Introductory Handbook for

DWI Court Program Evaluations

A publication by
The National Center
for DWI Courts

NCDC
National Center
for DWI Courts

Making Your Community A Safer Place

Introductory Handbook for DWI Court Program Evaluations

by Douglas B. Marlowe, J.D., Ph.D.
Chief of Science, Policy & Law, NADCP

About NADCP

The National Association of Drug Court Professionals (NADCP), a not-for-profit organization located in the Nation's Capitol, was founded in 1994 by a group of judicial visionaries to reverse the growing impact of drug-related crime. They created a court model using a combination of accountability and treatment to compel and support drug using offenders to change their lives. From that vision came the Drug Court movement and the NADCP. In 1997, NADCP and the White House Office of National Drug Control Policy (ONDCP) then partnered to create the National Drug Court Institute (NDCI), the professional services branch NADCP. Today, NADCP is the premier national membership, training and advocacy organization for Drug Courts, representing over 22,000 multi-disciplinary Drug Court professionals.

NADCP hosts the largest annual training conference on drugs and crime in the nation and provides 130 Drug Court training and technical assistance events, benefiting tens of thousands of drug court professionals each year. NADCP/NDCI continues to write, publish, and disseminate scholastic and practical publications that are critical to the ongoing growth and fidelity of the Drug Court model and works tirelessly on Capitol Hill and in state legislatures to transform the American justice system through policy, legislation and appropriations.

The National Center for DWI Courts

The National Center for DWI Courts is a professional services division of NADCP designed to:

Raise awareness of the success of DWI Courts;
Provide training, technical assistance, and research to DWI Courts; and,
Establish new DWI Courts nationwide.

Established in June of 2007 through a partnership with Beam Global Spirits and Wine, Inc, this division works toward eliminating hardcore impaired driving through the expansion of DWI Court programs nationwide.

It is through that expansion of DWI Courts that NCDC's message of "Making Your Community a Safer Place" is coming true.

NCDC is supported by the National Highway Traffic Safety Administration, U.S. Department of Transportation; Beam Global Spirits & Wine, Inc; The Century Council; and The Wine & Spirits Wholesalers of America, Inc.

Printed: August 2010 in Alexandria, Virginia

The National Center for DWI Courts

Executive Director
C. West Huddleston

Director
David J. Wallace

Training Coordinator
Cliff Jacobs

Training Coordinator
Elisa Fulton

NCDC
National Center for DWI Courts

About the author, Dr. Doug Marlowe:

Douglas B. Marlowe, J.D., Ph.D. is the Chief of Science, Policy & Law for the National Association of Drug Court Professionals, a Senior Scientist at the Treatment Research Institute, and an Adjunct Associate Professor of Psychiatry at the University of Pennsylvania School of Medicine. A lawyer and clinical psychologist, Dr. Marlowe has received numerous state and federal research grants to study coercion in drug abuse treatment, the effects of drug courts and other diversion programs for drug abusers involved in the criminal justice system, and behavioral treatments for drug abusers and criminal offenders. He is a Fellow of the American Psychological Association (APA) and has received proficiency certification in the treatment of psychoactive substance use disorders from the APA College of Professional Psychology. Dr. Marlowe has published over 125 professional articles and chapters on the topics of crime and substance abuse. He is the Editor-in-Chief of the Drug Court Review and is on the editorial board of Criminal Justice & Behavior.

Support for this handbook was provided by the National Highway Traffic Safety Administration (NHTSA). The views expressed are those of the author and do not necessarily reflect the position of NHTSA or its agents or employees.

NHTSA
www.nhtsa.gov

Table of Contents

Foreword 6

Part I Introduction 8

1 Why to Measure 8

Part II When to Measure 11

1 Real-Time Recording 11
2 Date Stamping 11
3 Shake-Out Year 12
4 Ongoing Evaluations 13
5 Starting the Clock 13
6 Follow-Up Windows 14

Part III Whom to Measure 17

1 Intent-to-Treat Analyses 17
2 Comparison Groups 17
Random Assignment 18
Quasi-Experimental Comparison Groups 18
Matched Comparison Groups 19
Invalid Comparison Groups 19
3 Cohorts 20

Part IV What to Measure 22

1 Logic Model for DWI Courts 22
2 Moderator Variables 23
3 Mediator Variables 25
The Problem of the Missing Denominator 25
Performance Indicators for Supervision Services 25
Performance Indicators for Treatment Services 26
Restrictive Conditions 27
4 Proximal Outcomes 28
Missing or Adulterated Specimens 28
Performance Indicators for Proximal Outcomes 29
5 Distal Outcomes 30

Part V Where to Measure 32

1 Recommended MIS Features 32

Part VI How to Measure 38

1 Target Population 38
2 Disparate Minority Impacts 39

3 Accounting for Baseline Differences 39
Propensity Score Analyses 40
4 Days At-Risk 41
5 Service Provision 42
6 Infrequent Events 42
7 Selecting Statistical Consultants 43

Part VII GLOSSARY 46

Part VIII References 55

Index 58

Foreword

Drug Courts have been in existence for 20 years, and during that time have proven their effectiveness. Three GAO studies and five meta-analyses have shown Drug Courts to be the most effective method of changing the behavior of a drug user and reducing recidivism long-term. Nationwide, 75% of Drug Court graduates remain arrest-free at least two years after leaving the program. Compare this to the typical re-arrest rates on those in a traditional court, in which 46% of probationers commit a new offense and over 60% commit a probation violation. The studies demonstrate that Drug Courts significantly reduce crime and save money for taxpayers by offsetting the costs of law enforcement, court case processing and victimization resulting from future criminal activity.

However, it took time for the evidence to be developed. Questions were raised regarding the initial studies. Critics still use old, out-dated studies to challenge the effectiveness of Drug Courts. Without the scientific documentation, it was easier to challenge the anecdotal stories of success, and argue that Drug Courts were not successful.

DWI Courts are based on the Drug Court model, and it is expected that DWI Courts will also face challenges. To respond to those challenges a DWI Court must evaluate its program. Guiding Principle #9 of the 10 Guiding Principles as set out by the National Center for DWI Courts discusses the importance of evaluating a DWI Court program. In part it says: “A credible evaluation is the only mechanism for mapping the road to program success or failure.” However, over the past few years, several evaluations done of DWI Courts have had serious methodological shortcomings.

A systematic review of 41 published and unpublished DWI Court program evaluations found that many of them had serious methodological shortcomings, including reporting outcomes only for graduates, failing to account for participant dropout, employing inadequate statistical techniques, and evaluating potentially immature programs. The review noted that while there are promising indications that DWI Courts work, there needs to be more studies – effective and scientifically valid studies – done of DWI Courts. [Drug Court Review (Volume VI, Issue 2)]

We must make every effort to ensure that the studies done on DWI Courts are thorough and valid, with supporting evidence. This handbook is designed to provide courts with the knowledge to understand what is needed to do a proper study. With a thorough evaluation, DWI Court programs will be able to demonstrate that for the hardcore DWI offender, DWI Courts, like Drug Courts, are an effective tool in the criminal justice system.

This handbook would not be possible without the hard work of Dr. Doug Marlowe, and I would like to express my appreciation for all that he has done to put this handbook together. In addition, the National Highway Traffic Safety Administration (NHTSA) should be acknowledged and thanked for their support of this project. Without the support of NHTSA on this handbook, and for DWI Courts in general, we would not be as far along as we are in changing the hardcore DWI offender’s behavior and saving lives on our nation’s roads.

Building a Safer Community,

David Wallace
Director
The National Center for DWI Courts

Introduction

Part I

1 Introduction

Driving While Impaired (DWI) Courts are a relatively new criminal justice program designed to improve DWI offenders' compliance with substance abuse treatment and other conditions of community supervision and refrain from further DWI behavior (e.g., Freeman-Wilson & Huddleston, 1999). Modeled after Drug Courts, DWI Courts require participants to attend on-going status hearings in court, complete an intensive regimen of substance abuse treatment and other indicated services, and undergo random or continuous biological testing for alcohol or drug use (NADCP, 2005). Participants receive sanctions for violations of program conditions and rewards for positive achievements that steadily increase in magnitude over successive infractions or accomplishments.

The large majority of DWI Courts are post-adjudication programs serving repeat DWI offenders or those with relatively high blood alcohol concentrations (BACs) at the time of their arrest. Many of these programs require participants to serve some portion of a jail sentence with the remainder of detention suspended pending completion of treatment and community supervision. Failure to graduate successfully from the DWI Court often results in a return to custody to complete the full sentence. As of the end of 2008, there were 144 dedicated DWI Courts and an additional 382 hybrid DWI/Drug Courts in the U.S (NDCI, 2009).

The purpose of this introductory handbook is to assist practitioners in DWI Courts to evaluate the effectiveness of their programs and use the research data to monitor and improve upon their operations. Staff members in these programs are busy professionals and face numerous day-to-day challenges to meet the demands of their clients, the criminal justice system, and the public at-large. As a result, they may not have substantial time or financial resources to commit to studying their program's operations or their clients' outcomes. Moreover, many of these professionals have not received formal training in research methods and may lack the requisite knowledge to perform scientifically defensible program evaluations.

Fortunately, there are relatively simple and moderate-cost steps that DWI Courts can take to collect the information they need to document their services and measure their outcomes. For more sophisticated types of analyses, local evaluators should be consulted to statistically analyze the data. However, collecting the proper information in the appropriate format will greatly reduce the time and expense that is required for evaluators to retrieve the data they need to perform the appropriate analyses. This will substantially reduce the cost of evaluations and provide program staff with the answers they need in real time rather than requiring them to wait months for evaluation results to become available.

1.1 Why to Measure

In times of decreasing fiscal resources and increasing demands for accountability, DWI Courts, like other programs, will need to be able to document their

effectiveness in order to survive and flourish. Knowing that one has an effective program is one thing but being able to prove it to a skeptical audience is quite another. Lacking definitive proof of success, even excellent programs may risk losing funding as well as political and public support. Failing to allocate reasonable resources and effort towards evaluation can therefore lead to greater financial losses in the long run.

The major challenge facing DWI Courts is to gather the appropriate data in the most efficient and cost-effective manner. The following chapters of this handbook are designed to help DWI Court professionals understand the when, whom, what, where and how of data collection. By learning how to properly measure what is transpiring within their own programs, DWI Court professionals can provide local evaluators with the information they need to do their jobs quickly and efficiently, and get needed findings out to their stakeholders and to the DWI Court field within a reasonable time frame.

For terminology that may be unfamiliar to non-researchers, a glossary is provided at the back of this handbook which defines the terms and offers some basic rules-of-thumb for addressing issues commonly confronted in program evaluations. When viewing an electronic version of this document, simply click on a word or term that is depicted as a **link** and your cursor will be brought to that term in the glossary.

When to Measure

Part II

2 When to Measure

2.1 Real-Time Recording

The best time to record information about services or events is when they occur. This is sometimes referred to as **real-time recording**. The typical staff member in a DWI Court is responsible for dozens of participants and each participant has multiple obligations in the program, including appearing at court hearings, attending treatment sessions, and complying with alcohol or drug testing. It would be the rare professional who could accurately recall what events transpired or should have transpired days or weeks in the past. Attempting to reconstruct those events from memory is fraught with peril and is likely to introduce unacceptable error into the program evaluation.

The general rule-of-thumb is that data should be recorded within no more than approximately 7 to 10 days after the respective events have occurred. Managers in DWI Courts are strongly encouraged to work with their supervisors to ensure adherence to this recommended best-practice for DWI Court programs.

2.2 Date Stamping

Another important issue in program evaluations is **date stamping** (e.g., Heck, 2006). This refers to the practice of connecting every service or event to the date that it occurred or was supposed to have occurred. For example, staff should indicate that a counseling session was attended on January 1, 2008 or that a status hearing was missed on January 15, 2008. The timing of events is often critically important for measuring the performance of both programs and clients. For example, if a client attended 12 counseling sessions during the course of 1 month, this might reflect an adequate **dosage of services**. On the other hand, if he or she attended 12 sessions over 6 months, that would reflect a relatively low dosage of services. By connecting the counseling sessions to the dates on which they occurred, it is possible to measure the **density of services** that were provided; that is, the amount of services that were provided per unit of time, such as per month or per phase of the program.

