National Resource Center on Justice Involved Women

Gender Responsive Interventions in the Era of Evidence-Based Practice: A Consumer's Guide to Understanding Research

Patricia Van Voorhis, Ph.D., Professor Emerita, University of Cincinnati

Acknowledgements

This document was written by Patricia Van Voorhis, Ph.D., Professor Emerita, University of Cincinnati. It was edited by Becki Ney and Rachelle Ramirez of the NRCJIW, with help from Debbie Smith.

This project was supported by Grant No. 2010-DJ-BX-K080 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the Department of Justice's Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Table of Contents

The imperative of **evidence-based practices** (EBP) governs the implementation of many public and private sector innovations. It follows that decision makers in the field of criminal justice should require that interventions be empirically tested and found to be effective prior to implementation. Additionally, the EBP mandate affords opportunities for non-tested practices to be piloted and tested under the auspices of the agency considering its use. Although some of the corrections EBP research is now more than 25 years old, it set forward several research-based principles (e.g., the principles of effective intervention; Andrews & Bonta, 1994) that are fervently followed and pertain to a wide array of correctional interventions and practices now in use throughout the world.

Evidence-based *practices are practices, programs, assessments, or policies that have been tested by methodologically rigorous research and found to be effective in reducing recidivism.*

Gender responsive *practices are practices, programs, assessments, or policies that account for the differences in characteristics and life experiences that women and men bring to the justice system AND that have been tested by methodologically rigorous research and found to be effective in reducing recidivism.*

Advancements in Gender Responsive Research

In the 1990s, a number of qualitative and quantitative studies called attention to important differences between male and female offenders. Evidence showed that women and girls had much higher incidences of trauma and mental health needs than men and boys. Women and girls were more likely to be involved in dysfunctional relationships, poverty, and unsafe environments, and to be single parents. Substance abuse, which characterized the majority of female offenders, was likely to be intertwined with problems associated with mental health and trauma (Bloom, Owen, & Covington, 2003). In time, **gender responsive** assessments and programs were developed to identify and address these needs.

The EBP mandate required evidence of the effectiveness of these gender responsive assessments and programs; yet, at the time, nearly all of the research was being conducted on men. Many felt that the overwhelming evidence garnered from studies of men and boys meant that the tested programs would apply to women as well, even though they had not been sufficiently tested on women. At the same time, emerging gender responsive approaches were rejected by some, because they had not been sufficiently researched.

More recent research now offers empirical support of the effectiveness of programs designed specifically for women, such as Seeking Safety (Najavits, Gallop, & Weiss, 2006; Najavits, Weiss, Shaw, & Muenz, 1998), Moving On (Duwe & Clark, 2015; Gehring, Van Voorhis, & Bell, 2010), Helping Women Recover (Messina, Grella, Cartier, & Torres, 2010), Beyond Violence (Kubiak, Fedock, Bybee, & Kim, in press; Messina, 2014; Messina, Braithwaite, Calhoun, & Kubiak, 2016), and at least two parenting programs (Grella, 2009; Olds et al., 2004). The Women's Risk/Needs Assessment (WRNA) has been validated in several probation (Van Voorhis, Bauman, & Bruschette, 2013a), pre-release (Van Voorhis, Bauman, & Bruschette, 2012), and prison sites (Van Voorhis, Bauman, & Bruschette, 2013b) (see also Van Voorhis, Wright, Salisbury, & Bauman, 2010). Moreover, a recent meta-analysis found that high fidelity women's programs are not only effective but they are more effective for women than high fidelity, gender neutral programs (Gobeil, Blanchette, & Stewart, 2016).

Purpose of this Monograph

Even with evidence to support the effectiveness of gender responsive programs, it is still necessary to be good consumers and supporters of research. This monograph underscores the need for policymakers and practitioners to understand the fundamentals of research in order to guide their work with justice involved women. This means being informed about existing research (i.e., findings from both gender neutral and gender responsive research) and understanding how studies are conducted (e.g. sample sizes, research designs, outcome measures). It also means being astute designers of studies they wish to conduct within their own agencies and ensuring – through the collection and evaluation of their own data – that policies, programs, and practices (i.e., the use of assessments) are being implemented with fidelity and achieving their intended outcomes.

This monograph is organized into four sections that discuss the four main types of studies that pertain to gender responsive approaches. Each type of study contributes a distinct form of evidence. Similarly, each type of study addresses some scientific questions but not others.

- **Qualitative population profiles:** Early studies that provided information about the characteristics and needs of female offenders.
- **Risk assessment or prediction studies:** Studies involving the construction of risk/needs assessments such as the Women's Risk/Needs Assessment (WRNA) (Van Voorhis et al., 2010) that identified the needs that are statistically correlated with recidivism for women.
- **Evaluation studies:** Studies of correctional interventions designed to treat and ameliorate a variety of offender needs. The fundamental question addressed by these studies is whether the intervention effectively addressed or stabilized a problem and, if so, if the improvement led to reductions in recidivism.
- **Meta-analyses:** Studies that combine the findings from many independent studies. Because of an historical lack of women-only studies, meta-analyses of gender responsive studies have only just begun as an increasing number of rigorous gender responsive studies become available (see Gobeil et al., 2016).

Each of these four sections includes a short discussion of the importance of the research, including the benefits, contributions, limitations, and controversies; a description of how such studies are designed and conducted; and a guide highlighting the salient points of which consumers of gender responsive research should be aware.

