spots policing experiments, at least two rival hypotheses are suggested to explain *why* crime went down:

a. The crimes that would have happened at the treatment group hot spots did not happen because the extra patrol prevented the crime.
b. The crimes that did not happen at the treatment group were displaced to some other location, so that no net crime prevention effect resulted from the extra patrols at the hot spots.

## 15. Conclusion/Causation

A conclusion about causation—a cause and effect relationship between output and outcome—depends on the capacity to eliminate rival hypotheses. Until rival hypotheses can be or have been eliminated, the causation of any difference in outcome between two groups remains uncertain. Under that condition, conclusions about causation cannot be drawn on the basis of the evidence.

## 16. Correlation

Correlation is a pattern of two variables moving together in a predictable way, either

a. Positive (direct)—if X goes up, then Y goes up—or
b. Negative (inverse)—if X goes up, then Y goes down.

Note: The degree of correlation can only vary between perfect (100% positive or 100% negative) and non-existent (correlation = zero.).

Note: the size of an effect is infinite, with no limit to how much crime can go up, and no limit but zero to the effect size of a reduction in crime.

Note: **Correlation does not prove causation.** Correlation is a necessary, but not sufficient, condition for demonstrating a causal relationship between two variables. Rival hypotheses (see above) must be eliminated to show causation.

## 17. Statistical Power

Statistical "power" is a feature of a test plan (research design) that determines

a. How likely the design is to detect a true effect as not due to chance, based on
b. The size of the sample, as well as
c. The amount of differences (variance) in characteristics across the units being assigned to both control and treatment groups, and
d. The actual size of the effect.

Note: power can be estimated in advance of a test by predicting, or assuming, different effect sizes, and even different degrees of variance within the samples actually obtained for the test.

## 18. Confidence Intervals/Range of Error

Depending on the power of a test, the range of error around any result is said to vary from large to small. This range is

a. called the 95% "confidence interval" between the highest and lowest points of
b. the *estimate* in the research design of the
c. larger universe of units from which the study sample was drawn.

## 19. Statistical Significance

Statistical significance is a term many statisticians apply to any result that has confidence intervals which do not include "zero" effect.

Note: What is "statistically significant" may be practically insignificant, if the effect size is small (or not cost effective) or when the result is due to a very powerful test, in which "significant" findings with small confidence intervals can occur by chance.

Note: What is not "statistically significant" may be very significant in a practical sense, such as program that is highly cost-effective. That is why many statisticians advise against making a 95% significance level into a "cult" of absolute good or worthlessness. Instead, they counsel acting on a cost-effective result with "borderline" significance and repeating the experiment if possible with a more powerful test.

## 20. Internal and External Validity

The validity of a conclusion based on research evidence can be assessed within the boundaries of any particular study, or it can be questioned as to its generalizability across the world. Validity depends largely on ruling out rival hypotheses (see above).

INTERNAL validity is the success of a study ruling out rival hypotheses within the study itself.

EXTERNAL validity is the success of a study in generalizing to other places and times from where and when the research was done.

NOTE 1: While internal validity can be assessed using principles of research design, external validity can only be assessed on the basis of replication.

NOTE 2: If a study is not internally valid, it is unlikely to be externally valid.

NOTE 3: Even an internally valid study may not be externally valid, since the initial findings may have depended on some unique features of the local context at the study site.

## 21. Cost-Effectiveness

The cost-effectiveness of a program is the difference between the cost of producing the program and the cost of the harm or crime it succeeds in preventing. "Cost" of crime can be estimated in various ways; the easiest is to estimate the amount of crime a particular activity can prevent, based on experimental evidence. The crime prevention benefit of hot spots policing, for

example, can be estimated by experiments. Any further experiments that do not reduce as much crime, or crime harm, as hot spots police patrol could be deemed to be cost-ineffective.

# PART I.B. TARGETING

## 22. Evidence-Based Targeting

Evidence-base targeting is a quantitative process of

a. Identifying every possible case on one kind of unit of analysis, such a 100% of residential addresses.
b. Identifying an outcome criterion, such as the number of burglaries in 2015 across all residential addresses.
c. Rank-ordering, from highest to lowest, all of the cases of this unit of analysis, in terms of the volume or seriousness of the chosen outcome—such as the address with the most burglaries in 2015, followed by the second-most, third-most, etc.
d. Assigning police resources in light of these rankings, as well as other considerations, including the availability of well-tested programs for preventing or reducing burglaries.