Date stamping is also necessary for measuring the time delay between important events. For example, research indicates that the faster sanctions or rewards are applied, the more effective they are (e.g., Marlowe, 2007; Marlowe & Kirby, 1999). Therefore, the average time delay between the occurrence of infractions and the imposition of sanctions, or between the occurrence of achievements and the imposition of rewards, is an important indicator of how well a program is responding to participants' conduct. Date stamping is required for an evaluator to be able to measure this critical performance indicator.

Unfortunately, date stamping can be unwieldy if evaluators use traditional forms of **statistical spreadsheets**. Most spreadsheets are two dimensional, meaning they

are comprised of rows and columns. If client names are listed by rows and services are listed by columns, how does one account for the dates on which the services were delivered? It is usually necessary to create separate columns for each service on each date, resulting in many hundreds of columns.

As will be discussed in greater detail later, newer generations of data-entry systems automatically date-stamp entries. For example, data-entry screens might appear just like a professional's own appointment calendar. Information is simply entered on the appropriate day in the calendar and is automatically date-stamped for analysis. DWI Courts are strongly encouraged to use data-entry systems that automatically date-stamp all entries.

2.3 Shake-Out Year

Generally speaking, data collected during the first year of operations should be used to inform programmatic modifications, and should ordinarily be included in a **process evaluation** as opposed to an outcome evaluation (e.g., Heck, 2006; Rempel, 2007). A process evaluation indicates, among other things, whether a program is meeting the basic performance benchmarks that are recommended for the field. For example, it is important to know whether a DWI Court is in compliance with the Ten Guiding Principles of DWI Courts (NADCP, 2005) or the Ten Key Components of Drug Courts (NADCP, 1997). In contrast, an **outcome evaluation** examines client-level impacts on such variables as recidivism or substance use. Because it typically takes several months to identify problems and make corrections, DWI Courts should be given ample time to pilot-test their operations and implement any needed modifications before formal outcome evaluations are conducted. By the end of the first year, which is often referred to as the **shake-out year**, a program could be expected to have fine-tuned its procedures sufficiently to ensure compliance with the basic standards for the field, and it would then be appropriate to move forward with a formal outcome evaluation.

Unfortunately, this has not been the practice in many DWI Court evaluations. Often, DWI Courts have been required to conduct outcome evaluations during their initial implementation year to satisfy sponsors or funding agencies. The result is that many DWI Court evaluations report outcomes during the program's shake-out year when it is just beginning to develop and improve its procedures. This can have the unfortunate effect of underestimating the potential impacts of DWI Courts once they have hit their stride.

A general rule-of-thumb is to reserve formal outcome evaluations for the second year of operations and repeatedly thereafter. This is not to suggest that client-level outcomes are unimportant during the first year, but only that they should ordinarily be used to improve the program rather than to suggest how well the program is likely to function in the future.

2.4 Ongoing Evaluations

It is not justifiable to perform an evaluation of a program at a single point in time and to assume that the results will be representative of what can be expected from that program in the future. Conditions are likely to change appreciably in a given jurisdiction over time and those changes can be expected to influence the effectiveness of DWI Courts. For one thing, a DWI Court might experience a substantial increase or decrease in funding, which could greatly improve upon or detract from the quality of the services it can offer. In addition, staff turnover is common in both the substance abuse treatment system and the criminal justice system, and new personnel may improve or impede the success of a program. Similarly, veteran staff members may enhance their skills through experience and continuing education or may become "stale" and less effective in their roles.

The client population for a program could also change over time. A DWI Court might, for example, begin to treat more serious recidivist DWI offenders as a result of changes in charging practices by the prosecution. Similarly, substance abuse patterns might change in a community. For example, the introduction of **methamphetamine** could require a DWI Court to treat a more severely addicted and impaired population.

It is essential, therefore, for program evaluations to be repeatedly conducted over time to detect changes in performance. Unfortunately, this can be cost-prohibitive if a program is required to periodically contract with independent investigators to perform the evaluations. A preferable course of action is to continuously collect and analyze performance information in an ongoing manner. Fortunately, there is a limited number of basic statistical analyses that are typically required for purposes of routine program evaluations. For example, it is often most essential to measure such variables as a program's graduation rate, **recidivism** rate, attendance rate in counseling sessions, and proportion of alcohol and drug-negative tests. As will be discussed later, with a suitably automated **management information system** (MIS), it should be possible to conduct these types of basic analyses at the "push of a button" using pre-programmed **computer syntax**. Selecting the right MIS can save a program considerable time and money by avoiding the need to hire external investigators to perform basic performance evaluations in an ongoing manner.

2.5 Starting the Clock

For most outcome analyses, the **starting clock** should be begun at the time participants entered the DWI Court. For example, outcomes for each participant might be measured over a period of 12 months starting from the date of entry into the program. One reason for this is that it gives all participants an equivalent **follow-up window**. Each participant would have 12 months in which to engage in relevant outcome behaviors, such as committing new DWI offenses, finding a job, or attaining sobriety.

Imagine, instead, what would happen if outcomes were measured over 12 months

from the date of graduation or termination. Some participants might have graduated in 12 months whereas others might have graduated in 18 months. As a result, outcomes for the former participants would be measured for 24 months after entry and outcomes for the latter participants would be measured for 30 months after entry. The latter individuals would have 6 additional months in which to commit new offenses or engage in other outcome behaviors. In other words, they would have more opportunities for those behaviors. This is referred to as having more **days at-risk**. If various participants have different days at-risk, outcome comparisons may be difficult to interpret.

The issue of days at-risk is particularly problematic if outcomes are to be contrasted against those of a **comparison group**, such as probationers. The period of probation supervision might be substantially shorter or longer than the length of enrollment in DWI Court. For example, probationers might be under probation supervision for an average of 24 months whereas DWI Court participants might be enrolled for an average of 18 months. Therefore, measuring outcomes from the point of completion or termination would lead to major differences in their days at-risk since arrest. This could seriously confound the results and make the conclusions invalid.

The problem of days at-risk is not solved altogether by starting the clock from the point of entry. As was noted, probationers might have been under supervision for a longer or shorter period of time than DWI Court participants. If so, they would have different numbers of days-under-supervision, which is similar to the issue of days at-risk because it restricts their opportunities to re-offend or abuse substances. As will be discussed later, it may be necessary to statistically adjust or control for differences in days-under-supervision in order to obtain valid and defensible results.

2.6 Follow-Up Windows

There is no one correct **follow-up window** for evaluating outcomes. The most appropriate follow-up period will depend upon what specific outcome is being assessed. For example, it may be very informative to measure counseling attendance during the first several weeks or months of a program. It is important to know whether a DWI Court is successful at getting participants engaged in treatment and providing an adequate **dosage of services**. However, because many DWI Courts steadily decrease the dosage of treatment as participants move through the phases of the program, treatment attendance may become a less important outcome after several months. Over time, it may become more important to know whether substance use has declined or whether engagement in pro-social activities, such as employment, has increased. **Date stamping** permits the evaluator to analyze different types of outcomes at different follow-up intervals.

Generally speaking, it is not very informative to measure re-arrest or new conviction rates during the first several months of a program. It typically takes substantial time for an offender to re-engage in DWI or other criminal behavior, be detected by the authorities for that behavior, and be formally charged or convicted in a legal

proceeding. As a result, re-arrest and conviction rates are apt to be low during the first several months of treatment (excluding technical violations, which may be more common). This does not necessarily reflect well on the program but rather may reflect limited opportunities to re-offend.

This issue can be particularly problematic if re-arrests or convictions are to be contrasted against that of a **comparison group**, such as probationers. If recidivism is low in both groups, it will be very difficult to detect **statistically significant differences**. This is true even if the DWI Court is, in fact, a superior program. For mathematical reasons, there needs to be sufficient **variance** in the data—that is, larger numbers of recidivism events overall—before statistically significant differences can be detected. For example, if 10% of the probationers recidivated during the first 6 months and 5% of the DWI Court clients recidivated, this difference would probably not be statistically significant unless there was a large number of participants in the study. However, it might be meaningful from a clinical or public policy perspective. The best course of action is to wait 12 months before reporting on recidivism outcomes. This should allow sufficient time to elapse for recidivism rates to diverge meaningfully between the groups.

Where feasible, recidivism should be tracked for at least 3 years post-entry, and ideally up to 5 years post-entry. Research suggests that the majority of recidivism events for drug and alcohol offenders occur during the first 3 to 5 years (e.g., Martin et al., 1999). Therefore, following participants for 3 to 5 years should ensure that most recidivism events are accounted for in the evaluation. Of course, recidivism analyses will still be informative after only the first or second year post-entry, and those interim analyses should be reported as the data become available. However, following participants for 3 to 5 years is likely to elicit the most stable and reliable long-term estimates of recidivism.

Whom to Measure

Part III

3 Whom to Measure

3.1 Intent-to-Treat Analyses

It is essential to analyze outcomes for *all* individuals who participated in the DWI Court, regardless of whether they successfully graduated or were unsuccessfully terminated from the program. This is called an **intent-to-treat analysis** because it examines outcomes on all individuals whom the program set out to treat. It is not appropriate only to report outcomes for graduates because this unfairly inflates the apparent success of the program. Individuals who graduated from the DWI Court might, for example, have had less severe drug or alcohol problems to begin with, higher motivation for change, or better social support systems than the average DWI offender. As a result, they might have been less likely to commit future DWI offenses regardless of the services they received in the DWI Court. The most important question is how the program fared for all participants.

This is particularly important when outcomes are contrasted against those of a **comparison group**, such as probationers. Selecting out the most successful DWI Court cases and comparing their outcomes to all of the probationers would be unfair. It would be akin to selecting out the A+ students from one classroom, comparing their academic success to all of the students in another classroom, and then concluding that the first class had a better teacher. This would clearly be a biased and unfair comparison.

This is not to suggest that outcomes for graduates are of no interest. Programs may, indeed, want to know what happens to individuals who received all of the services that are offered by the DWI Court. However, this should be a **secondary analysis** that is conducted after the intent-to-treat analyses have been completed. If it is first determined that the program achieved positive outcomes on an intent-to-treat basis, it may then be appropriate to go further and determine whether outcomes were even better for the graduates. However, if the intent-to-treat outcomes are not impressive, it is generally not acceptable to move on to evaluate outcomes for the graduates alone.

Importantly, if secondary analyses are performed on graduates, then the comparison sample should also be limited to successful completers. For example, DWI Court graduates should be compared to probationers who satisfied the terms of probation. Comparing DWI Court graduates to all probationers, including probation failures, would unfairly stack the deck in favor of the DWI Court.

3.2 Comparison Groups

There are several ways to select a suitable **comparison group** for a DWI Court program evaluation. Some comparison groups are far superior to others and some may be so unfairly biased that the results of the evaluation will be seriously suspect.

3.2.1 Random Assignment

By far the best strategy is to randomly assign eligible DWI offenders to the DWI Court or to an alternative disposition, such as probation. **Random assignment** provides the greatest assurance that the groups started out with an equal chance of success. Therefore, if there are better outcomes for the DWI Court group, these can be more confidently attributed to the effects of the DWI Court as opposed to other extraneous factors, such as differences in the severity or prognoses of the groups when they first entered the programs.

In many instances, however, random assignment may not be feasible. Some DWI Courts may have difficulty maintaining a full census and will not want to turn away eligible individuals. Moreover, some staff members or stakeholders may have ethical objections against denying services to otherwise eligible individuals. In fact, random assignment has been used in several research studies involving DWI Courts (Breckenridge et al., 2000; MacDonald et al., 2007) and Drug Courts (Festinger et al., 2002; Gottfredson et al., 2003; Marlowe et al., 2007; Marlowe et al., 2008) and is generally not considered to be unethical if appropriate safeguards are instituted. For example, participants might need to give their informed consent to be randomly assigned and an **Institutional Review Board (IRB)** or **Data & Safety Monitoring Board (DSMB)** might need to oversee the study. Regardless, it still might not "feel right" to some staff members or may be politically inexpedient to engage in random assignment.

