OXFORD

Article

Timeline Followback Self-Reports Underestimate Alcohol Use Prior to Successful Contingency Management Treatment

Brent A. Kaplan and Mikhail N. Koffarnus*

Fralin Biomedical Research Institute at VTC, Virginia Tech, 1 Riverside Circle, Roanoke, VA 24016, USA

*Corresponding author: Fralin Biomedical Research Institute at VTC, Virginia Tech, Roanoke, VA, USA. Tel: +5405262107; Fax: 5405262282; E-mail: mickyk@vtc.vt.edu

Abstract

Aims: The timeline followback (TLFB) is a retrospective self-report task that has been used successfully to measure prior alcohol consumption. The current study reanalyzed data from a recent successful demonstration of a remote contingency management trial for reducing alcohol consumption.
Methods: We first compared the accuracy of the TLFB and past-day self-reports to a biochemically verified measure of recent alcohol use (i.e., breathalyzer). We then compared the correspondence between the two self-report measures over two phases of the parent study: a phase immediately prior to and a phase including the treatment component.
Results: Our findings indicated that past-day self-reports displayed significantly higher accuracy with breathalyzer measures as compared to TLFB. In addition, we found only the experimental group, after reducing consumption, reported lower alcohol use on the TLFB prior to the treatment compared to their past-day self-reports and to the control group.
Conclusions: Our findings suggest daily monitoring techniques are more accurate than the TLFB for measuring alcohol consumption and when possible should be preferred over the TLFB. If the TLFB is the only viable method of measuring alcohol consumption, in order to maximize accuracy researchers and clinicians should obtain responses prior to the start of a procedure aimed at reducing alcohol consumption.

The Timeline Followback (TLFB) (Sobell and Sobell, 1992; Del Boca and Darkes, 2003) is one of several self-report tasks used to measure alcohol consumption, and is characterized by a retrospective daily self-report of alcohol use quantity for a period of time (often 30 days) preceding the assessment day. In the TLFB, respondents are usually provided with a calendar and are prompted to recall memorable events that occurred during the past duration of time. Respondents are then prompted to remember as best as they can the number of drinks they consumed each day. Alternative self-report measures that rely on shorter durations of recall have also been used to measure alcohol consumption. For example, Interactive Voice Response systems (Perrine *et al.*, 1995; Corkrey and Parkinson, 2002; Tucker *et al.*, 2012) and Ecological Momentary Assessments (Collins *et al.*, 2003;

Shiffman *et al.*, 2008; Shiffman, 2009) including physical or electronic diaries have been used to measure daily alcohol consumption. Interactive Voice Response systems are typically automated systems that call or text individuals daily and those individuals indicate how many drinks they have consumed that day by selecting numbers on the keypad. Similarly, Ecological Momentary Assessments traditionally measure whether and to what degree the individual is drinking at the very moment they are prompted. Completing physical and electronic diaries usually requires the individual to report the number of drinks consumed at the end of the day or have individuals report the day after. Nonetheless, measurement systems that reduce the duration of recall are typically preferred over more lengthy periods of recall. Therefore, an important difference between the TLFB and other

measures of retrospective self-reports is the proximity in time between engaging in the behavior and when the behavior is reported.

Previous studies comparing TLFB to daily self-reports have generally found good to excellent correspondence (Toll *et al.*, 2006; Tucker *et al.*, 2007; Simpson *et al.*, 2011) with correlation coefficients ranging from 0.51 (Perrine *et al.*, 1995) to 0.97 (Carney *et al.*, 1998). However, several studies have found recall on the TLFB tends to underestimate alcohol consumption overall, especially compared to daily monitoring techniques such as with Interactive Voice Response systems (Searles *et al.*, 2002; Toll *et al.*, 2006). In addition, several studies have compared biochemically verified measures of recent alcohol use to self-report measures (Sobell *et al.*, 1979; O'Farrell and Maisto, 1987; Perrine *et al.*, 1995; Whitford *et al.*, 2009). Overall, findings from these studies have concluded that when no drinking has occurred, biochemically verified measures tend to correspond well with self-report measures; however, when drinking has occurred self-reports tend to be less valid and under-report drinking. Almost unanimously, this research has suggested that biochemically verified measures should be used in addition to self-report instruments. Although research comparing correspondence between the two types of measures has been conducted among individuals in inpatient/outpatient treatment programs (Sobell *et al.*, 1979; Whitford *et al.*, 2009) or volunteers from the community (Perrine *et al.*, 1995), we are not aware of any research comparing breathalyzer outcomes to self-report measures within a contingency management procedure.

