The Reliability of Form 90: An Instrument for Assessing Alcohol Treatment Outcome*

J. SCOTT TONIGAN, PH.D., WILLIAM R. MILLER, PH.D., AND JANICE M. BROWN, PH.D.†

Department of Psychology & Center on Alcoholism, Substance Abuse and Addictions (CASAA), University of New Mexico, 2350 Alamo SE, Albuquerque, New Mexico 87131-1161

ABSTRACT. *Objective:* Project MATCH is a randomized clinical trial consisting of five outpatient and five aftercare units at nine sites. Of importance in this multisite trial examining the efficacy of client-treatment matching was the cross- and within-site reliability of the structured interview used to assess alcohol treatment outcomes, the Form 90. Evaluation of the reliability of Form 90 is the subject of this article. *Method:* The reliability of Form 90 was evaluated in two test-retest studies. The cross-site reliability study consisted of 70 paired test-retest interviews conducted by different interviewers. Clients for this study were recruited from inpatient, outpatient and college settings. The within-site reliability study had a total of 108 paired test-retest interviews, with 54 of the retests conducted by different interviewers and 54 by the same interviewer. Clients for this study were most often presenting for alcohol treatment at the nine sties and were selected to be representative of the larger Project MATCH sample. *Results:* Good-to-excellent reliability was found for all key summary measures of alcohol consumption and psychosocial functioning, and most frequently used illicit drugs had moderate reliability. No decay in consistency of self-reported drinking was found at more distal points from dates of test-retest interviews. Application of 68% confidence intervals for primary alcohol consumption measures suggests that trained researchers and clinicians can obtain consistent information regarding client drinking. *Conclusions:* Form 90 appears to be a reliable instrument for alcohol treatment assessment research when interviewers have received careful training and supervision in its use. (*J. Stud. Alcohol* **58:** 358-364, 1997)

ALCOHOL CONSUMPTION is a primary domain of dependent variables in the assessment of alcohol treatment outcomes (Litten and Allen, 1992). A variety of methods have been used to quantify drinking, including prospective self-monitoring, quantity-frequency questions, calendar-based timeline reconstruction, and retrospective grids representing a typical period (e.g., week) of drinking (Cervantes et al., 1994; Miller and Del Boca, 1994). Each of these methods has its advantages and drawbacks, and no single approach has emerged as the definitive strategy for outcome assessment.

The Project MATCH Research Group (1993) confronted this issue when selecting instrumentation to measure this critical outcome domain. The result of deliberations among this group of 23 senior investigators was an attempt to combine the strengths of prior approaches by creating a new family of structured interview instruments (Miller and Del Boca, 1994). The core interview protocol was named "Form 90" in part because of the number of assessment instruments being considered for the trial and in part because the baseline interview focused on reconstruction of drinking during a 90-day window.

The intake interview (90-AI) reconstructs daily alcohol consumption during the 90 days prior to the client's most recent drink. Thus the baseline assessment window is of variable length, with all days between the most recent drink and the interview recorded as abstinent days. The follow-up Form 90-AF reconstructs the period from the last day covered by the prior interview (90-AI or 90-AF) up to the day before the current interview. The intended result is a continuous calendar from baseline through the last day of follow-up (Miller and Del Boca, 1994). A detailed test manual, alternative interview protocols and supporting software are available elsewhere (Miller, 1996).

Form 90 combines two previously published methods for assessing alcohol consumption. A calendar base is used to ensure a continuous record for each day in the assessment period, in the manner of the timeline follow-back method (TLFB) (Sobell and Sobell, 1992). Because drinking patterns often manifest consistency from week to week or from episode to episode, a grid averaging method (Miller and Marlatt, 1984) was incorporated to capture efficiently such consistent patterns when they occur, inserting them into appropriate sections of the calendar. The result, as desired by the Project MATCH Research Group, is a flexible interviewing system that can be adapted to drinking styles that vary from a steady and predictable pattern (relying heavily on the grid method) to wholly idiosyncratic consumption (using day-by-day TLFB reconstruction). Although psychometric characteristics of both timeline and grid averaging methods have been documented (e.g., Miller et al., 1992; Sobell et al.,

Received: March 4, 1996. Revision: April 23, 1996.