Profiles of Women Offenders Resulting from Qualitative Research

The first credible glimpses into the nature of female offending were provided by feminist scholars who conducted in-depth, open-ended or semi-structured interviews with justice involved women and girls (e.g., Bloom, 1996; Chesney-Lind, 1997; Chesney-Lind & Sheldon, 1992; Daly, 1992; Owen, 1998; Richie, 1996). These revealed rich life stories of female delinquents and justice involved women. After listening to their stories and following the

> ***Qualitative research*** *refers to a research methodology that is used primarily for exploratory purposes. Data are collected through unstructured or semi-structured methods, often involving participant observation, focus groups, or interviews. Sample sizes are often small.*
>
> *Qualitative studies, unlike quantitative studies, do not involve statistical analyses. The observation and questioning strategies used in qualitative studies afford an opportunity to learn about relationships, motivations, trends, and other complex processes and their various interactions. Qualitative studies often generate hypotheses for quantitative studies.*

rigorous rules of **qualitative research**, these scholars identified repeated themes that they then used to differentiate between types of female offenders. The girls' and women's narratives often told of "pathways" to delinquent or criminal involvement; they often related how one tragedy (e.g., abuse) led to another (e.g., mental health problems) that led to the need to self-medicate (i.e., substance abuse) that led to being arrested for substance abuse related crimes. Other themes and pathways were observed, but it gradually became fairly clear that the narratives of female offenders were quite different than those of males.

On a political level, since the early authors were feminist, they were subject to rejection by those who rejected the feminist movement. Other criticisms faulted their choice of methodology: the authors of these studies provided few statistical analyses. The critics of qualitative research observed that the samples were too small and therefore could not adequately represent the broader population of female offenders. For example, in early editions of their highly regarded textbook The Psychology of Criminal Conduct, Andrews and Bonta (1994) ranked study methodologies according to their rigor and to the faith one could place in the veracity of their findings. Qualitative research was relegated to the lowest rung on the hierarchy of research methodologies. Furthermore, in early editions of their textbook, Andrews and Bonta suggested that female scholars were placing a feminist lens on their observations and therefore lacked objectivity.

On a political level, since the early authors were feminist, they were subject to rejection by those who rejected the feminist movement. Other criticisms faulted their choice of methodology: the authors of these studies provided few statistical analyses. The critics of qualitative research observed that the samples were too small and therefore could not adequately represent the broader population of female offenders. For example, in early editions of their highly regarded textbook The Psychology of Criminal Conduct, Andrews and Bonta (1994) ranked study methodologies according to their rigor and to the faith one could place in the veracity of their findings. Qualitative research was relegated to the lowest rung on the hierarchy of research methodologies. Furthermore, in early editions of their textbook, Andrews and Bonta suggested that female scholars were placing a feminist lens on their observations and therefore lacked objectivity.

A Consumer's Guide to Qualitative Research

- Contrary to the criticism stated above, qualitative research is a valid, scientifically rigorous approach. Qualitative researchers must: 1) follow certain guidelines to ensure that their samples adequately represent the populations of individuals that are the subject of their studies; 2) use various techniques (e.g., additional reviewers, consensus reviews) to ensure that their findings are reliable (i.e., that other researchers would interpret the findings in a similar manner); and 3) follow scientific guidelines for identifying themes to guard against criticisms that their observations are biased by their personal world views (see Silverman, 2010).

- Exploratory, qualitative research is essential, especially when there is limited existing knowledge. This was the case in the 1980s and 1990s, when many of these initial studies emerged; very little research had been conducted on female offenders before that time. In the absence of sufficient knowledge, open-ended interviews provide an excellent way to collect information on the unknown, to seek additional information when observations need further explanation, and to expose fine nuances and complex interactions among influencing factors.

- Qualitative research can unearth issues that can later be researched quantitatively. In most studies, issues emerge that cannot be answered by the data available to the researcher. In fact, good studies generate unanticipated questions. The studies cited above, for example, identified women's needs pertaining to abuse, mental

health, dysfunctional relationships, poverty, health, and child rearing. Some women also evidenced needs typical of male offenders, such as antisocial peers. However, precise questions about prevalence, or the percentage of the female offender population affected by these needs, could not be answered in the qualitative studies. Similarly, qualitative studies could not show whether identified needs were risk factors, and whether treatment could ameliorate them. These questions awaited longitudinal data, and prediction and evaluation research (see Sections 2 and 3).

- Quantitative studies later supported the results of these earlier qualitative studies of women. For example, subsequent quantitative surveys found high proportions of women offenders suffering from mental illness, abuse, poverty, dysfunctional relationships, and parental stress. Other studies showed that these prevalences were significantly greater among women than men (see Salisbury & Van Voorhis, 2009 and Van Voorhis et al., 2010 for reviews). Using various statistical clustering and path analysis techniques, studies quantitatively replicated some of the pathways (Brennan, Breitenbach, Dieterich, Salibury, & Van Voorhis, 2012) and showed that they were statistically and significantly related to later offense-related outcomes (McClellan, Farabee, & Crouch, 1997; Salisbury & Van Voorhis, 2009). Additionally, the prediction research discussed below later found most of the gender responsive needs identified in the qualitative studies to be predictive of prison misconducts and future offending (Van Voorhis et al., 2010).

Prediction Studies: The Creation of Gender Responsive Risk/Needs Assessments

The findings of **prediction studies** have changed the face of correctional practice. Prediction research resulted in assessments used to classify justice involved individuals according to their risk (high, medium, or low) of incurring offense-related

Prediction studies *seek to identify social and demographic background characteristics that, in corrections, predict offense-related outcomes. Depending upon the study, predictors typically include offender needs, demographic factors such as age and living environment, and criminal history and current offense attributes.*

outcomes such as pretrial failure, new offenses, prison misconducts, and failure to abide by the conditions of community supervision. Knowledge of risk then determined levels of community supervision, security of prison living situations, and identified those offenders most in need of treatment resources and other services.

Prediction studies are quantitative: they involve numbers and structured information-gathering procedures (e.g., forced choice surveys/interviews, reviews of offender records). The research typically takes place over a minimum of two time periods:

- Time period 1: Data for the variables that represent the predictors is obtained. Prior to collecting this information, researchers must construct data collection tools or computer programs that will record information on the predictor variables for each offender. The researchers' knowledge of criminological research determines the predictors to include in the study.
- Time period 2: Data is collected pertaining to whether the individual actually committed the offense-related behaviors that the research is designed to predict (e.g., new offenses, violations of the terms of correctional or court supervision, or prison misconducts).

