

Predicting ASI Interviewer Severity Ratings  
for a Computer-Administered Addiction Severity Index

Stephen F. Butler, Ph.D.

Innovative Training Systems, Inc, Newton, MA

John S. Cacciola, Ph.D.

University of Pennsylvania

Simon H. Budman, Ph.D.

Innovative Training Systems, Inc., Newton, MA

Sabrina Ford, Ph.D.

Swarthmore College

David Gastfriend, MD

Massachusetts General Hospital

Frederick L. Newman, Ph.D.

Florida International University

Arlene Frank, Ph.D.

Brookside Hospital, Nashua, NH

A. Thomas McLellan, Ph.D.

University of Pennsylvania

Jack Blaine, MD

National Institute on Drug Abuse

Karla Moras, Ph.D.

University of Pennsylvania

Ihsan M. Salloum, M.D., M.P.H.

Western Psychiatric Institute and Clinic

Manuscript of published article: Butler, S.F., Newman, F.L., Cacciola, J.S., Frank, A., Budman, S.H., McLellan, A.T., Ford, S., Blaine, J., Gastfriend, D., Moras, K., & Salloum, I.M. (1998).

Predicting ASI interviewer severity ratings for a computer-administered Addiction Severity Index. *Psychological Assessment, 10*, 399-407.

RUNNING HEAD: Predicting the ASI Severity Ratings

Abstract

The Addiction Severity Index (ASI) is a reliable and valid measure of problem severity among addicted patients. Concerns have been raised about the reliability of the Interviewer Severity Rating (ISR), a summary score for each of seven domains. As part of an effort to build a computer-administered ASI, regression equations were developed to predict the ISR. Repeated re-sampling of 1124 ASIs conducted by trained interviewers, permitted derivation of stable regression equations predicting the ISR for each ASI domain from patients' answers to pre-selected interview items. The resulting seven Predicted Severity Ratings (PSRs) were tested on eight, standardized vignettes, with "gold standard," expert-generated ISRs. Reliabilities compared well to those of intensively trained interviewers. PSRs could be an alternative to interviewer ratings and may make possible a computer-administered version of the ASI.

Predicting ASI Interviewer Severity Ratings  
for a Computer-Administered Addiction Severity Index

The Addiction Severity Index (ASI; McLellan, et al. 1980, 1985, 1990) is the most widely used measure of problem severity among addicted clients entering treatment. This semi-structured interview was developed to serve as a standardized, reliable and valid instrument for evaluating substance abusing clients. The interview is used frequently in both traditional research settings and as an outcome measure in clinical settings. Its use for determining need-for-treatment is now mandated, or highly recommended, by several governmental agencies (e.g., NIDA, VA, New Jersey), including a number of states, counties and cities for programs they fund. The ASI is used worldwide and has been translated into thirteen languages. It has been expanded to specialized populations such as cocaine abusing mothers, cocaine freebase users, opiate dependent persons, federal prisoners, psychiatrically ill substance abusers, homeless, and individuals with antisocial personality disorder (e.g., Cacciola, et. al., 1995; Hodgins & El-Guebaly, 1992; Rounsaville, et al., 1991 and Smith, et al., 1990).

The ASI covers the client's Medical Status, Employment and Support Status, Drug Use, Alcohol Use, Legal Status, Family and Social Relationships, and Psychiatric Status. Clients are asked to respond to specific questions about the problems they have experienced both recently--in the last 30 days--and over their lifetimes. Thus, both urgent concerns and longstanding, chronic problems are identified by the ASI. Clients are also asked to rate the extent of their difficulties and their need for treatment in each of the seven problem areas. Clients' answers to the ASI inquiries are summarized into Composite Scores (McLellan et al., 1992) which are

considered “objective” (in that they are largely verifiable) and are used to measure change over time in response to treatment.

In addition to these client-based ratings, the interviewer makes an independent rating of the severity of each problem area, based on the interviewer’s experience with the client during the interview. This rating starts at 0, meaning no problem exists and treatment is probably not necessary, and ends at 9, meaning an extreme problem exists and treatment is absolutely necessary. Because severity ratings are relatively easy to use and straightforward to understand, the seven Interviewer Severity Ratings (ISRs) have become widely used summary scores for making clinical decisions and for treatment planning.

The ISRs are fundamentally a subjective rating. As such, the issues of interviewer training and interrater reliability are of paramount importance. A manual has been written with guidelines for making the ISRs (McLellan et al., 1990). Initial research conducted by McLellan’s group, suggested that interrater reliability for the ISRs was generally quite good (McLellan et al., 1985) ranging from .84 (Employment) to .95 (Drug Use), using Spearman-Brown coefficients. However, efforts to replicate these high reliabilities have yielded somewhat lower estimates. For instance, using IntraClass Correlations (ICCs), Hodgins and El-Guebaly (1992) found reliabilities ranging from .30 for Employment to .96 for Legal and Alcohol Use areas. There is some disagreement regarding interpretation of different magnitudes of ICCs and different measures of reliability. For instance, Fleiss (1981) has noted that for most purposes, IntraClass R “values greater than .75 or so may be taken to represent excellent agreement beyond chance” (p. 218). Values between .40 and .75 represent fair to good agreement beyond chance. Landis and Koch

(1977), on the other hand describe ICCs greater than .80 as perfect reliability while ICCs as low as .21 reflect fair reliability. Nevertheless, Hodgins and El-Guebaly's results call into question the generalizability of the initial reliability coefficients of the ISRs.