*This report was supported in part by National Institute on Alcohol Abuse and Alcoholism grants K05-AA00133 and U10-AA08435.

†Janice M. Brown is with the Center for Alcohol and Drug Related Studies, Oklahoma City, Okla.

1988), the combination of these strategies represents a hybrid method for reconstructing alcohol consumption in treatment outcome research and requires evaluation.

In addition to reconstructing alcohol consumption, Form 90 contains structured questions for quantifying several related domains. Days of institutionalization, including incarceration and residential treatment, are recorded on the calendar. Residential status (place of abode), health care utilization (for medical, alcohol, drug and mental health care) and 12-step group participation are recorded by days during the assessment window. Days of engagement in employment, education and religious activities are also tabulated. Finally, Form 90 queries the number of days of use of other drugs during the assessment window, both for prescription medications and for illicit drugs. In the baseline (90-AI) version, lifetime duration of use is quantified by estimating for each drug class the total number of weeks during which the respondent used the drug at least once.

Method

Two studies were conducted to determine the psychometric characteristics of core instruments being used in a multisite clinical trial of treatment for alcohol-related problems (Project MATCH Research Group, 1993). A *cross-site* reliability study was conducted to determine if site variation in client self-report of drinking may be a function of between-site interviewer inconsistencies. Here, nine experienced interviewers (one from each MATCH site) interviewed a total of 82 heavy-drinking participants drawn from clinical and college populations. A *within-site* reliability study was also conducted wherein pairs of interviewers from each of the nine sites independently interviewed 6 clients (per site) using the baseline Form 90 instrument. All nine of the interviewers from the cross-site study also participated in the within-site reliability study. The objective of this study was to assess the consistency (reliability) of interviewers within sites. When each participant completed the interview in the within-site study, a debriefing session was conducted in which coordinating staff sought to discern the reasons for discrepancies between the two interviews (cross-site and within-site). Alcohol consumption was reported in standard drink units equal to 0.5 oz (15 ml) of absolute ethanol (Miller et al., 1991). More complete details of the methodology of the two reliability studies are reported by Del Boca et al. (in press).

Interviewers

Project MATCH included both outpatient and aftercare sites. For both reliability studies, outpatient and aftercare interviewer pairs were initially analyzed separately to evaluate the effects of interviewer experience with populations of different severity. As an example of how clients assigned to outpatient and aftercare interviewers differed in the two studies, on the *test* assessment of the cross-site study mean drinks per drinking day (DPD) was 20.1 for clients assigned to interviewers from aftercare sites and was 10.2 for those assigned to outpatient interviewers. In the within-site study, DPD for clients assessed by aftercare interviewers was 20.1 and DPD for clients interviewed by outpatient interviewers was 18.3.

The combined interviewer groups ($N = 18$) were 38 years of age, on average, with 19 years of education; just over half had prior experience as alcohol/drug counselors, and slightly fewer than half had participated in previous alcohol/drug studies.

Participants

In the within-site study, a total of 54 clients (six from each site) were tested by one interviewer and retested by a different interviewer, and 54 others were tested and retested by the same interviewer. Of these 108 clients, 84 (78%) were male. A majority (70%) were of white non-Hispanic origin. They reported a mean of 12.9 years of education, an average age of 39.4 years and 2.4 previous treatments for alcohol problems.

For the cross-site study, 10 of 82 participants were intentionally retested by the same interviewer (see Del Boca et al., in press), and in two cases the participant did not return for the second (retest) session. Thus 70 participants (57 male) are included in these analyses. These 70 participants reported a mean age of 31.2 years. Many had been treated previously for alcohol-related problems (mean of 1.7 times). By self-identified ethnicity, 44 were white non-Hispanic, 15 Hispanic, 2 black, 7 Native American and 2 of other identification.