Although reliability and validity of the TLFB, as with other self-report tasks, tend to be adequate to high (Sobell *et al.*, 1988; Sobell and Sobell, 1992), several factors may influence the accuracy of self-reports (Del Boca and Darkes, 2003; Vinson *et al.*, 2003). These factors include social context factors (e.g. assessment settings, cultural norms), respondent characteristics (e.g. dependence severity, recovery stage) and task attributes (e.g. mode of administration). Uncovering the potential factors that contribute to the accuracy of self-report measures is an important avenue of inquiry, especially for measurement of alcohol consumption. Notwithstanding the importance of accurate measurement devices in scientific research, how is one able to determine the efficacy of a treatment program aimed to reduce alcohol consumption if the factors influencing accuracy are unknown? One factor that may impact TLFB accuracy is a change in alcohol use patterns during the time covered by the TLFB, such as that coinciding with a successful treatment episode. However, we are not aware of any research that has examined the accuracy of the TLFB when the window of recall coincides with changes in typical drinking patterns.

We recently demonstrated the feasibility of a remote monitoring system for implementing contingency management (Koffarnus *et al.*, 2018). Briefly, contingency management procedures provide incentives (e.g. money, points) contingent upon certain objectively defined behavioral measures, such as verified abstinence from a substance (Higgins and Silverman, 1999; Petry *et al.*, 2000; Prendergast *et al.*, 2006; Barnett *et al.*, 2011). Contingency management procedures have been shown to be highly efficacious in behavior change studies, including substance use (Prendergast *et al.*, 2006) and medication adherence (Schmitz *et al.*, 1998; Carroll *et al.*, 2001). In our recent study (Koffarnus *et al.*, 2018), heavy-drinking (two heavy-drinking days during the Monitoring Only phase [see below]; ≥4/5 drinks per day for females and males, respectively), treatment-seeking participants with alcohol use disorder were assigned to either a Contingent or Noncontingent group and submitted daily breath alcohol samples using a breathalyzer. The onset of the study began with a 7-day Monitoring Only period where participants reported the number of alcoholic drinks consumed during the past day (past-day self-reports); both groups

received incentives for reporting, but incentive amount was not contingent on negative breath samples during this period. A 21-day Treatment phase followed the monitoring phase. During this phase, participants in the Contingent group received daily incentives for verified abstinence and those in the Noncontingent group received the same daily incentive amounts (the amount was yoked to a previous Contingent participant), but not contingent upon verified abstinence. Following this treatment phase, participants returned to the laboratory and completed an experimenter-led 30-day TLFB (Sobell and Sobell, 1992), which captured drinking recall during both of the previous phases.

Results of this parent study demonstrated the procedure was highly efficacious with 85% abstinence rates in the Contingent group, compared to only 38% in the Noncontingent group during the treatment phase and adherence to daily reporting was high, with rates exceeding 95%. Although drinks per day did not differ across the groups prior to study onset, inspection of daily self-reports and recall on the TLFB suggested potential discrepancies between the two groups. That is, participants in the Contingent group appeared to recall fewer drinks per day on the TLFB compared to their daily self-reports, but only for the one of the study phases, whereas the Noncontingent group displayed relatively consistent correspondence across both phases of the study. As mentioned earlier, several factors can influence the accuracy of self-reports of alcohol use and understanding the myriad variables is vitally important in both research and treatment settings as any intervention aimed at changing drinking patterns needs to reflect actual—not perceived—changes in drinking. In addition, using biochemical measures of recent alcohol use is not always possible, leading researchers to rely on the TLFB or similar self-report measures.

