Applied Issues in Treatment Outcome Assessment

J. Scott Tonigan, Ph.D.

Center on Alcoholism, Substance Abuse and Addictions (CASAA)
Albuquerque, NM

It is an exciting time to conduct alcoholism treatment outcome evaluation. Advancements in statistical software for personal computers, for example, have dramatically increased the type and complexity of techniques available to the evaluator. Although some concern has been raised about how the democratization of the tools of evaluation may precipitate their inappropriate use (e.g., Pedhazur 1982), indirect evidence suggests that increased accessibility has had an overall positive effect in the field. Miller et al. (1995*a*) found, for example, that the methodological quality of research outcome studies has improved significantly in the past 20 years, much of this due to selection of assessment instruments with known psychometric properties and the appropriate use of multivariate techniques. Software advances for personal computers have also spawned an audience-friendly revolution in how findings are presented, with time-to-event outcomes, hierarchical linear modeling findings, and structural equation modeling findings now presented in an understandable and graphic format.

It is also a critical time for doing rigorous outcome evaluation. In many States evaluation is now legislatively mandated, with future program appropriations tied to demonstration of treatment effectiveness. Programs and jobs can hinge on how well an evaluation report communicates findings to audiences unfamiliar with research methodology and the multifaceted nature of alcohol treatment outcome(s). Under these conditions the evaluator has a clear responsibility to select assessment tools with demonstrated reliability and validity that are also sensitive to, and theoretically consistent with, treatment program objectives.

The purpose of this chapter is to familiarize the reader with a variety of fundamental issues that arise in the conduct of outcome evaluation in alcoholism treatment. The relative merits of specific measures of alcohol consumption (see the chapter by Sobell and Sobell) and biological markers (see the chapter by Allen et al.) are reviewed elsewhere in this *Guide* and will not be reiterated. This chapter begins with a general discussion of the importance of using assessments with strong psychometric properties. Reliability theory is described from an applied perspective, with examples provided using assessment tools reviewed in this *Guide*. The next section briefly addresses the goals of summative and formative alcohol-related outcome evaluation, highlighting the differences between individual and group-based evaluation. This is followed by a section that reviews alternative perspectives of alcoholism, with attention directed to how these definitions of alcoholism suggest relevant measures of change; a section that discusses the measurement of behavior change across time, noting how commonly observed patterns of behavioral change differ across particular domains of functioning;

and a section that introduces the concept of *meaningful* changes in drinking behavior and then offers specific recommendations for clinicians and researchers on how to evaluate the magnitude of behavior changes associated with treatment. The final section outlines some practical considerations in alcohol outcome evaluation, including interviewer role and training, instrument consistency, and data entry.

The Value of Reliable Measures

Reliability refers to the extent that a measure is consistent and stable. In this regard, classical psychometric theory states that an observed score (O) is a function of the true score (T) and measurement error (E); O = T + E. Formally, reliability can be defined as

$$r_{xx} = 1 - (S_e^2 / S_x^2)$$

where r_{xx} is reliability, S_e^2 is error variance in a group of scores, and S_x^2 is variance in a group of observed scores. Reflection on the general meaning of the reliability formula reveals that a reliability coefficient (possible range 0 to 1.0) represents, in essence, the proportion of "true" score variance measured by a given instrument. Reliability coefficients approaching a value of 1.0 therefore indicate that nearly all variability in responses represents "true" or actual variability (no measurement error), while a reliability coefficient beneath 0.50 indicates that less than half of the variability in observed scores reflects "true" variability in the measured attribute (high measurement error).

To underscore the importance of reliability, imagine that a clinician is interested in the relationship between number of therapy sessions attended and days abstinent in a 60-day period. The question is not trivial for the clinician because of growing pressures to simultaneously enlarge caseloads and provide fewer sessions per client. Assume the reliability of the measure of sessions attended is good, 0.95, but the reliability of the days abstinent measure is poor, 0.50. Finally, assume the *real* correlation between days in therapy and days abstinent is 0.75. The net result of measurement error in this example is that the observed correlation *cannot* exceed 0.52 (0.95 x 0.50 x 0.75). Thus, although frequency of therapy accounts for more than half of the *real* variance in posttreatment abstinence (0.75^2 = 56 percent), the use of an unreliable measure in this example would lead the therapist to conclude that the relationship is not strong enough to warrant approval of a greater number of therapy sessions (0.52^2 = 27 percent).

As shown, the net effect of measurement error is to attenuate the magnitude of an observed correlation (Hunter et al. 1982). This is always the case. Unlike our example, however, the actual population correlation is rarely known and, as a result, the exact cost of measurement error is difficult to estimate. Measurement error, or lack of reliability, can therefore mask relationships of interest and, in some cases, may lead evaluators to draw too weak conclusions about treatment efficacy. A key point is that the relative importance of measurement error is inversely proportional to the anticipated magnitude of effect. As such, it is particularly important to use highly reliable measures when small effects are anticipated.

