

How to Design an Effective Treatment Outcome Study

Lisa M. Najavits

Harvard Medical School, Boston, MA and McLean Hospital, Belmont, MA

This paper provides a “how to” guide for designing an effective treatment outcome study. Three areas addressed are (1) a description of the three stages: treatment development, efficacy testing, and effectiveness testing; (2) suggestions for writing a grant proposal; and (3) optimal study design. The latter includes considerations of comorbidity, standard design features, treatment, assessment, patients, therapists, and replication. These are discussed with regard to the testing of psychosocial treatments, with examples from the area of pathological gambling. Finally, general questions applicable to any study are provided, including: Is there a rationale for all design decisions? Is the investigator aware of what can and cannot be inferred from the data, based on the design? Does the study design address research questions that are important and genuinely of benefit to the field?

KEY WORDS: methodology; outcome; treatment; research design.

Conducting an effective treatment outcome study is a challenge. It requires an investment of years, and is typically accomplished only via funding. It requires careful design decisions that will impact what can ultimately be known several years later. Yet it is also one of the most rewarding areas of work: it addresses the need to demonstrate that a treatment works; it moves the field forward by revealing what interventions are helpful within a given treatment, and it provides data

Please address all correspondence to Lisa M. Najavits, Proctor III, McLean Hospital, 115 Mill Street, Belmont, MA 02478; e-mail: info@seekingsafety.org.

to move beyond emotional allegiances into rational selections of treatment. Historically, the past two decades have seen a virtual explosion of outcome research as managed care has gained ascendancy. With decreased resources available for mental health and substance abuse treatment (e.g., a 50% decline in the past 11 years; Hay Group, 1999 in Bickman, 1999), there is now pressure to demonstrate that a treatment works before allocating scarce dollars toward funding it.

With the huge growth in outcome research, there has developed a substantial technology on how to conduct an effective treatment study. The field of pathological gambling, which is relatively newer to this area, can make use of this body of knowledge to avoid some of the mistakes learned in older fields of outcome research (for a historical overview, see Garfield & Bergin, 1994b; Garfield & Bergin, 1994a). As gambling continues to spread rapidly (Korn & Shaffer, 1999), the numbers of pathological gamblers, and thus need for effective treatments, increases commensurately. Thus far, few treatments have been developed and tested, and there is a clear need for innovative approaches.

In this paper, three topics will be addressed: (1) stages of treatment development; (2) suggestions for the grant proposal; and (3) optimal study design. These will be discussed in relation to the testing of psychosocial treatments, although there are strong parallels in the field of medication trials as well.

STAGES OF TREATMENT DEVELOPMENT

The first consideration in study design is understanding the stages of treatment development. Originally developed as a three-phase process under the Food and Drug Administration to evaluate medications, this model has also been applied to psychotherapy evaluation. The National Institute on Drug Abuse in the early 1990's was the first of the National Institutes of Health to initiate a Behavioral Therapies Development Program to spur the development and scientific testing of psychosocial treatments in the addictions. Since then other institutes, such as the National Institute of Mental Health and the National Institute on Alcohol Abuse and Alcoholism have created similar programs for their areas. While treatment studies were conducted prior to these initiatives, dating back to the 1960's, the funding of early-stage treatment development and advances in the technology to test treatments

has been greatly enhanced through these formal program announcements.

The three stages are summarized below (Rousanville, Carroll, & Onken, 2001):

Stage 1—Early Therapy Development. In this stage, the focus is on careful development of the treatment and basic scientific testing. A treatment is conceptualized, repeatedly refined by trying it with patients, and pilot tested. The pilot test may be a simple pre-post design, or may include a control condition. The usual products of this stage are a treatment manual, an adherence scale and training plan for therapists, relevant assessment instruments, and the results of the pilot test, which can provide such information as the effect size to use in the next stage. Stage 1 is sometimes conceptualized as two stages, 1-A and 1-B, where the former is focused solely on treatment development (e.g., developing the manual and associated materials), and the latter focused on pilot data. Stage 1 typically takes from 2–3 years.

Stage 2—Efficacy Testing. In this stage, the goal is to determine whether the treatment works under the best possible conditions; that is, with intensive training, supervision, careful selection of appropriate patients, and in-depth assessment. The study design is usually a randomized, controlled study to rigorously test the treatment, comparing it to either no-treatment, treatment-as-usual in the community, or an existing alternative treatment with known efficacy (e.g., in the field of addictions, comparing a new treatment to 12-step drug counseling). The usual products of this stage are the results of the randomized controlled trial, as well as more refined treatment materials than at stage 1 (e.g., adequate psychometric properties of the adherence scale and other treatment-specific measures, descriptive data on the patient and therapist samples, a final version of the treatment manual, and some type of “dismantling” test linking the theory and techniques of the treatment to outcomes). This stage usually takes 2–4 years.

Stage 3—Effectiveness. Also known as “generalizability” or “transferability,” the goal in this stage is to evaluate how well a

treatment performs in real-world conditions, rather than the highly controlled study of stage 2. Thus, the treatment might be implemented in a community setting with minimal training and supervision, and applied to a broad range of patients. The typical products of this stage are outcome data collected from the effectiveness study (which may involve multiple clinical sites), specification of dissemination strategies and their feasibility, and data on therapist outcomes. This stage typically takes 2–3 years.