Reliability analyses

Several approaches were used to evaluate the reliability of Form 90 interview data. Comparisons between test and retest administrations of the Form 90 are provided using both intraclass (ICC) and Pearson product-moment (r) correlation coefficients. The ICC calculations follow formula 2.1 summarized by Shrout and Fleiss (1979). It is important to highlight how ICC and r provide different perspectives of assessment 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 upon r are available and generally accepted (e.g., Cohen, 1988). There is less agreement, however, about interpretation of reliability when computing ICCs. Cicchetti (1994) has recommended the following ranges to interpret the reliability of clinical instruments when ICCs are evaluated: below .40 = poor, .40 to .59 = fair,

.60 to .74 = good, and .75 to 1.00 = excellent. These ICC interpretations were applied in this study.

Kappa coefficients were used to determine the extent of test-retest interviewer agreement on the presence of a discrete activity (e.g., use of illicit drugs). This index was selected because it partials out that portion of test-retest interviewer agreement based upon chance alone. Finally, for clinical purposes, we have provided 68% confidence intervals for central alcohol consumption measures. Based upon sample estimates of the standard error of the mean (SE), these confidence bands inform clinicians of the expected imprecision in measurement from Form 90 interviews.

Results

For purposes of analysis, variables collected during the Form 90 interview were divided into four classes: (1) measures of general adjustment; (2) indices of alcohol use; (3) measures of *lifetime* illicit drug use, separated into 11 drug classes; and (4) measures of *recent* illicit drug use (during the assessment window) in the same 11 categories. Detailed definitions for each of these variables are provided in the Form 90 administration manual (Miller, 1996).

Construction of measures within the four classes was straightforward. Each of the seven general functioning measures indicated the number of days an event had occurred

TABLE 1. Form 90 test-retest reliability coefficients for four groups of variables

	Within-site				Cross-site	
	Same interviewer (n = 54)		Different interviewers (n = 54)		Different interviewers (n = 70)	
Variables	r	ICC	r	ICC	r	ICC
General functioning						
Days worked for pay	.98	.98	.85	.53	.98	.90
Days in school	.99	.99	.93	.61	.94	.76
Days in own residence	.99	.99	.74	.31	.75	.41
Days religious attendance	.98	.97	.79	.40	.96	.82
Days medical care	.99	.99	.93	.62	.91	.69
Days psych. treatment	.99	.98	.63	.19	.81	.43
Days AA attendance	.62	.59	.92	.62	.87	.53
Alcohol use						
Total consumption	.96	.92	.91	.61	.97	.82
Drinks per drinking day	.93	.89	.88	.55	.95	.71
Percent abstinent days	.98	.97	.96	.76	.96	.85
Percent heavy days	.98	.95	.92	.60	.97	.89

	r	ICC	Kap.	r	ICC	Kap.	r	ICC	Kap.
Lifetime illicit drug use									
Tobacco	.97	.97	.79	.95	.81	.66	.94	.94	.85
Marijuana	.92	.91	1.0	.93	.66	.74	.66	.66	1.0
Hallucinogens	.82	.76	–[a]	.71	.25	.93	.68	.55	1.0
Stimulants	.47	.45	.81	.84	.35	.81	.95	.95	.97
Cocaine	.80	.75	1.0	.86	.46	.96	.77	.68	1.0
Tranquilizers	.90	.82	.93	.89	.49	.96	.75	.69	.84
Sedatives	.95	.95	.89	.98	.92	.89	.92	.88	.92
Steroids	.91	.90	.79	.30	.04	.66	.97	.96	1.0
Opiates	.93	.91	.93	.06	.02	.78	.69	.68	.82
Inhalants	.94	.85	1.0	.99	.99	.92	.94	.74	.92
Other drugs	–	–	–	–	–	–	–	–	–
Recent illicit drug use									
Tobacco	.97	.94	.93	.96	.99	.94	.91	.90	.84
Marijuana	.98	.97	.89	.96	.90	1.0	.71	.69	.81
Hallucinogens	–	–	–	.99	.68	1.0	.99	.98	1.0
Stimulants	.99	.99	1.0	.93	.52	.66	.94	.86	.88
Cocaine	.99	.98	1.0	.91	.84	.88	.99	.98	.55
Tranquilizers	.97	.96	.85	.18	.04	.24	.38	.33	.17
Sedatives	–	–	–	.99	.98	1.0	–	–	–
Steroids	–	–	–	–	–	–	–	–	–
Opiates	.99	.99	.91	.99	.94	.85	.37	.25	.74
Inhalants	–	–	–	–	–	–	–	–	–
Other drugs	–	–	–	–	–	–	–	–	–

[a]Rates were too low to calculate these reliability coefficients. Kappa coefficients reflect agreement regarding the presence or absence of use.