The standard error of measurement is defined as: $S_e = S_x \mid 1 - r_{xx}$. This statistic is an invaluable aid for researchers and practitioners for interpretation of individual scores. For example, the 25-item Alcohol Dependence Scale (ADS) is commonly used to screen individuals at risk of alcohol dependence. Generally, a score of 9 or higher (possible range is 0 to 47) is suggestive of DSM alcohol dependence. Skinner and Horn (1984) reported that, as part of a larger test-retest exercise, the 25-item ADS had a reliability coefficient of 0.92, and in a normative sample of problem drinkers (N = 225) the ADS had a standard deviation of 11. The standard error of measurement for

the ADS with problematic drinkers is therefore $S_e = 11 \mid 1 - 0.92$ or 3.11. What does this value of 3.11 mean? Applying the normal curve, we can develop a *band interpretation,* which states that a respondent's "true" score will be ± 3.11 its observed value 68 percent of the time, and 2 x 3.11 = 6.22 its observed value 95 percent of the time. From this example one can see that to have 68 percent certainty about a "true" ADS score of 9, one must consider potential observed scores that range between 5.89 and 12.11 (9 ± 3.11). In cases where cutoff values are used for screening or diagnostic purposes in alcohol treatment, it is especially important that the standard error of measurement be considered in making clinical decisions.

Three methods for investigating reliability are described in this section: stability, equivalency, and internal item consistency. An example of each method is presented using an assessment tool included in this *Guide*. The presentation is intentionally simplified and limited to those reliability statistics most commonly reported in alcohol-related literature. Readers interested in a more detailed account of these methods or a more comprehensive presentation of approaches to determine instrument reliability should refer to texts dedicated to the topic (e.g., Carmines and Zeller 1979; Aiken 2000).

Stability

This aspect of reliability refers to the extent that an observed score is consistent between two administrations (test-retest). Clearly, length of delay between administrations is an important consideration when assessing stability of measurement, with too short or too long of an interval introducing potential bias of recall and attribute instability effects, respectively. Ideally, length of delay between the two administrations balances attribute stability, measurement reactivity, and recall. Two of the most popular statistics to characterize the stability of two measurements are the Pearson product moment (*r*) and the intraclass correlations (ICCs). Because of their widespread use in assessing reliability, it is important to highlight how the ICC and the *r* coefficient provide different perspectives of stability.

The *r* coefficient expresses the degree to which paired values have similar rank orderings within their respective distributions. Absolute differences between paired values, however, are not considered in the computation of *r*. Thus, although the *relative* ranking of paired scores may be very similar, *absolute* values of the paired scores may be dissimilar. The ICC corrects for this limitation by indexing the absolute difference in agreement between paired scores as well as enabling partitioning of the variance of interest into several components. Standards to assess the reliability of instruments based on *r* are available and generally accepted. There is less agreement, however, about interpretation of ICCs. Cicchetti (1994) has recommended the following ranges to interpret the reliability of clinical instruments when ICCs are evaluated: below 0.40 = poor, 0.40 to 0.59 = fair, 0.60 to 0.74 = good, and 0.75 to 1.00 = excellent.

One example of the computational and interpretive differences arising between *r* and ICC was provided by Tonigan and colleagues (1997) in their evaluation of the test-retest reliability of Form 90. A test-retest study was conducted to investigate the reliability of primary measures used in Project MATCH, a large multisite study of client-treatment matching (Project MATCH Research Group 1997, 1998). A 2-day interval separated administration of the Form 90 interview conducted by different interviewers from different clinical sites (N = 70 pairs). The Pearson product moment correlation between test-retest counts of the frequency of days in which Alcoholics Anonymous (AA) was attended (90 days before the interview) was r = 0.87. This generally would be regarded as demonstrating good to excellent stability. In contrast, the ICC for frequency of AA days was ICC = 0.53, which

according to Cicchetti (1994) should be considered fair reliability. The important point is that the ICC will always yield a more conservative estimate of reliability relative to *r*.

Equivalency

This aspect of reliability examines the extent to which two different forms of the same test yield a consistent observed score. This kind of reliability also investigates equivalency among group means and the variance of two administrations of parallel tests. Theoretically, the split-half method of determining the internal item consistency of a test (discussed below) is a specialized aspect of equivalency testing. Statistics used to determine the equivalency of two parallel tests include the Pearson product moment and ICC coefficients. A unique advantage of a parallel test is that, in pre-post applications, the potential biasing effect of recall is minimized. In prevention research where knowledge gains following a school-based intervention are to be measured, the use of parallel tests with high reliability is worthy of consideration.

Babor (1996) offered an interesting variation in applying the equivalency approach to demonstrating instrument reliability. In the Project MATCH reliability study described earlier, two measures of alcohol dependence were collected, one a semi-structured interview based on DSM-III-R criteria (American Psychiatric Association 1987) and the other a 16-item self-report questionnaire (the Ethanol Dependence Syndrome [EDS] Scale) designed to parallel DSM-III-R criteria. Whereas the reliability of the semi-structured interview had received substantial attention, the 16-item "parallel" form had not. It is worth noting that the alternative forms also crossed method of data collection, that is, interview versus self-report. Pearson product moment correlations indicated that the two approaches yielded relatively consistent findings (range of *r*'s was 0.67 to 0.88) between the two assessments, with the EDS scale costing substantially less to administer.

Internal Item Consistency

Sometimes it is not possible to administer a test twice in a pre-post format to obtain reliability estimates, and for other reasons it may not be feasible or desirable to create parallel tests as is done in equivalency studies. It is still possible, nevertheless, to loosely assess the reliability of an assessment (using a single administration). Coefficients of internal item consistency, for example, identify the extent of item homogeneity in an assessment, which can inform one about the extent to which item content forms single or multiple *predicted* domains. As an example, the Drinker Inventory of Consequences (DrInC) (Miller et al. 1995*b*) was designed to measure adverse consequences associated with alcohol use. Miller and colleagues reasoned that such consequences could be grouped into discrete categories, including legal, health-related, interpersonal consequences, and the like. To this end, they developed an item pool representing each domain, had experts in the field review the items, and then administered the total pool of items to a sample of treatment-seeking clients (with items within each domain scattered in order). Logically, item responses within a domain ought to form a more homogeneous set than items combined across domains (or all items combined). Cronbach alpha is the most commonly reported statistic to reflect item homogeneity, which technically reflects the averaged extent to which each item correlates with its total set of items.