Thus, the investigator needs to determine what stage is appropriate, select a study design relevant to that stage, and identify the types of products that will emerge from that stage. Moving through all three stages typically lasts about a decade. It may be more or less depending on whether the goal is to develop a new treatment or test an existing one, the complexity of the treatment, and the level of staff training required. For a detailed description of the stages, see Rousanville et al. (2001); for a chapter on related issues see Linehan (1999).

SUGGESTIONS FOR THE GRANT PROPOSAL

Given the typical need for grant funding for outcome studies, several strategies are suggested.

1. *Read the web page of the institute or foundation to which you would like to apply for funding.* This allows you to better understand its priorities. For example, the National Institute on Drug Abuse (NIDA) web page (www.nida.nih.gov) shows what grants have been funded by the Institute, describes current research funding announcements (areas the Institute wants to pursue), and offers policy statements relevant to research and clinical practice.
2. *Model your study on exemplary outcome studies.* Take the time to identify high-quality studies that are as close as possible to the type you have in mind. For example, conduct a literature search on investigators who were funded by the institute you are applying to. If you are interested in designing a stage 1 study, locate published articles of stage 1 studies, and observe the design decisions that were made. Even if the treatment or

disorders are different, many design issues may be similar. (Some examples of stage 1 published studies include Najavits, Weiss, Shaw, & Muenz, 1998, and Weiss et al., 2000.) For a stage 2 study, some examples include the large multisite NIDA Collaborative Cocaine Treatment Study (Crits-Christoph et al., 1997); or, for a smaller-scale example, Linehan et al. (1999).

A caveat, however: modeling on exemplary studies is a start, but it is essential to carefully think out the issues particular to your own project. Each patient population, disorder, treatment, and context may require adaptation. A “copycat” grant can backfire—indeed, one grant applicant actually submitted three of the same applications, simply using a “search and replace” function on the computer to change the name of the treatment type; needless to say, none was funded.

3. *Consider embedding a gambling study within existing institute priorities.* In pathological gambling (PG), for example, it may be strategic to submit a proposal in such a way that both PG and some other issue, e.g., drug abuse or alcoholism, are addressed. Thus, the proposal can be submitted either to the National Institute on Alcohol Abuse and Alcoholism or to NIDA. Given the strong comorbidity between PG and other disorders, this is scientifically legitimate, while also helping find a home for your project.
4. *Know the field.* This means not just understanding the particular disorder and treatment you plan to study, but also having a wide-angle view of outcome research in general. Perhaps the best guide is Garfield and Bergin’s *Handbook of Psychotherapy and Behavior Change* (1994a), widely considered the “bible” of psychotherapy research. This book summarizes current knowledge as well as the development of psychosocial outcome and process research from its origin in the 1950’s to the present. It provides chapter topics such as “Research on client variables in psychotherapy” (Garfield, 1994), “Therapist variables” (Beutler, Machado, & Neufeldt, 1994), and “Assessing psychotherapy outcomes and processes” (Lambert & Hill, 1994).
5. *Read “how to” guides for funding.* One especially helpful article for new investigators is Rush, Gullion, and Prien (1996). They outline how NIH grant procedures work and tips for getting funded, including how to select research hypotheses, documenting feasibility, establishing effect sizes, and writing a

“reader-friendly” application. Another guide is Bauer (1999). A search at libraries and on the web will yield others.

OPTIMAL STUDY DESIGN

What is the ideal study design? The answer depends on the specific questions being addressed, and the realistic framework of what can be achieved. The most significant contributions to knowledge, however, take each area below into account in a careful manner. Study design may seem a rather technical area, but at a higher level, all empirical knowledge is linked to study design: each choice will determine what can ultimately be known from the project.

Comorbidity Design Issues

Typically, PG cooccurs with other mental health or substance use disorders. Thus, a question will be how to plan for the assessment and treatment of multiple disorders in your project. There are four basic strategies for the treatment of co-occurring disorders. Each of these is described here in relation to two disorders (e.g., an addictive disorder and a mental health disorder), but if a person has more than two disorders, the same principles may likely apply.

1. *Integrated treatment*: both disorders are treated at the same time by the same treater.
2. *Parallel treatment*: both disorders are treated at the same time by different treaters.
3. *Sequential treatment*: treatment for one disorder, then the other disorder.
4. *Single treatment*: treatment for only one disorder.

It is usually recommended that when two disorders co-occur, both are treated at the same time. Such integrated treatment has come to be viewed as the state-of-the-art for dual diagnosis treatment (Weiss, Najavits, & Mirin, 1998). However, there may be legitimate reasons to study some other model. For example, many programs do not have staff who are capable of treating both an addictive and a mental disorder. For licensed addiction counselors, it may be an ethical violation to

conduct mental health treatment. Thus, an investigator might choose to study how treatment fares when two treaters are provided, one for addiction and one for mental health. Or, given the relatively few outcome studies on dual diagnosis treatment, an investigator might want to study sequential treatment in contrast to integrated or parallel treatment. Most likely the only design that would be hard to justify would be single-model treatment, unless one has excluded dual diagnosis cases (e.g., in a study of PG., excluding those who have a mental or addictive disorder other than PG).