The parent study discussed above is well-suited to compare the accuracy of two different self-report measures (TLFB, past-day self-reports) to biochemically obtained measures of recent alcohol use (breathalyzers). As a secondary analysis of the parent study, the present report investigated the accuracy of retrospective recall on the TLFB compared to daily self-reports, as well as how these reports coincided with daily breathalyzer samples. We compared these variables across treatment group and phase of the parent study. We did not have *a priori* hypotheses about these analyses.

METHODS

Participants

Treatment-seeking adults with alcohol use disorder participated in the parent study (Koffarnus *et al.*, 2018). Of the 40 participants who met randomization criteria in the parent study, all 40 completed the parent study (i.e. no attrition during the Treatment phase). Of those 40, 39 were used in the analyses presented here. A total of 20 and 19 participants' data from the Contingent and Noncontingent groups, respectively, were included. One participant from the Noncontingent group did not complete the relevant TLFB assessment and was therefore not included.

Procedure

The parent study consisted of two phases: a 7-day Monitoring Only phase and a 21-day Treatment phase. Participants also completed three in-lab assessment sessions including one prior to the Monitoring Only phase, one immediately following the Treatment phase, and one occurring 1-month following the Treatment phase. Among other measures obtained during each of these in-lab assessment sessions, participants completed an experimenter-led TLFB for the past 30 days reporting the number of drinks consumed each day. TLFB

instructions suggested that participants recall drinks starting with the most recent day and proceed backwards, but participants could jump around the 30-day period if they desired. For the purposes of this paper, we reanalyzed results from the second assessment session. We encourage readers to consult the parent study for full details.

During both phases, all participants provided daily self-reports of previous-day drinking (i.e. past-day self-reports). Participants were prompted via text message to report the number of drinks they consumed the previous day and received a phone-call prompt if drinks were not reported by early evening. Fewer than 1% of past-day self-report opportunities were missed by participants in either group. Previous-day, as opposed to current-day, drinking was chosen as to reduce the variability between reporting time and drinking patterns throughout the day.

Data analysis

We first examined the correspondence between the self-report measures of drinking (past-day self-reports, TLFB) and breathalyzer results. Because the Monitoring Only phase did not include breathalyzer assessments, we compared these measures during the 3-week Treatment phase. For every participant and for each day, a 0 was assigned if either no drinking was self-reported or if all three breathalyzer samples were negative. In cases where any drinking was either reported or detected by the breathalyzer, a 1 was assigned. Missing reports or samples were analyzed both as missing (i.e. modeled as missing data) or as positive (i.e. interpolated with a value of 1). To determine the correspondence between measures, each day was recorded as a 'hit' (i.e. both self-report measure and breathalyzer were positive), a 'correct rejection' (i.e. both self-report measure and breathalyzer were negative), a 'miss' (i.e. self-report measure was negative and breathalyzer was positive), or a 'false alarm' (i.e. self-report measure was positive and breathalyzer was negative). Overall accuracy was calculated as a percentage by summing the number of hits and correct rejections and dividing by the total number of hits, correct rejections, misses and false alarms. Overall accuracy was left-skewed and not amenable to normalizing transformations. Therefore, we first calculated the difference in accuracy between the two measures and compared the Contingent and Noncontingent groups using a Mann–Whitney U test. Given the two groups did not differ with respect to accuracy, we collapsed the two groups and compared whether past-day self-reports were relatively more accurate compared to the TLFB using a Wilcoxon Signed-Rank test.

The number of drinks consumed each day as reported from the past-day self-reports was compared to the number of drinks reported on the TLFB. We analyzed the correspondence between reports by matching the drinking days across the two tasks and for each day we subtracted the number of drinks reported on the past-day self-reports from the TLFB. We then compared how this correspondence (difference in the number of drinks reported) differed across the Monitoring Only and Treatment phases. Data were analyzed in R Statistical Software Version 3.5.1 (R Core Team, 2018) using the *geepack* package (Halekoh *et al.*, 2006). Generalized estimating equation (GEE) was specified with an autoregressive(1) correlation structure to account for the intra-subject correlations across days (repeated measures; Liang and Zeger, 1986). GEEs are preferred over other repeated-measures analyses (e.g. ANOVA) because they can be fitted with specified correlation structures (e.g. AR(1)) and are relatively robust against model misspecification. GEE was chosen over a linear mixed-effects model because we were not interested in explicitly modeling correspondence across each day, but were interested in accounting for the intra-subject correlations across days. Unstandardized beta weights and sandwiched standard errors are reported. Using estimated marginal means (*emmeans* package; Lenth *et al.*, 2018), we examined main effects and interaction of group (Contingent vs. Noncontingent) and phase (Monitoring Only vs. Treatment) on the difference in drinks reported (i.e. drinks on the TLFB – drinks on the daily diaries). We also examined correspondence between the two measures using Spearman rank order correlations (ρ). Effects were considered significant at the $\alpha = 0.05$ level.