Time must elapse between the collection of the predictors in time period 1 and the outcome variables in time period 2. The time usually ranges from 6 months to 3 years.

When all of the data have been collected, coded, and prepared, statistical analyses are conducted in

order to determine which of the potential predictors are, in fact, significantly related to the offense-related outcomes. The analyses produce **measures of association**, most often **correlation coefficients** or **odds ratios (OR)**, which show how strongly a variable impacts the outcome behavior. For example, for women offenders, the likelihood of recidivism increases as scores on an anger scale increase. This typically produces a positive correlation between the anger scale and the recidivism measure. Negative correlations may also occur, as when increases in a scale reduce the likelihood of the offense-related outcome. For example, increases in certain forms of family support tend to reduce the likelihood that an inmate will incur serious prison misconducts. The research produces a second statistic that shows whether the correlation coefficient or odds ratio is statistically significant. Regardless of the size of the statistical measure of association, no confidence can be placed in it if **statistical significance** is not found.

The final risk scale is a single measure formed by combining all significant predictors of the outcome variable. When that scale is formed, it too is correlated with the outcome variables, and measures of the strength of the association and its significance are computed. At this point, a measure of the **area under the curve (AOC)** should also be calculated.

The task of constructing a risk assessment (or prediction tool) involves construction validation research and a round of revalidation studies. It is essential to revalidate the tool on different samples and settings. At some point in time, a researcher may decide to revise an assessment; when this is done, the revised assessment should also be revalidated.

In corrections, early prediction, or risk assessment, instruments were constructed using demographic measures (e.g., age), criminal history, and prior offense measures. Such static instruments could classify offenders into high, medium, and low risk categories, but they could not be used to assess needs (see Van Voorhis & Salisbury, 2014). Examples of such tools included the U.S. Parole Commission's Salient Factor Score (Hoffman, 1994) and a host of other correctional custody classification tools.

With the advent of the Level of Service Inventory-Revised (LSI-R; Andrews & Bonta, 1995) and the Northpointe COMPAS (Brennan, Dieterich, & Oliver, 2006), needs that were predictive of future offending were also included in the risk assessment scales. The new tools were referred to as dynamic risk/needs assessments. These tools performed one of the same functions as the earlier tools—they still classified justice involved

__"Measure of association"__ is a broad term encompassing all statistical coefficients that portray how strongly a measure of interest (e.g., a predictor) is associated with the outcome of interest.

__Correlation coefficients__ show the results of an analysis that associates a causal variable (e.g., total scores on a risk assessment) with an outcome of interest (e.g., number of prison disciplinary infractions).

An __odds ratio (OR)__ is a measure of association between an exposure to something (e.g., participation in a correctional treatment program) and an outcome (e.g., whether a new arrest occurred within a two year timeframe).

__"Statistical significance"__ refers to whether we can put any faith in the reported measures of association. Without sufficient statistical significance, we would conclude that the measure of association occurred by chance and does not reflect a true association.

__Area under the curve (AOC)__ is the "hit rate" or extent to which a prediction tool correctly identifies the individuals who (in this case) recidivate and those who do not recidivate.

individuals as high, medium, and low risk. However, they also served as needs assessments: each of the needs included in the risk scale had its own score. Therefore, practitioners could determine an individual's risk score and then identify the needs that contributed to that score. High risk offenders could be supervised more intensively and given high priority to attend programs targeted to those needs that received high scores. The second advantage of these assessments was that scores could change over time. For example, an offender completing high school, or found to have achieved abstinence as a result of substance abuse treatment, would score lower on a new assessment. The early static and dynamic risk/needs assessments were constructed on samples of male offenders and later revalidated on samples of female offenders and are referred to as gender neutral assessments.

Assessments specific to women offenders were developed in the late 1990s and included specific needs that were found to be predictive for women offenders. The additional predictors included mental health, anger, dysfunctional relationships, abuse and trauma, safety, and parental stress, and strengths pertaining to self-efficacy, family support, parental investment, and educational assets. The earlier gender neutral dynamic risk/needs assessments were predictive for women, but the new gender responsive tools proved to be more predictive in most of the samples studied (Van Voorhis et al., 2010; Van Voorhis et al., 2012; Van Voorhis et al., 2013a; Van Voorhis et al., 2013b).

A Consumer's Guide to Prediction and Risk Assessment Research

- Assessments should only be used on individuals similar to those who were studied in the original construction and revalidation research. The characteristics of the research samples determine for whom the assessment is appropriate. For example, if an assessment is created and validated on samples of men, it should be used on men, and not on women or juveniles. This is a well-known principle of research, called **external validity**, but it is often violated in the field of corrections. For example, most custody classification assessments were developed and revalidated on men and applied to women with little research; when they finally were validated on women, it was clear that they were less valid for women than for men (Hardyman & Van Voorhis, 2004). When gender neutral, dynamic risk/needs assessments were validated on women, they were found to be predictive/valid for women (see Smith, Cullen, & Latessa, 2009), but newer gender responsive assessments proved to be even more predictive (Van Voorhis et al., 2010).
- Repeated references to the validity of male-based assessment tools for women should not be used to refute new gender responsive assessments for women. The original gender neutral assessments were created for men and only later validated for women. The findings that the male-based, gender neutral assessments are valid for women should not be taken to mean that improvements are not warranted. The comparison is "apples" to "oranges," because the earlier validation studies, though finally conducted on women, did not study the additional gender responsive variables.
- The practical uses of the assessment should closely follow the research. The construction and validation research also determines the applications of an assessment. If an assessment is validated on a probation sample, for example, it should not be used on a pretrial or prison sample unless it is validated on such samples. If a custody classification tool is validated on a prison sample to predict prison misconduct, it should not be used to make prison release decisions unless

External validity *refers to the generalizability of research findings to individuals who are similar to those who have been studied. A study is not externally valid for groups it did not study.*

research validates whether the classification tool predicts community recidivism (most custody assessments do not; Hardyman & Van Voorhis, 2004). Perhaps the misapplication that creates the most apprehension for researchers is the use of a risk/needs assessment tool to make a sentencing decision regarding prison or probation, since most assessment validation studies are conducted on samples that have already been placed in a correctional option—prison, probation, or parole.

