Journal of Studies on Alcohol, Vol. 46, No. 3, 1985

Reliability of Self-Reported Alcohol Consumption in a General Population Survey*

GERALD D. WILLIAMS, SHERRIE S. AITKEN AND HENRY MALIN†

Alcohol Epidemiologic Data System, Suite 600, 1400 Eye Street NW, Washington, D.C. 20005

ABSTRACT. Reliability estimates for the measures of alcohol consumption used in a household survey were examined. Respondents who had not consumed any alcohol within the previous year and those under age 18 were excluded from the survey, yielding a sample of 1395 (48% men). With one survey that involved a 2-week recall period and another that involved a 4-week period, three methods for estimating reliability were employed: alternate forms, test-retest and a combined method. Validity was also examined using a 30-day drinking diary as a criterion. The findings indicate high levels of reliability, averaging .91 for the consumption measures of beer, wine and distilled spirits. Validity estimates were also fairly substantial but not uniformly so across the different beverage types. In general, the results suggest that these consumption measures can be used with considerable confidence. (*J. Stud. Alcohol* **46:** 223-227, 1985)

THERE ARE few, if any, feasible alternatives to self-report measures in assessing variation in alcohol consumption among individuals and groups in the general population. Concerns have been expressed (Armor et al., 1978; Ebon Research Systems, 1979; Polich, 1982) about the adequacy of such measures but relatively few reliability or validity studies on survey measures have been reported. The major purpose of the present study was to provide reliability estimates for measures of drinking used in a household survey. Also, validity was explored using a 30-day drinking diary as a criterion measure. The primary focus of the study was the statistical evaluation of two separate instruments used to assess the alcohol consumption of individuals over a period of 2 or 4 weeks.

The concepts of reliability and validity have been defined in various ways (Anastasi, 1968; Armor et al., 1978). Technical or statistical definitions, in particular, often vary according to the kinds of reliability and validity being investigated, or the definition of error variance or measurement error (Armor et al., 1978). Reliability typically refers to the consistency of scores obtained by the same individuals when retested with the same instrument on different occasions, with different sets of equivalent items or under varying examining conditions (Anastasi, 1968). Validity is commonly used in two senses: the usefulness of scores for predicting future behavior and the agreement between scores and some external criterion measuring the same behavior. In both cases, validity relates to the specific construct being measured, here alcohol consumption, and to the use to be made of the instrument.

Despite the somewhat limited literature on survey consumption measures, there are some studies that are relevant to the present investigation. In particular some encouraging results have been obtained with respect to reliability. In a sample of 80 drinkers in London, Edwards et al. (1973) obtained test-retest reliabilities ranging from .76 to .82 for various drinking measures. In a longitudinal study of drug use among high-school students in New York State, Single et al. (1975) obtained test-retest reliabilities of .79 for distilled spirits and .67 for wine or beer from two administrations of the same questionnaire. In a study of alcohol consumption among 78 pregnant women, Streissguth et al. (1976) conducted initial and follow-up interviews 1 week apart and obtained test-retest reliabilities of .90 for the period of pregnancy and .89 for the period prior to pregnancy.

Most of the studies of reliability have focused on problem drinkers (Davidson and Stein, 1983; Holland et al., 1979; Sobell et al., 1979). These and other studies (Cooper et al., 1981; Sobell and Sobell, 1975) usually report satisfactory levels of reliability or validity on self-reported consumption measures used with patients in alcoholism treatment programs. The notable exception was a test-retest reliability study (Summers, 1970) in which 14 of 15 men inpatients of an alcoholism treat-

Received: 2 September 1983. Revision: 19 July 1984.

* This research was supported by National Institute on Alcohol Abuse and Alcoholism contract ADM 281-81-0005.

† Dr. Williams is Senior Associate and Dr. Aitken is Vice President at CSR, Incorporated, Washington, D.C. Mr. Malin is Acting Branch Chief, Surveillance Branch, Division of Biometry and Epidemiology, National Institute on Alcohol Abuse and Alcoholism.