Summary

Measurement is the cornerstone of outcome evaluation. At least three benefits will accrue from struggling through the formulas, examples, and conceptual issues framed in this section. Foremost, knowledge of measurement reliability is necessary to be an educated consumer of the alcohol-related

assessment tools contained in this *Guide*. Second, understanding that "reliability" is a continuum in which instruments can be described as having less or more (as opposed to being inconsistent or consistent) is important for avoiding the pitfall of reifying measurements. Even measures considered as having "good" reliability (e.g., r_{xx} = 0.80), for example, do not fully account for, or precisely reflect, an individual's "true" score (e.g., 20 percent error in measurement). The third benefit is one of omission, having the knowledge *not* to follow the conventional practice of developing study-specific or clinician-derived assessment tools without any demonstrated reliability. Lack of reliability attenuates relationships of interest, whether they are investigated with correlational, analysis of variance (ANOVA)–based, or advanced statistical techniques such as multigroup structural equation modeling. Despite the argument that the need for content-specific assessments justifies "homegrown" assessments, meeting this need rarely compensates for the loss in measurement reliability.

Goals of Outcome Evaluation

The basic question in outcome evaluation is whether, and as the result of alcohol treatment exposure, a behavioral change has occurred. This change often refers to a reduction or cessation of alcohol consumption, although "harm reduction" models may place equal importance on changes in alcohol-related problems and high-risk–related behaviors. *Summative* evaluation addresses the question of programmatic value or the relative effectiveness of treatments; *formative* evaluation focuses on collection of information to improve existing treatment services. Generally, the unit of analysis in summative evaluation is aggregated, group-based data, whereas formative evaluation may include both individual-based and group-based information. This distinction is not firm, however, as summative evaluation may include case studies to illustrate group-based findings.

In defining the unit of analysis in evaluation, the core issue is to whom (or what) findings are to be generalized—to clients or to types of treatments. Typically, clinicians are concerned with the posttreatment functioning of *individuals*. Here, followup assessment identifies whether additional alcohol treatment may be indicated, whether an aftercare program is sufficiently meeting client needs, and/or if alternative or additional interventions may be indicated for non–substance abuse problems. Further, clinicians can evaluate client impressions of the therapeutic experience, noting how these services may be improved. These examples illustrate the major purposes of individual-based outcome evaluation, namely (1) therapeutic feedback to the client or therapist and/or (2) feedback to improve delivery of services.

Evaluation can also involve the examination of the relative changes in *groups* of individuals who have received alcohol treatment. Individuals' responses at followup are still recorded, with the important distinction that responses are aggregated to make decisions about the relative efficacy of treatment(s). In clinical settings, group-based evaluation generally is conducted to ascertain the extent of programmatic outcome evaluation of a single type of treatment, whereas in research settings programmatic outcome is conducted to determine the relative efficacy of different types of treatment. Several excellent texts are available that cover the topics of experimental and quasi-experimental design and potential threats to validity of findings (e.g., Cook and Campbell 1979).

Relevant Measures of Change

There is a historical appreciation of the importance of alcohol consumption as a criterion for judging treatment outcome, and most would regard assessment of outcome without such a measure as inade-

quate. There is less agreement, however, about the need to assess nondrinking domains to define outcome, and even less consensus about which domains may be particularly relevant. The recent attention to harm reduction models for evaluating outcome, which emphasize not the reduction of alcohol consumption per se but instead decreases in alcohol-related problems and risk-taking behaviors, has led to renewed interest in the issue of life functioning outcomes more generally.

Babor et al. (1988) summarized how differences in definition of outcome reflect two competing paradigms describing the phenomenon of alcoholism. One model views alcoholism as a *unitary* syndrome with abstinence as the sole marker of treatment response, or success. In this model, psychosocial functioning, employment, use of illicit drugs, and an array of other domains, although seen as important, are regarded as being so strongly associated with alcohol use that they can be inferred directly from changes in alcohol consumption; thus, they tend not to be considered extremely relevant for change measurement. On the other hand, a *multidimensional* model views alcoholism as a cluster of somewhat independent dimensions, with reductions in drinking as an important but not sole determinant (and indicator) of treatment efficacy. Because life functioning domains, such as physical health and social adjustment, are considered to fluctuate largely independently of one another, and because they also predict future alcohol consumption, proponents of the multidimensional model assert that outcome should be defined broadly, taking into account an array of domains (Longabaugh et al. 1994). It is important to note that, despite these differences between unitary and multidimensional models of alcoholism, the models intersect on the importance of measuring alcohol use using multiple measures that reflect various aspects of drinking (e.g., frequency and intensity).

The simplest analytical strategy to determine the viability of these two competing definitions of outcome is to correlate alcohol consumption with broader-based life functioning domain measures. Larger positive correlations would tend to support the unitary model, whereas modest to negligible correlations would support the multidimensional view of alcoholism. Table 1 summarizes the bivariate correlations between three measures of alcohol use for the 6-month period after alcohol treatment and five measures of client functioning also collected 6 months after treatment. The two samples in table 1 were recruited for Project MATCH, a study with high internal validity using only assessments with demonstrated reliability by highly trained and certified interviewers.

