Sources of Unreliability in the Diagnosis of Substance Dependence*

RICHARD FEINN, PH.D., JOEL GELERNTER, M.D.,[†] JOSEPH F. CUBELLS, M.D., PH.D.,[†] LINDSAY FARRER, PH.D.,[†] AND HENRY R. KRANZLER, M.D.[†]

Clinical Epidemiology Research Center, Veterans Affairs Connecticut Healthcare Center and Yale University School of Medicine, 950 Campbell Avenue, Building 35A, West Haven, Connecticut 06516

ABSTRACT. Objective: The Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) yields Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) diagnoses for a variety of psychiatric disorders, including alcohol and drug dependence. Using generalizability theory, we sought to ascertain the sources of unreliability for DSM-IV substance-dependence diagnoses and their criterion counts. **Method:** Two hundred ninety-three subjects (52.2% women) were interviewed twice over a 2-week period, and a generalizability coefficient and an index of dependability (with confidence intervals) were calculated for each dependence category. **Results:** Overall, there were good-to-excellent reliabilities for the more common diagnoses and criterion counts, including tobacco, alcohol, cocaine, and opioid dependence. The reliabilities were not as good for marijuana dependence and the less common diagnoses of stimulant, sedative, and other drug dependence.

There was greater variability between interviewers (inter-rater reliability) than occasions (test-retest reliability). However, for most diagnoses, the subject by occasion variability was larger than the subject by interviewer variability, indicative of greater consistency in the contribution by interviewers to the ordering of subjects than in the contribution by subjects themselves between the two interviews. **Conclusions:** These results are consistent with prior findings that the SSADDA yields reliable diagnoses and criterion counts for the more prevalent substance-dependence diagnoses. The present analysis extends these findings by showing that the greatest source of unreliability was the subjects' report. This underscores the need for efforts to increase the reliability of substance-dependence diagnoses (and by extension other self-reported phenotypic features) by enhancing the consistency of the information provided by the subjects interviewed. (*J. Stud. Alcohol Drugs* 70: 475-481, 2009)

THE RELIABILITY OF PSYCHIATRIC DIAGNOSES obtained using structured interviews varies considerably (Berney et al., 2002; Bryant et al., 1992; Bucholz et al., 1994; Burke, 1986; Easton et al., 1997; Griffin et al., 1987; Williams et al., 1992). There are many sources of unreliability in the diagnostic process, and most measures of reliability do not capture them all (Brennan, 2001). Classical test theory assesses the reliability of diagnoses using a procedure with a focus limited to a single external source of variance. Generalizability theory, which extends classical test theory, reduces the potential for invalid conclusions and

increases generalizability by partitioning the error variance into distinct sources (facets) (Brennan, 2001).

Specifically, the unreliability of a diagnosis could result from differences among subjects (the target of interest or object of measurement), which in generalizability theory is considered the universe score variance (true variance), as well as from the effects of interviewers, occasions, and the interactions among all of these latter facets. Whereas the absolute error variance stems from the difference in a subject's observed score and the subject's universe score, the relative error variance consists only of the variance terms for which the facets interact with the object of measurement. Thus, absolute error variance will always be at least as large as, and is usually larger than, the relative error variance, because the relative error concerns only deviations (i.e., ranking), whereas absolute error concerns the actual position of scores.

The traditional test-retest correlation for stability is analogous to using relative error, because a correlation coefficient disregards the mean location of a scale. For a criterion reference measure, such as diagnoses derived from an interview procedure, the absolute error should also be taken into account because decisions about diagnosis are made using the absolute score and not just the relative position of one subject to another (Brennan, 2000). The generalizability coefficient is the ratio of universe score variance to

Received: July 21, 2008. Revision: October 6, 2008.

*This research was supported by National Institutes of Health grants DA12690, DA12422, DA12849, DA12890, DA15105, DA018432, and AA13736.

