Research and Evaluation DATABITS
2012-10 October 5, 2012

What's the Evidence for Evidence-Based Practice?

by Jeffrey A. Butts

Youth justice practitioners need to understand the basics of evaluation research, including the statistical methods used to generate evidence of program effectiveness. A study that reports statistically significant results is not necessarily evidence of effectiveness, and being evidence-based does not mean a program is guaranteed to work. In today's youth justice system, understanding these basic principles of evaluation research is part of every practitioner's job.

The Limits of Evaluation Research

An evidence-based approach to youth justice is better than an approach based purely on faith or anecdote, but practitioners need to appreciate the limitations of evaluation research. First, the findings of existing evaluations are not a sufficient basis for making all of the choices involved in building and operating a modern youth justice system. Lawmakers who insist on irrefutable evidence for every policy or program will end up distorting the necessary balance of comprehensiveness and effectiveness.

Second, there is no such thing as a perfect study. Program evaluations are essentially studies of human behavior as well as the strategies for changing behavior. Human behavior, however, is enormously complex and not completely measurable. In a technical sense, researchers never prove that programs work. Their goal is to reduce uncertainty.

Third, no matter how strong evaluation results may be, some uncertainty always remains. To say that a program is evidence-based means that researchers are pretty sure that having the program is better than not having the program, or that the odds of the program achieving its outcomes are pretty good. Positive evaluation findings do not guarantee that a program will work every time, for every person, and in every situation. Practitioner judgment is still required.

Interventions that can be assessed by experimental methods attract the bulk of talent and resources, while promising activities that aren't built on a linear relationship between cause and effect and cannot be entirely contained and controlled in a laboratory-like setting will be disparaged and downgraded.

Katya Fels Smyth and Lisbeth B. Schorr, 2009

Research evidence comes in different forms. In fact, some evidence originates from qualitative studies, where data are maintained as stories or narratives and researchers conduct their investigations using interviews and direct observations. Qualitative studies have a role to play in the evaluation of youth justice programs, but they rarely achieve the same policy impact as do quantitative or statistical studies.

Statistical Significance

Even in quantitative studies, standards of evidence vary. Studies of basic, empirical questions (e.g., is drug court associated with less recividism?) may rely on statistical significance as their principal metric. Stated in terms of probability, or ***p*** values, a researcher might report that the use of a particular intervention is associated with lower recidivism, and the connection between the two is so strong that there is less than a one percent probability ($p < .01$) that the association would occur by chance alone. *(Note that this means such an association could be completely coincidental in one of every 100 tests.)*

JOHN JAY COLLEGE OF CRIMINAL JUSTICE

RESEARCH AND EVALUATION CENTER
Part of the Research Consortium of John Jay College
555 West 57th Street, Suite 605
New York, NY 10019
www.johnjayresearch.org/rec

CUNY

Researchers use probability values to describe the statistical significance of study results. The value of ***p*** indicates how unusual a particular finding is based upon the distribution of similar findings. The particular threshold used (i.e., 1%, 5%, or 10%) is chosen in advance, according to theory or the experience of other studies and similar programs. Research findings should never be described in levels, with one finding being termed more or less significant than another. The results of an analysis are either significant or not significant.

Significance is often misinterpreted as importance. The statistical significance of a particular finding is determined by the size of a difference in combination with the number of observations (or, ***N***) used to detect that difference. Even a large difference (e.g., 40% versus 60% recidivism) may fail to reach the level of statistical significance if the study relied on a small sample. Some studies, for example, may collect data on just 20 or 25 youth.

When researchers use very large samples involving thousands of cases, on the other hand, even a small difference (e.g., 50% versus 52% recidivism) may be statistically significant. Of course, few public officials would invest much in a program that reduced recidivism by just two percentage points.

Effect Size

In evaluations, "effect size" is often a better metric than statistical significance for assessing program impact. Effect size is calcuated as a range from +1.0 to –1.0, where more negative numbers indicate stronger reductions in recidivism. The most successful evidence-based programs usually have effect sizes between –.10 and –.30. Effect size measures a change in outcome, controlling for the variability of that outcome.

A program that reduces recidivism by 50 percent will have a larger effect size than a program that lowers recidivism by just 10 percent, but such comparisons are sensitive to the average level of recidivism. If expected recidivism is very low, such as when only five percent of youth in a prevention program are likely be re-arrested, a change of three points (from 5% to 2%) might be a large and valuable effect.

Effect size gauges the scale of measured change against the natural or expected variation in the same outcome. For example, if recidivism for a particular type of youthful offender is known to fluctuate widely, perhaps between 20 and 60 percent, a change of three percentage points would seem trivial and not worth the resources required to fund the program. On the other hand, if recidivism for a particular population rarely varies outside a five point range, say from 45 to 50 percent, a program able to produce a consistent decline of three percentage points would have a very strong effect size.

SUBSTANTIVE IMPORTANCE

Depends on scale. How much difference do we see between youth outcomes or between program outcomes? Is the degree of difference important or meaningful?

STATISTICAL SIGNIFICANCE

Depends on variability. How likely or unlikely is the difference we see? Could it be due to chance alone?

EFFECT SIZE

Combines importance and significance. Given how unlikely this difference would be to occur by chance alone, can we attribute the difference to the program, and is the difference valuable?

Of course, some low-cost programs with modest effect sizes may still merit the label "evidence-based" because they generate a positive return on investment. A program that costs very little to implement and operate (e.g., teen courts) might be a worthwhile investment even if it has a relatively small effect size. Cost is an increasingly important component of evaluation research in youth justice.

Conclusion

Evaluation research should—and always will—play a role in the youth justice system. That role, however, should not be absolute or controlling. Evidence is not simply discovered; it is purchased. Strong evidence requires the sustained efforts of researchers working in collaboration with practitioners, and these efforts require the investment of resources—sometimes substantial resources. There will never be enough funding to evaluate every single component of the justice system. Thus, not all programs can be evidence-based.

Research evidence does not emerge from a pristine and impartial search for the most effective practices. The evidence we have today is the fruit of our previous research investments—investments made by funders and policymakers with beliefs, values, preferences, and even self-interest. As long as this is the case, practitioners must exercise caution in how they interpret and apply the evidence produced by evaluation research. Evidence should inform, but never simply dictate the shape of policy and practice.

Funding for the preparation of this document was provided by the Pinkerton Fellowship Initiative and the Rudin Graduate Research Fellowships at John Jay College of Criminal Justice. Any opinions are the author's alone and do not necessarily represent the views of John Jay College or its funders.