(e.g., days in school). Measures of alcohol use represented the total number of standard drink units reported during the assessment period (total consumption), number of standard drink units consumed on a given day (drinks per drinking day), number of abstinent days divided by the total number of days in the assessment period (percent abstinent days) and number of days of heavy drinking (six standard drinks for men and four standard drinks for women) divided by the total number of days in the assessment window (percent heavy days). Finally, lifetime illicit drug use was measured as the number of weeks that a drug had been used at least once during a week, and recent illicit drug use reflected the reported number of days of drug use during the assessment period.

Table 1 presents the Pearson product-moment and intraclass correlations for each of these four groups of variables. The first column presents results from the within-site study when respondents were tested by *the same interviewers.* The second column provides reliability coefficients for these four variable sets when within-site study participants were tested by *different interviewers* from the same clinical site. The third column provides reliability coefficients from the cross-site study, where respondents were tested by different interviewers from different sites. Kappa coefficients were also calculated as indices of agreement between interviews with regard to the absolute presence or absence of use of each class of drugs (other than alcohol). Because very few participants reported using drugs other than those represented in the major drug classes, statistics could not be computed for the "other drugs" category.

Applying Cicchetti's (1994) classification scheme for those interviews conducted by different interviewers, two measures within the general functioning category had only poor-to-fair reliability (days in own residence and psychological treatment). The five other measures in this domain had at least one good reliability ICC, with three measures having excellent reliability in the cross-site study (days worked for pay, school, and religious attendance). Three of the four measures in the "alcohol use" category had good-to-excellent reliability (total consumption, percent abstinent days and percent heavy days). The fourth measure, drinks per drinking day, reflected greater variability and somewhat lower ICCs indicating *fair-to-good* reliability.

Response to alcoholism treatment is often evaluated across time. An important statistical assumption when simultaneously evaluating multiple follow-up data points is that the reliability of measurement is relatively constant as well as good. Figure 1 shows mean weekly alcohol