Reprint requests to Dr. Malin at the Alcohol Epidemiologic Data System.

ment program changed at least half of their responses to questions on a lifetime drinking history when reinterviewed after 2 weeks. Sobell and Sobell (1982) and Hesselbrock et al. (1983) have provided reviews of measurement problems and other problems related to this particular segment of the drinking population. However, the relevance of such studies for survey instruments designed to measure drinking in the general population is questionable.

The relevant literature pertaining to the validity of survey measures on alcohol consumption is even more limited. When Jessor et al. (1968) validated a quantity-frequency (Q-F) measure against the external criterion of the self-reported number of times intoxicated, they obtained correlations of .60 for adults and .71 for high-school students. Similarly, Garrett and Bahr (1974) validated a Q-F measure against a self-rating criterion, and obtained average correlations of .61 for men and .82 for women.

Other investigators have conducted studies to validate self-reported drinking with measures of blood alcohol level (BAL). Fine et al. (1978) reported a correlation of .11 between self-reported Q-F measures and BAL with a sample of 3837 men who had been arrested for driving while intoxicated. Polich (1982) reported a correlation of .36 between BAL and self-reported consumption for 91 subjects who reported drinking 24 hr prior to an interview. One finding of special interest in Polich's study was that overreporting of alcohol consumption appeared nearly equal to underreporting on the self-report measures.

Still other studies (Gregson and Stacey, 1982; Pernanen, 1974; Popham and Schmidt, 1981) have found little correspondence between self-report estimates and consumption estimates derived from sales data. Such estimates from sales data, however, are aggregate or ecological measures and are not very appropriate as an external criterion for validating an instrument measuring consumption by individuals (Robinson, 1950).

In general, then, many of the reliability and validity studies on self-reported consumption have either (*1*) focused on problem drinkers, e.g., patients in alcoholism treatment programs, (*2*) integrated general consumption items within an overall drinking history or other measures or (*3*) addressed the issues of the adequacy of population estimates of alcohol consumption. The present study was designed specifically to examine the properties of self-reported consumption measures used in a general population survey.

Method

Sample

This study involved a household survey in the Standard Metropolitan Statistical Areas of Greensboro, North Carolina, Baltimore and Chicago. Three hundred sample points were established in 34 different communities and Kish's (1965) selection procedures were employed for identification of all eligible household members and random selection of respondents. A set of demographic questions and screening questions was asked of all sampled respondents so that abstainers (those who had not had a drink of any kind of alcoholic beverage within the past 12 months) and persons under age 18 could be identified and excluded from the study.

The survey procedure yielded a total of 2109 respondents, 714 (34%) of whom were abstainers. A total of 48% of the drinking respondents were men; the age range was 18–94. Of the drinking respondents, 59% were married at the time of the survey, 22% had never been married and 19% were divorced, separated or widowed. Complete details of the sampling methods and respondent characteristics have been reported elsewhere (CSR, Incorporated, 1983).

Overview of study components

The present study compares different reporting periods for alcohol consumption, different questions for calculating alcohol consumption, and different methods for deriving reliability and validity estimates.

Survey instruments

The basic survey data were collected by trained interviewers. Two survey forms were used for past alcohol consumption, one with a 14-day and one with a 28-day recall period. (Self-administered versions of the survey were also used but are not included in the results reported here.) A respondent's date of last drink of an alcoholic beverage was the anchor point to determine the reference period for his self-reported drinking over a 2- or 4-week time interval.

