Limitations of the WSIPP Study: BIP Completers vs Non-Completers
Eric Mankowski, 12/23/12

Below, I offer a response to [the WSIPP study] by first summarizing the limitations of experimental research designs, then by describing how statistical controls are intended to address these limitations, then describing the limitations in turn of such statistical controls, and finally confirming your assessment of the logic of statistical control and its implications for our conclusions regarding treatment outcome research. In the end, my view and the view of most of my colleagues is that both methods have limitations, the limitations are different from each other, and consequently, both kinds of research designs complement each other in useful ways and together give us the most accurate picture.

1. Experimental designs as implemented typically fail to account for differences among programs in the study. The context of the programs may influence program outcomes, for example, differences in program curriculum or differences in the larger coordinated community response that surrounds the BIP program. Evaluations of BIP treatment in counties where there is poor CCR should be weighed cautiously.

2. Experimental designs must be implemented properly in order to conclude that the treatment worked or not. For example, if/when a judge overrides random assignment of offenders to treatment vs. jail (e.g., because of concerns about whether jail or treatment is appropriate for a given offender), the results of the study are not valid with regard to the BIP treatment program.

3. Most critically, cases assigned to treatment often do not get the treatment in an "intention to treat" experimental design. This occurs in a study where cases are assigned to BIP at conviction or sentencing (the intent is that they will get treatment), but the case never enrolls in BIP or enrolls but does not complete treatment. Such cases are still included in the analysis of the "treatment" group because it was intended that they would get treatment. However, this actually produces an evaluation of the effectiveness of the entire criminal justice system as a whole (including treatment), not an evaluation specifically of BIP treatment.

4. Many experimental studies of BIP have failed to measure anything but criminal recidivism as an outcome. This fails to capture whether the programs affected non-physical abuse, and because there can be bias in recidivism statistics (e.g., bias in arrest, prosecution, sentencing, etc. associated with race), we should not rely on this as a sole outcome on which to judge BIP effectiveness.

Several of these limitations of experimental designs can be partially addressed in studies where program completers are compared to cases that begin treatment but fail to complete it. This design addresses limitations #1, #2 and #3 above.

However, such statistical control (aka 'quasi' experimental) designs have their own limitations as well. Specifically, the problem is that cases that complete and fail to complete programs have been shown to be different from each other (e.g., demographic characteristics are different). Consequently, one cannot

infer with certainty that the treatment was responsible for differences in outcomes between the completers and non-completers. Instead, it may be that the different characteristics of the cases are partly or completely responsible for both differences in program completion and in outcomes (e.g., recidivism, partner reports of abuse). For example, men who are poor may be both more likely to drop out of a program because they cannot afford it and more likely to be rearrested or convicted for partner violence than wealthy men due to bias in the criminal justice system, which would produce the result that the treatment worked when in fact the lower recidivism of completers was due to their wealth.

One way to address this limitation is by using statistical controls for any known differences between the cases that complete and fail to complete BIP. This is what Gondolf and others have done (as well as many researchers in areas of study throughout the social sciences, where experimental assignment to treatments or other groups is unethical and/or impractical). And, it is these studies that Marna Miller's review selection criteria have excluded because she believes that statistical control of any differences between the groups is not a valid way to address this problem. For example, we can only statistically control for differences between the groups that we know about (that we have measured). If there is a real difference but we have not measured it, we cannot know about it and we cannot control for it.

Many researchers, journal editors, scientific expert panels, federal scientific granting agencies, etc. disagree with her dim view of statistical controls; however, many researchers concur with her – it is an area of ongoing debate. As our statistical modeling techniques have become more sophisticated, however, our ability to compensate for the lack of experimental control improves and becomes more effective.

As you point out, the entire dismissal of the results of a statistical control designs may be throwing out the baby with the bathwater. Certainly some of the difference in outcome between the groups may be due to statistically uncontrolled differences; in fact, we can assess exactly how much of the difference in outcome is due to known (measured) pre-treatment differences between the groups. In the case of Gondolf's studies, the 50% decrease in recidivism he found in those who completed BIP held true even after statistical control for many characteristics known to be associated with violence more generally (e.g., personality disorder, alcoholism, demographic differences, etc). In other studies, the apparent treatment effect may decrease after statistically controlling for such differences, but still remain reliable.

Thus, as you correctly point out, the treatment effect should not be assumed to be zero simply because of known or unknown differences between the treatment completers and non-completers. The logic you describe below in your email is exactly correct by my understanding, and has been done by Gondolf (and other researchers who have identified characteristics associated with both treatment drop out and recidivism, and controlled statistically for the ones they have measured)