Script for Understanding Juvenile Justice Evaluation Reports

[Title Slide 1 of 40]

Welcome to Understanding Juvenile Justice Evaluation Reports

This is one of a series of briefings prepared by the Justice Research and Statistics Association’s National Juvenile Justice Evaluation Center project. The purpose of this briefing series is to provide juvenile justice program managers with information that will help them to evaluate their programs. Each briefing addresses a topic that is of particular interest to juvenile justice program managers who are trying to determine the effectiveness of the programs they operate.

Note that this PowerPoint presentation is also available in PDF format on the NJJEC website at www.jrsa.org/njjec/

You’ll also find more information about the project overall, including resources and tools, on our website.

This project is funded through a grant from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, United States Department of Justice. Neither the United States Department of Justice nor any of its components operate, control, are responsible for, or necessarily endorse, this project presentation (including, without limitation, its content, technical infrastructure, and policies, and any services or tools provided).

[“Briefing Overview” Slide 2 of 40]

The purpose of this briefing is to provide information to juvenile justice program managers to help them build understanding of juvenile justice evaluation reports by discussing how to distinguish between good and poor quality evaluation reports. The brief is designed to introduce and explain the key concepts in outcome evaluation research. Our goal is to help practitioners

- Understand key evaluation terms and designs, and
- Recognize how to identify a well-written evaluation report.

This presentation does not explain how to identify evidence-based programs or “what works.” It is not intended to assist the reader with making overall judgments or determinations about specific programs or program types. This guide is focused on providing the reader with the basic information needed to identify high quality evaluation reports.
Evaluation is a systematic and objective process for determining the success or impact of a policy or program. Evaluation is usually considered a type of research and consequently uses many of the same approaches (such as research designs). Research typically asks questions about why or how something operates.

Evaluation addresses questions about whether, and to what extent, the program is achieving its goals and objectives and the impact of the intervention.

A good evaluation has several distinguishing characteristics relating to focus, methodology, and function. Evaluation:

- Assesses the effectiveness of an ongoing program in achieving its objectives, and
- Through its design, helps to distinguish the effects of a program or policy from those of other factors that may cause the outcomes.

With this information, practitioners can then implement program improvements through modifications to the current program model or operations.

Evaluations are generally categorized as being either process or outcome. Process evaluations focus on the implementation of the program or project. They may precede an outcome evaluation and can also be used along with an outcome evaluation to ensure that the program model or elements are being implemented with fidelity (or consistent with the model). Outcome evaluations (sometimes called impact evaluations) focus on the effectiveness of a program or project. This guide concerns itself with outcome evaluations and how to tell whether they are of high quality.

Evaluations vary widely in quality.

While most people understand this, fewer are comfortable with determining which are of high quality. Reading an evaluation report, particularly one that is full of statistics and technical terminology on research methods, can be overwhelming.

Nevertheless, there are some key issues that anyone can use to help distinguish between good and poor evaluations. Two primary issues are:

- The quality of the evaluation design; and
- How well is the evaluation carried out?

These issues are discussed in detail in this presentation.
Issue 1: Role of Evaluation Design

The basic question an evaluation asks is, did the program or policy have the intended effect? There are many ways to go about trying to answer this question. The research design is the most important piece of information to use in assessing whether the evaluation can answer this question.

To assess the effect of a program through an outcome evaluation, an evaluator must establish that a causal relationship exists between the program and any outcome shown. In other words, did the program (cause) lead to the outcome (effect)? In order to demonstrate a causal relationship between the program and the outcome, the evaluator must show:

1) that the cause preceded the effect,
2) the cause and effect are related to each other, and
3) that the effect was not caused by an outside factor.

Accounting For External Factors

Often, statements about the relationship between a program and outcomes are not warranted because the design of the evaluation does not take into account external factors that may be responsible for the outcomes. For example, a tutoring program seeks to improve academic performance. But the student’s grades could also be influenced by external factors, such as a successful anti-bullying campaign in the school and if the participants also joined a homework club.
The way in which an evaluator can demonstrate that a causal relationship exists is through the evaluation design, that is, the structure of the evaluation. Evaluators may choose from many different designs and some of these designs are better than others for demonstrating causality. Factors like resources to do the evaluation (time and money), and concerns about the appropriateness of an approach given what the program is trying to accomplish usually play a big role in what approach is selected.