Another examination of the reliability of the ISRs (Alterman et al., 1994) found similar reliability problems. These authors found ICCs between a low of .31 (Drug Use) to a high of .78 (Alcohol Use). In addition to Drug Use, two other areas, Employment (.53) and Legal (.48), achieved ICCs less than .60. While admitting that their training procedures were not as rigorous as those used in McLellan's original studies, the authors note that the training their raters received may reflect more accurately the level of ASI training in the field. Indeed, the intensity of the training and commitment by the raters in research settings may substantially exceed what occurs in clinical settings. Under certain clinical circumstances, for instance, the interviewer might unconsciously (or consciously!) experience pressure to ensure that a given client's score is severe enough to justify an admission. Rater "drift" can also account for unreliability of the ratings over time, requiring regular calibration and reliability checks that may not remain a priority in clinical settings.

Such considerations have resulted in general concern about widespread use of the ISR. Indeed, McLellan and his colleagues (1992), have expressed strong caution about the ISR because of concerns about the subjectivity of the rating and have suggested a focus on the Composite Scores. Despite this caution, use of ISRs as a summary score remains popular in research to provide patient profiles for treatment matching and in clinical settings to inform treatment planning (Cacciola et al., 1997). In this regard, the Composite Scores, though more

reliable, reflect problem status only for the prior 30 days; and there remains interest in more general measures of “lifetime problem severity.” The Composite Scores are arithmetically weighted summary scores derived from a set of items in each problem area. Composite Scores, usually used to measure change, are based only on items subject to change and therefore may not reflect “need for treatment” as well as the ISR. In addition, examinations of the relationship between Composite Scores and ISRs have yielded variable results (e.g., Alterman et al., 1994), suggesting that for at least some problem areas, ISRs and Composite Scores measure different things. Some authors have suggested that ISR reliability could be improved by providing guidance for anchoring points along the severity scale (Alterman et al., 1994).

In an attempt to address this concern, Cacciola and colleagues (1997) recently developed a set of eight standard ASI vignettes. Each vignette has ISRs for each problem area developed as a consensus of two expert ASI trainers. These authors compared three conditions: (1) ASI training alone; (2) training plus vignettes; and (3) training, plus vignettes, plus “annotated answers”—which operationalized the rationale for the expert consensus ISR within each problem area. Results revealed limited evidence that the ASI vignettes and feedback increased reliability of trainees’ ISRs with expert consensus ISRs. Despite the enhanced training with standardized vignettes, reliability levels were not uniformly high. ICCs were less than .60 for Alcohol and Family/Social sections of the ASI. These results led the authors to emphasize the need to improve ISR reliability or “to develop alternative methods for summarizing problem severity” (p. 9).

One such alternative is to derive mathematical equations which predict the seven ISRs. The impetus for the present effort to develop such a mathematical prediction of the ISRs arose as

part of a project to develop a completely computer-administered ASI. A computer-administered ASI might have a number of advantages over an interviewer-administered ASI. Standardization of the presentation of the questions is one potential advantage. In addition, there is some evidence that respondents to computer-presented questions are more honest than when talking to a person-interviewer. This may be especially true when the answers to questions might be embarrassing or involve actions that society might judge harshly. For example, Navaline, Snider, Petro, Tobin, Metzger, Alterman, and Woody (1994) found that Ss responded more honestly to questions about high-risk HIV/AIDS behaviors when the questions were presented by computer rather than a live interviewer. More recently, Turner, Ku, Rogers, Lindberg, Pleck, and Sonenstein (1998) found that adolescent males were more likely to admit acts of male-male sex, injection drug use, and sexual contact with intravenous drug users on a computer-administered “self-interview” than those using a more traditional written questionnaire.

A central problem for a computer-administered ASI is the absence of an interviewer to provide the Interviewer Severity Rating. Guidelines in the ASI manual for making the ISRs (McLellan et al., 1990) emphasize use of the respondent’s answers to the ASI questions. Consequently, it was logical to attempt to replicate this decision-making process using regression procedures to model the interviewer’s severity rating for each ASI domain based on respondents’ answers to the interview. This very practical need requires the computer to utilize responses from the interviewee to generate a “predicted severity rating” (PSR). Such a PSR must result in a rating that is reasonably comparable to a well-trained, interviewer’s rating.

It was therefore hypothesized that the equations predicting the severity rating (or the PSRs) would show acceptable “interrater” reliability with ISRs generated by expert ASI trainers. This procedure replicates the current training requirements for ASI interviewers. Another hypothesis addresses the differential utility of the PSR over the existing ASI Composite Score as a substitute for ISRs. Specifically, it was expected that the equation-generated PSRs would prove to be better predictors of interviewer-generated ISRs than would the corresponding ASI Composite Scores.

## Method

### Overview of Procedures

Two studies are described below. In the first study, a large dataset of ASIs administered by trained, reliable interviewers was assembled. Using this large dataset, a stable regression equation was derived for each ASI problem area. The second study involved deriving Predicted Severity Ratings (PSRs) for each ASI domain and calculating their “interrater” reliability with corresponding ISR ratings generated by expert ASI raters on the standardized vignettes created for ASI training (Cacciola, et al., 1997). This replicates the current training procedures for ASI interviewers and therefore, represents the state-of-the-art procedures for determining whether a new “rater” (in this case the equation-derived PSRs) demonstrates inter-rater reliability. Finally, we will examine whether derived PSRs predict the ISRs better than the existing mathematical summary or Composite Score.

