Published in final edited form as:

J Stud Alcohol. 2004 November ; 65(6): 774–781.

Temporal Stability of the Timeline Followback Interview for Alcohol and Drug Use with Psychiatric Outpatients

Kate B. Carey, Michael P. Carey, Stephen A. Maisto, and James M. Henson

Center for Health and Behavior Syracuse University

Abstract

Objective—The purpose of this study was to evaluate the test-retest reliability of the Timeline Followback (TLFB) interview for assessing daily alcohol and drug use with adults living with a severe mental illness.

Method—Participants were 132 psychiatric outpatients (64% male) with a confirmed schizophrenia-spectrum (52%) or major mood disorder (48%) and a lifetime history of substance use disorder. This sample completed a 90-day TLFB twice, separated by a mean of 5 days, and represents 55% of the participants who originally consented to be in the study.

Results—Test-retest reliability coefficients ranged from .73 to 1.00 (rounded) for 30-day TLFB, and from .77 to 1.00 (rounded) for the 90-day TLFB. Within-subject comparisons of means across the three 30-day windows revealed no significant differences, and no degradation of the magnitude of the reliability coefficients was observed with increasingly distal assessment periods.

Conclusion—The TLFB is a reliable method of assessing alcohol and drug use in outpatients diagnosed with severe mental illness.

The frequent co-occurrence of substance use and misuse with severe mental disorders requires assessment of alcohol and drug use behavior in psychiatric treatment settings. Epidemiological data indicate that one-third to on-half of persons with a mental disorder qualify for a diagnosis of alcohol or other drug abuse or dependence at some time in their lives (Kessler, Nelson, McGonagle, Edlund, Frank, & Leaf, 1996; Regier, Farmer, Rae, Locke, Keith, Judd, & Goodwin, 1990). Comorbidity rates are highest for persons living with bipolar disorder (61%) or schizophrenia (47%) (Regier et al., 1990). Because even low levels of substance use can have destabilizing effects on persons with severe mental disorders (Drake and Brunette, 1998), monitoring substance use over time is important for effective treatment of patients with co-occurring disorders.

A recent review of clinically useful assessments for measuring substance use in psychiatric populations identified the Timeline Followback (TLFB) technique (Sobell and Sobell, 1996) as a method useful for planning treatments, providing motivational feedback, and for monitoring change over time (Carey, 2002). Evidence for the reliability and validity of summary alcohol use variables derived from the TLFB has been obtained from studies using a variety of populations, including college students, community residents, and participants in alcohol treatment, for time frames of up to one year (Sobell and Sobell, 1996). Because the TLFB yields count data and lacks a latent factor structure, reliability has typically been reported in the form of test-retest stability. With intervals of 2–4 weeks, the temporal stability of one-month quantities (e.g., total drinks, drinks per day) and frequencies (e.g., drinking days) typically exceeds .85 (Sobell and Sobell, 1996; Sobell et al., 1988; Sobell et al., 1986).

Reliability coefficients of this magnitude indicate a high degree of consistency among the quantity and frequency summary variables derived from daily drinking assessments.

The TLFB format has also been expanded to assess drug use, with similarly strong reliability evidence reported from drug abuse treatment samples (Ehrman and Robbins, 1994; Fals-Stewart et al., 2000; Hersh et al., 1999; Sacks et al., 2003). For example, Fals-Stewart et al. (2000) reported the two-week test-retest reliability for seven different substance classes; participants were 113 outpatients in treatment for substance use disorders who were screened to exclude persons with psychotic disorders. Separate retest correlations were calculated for reporting intervals of 30, 90, and 365 days. Based on only the participants with non-zero pairs of responses, the retest correlations for cannabis use days were .89, .90, and .92 for the respective reporting intervals; similarly, cocaine use days revealed correlations of .95, .92, and .89. In the context of multiple drug use, the retest correlations for alcohol use days were also quite high (.94, .90 and .88). In sum, these data confirm that drug and alcohol reliability coefficients do not differ in magnitude, and are high across reporting intervals from one month to one year.

Although ample evidence supports the reliability of the TLFB in samples of substance users, these results cannot be generalized to clinical populations with major psychiatric disorders. Indeed, the importance of evaluating the psychometric properties of self-report based substance use assessments among persons with severe mental illness has been argued (Carey et al., 1997; Drake et al., 1995; Goldfinger et al., 1996). As Carey (2002) described, persons with severe mental disorders exhibit greater variability in their mental status than persons without such disorders. Less accurate self-reports may arise due to acute distress, exacerbated symptoms, impaired reality orientation, confusion and cognitive impairment, in addition to frequent and excessive substance use. However, the mere presence of a co-occurring psychiatric diagnosis does not compromise the psychometric soundness of substance-related assessments, as has been demonstrated in previous empirical reports (Carey et al., 2001; Cocco and Carey, 1998; Maisto et al., 2000; Teitelbaum and Carey, 2000).