RESULTS

Participant demographics

Table 1 displays participants' baseline demographic variables. No differences were observed for any of the variables between the two groups.

Correspondence between self-report and breathalyzer measures

Missing–missing analyses

We observed no group differences (Contingent vs. Noncontingent) with respect to differential accuracy of the self-report and breathalyzer measures ($W = 172.5$, $P = 0.628$). After collapsing across groups, correspondence between past-day self-reports was significantly more accurate than TLFB self-reports when compared to contemporaneous breathalyzer measures ($V = 337$, $P < 0.001$). Specifically, for the contingent group the median accuracy was 92.2% ($M = 69.5\%$, SD = 35.8%) and 70.7% ($M = 65.1\%$, SD = 34.8%) for the past-day self-reports and TLFB, respectively. For the Noncontingent group, the median accuracy was 83.3% ($M = 74.0\%$, SD = 29.4%) and 71.4% ($M = 67.7\%$, SD = 29.5%) for the past-day self-reports and TLFB, respectively.

Table 1. Participant demographics

Group (n) Demographic variable	Contingent (20)	Noncontingent (19) Mean (SD)/n (%)	P-value
Age (years)	46.55 (12.53)	45.32 (11.83)	0.75
Monthly Income ($USD)	2789.55 (2260.41)	2688.42 (2400.68)	0.89
Education (years)	14.50 (2.54)	14.58 (2.19)	0.92
Sex (Male)	13 (65.0)	14 (73.7)	0.81
Race (White)	16 (80.0)	18 (94.7)	0.37
Ethnicity (non-Hispanic)	20 (100.0)	19 (100.0)	–
Drinks per day	6.48 (2.77)	6.25 (4.25)	0.84
Heavy drinking (years)	21.10 (9.98)	20.62 (10.85)	0.89
Collection rate	97% (4%)	96% (4%)	0.40

Missing-positive analyses

When missing values were imputed as positive signals in the signal detection analyses, results were largely similar. We observed no group differences with respect to differential accuracy of the self-report and breathalyzer measures ($W = 170$, $P = 0.60$). We did observe that accuracy of past-day self-reports were greater than that of the TLFB ($V = 390$, $P = 0.002$). Specifically, for the Contingent group the median accuracy was 87.5% ($M = 68.9\%$, SD $= 34.5\%$) and 69% ($M = 64.8\%$, SD $= 33.5\%$) on the past-day self-reports and TLFB, respectively. For the Noncontingent group, the median accuracy was 80% ($M = 73.9\%$, SD $= 29.4\%$) and 71.4% ($M = 67.4\%$, SD $= 29.7\%$) on the past-day self-reports and TLFB, respectively.

Correspondence between self-report measures across phases

Figure 1 displays the number of drinks consumed per day as reported on the past-day self-reports compared to the TLFB during the two phases. During the Treatment phase, both groups consistently recalled their number of drinks consumed each day (Contingent $\rho = 0.80$; Noncontingent $\rho = 0.76$), despite the large difference in mean drinks per day between the groups (Koffarnus *et al.*, 2018); the Contingent group reported a mean of 1.87 (SD $=$ 2.86) and 1.75 (SD $= 2.96$) drinks per day on the past-day self-reports and TLFB, respectively, whereas the Noncontingent group reported a mean of 5.88 (SD $= 4.94$) and 5.08 (SD $= 4.61$) drinks per day on the past-day self-reports and TLFB, respectively. However, during the Monitoring Only phase, participants in the Contingent group displayed poorer accuracy due to reliably reporting fewer drinks per day on the TLFB as compared to their daily reports ($\rho = 0.47$). The Noncontingent group tended to have better consistency in their reporting across the two self-report measures

during this Monitoring Only phase ($\rho = 0.68$). Here, we also note that consistent with many contingency management applications, the Noncontingent group displayed greater variability in the number of drinks reported regardless of self-report method during the Treatment phase.

