

Combining psychometric and biometric measures of substance use

Richard Lennox^{a,*}, Michael L. Dennis^{b,1}, Christy K Scott^c, Rod Funk^b

^a *Psychometric Technologies, 2404 Western Park Lane, Hillsborough, NC 27278, USA*

^b *Chestnut Health Systems, 720 W. Chestnut, Bloomington, IL 61701, USA*

^c *Chestnut Health Systems, 720 N. Wells, Suite 300, Chicago, IL 60610, USA*

Received 6 April 2004; received in revised form 12 October 2005; accepted 13 October 2005

Abstract

This paper examines the need, feasibility, and validity of combining two biometric (urine and saliva) and three self-report (recency, peak quantity, and frequency) measures of substance use for marijuana, cocaine, opioids, and other substances (including alcohol and other drugs). Using data from 337 adults with substance dependence, we used structural equation modeling to demonstrate that these multiple measures are driven by the same underlying factor (substance use) and that no single measure is without error. We then compared the individual measures and several possible combinations of them (including one based on the latent factors and another based on the Global Appraisal of Individual Needs (GAIN) Substance Frequency Scale) to examine how well each predicted a wide range of substance-related problems. The measure with the highest construct validity in these analyses varied by drug and problem. Despite their advantages for detection, biometric measures were frequently less sensitive to the severity of other problems. Composite measures based on the substance-specific latent factors performed better than simple combinations of the biometric and psychometric measures. The Substance Frequency Scale from the GAIN performed as well as or better than all measures across problem areas, including the latent factor for any use. While the research was limited in some ways, it has important implications for the ongoing debate about the proper way to combine biometric and psychometric data.

© 2005 Elsevier Ireland Ltd. All rights reserved.

Keywords: Psychometric measures; Biometric measures

1. Introduction

Although frequently used in research, the accuracy of measures of substance use collected by self-report is often questioned and perceived as biased. A major point of contention remains regarding the veracity of self-report and, to a lesser degree, peer report. Whether as a result of conscious distortion, unconscious denial, or measurement artifact, the literature clearly demonstrates systematic differences in self-report and biometric measures (Amsel et al., 1976; Buchan et al., 2002; Cook et al., 1995; Darke, 1998; Dennis et al., 2003a; Harrison, 1995; Hersh et al., 1999; Katz et al., 2005; Landry et al., 2003; McNagny and Parker, 1992; Messina et al., 2000; Mieczkowski, 1990; Nelson et al., 1998; Preston et al., 1997; Skog, 1992; Stephens, 1972;

Stephens and Feucht, 1993; Weatherby et al., 1994; Wish et al., 1997). Biometric screening measures clearly identify people who denied recent use, but they also miss people who readily acknowledge use and can have limited utility for quantifying the extent of use of a long period of time—which is what is wanted in many clinical and clinical research settings.

A separate but closely related concern is how to best operationalize self-reported measures of substance use. Some of the options include days of use, amount of drugs consumed (over a period or on average), days of heavy use, times used, and peak use. Research comparing these measures has shown that they do not respond the same to demand characteristics but rather follow the logic of fuzzy number sets (Matt and Wilson, 1994; Matt et al., 2003). Lennox et al. (1996) combined quantity and frequency items into a single latent variable, using two quantity and two frequency items as measured indicators of the single heavy-drinking variables. This approach argues that the two types of measures can be considered fallible effect-indicators (Bollen and Lennox, 1991) of the same construct, that their intersection can be considered a more accurate measure of substance use, and that the latent variable can measure this alcohol use on a contin-

* Corresponding author at: Psychometric Technologies, 2404 Western Park Lane, Hillsborough, NC 27278, USA. Tel.: +1 919 245 0930/042 0448; fax: +1 919 245 0940/933 0797.

E-mail addresses: rlennox@psychometricstech.com (R. Lennox), mdennis@chestnut.org (M.L. Dennis).

¹ Tel.: +1 309 820 3805; fax: +1 309 829 4661.

uum of light to heavy drinking. Lennox et al. (1996) confirmed this structure in a sample from the National Household Survey on Drug Abuse and demonstrated that the latent variable model outperformed standard approaches to measuring alcohol use, including summed score of items and the multiplicative product of quantity and frequency in predicting work-related adverse consequences. Lennox et al. (1998) also used this approach to test the unique impact of heavy drinking and alcohol dependence on adverse alcohol-related consequences.

Compared to psychometric measures, biometric measures of substance use (e.g., from urine, saliva, blood, or hair) are often tacitly accepted as a more accurate assessment of use because they are not subject to various biases that often impact the accuracy of self-report. While they are certainly less likely to be biased due to individual demand characteristics, they too are impacted by several sources of error (Buchan et al., 2002; Cone and Weddington, 1989; Del Boca and Noll, 2000; Del Boca and Darkes, 2003; Dennis et al., 2003a,b, 2004; Feucht et al., 1994; Mieczkowski et al., 1991; Visher and McFadden, 1991; Bennett et al., 2003). The greatest threat lies in a systematic error that comes from using a static snapshot of the biological state to infer a more general consumption level over a period of time (Goldstein and Brown, 2003). For example, blood tests may only reliably detect drugs that were used in a 1-h period prior to the test, saliva tests for substances used during the prior 1–2 days, and urine tests for use during the prior 1–7 days. As such, these tests may be insensitive to longer term and more devastating chronic drug use patterns.