- The failure of an assessment to predict the outcomes during revalidation research does not always implicate the assessment. There are several alternative explanations that should be ruled out first:
  - The assessment may have been administered incorrectly, for example, when an assessment requiring an interview is completed in a group setting or through record data, or is administered by an untrained interviewer.
  - Poor validity (a failure to predict) may also implicate the outcome variable. When jurisdictions lose control of the quality of recidivism measures, it is highly likely that an assessment found in earlier research to be valid will be found to be invalid in a later study.
  - Sometimes valid assessments can appear invalid when the assessment is followed by the effective treatment of risk/need factors before these factors are themselves validated. This can occur when the treatment of risk factors changes them, and by the time the offense data is collected, the original risk measures may no longer validly describe the treated individual. In such cases, a post treatment assessment is the one that should be revalidated.

- Assessments don't do well at predicting the occurrence of extremely rare events or extremely frequent events. If base rates of such events are extremely low, it will be more accurate to predict that the event will never occur than to devise a prediction scale to perform the prediction. This fact of science often frustrates the construction of risk assessments to comply with the Prison Rape Elimination Act (PREA) as well as the construction of violence prediction tools.

- Beware of the vested interests in the assessment industry. The costs of using off-the-shelf risk/needs instruments vary from minimal (e.g., training, software, and photocopying) to extremely high (e.g., training, software, ongoing royalties, and software fees). Scientifically sound assessments should provide the means for other researchers to replicate the research. Moreover, an assessment should never be purchased without the examination of validation research available in published articles, test manuals, or elsewhere.

- Further research is needed to determine whether dynamic risk/needs assessments identify more high risk offenders than static assessments relying only upon criminal history measures; some studies suggest that dynamic risk/needs assessments over assess risk (Austin, 2006; Caudy, Durso, & Taxman, 2013). Typically, the overlap between the two is very high—that is, the high risk offender, as identified by the static instrument, also is highly likely to be high risk on a dynamic/risk needs instrument.

- Some maintain that all of the risk/need measures on an assessment should predict all of the time (Austin, 2006; Caudy et al., 2013). This assertion ignores sample variations that affect all assessment studies. In fact, correctional samples show a good deal of variation, especially on measures of needs. The predictive validity of individual scales across individual studies can vary by: 1) how much treatment is received prior to the assessment; 2) sentencing strategies (e.g., "get tough" versus the use of alternative sentences); 3) how troubled a population is; and 4) demographic variations across correctional populations and

geographic areas. Just the same, the total of the risk/needs factors should predict regardless of sample. Moreover, the variation across studies should not impact meta-analyses because they are designed to "smooth out" sample variations.

- The research on women's risk/needs assessments should not be faulted for not having male comparison groups. In truth, focused attention on specific populations is far from rare in science, and we are all beneficiaries of that fact. The absence of a comparison group in prediction or epidemiological studies is not the same fatal flaw that it is in evaluation research (see Section 3, below). It is true that a new wave of research should examine male offender populations for the impact of trauma, mental health, relationship dysfunction, abuse, and safety issues on their recidivism. It is also true that we should not create a new external validity problem by applying the gender responsive variables to men without appropriate research. But both issues can be addressed through new research, and both suggestions do not imply that it was wrong to study women apart from men.

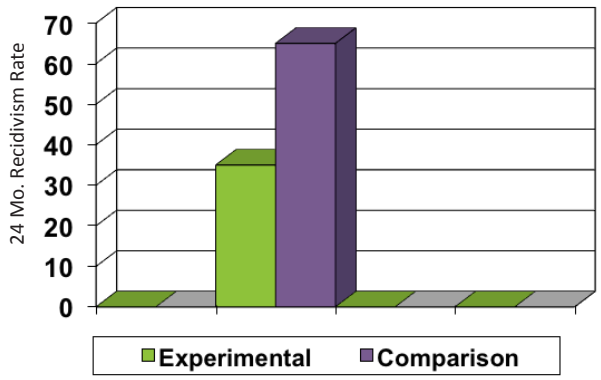
Evaluation Studies: Do Gender Responsive Programs Work to Reduce Recidivism?

In the age of evidence-based practices, it is very difficult to advocate for programs that have not been evaluated and found to be effective through **controlled studies**. "Controlled studies" are also referred to as "experimental studies." In these studies, there is an experimental group, which receives treatment, and at least one comparison group (also called a control group) which does not receive treatment. It is hoped that those in the experimental group have better outcomes than those in the comparison group. In Figure 1—a hypothetical example—a group of female offenders who participated in a cognitive behavioral program (i.e., the experimental group) had a two-year recidivism rate that was 30 percent lower than that of a similar group of female offenders who did not participate in the cognitive behavioral program (i.e., the comparison group).

Controlled studies *involve an experimental group (a group receiving some type of treatment) and a comparison group of similar individuals who do not receive the treatment. Controlled studies are also referred to as "experimental studies."*

Reductions in the recidivism rate (for the experimental group vs. the control group) is the sought after outcome of interest; however, researchers sometimes examine the experimental program's effect on intermediate outcomes, or changes that occur at the immediate conclusion of program participation (e.g., see Kubiak, Kim, Fedock & Bybee, 2012; Najavits, Weiss, Shaw, & Muenz,

Figure 1: Recidivism Rate for Hypothetical Experimental and Comparison Groups 24 Months Following Program Completion



1998). A number of the evaluations of gender responsive programs, for example, examine the program's impact on anxiety or depression at the conclusion of the program, as improvements in such mental health attributes are assumed to foreshadow reductions in recidivism.