[†]Henry R. Kranzler is with the Department of Psychiatry, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, CT 06030-2103. Correspondence may be sent to him at that address or via email at: kranzler@psychiatry.uchc.edu. Joel Gelernter is with the Departments of Psychiatry, Genetics, and Neurobiology, VA Connecticut Healthcare Center and Yale University School of Medicine, West Haven, CT. Joseph F. Cubells is with the Department of Human Genetics, Emory University School of Medicine, Atlanta, GA. Lindsay Farrer is with the Departments of Medicine (Genetics Program), Neurology, Genetics and Genomics, Epidemiology, and Biostatistics, Boston University Schools of Medicine and Public Health, Boston, MA.

universe score variance plus relative error variance. In generalizability theory, the measure of reliability of the absolute scores is the index of dependability, which is the ratio of the universe score variance to the universe score variance plus absolute error variance. The generalizability coefficient will be at least as large as, and usually larger than, the index of dependability, which makes intuitive sense in that it is more difficult to predict an absolute score than a deviation score (i.e., relative ranking).

Although test-retest reliability provides a measure of stability, it offers no information about the consistency between different raters. A measure of internal consistency (e.g., Cronbach's alpha) provides no information on the stability of diagnoses over time. An advantage of generalizability theory is its ability to estimate both inter-rater and intra-rater reliabilities from the same dataset simultaneously (Cronbach et al., 1963).

We developed a computerized diagnostic instrument, the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA), for use in studies of the genetics of substance use and associated disorders. Using the SSADDA, the test-retest and inter-rater reliabilities of most Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994), substance-dependence diagnoses (Pierucci-Lagha et al., 2005) and diagnostic criteria (Pierucci-Lagha et al., 2007) were good to excellent. This article extends earlier findings by applying generalizability theory to evaluate the internal consistency and stability of substance-dependence diagnoses obtained using the SSADDA and identifying the sources of error variance in their measurement for both categorical diagnoses and continuous diagnostic criterion counts. This information may enable researchers to increase the reliability of SSADDA diagnoses by addressing the identified sources of unreliability.

Method

Sample

A total of 293 subjects, including 159 individuals with substance dependence, 59 individuals with one or more primary psychiatric disorders, and 75 community controls, were recruited from substance-abuse treatment facilities affiliated with the University of Connecticut and Yale University, inpatient and outpatient services of the Department of Psychiatry at the University of Connecticut, and the community through advertisements in local media, respectively. The sampling scheme was chosen to provide a sufficient number of subjects within each diagnostic category to permit analysis, rather than to provide a sample representative of any particular population. Subjects were included in the study if they were willing and able to provide informed consent to participate and were at least age 18. They were excluded if they showed clinical evidence of a severe psychiatric or

medical disorder that could limit their capacity to provide accurate diagnostic information (e.g., schizophrenia, gross cognitive impairment) or if they were unable to read or write English (less than an eighth-grade reading level). Subjects were paid for their participation.

Interviewers

Twelve nonclinician interviewers participated in the study. Each of the sites provided trained interviewers with different levels of experience in the administration of the SSADDA. Before initiation of the study, the interviewers received standardized, intensive training involving practice interviews with individuals recruited for that purpose. Interviewers were then required to administer at least 10 practice interviews before they were allowed to begin data collection. To ensure criterion-level performance, the last three practice interviews were reviewed by the study coordinator, an interviewer with more than 15 years of experience in the administration of such interviews and with more than 10 years of experience in training other interviewers.

Statistical analysis

Both the generalizability coefficient and the index of dependability are reported for the eight substance-dependence diagnoses. In addition, inter-rater and intra-rater reliabilities are estimated by treating certain facets as fixed, which removes the variability in observed scores from facets not relevant to the particular reliability estimate.