A basic conclusion to be drawn in surveying the magnitude of the correlations in table 1 is that, with the exception of alcohol-related problems, none of the correlations provide sufficient support for the unitary definition of alcoholism. To be sure, lack of instrument reliability attenuates the correlations of interest. It seems unlikely, however, that correction for attenuation would increase the magnitude of the correlations to the point of being supportive of the unitary concept of alcoholism. These findings do not agree with Emrick's (1974) recommendation that abstinence is sufficient to indicate posttreatment improvement in broader psychosocial domains. It is therefore recommended that psychosocial functioning be measured directly rather than inferred by changes in alcohol consumption.

Table 1 also facilitates comparison of the magnitude (hence stability) of relationships between drinking and psychosocial functioning by severity of alcohol-related problems. Can a stronger case be made for the unitary view of alcoholism among more or less severely impaired individuals? Relative to the outpatient sample in Project MATCH, for instance, the aftercare sample reported at recruitment significantly more frequent and intense drinking, a greater number of alcohol-related consequences, higher number of prior treatment experiences, and less social stability. The values in

TABLE 1.—Correlations between three measures of alcohol use and five measures of general functioning: Project MATCH aftercare (*N* = 772) and outpatient (*N* = 952) samples

Measures of general functioning	Measures of alcohol use 6 months posttreatment		
	PDA	DDD	First drink
Aftercare Sample			
BDI	–0.31 (0.34)	0.34 (0.07)	–0.27 (0.01)
Purpose in life	0.29 (0.11)	–0.32 (0.26)	0.24 (0.28)
PFI	0.20 (0.35)	–0.28 (0.27)	0.25 (0.01)
Alcohol-related problems	–0.55 (0.15)	0.67 (0.03)	–0.45 (0.01)
Illicit drug use	–0.13 (0.28)	0.13 (0.04)	–0.12 (0.42)
Outpatient Sample			
BDI	–0.29	0.27	–0.16
Purpose in life	0.23	–0.29	0.21
PFI	0.22	–0.25	0.13
Alcohol-related problems	–0.51	0.61	–0.31
Illicit drug use	–0.16	0.22	–0.11

Note: For measures of alcohol use, PDA = percent days abstinent for the 6 months after treatment (months 4–9); DDD = drinks per drinking day for the 6 months after treatment (months 4–9); first drink = the number of days between first therapy session and the first reported use of any alcohol. For measures of general functioning, BDI = Beck Depression Inventory; PFI = Psychosocial Functioning Inventory.

parentheses in table 1 show the probability values associated with contrasting parallel correlations between the two samples. For example, the correlation between percent days abstinent (PDA) and the Beck Depression Inventory (Beck et al. 1961) score was –0.31 for the aftercare sample and –0.29 for the outpatient sample. The question posed by statistically contrasting these two correlations is whether the observed difference in their magnitude reflects simple sampling and measurement error or "true" differences in the strength of the relationship between abstinence and depression. The probability value of 0.34 indicates that the magnitude of the two correlations is relatively equivalent (e.g., stable) between the aftercare and outpatient samples.

No between-sample differences were found in the magnitude of relationships between PDA and the five measures of client functioning. In contrast, in the aftercare sample there was a significantly stronger relationship between drinks per drinking day (DDD) and alcohol-related consequences relative to outpatient clients, whereas outpatient clients reported a significantly stronger and positive relationship between DDD and illicit drug use relative to the aftercare sample. Finally, somewhat consistent sample differences (three of five tests) were found using the number-of-days-to-first-drink measure. Significantly stronger negative correlations between days to relapse and increased alcohol-related consequences and depression were reported in the aftercare sample relative to the outpatient sample.

CONCEPTUAL CONSIDERATIONS IN MEASURING BEHAVIOR CHANGE OVER TIME

The decision of what to measure followed by the selection of a reliable instrument are important steps in conducting outcome evaluation. This section addresses the equally salient topic of determining when to administer an assessment,