Results of the GEE suggested a significant interaction between group and phase on the difference in drinks reported between the TLFB and past-day self-reports (Fig. 2; $b = 2.45$, SE $= 0.78$, $\chi^2 =$ 9.82, $P = 0.002$). During the Monitoring Only phase, the Contingent group reported significantly fewer drinks per day on the TLFB as compared to the past-day self-reports by a mean of 2.39 (SE $= 0.64$) drinks per day compared to the Noncontingent group, who underreported drinks per day by a mean of only 0.43 (SE $=$ 0.39). This difference (-1.96, SE $= 0.72$) was statistically significant ($z = 2.71$, $P = 0.007$). Additionally, the Contingent group reported significantly fewer drinks on the TLFB compared to the past-day self-reports during the Monitoring Only phase ($M = 2.39$; SE $=$ 0.64) compared to the Treatment phase, when they reported fewer drinks by a mean of 0.10 drinks per day (SE $= 0.15$). This difference (-2.29, SE $= 0.70$) was also statistically significant ($z = 3.29$, $P = 0.001$).

DISCUSSION

Self-report measures have typically been shown to be valid and reliable in terms of measuring alcohol consumption. Here, we demonstrated two findings. First, we observed greater correspondence between the past-day self-reports and breathalyzer measures compared to the TLFB during the Treatment phase, regardless of whether participants were in the Contingent or Noncontingent group. To our knowledge, this is the first study to directly compare breathalyzer results to two self-report variants among individuals in a contingency management procedure. Second, participants who
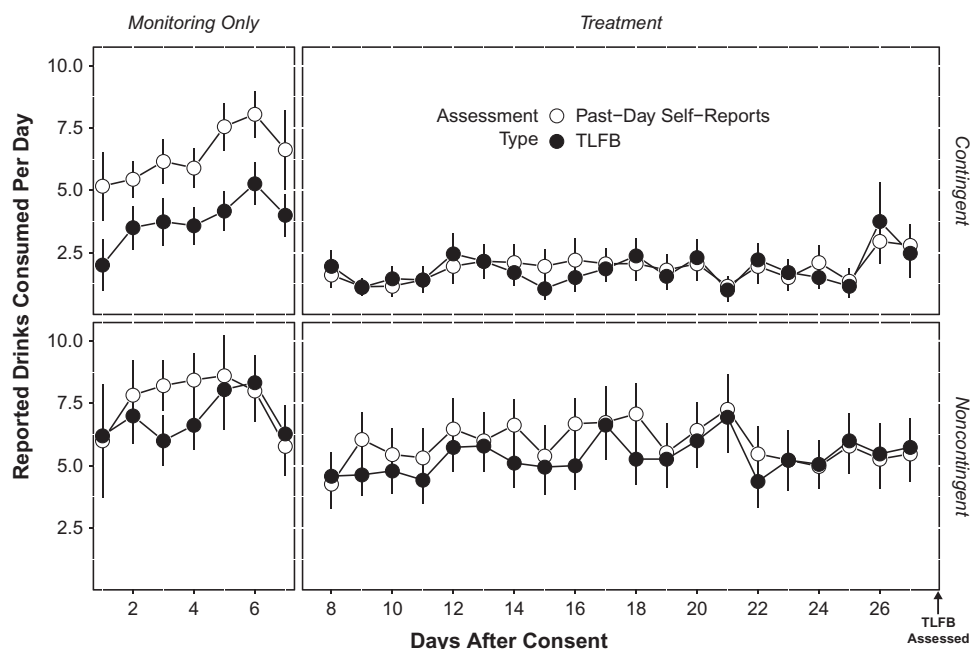