Both the 14-day and 28-day interview forms included quantity-frequency-variability questions about the consumption of beer, wine and distilled spirits. However, only Q-F measures of consumption are reported in this article. A simplified and an extended scoring procedure, respectively, were used to calculate average daily consumption for each beverage type: Estimate 1—total number of drinks of (beer, wine, distilled spirits) during the reporting period × number of ounces of (beer, wine, spirits) in typical drink/number of days in reporting period. Estimate 2—number of days on which respondent drank (beer, wine, spirits) during the reporting period × number of drinks typically consumed on days when respondent drank (beer, wine, spirits) × number of ounces of (beer, wine, spirits) in typical drink/number of days in reporting period. The ounces of each beverage type consumed were converted to absolute alcohol content using multipliers of .04 for beer, .15 for wine and .45 for distilled spirits (Armor et al., 1978). Estimates of total consumption of absolute alcohol from all beverage types combined were also derived from the survey information.

Reliability determination

Three different methods for investigating reliability were used: (*1*) The alternate-forms method involved the administration of two different forms during the initial or baseline

survey. For example, respondents who were administered an interview for a 14-day recall period were then requested to complete a 28-day interview immediately after the first interview, and vice versa. The date of last drink reported in the first survey was also used for the second survey to avoid confounding variation representing reliability with variation that might be attributed to different time periods. In other words, the use of different time periods, even over a period of 5–7 days, could involve changes in consumption reports that represent "true" changes in consumption (Armor et al., 1978; Heise, 1969). Reliability was estimated by computing the Pearson product-moment correlations between the measures of consumption derived from the two forms of the survey. (*2*) The test-retest method involved the administration of the same survey form on two occasions. The second survey was administered 5–7 days after the first survey, keeping the anchor point constant. Again, the Pearson product-moment correlation was used to estimate reliability. (*3*) The combined method involved the procedures of both the previous methods—that is, two different survey forms were administered on two occasions separated by 5–7 days.

Validity determination

The usefulness of a validity study depends heavily on the use of an appropriate external criterion. In the present case, external criterion data for alcohol consumption were collected by the diary method (Fuller et al., 1972). Respondents who consented to participate in the diary study were required to complete one of the drinking surveys and, subsequently, keep a record of their daily consumption of beer, wine and distilled spirits for a period of 30 days. It was assumed that the daily recording would be a more accurate measure of alcohol consumption than the 2- or 4-week recall measures. Respondents were provided with a 30-day diary form that requested the number of drinks and the quantity consumed of each beverage type by periods of the day, e.g., before lunch, during lunch, during dinner and after dinner. The division into time periods was used to facilitate daily recall. Interviewers called the diary respondents each week to ensure that drinking was being reported on a daily basis. After 29–31 days, interviewers returned to the respondent's residence to administer a follow-up drinking survey and to collect the diaries. Incentive payments of $20 were given to respondents who completed the 30-day diary. No other procedures employed in this study involved incentive payments.

Two types of validity were investigated—concurrent and predictive. Concurrent validity involves the extent to which a measure corresponds to an external criterion with respect to the same period of time. In the present study, concurrent validity was estimated by the Pearson product-moment correlations between the consumption estimates derived from the follow-up survey and the consumption data derived from the diary. Predictive validity involves the extent to which scores on a predictor measure predict the criterion behavior for some future period. In this study, the predictor measures were the estimates of consumption derived from the initial baseline survey. Criterion data from two sources were used: the measures of consumption derived from the diary and the estimates of consumption derived from the follow-up survey. In each case, validity was estimated by the correlations between the predictor measures and the criterion measures. The correlations between initial and follow-up survey measures are regarded as estimates of validity rather than reliability because the surveys involved consumption during two different time periods that were separated by a minimum of 30 days.

Results

The estimates of reliability computed by the alternate-forms, test-retest and combined methods are all high (Table 1). The test-retest reliabilities for consumption over a 28-day reporting period tend to be slightly higher than the corresponding reliabilities for a 14-day period. In general, however, the results suggest no appreciable differences in reliability between the two procedures for estimating alcohol consumption or between the 14-day and 28-day reporting periods. The average reliability across the different beverage types for each method of determining reliability is .91. In short, the reliability estimates suggest that the various measures of alcohol consumption can be used with considerable confidence and that they can be used interchangeably.