At this point, virtually all studies address only two disorders at most (e.g., an addictive disorder and the most prominent mental disorder). While it is known that patients may have three, four or more disorders, studies at this time rarely have the practical resources or theoretical knowledge base to address all of them at the same time. For example, a severe PG patient may have alcoholism, major depression, and PTSD. Most investigators currently focus on the addictive disorders plus one area of mental health. The best studies will monitor and assess all disorders even if they do not attempt to treat all of them.

Finally, the choice of treatments will also be affected by co-morbidity. A treatment that may work effectively in a single-diagnosis population may not be helpful in the context of co-morbidity. For example, Gamblers Anonymous may not be advisable for a patient with severe social phobia and PG until the social phobia is treated.

Treatment

Treatment Selection and Development. The choice of a treatment is clearly a key element of any outcome study, as well as the question of most interest: i.e., does this treatment work for this population? If the investigator plans to develop a new treatment, two questions should first be addressed:

1. Is there truly a need for a new treatment? A comprehensive literature search is essential to avoid “reinventing the wheel.” A treatment may already exist for the population, or a treatment may exist in a related area that could be adapted rather than designing a wholly new treatment. Given the close association between PG and other addictive disorders, for example, one might first want to try treatments that are known to

work for substance abuse, such as 12-step or cognitive-behavioral. If the choice is to develop a new treatment, a rationale (comparing and contrasting it with existing treatments), is essential.

2. If a new treatment is to be developed, how can it be done in the most careful way possible? For example, will the treatment be refined based on patient and therapist feedback, and perhaps expert consultation? Will there be focus groups to understand what issues the population most needs addressed? What processes will be included to make sure to truly listen to the patients and providers about the treatment (e.g., questionnaires? interviews? qualitative analysis of session tapes?) Will the investigator conduct at least part of the treatment, so as to experience it firsthand? Moreover, if the treatment changes over the course of the study based on such feedback, how will the data be affected? If the treatment is initially 12 sessions, for example, but gets expanded to 24 sessions, how will the outcome data be interpreted, when some patients have received much less of the treatment than others?

Treatment Integrity. One of the major advances in treatment research has been the advent of systematic ways to insure that the treatment is being conducted as intended. Standard methods are described below to ensure treatment adherence (is the treatment being conducted as planned?) and treatment purity (is the therapist providing *only* the planned treatment, and not other extraneous treatments?). See Carroll et al. (1998) for an example of how these issues were addressed in a major addiction study.

Use of a Treatment Manual. This is a required element of all outcome studies. The manual specifies, in detail, the theory, rationale, and methods of the treatment being studied. See Addis (1997) for a discussion of this topic, and Najavits, Weiss, Shaw, and Dierberger (2000) for a description of different elements within a manual.

Adherence Ratings (and Taped Sessions). In virtually all studies, some number of sessions are taped, either by audio or video, and rated to determine the degree to which the treatment was conducted as planned. Adherence ratings are usually divided into fidelity (how *much*

of the treatment was done?) and quality (how *well* was the treatment done?). See the section “Assessment” for the importance of interrater reliability and other psychometric qualities of the adherence scale, and the training and supervision of raters.

Limits on Uncontrolled, External Treatments. This refers to the need to ascertain, and often to control, the type and amount of treatment patients receive outside of the treatment provided on the study. Without this, it would be unclear how much of the outcome is based on the treatment being studied versus external treatments. For example, if a patient receives two weeks of inpatient psychiatric treatment during a study, how would this be handled? If a patient decides to seek out extra treatment while in the study how will this be handled (e.g., pursuing acupuncture detoxification for substance abuse, or obtaining a new medication)? Given that many patients attend 12-step self-help groups, how will this be monitored and taken into account in later data analysis? Typical strategies for dealing with external treatment include: (a) asking patients to limit the amount of external treatment, and making this consistent across study conditions (e.g., Crits-Christoph et al., 1997); (b) monitoring the amount and type of external treatment using standardized measures such as the Treatment Services Review (McLellan, 1989) throughout the study so that in later data analysis, it can be factored out; and (c) defining criteria for “protocol completers” and “protocol violators” (i.e., what amount or type of external treatment would be of such impact on outcome that patients would be considered a “protocol violator” and their data analyzed separately from those who completed the study as planned)? (See Siqueland et al., 1998 for an example of how protocol violators are classified in an addiction study.) Note that this issue will depend on whether the treatment is a stage 1 or 2 study (in which control over external treatments is key), or a stage 3 study (in which it is usually assumed that external treatments are part of the naturalistic design).

Intent-to-Treat and Completer Analyses. Most studies analyze the data in two ways. The first is an *intent-to-treat* analysis, meaning that all randomized patients are stratified according to their assignment (experimental versus control condition) and analyzed regardless of whether they attended even a single treatment session. The second type of analysis is the *completer analysis*, meaning the comparison of patients

who attended the treatment as planned (e.g., completing at least 25% of available treatment sessions) versus those who did not (dropouts from treatment).