## Study 1: Generation of the Severity Rating Equations

### Dataset and Subjects

Four independent datasets were merged to create one large dataset for generating the regression models. Three of the datasets were obtained from the PENN/VA archive. These datasets were provided by Dr. McLellan as ones in which he had confidence in the training of the raters. Each dataset was composed of a different client population: a cocaine dependent population, an alcohol dependent population and an opiate-addicted population. The samples were drawn from PENN/VA treatment centers and were, therefore, exclusively male. To increase the overall N and representation of women, we added a sample of initial ASIs conducted as part of the NIDA Collaborative Cocaine Treatment Study (Crits-Christoph et al., 1997).

The merged datasets included 1124 intake ASIs of substance abusers who applied for treatment. Only ASIs from treatment applicants were used since this is the most appropriate population for the ASI (McLellan et al., 1992). Characteristics of the subjects in each of these datasets and the combined sample are presented in Table 1. As can be seen, the majority of patients (81%) are in studies of cocaine abuse, with only 13% from an alcohol study and only 6% from an opiate study. Thus, the sample is heavily weighted in terms of clients in treatment for cocaine abuse. It is important to note, that abuse of multiple substances was the rule, rather than the exception for all the samples. For example, only 28% of the entire sample (303 Ss) reported no alcohol abuse. Another feature of the sample is the under-representation of women. Due largely to reliance on VA samples (three out of the four samples) that were exclusively male, the entire sample contained only 19% women, and all these were in the Collaborative Cocaine

Treatment Study. Furthermore, the combined sample consists of 59% African-American clients, with only 39% white clients and very few Hispanic clients (.5%). For the entire dataset, the mean age of respondents was 35 years, (SD = 7.38). Ages ranged from 19 to 61. Approximately 60% of the sample were currently employed and 20% were on welfare, while the average education level achieved was high school (although these data were not available for two of the samples). A substantial proportion of the samples had been arrested (51%) and 11% were on parole or probation. Only about 25% of the entire sample were married, while about a third (33%) were divorced or separated and 39% were single.

-----  
 Insert Table 1 about here  
 -----

Procedures

ASI Training Procedures. The ASI interviews in each dataset were conducted by highly trained interviewers under research conditions. The ASIs in the three PENN/VA datasets were conducted by research staff at the Center for Studies on Addiction at the University of Pennsylvania. Formal training of these ASI interviewers involved review of the ASI instructional manual and observation of an ASI interview. This was followed by a two-day ASI workshop conducted by an expert ASI trainer which included didactic presentation and role playing. Over several weeks, trainees conducted ASI interviews and were observed by senior ASI interviewers/trainers. During this time, the trainees received feedback on coding and interviewing. Although no formal reliability standards were employed, only after acceptable

coding and interviewing skills were demonstrated was the trainee allowed to conduct independent interviews. Interviewers were observed periodically to maintain quality and diminish rater drift. Also, interviews were regularly checked by a senior interviewer for errors and consistency, and any problems were discussed with the interviewer.

Similar training procedures were followed in the NIDA Collaborative Cocaine Treatment Study. The ASI interviewers were trained by experts at the Center for Studies on Addiction at the University of Pennsylvania. They participated in two reliability tests of administration and scoring of the ASI. The first test was of percent agreement among the interviewers within each of five research sites, resulting in percent agreement in excess of 95% at each site. The second test involved comparing site ratings with those of expert interviewers at the Center for Studies on Addiction at the University of Pennsylvania. This comparison yielded percent agreements of 93% or higher on most individual items; formal tests of reliability for ISR ratings were not conducted.

Plan for the analyses. A stepwise, linear regression with a planned  $R^2$  goal of .50 was performed for each of the seven ISRs. The plan for generating the regression models for estimating ISRs was inspired by the cross validation, jackknife and bootstrap methods for estimating prediction error (Efron, 1982). The basic idea is to repeatedly re-sample a dataset to arrive at a regression equation that is minimally dependent on particular features of any given array of data. To this end, we employed the following steps: (1) The entire dataset was randomly divided in half. (2) An equation for the first half was derived, followed by another equation, from the second half. Steps 1 and 2 were repeated until an arbitrary total of nine separate equations were derived for each of the ASI domains (medical, employment, alcohol use, drug

use, psychiatric, legal and family). Each equation was, therefore, derived from one half of the entire dataset. (3) The  $R^2$  values were arranged in ascending order for each of the nine repetitions. An odd number (nine) of equations were derived so that one equation would fall in the middle. Finally, (4) the equation with the central value (fifth largest)  $R^2$  was selected. It was reasoned that the equation resulting in this central  $R^2$  value represented the regression solution most likely to be replicated. Equations with the largest  $R^2$ s are likely to reflect overly optimistic predictions, and are therefore, unlikely to be replicable. One equation was selected for each ASI domain to generate a Predicted Severity Rating (PSR) for that ASI domain.

Variables to be entered into the regression analyses. The variables to be entered into the regression were determined a priori based on the following considerations. The first step was to conceptualize components of the severity ratings. The ASI Manual (McLellan et al., 1990) defines “severity” as the “need for treatment” and mandates that these ratings “should be based upon report of amount, duration, and intensity of symptoms within the problem area” (emphasis added, McLellan et al., 1990, p.9). Such ratings are to be made regardless of the patient’s potential for benefit or availability of treatment. The ISR is intended to capture the extent to which treatment is needed based exclusively on the client’s responses to the items in each problem area.