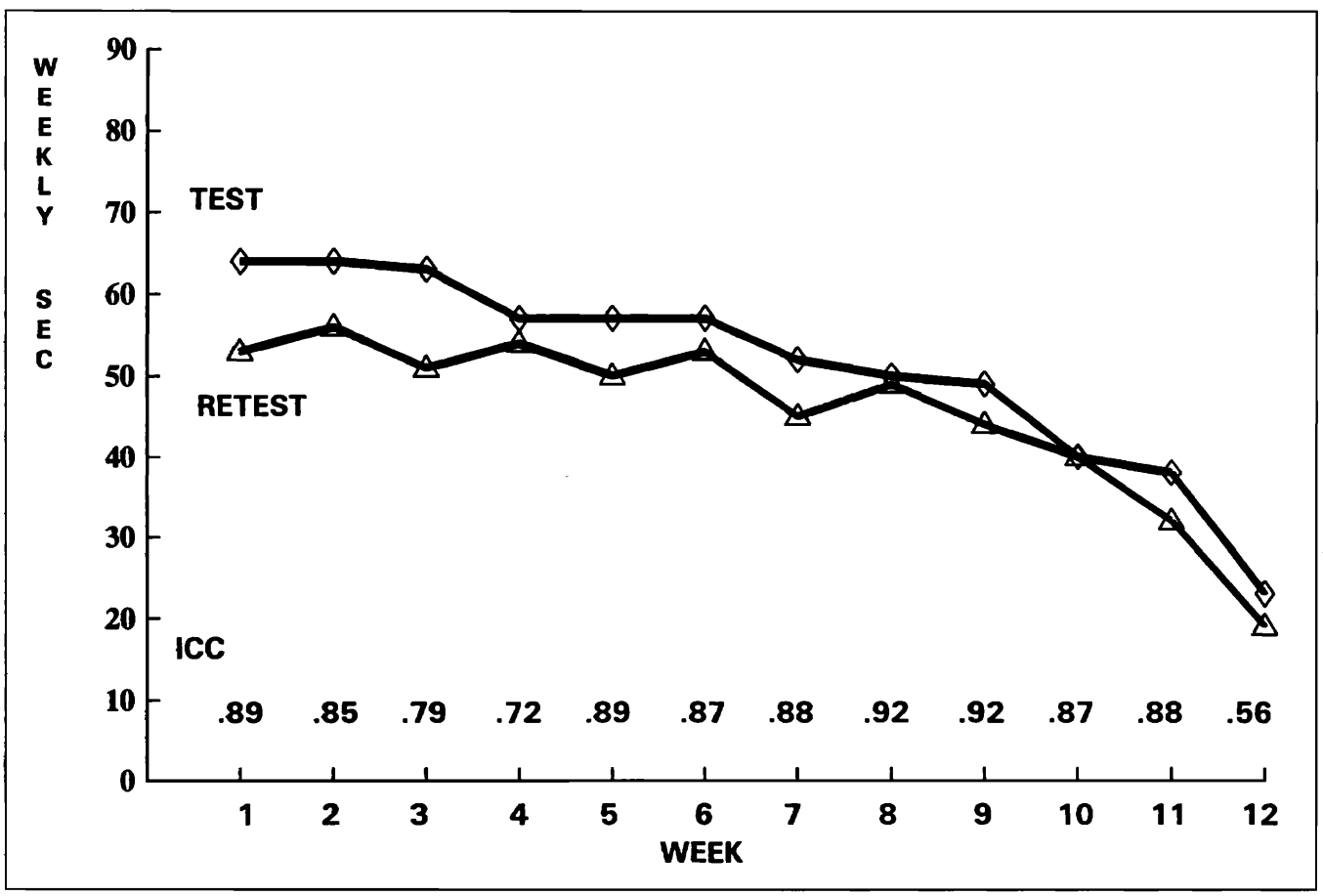

FIGURE 1. Form 90 cross-site reliability study: Mean weekly SEC by ICC and assessment week ($N = 70$).

TABLE 2. Test-retest means (±SD) for three interviewer samples on selected Form 90 measures: Most recent 90 days only