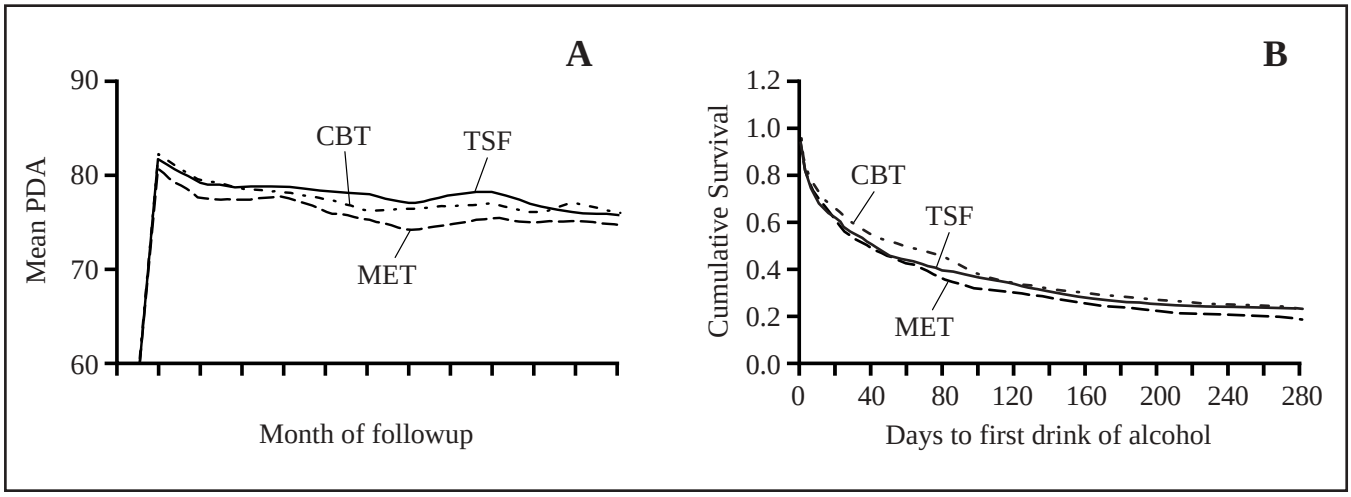
taking into account that changes in domains of individual functioning tend to occur at different rates after treatment (and with different patterns). The discussion that follows is based on findings of many studies of alcohol treatment–seeking adults, and it is important to emphasize that the measurement patterns described here may differ somewhat or a great deal from other populations of alcohol users, such as adolescents, treatment-resistant persons, and persons in natural recovery. With this caveat, a relatively common pattern of treatment outcomes across three domains of functioning can be described. First, typically the largest reduction in severity of alcohol-related problems will occur in the first 3 months after recruitment, a time during the delivery of the intervention. Only modest group changes in severity of alcohol problems, however, tend to be observed after this initial improvement. Counterintuitively, severity of medical problems tends to increase with reductions in alcohol severity and then begin to decline at the 6-month assessment (positively quadratic relationship). Legal problems, on the other hand, tend to be the most severe at baseline, decline to the 6-month assessment, and then begin to rise again (negative quadratic relationship).

Clearly, *when* an outcome is measured can be as important a decision as *what* is measured. In the case of severity of medical problems, for example, evaluation of pre-post changes using intake and 6-month data would lead to the erroneous conclusion that the intervention led to an increase in medical severity. Of course, the clinical interpretation is that with reductions in alcohol use persons begin to attend to acute and long-standing medical problems, both related and unrelated to alcohol use. This behavior appears to peak 6 months after treatment and then subsides.

Demonstration of treatment effectiveness based on drinking reductions over time may appear relatively straightforward. Such is not the case. Measures of alcohol use can offer alternative perspectives of treatment potency over time and, as such, can lead to conflicting conclusions about the relative effectiveness of treatment. As an example, figure 1 presents Project MATCH client outcome for 12 months after study recruitment using two oppositional measures of alcohol use: (1) mean PDA in monthly intervals (positive outcome) and (2) number of days of abstinence until relapse occurs as defined by taking one or more drinks between the first therapy session and the following 100 days (negative outcome). Panel A shows that significant gains in monthly abstinence rates were obtained in each treatment group, with an overall pre-post change in PDA between recruitment and 6-month followup of more than 100 percent (31 percent vs. 78 percent, effect size = 1.66). In contrast, the time-to-event analysis in panel B suggests that fully 75 percent of all clients had at least one drink of alcohol between the first therapy session and the following 100 days. The Pearson correlation between days to first drink and days to first heavy drinking day (six or more drinks at one time) was 0.81, suggesting that, for the 75 percent of the clients who did consume alcohol, the two events were the same or temporally close in time.

An even more complex and subtle picture arises when judging the relative effectiveness of alcohol treatments over time using alternative measures of alcohol use. Figure 1 shows that the 12-step facilitation therapy (TSF) group reported the highest *mean* rate of abstinence over 12 months, but cognitive-behavioral therapy (CBT) clients reported modestly fewer instances of relapse relative to TSF and motivational enhancement therapy (MET) clients during this same period. Alcohol use measures depicting the virtues of MET have also been identified. The question faced by an evaluator is, Which is the superior alcohol treatment, CBT, MET, or TSF? This dilemma highlights one of the fundamental measurement challenges facing treatment outcome evaluators. By design, treatments are generally qualitatively different, each having a unique orien-

FIGURE 1.—Project MATCH client outcome for aftercare and outpatient samples.



Note: (A) Percent days abstinent (PDA) by treatment assignment. (B) Survival analysis by treatment assignment. CBT = cognitive-behavioral therapy; MET = motivational enhancement therapy; TSF = 12-step facilitation therapy.

tation and strategy. While the abstract goal of treatments may be concordant, alcohol use measures are differentially sensitive to the active ingredients of a particular treatment. Such differential sensitivity can reflect, over time, different patterns of treatment outcome. Thus, TSF with its strong emphasis on total abstinence may appear most effective judged by overall, monthly abstinence rates, whereas CBT skill training in stressing recognition of personal "triggers" for alcohol use may differentially offset the initial use of alcohol.

Although consensus has yet to emerge on how to resolve this issue, three strategies are offered, each of which has distinct advantages and limitations:

- Develop a specific and narrow definition of treatment effectiveness, one that all treatments are intended to directly impact. Effectiveness may be determined by a single outcome measure, but qualitative differences among treatment approaches must necessarily be restricted.
- Apply multiple and oppositional measures to determine treatment effectiveness, acknowledging that, in all likelihood, all-purpose effectiveness cannot be demonstrated. This approach allows for unrestricted qualitative differences among treatments, but at the expense of interpretative clarity
- Characterize treatment effect in multidimensional terms, jointly and statistically considering multiple measures of outcome at one time.