Fig. 1 Reported number of alcoholic drinks consumed per day during the monitoring only and treatment phases. The top panel displays participants in the contingent group, whereas the bottom panel displays participants in the noncontingent group. White circles indicate past-day self-reports and black circles indicate timeline followback (TLFB), with associated standard errors. Participants in the contingent group systematically reported fewer drinks consumed each day on the TLFB compared to past-day self-reports, but only during the monitoring only phase.
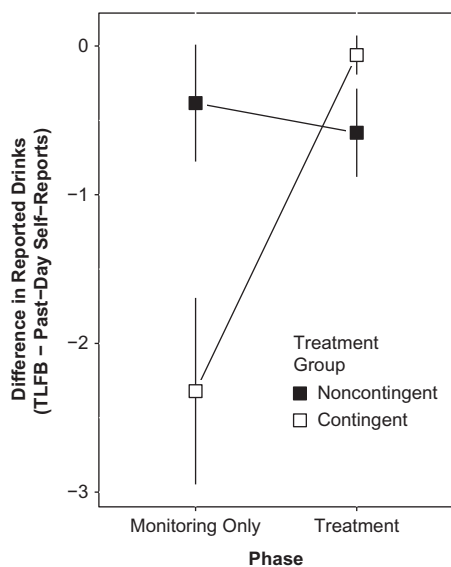
Fig. 2 Differences in the reported number of alcohol drinks consumed between the timeline followback (TLFB) and past-day self-reports across the monitoring only and treatment phases. Black squares indicate participants in the noncontingent Group, whereas white squares indicate participants in the contingent group. Error bars indicate standard errors. Participants in the contingent group significantly underreported the average number of drinks consumed each day during the monitoring only phase.

were provided incentives contingent upon abstinence from alcohol (Contingent group) consistently recalled drinking fewer drinks during the initial phase (Monitoring Only) of the study prior to the incentive intervention compared to those who were assigned to a Noncontingent group. Apart from this significant interaction, TLFB reports were moderately congruent with daily self-report measures, with weaker correlations observed during the Monitoring Only phase (i.e. more remote) compared to the Treatment phase (i.e. more recent), a finding consistent with the literature (Searles *et al.*, 2000; Vinson *et al.*, 2003; c.f. Searles *et al.*, 2002).

Another finding consistent with the literature is that participants, regardless of group assignment or condition, displayed a slight trend towards reporting fewer drinks on the TLFB as compared to the past-day self-reports (Searles *et al.*, 2002; Toll *et al.*, 2006). Our interpretations of the slight underreporting are consistent with those raised by Searles *et al.* (2002). Specifically, concealment of their drinking patterns is an unlikely explanation because all participants were aware that they were providing daily reports and breathalyzer samples. In addition, the relatively small trend to under-report on the TLFB may be a function of participants reporting daily drinking, and as such they were more aware of their daily drinking quantities resulting in greater correspondence between the two measures than what might be expected in a clinical setting. Overall, the moderate to strong correlations observed, albeit with some caveats discussed below, suggest the TLFB is an adequate self-report measure to be used within a contingency management procedure.

As mentioned previously, participants in the Contingent group systematically underreported their alcohol use during the Monitoring Only phase prior to any intervention, demonstrating that accuracy of retrospective recall on the TLFB may be influenced by changes in alcohol consumption patterns in the preceding 30 days. Social contextual factors and task variables are unlikely to

explain the significant interaction we observed here. One possible explanation for the discrepancy in recall is that participants in the Contingent group could have been not paying attention when completing the TLFB, misremembering when their drinking patterns changed. Remembering this shift in drinking patterns may have been especially difficult if participants were completing the TLFB by generalizing their typical or current pattern of drinking instead of attempting to remember drink quantity each individual day. The TLFB was administered by a trained research assistant who prompted participants to identify significant events that may have affected drinking patterns, reducing the likelihood that participants were inattentive to changes in drinking pattern. However, specific prompts catered to that participant (i.e. 'remember that you began the intervention phase of this experiment on X date') were not provided. Another contributor to these effects may be due to a memory bias or memory heuristic. Research has shown that present beliefs or behavior patterns can influence recall of past behavior, especially when memory of that past behavior is imprecise (Ross, 1989; Hammersley, 1994; Stone *et al.*, 2000). This pattern has been shown to include at least some addiction-related memories, with smokers who quit and subsequently relapse underestimating their previous dedication to quitting (Shiffman *et al.*, 1997). To our knowledge, this type of memory bias has not previously been assessed regarding drug use quantity preceding a successful quit attempt. Future research should more directly examine the mechanisms underlying this bias to retrospectively report fewer drinks per day after quitting.