	Within-site		Cross-site
Variables	Same interviewer	Different interviewers	Different interviewers
Days worked for pay			
Test	29.3 ± 27.4	23.5 ± 27.4	33.1 ± 27.6
Retest	30.1 ± 27.5	24.0 ± 25.7	32.2 ± 28.0
Days in school			
Test	1.2 ± 7.1	1.1 ± 6.1	26.8 ± 27.3
Retest	1.2 ± 7.1	1.7 ± 7.9	27.6 ± 28.0
Days in own residence			
Test	63.4 ± 34.5	63.9 ± 31.9	55.3 ± 39.3
Retest	63.0 ± 34.2	63.7 ± 32.6	52.1 ± 40.4
Days religious attendance			
Test	1.8 ± 3.8	2.7 ± 5.7	2.6 ± 4.3
Retest	1.9 ± 3.7	2.7 ± 4.8	2.7 ± 4.3
Days medical care			
Test	1.5 ± 4.8	1.5 ± 3.8	1.6 ± 2.9
Retest	1.6 ± 5.2	1.5 ± 3.1	1.6 ± 2.8
Days psych. treatment			
Test	0.6 ± 2.3	1.3 ± 4.6	1.8 ± 5.1
Retest	0.7 ± 2.5	1.4 ± 3.7	2.1 ± 7.2
Days AA attendance			
Test	11.1 ± 17.1	10.3 ± 15.5	3.8 ± 9.2
Retest	10.4 ± 11.8	11.2 ± 17.7	6.1 ± 11.3‡
Total alcohol consumption			
Test	982 ± 931	1078 ± 958	614 ± 801
Retest	944 ± 957	1234 ± 1066*	559 ± 671*
Drinks per drinking day			
Test	18.2 ± 13.0	20.7 ± 13.5	14.1 ± 14.1
Retest	17.6 ± 13.2	22.1 ± 15.6*	13.2 ± 11.3
Percent abstinent days			
Test	50 ± 32	48 ± 31	56 ± 26
Retest	51 ± 31	46 ± 31	56 ± 27
Percent heavy days			
Test	46 ± 33	47 ± 33	32 ± 29
Retest	45 ± 32	47 ± 32	32 ± 29
Tobacco use days			
Test	16.4 ± 11.1	19.4 ± 13.9	40.4 ± 41.8
Retest	16.0 ± 11.5	19.0 ± 13.8	40.6 ± 42.0
Marijuana use days			
Test	3.3 ± 11.7	9.4 ± 20.4	3.8 ± 12.1
Retest	2.9 ± 12.0	9.0 ± 18.6	5.8 ± 15.5
Hallucinogen use days			
Test	0.0 ± 0.0	0.0 ± 0.1	0.1 ± 0.7
Retest	0.0 ± 0.0	0.0 ± 0.1	0.1 ± 0.6
Stimulant use days			
Test	0.8 ± 4.9	0.1 ± 0.7	0.9 ± 4.2
Retest	0.9 ± 5.1	0.1 ± 0.7	0.8 ± 4.1
Cocaine use days			
Test	4.9 ± 16.0	10.0 ± 20.0	0.4 ± 2.1
Retest	4.6 ± 15.8	9.8 ± 18.9	0.4 ± 2.0
Tranquilizer use days			
Test	2.3 ± 12.7	0.4 ± 2.0	0.7 ± 3.4
Retest	2.0 ± 12.3	0.4 ± 1.6	1.3 ± 5.9
Sedative use days			
Test	0.1 ± 1.0	1.7 ± 12.1	0.0 ± 0.0
Retest	0.0 ± 0.0	1.7 ± 12.1	0.0 ± 0.0
Steroid use days			
Test	2.8 ± 14.6	0.0 ± 0.0	0.9 ± 7.2
Retest	2.5 ± 13.6	0.0 ± 0.0	0.4 ± 2.9
Opiate use days			
Test	4.6 ± 18.1	3.3 ± 15.1	2.1 ± 11.0
Retest	4.3 ± 17.9	3.2 ± 15.1	0.7 ± 4.3
Inhalant use days			
Test	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Retest	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

*$p < .05$ test vs retest. ‡$p < .001$ for test vs retest.

consumption (Standard Ethanol Content; SEC) for test and retest administrations for the cross-site reliability study. In the figure, Week 1 corresponds to Days 1 through 7 (furthest days from date of the test-retest interviews), and the most recent week, Days 77 through 84, is represented by Week 12. ICCs were computed separately for each week to compare test-retest Form 90 agreement. As shown, 10 (83%) of the 12 ICCs indicated excellent reliability, and more distal client report of drinking (Weeks 1, 2 and 3) was surprisingly good, reflecting no decay of consistency for further-removed weeks within this range.

Relatively high test-retest agreement (kappa) was found for the presence or absence of lifetime illicit drug use (11 measures). Substantial variation in reliability for reported weeks of use of specific illicit drugs, however, was also obtained when test-retest interviews were conducted by different interviewers. More frequently used drugs (tobacco, marijuana and cocaine) had ICCs that ranged from fair-to-excellent, and only days of opiate use (very low occurrence in this sample) reflected consistently poor reliability. A similar pattern of reliability findings was obtained for measures within the recent illicit drug use category. Findings here may be affected by the reported low frequency of daily use of specific drugs. Days of tobacco, marijuana, cocaine, hallucinogens and stimulant use had, at a minimum, at least one ICC in the excellent range. Other measures of specific illicit drug use lacked even fair reliability (e.g., tranquilizers) while still other drugs had insufficient endorsement to evaluate reliability via the ICC. As noted above, in the Form 90-AI version tested, days of drug use were not reconstructed on a day-by-day basis using the calendar, but as an aggregate number for the entire retrospective period. Timeline reconstruction of each drug category may produce more reliable estimates, and this approach has been incorporated into the polydrug version Forms 90-DI and 90-DF (Miller, 1996).

To facilitate comparison, Table 2 provides means and standard deviations for the variables shown in Table 1. Of the 63 pairs of variables, test and retest means were significantly different ($p < .05$) on four tests. Of three such alcohol consumption comparisons, two reflected higher retest consumption (within site, different interviewers) and one reflected lower retest consumption (cross-site, different interviewers). Reported days of AA attendance were significantly higher at retest in the cross-site study only. By chance, 3.2 significant differences (Type I errors) would be expected among 63 contrasts with alpha set at $p < .05$.