## 23. Predictions

A prediction, some statisticians say, is a precise statement of exactly where and when, or by or to whom, something is going to happen, made before the event in question can occur. A prediction has no confidence interval, no probability of error. It can only be 100% right or 100% wrong.

Note: A prediction can even be within a narrow range, such as

Example: There will be residential burglaries between 5th an 6th street on Avenue C between 11 am and 3 pm next Wednesday through Friday.

## 24. Forecasts

A forecast is a statement of the probability of an event happening within a range of error, giving estimated confidence intervals. The probability can range from near-zero to 100%, based on calculations from data on previous events of that kind.

Example: Based on data from 1,000 previous days of televised football matches in Sydney, Australia, we forecast a 36% chance of a domestic

homicide occurring during next week's football match, with a range of error from 16% to 56%.

## 25. Statistical Prediction

Statistical prediction of events in individual cases, such as specific offenders, is done by applying the forecast for a large group of people with similar characteristics to one member of that group—*while acknowledging the estimated rate of error*. It is based on large samples of similar units, using an "algorithm" or formula for taking each characteristic of the group members into account, such as their age, gender, prior crime histories, employment history, crime rate in the area of their residential address, age at first prosecution, etc.. The same approach can be used for screening applicants for police agencies, or assessing police officers attracting repeat complaints.

Note: Unlike the fictional "pre-crime" analysis of the people who are about-to-commit-murder in the next hour in the movie and Philip K. Dick novel MINORITY REPORT, statistical prediction is far more accurate over a multi-year period than it is over very short periods.

## 26. Clinical (Subjective) Prediction

Clinical prediction makes claims about individual cases either doing something or not doing something in the future, based on qualitative experience with similar cases in the past.

Note: The prior experience with *very* similar cases as the basis for these predictions may be small or zero, at least in comparison with statistical forecasting drawing on 30,000 or 50,000 cases, with hundreds of cases matched on 20 or more characteristics (such as age or gender or prior crimes). No one individual can ever experience as many cases as a "big data" set can analyze.

Note: Many serious offenders are still released from prison based on clinical (subjective) predictions that they will not repeat violent crimes. Many police officer applicants are hired or rejected on the basis of clinical predictions. Racial and ethnic biases in these predictions have rarely been analyzed.

# PART I C. TRACKING

## 27. Evidence-Based Tracking

Evidence-based tracking is a police activity that periodically reviews

a. Continuous *measurements* of
b. What police are doing, where, when and how, in relation to
c. Targeting analyses of shifting patterns of where when and how crime or harm is occurring, in order to
d. Decide whether police practices should be re-targeted or
e. What feedback should be given to police leaders and officers

Example: Tracking of police patrols in Trinidad increased daily patrol time in homicide hot spots from an average of 20 to 130 minutes per day, per hot spot, while homicide dropped from 9% to 45%.

## 28. Measurement

Measurement is a quantitatively coded observation of something happening in the real world, using consistent definitions and technologies. If measurement is not consistent and unbiased in its inputs and outputs,

a. it may be unreliable in its measurements
b. the unreliability may make findings and conclusions invalid;

Note: Experimental measures of both treatments and control units must be identical, in order to isolate differences in outcomes due to treatment effects.

## 29. Implementation

Implementation is the process by which a program designed on paper or in principle is launched into practice. Implementation is by definition successful when

a. Most of the program elements are found to have been put into place
b. Required activities of the program occur on a recurrent basis
c. The qualitative dimensions of program elements are delivered
d. The correct quantities of dosage of program elements are also delivered.

Implementation cannot be successful if there is no tracking in place to determine whether the program is being delivered as it was designed to be.

## 30. Feedback on Implementation

When good evidence shows that operational personnel do not carry out a program as instructed, corrections of these errors are unlikely to occur unless direct feedback is provided to the personnel in question. Feedback can include informal or formal conversations, written notices, or transfer of personnel. Feedback provided in an undiplomatic manner, however, make things worse, by causing labour relations issues or even lower levels of implementation.