In general, the estimates of the concurrent and predictive validity of the consumption measures are fairly substantial (Table 2). With the exception of beer consumption on the 14-day form, the concurrent validities average about .80. The differences of the concurrent validities on 14-day vs 28-day forms for total consumption reflect the low validities for beer consumption. Similarly, the predictive validities of the consumption measures were fairly good except for the measure of beer consumption on the 14-day form. It should be noted, however, that the predictions are based on a relatively short time lapse, i.e., a minimum of 30 days. Thus, the estimates of predictive validity are probably higher than would be the case for a longer time period. A correlation between the first and second 2 weeks of alcohol consumption reported on the diary indicated a

TABLE 1. Reliabilities computed by various methods for two procedures for estimating alcohol consumption (Estimates 1 and 2)

Period covered by first survey form	Type of reliability: Alternate forms Est.1	Alternate forms Est.2	Test-retest Est.1	Test-retest Est.2	Combined Est.1	Combined Est.2
14 Days	(N = 80)		(N = 79)		(N = 91)	
Beer	.95	.93	.88	.89	.96	.96
Wine	.91	.91	.90	.99	.85	.78
Distilled spirits	.84	.97	.88	.89	.94	.91
Total	.86	.93	.87	.84	.95	.92
28 Days	(N = 85)		(N = 100)		(N = 68)	
Beer	.97	.98	.94	.92	.87	.74
Wine	.94	.93	.99	.99	.86	.90
Distilled spirits	.91	.92	.99	.98	.79	.84
Total	.96	.97	.98	.98	.86	.84

TABLE 2. Concurrent and predictive validities for two procedures for estimating alcohol consumption [a]

Period covered by survey	Concurrent validity with diary as criterion		Predictive validity with Diary as criterion		Predictive validity with Follow-up survey as criterion	
	Est. 1	Est. 2	Est. 1	Est. 2	Est. 1	Est. 2
14 Days (N = 105)						
Beer	.36	.42	.38	.30	.80	.78
Wine	.80	.82	.95	.95	.87	.92
Distilled spirits	.83	.82	.80	.90	.68	.75
Total	.55	.57	.60	.56	.63	.68
28 Days (N = 98)						
Beer	.92	.77	.75	.76	.76	.71
Wine	.83	.72	.75	.71	.76	.78
Distilled spirits	.78	.79	.77	.78	.49	.40
Total	.72	.82	.66	.70	.48	.44

[a] Concurrent validities were estimated with the follow-up survey, which entailed a reporting period included in the diary. The predictive validities predicted alcohol consumption for a period of 1 month following the initial, or baseline, survey.

corrected, split-half reliability estimate of .99 for each type of beverage and total consumption.

Another issue in evaluating the reliability and validity of these measures is the extent to which they yield comparable estimates of average daily consumption. The consumption averages in Table 3 were rounded to the nearest tenth of an ounce. (It seems unlikely that these measures yield more precise estimates of consumption than the nearest tenth.) In general, the various survey estimates are fairly close to each other and to the diary measures. Moreover, the variability associated with these measures was quite large, and in most cases the standard deviations were about twice their corresponding averages. Consequently, for consumption of each type of alcoholic beverage and for total consumption, no significant differences were found among the various estimates.

TABLE 3. Average daily alcohol consumption as measured by various procedures, in ounces of absolute alcohol [a]

Period covered by initial survey	Initial survey form		Follow-up survey form		
	Est. 1	Est. 2	Est. 1	Est. 2	Diary
14 Days (N = 105)					
Beer	0.4	0.5	0.5	0.6	0.4
Wine	0.3	0.2	0.3	0.3	0.3
Distilled spirits	0.3	0.3	0.4	0.4	0.4
Total	0.9	1.0	1.2	1.2	1.0
28 Days (N = 98)					
Beer	0.4	0.4	0.3	0.4	0.4
Wine	0.1	0.1	0.1	0.1	0.1
Distilled spirits	0.4	0.4	0.3	0.3	0.3
Total	0.9	0.9	0.7	0.8	0.9

[a] Consumption was rounded to the nearest tenth of an ounce. The averages for all beverage types, therefore, do not always add up to the total average.