Evaluation studies require researchers to set forward a research plan that determines criteria for selecting participants. Decisions must be made about how the experimental and comparison groups will be formed (e.g., through **random assignment**, **matched comparison groups**, **ad hoc comparison groups**, or **propensity score matching**). Typically, all data collection documents and procedures must be designed in advance. As participants enter a program, data pertaining to their background and perhaps their scores on pretest measures of the treatment targets are collected. Program **process measures** are collected while participants attend the program. At program completion, the researchers may administer surveys or assessments of intermediate outcomes (e.g., reading level, self-efficacy, anxiety, problem solving) to determine whether changes were seen as a result of treatment. As with prediction studies, the collection of recidivism data takes place after a period of time elapses—6 months to 3 years, perhaps.

The final study results will focus on showing whether the intervention reduced recidivism (i.e., whether being in the comparison group or the experimental group made a difference). However results can be presented in a myriad of ways. Results can be expressed as the percent recidivism for each group, which can be shown in a bar graph such as the one in Figure 1. Results can also be expressed using a correlation statistic. In this case, the higher the correlation value, the stronger the effect the program had in reducing recidivism. Good results usually surpass .20; best results may surpass .30, but this is rarer. A correlation of .00 means that the program had no effect. A coefficient with a minus value usually means that the comparison group had better results than the experimental group—in other words, that the comparison group had a lower recidivism rate than the experimental group.

Regardless of how the results are presented, readers should also see a probability value showing whether the results were statistically significant. The issue here is not the strength of the relationship but whether, according to the rules of probability, we can place much faith in the results.

Random assignment *is the process of randomly assigning eligible participants to either the experimental or control group.*

Matched comparison groups *is the use of a comparison group that is matched to an experimental group based on characteristics that could affect research results. This is done case by case. By matching in this way, potential biasing factors are balanced between the two groups.*

Ad hoc comparison groups *are groups that are selected to be control groups without any matching or random assignment. Ad hoc comparison groups are only workable if the groups are similar in terms of key demographic and background attributes that could bias the comparison between experimental and control groups on outcome measures.*

Propensity score matching *is a modeling procedure that allows researchers to statistically mimic the similarities expected between participants randomly assigned to treatment and comparison groups.*

Process measures *measure various attributes of the program's quality and integrity, and may include, for example, attendance rates, program completion, facilitator skills, etc.*

When outcomes are intended to measure the differences between test scores, rather than simply answering yes or no to the question of whether recidivism was reduced, **t-tests** or **analysis of variance** statistics may be used. Readers will be able to observe the different mean scores and will be informed about whether the differences are statistically significant.

T-tests are tests of whether means are statistically significant.

Analysis of variance (ANOVA) are tests of whether the variances around means are statistically significant.

T-tests and ANOVAs do not measure the strength of an association (like a correlation or odds ratio), only statistical significance.

A Consumer's Guide to Evaluation Studies of Gender Responsive Programs

- Conduct more research on gender responsive programs. Although gender responsive programs have sometimes been unfairly criticized, it is nevertheless very true that more studies are needed.
- Program evaluations should not begin until the program has been piloted for a period of time and determined to be running with fidelity to the program design. Beginning evaluation studies too soon, before all procedures for quality assurance are in place, can make good programs look bad. For example, outcome studies of the Beyond Violence program (Covington, 2013) were preceded by in depth process evaluations of the program's fidelity and feasibility.
- An evaluation study must have a control or comparison group. There are many reports of interventions that describe a program or policy, but they are just that—descriptions that give us no basis for knowing whether the results are directly related to the intervention. These cannot be considered to be tests of the intervention unless there is a comparison group.
- The characteristics of the comparison group should be as similar as possible to those of the experimental group. This is achieved through the use of random assignment, matched comparison groups, ad hoc comparison groups, or propensity score matching.
- The failure of an experimental program to reduce recidivism does not always implicate the experimental program. There are alternative explanations that should be ruled out first:
  - Determine if the program was implemented with fidelity. Correctional programs take place in turbulent environments that can mar the quality of well-designed programs. When program integrity fails, the evaluation likely will show that the program had no effect in reducing recidivism. In this way, poor program quality can make a good program look ineffective. Measuring program quality (through process measures) and reporting whether the program was delivered as it was supposed to may help explain the program's poor outcomes. In studies where a program is implemented in multiple sites, researchers can reanalyze the data by separating sites where the program was implemented with high fidelity from sites where the program was implemented with low fidelity. In doing so, it is not unusual to find that sites where a program was implemented with high fidelity have better results than a site where the program was implemented with low fidelity.
  - Determine if the "wrong" individuals were admitted to the program. Most programs are designed for individuals with specific needs and should not be expected to work with individuals who do not have those needs (e.g., as when a parenting program is administered to a woman who has no children). Often, this error involves admitting low risk individuals into programs designed for high risk individuals. Researchers may admit the "wrong" individuals into a program because admission procedures are constrained by court requirements or the need to show program utilization, but doing so may mar or mask evaluation results. In such cases, good outcomes for appropriate program participants may be cancelled out by bad outcomes achieved by inappropriate participants. In such cases, separate analyses

can be conducted for each group in order to gain a more accurate understanding of the program effects.

A simple analysis of a program's effect on recidivism may raise more questions than it answers. Although a presentation of the outcomes for all participants is required, additional analyses should provide enough information to show what worked and what did not work. Enough information should be available to show how the program might be replicated and what shortcomings should be avoided in future uses of the program.