It has long been known in the Human Information and Judgment literature (see review by Newman, 1983) that the human judge will select three to five “chunks” of information that appear to be most highly correlated with a global subjective judgment as the basis for that judgment. This finding suggests that three to five client-generated item responses (or sets of

responses) within each of the seven domains of the ASI will predict the ISR. The logical basis for selecting items for the Predicted Severity Rating was simply those client-answered items that appeared to directly reflect problem severity consistent with the intentions of the ASI's creators. We arrived at four essential categories of predictors. These categories of predictors were: (1) client's expressed need for treatment, (2) the degree to which the client was troubled by the problems in that area, (3) current severity and (4) duration of the problems. No other items were found to cut across all seven domains and to have this logical link with the global rating of problem severity. ASI items reflecting these four categories of predictors were selected for each domain. Specifically, the predictors were characterized as follows:

1. ASI global ratings by the patient of his or her need for treatment/counseling.
2. ASI global ratings by the patient of how troubled or bothered he or she was by the problems in the past 30 days.
3. ASI items reflecting how severe the problem was currently.
4. ASI items reflecting the duration of the problem.

In most domains, the first two categories ("need" and "troubled") were drawn from the corresponding single ASI item as rated by the client. Only the Family and Social domain required summing variables for these two categories, since the patient rates family and social "need" and "degree troubled" separately. Thus, these ratings were added to form a "Need" variable and a "Troubled" variable. For current severity and duration, most domains had several items whose content was judged as relevant to these concepts. A priori clinical judgments were made by the first author as to which items should make up these summed variables.

Results for Study 1

Table 2 presents the multiple Rs and R<sup>2</sup>s for all 7 domains for each of the nine repetitions of the regression. As can be seen, the R<sup>2</sup>s for all the regression solutions exceed the a priori goal of .50 and were generally high. The R<sup>2</sup>s ranged from a low of .55 for Family/Social to a high of .80 for the Psychiatric problem area. For each ASI problem domain, the R<sup>2</sup>s were arrayed in order of magnitude from lowest to highest. The fifth or middle R<sup>2</sup> was used to select the equation for the Predicted Severity Rating (PSR). In Table 2 the equation used for the Medical problem area was equation B2 (R<sup>2</sup> = .78), Employment was equation E1 (R<sup>2</sup> = .61), equation C1 met the criteria for Alcohol Use (R<sup>2</sup> = .65) and Legal (R<sup>2</sup> = .64) problem areas, Drug Use was equation D1 (R<sup>2</sup> = .70), the equation for the Psychiatric problem area was A2 (R<sup>2</sup> = .77), and equation C2 was selected for Family/Social problems (R<sup>2</sup> = .59). In the case of the Medical, Employment and Alcohol problem areas, the R<sup>2</sup> values had to be carried out to four decimal places to determine the middle value. Alcohol Use, Drug Use and Family/Social problem areas did not enter all four variables in the stepwise regression models. Thus, the “troubled by problem” variable was not entered for the Alcohol Use and Family/Social problem areas, while the “troubled by problem” and “current problem” variables were not entered for the Drug Use problem area.

-----  
 Insert Table 2 about here  
 -----

## Study 2: Confirmation of the PSRs

Materials

Standard ISRs were obtained from the eight ASI vignettes developed by Cacciola et al. (1997). As described above, these vignettes are fictionalized narrative case summaries that reproduced the quantitative ASI information in case report format. Characteristics of the clients in these eight vignettes are presented in Table 3. As can be seen, a variety of client characteristics are represented. A range of ages (from 22 to 45) are included, along with men and women of African-American and white racial backgrounds. Major substance of abuse includes alcohol only, alcohol and drugs, and polydrug abuse.

-----  
Insert Table 3 about here  
-----

For every vignette, an ISR answer key was derived from a consensus of three expert ASI trainers. Each expert rater had at least five years of experience conducting ASIs. One was a Ph.D. psychologist and two were masters-level raters. All had extensive experience conducting training on use of the ASI, including making severity ratings. Each expert rated all vignettes individually, and the three raters met to discuss the ratings. On those items where disagreement existed, the item was discussed, and a consensus developed for the final rating. Current ASI training procedures use these consensus ISRs as the “gold standard” against which interrater reliability is judged. For this reason, we used this same criteria to test the interrater reliability of the equation-generated Predicted Severity Ratings (PSRs) as though the equations were new

trainees. Thus, the purpose of this test was to determine the extent to which the equations could serve, in a practical way, as a substitute interviewer who was highly trained and met or exceeded the interrater reliability of human interviewers.

The range of possible severity ratings (from 0 to 9) for each domain were generally represented by the eight vignettes. Specifically, the range of scores for the Medical Domain was 0 to 7, for Employment the scores ranged from 2 to 7, for the Alcohol Domain the range was 1 to 8, for the Drug Domain the range was 2 to 8, for Legal the range was 1 to 5, for Social/Family the range was 2 to 7, and for the Psychiatric Domain, the range was 3 to 8.