Scoring for substance-dependence diagnoses required that at least three of seven possible criteria be met. For each substance evaluated using the SSADDA, the number of criteria met by the subject was recorded on a tally sheet from which the diagnosis was ascertained. Two sets of reliability analyses were performed: one for the dichotomous substance-dependence diagnosis and another for the number of criteria met (tally score), neither of which is normally distributed. The method used to estimate the variance components was MINQUE (minimum norm quadratic unbiased estimation), which makes no distributional assumptions. This approach is recommended by Bell (1985) for the application of generalizability theory to unbalanced designs, which is the case here because each of the 12 interviewers interviewed only a small subset of the subjects (i.e., the subjects were not completely crossed with interviewers and occasions). All analyses were performed using SAS Version 9.1.

Results

Demographic description

Table 1 displays the demographic characteristics of the study participants. As can be seen, there were slightly more

TABLE 1. Demographics ($n = 293$)

Variable	% or mean (SD)
Gender	
Male	47.8%
Female	52.8%
Age, years	37.8 (10.6)
Race	
White	46.8%
Black	38.2%
Latino	7.5%
Other	7.5%
Education, years	13.0 (2.4)
Marital status	
Married	10.6%
D/S/W	29.6%
Never married	59.7%
Employment status	
Working	21.6%
Not working	78.4%
Household income	
<\$10,000	42.4%
\$10,000-19,999	20.6%
\$20,000-39,999	19.1%
\$40,000-74,999	11.6%
≥\$75,000	6.4%

Note: D/S/W = divorced, separated, widowed.

women (53%) than men in the sample, which was generally young (mean age = 38 years), predominantly white (47%) or black (38%) and never married (60%), with a mean of 13 years of education. The sample was largely unemployed (78%) and with an annual income below \$20,000 (63%).

Generalizability analysis

Subjects were interviewed on two separate occasions, either by the same interviewer on both occasions or by two (of 12) different interviewers, one on each occasion. On average (SD), approximately 2 weeks (14.73 [5.34] days) elapsed between the two interviews.

Overall analysis

Table 2 presents the sources of variance attributed to different facets for each of the eight dependence diagnoses. As shown, the possible sources of variance that contribute to a dependence diagnosis are the main effects of subject, interviewer, and occasion, as well as the pair-wise interactions between these three factors. For every substance-dependence diagnosis, except sedative dependence, subjects (the target of measurement) accounted for the most variability in the dependence rating, ranging from 36% (sedative) to 88% (opioid), with an average of 63% across the eight substance-dependence diagnoses. The next largest source of variance was attributed to the interaction between the subject and the occasion, with an average of 23% across the eight substance-dependence diagnoses. This was followed by the variability attributed to the interviewers, wherein the average was 8% across the diagnoses.

The variability stemming from the interaction between subjects and occasions indicates that the subjects were not entirely consistent in their responses between the two interviews, a source of error variance that is problematic to control from the researcher's standpoint. The generally smaller amount of variability attributed to the main effect of interviewers reflects the modest discrepancy in how different interviewers rated the same subject. This was not true for all substances, however, because the variability attributable to the interviewer was the second largest source of variance for both tobacco and cocaine dependence.

The second-to-bottom row of Table 2 lists the generalizability coefficient (Ep^2), and the bottom row lists the index of dependability (Φ). The Ep^2 is similar to the often-used internal reliability coefficient α and denotes the consistency in the relative scores of the subjects. The Ep^2 ranges from poor (sedative) to excellent (opioid), with an average of .70 across the eight dependence diagnoses, which is good

TABLE 2. Component variances and percentage of variance (%) for dependence diagnoses