MEANINGFUL CHANGES IN DRINKING BEHAVIOR

Satisfactorily addressing the inherent tension of comparing qualitatively different treatments using the same outcome measure(s), the evaluator then relies on inferential testing to assess the probability that observed treatment differences represent chance fluctuation. The clinician, too, is faced with this question, but does so considering the individual as the unit of analysis. Specific recommendations are made in this section to aid clinicians and researchers in making this determination.

Recommendations for Clinicians

At least three methods can be used to assess whether individuals demonstrate meaningful improvement in alcohol-related problems. The most obvious, of course, is the determination of whether clients achieve and can maintain treatment objectives. To make this determination, it is recommended that posttreatment assessment be done by an independent interviewer, and that the assessment be conducted at least 3 months after the cessation of treatment. Although it may not be feasible to have independent interviewers, such a practice is desired.

A second approach can be followed when assessment tools have published normative data. Clinicians can index individual pre-post scores to a normative sample, noting the extent of change in deciles, quartiles, and the like between pre- and posttest scores. With this approach, meaningful changes can be defined in relative terms (intraindividual) or in terms of a predetermined normative cutoff value (interindividual).

The third method distinguishes nonmeaningful and meaningful change, and its rationale draws on the earlier discussion of standard error of measurement. Pre-post changes in an individual's score that do not exceed the reported standard error of an instrument should be regarded as nonmeaningful changes. In this case it is uncertain whether observed pre-post changes reflect actual change in behavior or just error in measurement. In contrast, pre-post score changes that are at least 2 times the standard error of an instrument exceed measurement error substantially and also represent considerable improvement in functioning for an individual (95 percent).

Recommendations for Researchers

Rejection of the null hypothesis is a necessary but not sufficient condition to declare a meaningful effect. Blithely declaring meaningfulness because of rejection of the null hypothesis ignores the basic fact that as sample size increases the magnitude of effect required to reject the null hypothesis decreases. With large samples, woefully small effects can be reliably detected, but they may have little clinical meaning. In addition, while efforts to control for an inflated type I error rate (rejection of a true null hypothesis) ought to be applauded, these procedures only maintain a nominal alpha level (e.g., 0.05) and do not speak at all to the question of meaningfulness.

Measures of effect size should be routinely computed and reported beside the results of significance tests. They are crucial for a determination of the magnitude of an observed effect, and they can be reported in a variety of forms, such as variance accounted for or magnitude in mean difference. Several excellent texts in the areas of meta-analysis (e.g., Hunter et al. 1982; Hedges and Olkin 1985) and power analysis (e.g., Cohen 1988) are available to assist researchers in the calculation of effect sizes, and many of the major statistical software packages now offer the option to report measures of effect sizes along with inferential tests (e.g., SPSSpc and SAS). Finally, specialized software is now available—free of charge on the Internet—to correct effect sizes for small-sample bias and to assess whether effect size distributions are estimates of a single parameter.

Exact guidelines for what constitutes a large or meaningful effect is specific to an area of study and consideration of the costs involved in producing the effect. Small effect sizes associated with minimal costs, for example, may be considered meaningful from a public policy perspective, while moderate to large effect sizes requiring huge financial expenditures to be produced may be considered less meaningful. The important point regarding this cost-benefit definition of meaningfulness is that scientists have the responsibility to describe benefit in a systematic fashion that facilitates comparison across treatment approaches.

Practical Considerations in Measuring Behavior Change Over Time

This section reviews some practical aspects of outcome evaluation. In essence, a laundry list of considerations is presented, ranging from the importance of collecting representative baseline data to problems associated with using different versions of the same assessment over the course of a study.

Representative Baseline

For meaningful analysis of change, it is imperative that comparable pre- and posttreatment measures be collected. In fact, the importance of a detailed account of the effect of client pretreatment characteristics on severity measures cannot be overemphasized. Without such information, judgment of improvement following treatment is, at best, difficult. Detailed pretreatment assessment also allows for the search for prognostic indicators of outcome, some of which may be as powerful predictors of outcome as the treatment experience itself. Description of pretreatment drinking should take into account the nature of consumption of a clinical population and how consumption may vary in proximity to presentation for treatment. Adolescents, for example, tend to drink infrequently but at high intensity levels (e.g., binge). In this case a quantity-frequency (QF) measure may significantly underestimate salient drinking factors and, in the case of a typical 30-day assessment window, fail to characterize the full profile of drinking. In contrast, a QF measure may be appropriate for clinical populations characterized by steady drinking patterns over sustained periods of time. There is some evidence that client drinking immediately before presentation for treatment does not accurately mirror *typical* drinking. It is recommended, therefore, that assessment of pretreatment drinking elicit information for at least the 90 days prior to treatment. The chapter by Sobell and Sobell in this *Guide* highlights several advantages and disadvantages of particular consumption measures and selection of a pre-post drinking measure.