Evaluation of Active Ingredients of Treatment. This is also known as dismantling the effects of the treatment; that is, trying to understand what about the treatment might account for its impact. It is typically addressed as part of a stage 2 study. For example, if the treatment consists of two components (perhaps a cognitive-behavioral skills training intervention for 10 sessions and an emotional processing intervention for 12 sessions), the outcome results may be accounted for mostly by the emotional processing component rather than the CBT component. Investigators usually develop a scale to explore what aspects of the treatment patients and clinicians find most helpful, or design other ways to evaluate this question. For an example of how the active-ingredients question was addressed for eye movement desensitization reprocessing therapy see Devilly, Spence, and Rapee (1998).

Comparison of Treatment-Specific Versus Nonspecific Variables. It is widely accepted that all therapies comprise both treatment-specific and non-specific elements; the former are unique to the treatment while the latter occur in all treatments, such as support, alliance, and information (Garfield & Bergin, 1994a). It can be useful to explore how much of the outcome in a study derives from each. Adherence scales are one way to do this if the scale is designed to rate both treatment-specific and non-specific variables. For examples in the substance abuse field, see Barber, Mercer, Krakauer, and Calvo (1996) and Carroll et al. (2000).

Follow-Ups After Treatment Has Ended. Follow-up assessments measure the treatment's degree of lasting impact. Also, some treatments have "delayed-emergence effects"; that is, they show increased impact *after* the treatment has ended (Carroll et al., 1994). Stage 1 studies may have a short-term follow-up, such as three months post-treatment, while stage 2 or 3 studies might have one or more years of follow-up.

Evaluation of Costs of Treatment. A true cost-effectiveness study is very expensive to conduct as it requires comprehensive evaluation of costs outside the study (days lost from work, medical costs, etc.). None-

theless, a researcher can provide basic data on costs that may help future researchers and consumers; such data might include the cost of therapy sessions, training, and supervision.

Description of Any Features That Are Not Typically Provided in a Clinical Setting. Investigators often include incentives to encourage patients to attend assessments and/or treatment sessions (e.g., babysitting, transportation vouchers, food). Unless these are disclosed, it is likely that the results of the study will be biased in a positive direction (i.e., better outcomes will be found than would be possible in real-world settings that do not provide such incentives). One incentive deserves particular note: investigators sometimes pay patients just to attend treatment (e.g., \$15 for every therapy session attended). This is a treatment intervention in and of itself (see Higgins et al.), and must be reported in any published articles on the study.

Standard Design Features

Several issues need to be addressed in any study, although how they are resolved will, again, depend on existing parameters such as the stage of treatment development, the population, and the type of treatment. The points here are drawn largely from Najavits & Weiss (1994a), but the reader is referred to a classic work on the subject by Campbell and Stanley (1966). Other more recent books on study design may also be helpful. Most of the considerations below are required for a stage 2 or 3 study; they may or may not be needed in a stage 1 study.

Ideally, the study will be *a priori*, which means that the hypotheses are conceptualized prior to data collection, thus obtaining data that are most relevant to the question of interest. In contrast, a *post hoc* analysis means re-analyzing existing data to explore new questions that were not originally hypothesized. Such a re-analysis increases the likelihood of type II error (i.e., the chance that one will find a positive result that is not really true).

The optimal study also typically includes a *control or comparison group*. The former is a group that does not receive systematic treatment (e.g., a no-treatment control, a wait-list control, or a treatment-as-usual control in which patients obtain whatever they normally would in the community); the latter is defined as a group that receives an

alternative systematic treatment different than the study treatment (such as a study in which cognitive-behavioral therapy is compared to drug counseling, both of which use treatment manuals and equivalently trained staff). Note that selecting an appropriate control or comparison group can be a challenge; see Borkovec (1993) for a helpful discussion of this issue.

The study also will ideally include *random assignment to treatment conditions*, such that patients have an equal chance of being assigned to the treatment or the control/comparison conditions. The best studies also have *random assignment to therapist within treatment conditions*, as assignment to therapist may have as much or more of an impact on outcome as treatment type (Najavits & Weiss, 1994b). A solid study also uses *balanced design*, that is, equal amounts (“dose”) of treatment for each condition in the study. The sample size needs to be determined by *power analysis* (Cohen, 1977).

Any non-standard design elements (e.g., use of within-subject controls) need to be justified. Moreover, all design choices should be clearly related to the end-goals of the project: what questions can legitimately be addressed, given the design that is selected? For example, one proposal sought to study a new medication in the treatment of PG compared to a placebo control. This would be fine, except a manualized psychosocial treatment was also included for all patients, which would, in the end, make it impossible to determine whether outcomes were due to the medication or placebo, the psychosocial treatment, or an interaction (as the combination of medication or placebo plus a psychosocial treatment may be more powerful than either alone). If the key question were the impact of the medication plus the psychosocial treatment, such a design would be legitimate; if the key question were the impact of medication alone, it would not. The bottom line is that there is no one right method, but rather the careful selection of methods that accurately test a question of interest.

Assessment

DSM-IV Diagnoses. DSM-IV diagnoses are standard in outcome studies for deciding which patients to include or exclude. The best studies, in addition, provide a broad assessment of other Axis I and, if possible, Axis II disorders. Specifying whether the disorders are current, past, or both is essential, and the level of severity is also helpful.