Discussion

The reliability estimates provided in this article give a fairly comprehensive view of the reliability of self-reported drinking, at least for the survey measures employed here. The overall results relating to the reliability of these measures should be encouraging for researchers interested in correlational studies on alcohol use or epidemiological studies. The validity estimates are also generally within the upper range of validity coefficients found in much of the psychometric research, especially research on predictive validities. However, the validity estimates are somewhat variable and may be a bit smaller than would be optimal. Also, the concurrent validity estimates may have been influenced by subjects who were sensitized to their drinking while completing the diary.

In terms of the reliability and validity of the different procedures employed in this study, no appreciable differences were found that were not likely to be the product of some artifact, e.g., missing data on one or more of the drinking measures. Further, most of the results for the 28-day forms—contrary to expectations—did not appear appreciably different from those for the 14-day forms. The notable exceptions were the relatively low validity estimates for reported beer consumption on the 14-day form. Also, the slightly lower predictive validities using the follow-up survey as a criterion measure may have been influenced by the reactivity of the diary respondents to the monitoring of their daily drinking. Overall, however, these findings fail to support the advantages that are usually cited to advocate the use of a longer reporting period in assessing individual alcohol consumption. Since the difference between 2 and 4 weeks, however, is relatively small, such a conclusion may not generalize to longer time intervals, e.g., 6 months or a year.

In interpreting the results of this study, it must be remembered that the standard deviations on the self-reported measures of alcohol consumption tended to be quite large (on the average about twice the size of the mean values). Consequently, very large samples are required to find statistically significant differences among various samples of drinkers in the general population. The relatively small samples used in this multifaceted approach to investigating reliability and validity may have prevented the detection of real differences among the survey forms and the methods used for determining reliability and validity.

Finally, the gist of these findings is that the survey instruments used tend to be relatively reliable and valid for a general survey on alcohol consumption. Overall, these findings should be viewed positively because they suggest that the survey instruments (and the consumption measures derived from the instruments) can be

used with considerable confidence and that they can be used interchangeably in accordance with the demands of a given investigation.

References

ANASTASI, A. Psychological Testing, New York: Macmillan Publishing Co., Inc., 1968.

ARMOR, D. J., POLICH, J. M. AND STAMBUL, H. B. Reliability and validity of self-reported drinking behavior. In: ARMOR, D. J., POLICH, J. M. AND STAMBUL, H. B. Alcoholism and Treatment, New York: John Wiley & Sons, Inc., 1978, pp. 173–211.

COOPER, A. M., SOBELL, M. B., SOBELL, L. C. AND MAISTO, S. A. Validity of alcoholics' self-reports: Duration data. Int. J. Addict. **16:** 401–406, 1981.

CSR, INCORPORATED. Validity/Reliability Study of Self-Reported Drinking. Final report. Prepared for the National Institute on Alcohol Abuse and Alcoholism, NTIS No. PB84-181429, Springfield, Va.: National Technical Information Service, 1983.

DAVIDSON, R. S. AND STEIN, S. Reliability of self-report of alcoholics. Behav. Mod. **6:** 107–119, 1983.

EBON RESEARCH SYSTEMS. Summary of the Proceedings of the Alcohol Epidemiology Questionnaire Item Review and Evaluation Meeting. Prepared for the National Institute on Alcohol Abuse and Alcoholism, Contract No. ADM 281-79-0015, Washington, D.C.: U.S. Government Printing Office, 1979.

EDWARDS, G., HENSMAN, C. AND PETO, J. Drinking in a London suburb: Reinterview of a subsample and assessment of response consistency. Q. J. Stud. Alcohol **34:** 1244–1254, 1973.