The current findings may be especially relevant for researchers or clinicians who need to maximize the accuracy of measuring alcohol consumption via self-reports, especially when daily monitoring is not an option. First, our results show that daily self-reports should be used when possible, as these were significantly more accurate than the TLFB when compared with a biochemically verified measure of recent alcohol consumption (i.e. breathalyzer). Second, if daily monitoring techniques are not possible, then in the context of a study that is expected to capture changes in drinking patterns, our findings suggest that retrospective recall (e.g. TLFB) be conducted prior to expected changes in drinking. After a successful cessation attempt or other change in drinking, individuals may not be able to accurately recall their drinking patterns prior to the change in drinking patterns.

Limitations and future directions

Several limitations of the current study should be noted. First, this study was not designed to assess the validity of the TLFB and we did not obtain TLFB responses immediately prior to the start of treatment. Second, because our relatively modest sample size ($n = 39$) limits robust generalizations, future research should replicate the current findings among a broader sample of participants, perhaps with additional measures to explore the possibility of memory biases or heuristics affecting retrospective recall. Relatedly, the parent study recruited heavy-drinking, treatment-seeking individuals with alcohol use disorder. Future studies would benefit from recruiting drinkers who do not display alcohol use disorder and/or are not seeking treatment.

In summary, correspondence between TLFB and past-day self-report measures support the general utility of the TLFB as a method of obtaining retrospective measures of alcohol use, although the TLFB may not be a reliable indicator of alcohol use prior to a substantial change in use patterns, such as that coinciding with a successful treatment episode. When possible, daily monitoring techniques

should be used as our results suggest high correspondence between daily reporting and a biochemically based measure (i.e. breathalyzer).

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

Barnett NP, Tidey J, Murphy JG, et al. (2011) Contingency management for alcohol use reduction: a pilot study using a transdermal alcohol sensor. *Drug Alcohol Depend* **118**:391–9.

Carney MA, Tennen H, Affleck G, et al. (1998) Levels and patterns of alcohol consumption using timeline follow-back, daily diaries and real-time 'electronic interviews'. *J Stud Alcohol* **59**:447–54.

Carroll KM, Ball SA, Nich C, et al. (2001) Targeting behavioral therapies to enhance naltrexone treatment of opioid dependence. *Arch Gen Psychiatry* **58**:755.

Collins RL, Kashdan TB, Gollnisch G. (2003) The feasibility of using cellular phones to collect ecological momentary assessment data: application to alcohol consumption. *Exp Clin Psychopharmacol* **11**:73–8.

Corkrey R, Parkinson L. (2002) Interactive voice response: review of studies 1989–2000. *Behav Res Methods Instrum Comput* **34**:342–53.

Del Boca FK, Darkes J. (2003) The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction* **98**:1–12.

Halekoh U, Højsgaard S, Yan J. (2006) The R Package geepack for generalized estimating equations. *J Stat Softw* **15**:1–11.

Hammersley R. (1994) A digest of memory phenomena for addiction research. *Addiction* **89**:283–93.

Higgins ST, Silverman K (eds). (1999) *Motivating Behavior Change Among Illicit-drug Abusers: Research on Contingency Management Interventions*. Washington: American Psychological Association.

Koffarnus MN, Bickel WK, Kablinger AS. (2018) Remote alcohol monitoring to facilitate incentive-based treatment for alcohol use disorder: a randomized trial. *Alcohol Clin Exp Res* **42**:2423–31.

Lenth R, Love J, Herve M. (2018) Estimated Marginal Means, aka Least-Squares Means.

Liang K, Zeger SL. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**:13–22.