The State of the Art of Gender Responsive Approaches as Depicted in Meta-analysis

The statistical technique known as **meta-analysis** contributed significantly to the evolution of evidence-based practice. "Meta-analyses" can be defined as "studies of the studies." They use methodologically rigorous strategies to statistically synthesize findings from numerous experimental studies. Researchers assemble all available controlled studies of a specific intervention—cognitive behavioral programs, for example—and then calculate a summary "effect size" statistic that shows how effective the intervention is across studies. More confidence can be placed in the results of a synthesis of studies than the results of a single study or the results of a simple tally of outcomes across studies (e.g., vote counting). Meta-analyses show the effect size using a number or statistical values, including various correlation coefficients or odds ratios, or confidence intervals.

Studies may be coded in several ways, for example, according to:

- their characteristics (e.g., published vs. unpublished)
- the setting (e.g., community vs. institution)
- program characteristics (e.g., group size)
- characteristics of the study participants (e.g., proportion of medium to high risk subjects).

***Meta-analysis** is a statistical technique for combining the results of numerous independent controlled studies. The combined effect size gives a truer picture of actual outcomes and smooths out differences between studies.*

In reporting the results, the researcher can then partition findings according to these various attributes, for example, the effect size for the entire sample versus the effect size for high risk offenders. Of course, the researcher cannot account for any attributes that are not discussed in the individual studies; this places some limits on how detailed the meta-analysis can be.

The ability to evaluate program qualities for their impact on participant outcomes led to efforts to more finely isolate the program attributes that separate effective programs from ineffective ones. These in turn led to program assessment tools such as the Correctional Program Assessment Inventory (CPAI 2000; Gendreau & Andrews, 2001) and the Correctional Program Checklist (CPC; University of Cincinnati Corrections Institute, 2006). Authors relied on more research than the meta-analyses alone in constructing these tools, but the meta-analyses were highly influential. These early program assessments have been validated; programs scoring higher on the assessments are more likely to have better participant outcomes than those receiving low scores. Again, however, the assessment tools are based largely on studies of male offenders.

The classic meta-analyses in corrections (e.g., Andrews et al., 1990; Lipsey, 1992) were published in the early 1990s and are now approximately 20 to 25 years old. Even so, they ushered in the evidence-based movement in corrections and created incontrovertible support for the effectiveness of correction treatment in comparison to other correctional paradigms such as deterrence, incapacitation, or punishment. Unfortunately these early meta-analyses were based primarily on studies

of male offender populations, because there were so few evaluations involving women participants.

In order to be included in a meta-analysis, several studies of the program or similar programs must exist. Programs that are too new to have been evaluated or to have been evaluated several times cannot be included. To date, this has been the case with most gender responsive programs; however, the situation is changing with the emergence of new research (e.g., Gobeil, Blanchette, & Stewart, 2016).

A Consumer's Guide to Understanding the Relevance of Meta-analyses

Meta-analyses are considered the "gold standard" of EBP; however, they should be scrutinized for the following issues:

- Some meta-analyses may be affected by sampling biases; that is, their findings may be affected by the choice of studies included in the research. If a meta-analysis is restricted to only published studies, for example, most of the studies will be those that showed the intervention to be effective. Studies that produce insignificant findings are often not submitted for publication. Similarly, good meta-analyses use studies found in various sources rather than ones found in a single source (e.g., a single journal).
- Coding biases may occur when documenting study and program characteristics. Definitions for how a study should be coded should be agreed upon in advance. The meta-analysis should describe the procedures for ensuring the reliability of the codes.
- The fact that a gender responsive program is not studied in a meta-analysis should not be the basis for excluding the program from implementation. If only those programs studied in meta-analyses were implemented, the field of corrections would not benefit from the vital interventions discovered in single, well executed, experimental studies, just as the field of medicine would not benefit from the findings of a single, well executed, experimental study of a new cancer treatment until it had amassed enough additional studies to be included in a meta-analysis.
- Don't look to older meta-analyses for evidentiary support for the implementation of new programs. If the program could not have been included in the meta-analysis, the meta-analysis is irrelevant to the task of providing evidentiary support. It would be appropriate in such cases to look for single studies instead. These studies appear in journal articles and in program and assessment websites. Government databases such as the Substance Abuse and Mental Health Services Administration's (SAMHSA) National Registry of Evidence-Based Programs and Practices (NREPP; www.samsha.gov/nrepp) or the Office of Justice Programs' (OJP) Crimesolutions.gov are excellent resources in this regard. These websites provide ratings of the quality or rigor of each study. The Washington State Institute for Public Policy is another resource that tracks a wide variety of programs according to cost effectiveness (www.wsipp.wa.gov), including many gender responsive programs.
- Program assessment tools that are based on studies of male offenders are not relevant to gender responsive programs. Such program assessments place the wrong template on gender responsive programs and do not account for the importance of trauma programs, parenting classes, healthy relationship classes, and gender responsive substance abuse treatment for women. Although they are in an early stage of development, several alternative gender responsive program assessment tools exist. These include the Gender Informed Practice Assessment (GIPA; Center for Effective Public Policy, 2010) and a self-administered tool called the Gender-Responsive Policy and Practice Assessment (Bloom, Covington, Messina, Owen & Selvagi, 2014)—both supported by the National Institute of Corrections—as well as the Gender-Responsive Community Programs Inventory (Van Voorhis, 2015).

- Agencies should continue to conduct their own controlled studies. This may involve placing faith in an untested program for the period of time needed to pilot and evaluate the program, but the end result could be an evidence-based option. In the case of gender responsive programs, this would be far from a leap of faith since most gender responsive approaches use modalities that have been proven effective in other disciplines, such as psychology, mental health, medicine, education, and social work (see Bloom et al., 2003).

Conclusion: Being an Educated Consumer of Research to Ensure that Gender Responsive Programs Are Evidence-Based

Clearly, the volume of evidence supportive of gender responsive approaches is growing. In this regard, the importance of Gobeil and colleagues' 2016 meta-analysis cannot be overstated. Critics may wish to argue that the male-based, meta-analytic studies contain more studies and benefit from more years of replicated research; however, as the number of gender responsive studies increases, the results are becoming irrefutable: gender responsive approaches do make a difference in terms of improving outcomes for justice involved women. Nonetheless, it is imperative to garner additional evidence to support gender responsive approaches; further studies are required.