3.2.2 Quasi-Experimental Comparison Groups

The next best approach after random assignment is to use a **quasi-experimental comparison group**. This involves comparing outcomes to individuals who did not enter the DWI Court *for reasons that are unlikely to have affected outcomes*. An excellent example of such a comparison group would be:

- DWI offenders who were eligible for the program but were denied access because there were no empty slots available.

The mere happenstance of a full census is unlikely to lead to the systematic exclusion of individuals with more severe problems or poorer prognoses, and therefore is unlikely to bias the results. Less optimal, but still potentially acceptable, quasi-experimental comparison groups include the following:

- DWI offenders who would have been eligible for the DWI Court but were arrested in the year or so before the DWI Court was established.
- DWI offenders who would have been eligible for the DWI Court but were arrested in an adjacent jurisdiction that does not have a DWI Court.

Because such individuals were arrested at an earlier point in time or in a different geographic region, there still might be systematic differences between the groups that could bias the outcomes. For example, socioeconomic conditions or police

practices might have changed between the year before and the year after the establishment of the DWI Court, or might differ between neighboring counties. However, the likelihood of this occurring is usually not substantial and these may be the only practical and reasonable comparison groups that can be used for purposes of many program evaluations.

Importantly, when using a quasi-experimental comparison group, it is essential to check for possible pre-existing differences between the groups that could have affected the results. For example, if outcomes from a DWI Court are compared against those of DWI probationers from a neighboring county, it could turn out that the comparison probationers had a more serious criminal history. This, in turn, might have put them at greater risk for DWI recidivism. If so, superior outcomes for the DWI Court might not have been due to the effects of the program but rather to the fact that it treated a less severe population. This could confound the results and make the comparison invalid.

It is necessary to statistically control or adjust for this difference in prior criminal history and for any other relevant differences that might have existed between the two groups before they entered the programs. As will be discussed in greater detail later, there are a number of statistical procedures that one can employ to take account of such differences and still obtain defensible results.

3.2.3 Matched Comparison Groups

Another approach to creating an acceptable non-randomized comparison sample would be to use a **matched comparison group**. This involves systematically selecting individuals from out of a larger sample of potential candidates who are similar to the DWI Court participants on characteristics that would be expected to affect outcomes. For example, an evaluator might pair DWI Court participants with DWI probationers who are equivalent in terms of their criminal histories, demographic characteristics, and/or substance abuse problems. Because the evaluator will be choosing only those comparison individuals who are similar to the DWI Court participants, it is often necessary to start out with a relatively large sample of potential candidates from which to select the most comparable individuals.

Importantly, the success of any matching strategy will depend largely on whether the evaluator has adequate information about the comparison candidates to make valid matches. If data are not available on such important variables as their criminal histories or substance abuse problems, it will not be possible to place confidence in the validity of the matches. It is not sufficient to simply match the groups on variables that are easily measurable and available, such as gender or race, because the samples might still differ on other important characteristics that were never taken into account.

3.2.4 Invalid Comparison Groups

There are a number of comparison groups that have been used in some DWI Court

program evaluations that are quite likely to have biased the results. As a general rule, it is ***not*** appropriate to compare outcomes from a DWI Court to those DWI offenders who:

- Refused to enter the DWI Court;
- Were denied access to the DWI Court due to their clinical or criminal histories;
- Dropped out of the DWI Court; or
- Were terminated from the DWI Court.

There is a serious risk that such individuals are apt to have had relatively poorer prognoses to begin with, due to such negative characteristics as lower motivation for change, lesser social supports, or more serious substance abuse histories. Therefore, comparing them to the DWI Court sample could unfairly stack the deck in favor of the DWI Court. Given the high likelihood that these groups will have been seriously disadvantaged on important factors, statistical adjustments usually cannot be confidently relied upon to overcome the biased differences.

3.3 Cohorts

As was discussed earlier, it is not justifiable to perform an evaluation at a single point in time and to assume that the results will be representative of what can be expected from that program in the future. It is necessary to perform evaluations on a continuous basis to assess possible changes in program performance or client outcomes over time.

The National Center for DWI Courts (NCDC) recommends evaluating **cohorts** of participants. A cohort is defined here as a group of individuals who entered the program during the same general time period, usually over an interval of approximately 6 or 12 months (Heck, 2006; Rubio et al., 2008). For example, all participants who entered the DWI Court between January 1st and December 31st of a given year might be defined as a cohort. Outcome analyses could then be conducted separately for cohorts defined as having entered the program between 1/1/06 and l2/31/06, between 1/1/07 and 12/31/07, and so forth. This would provide a series of “snapshots” indicating how well the program performed during calendar year 2006, calendar year 2007, and so on.

What to Measure

Part IV

4 What to Measure

4.1 Logic Model for DWI Courts

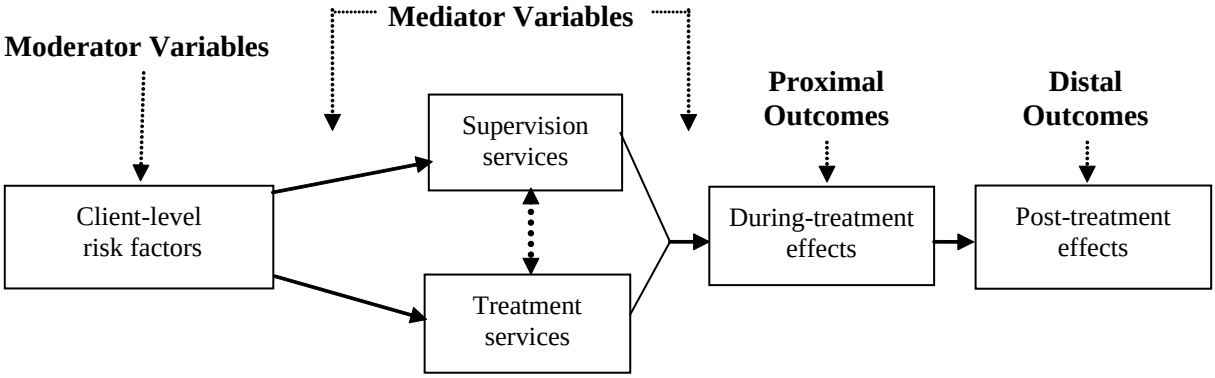
In 2005, The National Drug Court Institute (NDCI) convened a nationally recognized expert panel of researchers and evaluators to develop a logic model and performance indicators for DWI Court program evaluations. A **logic model** is a depiction of how a program is believed to exert its effects. DWI Courts are hypothesized to improve outcomes for DWI offenders by combining mandatory substance abuse treatment with a strict program of behavioral monitoring and accountability. The essential components of a DWI Court include (NADCP, 2005):

- continuous judicial supervision via regularly scheduled status hearings in court;
- mandatory completion of substance abuse treatment and other indicated services;
- continuous or random biological testing for alcohol and other drug ingestion;
- imposition of a progressively escalating sequence of punitive sanctions for infractions and positive incentives for achievements;
- satisfaction of applicable legal restrictions and obligations, such as installation of ignition interlock devices, sales of relevant vehicles or payment of fines and fees.

Performance indicators are quantifiable measures of each component of the logic model. Program-level performance indicators reveal what services a program is actually providing and client-level performance indicators reveal how well participants in the program are faring. Examples of program-level performance indicators might include how often status hearings are held or how often participants receive substance abuse treatment services. Examples of client-level performance indicators might include how often participants test negative for alcohol and other drugs or graduate successfully from the program.

Some logic models may be quite elaborate and may hypothesize specific mechanisms of action. For example, a researcher might hypothesize that DWI Courts exert their effects, first, by enhancing participants' self-esteem or motivation for change. This, in turn, might lead participants to commit themselves more fully to substance abuse treatment, thus ultimately bringing about long-term reductions in DWI conduct. Unfortunately, it is difficult to test such complicated logic models during the course of routine program evaluations because they require evaluators to conduct frequent and repeated assessments of participants using such measures as structured interviews or self-report tests.

A **skeletal logic model** depicts the minimum components of a program without which it would not be expected to achieve its effects. For example, if a program did not provide substance abuse treatment, it would not be a DWI Court regardless of what it called itself because it lacks a minimally required component of a DWI Court. Similarly, if a program did not require ongoing court appearances, it would not be a DWI Court regardless of its name. The following Figure depicts a skeletal logic model for DWI Court programs:



4.2 Moderator Variables

Client-level **moderator variables**, also known as **risk factors**, are included in the logic model because DWI Courts are not necessarily intended for use with the entire population of DWI offenders. Roughly two thirds of first-time DWI arrestees do not go on to commit a further DWI offense (e.g., Cornish & Marlowe, 2003). Therefore, providing all DWI offenders with an intensive court-managed intervention might be unlikely to improve outcomes for the population as a whole. For statistical reasons, it is difficult to detect improvements on an outcome such as DWI recidivism if the probability of that outcome is relatively limited in the population to begin with.

In this light, an individual's risk level is seen to moderate—that is, influence or interact with—the effects of the intervention. By including relevant moderator variables in the statistical model, an evaluator can determine which participants were helped by the DWI Court and which ones were not. This helps to avoid a possibly wrongful conclusion that the DWI Court "did not work" when the real issue might have been that the wrong **target population** was treated in the first instance.

According to the criminological theory of the **Risk Principle**, intensive interventions such as DWI Courts are theorized to exert the greatest effects for **high-risk** individuals who are characterized by relatively more severe antisocial propensities or treatment-refractory histories (e.g., Taxman & Marlowe, 2006). In contrast, intensive interventions may be unnecessary for **low-risk** individuals who may be apt to improve their behavior almost regardless of what interventions are provided to them (e.g., DeMatteo et al., 2006). Importantly, in this context the term "risk" does not

relate to a risk for violence or dangerousness. Rather, it refers to a poor prognosis in standard treatment or supervision programs unless more intensive services are provided. High-risk DWI offenders are less likely to perform adequately without the additional structure and discipline afforded by a DWI Court. In contrast, low-risk DWI offenders may be apt to perform equivalently in less intensive interventions, such as standard treatment or probation.

Several moderator variables have been consistently identified by researchers in numerous studies involving DWI offenders (Beerman et al. 1988; C'de Baca et al., 2001; Donovan et al., 1990; Fell, 1994; Lapham et al., 1997; Lapham et al., 2000; Marowitz, 1998; Peck et al., 1994; Perrine et al., 1989; Schell et al., 2006; Tashima & Marelich, 1989; Wells-Parker et al., 1985; Yu & Williford, 1993). These moderator variables should ideally be measured in any DWI Court program evaluation:

- Current age
- Gender
- Education (number of years completed)
  - Do not count a graduate equivalency degree (GED) as 12 years but note separately that a GED was obtained
- Marital status
- Number of prior DWI convictions
- Number of prior moving traffic violations
- Number of other prior criminal convictions
- Blood alcohol content (BAC) at arrest
- **DSM-IV** diagnosis of **substance abuse**
- **DSM-IV** diagnosis of **substance dependence**
- **DSM-IV** diagnosis of other major **Axis I psychiatric disorder** (yes or no)
- Number of prior substance abuse treatment attempts
  - Do not include self-help groups but note separately that they were attended
- Age of onset of substance abuse
- Age of onset of delinquent or criminal activity
- Chronic unemployment or unstable employment (yes or no)
- Unstable living arrangements (yes or no)
- Proportion of time spent interacting with other substance abusers or

offenders

- o May be measured on a **Likert scale** (e.g., ranging from “none” to “most”)

4.3 Mediator Variables

DWI Courts provide close and continuous supervision of DWI offenders combined with evidence-based treatment services. Supervision and treatment services are called **mediator variables** because the effects of DWI Courts are believed to be mediated by—that is, directly caused by—these services. Unlike moderator variables (discussed above), which indicate *whom* the intervention works for, mediator variables indicate *how* the intervention works. Mediator variables are also sometimes referred to as **performance indicators** because they reveal how well a program is functioning and what types of services are actually being delivered (as opposed to what services were *intended* to be delivered).