### Procedure

This confirmation was conducted by comparing the equation-generated Predicted Severity Ratings (PSRs) with ISRs made by expert ASI raters on the set of eight standardized ASI vignettes. Comparisons were made using IntraClass correlations (ICCs). In these analyses, the PSRs derived from the equations were rounded to the nearest whole number before the ICCs were calculated to reflect the fact that ISRs are always whole number ratings from 0 to 9. It was expected that the PSRs would meet acceptable standards of interrater reliability with expert-generated ISRs and would be better at predicting ISRs than the mathematically derived Composite Scores.

### Results of Study 2

Comparison of the PSRs and expert interviewers' ISRs. Using the "client's" responses from the eight standard vignettes, the equations were used to generate a PSR for each of the seven ASI domains for all eight vignettes. The PSRs were then tested for their reliability with

the ratings of expert trainers. Table 4 presents the ICCs calculated for the PSRs and the expert-generated ISRs. As can be seen, all ICCs for the PSRs were above .60, ranging from a high of .96 for Drug Use to .64 for Family/Social.

As a reference point for interpreting these ICC scores, Table 4 also presents mean ICCs for ISRs obtained under the most rigorous training condition studied by Cacciola and colleagues (1997). These data were obtained on ratings of five videotapes, so they should not be viewed as directly comparable to the PSR data. Nevertheless, they do illustrate the magnitude of ICCs obtained by human interviewers who have undergone ASI standard training plus reviewing videotapes of the standard case narratives (role played by Center staff), plus ISR feedback. In each case, the reliability figures for the PSRs compared favorably with the mean ISRs achieved by this extensive training program (Cacciola et al., 1997). To examine how well the PSR did

-----  
 Insert Table 4 about here  
 -----

compared to the best possible interviewer ratings (though these were not necessarily generated by a single rater), we compared the ICCs obtained by the PSR equations to the highest ICCs achieved in each domain by the interviewers. The PSR was superior than the best rating in three of the seven domains: Medical (.92 compared to .75), Employment (.72 to .69), and Alcohol Use (.74 to .71). It was virtually equal for the Drug Use Domain (.96 to .99), and lower, though above .60, in Psychiatric (.73 to .89), Legal (.65 to .81), and Family/Social (.64 to .68). Thus, the

PSR equations had consistently good ICCs, but not better than the best rating taken from the pool of interviewer trainees.

PSR versus ASI Composite Scores. Table 5 presents Pearson correlations between Composite Scores, PSRs and ISRs for the eight standard vignettes. The Pearson correlations for the PSR with the expert-generated ISR range from a high of .98 for Drug Use to a low of .66 for Legal problems. The Composite Scores correlated highly with ISRs on some scales. However, for two scales, the correlations were substantially lower for the Composite Scores than the PSRs: Employment,  $r = .31$  for Composite Score,  $r = .84$  for PSR and Family/social  $r = .43$  for Composite Score,  $r = .75$  for PSR. It is interesting at this point to consider the possibility that rather than using regression equations, unit or equal weighting of the variables entered into the regression equation could produce equally good predictions of the interviewer rating. Indeed, the literature suggests that simple sums of variables can produce superior predictions in new samples than using least squares regression coefficients (e.g., Wainer, 1978). This phenomenon results from sensitivity of the least squares regression method to the idiosyncrasies of the original sample. In a new sample, when these idiosyncrasies are not present, regression models tend to decrease in their predictive effectiveness, especially when the initial sample is small (Schmidt, 1971). It was to minimize this effect that we sought to put together a large dataset, to resample repeatedly from that dataset, and to choose those equations whose  $R^2$  was in the middle of the range of  $R^2$ s obtained for each domain. To further address this issue, Table 5 includes correlations with unit weighted sums of the ratings from each variable entered into the regression equation. Clearly, these data suggest that in some cases the unit weighted sums are as good as or

better than the prediction provided by the regression model. The notable exception is the Employment Domain, where the unit weight produces a markedly poor prediction (-.05) of the interviewer severity rating. Such variability in predictive accuracy is problematic for scores that are to be used for clinical decision making.

-----  
Insert Table 5 about here  
-----

### **Discussion**

Despite widespread use of ISRs for clinical decision making, unreliable ISRs remain an impediment to the appropriate use of the ASI in the clinical setting. Thus, the main goal of this investigation was to generate equations that could provide calculated estimates (called Predicted Severity Ratings or PSRs) of interviewer severity ratings (ISRs) for the seven domains of the ASI. Such PSRs, if found to produce ISRs with consistent interrater reliability, could constitute the alternative method for summarizing problem severity on the ASI called for by Cacciola and colleagues (1997).

Our approach has assumed that the PSR must meet the same criteria as a trainee-interviewer, namely that it achieve acceptable interrater reliability with expert ASI interviewers. Employing a large database of ASI interviews, we generated regression models of the ISR for each of the seven ASI domains. The models appear to be stable and replicable and show potential for generating individual severity ratings that are as reliable as ISRs made by ASI interviewers who have undergone extensive training. Importantly, these PSRs are easily derived

directly from data obtained during the ASI interview with a relatively simple calculation.

Compared to the expense of training and maintaining reliable raters, use of the PSRs can be a highly cost-effective way to obtain severity ratings with known and consistent interrater reliability.