Client attrition during and after treatment is an unfortunate fact in outcome evaluation. Detailed measurement of alcohol consumption at pretreatment is essential for understanding how, if at all, such attrition may bias study findings. Typically, attrition (yes/no) is crossed with treatment assignment via a chi-square test to assess whether attrition was random or systematically related to the kind of treatment offered. This is an important first step, but it does not address whether *severity* of alcohol-related problems (at intake) was prognostic of attrition, which (if this is the case) can have serious consequences for study internal and external validity. Two analyses can investigate these potential biases, both of which rely on detailed pretreatment measurement of alcohol consumption. Attrition can bias the external validity of a study when more (or less) severe clients systematically drop out, disregarding group assignment. The nature of the sample recruited and the nature of the sample actually available for outcome analyses differ, with the net effect that study findings may not generalize to the intended population. Logistic regression and discriminant function analyses with attrition status as the dependent measure (yes/no) and alcohol severity measures as predictors are two techniques especially well suited to investigate this threat to external validity. In comparative studies, internal validity can be compromised when more (or less) severe clients systematically drop out of one treatment. In this situation, the sheer number of dropouts may (or may not) be relatively equivalent between treatments, but factors predicting attrition differ by treatment condition. Causal statements about the relative effectiveness of the treatments can become problematic under this condition.

Two considerations should guide pretreatment assessment of nondrinking severity characteris-

tics. First, is assessment of this characteristic distorted by recent drinking? Failure to take this type of problem into account may result in erroneous conclusions about client posttreatment improvement. For example, depression (e.g., as measured by the Beck Depression Inventory score) tends to be artificially elevated in conjunction with heavy drinking, whereas measures of cognitive functioning (e.g., as measured by the Trail Making Tests Forms A and B) tend to be underestimated following heavy drinking. Confounded assessment of these domains and subsequent comparison with posttreatment measures may lead to the conclusion that treatment favorably reduced depression and increased cognitive functioning. A second consideration in pretreatment measurement involves selection of an appropriate timeframe for assessment. In cases where an event has a low probability of occurrence, it is important that pretreatment assessment sample a longer period of time. Examples of domains that may require longer timeframes are legal, health care utilization, and employment.

Assessment Order Effects

This section highlights issues raised when assessing multiple domains by integrating individual instruments. Although these concerns more often arise in research assessment lasting several hours, they may also apply to relatively short assessment protocols conducted for the purpose of case management. Described by Connors et al. (1994), care should be exercised in the use and sequencing of assessment batteries to take into account potential assessment order effects.

Assessment order effect refers to the influence that answering one set of questions has on answers to the next set of questions. Frequently, the effect of answering the first set of questions is referred to as *priming*. To illustrate these carryover effects, imagine that a clinician is interested in the relationship between posttreatment drinking (QF) and involvement in self-help programs (e.g., AA). Three months after cessation of treatment he or she contacts clients and routinely administers first the self-help and then the QF questions. It seems likely that those clients invested in AA but who are also drinking may underreport drinking. One method to eliminate potential order effects is to rotate the sequence of assessment instruments. The advantage of controlling for order effects, however, should be balanced with the need—at times—for an integrated assessment process wherein one assessment naturally leads to subsequent questions.

Interviewer Role and Training

This section addresses who ought to conduct followup interviews and what skills are important for collecting reliable and valid measurements. The recommendation of who ought to conduct followup interviews hinges, in part, on the purpose of evaluation. When followup is conducted in the formative context with the assumption that followup assessment has therapeutic benefit, a strong case can be made that either the client's therapist or a trained interviewer can collect reliable and valid data. In the case of summative evaluation, however, there are compelling reasons for therapists not to conduct followup interviews. Interviewers in summative evaluation should be blind to the type of treatment clients received so that the measures are not unintentionally biased.

Given appropriate matching of organizational role and purpose of evaluation, the importance of adequate interviewer training cannot be overemphasized. In the case of structured interviews (e.g., Addiction Severity Index, Alcohol Timeline Followback, and Form 90), interviewer training should consist of several modules that sequentially train to a predetermined standard of accuracy and then monitor for interviewer "drift" across the course of the evaluation. As an example of the training sequence, initial training may consist of observing a videotape of an interview. Standard probes to

ambiguous client responses are modeled, and trainees can be debriefed about the intent of the interview. Again using videotape, trainees can then observe and code the instrument as a model interview is conducted. Comparisons can be made among the trainees to discern why trainees may have scored a particular item differently. This procedure facilitates standardization in scoring among interviewers. When trainees can confidently master these steps, they perform a videotaped interview. Along with the hard-copy assessment instrument, this tape is reviewed and approved by the trainer before the trainee is certified to conduct actual followup assessments. Periodically, the trainer may choose to observe interviewers to ensure that the protocol is maintained or, when feasible, review videotaped interviews with interviewers to highlight strengths and weaknesses in an assessment.

A final reason for adequate training of interviewers is personnel turnover during the progress of an outcome evaluation. Research assistants and therapists tend to migrate to other jobs. Ironically, such turnover is often used as justification not to invest in training when, in fact, training should be even more intensive to maintain the integrity of assessment. It is acknowledged that the training sequence described is an ideal and may be difficult to follow with limited resources in field settings. Approximations to this ideal, however, will enhance the reliability of assessment significantly and thus increase the sensitivity of the outcome evaluation to detect relationships of interest.

Instrument Consistency

There are several possible explanations for the use of different versions of the same assessment instrument in a single evaluation study: changes in item content in copyrighted instruments during the course of the trial (items under test development get dropped and new items are included), duplication errors in photocopying, and miscommunication among interviewers about which version is to be used (this is especially likely when assessment is conducted at multiple sites). Regardless of the reason for lack of consistency in instrument use, the result, unfortunately, is that valuable information is lost or never collected for some clients.

When feasible, this problem can be minimized by preparing all client followup assessment packets in advance. Advance packaging enables rotation of self-assessment instruments to minimize systematic order effects, as well as ensuring identical assessments for all clients.