Evaluation designs are commonly divided into three major categories based on their characteristics:

- Experimental
- Quasi-experiment, and
- Non-experimental

Definitions and examples of these different approaches are covered in the next slides.

Experimental designs are distinguished by the random assignment of subjects into treatment group (those who received the program or policy) and control group (those who did not receive the program or policy).

To do a random assignment, think about flipping a coin to decide who goes into each group. The term “randomized control trial” (or RCT) is often used to refer to an experimental design.

Experimental designs are often referred to as the “gold standard,” because with random assignment to either group it is assumed that the only difference between the two groups is that one group had the treatment and the other did not. Thus, experimental designs provide the greatest level of confidence in the results.

Once individuals are randomly assigned into either the treatment or the control group, the intervention is delivered to the treatment group, and outcomes for each group are compared. Although designs will vary from experiment to experiment, a common variation on this design is that the same data are collected for each group before the intervention (a pre-test) and again after the intervention (post-test).

The evaluator examines whether there are differences in the change from the pre-test to the post-test for each group.
[“Strength of RCT” Slide 11 of 40]

With an RCT design, there is little question about whether the program or policy caused the change because this approach controls for external factors. Recall that external factors are factors other than the program or policy that may have caused the outcomes (such as the school anti-bullying campaign, which could account for part or all of a program participant’s improved academic performance).

An experimental design addresses these types of concerns, since both the tutoring program participants and non-participants would be exposed to the external factors (the school anti-bullying campaign.)

[“Understanding the Report: Tip #1” Slide 12 of 40]

Of course, to have this confidence requires that the design is implemented well. Evaluators must ensure that:

- Assignment to the treatment and control groups is actually random. For example, deciding to put all high-risk offenders in the treatment group, but randomly putting others in the treatment and control groups is not random assignment).
- Secondly, the evaluator needs is to ensure that the control group does not receive the intervention, and finally,
- the evaluation must examine outcomes for everyone, even those in the treatment group who don’t finish the intervention.

A good evaluation report on an experimental design should discuss all of these issues.

[“Quasi-Experimental Designs” Slide 13 of 40]

The major difference between experimental designs and quasi-experimental designs is that quasi-experimental designs do not employ a randomly assigned control group. Instead, quasi-experimental designs may use a comparison group which consists of a group of individuals/cases that are considered similar to those who received the treatment. The evaluator may attempt to ensure comparability between the two groups, to the extent possible, by matching the individuals in the groups on factors that are considered relevant, such as age, gender, or prior history with the criminal justice system.

For example, if gender and prior out of home placement are considered relevant factors, one way to match is to ensure that the treatment and comparison groups have similar proportions of males and subjects without previous out of home placements. If one group had more individuals without placements, then its members may be less likely to engage in delinquency than the members of the other group. In a quasi-experimental design, the more confident we can be that the two groups are similar in all key characteristics other than program participation; the more confident we can be that the program caused any observed differences in the outcomes.
As with experiments, designs of quasi-experiments will vary, but perhaps the most common quasi-experimental design used in evaluation is a pre-post design with a comparison group. Here again the same data are collected for each group before the intervention (pre-test) and again after the intervention (post-test). The evaluator examines whether there are differences in the change from the pre-test to the post-test for each group.

Another commonly used quasi-experimental design in juvenile justice is the time series analysis. With this method an evaluator typically studies something like the effect of a new policy or legislation in a jurisdiction. The evaluator conducts a series of observations prior to the new policy, then conducts another series of observations following the implementation of a policy. Let’s say that the school board implemented a new policy that was expected to decrease truancy. The evaluator may look at the number of unexcused absences from school for several quarters before the school policy changed, and then for several quarters following the policy to see whether there was a change in the number of absences.