Consumers should also beware of adopting/purchasing programs or assessments from vendors who claim to be evidence-based but who do not produce evidence (or whose evidence actually reflects the results of tests conducted on similar programs or assessments). Vendors of programs and assessments that are truly evidence-based often post validation or controlled tests on their websites. When these are not available, it would be wise to contact the vendor.

What if a decision maker sees the value in a practice but cannot locate evidence? This was, after all, the state of science at the very beginning of the move to gender responsive programming, and it may still pertain to some types of policy and programmatic choices. In these circumstances, agencies should be encouraged to conduct their own evaluations. These evaluations can be done through partnerships with universities or research firms, or by developing internal research capabilities. However, two precautions are recommended:

- The approach should be based in psychological, educational, or mental health theory or practice. To their credit, most gender responsive approaches are.
- The program should be piloted for a period of time, allowing the opportunity for start-up problems to be resolved and for participants to implement the program with fidelity (see Kubiak et al., 2014).

While a program may not be considered evidence-based at the beginning of the research study, it may be by the end. Moreover, such studies build agency capacities in ways that extend beyond the evaluation itself. Evaluations provide information on agency processes and many show how these attributes can affect participant outcomes and organizational functioning. Such evidence can then begin to inform organizational development and decision making. In short, a study seeking evidence on a particular program could lead to an evidence-based organization.

As with most public sector endeavors, gender responsive programs and services will continue to require supporting, empirical evidence. Decision makers must become increasingly adept at assembling evidence needed to advocate for these interventions. This will require that they be well read and knowledgeable of available research, astute consumers of packaged programs, and informed designers of research they may need to conduct in their own agencies. It is hoped that this monograph offers initial support to these ongoing initiatives to help practitioners achieve better outcomes for justice involved women.

References

Andrews, D., & Bonta, J. (1994). *The psychology of criminal conduct.* Cincinnati, OH: Anderson.

Andrews, D., & Bonta, J. (1995). *Level of Service Inventory-Revised.* North Tonawanda, NY: Multi-Health Systems.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, *28*, 369–401. http://dx.doi.org/10.1111/j.1745-9125.1990.tb01330.x

Austin, J. (2006). How much risk can we take? The misuse of risk assessment in corrections. *Federal Probation*, *70*, 58–63.

Bloom, B. E. (1996). *Triple jeopardy: Race, class and gender as factors in women's imprisonment.* Riverside: University of California.

Bloom, B., Covington, S., Messina, N., Owen, B. & Selvaggi, K. (2014). *Gender-Responsive Policy & Practice Assessment.* Retrieved from National Institute of Corrections website: http://nicic.gov/grppa/

Bloom, B., Owen, B., & Covington, S. (2003). *Gender-responsive strategies: Research, practice, and guiding principles for women offenders.* Retrieved from National Institute of Corrections website: https://s3.amazonaws.com/static.nicic.gov/Library/018017.pdf

Brennan, T., Breitenbach, M., Dieterich, W., Salisbury, E., & Van Voorhis, P. (2012). Women's Pathways to Serious and Habitual Crime: A Person-Centered Analysis Incorporating Gender Responsive Factors. *Criminal Justice and Behavior, 39*(11), 1481-1508.

Brennan, T., Dieterich, W., & Oliver, W. (2006). *COMPAS: Technical manual and psychometric report (Version 5.0)*. Traverse City, MI: Northpointe Institute.

Caudy, M. S., Durso, J. M., & Taxman, F. S. (2013). How well do dynamic needs predict recidivism? Implications for risk assessment and reduction. *Journal of Criminal Justice, 41*, 458–466. http://dx.doi.org/10.1016/j.jcrimjus.2013.08.004

Center for Effective Public Policy. (2010). *Gender Informed Practice Assessment (GIPA)*. Silver Spring, MD: Author.

Chesney-Lind, M. (1997). *The female offender: Girls, women, and crime.* Thousand Oaks, CA: Sage.

Chesney-Lind, M., & Shelden, R. G. (1992). *Girls, delinquency and juvenile justice.* Belmont, CA: Thomas Wadsworth.

Daly, K. (1992). Women's pathways to felony court: Feminist theories of law breaking and problems of representation. *Southern California Review of Law and Women's Studies, 2*, 11–52.

Duwe, G., & Clark, V. (2015). The importance of program integrity: Outcome evaluation of a gender responsive, cognitive behavioral program for female offenders. *Criminology & Public Policy, 14*, 301–328. http://dx.doi.org/10.1111/1745-9133.12123

Gehring, K. S., Van Voorhis, P., & Bell, V. R. (2010). *"What Works" for female probationers? An evaluation of the Moving On program.* Retrieved from University of Cincinnati website: http://www.uc.edu/content/dam/uc/womenoffenders/docs/MOVING%20ON.pdf

Gendreau, P., & Andrews, D. (2001). *Correctional Program Assessment Inventory (CPAI-2000).* St. John, Canada: University of New Brunswick.

Gobeil, R., Blanchette, K., & Stewart, L. (2016). A meta-analytic review of correctional programs for women offenders: Gender-neutral versus gender-informed approaches. *Criminal Justice and Behavior, 43*, 301–322. http://dx.doi.org/10.1177/0093854815621100

Grella, C. (2009). *Female Offender Treatment and Employment Project (FOTEP): Summary of evaluation findings*. Los Angeles, CA: UCLA Integrated Substance Abuse Programs.

Hardyman, P. L., & Van Voorhis, P. (2004). *Developing gender-specific classification systems for women offenders*. Retrieved from National Institute of Corrections website: http://static.nicic.gov/Library/018931.pdf

Hoffman, P. B. (1994). Twenty years of operational use of a risk prediction instrument: The United States Parole Commission's Salient Factor Score. *Journal of Criminal Justice, 22*, 477–494. http://dx.doi.org/10.1016/0047-2352(94)90090-6

Kubiak, S., Kim, W.J., Fedock, G., & Bybee, D. (2012). Assessing short-term outcomes of an intervention for women convicted of violent crimes. *Journal of the Society of Social Work and Research*, 3, 197-212.