To our knowledge, only two studies examined the test-retest reliability of the TLFB when used with persons with a severe psychiatric disorder. Carey (1997) sampled psychiatric outpatients and reported evidence of test-retest stability of responses to two alcohol use variables on 1-month and 6-month TLFBs. With a retest interval that averaged 6 days, test-retest correlations for number of alcohol use days ranged from .82 (for the most recent month) to .62 (over 6 months). Test-retest correlations for maximum quantity consumed on a single occasion were consistent across both intervals, $r = .88$ and $r = .92$, respectively. However, the sample used in this study was small ($n = 17$).

Sacks, Drake, Williams, Banks, and Herrell (2003) also evaluated the test-retest reliability of the TLFB over a 1–2 week retest interval with 158 participants in a homelessness prevention program. Although participants were not recruited from a psychiatric treatment setting, two-thirds of the sample reported receiving inpatient psychiatric treatment at some time in their lives. The TLFB interview used in this study assessed alcohol and composite drug use over 6 months. Excluding zero-zero pairs, the intraclass correlations for the most recent and the most distal months were .76 and .73 for alcohol use days ($n = 80$), .68 and .81 for number of drinks ($n = 80$), .64 and .91 for marijuana use days ($n = 40$), and .77 and .89 for any drug use days ($n = 64$). Thus, the responses across time for both alcohol and drug use were stable, with no decrement for the most distal month relative to the most recent month. Furthermore, these authors noted that reliability was not related to gender, age, or severity of psychiatric symptoms.

The two available studies that have evaluated the TLFB among persons with severe mental illness have produced preliminary support for its reliability. However, the evidence is limited

because the (a) samples have been either small or not focusing exclusively on the population of interest, and/or (b) only a limited number of substance use variables were assessed. TLFB assessments are used to gather substance use information from patients with co-occurring substance use and psychiatric disorders (Carey et al., 2002; Drake et al., 2000; el-Guebaly et al., 1999). However, there is little empirical information supporting the psychometric quality of the data collected. Accordingly, it is important to establish the reliability of the TLFB for a range of variables likely to be of interest to researchers, over varying time periods.

This study had three primary aims. First, we replicate findings of Carey (1997) with a larger sample, and extend the reliability analyses to a larger collection of pertinent alcohol and drug use variables derived from the TLFB. In this regard, we assess test-retest stability for (a) maximum quantity consumed in the last month, (b) total number of drinks, (c) number of drinking days, (d) number of heavy drinking days, (e) number of marijuana use days, and (f) number of days in which any other drug was used. Because the inclusion of non-users who provide zero-zero responses to repeated assessments can artificially inflate reliability estimates, for each variable we report coefficients based only on participants who reported using that substance on at least one of the two assessments.

Second, we compare the reliability coefficients of the most recent month to the most distal month within a 90-day assessment window. Comparison of individual months will address whether response stability decreases as the recall interval (and the memory demand) increases. In addition, we calculate and compare stability coefficients for both the 30-day and the 90-day time frames, to guide researchers who require longer assessment intervals.

Third, because reports of the frequency of different substance use behaviors tend not to be distributed normally, we examine the benefits of two outlier management strategies. Tabachnik and Fidell (2001) describe procedures to reduce the influence of extreme responses on correlational statistics. For example, extreme (high) values can be reduced to more plausible but still improbable levels by transforming the distribution, or by truncating scores at ± 3 SD. Another method is to remove the outlier values, which would be appropriate if they are likely to have originated from a different population other than the population of interest, or if their influence cannot be reduced in other ways. We will evaluate the effect these methods have on the test-retest correlations.

Method

Participants and Recruitment

Participants were recruited from two state-funded psychiatric outpatient clinics as part of a larger study that focused on the relationship between readiness to change and substance use in psychiatric outpatients (Carey et al., 2001). To be eligible for participation, a patient had to (a) be at least 18 years of age, (b) have a documented diagnosis of schizophrenia or a major mood disorder, and (c) have a lifetime history of a substance use disorder. The presence of a substance use disorder was initially determined by one or more of the following markers in a patient's medical chart: a score of 8 or greater on the Alcohol Use Disorders Identification Test (AUDIT; Bohn et al., 1995), a score of 3 or greater on the short version of the Drug Abuse Screening Test (DAST; Skinner, 1982), or an affirmative response to the question "Have drugs or alcohol ever been a problem for you?" Psychiatric diagnoses and lifetime substance abuse or dependence were then confirmed by a Structured Clinical Interview for DSM-IV (SCID; First et al., 1995).