4.3.1 The Problem of the Missing Denominator

When measuring mediator variables or performance indicators, it is essential to record events that should have transpired but did not. For example, if a DWI offender attended 12 status hearings in court, this would reflect perfect attendance if 12 hearings had originally been scheduled. If, however, he or she was expected to have attended 24 hearings, then this 50% attendance rate might be considered unacceptable. Unfortunately, the denominator (i.e., the number of hearings that were originally scheduled) is often missing from client records, which makes the results of an evaluation difficult to interpret. This is sometimes referred to as the problem of the **missing denominator**. It is essential for program staff to continuously record information about appointments or other events that were scheduled to occur and to indicate whether each appointment was:

- Kept
- Not kept
- Excused, or
- Rescheduled.

4.3.2 Performance Indicators for Supervision Services

DWI Courts closely monitor participants and administer consequences—both rewarding and punitive—contingent upon their behaviors. This close level of supervision is believed to elicit superior outcomes through traditional principles of behavioral change known as **operant conditioning** or **contingency management**. Put simply, the rapid and certain detection of infractions and achievements, coupled with progressively escalating rewards or sanctions, has been demonstrated to

improve outcomes for both substance abusers and criminal offenders (e.g., Marlowe & Wong, 2008). The efficacy of any DWI Court will depend, in large part, on how well it applies these scientifically proven principles of behavioral change.

The basic components of supervision utilized in DWI Courts include status hearings in court, probation contacts, biological testing for alcohol and drug use, and graduated sanctions and incentives. Recommended **performance indicators** for evaluating these supervision services include the following:

Status Hearings

- Total number of status hearings attended
- Percentage of scheduled status hearings attended

Probation Contacts

- Total number of probation contacts
    - categorized by type of contact (e.g., in-office, home visit, bar check)
- Percentage of scheduled contacts attended

Biological Testing

- Length of time on continuous alcohol monitoring (e.g., **SCRAM** or sleep tether) (start & end dates for each episode)
- Total number of breathalyzer, urine or saliva tests administered
- Percentage of scheduled breathalyzer, urine or saliva tests administered

Sanctions & Incentives

- Total number of sanctions imposed
    - categorized by magnitude (e.g., verbal reprimand = low magnitude, increased urine testing = moderate magnitude, jail detention = high magnitude)
- Total number of incentives provided
    - Categorized by magnitude (e.g., verbal praise = low magnitude, reduced status hearings = moderate magnitude, graduation = high magnitude)

4.3.3 Performance Indicators for Treatment Services

The provision of substance abuse treatment and relevant **adjunctive services** is also considered to be integral to DWI Courts. The basic assumption underlying DWI

Courts is that addiction and related psychosocial impairments contribute substantially to recidivist DWI conduct; therefore, it is believed to be essential to address those clinical symptoms in order to achieve sustained behavioral change. Recommended **performance indicators** for evaluating the provision of treatment services include the following:

Substance Abuse Treatment

- Time-delay between entry into the DWI Court and the first treatment session (days or weeks)
- Total number of sessions attended
- Percentage of scheduled sessions attended

Modality of Treatment

- Length of time in each modality of treatment (e.g., detox, residential, outpatient) (start & end dates for each modality)

Medication

- Length of time prescribed medication for the treatment of addiction or substance abuse (start & end dates for each prescription)
    - Categorized by type of medication (e.g., **disulfiram** or **naltrexone**)

Adjunctive Services

- Total number of sessions attended
    - Categorized by type of service (e.g., psychiatric, educational, vocational)
    - Include only those participants with an *identified need* for the service
- Percentage of scheduled sessions attended
    - Categorized by type of service (e.g., psychiatric, educational, vocational)

4.3.4 Restrictive Conditions

It is also important to account for the imposition of **restrictive conditions**, such as home detention, residential treatment, or jail detention. The primary reason for this is that restrictive conditions can reduce participants' **days at-risk** in the community. For example, if a participant is detained in residential treatment or jail, he or she is likely to have substantially fewer opportunities to use alcohol or drugs or engage in DWI behavior. Therefore, it may be necessary to statistically account for the length of time that participants were in restrictive settings. This may be measured as follows:

- Length of time on restrictive community confinement (e.g., home detention) (start & end dates for each episode)
- Length of time in residential treatment or recovery housing (start & end dates for each episode)
- Length of time in jail detention (start & end dates for each episode)

4.4 Proximal Outcomes

It is typically easiest for evaluators to measure outcomes during participants' enrollment in the DWI Court. Although some might argue that the most important effects of DWI Courts are those occurring after participants are no longer under the supervision of the criminal justice system, there is substantial evidence that better during-treatment outcomes do predict better post-treatment outcomes (e.g., Rempel et al., 2003). Therefore, much is to be gained by examining **proximal outcomes** (or short term outcomes) while participants are still enrolled in the DWI Court.

4.4.1 Missing or Adulterated Specimens

It is first necessary to say a few words about missing or **adulterated biological specimens**. It is usually not appropriate to treat missing or adulterated specimens as missing data. It is reasonable to assume that participants will be less likely to provide a valid specimen if they anticipate that it will be positive for drugs or alcohol, and thus likely to elicit a punitive sanction. Therefore, the failure to provide a valid specimen is not a random or ignorable event, but rather may reflect a systematic effort at concealment of substance abuse. Treating missed specimens as missing data could be apt to systematically underestimate substance use and overestimate the beneficial effects of the program.

In fact, DWI Courts often view the failure to provide a valid specimen as a separate infraction from substance use, and one that may be seen as a more serious violation than providing a positive specimen (e.g., Marlowe & Wong, 2008). A participant who fails to provide a valid specimen may be assumed both to have engaged in substance use and also to have attempted to conceal that use or failed to behave responsibly by owning up to the use. Thus, the participant might receive two separate sanctions or a more severe sanction for making a bad situation worse (Marlowe, 2008).

The generally recommended course of action is to assume missing or adulterated specimens to be substance-positive and possibly to treat them as a second type of infraction as well. Of course, if a participant is *excused* by staff from providing a specimen for an acceptable reason (e.g., because of illness or employment obligations), then it would certainly be appropriate to treat it as missing data.

There are also advanced statistical **imputation procedures** that can sometimes be

employed to compensate for missing data. Some imputation procedures take into account the pattern of test results immediately before and after a missing specimen. For example, if a participant had several negative specimens immediately before and after a missed specimen, the missing specimen might be assumed to be negative. Alternatively, some imputation procedures assume the average or most prevalent result for the population to be the likely outcome of a missing specimen. Selecting an appropriate imputation procedure is complicated and expert statistical consultation is generally required to apply such procedures correctly in a given case.

4.4.2 Performance Indicators for Proximal Outcomes

According to the NDCI (Heck, 2006), the minimum set of **performance indicators** for evaluating **proximal outcomes** in Drug Courts should cover the following general outcome dimensions: (1) retention or completion status, (2) sobriety or substance use, and (3) recidivism. These outcome dimensions are also suitable for DWI Courts with minor modifications. Recommended performance indicators for evaluating proximal outcomes in DWI Courts are as follows:

Retention

- Graduation status (graduated, terminated, absconded, or other)
- Length of stay (date of entry to date of last physical contact)

Sobriety

- Longest interval of sustained sobriety while on continuous alcohol monitoring (e.g., **SCRAM** or sleep-tether) (start & end date)
- Total number of biological screens that were substance-negative
    - Count unexcused failures to provide a specimen as positive
- Percentage of scheduled biological screens that were substance-negative
    - Count unexcused failures to provide a specimen as positive
- Largest consecutive number of biological screens that were substance-negative
    - Count unexcused failures to provide a specimen as positive

Recidivism

- Number of new arrests for:
    - DWI
    - moving traffic violations

- other criminal charges

- Date of first arrest for the above

Payment of Fines & Fees

- Satisfaction of fines and fees (complete, partial, excused, or other)

4.5 Distal Outcomes

Ideally, **distal outcomes** (or long-term outcomes) would be assessed by re-contacting participants, interviewing them about their functioning, and performing biological testing. Unfortunately, this is seldom feasible because participants tend to be difficult to track down after they have left the program, are often reluctant to return for a follow-up appointment, and may be reticent to acknowledge information that could get them into trouble with the law. As a result, evaluators are often required to concentrate on examining criminal justice and motor vehicle databases for information pertaining to new arrests and convictions, traffic violations, and licensing actions. Recommended **performance indicators** for these distal outcomes are as follows:

Performance Indicators for Distal Outcomes

Recidivism

- Number of new arrests for:
  - DWI
  - moving traffic violations
  - other criminal charges
- Date of first arrest for the above
- Number of new convictions for:
  - DWI
  - moving traffic violations
  - other criminal charges
- Date of first conviction for the above
- Drivers license reinstatement (yes or no)

Where to Measure

Part V

5 Where to Measure

DWI Court professionals will undoubtedly find the above lists of **risk factors**, **performance indicators**, **proximal outcomes** and **distal outcomes** quite daunting, especially when issues of **date stamping** and the problem of the **missing denominator** are considered. The unfortunate reality is that few programs collect these data in a reliable manner. And, when programs do capture the appropriate data elements, they rarely systematically account for missed appointments, connect events to the dates they occurred or were supposed to have occurred, or enter the data in a format that permits immediate statistical analyses. In many instances, evaluators are required to extract data from written records or unwieldy **statistical spreadsheets** with little recourse for reconciling inconsistencies or accounting for missing entries. The unfortunate result is that evaluations are often completed months or years after the fact—when they may no longer reflect what is going on in the program—and there may be so many holes or caveats in the data that the conclusions that can be drawn are tentative at best.

The secret to valid, timely, and cost-efficient program evaluations lies in the selection of a suitable **management information system (MIS)**. If an MIS is not simple to use and does not provide immediate and easily understandable basic reports, then perhaps it should be replaced. The up-front costs of instituting a useful MIS will be offset many times over by providing greater efficiencies in staff operations and yielding the type of performance feedback that is necessary to continuously improve and fine-tune one's program.

5.1 Recommended MIS Features

Newer generations of **MIS** systems are capable of automating program evaluations and streamlining the burden on staff members and participants. Where feasible, it is recommended that DWI Courts purchase MIS systems with the following characteristics:

- ✓ **Web-Enabled.** Staff members in DWI Courts are frequently employed by different agencies and may have offices in several locations. They may also be required to visit participants at home, in jail, or in a residential treatment facility to conduct assessments or deliver services. This requires staff to have access to the MIS on the road and from multiple locations. If the system is web-enabled, it can be accessed from any location that has internet access and a web browser, including through the use of personal palm devices or laptop computers.

- ✓ **Security Protected.** The data should be stored and transferred using industry-standard 128 bit SSL encryption or better.

- ✓ **Regulated Access.** DWI Courts are comprised of multidisciplinary teams and all staff members may require access to at least some data elements.

It is necessary for various users to be able to log-on to the system; however, authorized levels of access should be carefully regulated. Only under relatively rare circumstances should one staff member be capable of altering data that were entered by another staff member. For example, the judge should not ordinarily be able to alter information entered by a treatment provider. The judge might, however, have read-only access to certain information from the treatment program, such as participants' counseling attendance and completion of treatment plan goals. The authorized level of access for each staff member should be specified by an MIS Administrator and correspond to that staff member's password and username.

- ✓ **Less-is-More Data Entry.** Staff members should only be required to view those data-entry screens that are relevant to their jobs. For example, a treatment provider ordinarily should not be faced with data-entry screens relating to probation contacts or status hearings. The treatment provider might be authorized to view summary reports relating to probation contacts or status hearings, but should not be required to scroll through that material if it is not on-point to the task at hand.

