

Buyer Beware: A Consumer's Guide to Reading and Understanding Correctional Research

A good consumer asks certain questions before buying something, and we should treat the products of research no differently than we would treat a VCR, a television or an investors' group. After all, we may well end up investing time and money in projects on the basis of what we read.

This article briefly reviews some important, but often neglected, questions that we should ask of correctional research in general, with some specific examples related to recidivism.

Selecting Study Samples: How Does who We Pick Affect what We Find?

Consider a hypothetical study with two groups of sex offenders in which each group completed one of two different treatment programs. Offenders were admitted to the programs on the basis of a risk assessment; moderate-risk offenders were placed in program A and high-risk offenders in program B.

If we observe lower recidivism among Group A subjects, is it because the program was more effective? Or is it because Group A subjects were a lower risk to begin with?

When we observe a significantly lower recidivism rate for one group of offenders, the question is: how much of this can be explained by pre-existing differences between the groups?

To offset this problem, we try to assemble sample groups that reflect the population being studied as much as possible, particularly with regard to important characteristics such as criminal history. An important part of this process has to do with random assignment. Where possible, we try to assign people to groups randomly to offset systematic bias. Sometimes we can't - for example, when we want to look at male-female differences. But even when we cannot control membership in a particular group (such as male/female), we still try to select study subjects in a random manner. While random selection offsets some sources of bias, the analysis of results remains open to interpretation.

Even when using random samples, we can still run into thorns when we try to make decisions such as should the sample be arranged using proportions that reflect the general population or the prison population? Consider the case of aboriginal people, for example, who represent 2% of the general population but 17% of the prison population. Should we arrange our study sample so that aboriginal offenders comprise 2% or 17% of the sample? Our decision depends on what we want to say about which population, so the breakdown we choose must be defined by our research questions. Naturally, the choice we make could affect our results.

This scenario raises the question how far can we generalize our results? If we find, for example, that aboriginal offenders with serious substance abuse problems are generally younger than non-aboriginal offenders with similar problems, does this mean that all aboriginal people with serious substance abuse problems are probably younger than non-aboriginal people with similar problems? Or can we only apply these results to the offender population?

To evaluate research we must ask some questions. How was the sampling done? Were the subjects

randomly selected? To what population do the study results apply?

Of course, in the real world we rarely can have perfect sampling.

Let's look at some more questions related to this problem.

Who's In and Who's Out: Problems of Participation and Non-participation

The types of information we gather in our studies, and how we gather that information, can skew our results. For example, if we are examining whether resistance to authority predicts recidivism and our sample is made up only of offenders who volunteered to participate, we are probably shooting ourselves in the proverbial foot by allowing the ones who truly resist authority to express that resistance by refusing to participate!

Similarly, we may have trouble getting at the information we consider important for a particular study. Again, let's pretend we are doing a study on recidivism. Some variables, such as the level of community and family support, though important to our study, cannot be tracked after offenders' sentences expire. We are thus limited to police reports about future criminal activity and to offenders' explanations for their recidivism. The challenge is to come up with creative study designs that accurately evaluate such postrelease risk factors.

The problem of postrelease follow-up can also make gathering statistics difficult. For example, the fact that some offenders spend only weeks on parole while others may spend many months can be a problem. How do we define success? Suppose we define it as no readmission within one year of release. Since our offenders are not released all at once, they may have been out for anywhere from a few weeks to a few years by the time we gather the data. They have not, therefore, all had the same number of chances to reoffend. If there is any connection between time outside and our treatment program (or other variables of interest), there may be a profound error in our results.

Further, some of our subjects may reoffend on the day we analyze the data, others may reoffend the day after we gather the data, while others may take 6, 60 or 600 weeks before reoffending. Others will never come back. The only ones we really know about are the ones who reoffend and were caught by the time we gathered the data, because they are back within range of our data-gathering machinery.

Birds of a Feather: The Restricted Range Problem

If we compare two groups of offenders and the offenders share a number of common characteristics, then it will be harder to find differences between them. We may therefore be unable to get valuable results simply because the two groups are too similar.

Let's consider another hypothetical example. We wish to examine the relationship between two variables - one variable is the severity of crimes committed by offenders who reoffend and the other variable is antisocial attitudes. We have data on several hundred offenders who reoffended with serious crimes, but none on offenders with relatively minor offences.

We plot the data, calculate some kind of statistic and decide that no relationship exists between the severity of crimes committed by offenders who reoffend and antisocial attitudes. Years later, somebody does the same study using **all** recidivists and discovers there **is** a relationship. Why?

The important point here is that if we limit ourselves to a group with a common background, there will be a lot of shared characteristics (such as antisocial attitudes). The offenders in our study who reoffended with serious crimes may only differ slightly among themselves, so this difference may be harder to detect. This may be familiar to some readers as the "restricted range problem" from Stats 100, which is illustrated in Figure 1.