The screening procedure identified 333 patients who were invited to be part of the study, and 240 consented to participate. Of the 240 that provided written informed consent, 132 participants completed all four sessions of the study and provided complete TLFB data on

initial test and retest. Participants ranged in age from 22 to 71 years ($M = 44.1$); 84 (64%) were male; 77% reported their race as Caucasian/White, 17% as African American, 2% as Native American, 2% as Hispanic, and 3% as other. Primary diagnoses were schizophrenia (34%) and major depressive disorder (33%), schizoaffective disorder (16%), bipolar disorder (15%), and other psychotic disorder (2%). The mean AUDIT score was 9.5 ($SD = 9.6$) and the mean DAST score was 2.0 ($SD = 2.9$). The mean Global Assessment of Functioning score, as assessed with the SCID, was 44.9 ($SD = 10.4$), indicating that most participants displayed serious psychiatric symptoms and/or impairment in social or occupational functioning. Moreover, the Positive and Negative Symptom Scale (PANSS) (Kay et al., 1987), revealed moderate levels of positive ($M = 13.54$, $SD = 5.0$) and negative ($M = 15.68$, $SD = 5.0$) symptoms. Although the majority of participants did live in their own house or apartment ($n = 88$; 67%), a substantial proportion relied on other's for shelter (e.g., 12% in group homes/halfway house, 15% in their parent's or other relation's home, 5% in a shelter or hotel). The average income reported by participants was \$514 a month ($SD = \362). Most participants received their income from disability ($n = 81$; 62%), welfare ($n = 23$; 18%), or both ($n = 13$; 10%); only seven participants received income solely from paid employment and eight participants did not provide this information.

Materials

AUDIT—The Alcohol Use Disorders Identification Test (Bohn et al., 1995) is a 10-item screening instrument designed to identify drinkers at risk for alcohol abuse and dependence. Internal consistency estimates have ranged from .75 - .94 in a variety of populations (Allen et al., 1997; Dawe et al., 2000). In a psychiatric setting, obtaining a score of 8 or greater identifies persons at high risk for alcohol use disorders with a high degree of sensitivity and specificity (Maisto et al., 2000).

DAST-10—The 10-item version of the Drug Abuse Screening Test (Skinner, 1982) was used to identify drug-use related problems in the past year. The DAST-10 is internally consistent ($\alpha = .86$), temporally stable ($ICC = .71$), and able to discriminate between psychiatric outpatients with and without current drug abuse/dependence diagnoses (Cocco and Carey, 1998). Sensitivity and specificity in the discrimination of drug use disorders with this population are optimized with a score of ≥ 3 (Maisto et al., 2000).

Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental Disorders—(SCID; First et al., 1995). To confirm participants' DSM-IV diagnosis, we used the SCID-Patient Version (SCID-I/P, Version 2), which is suited for use with psychiatric patients for whom differential diagnosis of psychotic disorders is often necessary. The sections on mood, psychotic, and substance-related disorders were used to confirm that participants met the diagnostic eligibility criteria. As recommended, information from participants' psychiatric charts was used to corroborate and supplement the SCID data. All diagnostic interviews were videotaped for purposes of assessing reliability; 28 videotaped interviews were reviewed independently by a second assessor who made parallel ratings. Primary psychiatric diagnosis was assessed reliably ($\kappa = 0.88$), as was the determination of a lifetime substance use diagnosis (Kuder-Richardson-20 = 0.86).

Timeline Follow-Back—(TLFB; Sobell and Sobell, 1996). The TLFB is well-suited to obtaining data on the frequency of occurrence of specified behaviors, and it reduces the likelihood of under-reporting relative to quantity-frequency averaging methods (Sobell et al., 1982). Following the TLFB administration manual (Sobell and Sobell, 1996), the TLFB interviewer uses an annotated calendar that is personalized for each participant. This calendar serves as a memory cue for participants as they try to recall daily drinking and/or drug use. For this study, participants viewed calendars representing the previous 90 days. Assessors highlighted major holidays over the three-month period, and then asked the participant to

identify their own personal holidays or days of importance. Other interview strategies were used to prompt recall, such as identifying extended abstinent periods and recording regular patterns around weekends or check days. If the participant had a personal calendar or date book, the assessor encouraged the participant to use it to reconstruct the events of the last 90 days.