This figure provides a visual example of how an evaluator may present data in a time series analysis. The policy was expected to decrease unexcused school absences, and it appears to have had that effect; absences decreased over the time examined. A good time series design would also examine whether other factors occurring around the time of the new policy, say the hiring of additional truant officers, appeared to contribute to the number of absences.
[“Understanding the Report: Tip #2” Slide 16 of 40]

Given the program or policy under study, it may be impossible or undesirable to use an RCT design. For example, evaluators cannot randomly assign attributes of youth such as age, gender and race, so questions related to these attributes cannot be studied using RCTs. In some cases, it may also be difficult to identify an appropriate comparison or control group.

In criminal justice this is particularly the case with programs/policies that affect entire communities or jurisdictions. Keep this in mind when considering the strength of the design.

**Always look for an explanation of why the evaluator chose a particular research design.**

[“Non-Experimental Designs” Slide 17 of 40]

Non-Experimental Designs provide the least amount of guidance on program or policy impact, is considered the weakest of all the designs, and raises the greatest number of questions about whether the program or policy caused the outcomes.

In this approach, the evaluator collects data on the group receiving the treatment or affected by the program or policy at one point only—after the program or policy has been implemented. The evaluator is not able to assess whether change occurred.

An example of a non-experimental design would be a survey of residents to assess satisfaction with police practices after the introduction of a new community policing effort. If the questions address only current satisfaction, there is no way to know whether satisfaction increased after community policing began, let alone whether community policing improved satisfaction.

[“Issue 2: Evaluation Execution” Slide 18 of 40]

For any design, the evaluator must ensure that the evaluation is conducted in such a way that the basic design is not undermined and that other elements of the evaluation, from the data collection to the analysis and the writing up of the results, are carried out sufficiently well so that one can trust the results.

Here we discuss the major issues (in addition to the research design) that one needs to identify when assessing how well the evaluation was carried out. It is difficult to say in the limited discussion provided here at what point the evaluation report should be considered seriously flawed if one or more of these issues is not addressed. However, the more of these items appropriately addressed in the evaluation report, the more confidence one should have in the results.
It is important that the number of subjects (or individuals, or cases) selected for the study IS large enough to support the analyses required to answer the evaluation questions and to raise little doubt that the results adequately represent the population. (The population is the entire group from which the sample was drawn.) This issue of sample size sufficiency is important when concerns such as resource constraints do not permit data collection from all subjects (the population) who may be affected by the program or policy.

This figure provides a simple guide for selecting a sufficient sample size. As you can see, as the population increases, the number of subjects required becomes relatively smaller; and with smaller populations, you need a greater proportion of subjects.

<table>
<thead>
<tr>
<th>Number in Population</th>
<th>Minimum Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>50</td>
<td>44</td>
</tr>
<tr>
<td>70</td>
<td>59</td>
</tr>
<tr>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>150</td>
<td>108</td>
</tr>
<tr>
<td>200</td>
<td>127</td>
</tr>
<tr>
<td>500</td>
<td>217</td>
</tr>
<tr>
<td>1000</td>
<td>278</td>
</tr>
<tr>
<td>1500</td>
<td>306</td>
</tr>
<tr>
<td>2000</td>
<td>322</td>
</tr>
</tbody>
</table>


A related issue to sample size is attrition. Attrition refers to subjects dropping out of the evaluation before data collection has concluded. Attrition is an important issue because individuals usually don’t drop out of evaluations randomly; for example, youth with lower grades are more likely to skip school and perhaps to drop out of the tutoring program.

One result of this non-random attrition is that treatment and comparison groups that started out looking similar may not end up that way.
Good evaluation studies track who drops out, when they drop out, and (ideally) why they drop out. The number of subjects who drop out as well as descriptive information about the dropouts should be included in the evaluation report. Outcome information should be included on these subjects (such as recidivism rate for dropouts) and should be a consideration in assessing the sufficiency of the sample size.

Always look for a discussion in the paper regarding the sample size, the degree of subject attrition, and if outcome analysis includes subjects who drop-out of the study.

Besides attrition, sometimes problems with the implementation of an evaluation, such as recruitment of subjects for the study or assignment to the treatment and control groups not occurring as planned, may result in the treatment group being substantially different on important characteristics from the comparison group or control group (for instance, age of youth, and school attendance history).