O'Farrell TJ, Maisto SA. (1987) The utility of self-report and biological measures of alcohol consumption in alcoholism treatment outcome studies. *Adv Behav Res Ther* **9**:91–125.

Perrine MW, Mundt JC, Searles JS, et al. (1995) Validation of daily self-reported alcohol consumption using interactive voice response (IVR) technology. *J Stud Alcohol* **56**:487–90.

Petry NM, Martin B, Cooney JL, et al. (2000) Give them prizes and they will come: contingency management for treatment of alcohol dependence. *J Consult Clin Psychol* **68**:250–7.

Prendergast M, Podus D, Finney J, et al. (2006) Contingency management for treatment of substance use disorders: a meta-analysis. *Addiction* **101**: 1546–60.

R Core Team. (2018) R: A Language and Environment for Statistical Computing.

Ross M. (1989) Relation of implicit theories to the construction of personal histories. *Psychol Rev* **96**:341–57.

Schmitz JM, Rhoades HM, Elk R, et al. (1998) Medication take-home doses and contingency management. *Exp Clin Psychopharmacol* **6**:162–8.

Searles JS, Helzer JE, Rose GL, et al. (2002) Concurrent and retrospective reports of alcohol consumption across 30, 90 and 366 days: interactive voice response compared with the timeline follow back. *J Stud Alcohol* **63**:352–62.

Searles JS, Helzer JE, Walter DE. (2000) Comparison of drinking patterns measured by daily reports and timeline follow back. *Psychol Addict Behav* **14**:277–86.

Shiffman S. (2009) Ecological momentary assessment (EMA) in studies of substance use. *Psychol Assess* **21**:486–97.

Shiffman S, Hufford M, Hickcox M, et al. (1997) Remember that? A comparison of real-time versus retrospective recall of smoking lapses. *J Consult Clin Psychol* **65**:292–300.

Shiffman S, Stone AA, Hufford MR. (2008) Ecological momentary assessment. *Annu Rev Clin Psychol* **4**:1–32.

Simpson CA, Xie L, Blum ER, et al. (2011) Agreement between prospective interactive voice response telephone reporting and structured recall reports of risk behaviors in rural substance users living with HIV/AIDS. *Psychol Addict Behav* **25**:185–90.

Sobell LC, Sobell MB. (1992) Timeline follow-back: a technique for assessing self-reported alcohol consumption. In Litten RZ, Allen JP (eds). *Measuring Alcohol Consumption: Psychosocial and Biological Methods*. Totowa, NJ: Humana Press, 41–72.

Sobell LC, Sobell MB, Leo GI, et al. (1988) Reliability of a Timeline Method: assessing normal drinkers' reports of recent drinking and a comparative evaluation across several populations. *Br J Addict* **83**:393–402.

Sobell MB, Sobell LC, VanderSpek R. (1979) Relationships among clinical judgment, self-report, and breath-analysis measures of intoxication in alcoholics. *J Consult Clin Psychol* **47**:204–6.

Stone AA, Turkkan JS, Bachrach CA, et al (eds). (2000) *The Science of Self-report: Implications for Research and Practice*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Toll BA, Cooney NL, McKee SA, et al. (2006) Correspondence between Interactive Voice Response (IVR) and Timeline Followback (TLFB) reports of drinking behavior. *Addict Behav* **31**:726–31.

Tucker JA, Foushee HR, Black BC, et al. (2007) Agreement between prospective interactive voice response self-monitoring and structured retrospective reports of drinking and contextual variables during natural resolution attempts. *J Stud Alcohol Drugs* **68**:538–42.

Tucker JA, Roth DL, Huang J, et al. (2012) Effects of interactive voice response self-monitoring on natural resolution of drinking problems: utilization and behavioral economic factors. *J Stud Alcohol Drugs* **73**: 686–98.

Vinson DC, Reidinger C, Wilcosky T. (2003) Factors affecting the validity of a Timeline Follow-Back interview. *J Stud Alcohol* **64**:733–40.

Whitford JL, Widner SC, Mellick D, et al. (2009) Self-report of drinking compared to objective markers of alcohol consumption. *Am J Drug Alcohol Abuse* **35**:55–8.