There are some limitations to the present study. This study subjected the PSR equations to a formal confirmation, using the standard vignettes rated by expert ASI interviewers. While the results were promising, the low N remains a concern. Nevertheless, use of these eight cases was an important first step, since these cases were the only standardized vignettes whose ratings are, in practice, considered to be the “gold standard” ISRs. A more extensive examination of interrater reliability with a larger N would be necessary to ensure that the ICCs reported here are stable. Finally, the present study does not address the validity of the PSR. It was not the purpose of this study to investigate the validity of the ISRs themselves. Considerable data exist which address the validity of the ISRs. For example, McLellan et al. (1985) found good evidence supporting the concurrent and discriminant validity of the severity ratings, suggesting that the ISRs do not yield simply a generalized “cry for help.” Presumably, in the present study, high correspondence with expert-generated ISRs suggests that the PSR would be valid, or at least as valid as any well-trained rater. Nevertheless, it will be important in future work to empirically validate the PSRs as well as their utility as outcome measures.

The sample used to generate and confirm the PSRs appears to be generally representative of the larger population seeking substance abuse treatment. The principal substance abuse diagnoses were all represented, including primary alcohol, cocaine and opiate addiction.

However, it should be highlighted that the sample over-represents subjects in studies of cocaine treatment and the generated equations should be confirmed on samples containing more subjects in alcohol or opiate treatment. The distribution of gender under represents women substance abusers somewhat with about 20% women in the present sample as compared to a national representation of about 26.5% female substance abusers in treatment (SAMHSA, 1993). In the sample used by McLellan et al. (1985) for validation of the ASI, 64% were men and 36% were women. On the other hand, African-Americans were over represented in the present sample, comprising about 60% of the dataset. Surveys (SAMHSA, 1993) suggest that about 60% of substance abusers nationally are white, not Hispanic with only 21% African-American and 14% Hispanic. Hispanic substance abusers are not represented in the present sample. Validation with a Hispanic population of the PSRs generated in this study is necessary.

A final consideration involves the decision to rely on the regression equations to predict the severity rating scores. As discussed above, the other options include the existing the Composite Scores and the unit weighted sums of the variables entered into the equation. There are empirical and conceptual reasons for relying on the regression equations. Conceptually, as discussed above, the Composite Scores include only information that reflects the prior 30 days. Decisions about problem severity might involve longer-term, historical information, such as duration of the problem, which is included in the PSR equations. Also, the PSR calculations result in a number between “0” and “9” which people are used to and comfortable with. Although unit-weighted scores would require some mathematical transformations to place them into the 0 to 9 ISR framework, the calculations for these scores would be relatively simple and

straightforward. Consistent with the literature (e.g., Dawes & Corrigan, 1974; Wainer, 1976), the unit-weighted sums provided, in most cases, excellent correspondence with the ISRs, so that the decision of which strategy to select for generating substitute ISRs is ultimately up to the user. The argument for using unit-weighted sums is their simplicity. Probably the strongest argument for using the PSR is that, based on the limited confirmatory analysis presented here, it appears the regression generated PSRs provided a consistently better prediction of expert-generated ISRs across all domains. Further testing of this correspondence along with examination of the PSRs with other substance abusing populations is warranted.

Overall, this study has demonstrated the feasibility of creating PSRs for estimating ISRs based solely on client responses for all domains of the ASI. If these PSRs prove stable over subsequent trials and withstand further tests of interrater reliability, they may go some way toward reducing the susceptibility to rater drift and rater bias that plagues ASI usage in clinical settings and is costly to researchers.

As use of multimedia computer technology continues to proliferate, and hardware costs continue to decrease, development of a computer-administered version of the ASI will become increasingly feasible. If such feasibility can be demonstrated, a mathematical estimate of the ISR would be a requisite feature of any such computer program. Naturally, with no interviewer to supply an ISR, a computerized version of the ASI would be unable to provide the Interviewer Severity Rating. Computer administration of the ASI could have several advantages over interviewer administration, including greater interview uniformity, significantly less expense, and possibly more honest answers than those provided to a person-interviewer (see for example,

Navaline et al., 1994). Development and testing of a computer-administered ASI is currently underway. Generation of the PSR represents an initial step toward development of a fully automated addiction severity assessment.

### References

- Alterman, A.I., Brown, L.S., Zaballero, A., & McKay, J.R. (1994). Interviewer severity ratings and composite scores of the ASI: a further look. Drug and Alcohol Dependence, 34, 201-209.
- Cacciola, J., Alterman, A.I. Fureman, M.A., Parikh, G.A., Rutherford, M. J. (1997). The use of vignettes for the addiction severity index training. Journal of Substance Abuse Treatment, 14, 439-443.
- Cacciola, J., Alterman, A.I. and Rutherford, M. (1995). Treatment response of antisocial substance abusers. Journal of Nervous and Mental Disease, 183, 166-171.
- Crits-Christoph, P., Siqueland, L., Blaine, J., Frank, A., Luborsky, L., Onken, L.S., Muenz, L., Thase, M.E., Weiss, R.D., Gastfriend, D.R., Woody, G., Barber, J.P., Butler, S.F., Daley, D., Bishop, S., Najavits, L.M., Lis, J., Mercer, D., Griffin, M.L., Moras, K. & Beck, A. (1997). The National Institute on Drug Abuse Collaborative Cocaine Collaborative Treatment Study: Rationale and methods, Archives General Psychiatry, 54, 721-726.
- Dawes, R.M. & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.
- Efron, B. (1982). The Jackknifing, The Bootstrap, and Other Resampling Plans. The Society for Industrial and Applied Mathematics, Bristol, England. Cited in Selvin, S. (1991). Statistical Analysis of Epidemiological Data. New York: Oxford University Press.
- Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed). Wiley & Sons, N.Y.