Kubiak, S., Fedock, G., Bybee, D., & Kim, W.J (in press). Long-term outcomes of a RCT Intervention Study for Women with Violent Crime. *Journal of the Society of Social Work and Research*.

Kubiak, S., Fedock, G., Tillander, E., Kim, W.J., & Bybee, D. (2014). Assessing the feasibility and fidelity of an intervention for women with violent offenses. *Evaluation and Program Planning*, 42, 1-10.

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartman, L. V. Hedges, R. J. Light, … F. Mosteller. (Eds.). *Meta-analysis for explanation: A casebook* (pp. 83–127). New York, NY: Russell Sage Foundation.

McClellan, D. S., Farabee, D., & Crouch, B. M. (1997). Early victimization, drug use, and criminality: A comparison of male and female prisoners. *Criminal Justice and Behavior, 24*, 455–476. http://dx.doi.org/10.1177/0093854897024004004

Messina, N. (2014). Beyond violence: *Final report.* Sacramento: California Department of Corrections and Rehabilitation. Retrieved from http://www.stephaniecovington.com/assets/files/BV-FINAL-Report-for-CDCR_FOPS-lw-edits.pdf

Messina, N., Braithwaite, J., Calhoun, & Kubiak, S. (2016). Examination of a Violence Prevention Program for Female Offenders. *Violence and Gender*, 3(3): 143-149.

Messina, N., Grella, C. E., Cartier, J., & Torres, S. (2010). A randomized experimental study of gender-responsive substance abuse treatment for women in prison. *Journal of Substance Abuse Treatment, 38*, 97–107. http://dx.doi.org/10.1016/j.jsat.2009.09.004

Najavits, L. M., Gallop, R. J., & Weiss, R. D. (2006). Seeking safety therapy for adolescent girls with PTSD and substance abuse disorder: A randomized controlled trial. *The Journal of Behavioral Health Services & Research, 33*, 453–463. http://dx.doi.org/10.1007/s11414-006-9034-2

Najavits, L. M., Weiss, R. D., Shaw, S. R., & Muenz, L. R. (1998). "Seeking safety": Outcomes of a new cognitive-behavioral psychotherapy for women with posttramatic stress disorder and substance dependence. *Journal of Traumatic Stress, 11*, 437–456. http://dx.doi.org/10.1023/a:1024496427434

Olds, D. L., Robinson, J., Pettitt, L., Luckey, D. W., Holmberg, J., Ng, R. K., … Henderson, C. R. (2004). Effects of home visits by paraprofessionals and by nurses: Age 4 follow-up results of a randomized trial. *Pediatrics, 114*, 1560–1568. http://dx.doi.org/10.1542/peds.2004-0961

Owen, B. (1998). In the mix: *Struggle and survival in women's prisons*. Albany, NY: SUNY Press.

Richie, B. (1996). *Compelled to crime: The gendered entrapment of battered black women.* New York, NY: Routledge.

Salisbury, E. J., & Van Voorhis, P. (2009). Gendered pathways: An empirical investigation of women probationers' path to incarceration. *Criminal Justice and Behavior, 36*, 541–566. http://dx.doi.org/10.1177/0093854809334076

Silverman, D. (2010). *Qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Smith, P., Cullen, F. T., & Latessa, E. J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy, 8*, 183–208. http://dx.doi.org/10.1111/j.1745-9133.2009.00551.x

University of Cincinnati Corrections Institute. (2006). *Correctional Programs Checklist.* Cincinnati, OH: University of Cincinnati Criminal Justice Research Center.

Van Voorhis, P., Bauman, A., & Brushett, R. (2012). Revalidation of the Women's Risk Needs Assessment, Prerelease Results, Final Report. Cincinnati OH: University of Cincinnati Criminal Justice Research Center.

Van Voorhis, P., Bauman, A., & Brushett, R. (2013a). Revalidation of the Women's Risk Needs Assessment, Probation Results, Final Report. Cincinnati OH: University of Cincinnati Criminal Justice Research Center.

Van Voorhis, P., Bauman, A., & Brushett, R. (2013b). Revalidation of the Women's Risk Needs Assessment, Prison Results, Final Report. Cincinnati OH: University of Cincinnati Criminal Justice Research Center.

Van Voorhis, P. (2015). *Gender-Responsive Community Programs Inventory.* Cincinnati, OH: University of Cincinnati Criminal Justice Research Center.

Van Voorhis, P., & Salisbury, E. J. (2014). *Correctional counseling and rehabilitation* (8th ed.). Waltham, MA: Anderson.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to existing risk/needs assessment: The current status of a gender-responsive supplement. *Criminal Justice and Behavior, 37,* 261–288. http://dx.doi.org/10.1177/0093854809357442

Additional Resources

National Institute of Corrections, Women Offenders (topic page): http://nicic.gov/womenoffenders

National institute of Corrections Library: http://nicic.gov/library/

National Resource Center on Justice Involved Women, Resources: http://cjinvolvedwomen.org/resources/

Office of Justice Programs' (OJP) Crime Solutions: https://www.crimesolutions.gov/

Substance Abuse and Mental Health Services Administration's (SAMHSA) National Registry of Evidence-Based Programs and Practices (NREPP): http://nrepp.samhsa.gov/01_landing.aspx

Van Voorhis, P. (2012. Volmer Award Address: On Behalf of Women Offenders: Women's Place in the Science of Evidence-Based Practice. Criminology and Public Policy, 11, 111-145.

Washington State Institute for Public Policy: http://www.wsipp.wa.gov/

Women's Risk and Need Assessment: http://www.uc.edu/womenoffenders.html