- ✓ **Need-to-Know Data Entry.** DWI Court professionals have a right to know why they are being asked to collect information and should be permitted to avoid duplicated effort. If there is no immediately obvious and empirically defensible reason why particular information is being collected, then perhaps it is unnecessary to collect that information. Redundancies should also be eliminated. For example, once a participant's age is entered into one data-entry screen, it should automatically populate the respective fields in other data-entry screens. Of course, if there is an error in data entry or the information changes, it should be possible to override incorrect entries.

- ✓ **Quick Data Entry.** It should generally take no more than about 2 minutes to enter all of the required weekly data elements for a given client.

- ✓ **Intuitive Data Entry.** Most professionals are accustomed to using the Internet for such things as paying bills, purchasing goods and services, or gathering information. A great deal of effort has gone into developing commercial websites that are intuitive and simple to use. There is no reason why MIS systems for DWI Courts should be organized differently. If a system does not have similar features and does not follow a similar intuitive design as many commercial websites, then perhaps it's time to upgrade.

- ✓ **Automatic Date Stamping.** Some of the better data-entry screens appear just like a professional's own appointment calendar. Information is entered on the appropriate day in the calendar and is automatically **date stamped** for analysis.

- ✓ **Ticklers for Missing Data.** Modern MIS systems routinely prod or "tickle" users for missing or incomplete data. For example, if a probation officer enters attendance information for the week of 10/15, the system should alert the probation officer that data have not yet been entered for the previous week of 10/8. Ticklers make it less likely that missing data will become so stale that it is difficult to reliably reconstruct the events from memory or handwritten records.

- ✓ **Flexible Input Screens.** It is often necessary to add new items, delete or "gray out" unwanted items, or change the wording of items to meet new challenges or answer new questions. For example, a DWI Court might contract with a new treatment program to provide mental health services. This might require a new data-entry screen to be added with items relating to attendance at mental health counseling sessions. If adding new items requires re-programming of the system, perhaps it is time to upgrade. It should be possible to add new items, delete items and change the wording of items in no more than a few hours or a few days.

- ✓ **Behind-the-Scenes Database.** DWI Court staff members should not be required to wade through columns of numeric data. However, behind the scenes, data should be stored in an analyzable format that permits immediate statistical testing. It should not be necessary to reenter data into a **statistical spreadsheet**. The data should be stored in specified fields that can be readily selected and entered into statistical analyses.

- ✓ **Longitudinal Database.** Under no circumstance should the system overwrite previous data. For example, if a participant is unemployed when he or she enters the DWI Court, and then obtains a job a few months later, the participant's employment status should not simply be overwritten from unemployed to employed. Doing so would make it very difficult to determine at a later date whether the participant had obtained a job while he or she was in the DWI Court or had originally entered the program with a job. Moreover, it would be difficult to determine how long it took the participant to find the job. Instead, the database should be arranged longitudinally, meaning that new events should be appended alongside older events. If the database is arranged longitudinally and the entries are date stamped, it will be possible to determine whether participants' status changed over time and how long it took for those changes to occur.

- ✓ **Automated Descriptive Reports.** There is almost no limit to the number of research questions that DWI Court staff members and their stakeholders might ask. However, there are a limited number of basic **descriptive analyses** that most programs will want to conduct. For example, virtually every program will want to know its graduation rate, recidivism rate, attendance rate in counseling sessions, and proportion of alcohol or drug-positive tests. In addition, many programs will want to know how these outcomes might differ, if at all, between males and

females, between various racial or ethnic groups, or as a function of participants' **risk factors**. It is possible to write **computer syntax** in advance for these common types of analyses so that the analyses can be performed virtually at the push of a button. For more sophisticated research questions, it may still be necessary to transfer the data to a statistician for analyses; however, for many run-of-the mill questions it should be possible to generate reports nearly instantly.

- ✓ **Graphic and Tabular Capabilities.** In addition to generating statistical results, the reports should include easy-to-interpret graphs, tables, pie charts and the like. Following the adage that "a picture is worth a thousand words," it should be possible to generate informative graphics at the push of a button.

- ✓ **Continuous Performance Feedback.** The most important reason for evaluating a program is to improve its operations and enhance participants' outcomes. If an intervention is not working or a participant requires a change to his or her treatment plan, it is essential to find this information out quickly while there is still time to make an adjustment. This requires the **MIS** to continuously monitor the data and issue automated alerts to staff whenever a particular course of action might be called for. For example, the system might automatically alert staff whenever a participant misses two counseling sessions in a row or provides two drug-positive urine specimens. This would prompt the staff to discuss the participant at the next case review and decide upon a suitable response. Research indicates that generating automated feedback in this manner can substantially improve the efficiency of staff operations and enhance client outcomes (Marlowe et al., 2008).

- ✓ **User Accountability Reports.** One of the biggest threats to valid program evaluations lies in the failure of some staff members to enter their data in a timely manner. The **MIS** should be capable of generating user accountability reports, which indicate how long it takes for data to be entered after the relevant events have transpired. For example, management staff should be made aware that a counselor has been entering attendance information an average of 4 weeks after the sessions have occurred, or that alcohol-screen results are being listed as "pending" by the lab for an average of more than 2 weeks. This would enable management to intervene not only to improve the quality of the evaluation, but more importantly to improve the quality of the program itself.

It can never be stressed enough that the key to a successful program evaluation lies in the selection of a proper **MIS**. If the system is easy to use, captures the essential performance indicators, and stores the data in an analyzable format, then the battle is virtually won. And the cost of purchasing such an MIS would be offset many times over by the fact that there is no longer a need for external researchers to spend hundreds of hours attempting (often with limited success) to extract information from

written charts, spreadsheets, or disparate databases that cannot be easily merged. If the appropriate data can be handed off to researchers in the proper format, useful findings should be obtainable in a reasonable period of time and at a reasonable cost to the program.

How to Measure

Part VI

6 How to Measure

It is beyond the scope of this introductory handbook to discuss the specifics of data analyses and statistical techniques. As was noted earlier, most **MIS** systems should be capable of generating basic **descriptive analyses** using pre-programmed **computer syntax**. For more sophisticated analyses, it will typically be necessary to obtain expert statistical consultation. Although it might seem straightforward how to analyze data, there are numerous logical traps and pitfalls that can lead the unwary evaluator to the wrong conclusions. It is best to rely on experienced evaluators to perform the proper statistical analyses and correctly interpret the results.

Assuming that the data have been properly collected and entered, the costs associated with obtaining expert statistical consultation should not be prohibitive. If researchers are handed the relevant data in the proper format, it should generally take no more than a few days or weeks to complete the analyses and summarize the findings. The reason that it usually takes much longer for researchers to complete their work is that the data are often full of holes and inconsistencies or are missing altogether. This requires researchers to spend months trying (and not always succeeding) to track down information, reconcile errors, and fill in the gaps.

This chapter will focus on common analytical errors that have been made in some DWI Court program evaluations, and which have laid waste to hundreds of hours of work and tens of thousands of dollars in consulting fees. Guarding against these common errors can avoid substantial distress stemming from indefensible evaluation results.

6.1 Target Population

It is common practice to correlate client-level characteristics—such as age, race or number of prior DWI convictions—with outcomes in DWI Courts. Determining which types of offenders perform best in the program can help to identify the optimal **target population** for the DWI Court. It can also help to determine which clients will require more enhanced services in order to succeed.

Importantly, however, the mere fact that a client-level characteristic predicts better success in a DWI Court does not necessarily mean that it represents the best target population for that program. Many variables predict better outcomes for DWI offenders as a whole but do not necessarily reveal who is best suited to a particular intervention.

Assume, for example, that outcomes are found to be better for older DWI Court participants and for those with fewer prior DWI convictions. This could lead a DWI Court to conclude, perhaps wrongly, that it should focus its efforts on treating older first-time DWI offenders. However, a younger age and prior offense history are common risk factors for poorer outcomes in most types of criminal justice programs (e.g., Marlowe et al., 2003). Regardless of whether such individuals are sentenced

to pre-trial diversion, probation, DWI Court or jail, they tend to do worse than older offenders and those with less severe criminal histories.

The important question is whether individuals with these high-risk characteristics perform significantly better in DWI Court as opposed to an alternative disposition, such as probation. This requires the evaluator to conduct **interaction analyses**. Interaction analyses indicate not only which program had better outcomes (e.g., DWI Court vs. probation) and which types of DWI offenders had better outcomes (e.g., younger vs. older offenders); more importantly, they indicate which types of offenders performed better in which program. It might turn out, for example, that younger participants performed better in DWI Court whereas older participants performed better on probation or equally well in either program. Thus, an interaction analysis suggests the best way to match DWI offenders to the most effective programs and indicates how a jurisdiction can marshal its resources most efficiently and cost-effectively.

6.2 Disparate Minority Impacts

It is important to determine whether certain gender, racial, ethnic or cultural sub-groups might be performing unusually poorly, or unusually successfully, in a DWI Court program. If, for example, racial minorities were found to be failing out of a program at a significantly higher rate than other participants, this could raise serious concerns about whether the program is providing culturally competent and sensitive services. It could also raise equal protection or due process objections.

It is not sufficient, however, merely to determine whether outcomes differ between gender, racial, ethnic or cultural sub-groups. Other variables might be correlated with participants' gender, racial, ethnic or cultural identity, and those variables might be the ones that are truly responsible for any differences in outcomes. For example, in some communities Caucasians might be more likely than other racial groups to abuse **methamphetamine**. Because this particular drug is so highly addictive and potentially destructive, individuals who are abusing it could be expected to have relatively worse outcomes. Therefore, poorer outcomes might be explained not by race per se, but rather by the fact that race is correlated with the drug of choice. The drug of choice might be the real culprit in explaining poorer outcomes.

This requires evaluators to use slightly more advanced statistical procedures that first take into account potentially competing explanations, and then determine whether the variable of interest still remains correlated with outcomes. For example, an evaluator might determine whether race is still associated with outcomes after the influence of the drug of choice has first been taken into account. Only then would it be defensible to conclude that there may be disparate racial impacts in the program.

6.3 Accounting for Baseline Differences

Many evaluators will want to compare outcomes between DWI Court participants

and those of a **comparison group**, such as probationers. Prior to making such comparisons it is necessary to rule out other competing explanations that could have accounted for differences in outcomes apart from the effects of the programs. For example, the probationers might have had more severe criminal histories. If so, then superior outcomes for the participants in DWI Court might have had nothing whatsoever to do with the effects of the program. Rather, it might simply have been a result of the fact that the DWI Court treated an "easier" population to begin with. Even when an evaluator uses **random assignment**, it is still possible (although much less likely) that the groups may differ on important dimensions at baseline.

If the groups do differ at baseline, all is not necessarily lost. There are certain procedures that a statistician can follow to adjust for the differences and still obtain defensible results. This typically involves a three-step process:

1. First, the groups are compared on relevant **risk factors** that have been associated with outcomes among DWI offenders. [These risk factors were listed in the earlier chapter entitled "What to Measure".] If the groups do differ on some risk factors, it *may* become necessary to adjust for those risk factors in the outcome comparisons. Unfortunately, in many instances the program may not have measured all of the relevant risk factors. For example, the evaluator might not have access to data on whether the participants had previously received substance abuse treatment or had prior criminal convictions. The more risk factors that are measured, the more confident one can be that alternative explanations for the results were appropriately ruled out.

2. Second, the evaluator determines whether, in fact, these risk factors predicted outcomes in the current sample. For example, although a younger age ordinarily predicts poorer outcomes among DWI offenders, it might not have done so in the present study. If it does not predict poorer outcomes in this study, then the evaluator does not need to be concerned about this variable going forward.

3. Third, any variables that BOTH differ between the groups AND predict outcomes must be statistically controlled for in the outcome analyses. For example, if the groups differed by gender and gender predicted outcomes, then gender must be entered as a covariate in the analyses. How this is accomplished will depend upon the particular statistical analysis being performed. For example, the evaluator might use an analysis of covariance (ANCOVA), a hierarchical regression analysis, or another statistical approach.