Note: One study of feedback (Makkai and Braithwaite 1994) found that a firm statement of higher expectations—"You are better than that"—evoked better compliance with program elements than either a too-soft or too-harsh interaction.

# Part II: Key Concepts for Policing

## 31. Deterrence and Prevention

Deterrence is a sub-category of crime prevention. Deterrence is the prevention of crime by making potential offenders too afraid of being caught and punished to commit a crime.

## 32. Situational Prevention

Situational prevention is any method of making it more difficult for people to do something harmful.

Example: ignition interlocks on autos owned by convicted drink drivers requires an automatic breath test to yield a legal-to-drive blood alcohol reading before the car will start. This makes drink driving more difficult, while not impossible, and has lowered rates of repeat drink driving.

## 33. General Deterrence

General deterrence is the effect of any threat of punishment on the entire population, including past offenders and those who have never offended.

## 34. Certainty of Punishment

The probability that a given crime will be followed by any unpleasant consequence administered by the state is generally called "certainty," as in "HOW" certain or uncertain punishment may be.

NOTE: Certainty is widely credited as being the most important dimension of both general and specific deterrence.

## 35. Severity of Punishment

The extent of pain or suffering inflicted on an offender as punishment is usually measured by the number of days of imprisonment associated with the offence, or even years of suspended sentences or community punishment hours.

## 36. Celerity (speed) of punishment

The speed of punishment is measured by the number days or even minutes between an offender committing a crime and a state invoking legal punishment of some sort, including arrest. The shorter the time period, the greater the "celerity."

Note: This dimension of deterrence is the least-studied element of criminal justice, but may be very promising. Speedy punishment was a key element of the Turning Point experiment in Birmingham UK.

## 37. Sword of Damocles Effect

A "sword of Damocles" effect is the deterrence of crime based upon the prospect of an instant punishment occurring if the offender is arrested for having committed another offence. This concept applies equally to suspended sentences, to random drug testing of offenders on probation, and to diversionary programs such as Turning Point.

## 38. Crackdown

A police crackdown is a sudden and substantial increase in the probability or severity of sanctions imposed by police on specific offences, or in specific places, at specific times.

## 39. Initial Deterrence

A crackdown that is followed by a sharp drop in targeted crime is said to have created "initial deterrence," insofar as crime is down for the moment, although past evidence suggests it will slowly return towards pre-crackdown levels—even if the crackdown is maintained.

## 40. Deterrence Decay

A gradual increase in crime after an initial post-crackdown drop is called a deterrence decay, as long as the crackdown itself remains in effect. If the crackdown is terminated, a crime increase would just deterrence lost.

## 41. Residual Deterrence

Residual deterrence is a period after a crackdown has ended, but initial deterrence has not been lost. The period in which crime remains below pre-crackdown levels can be called a "residue" of deterrence, in which the deterrent threat of the recent crackdown still lingers in the perceptions of those contemplating criminal acts.

## 42. Koper Curve

The Koper Curve was discovered by Professor Christopher Koper in his 1995 analysis of the systematic observations of police presence and crime in the Minneapolis Hot Spots Patrol Experiment. Across over 7,000 observations, the longer the police remained at a crime hot spot, the longer it took for the first incident of crime or disorder to be observed after they left. This relationship was observed for the first 15 minutes after a police unit first arrived, and has been called a period of residual deterrence. After 15 minutes, the period of residual deterrence actually declined the longer police remained in the hot spot. Thus the period of 10 to 15 minute patrols was deemed optimal in those locations.

Note 1: This finding was a non-experimental correlation; prediction, not causation.

Note 2: The Koper Curve analysis has never been replicated.

## 43. Hot Spots of Crime

There is no standard definition of the unit of analysis for crime hot spot. Various studies have used street addresses, street segments, circles of various radius size, or street-based grids. The one common denominator is that hot spots are almost always smaller than a full police beat, AND they all have far higher than average amounts of crime per square foot. Depending on the city, crime in hot spots can total 50% or more in less than 10% of the land mass.

Note: The pattern has been found in cities, both large and small, and in many countries, but lower-population density areas have rarely been studied.