Subscale	Tobacco	Alcohol	Marijuana	Cocaine	Opioid	Stimulant	Sedative	Other
Subject	.195 (77)	.159 (68)	.096 (58)	.217 (70)	.117 (88)	.016 (52)	.011 (36)	.041 (55)
Interviewer	.037 (15)	.013 (6)	.007 (4)	.071 (23)	.007 (5)	.001 (3)	.000 (0)	.004 (5)
Occasion	.000 (0)	.001 (0)	.000 (0)	.000 (0)	.000 (0)	.000 (0)	.000 (0)	.000 (0)
Subj. $\times$ Inter.	.015 (6)	.026 (11)	.000 (0)	.000 (0)	.000 (0)	.000 (0)	.000 (0)	.022 (29)
Subj. $\times$ Occas.	.007 (3)	.031 (13)	.059 (36)	.021 (7)	.004 (3)	.014 (45)	.020 (65)	.008 (11)
Inter. $\times$ Occas.	.000 (0)	.005 (2)	.003 (2)	.000 (0)	.005 (4)	.000 (0)	.000 (0)	.000 (0)
Ep^2	.899	.736	.619	.911	.967	.533	.355	.577
Φ	.768	.677	.582	.700	.880	.516	.355	.547

Notes: Subj. $\times$ Inter. = Subject $\times$ Interviewer; Subj. $\times$ Occas. = Subject $\times$ Occasion; Inter. $\times$ Occas. = Interviewer $\times$ Occasion; Ep^2 = generalizability coefficient; Φ = index of dependability.

(particularly in view of the dichotomous nature of the diagnoses). The Φ , which is useful for making absolute decisions rather than relative comparisons, ranges from poor (sedative) to very good (opioid), with an average of .63 across the eight dependence diagnoses. Both measures (Ep^2 and Φ) indicate that the SSADDA diagnoses of tobacco, alcohol, cocaine, and opioid dependence show good to excellent reliability, whereas the diagnoses of marijuana, stimulant, and other drug dependence are fair, and the diagnosis of sedative dependence is poor.

Table 3 displays the sources of variance in the sum scores from the tally sheet, which range from 0 to 7. The sum scores provide a more informative measure of the reliability of the SSADDA than the dichotomous dependence classifications and are useful for developing confidence intervals. Not surprisingly, as Table 3 shows, there is more variability in scores than in the dependence classification measures from Table 2. Although the proportion of variability attributed to each source is similar to the results from Table 2, the measures of reliability are generally higher. This is because the sum—which scores the proportion of variability attributed to subjects—is slightly higher (an average of 66% across the eight substances), and the proportion attributed to the interaction between subjects and occasions is smaller (14% across substances). Similar to that for the dependence diagnoses, the average amount of variability attributable to interviewers is 8%. As the bottom two rows in Table 3 show, the generalizability coefficients and the indices of dependability are higher than the analogous measures from the dependence diagnoses in Table 2. The Ep^2 ranges from .47 (sedative) to .98 (cocaine), with an average of .77 across substances, whereas the Φ ranges from .46 (sedative) to .85 (opioid), with an average of .68 across substances. The two indices indicate good reliability for tobacco, alcohol, marijuana, cocaine, opioid, and other substance dependence but only fair reliability estimates for stimulant and sedative dependence.

The error term from the index of dependability can be used to construct confidence intervals, reflecting the confidence that a subject truly meets the diagnosis of substance dependence. The absolute error variance (σ^2_{Δ}) ranged from 0.40 for sedatives to 2.8 for cocaine, averaging 1.2 across the eight substances. This corresponds to standard errors ranging from 0.63 to 1.67, with an average of 1.05 across substances. Thus, assuming normality for the sum scores, one can be 68% confident that the number of criteria met for a subject is the obtained score ± 1 standard error, and 95% confident using ± 1.96 standard errors. The 95% confidence intervals range from ± 1.2 for sedative dependence to ± 3.3 for cocaine dependence, with an average of approximately 2 across substances. A confidence interval of 2 indicates that one can be 95% confident that a subject meets dependence if he/she endorses five or more of the seven items or that a subject does not meet the criteria for dependence if he/she endorses fewer than two items. Naturally, the greatest risk of misclassification occurs at the cut-score demarcating two and three endorsed items.