Using structured clinical interviews to obtain the diagnoses is standard (such as the Structured Clinical Interview for DSM-IV by Spitzer, Williams, & Gibbon, 1997; or the Diagnostic Interview Schedule by Robins, Helzer, Croughan, & Ratcliff, 1981); chart review diagnoses or unstructured interview are considered unreliable. Occasionally, self-report measures are used, but usually only for diagnoses not central to the study.

Varied Measures. The first step is the selection of outcome measures to address the central outcomes desired in the study (for a study on PG, the amount and type of gambling, for example). It is now common for studies to re-evaluate the central diagnoses at the end of treatment and follow-up to determine not just whether there was a reduction in the severity of symptoms but whether the patient still has the diagnosis. In addition, other measures are included to tap a wide range of symptom and functioning variables. For example, PG treatment may impact substance use, family functioning, legal problems, financial problems, and depressive symptoms. In addition to outcome measures, the inclusion of process variables is also highly desirable. This might include ratings of therapeutic alliance (Luborsky et al., 1996), emotional responses of the therapists toward their patients (Najavits et al., 1995), and satisfaction with the treatment (Attkisson & Zwick, 1982).

Urinanalysis Testing. For any study in which substance use is assessed, drug testing is important. It gives the investigator a check on the veracity of patients' self-report. Such testing might be through a local laboratory, or, increasingly common, using on-site test kits.

Independent Outcome Ratings. At least some measures should be conducted by independent assessors, rather than solely self-report by patient and therapist. The assessors will typically be blind to treatment condition, and will be the same assessors at each time point per patient to promote the most consistent possible ratings.

Iatrogenic Effects. The majority of studies evaluate positive and neutral outcomes, but rarely assess for whether the treatment was *harmful* to patients. Yet any treatment that has the power to help also has the power to harm. Considering that most studies are conducted

by investigators who have an investment in obtaining positive outcomes, there thus appears to be a bias, whether conscious or unconscious, to not want negative feedback about one's treatment (Mohr, 1995; Najavits, 2001). To remedy this, scaling on measures should allow the full range of negative, neutral, and positive ratings (e.g., ratings from -3 , "extremely harmful" through 0 "neutral," to $+3$ "extremely helpful"). Much scaling only ranges from zero upwards, which omits harmful impact. Also, items should seek to solicit negative feedback (e.g., "Did the treatment make your PG *worse* in any way? If so, how?").

Statistical Testing of Outcome Results. While most studies include statistical testing, the quality varies. Consultation with a statistician is always advised. Some of the most common errors in outcome studies include "fishing expeditions" in which too many statistical tests are conducted, heightening the likelihood of false-positive findings (type II error); reporting insufficient statistical information such that later investigators will find it difficult to compare results, and using the wrong statistic (e.g., using percent agreement rather than kappa when assessing interrater reliability).

Psychometric Properties of Instruments. Reporting basic psychometric properties of instruments is standard. If a measure is not standardized, that information should be stated.

Training and Reliability of Assessors. The training and reliability of the interviewers should be included, particularly for those who conduct the diagnostic assessment.

Data on Dropouts. Data needs to be collected on all patients throughout the study, regardless of whether or not they drop out of the treatment. Dropouts are typically the more impaired patients, or those who do not like or benefit from the treatment. Their data is important to include in the outcome analysis; otherwise, results may be biased in a positive direction (making the treatment look better than it really is). (See the section "Intent-to-Treat and Completer Analysis" above.) After the study is over, it is also standard to statistically compare dropouts from completers to understand whether some systematic factor may have led patients to drop out. For example, if the drop-

outs had significantly lower education levels than completers, it could be that the treatment was too abstract or required too high a reading level.

Locating patients who have dropped out often requires “detective work” by the staff, and involves the collection of names and phone numbers of contacts from the start of the project, in case the patient is difficult to locate. An excellent guide to locating dropouts from addictions treatment is provided by Zweben et al. (1998); it also provides practical information on how to conduct an outcome study (e.g., patient tracking and data entry procedures). It can be ordered free through the National Clearinghouse for Alcohol and Drug Information (800-729-6686 or via the web at www.health.org). See also Cottler et al. (1996).

Therapist Variables. Most studies provide ample assessment of patients, but only very limited therapist assessment (e.g., degree, experience level, age, and gender). Expanded therapist measures might include knowledge base (e.g., knowledge of the treatment, knowledge about PG), attitudes (e.g., views on controversial issues such as the legalization of gambling; whether patients need to attend 12-step groups, the use of ultimatums in treatment, etc.), and personality characteristics.

Improper Use of Materials. It is not uncommon to read of a treatment manual or measure being “adapted” for a study without proper permission and/or citation of the original author. Intellectual property is a central issue in academia and the correct procedure is to contact the author to clarify copyright issues, obtain the right to adapt the work, and agree on how the adapted version will be cited.