Alcohol use data were collected on the first pass through the calendar, and drug use data were obtained on subsequent passes. A standard drink was defined as 12 oz of beer, 4 oz of wine, or 1 oz of hard liquor; these definitions were graphically illustrated on a card for participants to refer to throughout the interview. The assessor then guided the participant to reconstruct the number of standard drinks consumed each day, starting from the day prior to the assessment and proceeding backwards for 90 days. Participants were instructed to be as accurate as possible but when unable to remember, they were asked to provide their best guess. Unique codes were used to represent separate drug categories. Participants indicated which drugs they had used in the last 90 days, and starting with the drug used most frequently, they indicated the days on which that drug was used. Thus, the following variables could be derived from the calendars: (a) total standard drinks, (b) drinking frequency, (c) heavy drinking frequency (5+ standard drinks), (d) maximum quantity consumed, number of days using (e) marijuana, (f) cocaine, (g) amphetamine, (h) over-the-counter (OTC) stimulants, (i) sedative-hypnotics, (j) opiates, (k) hallucinogens, (l) inhalants, (m) injection drugs, and (n) total days on which any alcohol or drug was used.

Procedure

Assessment sessions took place in private offices in the clinics or in an adjacent building. A trained research assistant (RA) administered a breath analyzer test at the beginning of each session, and meetings were rescheduled if blood alcohol level was ≥ 0.02 . Data were collected during four sessions. The first session served to obtain informed consent and demographic and locator information to facilitate future contacts. The second assessment session was used to confirm the participants' primary diagnoses. SCIDs were administered by masters-level or PhD-level clinical psychologists. The third and fourth assessments were the test and retest sessions for the TLFB, administered by a well-trained and supervised RA. The average test-retest interval was five days ($SD = 4.1$). Consistent with methodology used in prior research, instructions for the retest session included reconstruction of the past 90 days prior to the first session. Those who completed the study received \$30 for their participation. Upon completion of the TLFB interviews, the RA compiled the daily alcohol and drug use data from the calendars into the respective composite variables.

Results

Preliminary Analyses

Forty-nine participants reported a complete absence of drug and alcohol use at both assessments. In order to ensure that these data would not artificially inflate the test-retest correlations, they were removed from further analyses, leaving 83 cases for reliability analysis. Chi-square analyses (or *t*-tests) demonstrated no differences between users and non-users with regard race, age, gender, and psychiatric diagnosis (all $ps > .10$).

The descriptive statistics for the full 90-day assessment window are reported in Table 1 for both the first and second TLFB assessments. Maximum drinks refers to the number of standard drinks consumed on the heaviest drinking day across the 90-day period, and total drinks represents a sum of standard drinks across the designated time period. Examination of Table 1 reveals that few people used drugs other than alcohol and marijuana, but those who did tended to use those substances regularly. The number of users for various drugs ranged from 15 cocaine

users to only 1 user in each of the amphetamine, inhalant, and injected drug use categories. The average use of these substances ranged from once every three days for opiate and sedative users, to once in a three-month period for hallucinogens. Because use of these drugs was infrequent, drug use excluding marijuana and alcohol was summed across all other drugs to form a composite variable of drug use labeled “Composite drug use days” in Table 1. It is important to note that the composite drug use variable does *not* include information regarding marijuana use.

Initial analyses explored if there was a systematic trend of under- or over-reporting at the retest session or at more distal months. Table 2 displays the drug and alcohol use averages from the first and second assessments for each 30-day time interval. A series of paired-samples *t*-tests revealed no differences between the first and second assessment across alcohol and drug use variables (all *ps* > .10). Thus, on average, retest values were not significantly different than the original test values.

One concern about retrospective daily assessments such as these is that participants may tend to under-report their substance use farther back in time, because they are unable to remember, and therefore reconstruct, specific events. To test this, a series of within-subject ANOVAs were conducted on the patients’ average reported monthly use for the three time intervals in the first TLFB assessment. These tests demonstrated no significant differences among the means of the three assessment months (all *ps* > .10). As summarized in Table 2, there was no differences in average reported use as participants reported use further back in time. In this sample, substance use appeared to be rather stable across the 90-day assessment window. Taken together, these analyses reveal no systematic effect on the participants’ average reported use from receiving multiple assessments or from reporting on increasingly distant time intervals.

Reliability Analyses

Test-retest reliability coefficients are reported in the form of the Pearson product-moment correlations in the first four columns of Table 3. Intraclass correlations were also computed, but they were nearly identical to the Pearson estimates (mode and median of difference = .00; mean difference = .008), and so are not reported.. These correlations were computed on the raw data (less zero-zero pairs) and are presented for each 30-day time interval as well as the data aggregated across the entire 90-day assessment window. For the 30-day time intervals, the correlations ranged from .73 for heavy drinking days, to 1.00 (rounded) for total drinks. The sums for the 90-day aggregated interval ranged from .77 to 1.00.