Inter-rater and intra-rater reliabilities

Inter-rater reliability is concerned only with the consistency of measurement between interviewers, whereas intra-rater reliability is concerned with the stability of measures over occasions (Time 1 and Time 2 in this case). When the focus of reliability is on one particular factor, the other factor is fixed and the accompanying variability associated with the interaction of the latter factor with subjects becomes part of the true score variance. This leads to larger reliability estimates, but a smaller universe of generalization, than when both factors were considered error variance, as in the previously reported analyses.

Table 4 displays the inter-rater and intra-rater reliabilities for the dependence measures. For most of the dependence categories, the inter-rater reliabilities measuring consistency

TABLE 3. Component variances and percentage of variance (%) for the criterion scores

Subscale	Tobacco	Alcohol	Marijuana	Cocaine	Opioid	Stimulant	Sedative	Other
Subject	5.27 (84)	5.33 (79)	3.32 (71)	8.03 (74)	4.77 (85)	0.42 (46)	0.34 (46)	1.81 (59)
Interviewer	0.85 (14)	0.53 (8)	0.34 (7)	2.55 (24)	0.28 (5)	0.03 (3)	0.01 (1)	0.17 (6)
Occasion	0.00 (0)	0.01 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Subj. $\times$ Inter.	0.14 (2)	0.44 (7)	0.00 (0)	0.00 (0)	0.00 (0)	0.16 (17)	0.00 (0)	1.00 (33)
Subj. $\times$ Occas.	0.00 (0)	0.21 (3)	1.03 (22)	0.19 (2)	0.34 (6)	0.31 (34)	0.39 (53)	0.00 (0)
Inter. $\times$ Occas.	0.05 (1)	0.21 (3)	0.00 (0)	0.07 (1)	0.21 (4)	0.00 (0)	0.00 (0)	0.07 (2)
Ep^2	.974	.891	.763	.977	.933	.472	.466	.644
Φ	.835	.792	.708	.741	.852	.457	.459	.593

Notes: Subj. $\times$ Inter. = Subject $\times$ Interviewer; Subj. $\times$ Occas. = Subject $\times$ Occasion; Inter. $\times$ Occas. = Interviewer $\times$ Occasion; Ep^2 = generalizability coefficient; Φ = index of dependability.

TABLE 4. Estimated consistency and stability coefficients for dependence diagnoses

Subscale	Inter-rater		Intra-rater	
	Ep ²	Φ	Ep ²	Φ
Tobacco	.931	.795	.968	.968
Alcohol	.880	.830	.856	.853
Marijuana	1.00	.956	.619	.619
Cocaine	1.00	.769	.911	.911
Opioid	1.00	.945	.967	.967
Stimulant	1.00	.968	.533	.533
Sedative	1.00	1.00	.355	.355
Other	.690	.653	.887	.887

are higher than the intra-rater reliabilities measuring stability. Recall from Table 2 that the largest variance components not associated with the main effect of subjects was usually the subject by occasion interaction and that the variance attributed to the interaction between interviewers and subjects was small. This explains the high inter-rater estimates and the lower intra-rater estimates. The perfect generalizability coefficients obtained for some of the inter-rater reliabilities were derived from the absence of variability between subjects and interviewers, whereas the lower index of dependability coefficients for the same substances is from the main effect of interviewers.

The inter-rater and intra-rater reliabilities for the sum scores are shown in Table 5. Similar to the overall analysis, the reliabilities for the sum scores are higher than the dependence classifications for both the inter-rater and intra-rater measures. Again, the inter-rater reliabilities are generally higher than the intra-rater reliabilities, and a maximum generalizability coefficient is obtained for some of the substances, indicating perfect agreement in the relative ranking of subjects among the interviewers. However, for both the dependence measures and the sum scores, both tobacco and other substance dependence had higher intra-rater reliabilities than inter-rater reliabilities. From Table 2, it is evident that this finding was derived from the larger variance attributed to the subject by interviewer interaction than the subject by occasion interaction.