Patients

Patient Selection. It is important to include a wide, representative range of patients and to avoid exclusionary rules that create a biased sample. Some studies will automatically rule out any patients who have suicidal ideation, a current crisis, medical problems, substance abuse, a personality disorder, are taking psychiatric medication, or are unwilling to attend 12-step groups. The field of outcome research has been criticized for such narrow selection, which ultimately can make the

research irrelevant. In the real world, clinicians cannot refuse such patients. Each exclusionary criterion needs to be explained and carefully justified in light of the study's goals, and needs to be reported in published articles on the treatment. In general, the broader the inclusion, the more generalizable the results.

Emergency Clinical Situations. Discussion of how patients will be treated in clinical emergencies is needed. This includes: will patients have access to after-hours help if they deteriorate? How will patients be treated if they become at risk of harm to self or others? What will happen if patients show up high or drunk to a treatment or assessment session? (For the latter, it is standard to let patients know in advance that sessions cannot be held if they arrive intoxicated; the session is rescheduled and the patient is sent home safely, such as via taxi or a friend.)

Therapists

This is perhaps the most neglected area of outcome research. It is often assumed that therapists are equivalent, and thus attention is directed almost exclusively toward the treatment and the patients being studied. However, recent reviews indicate that, empirically, the therapist is one of the most powerful determinants of outcome, often beyond the impact of treatment and patient factors (Najavits & Weiss, 1994b; Najavits, Crits-Christoph, & Dierberger, 2000). Issues include the following.

Therapist Selection. The best studies describe fully their therapist selection criteria. For example, some studies require therapists to provide two initial work samples prior to hire (audiotapes of their typical work with real patients). Having therapists conduct a sample case of the full treatment and being rated for adherence prior to being certified as a study therapist is also a useful method (e.g., Crits-Christoph et al., 1998). It is noteworthy that easy-to-measure therapist professional characteristics (e.g., degree, training, experience level, and in addiction studies, recovery status) are often presumed to be useful selection criteria; however, a large body of research indicates that these do not show a strong relationship to outcome (Zlotnick, 1996; Najavits et al., 2000). Thus, requiring therapists to have a particular

degree or experience level should be justified, and at the least, therapists should be assessed for performance once hired (through adherence and actual outcome and retention of patients). The goal in most outcome studies, particularly those in stages 1 and 2, is for therapists to be as homogeneous as possible in their performance: performing highly and consistently both within therapists (i.e., within each therapist's caseload) and across therapists (i.e., among therapists in the study). In a stage 3 study, therapists may be much more variable as the goal is to explore use of the treatment under real-world conditions.

Description of Training. Training and supervision of therapists on the study need to be described for future replication by others.

Therapist Effects in All Studies. Investigators often fail to examine clinician differences in outcomes (Martindale, 1978; Crits-Christoph, 1991). Yet it is possible, indeed often the case, that within a treatment therapists differ in outcomes. For example, a study has six CBT therapists and six psychodynamic therapists. Most studies simply compare the CBT treatment versus the psychodynamic treatment in outcome results. But a first step is to ascertain whether the CBT therapists differed among themselves (e.g., perhaps one therapist achieved outstanding results within her caseload while another did poorly). If they differ, interpretation of the results will need to take this into account. Note that therapists may need to be analyzed as a random rather than fixed factor. Moreover, it is important to analyze both therapist retention and outcome. For examples of analyzing therapist differences see (Crits-Christoph, 1991; Najavits & Strupp, 1994).

Replication

A known feature of outcome research is the "wild card phenomenon": that is, studies by investigators who developed the treatment show more positive outcomes than other investigators, even when the same procedures are used (Luborsky & Diguier, June, 1993). This is believed to result from the extra effort and perhaps unconscious biases of investigators who pioneered a treatment. Replication by investigators who did not develop the treatment is at some point a necessary step.

SUMMARY

Designing an effective study requires consideration of many issues. While there is no one right design, there are elements that can make a study more rigorous, and thus more able to answer the questions of interest. Decisions about design will depend on parameters such as the stage of treatment development, the population, the type of treatment, and the realistic strengths and weaknesses of one's context (staff availability and training, ability to recruit patients, etc). Several general questions can also help guide design decisions:

1. *Have all or most of the issues above been addressed?* Consideration of such issues is the first step in solid design. While few studies will be able to provide all features, the closer one can move toward them, the better.
2. *Is there a rationale for all design decisions?* The investigator should be clear and explicit on why decisions were made. Also, if there is a decision not to include a feature that is standard in comparable studies, a rationale for its exclusion is needed.
3. *Is the investigator aware of what can and cannot be inferred from the data, based on the design?* All design decisions lead to strengths and limitations about what ultimately can be asked of the data. Thus, for example, if a treatment is compared to a no-treatment control rather than an active-treatment control, the interpretation of results will differ as well.
4. *Does the study design address research questions that are important, relevant, and genuinely of benefit to the field?* Unless the research is of real value, there is no point in conducting it, particularly in the area of outcome research where studies are expensive and take many years. It is helpful to justify whether the research question is truly novel, addresses an area of public health relevance, or will be able to provide some new understanding of an existing question. Ultimately, the hope is that good outcome research can powerfully improve clinical treatment.