As evident from the descriptive statistics in Table 1, a small number of extreme responses (i.e., outliers) produced skewed distributions, which may artificially inflate the test-retest correlation. Although skewed distributions tend not to be problematic when computing correlations unless they are skewed in opposite directions, outliers can create spurious correlations. In order to assure that the high reliability coefficients were not an artifact of extreme data points, outliers were examined in two ways.

The first method was to convert the data to standardized scores for each assessment separately, and reduce values > |3| SDs to 3 or -3, respectively. The test-retest correlations were recomputed using the adjusted standardized scores. This method allowed the use of all available data, but limited the influence of extreme responses on the correlations. The second method of diminishing the influence of outliers on the correlations was to simply remove them, and to recalculate the test-retest correlations on the reduced sample.

The decision rules for identifying outliers were based on the sample distributions. Cutoffs were chosen (a) to minimize the number of cases excluded and (b) to identify values that were clearly disjunct from the distribution of the other scores. For maximum drinks on one occasion, an

outlier was defined as 40 or more drinks, and for number of drinks per month, the maximum value was 190 drinks. For the remaining variables, all of which measure number of use days, an outlier was defined as using 20 or more days per month. The exception to this latter rule was composite drug use days, for which an outlier was defined as 30 or more days, because it was a summed score over multiple drug categories.

The resulting test-retest correlations from these two methods of controlling for extreme scores are displayed in the last four columns of Table 3 for the most recent 30-day interval and the aggregate 90-day interval. Columns 5 and 6 display the correlations when extreme scores have been trimmed to a less extreme value to limit their influence, whereas columns 7 and 8 display the correlations with the extreme scores removed. Examination of these adjusted reliabilities confirms that these extreme data points do tend to inflate the correlations; the pattern of correlations shows that they decrease in magnitude more when extreme points are removed than when they are trimmed. For example, removing three people reduced the most recent month correlation for total drinks from 1.00 to .91. Another variable that was influenced by outliers was the aggregated 90-day correlation for number of drinking days. This correlation decreased from .97 to .88 when the extreme scores are removed. Nevertheless, the pattern of correlations adjusted for outliers shows that they still are moderate to high, ranging from .68 – .98 (see Table 3, columns 5 through 8).

Discussion

The primary goal of this study was to evaluate the reliability of the TLFB method for assessing alcohol and drug use among psychiatric outpatients. The test-retest reliability coefficients obtained in this sample support and extend the prior results reported by Carey (1997), who used a much smaller sample, fewer variables, and only a one-month TLFB. Furthermore, the reliability of alcohol and drug use did not degrade across the three months prior to assessment. Thus, data from the current study suggest that the TLFB can be used for reliable measurement of a wide range of substance use behaviors as far back as 90 days.

It might seem intuitively obvious that persons with severe mental illness give less reliable self-reports of their past behaviors than those who do not have such disorders. However, our data and previous research do not support this hypothesis. The test-retest reliability coefficients obtained in this study compare favorably to those reported in samples of college students (Sobell et al., 1986), drinkers from the general population (Sobell et al., 1988), and persons with drug use disorders (e.g., Fals-Stewart et al., 2000). Notably, all of these studies excluded participants with severe psychiatric disorders. Taken together with the results reported by Carey (1997) and Sacks et al. (2003), the findings of this study establish that the TLFB can be reliably used in samples with severe psychiatric disorders.

Research participants in clinical samples often report extreme values of a target behavior. Such “outliers” might be interpreted as invalid and they can distort parametric inferential statistics (Schroder, Carey, & Vanable, 2003). Although researchers frequently remove outliers prior to data analyses, that strategy may misrepresent the sample’s behavior and important information – from a public health perspective – may be lost. Indeed, it is the presence of extreme patterns of behavior that characterize an individual’s membership in a clinical population.

In the present sample, a significant percentage of participants reported extreme drug and alcohol use behavior; however, extreme scores did not necessarily reflect unreliable reporting. As an illustration, our data contained multiple patients who reliably reported extreme drinking behavior on both measurement occasions. However, our data also contained patients who reported minimal or no drinking on one measurement occasion and extreme drinking behavior on the other. This indicates that extreme data, or outliers, are not necessarily indicators of

unreliability. Using terminology from Tabachnik and Fidell (2001), an outlier exhibits leverage (distance from other scores) but does not necessarily exhibit discrepancy (deviation from the pattern of the others). We suggest that researchers should be very cautious in removing outliers, and instead should utilize other methods that minimize the influence of extreme responses on correlational statistics. The data in Table 3 show that the summary scores derived from the TLFB retain a high level of consistency regardless of the method of handling extreme cases, and trimming outliers can be as effective as removing them. Patients being treated for a severe mental illness are likely to give extreme responses; our data show that trimming the distribution produces reliability coefficients nearly identical to those obtained from the original distribution of scores with no loss of data.