- Hendricks, V.M., Kaplan, C.D., VanLimbeek, J., & Geerlings, P. (1989). The Addiction Severity Index: Reliability and validity in a Dutch addict population. Journal of Substance Abuse Treatment, 6, 133-141.
- Hodgins, D.C. and El-Guebaly, N. (1992). More data on the Addiction Severity Index: Reliability and validity with the mentally ill substance abuser. Journal of Nervous and Mental Disease, 180, 197-201.
- Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.
- McGahan, P.L., Griffith, J.A., Parente, R. McLellan, A.T. (undated manuscript). Composite Scores from the Addiction Severity Index. Available from the senior author.
- McLellan, A.T., Luborsky, L., Cacciola, J., Griffith, J.E., Evans, F., Barr, H.L., & O'Brien, C.P. (1985). New data from the Addiction Severity Index: Reliability and validity in three centers. Journal of Nervous and Mental Disease, 173, 412-423.
- McLellan, A.T., Luborsky, L., Woody, G. E., & O'Brien, C.P. (1980). An improved evaluation instrument for substance abuse patients: The Addiction Severity Index. Journal of Nervous and Mental Disease, 168, 26-33.
- McLellan, A.T., Kushner, H., Metzger, D., Peters, R., Smith, I., Grissom, G., Pettinati, H., and Argeriou, M. (1992). The fifth edition of the Addiction Severity Index. Journal of Substance Abuse Treatment, 9, 199-213.
- McLellan, A.T., Parikh, G., Bragg, A. Cacciola, J., Fureman, B., Incmikoski, R. (1990). Addiction Severity Index, Fifth Edition with preface: A guide to training and supervising

- the ASI Interviews based on the past ten years. Unpublished manuscript, University of Pennsylvania, Center for Studies of Addiction.
- Navaline, H.A., Snider, E.C., Petro, C.J., Tobin, D., Metzger, D. Alterman, A.I., & Woody, G.E. (1994). An automated version of the risk assessment battery (RAB): Enhancing the assessment of risk behaviors. AIDS Research and Human Retroviruses, *10*, S281-S283.
- Newman, F. L. (1983). Therapists' evaluations of psychotherapy. In M. Lambert, E. Christensen, & R. DeJulio (Eds.), The assessment of psychotherapy outcome. New York: Wiley. pp: 497-534.
- Rounsaville, B.J., Anton, S.F., Carroll, K.M., Budde, D., Prusoff, B.A., & Gavin, F.I. (1991). Psychiatric diagnosis of treatment seeking cocaine abusers. Archives of General Psychiatry, *45*, 43-51.
- SAMHSA (1993). National drug and alcoholism treatment unit survey-1991: Main findings report. DHHS Pub. No. 93-2007, US Dept of Health & Human Services.
- Schmidt, F.L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, *31*, 699-714.
- Smith, I.E., Moss-Wells, S., Moeti, R., & Coles, C. (1990). Characteristics of non-referred cocaine abusing mothers. In L.S. Harris (Ed.), Problems of drug dependence 1989 (NIDA Research Monograph #95, p. 330). Rockville, MD: National Institute on Drug Abuse.

Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., & Sonenstein, F.L. (1998).

Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. Science, 280, 867-873.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind.

Psychological Bulletin, 83, 213-217.

Wainer, H. (1978). On the sensitivity of regression and regressors. Psychological Bulletin, 85,

267-273.

Zanis, D.A., Petro, C., Fureman, E., & McLellan, A.T. (1996). An examination of the ASI

Interviewer severity rating system. Unpublished manuscript, University of Pennsylvania.

Predicting the ASI Severity Ratings

Table 1  
Characteristics of Patients by Database and Total

	Alcohol VA Sample N = 144	Cocaine VA Sample N = 230	Opiate VA Sample N = 72	Multisite Cocaine Sample N = 678	Total N=1124
Age	41.6 (8.0)	34.9 (5.7)	42.5 (4.9)	32.7 (6.5)	35.0 (7.4)
% Male	100%	100%	100%	67.9%	80.8%
% Female				32.1%	19.2%
% White	31.3%	2.6%	21.1%	54.8%	39.0%
% African-American	66.7%	96.5%	76.1%	42.8%	59.1%
% Hispanic	2.1%	.4%	2.8%		.5%
% Other				2.3%	1.5%
Years Education	Data Missing	12.34 (1.5)	Data Missing	12.5 (2.5)	12.4 (2.2)
% Working	51.4%	60.0%	58.3%	62.1%	60.1%
% on Welfare	18.8%	10.0%	23.6%	23.5%	20.1%
Years Alcohol Abuse	19.1(8.8)	8.4 (7.9)	12.0 (10.0)	9.9 (8.1)	11.0 (9.0)
Years Opiate Abuse		.6 (2.3)	17.4 (7.7)	.5 (2.2)	1.7 (5.1)
Years Cocaine Abuse	.3 (.7)	3.6 (3.9)	4.7 (5.7)	6.2 (4.5)	4.8 (4.6)
Years Other Drug Abuse	.7 (2.2)	9.8 (9.4)	13.6 (15.7)	9.2 (9.9)	8.8 (10.1)
% Arrested	41.4%	41.7%	73.6%	53.1%	50.6%
% Parole/probation	11.8%	9.2%	13.9%	10.7%	10.8%
% Married/remarried	20.1%	21.7%	72.2%	22.9%	25.5%
% Divorce	27.8%	16.1%		16.2%	16.6%
% Separated	23.6%	29.1%		12.4%	16.5%
% Widowed	2.1%	.9%	27.8%	.4%	2.5%
% Single/never married	26.4%	32.2%		48.1%	38.9%
% Treated for Psychiatric Problems	30.6%	23.9%	29.2%	28.2%	27.7%
% Hospitalized for Psychiatric Problems	14.6%	12.6%	18.1%	12.8%	13.3%