TABLE 5. Estimated consistency and stability coefficients for criterion scores

Subscale	Inter-rater		Intra-rater	
	Ep ²	Φ	Ep ²	Φ
Tobacco	.974	.842	1.00	1.00
Alcohol	.926	.851	.965	.963
Marijuana	1.00	.928	.763	.763
Cocaine	1.00	.763	.977	.977
Opioid	1.00	.948	.933	.933
Stimulant	.820	.793	.652	.652
Sedative	1.00	.986	.466	.466
Other	.644	.607	1.00	1.00

With respect to the consistency of the SSADDA, the inter-rater reliabilities suggest that the eight substance-dependence diagnoses are measured satisfactorily. With respect to diagnostic stability, the intra-rater reliability for the diagnosis of sedative dependence was less than satisfactory and that for stimulant dependence was only marginally satisfactory.

Discussion

This study applied generalizability theory to examine the reliability of diagnoses and the number of criteria endorsed for eight substance-dependence diagnoses, using data obtained with the SSADDA. The overall reliabilities of the diagnoses of tobacco, alcohol, cocaine, and opioid dependence were good to excellent, whereas the reliabilities were fair for diagnoses of marijuana, stimulant, and other substance dependence and poor for sedative dependence. The reliabilities for the criterion counts were nearly always better than the diagnoses themselves and ranged from very good to excellent for tobacco, alcohol, cocaine, and opioid dependence, and were good for marijuana dependence, fair for other substance dependence, and poor for stimulant and sedative dependence. These findings are similar to those reported by Pierucci-Lagha et al. (2007), who used Cohen's kappa to measure diagnostic concordance, concluding that reliability was good for nicotine, alcohol, opioid, and marijuana dependence and fair or poor for stimulant and sedative dependence.

The inter-rater reliability estimates for dependence diagnoses were excellent for all substance measures except other substance dependence, which was good. The generalizability estimates obtained were the maximum possible value (i.e., 1.00) for marijuana, cocaine, opioid, stimulant, and sedative dependence, and were approximately 0.90 for tobacco and alcohol dependence. Furthermore, the index of dependability measures was very good to excellent for most substances, varying between 0.80 and 1.00 for all substances except other substance dependence. This suggests that the interviewers typically concurred when assigning substance-dependence diagnoses to subjects. Inter-rater reliabilities for the criterion counts were mostly excellent, with only other substance dependence showing fair reliability. Again, perfect generalizability estimates were obtained for marijuana, cocaine, opioid, and sedative-dependence criterion counts, and estimates for tobacco, alcohol, and stimulant dependence criterion counts ranged from good to excellent. Similar to what was found for the dependence diagnoses, the index of dependability estimates varied from good to excellent for all substances except other substance dependence, where it was fair.

The test-retest reliability estimates, a measure of agreement between the two interview sessions, were not as high as the inter-rater reliability estimates. Even so, they were generally very good to excellent for most substance-dependence

diagnoses, with generalizability coefficients ranging from 0.86 to 0.97 for tobacco, alcohol, cocaine, opioid, and other substance dependence. The generalizability coefficient was fair for marijuana and stimulant dependence, and poor for sedative dependence. The index of dependability coefficients for the intra-rater dependence diagnoses was the same as the generalizability coefficients, except for alcohol dependence, which was slightly lower. For the inter-rater estimates, the absolute error variance, used for the index of dependability, contains the main effects for occasion, which is not part of the relative error variance, used for the generalizability coefficient. Thus, none of the variability in diagnoses was attributed to the main effect of occasion, meaning on average the mean diagnosis did not change from the first to second interview. Analogous to what was found regarding the inter-rater reliabilities, the test-retest reliabilities were higher for the criterion counts than the diagnoses. The generalizability coefficients were excellent for tobacco, alcohol, cocaine, opioid, and other substance-dependence diagnoses; fair to good for marijuana and stimulant dependence; and poor for sedative dependence. Similar to the findings for dependence, the test-retest index of dependability had the same values as the generalizability coefficients, except for alcohol dependence. Consequently, the main effect of occasion did not contribute at all to the variability in the criterion counts.