6.3.1 Propensity Score Analyses

An evaluator might also use a more sophisticated statistical procedure called a **propensity score analysis** to adjust for baseline differences between the groups. This is an advanced analysis that involves mathematically calculating the probability

that an individual with a given set of baseline characteristics would be in the DWI Court group as opposed to the comparison group—in other words, the relative similarity of that individual to one group as opposed to the other. The analysis then statistically accounts for this relative similarity when comparing outcomes between the groups. Advanced statistical expertise is ordinarily required to implement and interpret this procedure.

As is the case with other statistical adjustments, the success of this procedure will depend in large measure on whether the appropriate risk factors were measured. A propensity score analysis utilizes risk information to calculate the relative probability of a subject's membership in either group. If those data are not available or were not well measured, the analysis will be fraught with error.

6.4 Days At-Risk

As was discussed earlier, a common confound that can undermine the validity of DWI Court evaluations is the issue of **days at-risk**. This refers to the proportion of time that participants were free to engage in substance abuse, DWI conduct or other behaviors. For example, while participants are detained in jail, it would be difficult for them to drive a car and it would also be difficult (albeit not impossible) to use alcohol or other drugs. It would not be very interesting to find that DWI recidivism was low if participants had been incarcerated during most of the follow-up period.

This is particularly problematic if outcomes are contrasted against a **comparison group** and the two groups had significantly different days at-risk. Assume, for example, that probationers who did not enter the DWI Court were more likely to serve a longer jail term prior to being placed on probation. As a result, they would have had lesser opportunities to be arrested for a new DWI. This could lead to the erroneous conclusion that probationers did better than participants in DWI Court, when the real issue was that they were simply kept off the streets longer.

If the groups did differ by days at-risk, all is not necessarily lost. An approach similar to the one discussed above can sometimes be used to compensate for this issue. The number of days that participants were in jail or in another controlled environment would first be correlated with outcomes. If this variable was significantly correlated with outcomes, then it would be entered as a covariate in subsequent outcome analyses. This would reveal whether the DWI Court participants had better outcomes after the influence of days at-risk was statistically taken into account.

As was noted earlier, participants might also differ in terms of **days-under-supervision**. For example, probationers might be under probation supervision for an average of 24 months whereas DWI Court participants might be supervised for an average of 18 months. It may also be necessary to statistically control for this difference in the length of supervision in the outcome analyses.

6.5 Service Provision

DWI Courts will often want to know whether specific services within their programs are helping or hindering outcomes. For example, it may be important to know whether administering punitive sanctions or referring clients to a new counseling group improves outcomes or perhaps makes outcomes worse.

Importantly, such analyses should not be conducted post hoc, meaning after the fact. Assume, for example, that an evaluator correlates the number of sanctions that participants received with their ultimate outcomes in the program. It is determined from this analysis that participants who received more sanctions were less likely to graduate successfully from the program. This could lead to the erroneous conclusion that sanctions made outcomes worse. But such a conclusion might have confused cause with effect. It is likely that poor performance led the staff to apply more sanctions. If so, then more sanctions did not make outcomes worse, but rather poor outcomes led to more sanctions.

Similarly, participants who have more severe problems or who are not responding to standard services might be required to attend more counseling sessions or might be referred to a more intensive modality of care, such as residential treatment. Simply correlating treatment services with outcomes could lead to the erroneous conclusion that more treatment led to poorer outcomes. In fact, poorer performance often leads to a referral for more treatment.

The proper approach is to conduct these analyses a priori; that is, ideally participants should be randomly assigned in advance to different types of interventions or different dosages of services. For example, some participants might be required to attend 50 hours of counseling whereas others might be required to attend 10 hours of counseling. Then, differences in outcomes can be reasonably attributed to the **dosage of services** and not merely to biased referrals to receive those services.

6.6 Infrequent Events

Some events occur relatively infrequently in DWI Court programs. For example, if outcomes are examined over the first 6 to 12 months, re-arrests for new DWI offenses might not be expected to occur at very high rates. It usually takes some time for participants to re-engage in DWI conduct after an arrest, be detected by law enforcement for that conduct, and have criminal charges brought against them. Even those participants who do quickly return to DWI behavior would not be expected to have more than about one or two new arrests during the first 6 to 12 months.

A common mistake in program evaluations is to use statistical tests that examine the average number of infrequent events, such as the average number of DWI arrests. Because the number of re-arrests is likely to be zero for many participants, the average number of arrests across the entire sample might be less than about one or two. This creates what is called a skewed distribution. For mathematical reasons, it

is unlikely that the analysis will detect **statistically significant differences** when the distribution is skewed in this manner. That is, even if there is some meaningful difference between the groups, it is unlikely to be detected by the statistical test.

For infrequent events, it is recommended to transform the data into a binary or dichotomous variable (e.g., any new arrest vs. no new arrest) and to use what are called nonparametric statistical tests. For example, rather than analyzing the average number of new arrests, the evaluator might analyze the percentage of participants who had at least one new arrest. Nonparametric tests can help to avoid the problems associated with infrequent events and can lead to more sensitive and powerful findings.

6.7 Selecting Statistical Consultants

Apart from relatively simple **descriptive analyses** and routine comparisons (e.g., males vs. females), it is generally best to rely on experienced researchers to perform the appropriate statistical tests and interpret the findings. If the data have been properly recorded and stored in the right format, it should not take long for a statistician to analyze the results. The big question, however, is how DWI Court staff members—who have typically been trained as lawyers, clinicians, or criminal justice professionals—should be expected to be able to identify competent statistical consultants.

The first step in selecting the right team of consultants is to review their prior evaluation reports, especially any evaluations related to problem-solving court programs. Familiarity with how problem-solving courts operate is often essential for knowing what **risk factors**, **performance indicators**, **proximal outcomes** and **distal outcomes** should be examined. It is also important for knowing how to interpret the results and describe the findings in a practical and understandable manner to court and treatment professionals.

When reviewing prior evaluation reports, it is useful to consider the questions listed below. These are not presented in a particular order of importance and some questions might not be relevant for a particular evaluation. For example, if the evaluators did not contrast outcomes against those of a **comparison group**, then there would be no need to control for baseline differences between groups. However, under such circumstances the evaluators would be limited in terms of the conclusions they could draw from the study. Lacking a comparison sample, it would not be possible to conclude how well the DWI Court performed in relationship to other programs. If the evaluators went beyond the data in making such an impermissible interpretation, that might bode poorly for their competence as researchers.

- ✓ Did the evaluation team measure the relevant risk factors for DWI recidivism and examine outcomes according to participants' risk level?
- ✓ Did the evaluation team analyze important and relevant performance

indicators, proximal outcomes and distal outcomes?

- ✓ Did they properly account for missing data and adulterated specimens?
- ✓ Were the analyses conducted on an intent-to-treat basis?
- ✓ Did they use a reasonable follow-up window given the outcome variable being measured? For example, did they examine recidivism over a long enough time period for re-arrests or new convictions to be expected to occur?
- ✓ Did they select an unbiased comparison group?
- ✓ Did they check for baseline differences between groups at entry and statistically control for them where necessary?
- ✓ Did they start the clock running at an equivalent time for all participants? For example, did they start it running at the point of entry into either DWI Court or an alternative disposition such as probation?
- ✓ Did they statistically control for days at-risk in the community and days-under-supervision?
- ✓ Did they identify the best target population for the DWI Court by examining interaction effects rather than merely by examining predictors of outcomes?
- ✓ Did they rule out other possible explanations before identifying disparate impacts by gender, race, ethnic or cultural groups?
- ✓ Did they transform the data or use nonparametric tests for infrequent events?
- ✓ Were the interpretations of the findings defensible given the limitations of the research design?

The most important question of all is whether the evaluators' interpretations of the findings pointed to concrete actions that the program could take in the future to improve its performance and enhance clients' outcomes. The primary goal of a program evaluation is to improve outcomes and continuously move forward. If an evaluation report simply states whether or not a program worked in the past but does not indicate how it should work in the future, it will be of limited utility.

GLOSSARY

Part VII

7 GLOSSARY

Adjunctive services: Services that address problems or issues not directly related to substance abuse or addiction, although they may influence or be influenced by substance abuse or addiction. Common examples of adjunctive services in DWI Courts might include mental health services, vocational services, or family counseling. Not all participants in a DWI Court may require all of these services and the services may be reserved for participants with a demonstrated need.

Adulterated biological specimens: Bodily specimens such as urine or saliva that show evidence of having been tampered with or of being fraudulent or unreliable. Most labs routinely evaluate specimens for evidence of adulteration by examining such indicators as ph, creatinine, and specific gravity. The temperature of a specimen may also be tested at the time of delivery to be certain it is close to average body temperature and thus was not likely to have been delivered at an earlier point in time. Many DWI Courts treat adulterated specimens as substance-positive or require that a new specimen be delivered. Evaluators should ordinarily treat adulterated specimens as substance-positive or as not having been delivered as directed (i.e., as an unexcused failure to provide a scheduled sample).

Axis I psychiatric disorder: A serious or major psychiatric disorder that is recorded on a separate primary diagnostic "axis" according to the DSM-IV. Examples include but are not limited to major depression, bipolar disorder (also known as manic depression), schizophrenia, dementia, anxiety disorders, and post-traumatic stress disorder (PTSD). Substance abuse and substance dependence are also Axis I disorders, although they are generally discussed separately from other psychiatric disorders in this handbook.

Cohort: A group of individuals who entered the program during the same specified time period, often defined as over an interval of 6 or 12 months. For example, all participants who entered a DWI Court between January 1st and December 31st of a given year might be defined as a cohort.

Comparison group: A sample of individuals who did not participate in the program or who participated in a different program, against whom outcomes are contrasted. In DWI Court evaluations, the comparison group is often comprised of DWI offenders sentenced to probation or adjudication as usual. The comparison group should be as equivalent as possible to the DWI Court group on variables that could be expected to affect outcomes, such as DWI charges, criminal backgrounds, the severity of their substance abuse problems, and other risk factors. Without a comparison group, it cannot be determined whether outcomes were affected by the DWI Court or might have occurred anyway even if the participants had not entered the DWI Court.

Computer syntax: Pre-programmed statistical equations instructing how data are to be analyzed. For example, a computer system might have a pre-programmed statistical formula for calculating the average age of participants in a program.

Contingency management: A body of behavior therapy techniques demonstrated to improve outcomes when applied properly in substance abuse treatment programs and other settings, such as schools or mental health facilities. Much of the research on contingency management has focused on the use of rewards to positively reinforce clients' accomplishments, although it may also include sanctions to punish undesirable behaviors. DWI Courts use such strategies when they apply graduated rewards and sanctions for participants' achievements and infractions in the program.

Data & Safety Monitoring Board (DSMB): A multidisciplinary group of professionals with expertise in research methods, research ethics and statistical analyses who are responsible for overseeing the integrity of data collection and analyses during most federally sponsored randomized studies. Among other duties, the DSMB may perform or oversee interim analyses to determine whether any participants in the study are being disadvantaged or suffering adverse events related to the interventions.

Date stamping: The practice of connecting services or events to the dates on which they occurred or were supposed to have occurred. For example, indicating that a counseling session was attended on January 1, 2008 is a form of date-stamping. Some computerized data-entry systems automatically date-stamp entries.

Days at-risk: The number of days that a participant was free in the community to engage in relevant outcome behaviors, such as using alcohol or committing a new DWI offense. Days spent in a restrictive setting, such as jail or residential treatment, would ordinarily be excluded from a participant's days at-risk. Generally speaking, a longer follow-up period for a study is associated with greater days at-risk.

Days-under-supervision: The number of days that a participant was under the authority of a criminal justice agency, such as a DWI Court or probation. Because offenders tend to engage in fewer instances of substance use and crime while they are under criminal justice supervision, it is often important to examine outcomes both while they were under supervision and when they were no longer under supervision.

Density of services: The amount of treatment or other services that were delivered per unit of time, such as per month or per phase of the program. For example, attendance at 5 sessions in one month ($5 \div 1 = 5$) reflects a greater density of services than does attendance at 15 sessions in 6 months ($15 \div 6 = 2.5$) even though 15 sessions are more than 5 sessions.