The main question in an evaluation of a program or policy delivered to a group of subjects is usually something like:

1) Did the treatment group (those who received the program/ or were impacted by the policy) have better outcomes than those who did not receive the program or were not affected by the policy? or
2) Did the group that received the program or policy improve/change since the intervention was delivered?

However, when a program or policy is implemented, it is common to expect that certain subgroups of subjects will have better (or worse) outcomes than others. For example, one might reasonably expect that individuals completing the program would have better outcomes than individuals not completing the program.

The evaluation report should specify the important characteristics of similarity between the treatment and comparison groups and report on whether the two groups were actually similar on these characteristics.

The evaluation report should indicate whether or not the analyses was conducted which broke out outcomes by subgroups and the rationale for doing so.
[“Outcome Measures” Slide 26 of 40]

The measures (or indicators) selected for assessing the outcomes of the program or policy should:

- Fit well with the program or policy objectives and
- Accurately measure the concepts of interest

If, for example, the concept of interest is delinquency, delinquency could be measured in different ways such as:

- Whether the youth had ever been adjudicated; or
- The number of times the youth self-reported illegal behavior such as underage drinking or stealing.

[“Reliability and Validity” Slide 27 of 40]

One factor the evaluator considers when deciding how to measure the outcomes of interest (such as delinquency) is how good of a job a particular indicator will do in measuring the outcome. Evaluators refer to this as the reliability and validity of the measures. Reliability is about consistency and validity is about accuracy.

An example often used to explain the difference between reliability and validity is a bathroom scale. If the scale is off by 5 pounds, provided the person’s weight had not changed, it would read the same weight every time the person got on the scale. The resulting weight measures produced by the scale would be reliable. However, as the scale is off by 5 pounds, it is not accurate; the measures produced by the scale are not valid.

Without getting overly complex here, the evaluation report should, at a minimum, explain why the measures were selected, the process for selecting the measures, and the time frame measured by the indicator (for example, whether employment was obtained within six months of program discharge).

[“Data” Slide 28 of 40]

When reviewing the reliability and validity of the measures in the evaluation report, one should also consider the impact of the data used for the measure on the outcome(s). For example, if you are using a data source as a juvenile’s criminal history records, and if you look at the number of times a juvenile is arrested as a measure of crime, this will produce a higher delinquency count than using the number of times a youth was sentenced to out of home placement. This is because arrests occur at the beginning of the process and will happen more frequently than those adjudicated delinquent and then sentenced to out of home placement.

Further, think about whether the measures used are sufficiently precise. For example, if the report is an evaluation of a truancy program, does it track absences from school in terms of whether or not the youth has ever had an unexcused absence, or the number of times the
Youth had an unexcused absence? The second indicator would permit an examination of decreases (or increases) in truancy rather than simply the presence or absence of truancy. Since it is not uncommon for children to fluctuate in the number and times absent from school, the second measure would be a better choice for a truancy program evaluation.

[“Timing of Follow-Up” Slide 29 of 40]

When looking at the outcomes reported, consider the timing used for follow-up, if any.

For example, an evaluation that has reported recidivism occurring three months post-adjudication placement may look more successful than if the recidivism checks occurred 12 months post-adjudication.

Evaluators should account for why they chose that particular follow-up time and address the likely implications of using a different follow-up time. The follow-up time should make sense given the design of the program, and the time should be comparable for all study participants. For example, if they used a 12-month follow up time for recidivism, they should identify and report on when the recidivism occurred within those 12 months.

[“Understanding the Report: Tip #5” Slide 30 of 40]

Are measures suitable? The report should explain why the measures were selected, the process for selecting the measures, and the time frame measured.

Evaluators should also account for why they selected that particular follow-up period and the time should be comparable for all study participants.

[“Statistical Tests” Slide 31 of 40]

Perhaps one of the areas most difficult for individuals without a strong research background to assess is whether the statistical tests used were appropriate and interpreted correctly. (This assumes that the report is a quantitative study -- and collects numerical data).

An often used phrase here is statistical significance. In an evaluation, the term usually refers to the size of the observed difference(s) between the program (treatment) and non-program (comparison or control) groups on the outcome(s) measured. When the authors note that a finding was statistically significant, they are saying that the observed difference (whether or not it showed the program to have the desired result) was large enough that it is unlikely to have occurred by chance.