A limitation of this study was the high attrition rate of the patients who gave consent to be in this study. Slightly over half (55%) of the participants who provided consent completed the multiple assessments. Because we do not have access to the data on these subjects, we were unable to explore differences between those who did and did not complete the study. Therefore, these findings should be interpreted as representative of psychiatric outpatients who are able to complete a four-session assessment study.

These findings do, however, suggest several directions for future research. Although reliability is a prerequisite for validity, the high reliability coefficients found in this study do not establish the accuracy of the self-reports. Providing empirical support for the validity of self-reported drug use has its challenges (Maisto, McKay, & Connors, 1990), but precedents for validating self-reported substance use in psychiatric samples using collateral reports (Carey & Simons, 2000) and urinalysis tests (Weiss et al., 1998) have been reported. Another important direction for future research involves identifying predictors of unreliable or invalid self-reports. In this regard, Babor, Brown, and DelBoca (1990) presented a conceptual framework for understanding factors influencing the accuracy of self-reports; this framework suggests that respondent characteristics, motivational factors, the nature of the assessment task, and the social context of the information gathering all may influence the accuracy of self-report data. Few studies have addressed the predictors of self-report reliability in psychiatric patients. One exception is a study by Teitelbaum and Carey (2000), which found that male gender and the presence of a lifetime substance use disorder independently predicted test-retest discrepancies in responses to an alcohol screening instrument, whereas memory impairment and psychological symptom distress did not. Identification of factors associated with unreliable responding to the TLFB may allow researchers to maximize the effective use of this data collection procedure.

In conclusion, the TLFB is an assessment technique that is relatively easy to administer and can be used to obtain event-level data on a wide variety of substance use behaviors. This study supports the use of the TLFB outpatients diagnosed with severe mental illness to obtain information on the quantity and frequency of alcohol use and the frequency of drug use. Furthermore, these results enhance confidence that alcohol and drug use patterns can be assessed reliably for up to 90 days prior to assessment.

Acknowledgements

This work was supported in part by NIDA grants DA10010 and DA00426 to Kate B. Carey. The authors gratefully acknowledge Dan Purnine and Adrienne Williams for their assistance with data collection and management.

References

- Allen JP, Litten RZ, Fertig JB, Babor TF. A review of research on the Alcohol Use Disorders Identification Test (AUDIT). *Alcoholism: Clinical and Experimental Research* 1997;21:613–619.