Descriptive analyses: Statistical information about a group of subjects that does not seek to infer cause and effect or to attribute the findings to a larger population. For example, calculating the percentage of women in a given program or the average number of treatment sessions attended by participants in that program are descriptive analyses.

Distal outcomes: Long-term outcomes often occurring after participants are no longer enrolled in the program.

Disulfiram: A prescription medication that causes uncomfortable physical symptoms such as sweating and heart palpitations if alcohol is subsequently ingested. It is also known as Antabuse.

Dosage of services: The amount of a particular service or treatment that a participant actually received as opposed to what he or she was intended to receive. For example, if a participant attended 15 counseling sessions, the dosage of counseling sessions would be 15 sessions regardless of whether the program requires attendance at 10 sessions or at 20 sessions.

DSM-IV: Fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* published by the American Psychiatric Association. The DSM-IV contains the official diagnostic criteria for psychiatric disorders in the U.S. including substance abuse and substance dependence.

Follow-up window: The time period during which outcomes are measured. For example, if outcomes are measured for 12 months after each participant's entry into a DWI Court, then the follow-up window would be 12 months in length. In most analyses, follow-up windows should be equivalent in length for all participants.

High-risk offenders: Individuals with relatively more severe criminal backgrounds or treatment-refractory histories, who have a poorer prognosis in standard correctional rehabilitation programs. In this context the term "risk" does not relate to a risk for violence or dangerousness. Rather, it refers to a poor prognosis in standard treatment or supervisory interventions unless more intensive services are provided.

Imputation procedures: Statistical procedures that may be employed to compensate for missing data. Some imputation procedures take into account the pattern of results immediately before and after a missing result. Others assume the average or most prevalent result for the population to be the likely outcome of a missing specimen. Expert consultation is generally required to decide whether and how it is appropriate to use these procedures.

Institutional Review Board (IRB): A multidisciplinary group of professionals and community representatives with knowledge or expertise in research methods, research ethics, and the subject matter of a study. The IRB is responsible for determining whether a study is ethical and safe with regard to such matters as informed consent, confidentiality, and the risk/benefit ratio for participants. Researchers are usually required to receive approval from an independent and duly constituted IRB for most federal and state sponsored studies prior to beginning the research and must submit annual reports to the IRB during the course of the study. The IRB is empowered to stop a study or require changes to the methods as a condition of pre-approval or annual re-approval. Federal guidelines govern the basic structure, policies and procedures of most IRBs.

Intent-to-treat analysis: An analysis that includes data on all individuals who entered a program or intervention, regardless of whether or not they successfully

completed the program, dropped out, or were terminated. Intent-to-treat analyses should generally be reported as the primary analysis for most program evaluations.

Interaction analyses: Statistical procedures that examine more than one variable at a time and determine how those variables may influence outcomes alone and in combination. In DWI Court evaluations, interaction analyses might be used to determine which types of participants had better outcomes in DWI Court as opposed to a comparison condition. For example, an evaluator might determine whether high-risk participants performed better in DWI Court as opposed to probation by examining the interaction of risk level with the type of program.

Likert scale: An item format commonly used in self-report tests and structured interviews that asks the respondent to rate his or her answer on a numeric scale representing increasing or decreasing magnitude. For example, a participant might be asked to rate his or her satisfaction with a DWI Court on a scale ranging from 1 ("I dislike the program") to 5 ("I like the program very much").

Logic model: A diagram depicting how a program is believed to exert its effects. A typical logic model might indicate the types of clients the program is intended to treat, the types of services that are intended to be provided, and the types of outcomes that are predicted to be affected.

Low-risk offenders: Individuals with relatively less severe criminal and substance abuse treatment histories that have a relatively good prognosis for success in standard substance abuse treatment or correctional rehabilitation programs. Such individuals may not require intensive programs such as DWI Courts to have positive outcomes.

Management information system (MIS): An automated computer system that collects standardized data elements and may run statistical analyses on the data and yield output reports. In DWI Courts, an MIS might collect and analyze data on the services being delivered in the program and participants' performance during and after their enrollment in the program.

Matched comparison group: Individuals selected for a comparison sample by pairing them with experimental group subjects on characteristics that would be expected to affect outcomes. For example, an evaluator might match or pair DWI Court participants with DWI probationers who are equivalent in terms of their criminal histories, demographic characteristics, and/or substance abuse problems. The groups should be matched on variables that predict outcomes and not merely variables that are easy to measure, such as basic demographics.

Mediator variables: Services or interventions that are empirically demonstrated to influence outcomes. For example, if outcomes in DWI Courts depend in part on whether participants receive substance abuse treatment, then substance abuse treatment is a mediator variable. In contrast, if outcomes are the same regardless of whether or not participants receive substance abuse treatment, then substance abuse treatment is not a mediator variable.

Methamphetamine: A highly addictive stimulant drug that has been increasing in prevalence in the U.S. especially in western and rural communities. It is often prepared in homegrown "labs" and is not available by prescription and has no legitimate medical indication. It is often synthesized by altering the chemical composition and potency of legally prescribed stimulants or amphetamines.

Missing denominator: Refers to a problem commonly encountered in program evaluations in which there is a failure to faithfully record information about events that should have transpired but did not. For example, data might not be recorded on treatment sessions that were scheduled to occur but were not attended. This can complicate the interpretation of findings. It is important for staff to record information on whether appointments were kept, not kept, excused or rescheduled.

Moderator variables: Characteristics of participants that are empirically shown to influence or interact with the effects of an intervention. Moderator variables may be included in statistical analyses called moderator analyses or interaction analyses to assist in determining which types of participants were helped by a program and which ones might not have been.

Naltrexone: A prescription medication that can block opiate receptors in the brain and provide a partial blockade against the effects of alcohol. It has been empirically shown to reduce the length and intensity of relapse episodes among some alcoholics. It does not cause intoxication and is non-addictive. A long-acting form called Vivitrol® can be administered by injection and lasts approximately 30 days.

Operant conditioning: A body of scientifically demonstrated techniques for modifying behavior. The essential principles involve rewarding good behavior and punishing undesirable behavior. DWI Courts employ these principles, in part, through the application of graduated sanctions and rewards.

Outcome evaluation: A systematic study of how a program affects client-level performance, such as substance use or crime.

Performance indicators: Quantifiable measures of what services a program is providing and how participants are faring in the program. Examples of program-level performance indicators might include how often status hearings are held and how often participants attend substance abuse treatment. Examples of client-level performance indicators might include how often participants test negative for alcohol and other drugs and how often they graduate successfully from the program.

Process evaluation: A systematic study of how a program was developed, how services are organized and delivered within the program, and how staff members interact with each other and function within their respective roles.

Propensity score analysis: A statistical procedure that may be used to control for differences in baseline characteristics between groups. It involves mathematically calculating the probability that a particular individual would be in the DWI Court group as opposed to the comparison group, or the relative similarity of the individual

to one group as opposed to the other. The analysis then statistically accounts for this relative probability when comparing outcomes between groups. Advanced statistical expertise is often required to implement and interpret the results from this procedure.

Proximal outcomes: Short term outcomes usually occurring while participants are still enrolled in the program. Examples might include counseling attendance rates or graduation rates.

Quasi-experimental comparison group: A comparison sample of individuals who did not enter the program for reasons that are unlikely to have affected outcomes. An example might be DWI offenders who were eligible for the DWI Court but could not get in because there were no available slots.

Random assignment: A procedure for assigning subjects to different groups in an unsystematic and unbiased manner, such as by the flip of the coin. Random assignment provides the greatest assurances that the groups started out with an equal chance of success, and thus that any positive outcomes can be fairly attributed to the effects of the program and not to extraneous factors.

Real-time recording: The practice of recording information about events at or near the time that they occurred. Ideally, data should be recorded within no more than approximately 7 to 10 days after the respective events have occurred.

Recidivism: The incidence of new criminal activity occurring after participants entered the program. This includes new criminal activity occurring while participants are enrolled in the program and after they have graduated from or been terminated from the program. Recidivism is most commonly measured by the number or percentage of new arrests or new convictions after entry.

Restrictive conditions: Supervision or treatment requirements that make it difficult for participants to engage in substance use or other negative behaviors because they are in a residential setting or other secure environment.

Risk factors: Characteristics of offenders that have been empirically shown to predict poorer outcomes in standard correctional rehabilitation programs or treatment programs. Examples include a younger age, prior failures in treatment or rehabilitation, and a more serious criminal history or substance abuse history. Individuals with high-risk factors generally require more intensive and more structured interventions in order to succeed in treatment and refrain from recidivism.

Risk Principle: An evidence-based theory of correctional rehabilitation, which posits that intensive interventions such as DWI Courts are likely to achieve the greatest effects for high-risk offenders characterized by more severe antisocial propensities or treatment-refractory histories.

SCRAM: Secure Continuous Remote Alcohol Monitor. An anklet monitoring device that can detect alcohol in sweat vapors and transmits data wirelessly to a central

monitoring station. SCRAM devices and other continuous monitoring devices are increasingly being used in DWI Court programs.

Secondary analyses: Data analyses that examine particular subsets of participants (e.g., only women or only graduates) or that examine outcomes of secondary focus (e.g., education). Assuming that a DWI Court set out to treat all participants and to focus primarily on substance abuse and DWI recidivism, the primary analyses should generally reflect these principal aims. Secondary analyses are more likely to turn up unreliable or chance findings and thus can not be as confidently relied upon.

Shake-out year: The first year of operations for a new program, during which it is in the process of developing and trying out its procedures. It is generally best to reserve formal outcome evaluations of client-level impacts until after the shake-out year has been completed.

Skeletal logic model: A diagram depicting the minimum components of a program that are believed to be necessary for it to be effective. For example, if a program did not provide substance abuse treatment or court hearings, it would not be providing the minimum components that are considered to be necessary for a DWI Court.

Starting clock: The time point from which data collection begins. For program evaluations the clock should ordinarily be started on the date of entry.

Statistical spreadsheets: Computer files containing numerical data entered in rows and columns. Often, but not always, the column headings reflect the variables being measured (e.g., age or gender) and the row headings identify individual subjects (e.g., John Doe or Mary Roe). Newer computer systems can store data in a more easily usable format that is not so two-dimensional.

Statistically significant differences: Differences between groups that have a high mathematical probability (usually 95% or better) of being reproducible in the future. Some differences in outcomes may simply be due to chance. If an evaluator were to conduct the study again at another point in time, those differences might disappear or fail to be replicated. Statistically significant differences permit greater confidence in the reliability of one's findings.

Substance abuse: An official psychiatric diagnosis reflecting repetitive misuse of alcohol or other drugs under dangerous or inappropriate circumstances, including while driving an automobile or other vehicle.

Substance dependence: An official psychiatric diagnosis reflecting compulsive use of alcohol or other drugs. Characteristic symptoms may include cravings for the substance, uncontrolled usage, or uncomfortable withdrawal symptoms when levels of the substance decline in the bloodstream or the central nervous system.

Target population: The sub-group of participants who have been demonstrated to perform best in a particular type of program as opposed to in other programs, and thus who should be prioritized for entry into that program.

Variance: The degree to which participants have a range of different scores on a measure. For example, if all participants are between the ages of 21 and 23 years, then the variance in age is low. On the other hand, if participants have a wide distribution of ages ranging from 18 to 65 years, the variance would be high. For mathematical reasons, it is easier to detect statistically significant differences between groups if the variance on a measure is high.

References

Part VIII

8 References

Beerman, K. A., Smith, M. M., & Hall, R. L. (1988). Predictors of recidivism in DUI. *Journal of Studies on Alcohol, 49*, 443-449.

Breckenridge, J. F., Winfree, L. T., Maupin, J. R., & Clason, D. L. (2000). Drunk drivers, DWI "drug court" treatment, and recidivism: Who fails? *Justice Research & Policy, 2*, 87-105.

Cornish, J., & Marlowe, D. B. (2003). Alcohol treatment in the criminal justice system. In B. Johnson, P. Ruiz, & M. Galanter (Eds.), *Handbook of clinical alcoholism treatment* (pp. 197-207). Baltimore, MD: Lippincott, Williams & Wilkins.