ACKNOWLEDGMENT

Supported by grants RO1 DA-08631 and K02 DA-00400 from the from the National Institute on Drug Abuse, R21 AA-12181 from the

National Institute on Alcohol Abuse and Alcoholism, and the Dr. Ralph and Marian C. Falk Medical Research Trust

REFERENCES

- Addis, M. (1997). Evaluating the treatment manual as a means of disseminating empirically validated psychotherapies. *Clinical Psychology: Science and Practice*, 4, 1–11.
- Attkisson, C. C., & Zwick, R. (1982). The Client Satisfaction Questionnaire: Psychometric properties and correlations with service utilization and psychotherapy outcome. *Evaluation and Program Planning*, 5, 233–237.
- Barber, J. P., Mercer, D., Krakauer, I., & Calvo, N. (1996). Development of an adherence/competence rating scale for individual drug counseling. *Drug and Alcohol Dependence*, 43, 125–132.
- Bauer, D. (1999). The 'How To' Grants Manual: Successful Grantseeking Techniques for Obtaining Public and Private Grants. New York: American Council on Education, Oryx Press.
- Beutler, L. E., Machado, P. P., & Neufeldt, S. (1994). Therapist variables. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of Psychotherapy and Behavior Change* (pp. 229–269). New York: John Wiley & Sons, Inc.
- Bickman, L. (1999). Practice makes perfect and other myths about mental health services. *American Psychologist*, 54, 965–977.
- Borkovec, T. (1993). Between-group therapy outcome research: Design and methodology. NIDA Research Monographs, 137, 249–289.
- Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Carroll, K., Connors, G., Cooney, N., DiClemente, C., Donovan, D., Kadden, R., Longabaugh, R., Rounsaville, B., Wirtz, P., & Zweben, A. (1998). Internal validity of project MATCH treatments: Discriminability and integrity. *Journal of Consulting and Clinical Psychology*, 66, 290–303.
- Carroll, K., Nich, C., Sifry, R., Nuro, K., Frankforter, T., Ball, S., Fenton, L., & Rounsaville, B. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence*, 57, 225–238.
- Carroll, K. M., Rounsaville, B. J., Nich, C., Gordon, L. T., Wirtz, P. W., & Gawin, F. (1994). One-year follow-up of psychotherapy and pharmacotherapy for cocaine dependence: delayed emergence of psychotherapy effects. *Archives of General Psychiatry*, 51(12), 989–997.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. (revised edition ed.). New York: Academic Press.
- Cottler, L., Compton, W. M., Ben-Abdallah, A., Horne, M., Claverie, D. (1996). Achieving a 96.6% completer rate among substance abusers. *Drug and Alcohol Dependence*, 41(3): 209–217.
- Crits-Christoph, P. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, 1, 81–91.
- Crits-Christoph, P., Siqueland, L., Blaine, J., Frank, A., Luborsky, L., Onken, L., Muenz, L., Thase, M., Weiss, R., Gastfriend, D., Woody, G., Barber, J., Butler, S., Daley, D., Bishop, S., Najavits, L. M., Lis, J., Mercer, D., Griffin, M., Beck, A. T., & Moras, K. (1997). The NIDA Cocaine Collaborative Treatment Study: Rationale and methods. *Archives of General Psychology*, 54, 721–726.
- Crits-Christoph, P., Siqueland, L., Chittams, J., Barber, J. P., Beck, A. T., Liese, B. S., Luborsky, L., Mark, D., Mercer, D., Woody, G., Onken, L. S., Frank, A., Thase, M., & Najavits, L. M. (1998). Training in cognitive therapy, supportive-expressive therapy, and drug counseling treatment for cocaine dependence: Results from the NIDA Cocaine Collaborative Treatment Study. *Journal of Consulting and Clinical Psychology*, 66, 484–492.
- Devilly, G., Spence, S., & Rapee, R. (1998). Statistical and reliable change with eye movement desensitization and reprocessing: Treating trauma within a veteran population. *Behavior Therapy*, 29, 435–455.