- Bohn MJ, Babor TF, Kranzler HR. The Alcohol Use Disorders Identification Test (AUDIT): Validation of a screening instrument for use in medical settings. *Journal of Studies on Alcohol* 1995;56:423–432. [PubMed: 7674678]
- Carey KB. Reliability and validity of the Timeline Follow-Back Interview among psychiatric outpatients: A preliminary report. *Psychology of Addictive Behaviors* 1997;11:26–33.
- Carey KB. Clinically useful assessments: Substance use and comorbid psychiatric disorders. *Behavior Research and Therapy* 2002;40:1345–1361.
- Carey KB, Carey MP, Maisto SA, Purnine DM. The feasibility of enhancing psychiatric outpatients' readiness to change their substance use. *Psychiatric Services* 2002;53:602–608. [PubMed: 11986511]
- Carey KB, Cocco KM, Correia CJ. Reliability and validity of the Addiction Severity Index among outpatients with severe mental illness. *Psychological Assessment* 1997;9:422–428.
- Carey KB, Maisto SA, Carey MP, Purnine DM. Measuring readiness-to-change substance misuse among psychiatric outpatients I. Reliability and validity of self-report measures. *Journal of Studies on Alcohol* 2001;62:79–88. [PubMed: 11271968]
- Cocco KM, Carey KB. Psychometric properties of the Drug Abuse Screening Test in psychiatric outpatients. *Psychological Assessment* 1998;10:408–414.
- Dawe S, Seinen A, Kavanagh D. An examination of the utility of the AUDIT in people with schizophrenia. *Journal of Studies on Alcohol* 2000;61:744–750. [PubMed: 11022815]
- Drake, RE.; Brunette, MF. Complications of severe mental illness related to alcohol and drug use disorders. In: Galanter, M., editor. *Recent developments in alcoholism*. 14. New York: Plenum; 1998. p. 285-299.
- Drake RE, Mchugo GJ, Biesanz JC. The test-retest reliability of standardized instruments among homeless persons with substance use disorders. *Journal of Studies on Alcohol* 1995;56:161–167. [PubMed: 7760561]
- Drake RE, Xie H, Mchugo GJ, Green AI. The effects of clozapine on alcohol and drug use disorders among patients with schizophrenia. *Schizophrenia Bulletin* 2000;26:441–449. [PubMed: 10885642]
- Ehrman RN, Robbins SJ. Reliability and validity of 6-month Timeline reports of cocaine and heroin use in a methadone population. *Journal of Consulting and Clinical Psychology* 1994;62:8433–8450.
- El-Guebaly N, Hodgins DC, Armstrong S, Addington J. Methodological and clinical challenges in evaluating treatment outcome of substance-related disorders and comorbidity. *Canadian Journal of Psychiatry* 1999;44:264–270.
- Fals-Stewart W, O'farrell TJ, Freitas TT, Mcfarlin SK, Rutigliano P. The Timeline Followback reports of psychoactive substance use by drug-abusing patients: Psychometric properties. *Journal of Consulting & Clinical Psychology* 2000;68:134–144. [PubMed: 10710848]
- First, MG.; Spitzer, RL.; Gibbon, M.; Williams, JBW. *Structured clinical interview for DSM-IV--Patient version (SCID-I/P, Version 2.0)*. New York: New York State Psychiatric Institute, Biometric Department; 1995.
- Goldfinger SM, Schutt RK, Seidman LJ, Turner WM. Self-report and observer measures of substance abuse among homeless mentally ill persons in the cross-section and over time. *Journal of Nervous & Mental Disease* 1996;184:667–672. [PubMed: 8955679]
- Hersh D, Mulgrew CL, Van Kirk J, Kranzler HR. The validity of self-reported cocaine use in two groups of cocaine abusers. *Journal of Consulting & Clinical Psychology* 1999;67:37–42. [PubMed: 10028207]
- Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin* 1987;13:261–276. [PubMed: 3616518]
- Kessler RC, Nelson CB, McGonagle KA, Edlund MJ, Frank RG, Leaf PJ. The epidemiology of co-occurring addictive and mental disorders: Implications for prevention and service utilization. *American Journal of Orthopsychiatry* 1996;66:17–31. [PubMed: 8720638]
- Maisto SA, Carey MP, Carey KB, Gleason JG, Gordon CM. Use of the AUDIT and the DAST-10 to identify alcohol and drug use disorders among adults with a severe and persistent mental illness. *Psychological Assessment* 2000;12:186–192. [PubMed: 10887764]
- Regier DA, Farmer ME, Rae DS, Locke BZ, Keith SJ, Judd LL, Goodwin FK. Comorbidity of mental disorders with alcohol and other drug abuse. *Journal of the American Medical Association* 1990;264:2511–2518. [PubMed: 2232018]

- Sacks JAY, Drake RE, Williams VF, Banks SM, Herrell JM. Utility of the Time-Line Follow-Back to assess substance use among homeless adults. *Journal of Nervous & Mental Disease* 2003;191:145–153. [PubMed: 12637840]
- Skinner H. The Drug Abuse Screening Test. *Addictive Behaviors* 1982;7:363–371. [PubMed: 7183189]
- Sobell LC, Cellucci T, Nirenberg TD, Sobell MB. Do quantity-frequency data underestimate drinking-related health risks? *American Journal of Public Health* 1982;72:823–828. [PubMed: 7091478]
- Sobell, LC.; Sobell, MB. Timeline FollowBack user's guide: A calendar method for assessing alcohol and drug use. Toronto: Addiction Research Foundation; 1996.
- Sobell LC, Sobell MB, Leo GI, Cancilla A. Reliability of a timeline method: Assessing normal drinkers' reports of recent drinking and a comparative evaluation across several populations. *British Journal of Addiction* 1988;83:393–402. [PubMed: 3395719]
- Sobell MB, Sobell LC, Klajner F, Pavan D. The reliability of a timeline method for assessing normal drinker college students' recent drinking history: Utility for alcohol research. *Addictive Behaviors* 1986;11:149–161. [PubMed: 3739800]
- Tabachnick, BG.; Fidell, LS. Using multivariate statistics. 4. Boston, MA: Allyn and Bacon; 2001.
- Teitelbaum L, Carey KB. Temporal stability of alcohol screening measures in a psychiatric setting. *Psychology of Addictive Behaviors* 2000;14:401–404. [PubMed: 11130159]

Table 1
Summary Statistics of the Aggregated 90 Day Assessment Period for the First and Second TLFB Measurement Occasions