C'de Baca, J., Miller, W. R., & Lapham, S. C. (2001). A multiple risk factor approach for predicting DWI recidivism. *Journal of Substance Abuse Treatment, 21*, 207-215.

DeMatteo, D. S., Marlowe, D. B., & Festinger, D. S. (2006). Secondary prevention services for clients who are low risk in drug court: A conceptual model. *Crime & Delinquency, 52*, 114-134

Donovan, D. M., Umlauf, R. L., & Salzberg, P. M. (1990). Bad drivers: Identification of a target group for alcohol-related prevention and early intervention. *Journal of Studies on Alcohol, 51*, 136-141.

Fell, J. C. (1994). Current trends in drivers with repeat convictions or arrests for driving while impaired – United States. *Morbidity and Mortality Weekly Report: MMWR, 43*(41), 759-761.

Festinger, D. S., Marlowe, D. B., Lee, P. A., Kirby, K. C., Bovasso, G., & McLellan, A. T. (2002). Status hearings in drug court: When more is less and less is more. *Drug & Alcohol Dependence, 68*, 151-157.

Freeman-Wilson, K., & Huddleston, C. W. (1999). *DWI/Drug Courts: Defining a national strategy*. Alexandria, VA: National Drug Court Institute.

Gottfredson, D. C., Najaka, S. S., & Kearley, B. (2003). Effectiveness of drug treatment courts: Evidence from a randomized trial. *Criminology & Public Policy, 2,* 171-196.

Heck, C. (2006, June). *Local drug court research: Navigating performance measures and process evaluations* [Monograph Series No. 6]. Alexandria, VA: National Drug Court Institute.

Huddleston, C. W., Marlowe, D. B., & Casebolt, R. (2008, May). *Painting the current picture: A national report card on drug courts and other problem solving court programs in the United States* [Vol. II. No. 1]. Alexandria, VA: National Drug Court Institute.

Lapham, S. C., Skipper, B. J., Hunt, W. C., & Chang, L. (2000). Do risk factors for re-arrest differ for female and male drunk-driving offenders? *Alcoholism: Clinical and Experimental Research, 24*(11), 1647-1655.

Lapham, S. C., Skipper, B. J., & Simpson, G. L. (1997). A prospective study of the use of standardized instruments in predicting recidivism among first DWI offenders. *Journal of Studies on Alcohol, 58*(5), 524-530.

MacDonald, J. M., Morral, A. R., Raymond, B., & Eibner, C. (2007). The efficacy of the Rio Hondo DUI Court: A 2-year field experiment. *Evaluation Review, 31,* 4-23.

Marlowe, D. B. (2007). Strategies for administering rewards and sanctions. In J. E. Lessenger & G. F. Roper (Eds.), *Drug courts: A new approach to treatment and rehabilitation* (pp. 317-336). New York: Springer.

Marlowe, D. B. (2008). Application of sanctions. In C. Hardin & J. N. Kushner (Eds.), *Quality improvement for drug courts: Evidence-based practices* (pp. 107-114) [Monograph Series No. 9]. Alexandria, VA: National Drug Court Institute.

Marlowe, D. B., Festinger, D. S., Arabia, P. L., Dugosh, K. L., Benasutti, K. M., Croft, J. R., & McKay, J. R. (2008). Adaptive interventions in drug court: A pilot experiment. *Criminal Justice Review, 33*, 343-360.

Marlowe, D. B., Festinger, D. S., Dugosh, K. L., Lee, P. A., & Benasutti, K. M. (2007). Adapting judicial supervision to the risk level of drug offenders: Discharge and six-month outcomes from a prospective matching study. *Drug & Alcohol Dependence, 88S*, 4-13.

Marlowe, D. B., & Kirby, K. C. (1999). Effective use of sanctions in drug courts: Lessons from behavioral research. *National Drug Court Institute Review, 2,* 1-31.

Marlowe, D. B., Patapis, N. S., & DeMatteo, D. S. (2003). Amenability to treatment of drug offenders. *Federal Probation*, *67,* 40-46.

Marlowe, D. B., & Wong, C. J. (2008). Contingency management in adult criminal drug courts (pp. 334-354). In S. T. Higgins, K. Silverman, & S. H. Heil (Eds.), *Contingency management in substance abuse treatment*. New York: Guilford Press.

Marowitz, L. A. (1998). Predicting DUI recidivism: Blood alcohol concentration and driver record factors. *Accident Analysis and Prevention, 30*(4), 545-554.

Martin, S. S., Butzin, C. A., Saum, S. A., & Inciardi, J. A. (1999). Three-year outcomes of therapeutic community treatment for drug-involved offenders in Delaware. *Prison Journal, 79,* 294-320.

National Association of Drug Court Professionals. (1997). *Defining drug courts: The key components*. Washington, DC: Office of Justice Programs, U.S. Dept. of Justice.

National Association of Drug Court Professionals, National Drug Court Institute. (2005). *Guiding principles for DWI courts*. Alexandria, VA: National Drug Court Institute.

National Drug Court Institute. (2009). Unpublished Count of DWI Courts maintained by the National Drug Court Institute. On file with NADCP.

Peck, R. C., Arstein-Kerslake, G. W., & Helander, C. J. (1994). Psychometric and biographical correlates of drunk-driving recidivism and treatment program compliance. *Journal of Studies on Alcohol, 55,* 667-678.

Perrine, M. W., Peck, R. C., Fell, J. C. (1989). Epidemiologic perspectives on drunk driving. In *U.S. Department of Health and Human Service Surgeon General's Workshop on Drunk Driving: Background Papers*, (pp.35-76). U.S. Government Printing Office, Washington, DC.

Rempel, M. (2007). Action research: Using information to improve your drug court. In G. Berman, M. Rempel & R. V. Wolf (Eds.), *Documenting results: Research on problem-solving justice* (pp. 101-122). New York, NY: Center for Court Innovation.

Rempel, M., Fox-Kralstein, D., Cissner, A., Cohen, R., Labriola, M., Farole, D., Bader, A., & Magnani, M. (2003). *The New York State adult drug court evaluation*. New York, NY: Center for Court Innovation.

Rubio, D., M., Cheesman, F., & Federspiel, W. (2008, July). *Performance measurement of drug courts: The state of the art* [State Technical Assistance Bulletin Vol. 6]. Williamsburg, VA: National Center for State Courts.

Schell, T. L., Chan, K. S. Morral, A. R. (2006). Predicting DUI recidivism: Personality, attitudinal, and behavioral risk factors. *Drug and Alcohol Dependence, 82,* 33-40.

Tashima, H. N., & Marelich, W. D. (1989). A comparison of the relative effectiveness of alternative sanctions for DUI offenders, Vol. 1, *Development of a DUI Accident and Recidivism Tracking System.* Department of Motor Vehicles, Sacramento, CA.

Taxman, F. S., & Marlowe, D. B. (Eds.) (2006). Risk, needs, responsivity: In action or inaction? [Special Issue]. *Crime & Delinquency, 52(1)*.

Wells-Parker, E., Landrum, J. W., & Cosby, P. J. (1985). *Classifying the DUI Offender: A Cluster Analysis of Arrest Histories*. Alcohol Safety Education Program, Mississippi State University, Mississippi.

Yu, J., & Williford, W. R. (1993). Problem drinking and high-risk driving: an analysis of official and self-reported drinking driving in New York State. *Addiction, 88*, 219-228.

Index

- A -

Adjunctive Services 26
Adulterated Bological Secimens 28
Axis I Psychiatric Disorder 23

- C -

Cohorts 20
Comparison Group 14, 17, 39, 41, 43
Computer Syntax 13, 32, 38
Contingency Management 25

- D -

Data & Safety Monitoring Board (DSMB) 18
Date Stamped 11, 14, 32
Days At-risk 13, 27, 41
Days-under-supervision 41
Density of Services 11
Descriptive Analyses 32, 38, 43
Distal Outcomes 30, 32, 43
Disulfiram 26
Dosage of Services 11, 14, 42
DSM-IV 23

- F -

Follow-up Window 13, 14

- H -

High-risk 23

- I -

Imputation Procedures 28
Institutional Review Board (IRB) 18
Intent-to-treat Analysis 17
Interaction Analyses 38

- L -

Likert Scale 23
Logic Model 22
Low-risk 23

- M -

Management Information System (MIS) 13, 32, 38
Matched Comparison Group 19
Mediator Variables 25
Methamphetamine 13, 39
Missing Denominator 25, 32
Moderator Variables 23

- N -

Naltrexone 26
National Center for DWI Courts 20
National Drug Court Institute 22

- O -

Operant Conditioning 25
Outcome Evaluation 12

- P -

Performance Indicators 22, 25, 26, 30, 32, 43
Process Evaluation 12
Propensity Score Analysis 40
Proximal Outcomes 28, 32, 43

- Q -

Quasi-experimental Comparison Group 18

- R -

Random Assignment 18, 39
Real-time Recording 11
Recidivism 13
Restrictive Conditions 27
Risk Factors 23, 32, 39, 43

Risk Principle 23

- S -

SCRAM 25
Secondary Analysis 17
Shake-out Year 12
Skeletal Logic Model 22
Starting Clock 13
Statistical Spreadsheets 11, 32
Statistically Significant Differences 14, 42
Substance Abuse 23
Substance Dependence 23

- T -

Target Population 23, 38
Ten Guiding Principles of DWI Courts 12
Ten Key Components of Drug Courts 12

- V -

Variance 14

What Is A DWI Court?

To date, it has been left to the traditional courts and criminal justice system to deal with DWI cases, and it has become clear that the traditional process is not working for hardcore DWI offenders. (Hardcore DWI offenders are defined as individuals who drive with a BAC of 0.15 percent or greater, or who are arrested for or convicted of driving while intoxicated after a prior driving while impaired (DWI) conviction.) Punishment, unaccompanied by treatment and accountability, is an ineffective deterrent for the hardcore offenders. The outcome for the offender is continued dependence on alcohol; for the community, continued peril. A new strategy exists to fight these hardcore impaired drivers, generally called "DWI Courts" or "DWI/Drug Courts."

A DWI Court is an accountability court dedicated to changing the behavior of the hardcore offenders arrested for DWI. The goal of DWI Court or DWI/Drug Court is to protect public safety by using the highly successful Drug Court model that uses accountability and long-term treatment to address the root cause of impaired driving: alcohol and other substance abuse.

Recognizing that treating hardcore DWI offenders is complex and requires a combination of countermeasures is just as important as understanding that the type and timing of the intervention is critical to curbing these offenders' illegal and dangerous behaviors (National Association of State Judicial Educators, 2004). This is consistent with a National Traffic Safety Board report which suggests the importance of quickly identifying and intervening with those drivers having the highest rates of alcohol-impaired driving (Quinlan et al., 2005).

With the hardcore offender as its primary target population, DWI Courts follow Defining Drug Courts: The Key Components (NADCP, 1997) and the more recent Guiding Principles of DWI Courts. Unlike Drug Courts, however, DWI Courts operate within a post-conviction model. This notion is supported in a resolution by National Mothers Against Drunk Driving (MADD) stating "MADD recommends that DUI/DWI Courts should not be used to avoid a record of conviction and/or license sanctions."

In addition to MADD, the following organizations have also passed resolutions in support of DWI Courts:

- The Governor's Highway Safety Association (GHSA);
- The Highway Safety Committee for the International Association of Chiefs of Police (IACP);
- The National Alcohol Beverage Control Association (NABCA);
- The National District Attorneys Association;
- The National Association of Prosecutor Coordinators; and,
- The National Sherriff's Association (NSA).

As of December 2009, there are 344 Hybrid DWI/Drug Courts in operation. (A Hybrid DWI/Drug Court is one that started out as a Drug Court that now also takes DWI Offenders) In addition, there are another 172 designated DWI Courts bringing the total number of specialized courts dealing with hardcore impaired drivers to 526.

For more information, please contact the National Center for DWI Courts.