Figure 1

Figure 1
Correlation of X and Y - Full Range of Values

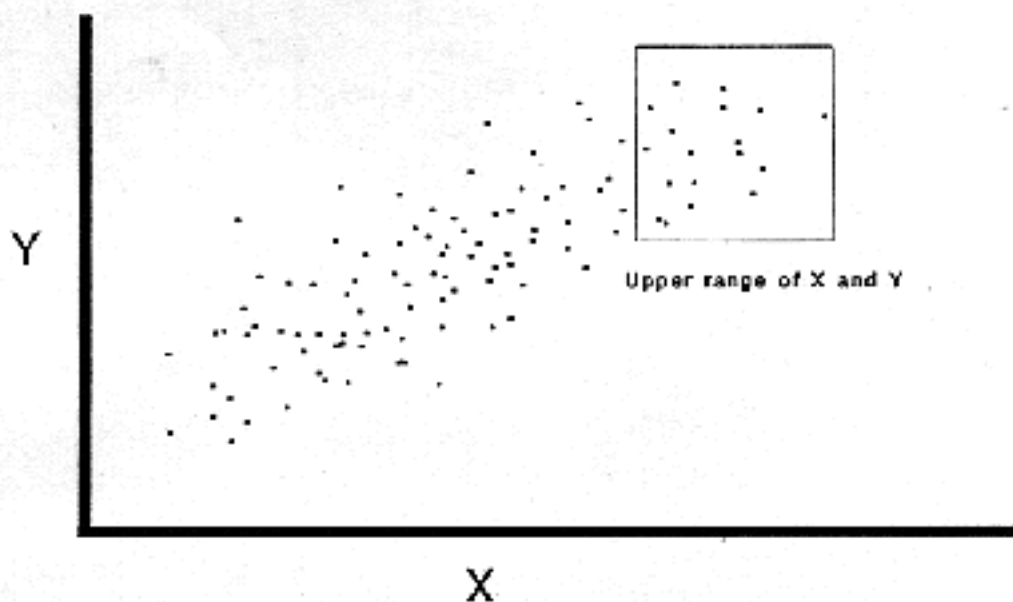


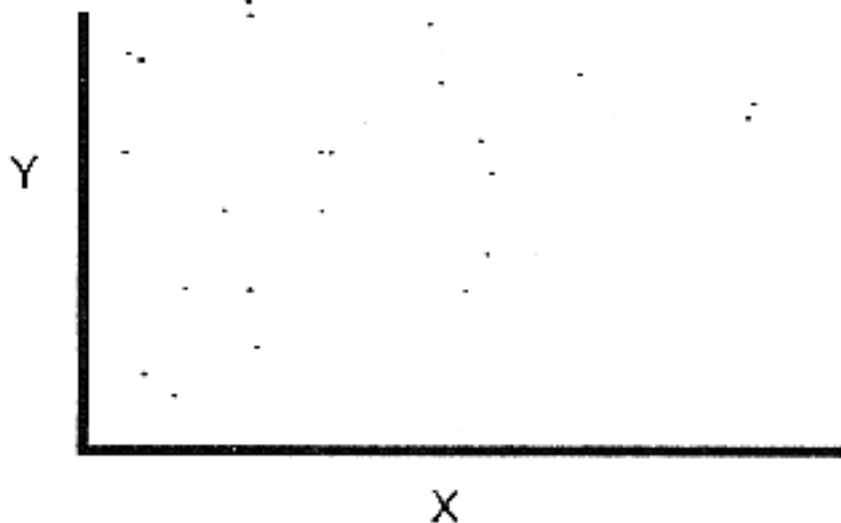
Figure 1 shows a plot of points reflecting the strong, "real-life" relationship between the severity of the new offence and some variable Y. Pretend for now that Y is a score on a measure of antisocial attitudes. If we were able to sample **everybody** and plot the scores for the severity of the new offence versus their antisocial attitudes, we would see that as antisocial attitude scores increased, so did the severity of the new offence. This would suggest that recidivists committing more serious offences are more likely to have antisocial attitudes (See Figure 1).

The points in the square at the top of Figure 1 are those of recidivists committing serious new offences. Figure 2 is a close-up of the data on these offenders. This is what the researcher who was restricted to only serious recidivists would have seen. Based on what is contained in the square, there is no strong evidence of a relationship between the severity of the new offence and Y, despite the strong relationship

shown in Figure 1 There was a lack of evidence because the sample was too homogeneous (See Figure 2).

Figure 2

Figure 2
Correlation of X and Y - Restricted Range of Values



Statistical Power: Hunting Rabbits with a Tank

The statistics we use in research are just ways of deciding whether or not a difference is big enough to warrant discussion.

Continuing with the above example, if the severity of the new offence is related to antisocial attitudes, then by sampling a range of individuals (to avoid the restricted range problem) we should get a plot that looks like Figure 1, with X representing scores for severity of the offence and Y representing antisocial attitude scores. The shape of that plot tells us the relationship: a high score for offence severity equals a high score for antisocial attitudes.