Data Entry

It is unfortunate that so little attention is given to the integrity of data entry procedures. In addictions research, it is not uncommon to hear of data entry keystroke errors in the range of 5 to 8 percent. In such cases, keystroke error may account for more error variance than interviewers. It is highly recommended that all data, and especially data pertaining to the central outcome measures, be double entered and verified. Many software packages are specifically designed for data entry (e.g., SAS and SPSSx). These packages have the advantage of defining out-of-range values in advance as well as defining Boolean functions to eliminate inconsistent responses across items. Although direct entry of data into spreadsheets for analyses or entry into word processing packages to be ASCII filed for later use in a statistical software package may be necessary because of limited resources, these practices are discouraged.

Summary

This chapter reviewed selected theoretical and applied issues in conducting alcohol treatment outcome evaluation. A strong case was made for the use of measures with demonstrated reliability, and examples of commonly reported reliability statistics were provided to assist readers in the evaluation and selection of assessments included

in this *Guide*. A general theme in the chapter was that the effectiveness of a treatment ought not be judged on the basis of a single measure of drinking collected at an arbitrary point after alcohol treatment. Different measures of alcohol use provide alternative perspectives of treatment effectiveness, and measures of general functioning may not correlate highly with changes in drinking. Illustrations were offered to show that the issue is made more complex because the topography of change across time differs between domains of interest. One of the most challenging aspects of outcome evaluation is the communication of findings to policymakers, treatment providers, and the scientific community. Here, the meaningfulness of findings becomes a primary consideration, and several strategies were presented to aid the clinician and evaluator in making this determination.

REFERENCES

Aiken, L.R. *Psychological Testing and Assessment*. 10th ed. Boston: Allyn & Bacon, 2000.

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised*. Washington, DC: the Association, 1987.

Babor, T.F. Reliability of the Ethanol Dependence Syndrome scale. *Psychol Addict Behav* 10(2):97–103, 1996.

Babor, T.F.; Dolinsky, Z.; Rounsaville, B.; and Jaffe, J. Unitary versus multidimensional models of alcoholism treatment outcome: An empirical study, *J Stud Alcohol* 49:167–177, 1988.

Beck, A.T.; Ward, C.H.; Mendelson, M.; Mock, J.; and Erbaugh, J. An inventory for measuring depression. *Arch Gen Psychiatry* 4:561–571, 1961.

Carmines, E.G., and Zeller, R.A. *Reliability and Validity Assessment*. Newbury Park, CA: Sage, 1979.

Cicchetti, D.V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Special section: Normative assessment *Psychol Assess* 6:284–290, 1994.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2d ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

Connors, G.J.; Allen, J.; Cooney, N.L.; DiClemente, C.C.; Tonigan, J.S.; and Anton, R. Assessment issues and strategies in alcoholism treatment matching research. *J Stud Alcohol Suppl* 12: 92–100, 1994.

Cook, T.D., and Campbell, D.T. *Quasi-Experimentation: Design and Analysis for Field Settings*. Chicago: Rand McNally College Publishing, 1979.

Emrick, C.D. A review of psychologically oriented treatment of alcoholism: I. The use and interrelationships of outcome criteria and drinking behavior following treatment. *Q J Stud Alcohol* 35:523–549, 1974.

Hedges, L.V., and Olkin, I. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press, 1985.

Hunter, J.E.; Schmidt, F.L.; and Jackson, G.B. *Meta-Analysis: Cumulating Research Findings Across Studies*. Beverly Hills, CA: Sage Publications, 1982.

Longabaugh, R.; Mattson, M.E.; Connors, G.J.; and Cooney, N.L. Quality of life as an outcome variable in alcoholism treatment research. *J Stud Alcohol Suppl* 12:119–129, 1994.

Miller, W.R.; Brown, J.M.; Simpson, T.S.; Handmaker, N.S.; Bien, T.H.; Luckie, L.F.; Montgomery, H.A.; Hester, R.K.; and Tonigan, J.S. What works? A methodological analysis of the alcohol treatment literature. In: Hester, R.K., and Miller, W.R., eds. *Handbook of Alcoholism Treatment Approaches*. 2d ed. Needham Heights, MA: Allyn & Bacon, 1995*a*.

Miller, W.R.; Tonigan, J.S.; and Longabaugh, R. *The Drinker Inventory of Consequences (DrInC): An Instrument for Assessing Adverse Consequences of Alcohol Abuse*. Project MATCH Monograph Series, Vol. 4. DHHS Pub. No. 95–3911. Rockville, MD: National Institute on Alcohol Abuse and Alcoholism, 1995*b*.

Pedhazur, E.J. *Multiple Regression in Behavioral Research: Explanation and Prediction*. 2d ed. New York: Holt, Rinehart, & Winston, 1982.

Project MATCH Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. *J Stud Alcohol* 58(1):7–29, 1997.

Project MATCH Research Group. Matching alcoholism treatments to client heterogeneity: Project MATCH three-year drinking outcomes. *Alcohol Clin Exp Res* 22:1300–1311, 1998.

Skinner, H.A., and Horn, J.L. *Alcohol Dependence Scale: Users Guide*. Toronto: Addiction Research Foundation, 1984.

Tonigan, J.S.; Miller, W.R.; and Brown, J.M. The reliability of Form 90: An instrument for assessing alcohol treatment outcome. *J Stud Alcohol* 58:358–364, 1997.