The findings from this study confirm previous studies that investigated the reliability of diagnoses obtained using the SSADDA (Pierucci-Lagha et al., 2005, 2007). The reliabilities for tobacco, alcohol, and opioid dependence were good to excellent, whereas the reliabilities for sedatives and stimulants were only fair. As an extension of the prior studies, this study focused on identifying the sources of error. The largest sources of error were attributable to the main effect of interviewers and the interaction between subjects and occasions. The variance stemming from the main effect of interviewers reveals a systematic difference in assigning scores that is consistent across the two occasions, and the variance arising from the interaction of subject with occasion indicates the subjects do not provide perfectly consistent responses between the two interview occasions. However, it must be noted that the error variance from interviewers was still small in comparison with the variability from true scores (i.e., subjects or the object of measurement). Averaged across the eight substance-dependence scales, subjects accounted for nearly two thirds of the variability in responses, whereas interviewers accounted for just 8%. The interaction between subjects and occasions was the greatest source of error variance and is likely to be the most difficult to remedy. The extensive training received by the interviewers who participated in this study is reflected in the small contribution to variability attributed to them. Additional training and practice before initiating interviews could potentially lower the variability between interviewers. The variability in subject responses from one occasion to the next may be ame-

nable to greater preparation of the subjects for the interview by emphasizing the need for careful consideration before providing responses to questions, additional probes aimed at identifying internally inconsistent responses, and greater attention to the "demand characteristics" of the interview setting. When feasible, averaging scores from two interviewers could also increase reliability, because it reduces all facets of variance connected with the interviewer. This approach, however, is more resource intensive, because it requires a second interviewer and twice the time of the interviewee. The differences in reliability estimates obtained for different substances suggest that different approaches may be required to yield highly reliable diagnoses and criterion counts for all substances.

Based on the 95% confidence intervals for the criterion counts for most subscales, we are fairly confident that participants who endorsed fewer than two items were correctly classified as not dependent and that those endorsing more than four items were correctly classified as dependent. However, the results suggest that persons who endorse two items may be misclassified as dependent and those endorsing three or four items may be misclassified as not dependent. The proportion of criterion counts that equaled two or three ranged from 1% (sedative) to 18% (alcohol). Not surprisingly, the higher dependence frequencies may be more prone to classification error. But, despite this, it is encouraging that more than 80% of the alcohol and 99% of the sedative classifications are almost certainly correct (this is most likely a lower bound because it is assuming the borderline cases are misclassified).

The method of generalizability theory has many advantages, as described in this article. However, because of the sampling variability inherent in sample data, the estimation method of MINQUE, which was used here, can sometimes give estimates for variance components that are negative, which is theoretically impossible. The likelihood of this happening increases when the design is unbalanced, and there are small frequencies in some cells, which was the case in the present study, especially for the rarer substance-dependence diagnoses. There were instances when the estimated variance component was negative. When this occurred, it was set equal to zero, which is the recommended approach (Cronbach et al., 1972). However, the few occasions of negative variances resulted in very small absolute values, which did not substantially affect the generalizations made.

In summary, the results from the generalizability analyses support the view that the SSADDA is a reliable instrument with which to diagnose dependence for most substances. The agreement among interviewers was good, and the greatest source of error was from subjects responding differently between occasions. Findings from the present study are consistent with previous evidence of the reliability of the SSADDA as an instrument for the diagnosis of substance dependence (Pierucci-Lagha et al., 2005, 2007).

Acknowledgments

The authors thank all who contributed to the development of the SSADDA and the interviewers and subjects who participated in the study of its reliability.