FINE, E. W., STEER, R. A. AND SCOLES, P. E. Relationship between blood alcohol concentration and self-reported drinking behavior. J. Stud. Alcohol **39:** 466–472, 1978.

FULLER, R. K., BEEB, H. T., LITTELL, A. S., HOUSER, H. B. AND WITSCHI, J. C. Drinking practices recorded by a diary method. Q. J. Stud. Alcohol **33:** 1106–1121, 1972.

GARRETT, G. R. AND BAHR, H. M. Comparison of self-rating and quantity-frequency measures of drinking. Q. J. Stud. Alcohol **35:** 1294–1306, 1974.

GREGSON, R. A. M. AND STACEY, B. G. Self-reported alcohol consumption: A real psychophysical problem. Psychol. Rep. **50:** 1027–1033, 1982.

HEISE, D. R. Separating reliability and stability in test–retest correlation. Am. Sociol. Rev. **34:** 93–101, 1969.

HESSELBROCK, M., BABOR, T. F., HESSELBROCK, V., MEYER, R. E. AND WORKMAN, K. "Never believe an alcoholic"? On the validity of self-report measures of alcohol dependence and related constructs. Int. J. Addict. **18:** 593–609, 1983.

HOLLAND, R., DATTA, K., IZADI, B. AND EVENSON, R. C. Reliability of an alcohol self-report instrument. J. Stud. Alcohol **40:** 142–144, 1979.

JESSOR, R., GRAVES, T. D., HANSON, R. C. AND JESSOR, S. L. Results: Establishing the criterion variables. In: Society, Personality and Deviant Behavior: A Study of a Tri-Ethnic Community, New York: Holt, Rinehart & Winston, Inc., 1968, pp. 166–188.

KISH, L. Survey Sampling, New York: John Wiley & Sons, Inc., 1965.

PERNANEN, K. Validity of survey data on alcohol use. In: GIBBINS, R. J., ISRAEL, Y., KALANT, H., POPHAM, R. E., SCHMIDT, W. AND SMART, R. G. (Eds.) Research Advances in Alcohol and Drug Problems, Vol. 1, New York: John Wiley & Sons, Inc., 1974, pp. 355–374.

POLICH, J. M. The validity of self-reports in alcoholism research. Addict. Behav. **7:** 123–132, 1982.

POPHAM, R. E. AND SCHMIDT, W. Words and deeds: The validity of self-report data on alcohol consumption. J. Stud. Alcohol **42:** 355–358, 1981.

ROBINSON, W. S. Ecological correlations and the behavior of individuals. Am. Sociol. Rev. **15:** 351–357, 1950.

SINGLE, E., KANDEL, D. AND JOHNSON, B. D. The reliability and validity of drug use responses in a large scale longitudinal survey. J. Drug Issues **5:** 426–443, 1975.

SOBELL, L. C., MAISTO, S. A., SOBELL, M. B. AND COOPER, A. M. Reliability of alcohol abusers' self-reports of drinking behavior. Behav. Res. Ther. **17:** 157–160, 1979.

SOBELL, L. C. AND SOBELL, M. B. Outpatient alcoholics give valid reports. J. Nerv. Ment. Dis. **161:** 32–42, 1975.

SOBELL, L. C. AND SOBELL, M. B. Alcoholism treatment outcome evaluation methodology. In: NATIONAL INSTITUTE ON ALCOHOL ABUSE AND ALCOHOLISM. Prevention, Intervention and Treatment: Concerns and Models, Alcohol and Health Monograph No. 3, Washington, D.C.: U.S. Government Printing Office, 1982, pp. 293–321.

STREISSGUTH, A. P., MARTIN, D. C. AND BUFFINGTON, V. E. Test-retest reliability of three scales derived from a quantity-frequency-variability assessment of self-reported alcohol consumption. Ann. N.Y. Acad. Sci. **273:** 458–466, 1976.

SUMMERS, T. Validity of alcoholics' self-reported drinking history. Q. J. Stud. Alcohol **31:** 972–974, 1970.