We can then calculate a statistic called "R" which tells us how strong that relationship is. If $R=0$, there is no relationship, and we would only see a cloud of points similar to Figure 2. However, if $R=1$, then the points in Figure 1 would fall in a straight line, and we could perfectly predict the severity of the new offence from the antisocial attitude score (and vice versa). However, the world is not perfect, so for Figure 1, R probably equals about .85, which is considered to be a strong correlation.

One thing to keep in mind, though, is sample size. Sometimes there is no meaningful relationship between two variables, but statistics tell us there is. The fuzzy cloud in Figure 2 appears to have about 50

points in it. That represents 50 subjects, 50 offenders on whom we have data. That is a fairly large sample, and it is possible that our statistics would tell us that even though $R=.25$ (generally viewed as a small relationship), it is worth looking into further.

This is the point at which statisticians and theoreticians differ. Theoreticians like statistics that say "it's worth discussing." Statisticians, on the other hand, protest "but, it's so small!" This argument comes from one simple fact: the statistics we commonly use become more and more sensitive to smaller and smaller differences as our sample size increases.

Without going into the gory technical details, this is the implication for correctional research: because we have access to huge samples, we can apply our statistics and find tiny differences that a computer program will tell us are worth discussing. Used uncritically, these computer programs can be hazardous, permitting one huge leap for theoretician-kind on the basis of one very small step (difference) in a large sample.

Using a large data base to hunt down trivial differences is like hunting rabbits with a tank - it's overkill. We must decide what size of difference is meaningful and then proceed.

To the reader, the relationship between the numbers and the size of a difference is not always clear. However, we should expect researchers to be able to say, in terms that anyone can understand, whether a difference is big, medium or small. The reporting of this relationship is often ignored in write-ups, an omission that can distort the **meaning** of the results⁽²⁾.

So we have another question for which we should expect an answer. What do these results mean? The answer will not always be given, but when it is not, we should ask "why not?" Armed with these questions, we may now look at some of the instruments commonly used in corrections.

Instrumentation: Blessings and Curses

With the amount of information routinely collected on all sorts of offenders, it is easy to rely on what is at hand and assume that it measures what it claims to measure. We forget that data can be thrown off by any number of factors, including the instrument not doing what it is supposed to be doing. Therefore, the validity of an instrument must be questioned. Do different psychopathy measures classify the same people as psychopaths? If two different measures of the same thing give us different answers, then which (if either) is correct? Or, is there something wrong with what is being measured (that is, the construct)? Does "psychopathy" even exist, and if so, can it be measured?

We may also ask if the instrument is reliable. Would we get the same score if we gave the test to the same person a little while later (assuming nothing has been done to affect how he or she would respond)? And if it is reliable, to what extent is this because the same file information has been used twice?

Of course, some tests are so well documented that we assume they are valid and reliable. But, if a study uses a new test or instrument without answering these questions for us (for example, by citing reliability and validity studies), then we can rightly be suspicious of the results.

"Statistics Are Always Right"

The assumption that statistics are always right is probably the reason why many of the questions people have about research remain unasked.

One of my favourite quotes (from the old sage Anonymous) reads: "Recent figures indicate that 43% of all statistics are utterly worthless." A statistic is just a number which, if calculated correctly, tells us something about a group of numbers.

But there is error in every single measurement. Sometimes it's small, sometimes it isn't. It would be a crowning achievement to predict recidivism with 90% accuracy, but there is always that 10% we don't anticipate. Unfortunately, we hear more about the exceptions (because of some heinous crime committed by one) than the rule. While the importance of an exception is probably greater in the field of corrections than in many other social sciences, there is little we can do about it.

For now, we must be content with imperfect instruments that make our ability to predict better than just guessing. We can learn from experience and try not to make the same mistakes again. While we cannot deny the consequences of inaccuracy, we may be less disheartened if we remember that - based in part on our assessments and instruments - hundreds of people did **not** become victims because parole was denied to offenders who would have committed a crime. Alas, we have no statistics on crimes that never were.

Summary

We have suggested that the consumer of research papers ask certain questions about the research methods and the results of studies. Not all research will answer the questions equally well, and of course, some answers will be referrals to other materials.

However, in social science research in general, and correctional research in particular, much ado can easily be made of nothing. Mountains may spring up from statistical mole-hills, and therefore, both the producers and the consumers of research should be aware of the need for clear communication. Leaving questions unanswered or unasked is not the way to achieve valid research.

(1) *Travis Gee, Research and Statistics Branch, Correctional Service of Canada, 340 Laurier Avenue West, Ottawa, Ontario K1A 0P9.*

(2) *See R.P. Carver, "The Case Against Statistical Significance Testing," Harvard Educational Review, 48(1978): 378-399. See also J. Cohen, Statistical Power Analysis for the Behaviour Sciences (revised Edition), (New York: Academic Press, 1977). And See A.W. MacRae, "Measurement Scales and Statistics: What Can Significance Tests Tell Us About the World?," British Journal of Psychology, 79 (1988): 161-171.*