Table 2

Predicting the ASI Severity Ratings

Multiple Rs and R<sup>2</sup>s for Each of 9 Split-half Repetitions of Stepwise Multiple Regression

	Random Sample Split Halves A				Random Sample Split Halves B				Random Sample Split Halves C				Random Sample Split Halves D				Random Sample Split Half E	
ASI Problem Domain	A1		A2		B1		B2		C1		C2		D1		D2		E1	
	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ	MUL R	R SQ
Medical	0.870	0.749	0.884	0.781	0.873	0.762	0.881	0.775	0.888	0.789	0.864	0.747	0.890	0.792	0.890	0.792	0.880	0.775
Employment	0.791	0.625	0.769	0.591	0.773	0.597	0.806	0.649	0.780	0.608	0.784	0.614	0.781	0.609	0.783	0.614	0.780	0.608
Alcohol Use	0.809	0.654	0.780	0.639	0.810	0.660	0.812	0.660	0.807	0.652	0.800	0.640	0.802	0.643	0.807	0.652	0.796	0.634
Drug Use	0.857	0.734	0.814	0.662	0.775	0.600	0.842	0.709	0.829	0.687	0.842	0.709	0.837	0.700	0.836	0.698	0.850	0.723
Psychiatric	0.878	0.771	0.879	0.773	0.865	0.748	0.893	0.798	0.872	0.760	0.884	0.781	0.869	0.756	0.887	0.787	0.883	0.779
Legal	0.781	0.610	0.810	0.655	0.833	0.694	0.803	0.645	0.797	0.635	0.790	0.624	0.798	0.637	0.787	0.619	0.793	0.629
Family	0.757	0.574	0.778	0.605	0.768	0.590	0.791	0.626	0.767	0.589	0.766	0.586	0.763	0.583	0.765	0.585	0.739	0.547

Table 3

Characteristics of the Patients Represented in the Eight Vignettes (Cacciola et al., 1997)

<b>Vignette No.</b>	<b>Age</b>	<b>Race</b>	<b>Marital Status</b>	<b>Education</b>	<b>Work Status</b>	<b>Legal Involvement History</b>	<b>Psychiatric Treatment History</b>	<b>Primary Substance of Abuse</b>
<b>1</b>	36	African-American	Single	16 Years	Employed	Yes	No	Alcohol and Drug
<b>2</b>	26	African-American	Single	9 Years	Unemployed on Welfare	Yes	Yes	Polydrug
<b>3</b>	40	African-American	Separated	13 Years	Unemployed on Welfare	Yes	No	Polydrug
<b>4</b>	22	White	Single	12 Years	Employed	Yes	No	Alcohol and Drug
<b>5</b>	23	White	Single	9 Years	Unemployed on Welfare	No	No	Alcohol
<b>6</b>	27	White	Married	12 Years	Employed	Yes	Yes	Alcohol
<b>7</b>	24	White	Divorced	12 Years	Unemployed on Welfare	No	Yes	Polydrug
<b>8</b>	45	White	Single	12 Years	Unemployed on Pension	No	Yes	Alcohol

Table 4

IntraClass Rs (ICCs) between Expert-generated ISRs

and PSRs along with Comparison Training Data (Cacciola et al., 1996)

<b>ASI Problem Areas</b>	<b>Predicted Severity Rating IntraClass Correlations for 8 Vignettes</b>	<b>Training + STD. Video and Interviewer Severity Rating Feedback 5 Raters 5 Vignettes N = 25</b>
Medical	.92	.71 ± 0.04
Employment	.72	.50 ± 0.19
Alcohol Use	.74	.51 ± 0.19
Drug Use	.96	.96 ± 0.03
Psychiatric	.73	.77 ± 0.12
Legal	.65	.53 ± 0.28
Family/social	.64	.47 ± 0.21

Table 5

Relationship of the PSR and Expert-Generated ISRs versus Composite

Scores and Expert-Generated ISRs and Unit Weighted Sum of ASI Variables

<b>ASI Problem Areas</b>	<b>PSR Pearson r with ISR</b>	<b>Composite Score Pearson r with ISR</b>	<b>Unit Weighted Sum of Variables Pearson r with ISR</b>
Medical	.95	.96	.91
Employment	.84	.31	-.05
Alcohol Use	.89	.86	.82
Drug Use	.98	.96	.94
Psychiatric	.87	.96	.74
Legal	.66	.63	.70
Family/social	.75	.43	.82

Author Notes

This project was supported, in part, by an SBIR grant to the first author #R43-DA09938-

01. Some of the data analyzed was made available by the NIDA Collaborative Cocaine Treatment Study, NIDA grant # U01 DA-07663.

Correspondence should be directed to Stephen F. Butler, Ph.D., Innovative Training Systems, Inc. 199 Wells Ave., Suite 108, Newton, MA 02159