- Garfield, S. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of Psychotherapy and Behavior Change* (4th ed., pp. 190–228). New York: John Wiley & Sons, Inc.
- Garfield, S., & Bergin, A. (1994a). *Handbook of Psychotherapy and Behavior Change*. New York: John Wiley & Sons, Inc.
- Garfield, S., & Bergin, A. (1994b). Introduction and historical overview. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of Psychotherapy and Behavior Change* (4th ed., pp. 3–18). New York: John Wiley & Sons, Inc.
- Hay Group (1999). Health care plan design and cost trends—1988 through 1998. Arlington, VA: National Association of Psychiatric Health Systems and Association of Behavioral Group Practices.
- Higgins, S. T., Budney, A. J., Bickel, W. K., Foerg, F. E., Donham, R., & Badger, G. J. (1994). Incentives improve outcome in outpatient behavioral treatment of cocaine dependence. *Archives of General Psychiatry*, 51, 568–576.
- Korn, D., & Shaffer, H. (1999). Gambling and the health of the public: Adopting a public health perspective. *Journal of Gambling Studies*, 15, 289–365.
- Lambert, M. J., & Hill, C. E. (1994). Assessing psychotherapy outcomes and processes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of Psychotherapy and Behavior Change* (4 ed., pp. 72–113). New York: John Wiley & Sons, Inc.
- Linehan, M. (1999). Development, evaluation, and dissemination of effective psychosocial treatments: Levels of disorder, stages of care, and stages of treatment research. In M. Glantz & C. Hartel (Eds.), *Drug abuse: Origins & interventions* (pp. 367–394). Washington, DC: American Psychological Association.
- Linehan, M. M., Schmidt, H., Dimeff, L. A., Craft, J. C., Kanter, J., & Comtois, K. A. (1999). Dialectical behavior therapy for patients with borderline personality disorder and drug-dependence. *American Journal on Addictions*, 8, 279–292.
- Luborsky, L., Barber, J. P., Siqueland, L., Johnson, S., Najavits, L. M., Frank, A., & Daley, D. (1996). The revised Helping Alliance questionnaire (HAQ-II): Psychometric properties. *Journal of Psychotherapy Practice and Research*, 6, 260–271.
- Luborsky, L., & Diguier, L. (June, 1993). A 'wild card' in comparative psychotherapy studies—The researcher's therapeutic alliance. Paper presented at the Paper presented at the annual meeting of the Society for Psychotherapy Research, Pittsburgh PA.
- Martindale, C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 46(6), 1526–1530.
- McLellan, A. T. (1989). Treatment Services Review. Unpublished measure, University of Pennsylvania.
- Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology: Science and Practice*, 2, 1–24.
- Najavits, L. M. (2001). Early career award paper: Helping difficult patients. *Psychotherapy Research*, 11, 131–152.
- Najavits, L. M., Crits-Christoph, P., & Dierberger, A. E. (2000). Clinicians' impact on substance abuse treatment. *Substance Use and Misuse*, 35, 2161–2190.
- Najavits, L. M., Griffin, M. L., Luborsky, L., Frank, A., Weiss, R. D., Liese, B. S., Thompson, H., Nakayama, E., Siqueland, L., Daley, D., & Simon Onken, L. (1995). Therapists' emotional reactions to substance abusers: A new questionnaire and initial findings. *Psychotherapy*, 32, 669–677.
- Najavits, L. M., & Strupp, H. H. (1994). Differences in the effectiveness of psychodynamic therapists: A process-outcome study. *Psychotherapy*, 31, 114–123.
- Najavits, L. M., & Weiss, R. D. (1994a). The role of psychotherapy in the treatment of substance use disorders. *Harvard Review of Psychiatry*, 2, 84–96.
- Najavits, L. M., & Weiss, R. D. (1994b). Variations in therapist effectiveness in the treatment of patients with substance use disorders: An empirical review. *Addiction*, 89, 679–688.
- Najavits, L. M., Weiss, R. D., Shaw, S. R., & Dierberger, A. E. (2000). Psychotherapists' views of treatment manuals. *Professional Psychology: Research and Practice*, 31, 404–408.
- Najavits, L. M., Weiss, R. D., Shaw, S. R., & Muenz, L. R. (1998). "Seeking Safety": Outcome of a

- new cognitive-behavioral psychotherapy for women with posttraumatic stress disorder and substance dependence. *Journal of Traumatic Stress*, 11, 437–456.
- Robins, L.N., Helzer, J.E., Croughan, J., & Ratcliff, K.S. (1981). National Institute of Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry*, 38, 381–389.
- Rousanville, B., Carroll, K., & Onken, L. (2001). A stage model of behavioral therapies research: Getting started and moving on from stage I. *Clinical Psychology: Science and Practice*, 8, 133–142.
- Rush, J., Gullion, C., & Prien, R. (1996). A curbstone consult to applicants for National Institute of Mental Health grant support. *Psychopharmacology Bulletin*, 32.
- Siqueland, L., Frank, A., Gastfriend, D., Muenz, L., Crits-Christoph, P., Chittams, J., Thase, M., Mercer, D., & Blaine, J. (1998). The protocol deviation patient: Characterization and implications for clinical trials research. *Psychotherapy Research*, 8, 287–306.
- Spitzer, R. L., Williams, J. B. W., & Gibbon, M. (1997). *Structured Clinical Interview for DSM-IV-Patient Version*. New York: Biometrics Research Institute.
- Weiss, R. D., Griffin, M., Greenfield, S., Najavits, L. M., Wyner, D., Soto, J., & Hennen, A. (2000). Group therapy for patients with bipolar disorder and substance dependence: Results of a pilot study. *Journal of Clinical Psychiatry*, 61, 361–367.
- Weiss, R. D., Najavits, L. M., & Mirin, S. M. (1998). Substance abuse and psychiatric disorders. In R. J. Frances & S. I. Miller (Eds.), *Clinical Textbook of Addictive Disorders* (2nd ed., pp. 291–318). New York: Guilford.
- Zlotnick, C. (1996). Therapists' differences in experience. *Professional Psychology*, 9, 28–34.
- Zweben, A., Barrett, D., Carty, K., McRee, B., Morse, P., & Rice, C. (Eds.). (1998). *Strategies for Facilitating Protocol Compliance in Alcoholism Treatment Research*. (Vol. 7). Bethesda, MD: U.S. Department of Health and Human Services.

Received July 6, 2001; final revision July 19, 2002.