References

- AMERICAN PSYCHIATRIC ASSOCIATION. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV), Washington, DC, 1994.
- BELL, J.F. Generalizability theory: The software problem. *J. Educ. Stat.* **10**: 19-29, 1985.
- BERNEY, A., PREISIG, M., MATTHEY, M.-L., FERRERO, F., AND FENTON, B.T. Diagnostic Interview for Genetic Studies (DIGS): Inter-rater and test-retest reliability of alcohol and drug diagnoses. *Drug Alcohol Depend.* **65**: 149-158, 2002.
- BRENNAN, R.L. Performance assessments from the perspective of generalizability theory. *Appl. Psychol. Meas.* **24**: 339-353, 2000.
- BRENNAN, R.L. Generalizability Theory: Statistics for Social Science and Public Policy, New York: Springer-Verlag, 2001.
- BRYANT, K.J., ROUNSAVILLE, B., SPITZER, R.L., AND WILLIAMS, J.B.W. Reliability of dual diagnosis: Substance dependence and psychiatric disorders. *J. Nerv. Ment. Dis.* **180**: 251-257, 1992.
- BUCHOLZ, K.K., CADORET, R., CLONINGER, C.R., DINWIDDIE, S.H., HESSELBROCK, V.M., NURNBERGER, J.I., JR., REICH, T., SCHMIDT, I., AND SCHUCKIT, M.A. A new semi-structured psychiatric interview for use in genetic linkage studies: A report on the reliability of the SSAGA. *J. Stud. Alcohol* **55**: 149-158, 1994.
- BURKE, J.D., JR. Diagnostic categorization by the Diagnostic Interview Schedule (DIS): A comparison with other methods of assessment. In: BARRETT, J.E. AND ROSE, R.M. (Eds.) *Mental Disorders in the Community: Progress and Challenge*, New York: Guilford Press, 1986, pp. 255-279.
- CRONBACH, L.J., GLESER, G.C., NANDA, H., AND RAJARATNAM, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Hoboken, NJ: John Wiley & Sons, 1972.
- CRONBACH, L.J., RAJARATNAM, N., AND GLESER, G.C. Theory of generalizability: A liberalization of reliability theory. *Brit. J. Stat. Psychol.* **16**: 137-163, 1963.
- EASTON, C., MEZA, E., MAGER, D., ULUG, B., KILIC, C., GOGUS, A., AND BABOR, T.F. Test-retest reliability of the alcohol and drug use disorder sections of the schedules for clinical assessment in neuropsychiatry (SCAN). *Drug Alcohol Depend.* **47**: 187-194, 1997.
- GRIFFIN, M.L., WEISS, R.D., MIRIN, S.M., WILSON, H., AND BOUCHARD-VOELK, B. The use of the Diagnostic Interview Schedule in drug-dependent patients. *Amer. J. Drug Alcohol Abuse* **13**: 281-291, 1987.
- PIERUCCI-LAGHA, A., GELERNTER, J., CHAN, G., ARIAS, A., CUBELLS, J.F., FARRER, L., AND KRANZLER, H.R. Reliability of DSM-IV diagnostic criteria using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend.* **91**: 85-90, 2007.
- PIERUCCI-LAGHA, A., GELERNTER, J., FEINN, R., CUBELLS, J.F., PEARSON, D., POLLASTRI, A., FARRER, L., AND KRANZLER, H.R. Diagnostic reliability of the Semi-Structured Assessment of Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend.* **80**: 303-312, 2005.
- WILLIAMS, J.B.W., GIBBON, M., FIRST, M.B., SPITZER, R.L., DAVIES, M., BORUS, J., HOWES, M.J., KANE, J., POPE, H.G., JR., ROUNSAVILLE, B., AND WITTCHEN, H.-U. The Structured Clinical Interview for DSM-III-R (SCID): II. Multi-site test-retest reliability. *Arch. Gen. Psychiat.* **49**: 630-636, 1992.