Item	First Assessment				Second Assessment					
	Mean	Median	SD	Max	n	Mean	Median	SD	Max	n
Maximum drinks	12.0	9.0	13.3	80	70	11.2	8.0	11.2	70	70
Drinking days	16.1	7.0	22.1	88	68	16.3	8.0	21.6	89	69
Heavy drinking days ^a	13.4	6.0	20.0	88	49	11.5	5.0	16.5	89	48
Total drinks	152.8	37.0	521.3	4251	68	147.9	38.0	524.5	4331	69
Marijuana use days	23.8	9.0	30.5	90	44	26.2	11.0	30.8	90	42
Cocaine use days	14.8	5.5	22.1	81	14	13.1	4.0	16.4	46	15
Amphetamine use days	14.0	14.0	0.0	14	1	7.0	7.0	0.0	7	1
OTC stimulant days	24.7	12.0	31.0	60	3	20.3	1.0	33.5	59	3
Sedative use days	27.9	21.0	28.3	81	8	29.3	18.0	30.5	87	7
Opiate use days	29.3	11.5	40.7	90	4	21.6	14.5	29.2	90	8
Hallucinogen use days	1.3	1.0	0.5	2	4	1.2	1.0	0.4	2	6
Inhalant use days	6.0	6.0	0.0	6	1	21.0	21.0	0.0	21	1
Injection drug use days	18.0	18.0	0.0	18	1	8.0	8.0	0.0	8	1
Composite drug use days ^b	23.4	8	31.8	116	34	23.7	7.0	33.7	140	34
Drug or alcohol use days	30.4	17.0	31.2	90	80	31.0	19.5	31.0	90	80

Note. Statistics are computed on users only. TLFB = Timeline Follow Back; OTC = over-the-counter

^a Heavy drinking day is defined as a day when 5 or more drinks are consumed.

^b Composite variable created by summing drug use over all substances, excluding alcohol and marijuana

Table 2

Individual Time Interval Means for the First and Second TLFB Assessment

Item	First assessment			Second assessment		
	0–30 days prior	31–60 days prior	61–90 days prior	0–30 days prior	31–60 days prior	61–90 days prior
Maximum drinks	9.1	9.4	8.3	8.4	8.8	8.5
Total drinks	59.4	58.0	59.8	59.0	55.1	59.8
Drinking days	6.3	6.2	6.1	6.4	6.1	6.6
Heavy drinking days	5.9	5.3	5.4	4.8	4.2	4.8
Marijuana use days	9.7	10.5	9.7	10.2	11.1	10.1
Composite drug use days ^a	9.5	8.7	9.6	8.4	9.1	10.4
Drug or alcohol use days	11.1	11.1	11.2	10.9	10.9	11.6

Note. Means based on participants who reported use of the substance of interest during indicated time period on test or retest. Time intervals are based on the number of days prior to the date of the first assessment. TLFB = Timeline Follow Back.

^aComposite drugs use variable that sums all drug use data except alcohol and marijuana

Table 3

Test-retest Correlations for 30 and 90 Day Time Intervals

Items	Raw Data ^a			Trimmed Data ^b			Reduced Data ^c		
	0– 30 days prior	31– 60 days prior	61– 90 days prior	0– 90 days prior	0–30 days prior	0–90 days prior	0–30 days prior	0–90 days prior	0–90 days prior
Maximum drinks	.93 (.58)	.82 (.59)	.92 (.59)	.97 (.70)	.88 (.58)	.92 (.70)	.78 (.56)	.90 (.67)	.90 (.67)
Total drinks	1.00 (.58)	.99 (.59)	1.00 (.59)	1.00 (.70)	.98 (.58)	.99 (.70)	.91 (.55)	.95 (.68)	.95 (.68)
Drinking days	.99 (.58)	.95 (.59)	.92 (.59)	.97 (.70)	.98 (.58)	.97 (.70)	.93 (.53)	.88 (.65)	.88 (.65)
Heavy drinking days	.73 (.37)	.72 (.41)	.68 (.42)	.77 (.51)	.68 (.37)	.76 (.51)	.73 (.34)	.84 (.48)	.84 (.48)
Marijuana use days	.94 (.37)	.88 (.33)	.84 (.35)	.91 (.44)	.94 (.37)	.91 (.44)	.92 (.28)	.88 (.35)	.88 (.35)
Composite drug use days	.94 (.25)	.94 (.30)	.90 (.31)	.94 (.38)	.92 (.25)	.94 (.38)	.88 (.23)	.90 (.36)	.90 (.36)
Drug or alcohol use days	.97 (.75)	.95 (.71)	.92 (.73)	.96 (.83)	.97 (.75)	.96 (.83)	.91 (.53)	.94 (.64)	.94 (.64)

Note. Pearson correlations were computed omitting participants who reported no use at both occasions. Time intervals are based on the number of days prior to the first data collection. Adjusted sample sizes are in parentheses.

^a Pearson correlations computed including all cases.

^b Pearson correlations computed on data where outliers were reduced to three standard deviations from the mean.

^c Pearson correlations computed with extreme raw scores omitted.