An important issue is the consistency and, hence, confidence that may be placed on Form 90 assessments of alcohol consumption. Band interpretation, based on standard error of measurement (SE), is one method to illustrate the extent of measurement imprecision at the individual level. For this purpose, we developed regions in which an observed score would fall around its true value. It should be underscored that true score is used here in the psychometric sense, indicating that part of the observed score that is unaffected by random error. Table 3 illustrates the band within which an observed score will approximate a true score with 68% confidence ($1 \pm SE$). Entries for standard drinks per drinking day (DPD) in Table 3 indicate that the reported DPD for a client may vary as much as 1.84 above or below its true value. Applied to a hypothetical case where 12.5 DPD was reported, one can have 68% confidence that the actual DPD for this case ranged between 14.34 and 10.66.

TABLE 3. Standard error in measurement ($1 \pm SE$) for selected Form 90 alcohol consumption measures

Alcohol consumption measures	Within-site		Cross-site
	Same interviewer	Different interviewers	Different interviewers
Total drinks in 90 days	126.73	130.39	95.75
Drinks per drinking day	1.76	1.84	1.68
Percent abstinent days	0.04	0.04	0.03
Percent heavy days	0.05	0.04	0.03

Discussion

In two separate test-retest reliability samples, the Form 90 interview yielded relatively consistent outcome measures of drinking, illicit drug use and psychosocial functioning ($r \geq .90$ in 57 of 81 comparisons). For drinking outcome indices in particular (the variables that the Form 90 interview was originally constructed to measure), reliability was consistently excellent by a more liberal correlational standard (all r values $\geq .88$) and ranged from fair to excellent using a more conservative (ICC) standard of reliability. Comparison of weekly ICCs also indicated that the Form 90 is a reliable instrument to measure alcohol consumption retrospectively *across time.*

Most measures of general psychosocial functioning in the Form 90 also demonstrated satisfactory reliability, indicating that Form 90 can be used when outcome evaluation includes assessment of posttreatment functioning in the areas of employment, health care utilization and educational activities. With few exceptions, kappa coefficients of consistency were high, reflecting good interinterviewer agreement regarding the presence or absence of specific types of drug use. Recent (90-day) drug use was likewise reliably measured by a liberal standard (r) and the most frequently used drugs (tobacco, marijuana, cocaine and stimulants) were reliably measured by a more conservative reliability standard (ICC). The less consistently reliable assessment of opioids and tranquilizers may be due in part to the infrequency of their use in this sample, and reliability could likely be improved by placement of drug use days, like alcohol days, in the base calendar rather than having respondents simply estimate the total number of days of use for each drug class within the assessment window.

Findings from these two studies support the reliability of Form 90, comparable to that reported for the timeline follow-back method. Comparison of current study findings with five

meta-analytically combined test-retest studies on the timeline follow-back (Sobell et al., 1988) indicate that both methods provide reliable estimates of alcohol consumption. For example, for number of abstinent and heavy-drinking days in a 90-day interval, the two approaches yielded sample weighted and averaged *r*'s of .94 (Timeline) and .96 (Form 90) and .90 (Timeline) and .95 (Form 90), respectively. Both methods also seem to provide reliable estimates of alcohol consumption *across* an assessment interval (Sobell et al., 1986).

It should be noted that results reported in this study were obtained with intensively trained interviewers, who followed manual-guided procedures and who were regularly monitored for drift from assessment protocols. Under conditions of less control and greater interviewer variability, similar reliability might not be obtained. The minimal extent and optimal methods of training to establish interviewer reliability need to be clarified for Form 90 and other interview procedures to assess alcohol consumption.