Without a statistical test, we cannot know how much importance (or significance) to place on the size of the observed differences (presumably) produced by the program being evaluated.
As with the selection of measures, the evaluation report should clearly indicate why a particular statistical test was selected and how it was interpreted.

After assessing program effectiveness, some evaluators take an additional step to assess the economic implications of the program. Cost-effectiveness and cost-benefit analysis (or CBA) are two approaches to doing this.

Cost-effectiveness analysis examines the monetary costs and outcomes of the program. CBA assesses not only the costs and effects of the program, but also the benefits. CBA considers whether the benefits outweigh the costs. For more information on CBA, see the Cost-Benefit Knowledge Bank for Criminal Justice: [http://cbkb.org](http://cbkb.org)

Reports of both meta-analyses and systematic reviews synthesize the results of studies on a particular topic in order to produce a summary statement on the question being examined (such as do drug courts reduce substance abuse?). They typically only include the studies with strong research designs that were carried out well. These studies are important because they can be used to determine whether a program is evidence-based.

Similar to other research designs, these two approaches spell out in advance the procedures they use to compare studies included in the analysis. In this way, the approach should be able to be replicated by others. Though the terms are not necessarily interchangeable because of the methods used, meta-analyses and systematic reviews generally have the same purpose. The methods used in both are quite involved and discussion of the methods and how to assess the quality of these types of studies are beyond the scope of this guide, but the reader should be aware that both meta-analyses and systematic reviews include a much more complex, thorough, and methodical review than is conducted by a literature review.
It takes time to assess the quality of an evaluation report. This investment of time is worthwhile, however, because knowing whether a report is of good quality will help with issues that frequently come up. Why does this evaluator say this program works, for example, but that evaluator says it doesn’t? The quality of the evaluation report often contributes to these conflicting conclusions. Being able to assess the quality of an evaluation report will help one determine whether conflicting conclusions are related to the quality of the report.

This guide is by no means comprehensive; there is far more that can be done to assess an evaluation’s quality. Nevertheless, the guide should help practitioners make informed decisions about how much to rely on a particular evaluation.

Next we would like to review a few key evaluation terms

- **Sample Size:** is the number of subjects in the study.
- **Sufficient sample size:** Is the number of subjects large enough that there is confidence that the results represent the population being studied?
- **Statistical significance:** is whether the program or policy is likely to have caused the desired result or change.
- **Effect size:** is how much of a change the program or policy caused.

For more information on research designs and other relevant materials, see:

- For the OJP Crime Solutions at [http://crimesolutions.gov/](http://crimesolutions.gov/)
- And the Research Methods Knowledge Base at [http://www.socialresearchmethods.net/kb/](http://www.socialresearchmethods.net/kb/)
The checklist below will help users apply the information provided in this guide to a particular evaluation. Although the checklist does not provide a final score on evaluation quality, checklist users can be sure that they have examined the report for all applicable issues. Issues such as

<table>
<thead>
<tr>
<th>Understanding Juvenile Justice Evaluation Reports: Issues To Consider</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the research design (is it experimental, quasi-experimental, non-experimental)?</td>
</tr>
<tr>
<td>If applicable:</td>
</tr>
<tr>
<td>Did random assignment go as planned?</td>
</tr>
<tr>
<td>Did anyone in the control or comparison group receive the intervention?</td>
</tr>
<tr>
<td>Did the evaluation examine outcomes for program dropouts?</td>
</tr>
<tr>
<td>Is the sample size sufficient?</td>
</tr>
<tr>
<td>Does the report address attrition well?</td>
</tr>
</tbody>
</table>

This concludes this presentation.

NJEC is a project of the Justice Research and Statistics Association funded by the Office of Juvenile Justice and Delinquency Prevention (OJJDP). NJEC’s purpose is to improve the evaluation capacity of states, localities, and tribes and facilitate the use of evidence-based programs and practices in juvenile justice.

Notice of Federal Funding and Federal Disclaimer
This presentation was prepared under Grant Number 2010-JF-FX-0063 from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, United States Department of Justice. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.