## 44. Crime Harm Index

A crime harm index uses sentencing policy to assign a "weight" of severity to each offence type. It then creates an index out of annual crime counts, by

a. Multiplying the number of days in prison for each offence type times
b. the number of offences of that type, and
c. summing the products of total days of theoretical imprisonment for all types of offences.

This Index can be calculated within societies or cities over time, across areas within cities, and across offenders and victims in their individual crime histories.

## 45. Police Legitimacy

Police legitimacy the degree to which both public and police perceive (or feel) that police have a moral right to impose force on the citizenry. It is not about legality, or even popularity, of police actions. The causes of police legitimacy are complex, even when legitimacy itself can be measured (usually by public opinion surveys). One cause has been shown by the Canberra restorative justice experiments to be procedural justice. It is not the only cause of legitimacy, which is also driven by perceived effectiveness of police in protecting the public, and in the perceived integrity of the police.

## 46. Procedural Justice

Procedural justice is a theory that people are more willing to obey the law if they think they have been treated by legal agents with *fair procedures*, in their own personal interactions with those agents. Fair procedures include explanations of what is happening and why, the provision of an opportunity to have a suspect explain his or her side of a case, an opportunity to appeal for correction of errors in factual assumptions, and general treatment as if one is a citizen welcome to live in a community rather than an outsider whose presence is unwelcome. The theory notably holds fair *procedures* to be more important than fair *outcomes*. For example, there is evidence that *how* a suspected was treated when arrested matters more than *whether* the suspect was arrested at all.

## 47. Restorative Justice Conference

A restorative justice conference is a consenting meeting in private, over 2-3 hours, of one or more crime victims, the suspects or convicted offenders against those same victims, and the family or friends of both victims and offenders.

a. The conference is arranged in advance, often by a police officer, who
b. meets privately with the principals to explain what will happen at the conference, when the facilitator of the conference makes sure everyone can say as much as they want to say (without undue repetition) about
    1. What happened?
    2. Who was affected by the crime and how?
    3. What should the offender do to try to repair the harm?

Note: Police-led restorative justice conferences have been found effective, especially with violent crime.

## 48. Offender-Desistance Policing

Policing for the goal of encouraging offenders to desist is meant to reduce the frequency, seriousness or persistence of the offender in any crime at all. Offender-desistance policing (ODP) takes a wide range of forms, including "Problem-Oriented Policing" (POP), covert surveillance to catch serious offenders in the act, restorative justice conferences, and deferred prosecution with a "Sword of Damocles" of an approved charge file hanging over the head of a suspect who agrees to an instant "Turning Point" plan of pre-prosecution police probation or rehabilitation.

## 49. Solvability Factors

A solvability factor is an element of evidence that is present in an investigative file at the close of the preliminary investigation that predicts that the case is more likely to yield a suspect and a charge than the average investigation for that offence. Solvability factors can only be identified by systematic research, using thousands of cases. Police have successfully separated solvable from unsolvable cases using these methods for caseloads of burglary, non-domestic violence, and metal theft from railroads.

## 50. Displacement

Displacement is a rival hypothesis to deterrence in many crime prevention experiments. The hypothesis claims that crime may have gone down in one group of measured targets, relative to control group targets, but it did not; the same crimes were actually committed elsewhere. The scientific problem

with the displacement hypothesis is that it is often un-testable, since all hypotheses should be testable, at least in theory. But there is no theory of testing the claim that crime went "elsewhere," unless the where can be specified. Repeated tests of displacement in the immediate vicinity of a crime prevention success have shown more reduction in crime than increases, or what is called "diffusion of benefit." One test of offenders arrested before a crackdown occurred to see if they were arrested elsewhere more often showed that they were not, relative to a similar test with offenders in control areas for the crackdown evaluation (Mazeika, 2013).

PROPOSED DEVELOPMENT PLAN FOR EVIDENCE-BASED POLICING:

**Abecedarian** of Evidence-Based Policing **(A.E.B.P.)** By examination.

**Master of Evidence-Based Policing (M.E.B.P.)** By examination or accredited University Master's degree.

**Fellow of the Society of Evidence-Based Policing (F.S.E.B.P).** By election based on contributions to policing knowledge.