In sum, under the conditions of these studies, Form 90 provided reliable estimates of alcohol consumption and related variables and represents one of several structured assessment procedures to quantify drinking behavior. Its complexity and the level of detail that it provides will be more than is needed for some purposes, where simpler quantity-frequency estimates suffice (Grant et al., 1995). In essence, Form 90 represents a modified TLFB, incorporating the grid method to expedite reconstruction of repetitive patterns, and including additional outcome variables. Like the TLFB, Form 90 yields a continuous record of behavior suitable for analyses (such as time-to-event) that require the location of events in real time. In some applications, a day-by-day reconstruction of all drug use would be desirable, for which adapted protocols (Forms 90-DI and 90-DF) are available (Miller, 1996). The relative advantages of TLFB and Form 90 cannot be determined from the present study. The intent and consensus of the Project MATCH Research Group was to produce in Form 90 a reliable instrument that was efficiently adaptable to a broad range of drinking patterns and that used the calendar base to reconstruct other outcome variables of interest in the trial. This report demonstrates the reliability of the instrument that resulted from these efforts.

Acknowledgment

The authors gratefully acknowledge the collaboration of the Project MATCH Research Group in preparation of this report.

References

CERVANTES, E.A., MILLER, W.R. AND TONIGAN, J.S. Comparison of timeline follow-back and averaging methods for quantifying alcohol consumption in treatment research. Assessment **1:** 23-30, 1994.

CICCHETTI, D.V. Guidelines' Criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. **6:** 284-290, 1994.

COHEN, J. Statistical Power Analysis for the Behavioral Sciences, 2d Edition, Hillsdale, N.J.: Lawrence Erlbaum Assocs. Inc., 1988.

DEL BOCA, F.K., BABOR, T.F., MCREE, B. AND WIRTZ, P.W. Assessment of reliability in multisite clinical research: An empirical assessment of four interviews at nine sites, in press.

GRANT, K.A., TONIGAN, J.S. AND MILLER, W.R. Comparison of three alcohol consumption measures: A concurrent validity study. J. Stud. Alcohol **56:** 168-172, 1995.

LITTEN, R.Z. AND ALLEN, J.P. (Eds.) Measuring Alcohol Consumption: Psychosocial and Biochemical Methods, Totowa, N.J.: Humana Press, 1992.

MILLER, W.R. Form 90: A structured assessment interview for drinking and related behaviors, Volume 5, NIAAA Project MATCH Monograph Series, NIH Publication No. 96-4004, Washington: Government Printing Office, 1996.

MILLER, W.R. AND DEL BOCA, F.K. Measurement of drinking behavior using the Form 90 family of instruments. J. Stud. Alcohol, Supplement No. 12, pp. 112-118, 1994.

MILLER, W.R., HEATHER, N. AND HALL, W. Calculating standard drink units: International comparisons. Brit. J. Addict. **86:** 43-47, 1991.

MILLER, W.R., LECKMAN, A.L., DELANEY, H.D. AND TINKCOM, M. Long-term follow-up of behavioral self-control training. J. Stud. Alcohol **53:** 249-261, 1992.

MILLER, W.R. AND MARLATT, G.A. Manual for the Comprehensive Drinker Profile, Odessa Fla.: Psychological Assessment Resources, 1984.

PROJECT MATCH RESEARCH GROUP. Project MATCH: Rationale and methods for a multisite clinical trial matching patients to alcoholism treatment. Alcsm Clin. Exp. Res. **17:** 1130-1145, 1993.

SHROUT, P.E. AND FLEISS, J.L. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. **86:** 420-428, 1979.

SOBELL, L.C. AND SOBELL, M.B. Timeline follow-back: A technique for assessing self-reported alcohol consumption. In: Litten, R.A. and Allen, J.P. (Eds.), Measuring Alcohol Consumption: Psychosocial and Biochemical Methods, Totowa, N.J.: Humana Press, 1992, pp. 41-72.

SOBELL, L.C., SOBELL, M.B., LEO, G.I. AND CANCILLA, A. Reliability of a timeline method: Assessing normal drinkers' reports of recent drinking and a comparative evaluation across several populations. Brit. J. Addict. **83:** 393-402, 1988.

SOBELL, M.B., SOBELL, L.C., KLAJNER, F., PAVAN, D. AND BASIAN, E. The reliability of a timeline method for assessing normal drinker college students' recent drinking history: Utility for alcohol research. Addict. Behav. **11:** 149-161, 1986.