A second problem is that researchers frequently ignore the large individual differences in the rates at which various drugs are metabolized. Published cut points are typically based on the 50th or 80th percentiles for how long people will continue to be positive after a given use. In some cases the actual distribution can go out days or weeks further (Buchan et al., 2002). Thus, there is no perfect way of comparing these measures of “metabolites” with self-reported measures of recency, frequency, or quantity of use. Other sources of error include relying on less expensive and/or faster screening tests that do not always agree with gas chromatography/mass spectrometry (or GC/MS, the gold standard) and often suffering from processing delays (particularly with unfrozen samples), handling errors, and participants tampering with the sample.

Rather than accepting one approach or the other, most researchers favor combining the results of multiple methods (Campbell and Fiske, 1959; Cone and Weddington, 1989; Hasin et al., 2003; Lennox and Dennis, 1994; Carroll, 1995; Kranzler et al., 1997). However, there is little research comparing the construct validity of different approaches for combining self-reported and biometric measures of substance use. Some alternatives include: (a) requiring each test to indicate substance use, (b) allowing any indication of substance use to be sufficient, and (c) combining the data into a dimensional measure (via structural equation modeling or a scale) that increases as there are more (and stronger) indications of use. Since no single approach can be considered a “criterion,” it is necessary to focus on the construct validity of the alternative approaches (Cronbach and Meehl, 1955). The construct validity of a measure is established

by demonstrating that the measure is related to other measures in ways that are consistent with theory. No single test or relationship is sufficient to establish the validity of a measure, but rather a pattern of relationships is required to eliminate alternative interpretations to the measure.

A tangential issue that has been recently debated in this journal is whether information should be combined by substance or across multiple substances, especially when using the measure for prognosis and to track outcomes (see Rounsaville et al., 2003; O'Brien and Lynch, 2003; Strain, 2003; Conway et al., 2003). This is an important issue because over one third of the people with abuse or dependence in the community and two thirds of those in treatment have multiple substance use disorders (Substance Abuse and Mental Health Services Administration, 2003a,b). While it may be useful to identify individual patterns of use (e.g., offering methadone to an opioid user), it is likely that symptoms of withdrawal, abuse/dependence, and other substance-related problems are multiply determined by all of the substances being used. Thus, in addition to looking at how to combine psychometric and biometric information within single substances, it is also important to consider what happens when this information is combined across substances. In this article we will use data from a large cohort of chronic substance users to empirically examine the relationship of multiple psychometric and biometric measures of substance use and the construct validity of different approaches for combining them.

2. Method

2.1. Data source

The data from this paper came from the 12-month post-intake wave of the early re-intervention (ERI) experiment² (Dennis et al., 2003a,b; Scott et al., 2005), which was designed to evaluate the effectiveness of quarterly monitoring, checkups, and early re-intervention on long-term outcomes of persons with lifetime substance use dependence. As part of this experiment 448 adults with substance dependence were recruited at intake and then interviewed quarterly for 24 months. Half were randomly assigned to receive recovery management checkups (RMC) and half were randomly assigned to a control group. Briefly, RMC involves the following steps: (1) determine “eligibility” (i.e., verify that the person is not already in treatment or jail and is living in the community), (2) determine “need” for treatment based on self-report, (3) transfer the participant to the linkage manager, and (4) have the linkage manager complete the intervention. The intervention utilized motivational interviewing techniques to: (a) provide personalized feedback to participants about their substance use and related problems, (b) help the participant recognize the problem and consider returning to treatment, (c) address existing barriers to treatment, and (d) schedule an assessment. The linkage manager provided both the motivational interviewing and linkage assistance for those

² Questions on the ERI study can be addressed directly to the second author of this paper, Dr. Dennis.

individuals who agreed to participate. Linkage assistance often included scheduling, reminder calls, transportation, and in some cases escorting the participant to the intake appointment.

In any given quarter, 60–80% of the participants self-reported being eligible (i.e., in the community and not in treatment), and 25–35% reported being “eligible and in need” of treatment. From any given quarter to the next, the “need for treatment” status changed for 25% or more of the participants. Sessions were audio taped, and the results of the screener (to determine eligibility and need), motivational interview, and linkage assistance were documented on the RMC worksheet; both were reviewed by the protocol supervisor for staff adherence and used to generate performance measures to monitor the protocol’s overall implementation.

A post-intake cohort was used to ensure a mix in terms of the percent of people using in the community (52%), incarcerated (5%), in treatment (10%), and in the community without using (33%) (see Scott et al., 2005). Of the 448 people randomly assigned in ERI, 424 (95%) completed 12-month interviews. Of these 424, interviews were conducted in person on-site for 338 (80%), in person off-site at a jail, prison, or other institution for 57 (13%), and by phone for 29 (7%). Due to logistical constraints, no attempt was made to collect biometrics from people being interviewed in institutions or by phone. Of the 338 interviewed on-site, valid saliva and urine data were collected from 337 (99.7%) participants during the 12-month interviews (79% of all 12-month interviews, and 75% of those randomized). Thus, the findings here are targeted only to people in the community (not the 13% incarcerated) and have the potential for a small unknown bias associated with excluding the 7% interviewed by phone. There were no significant differences by condition in the kappa between self-reported and biometric measures, so assignment is not considered in this analysis.

2.2. Sample composition

The 337 participants were 59% female, 85% African American, 8% Caucasian, 6% Hispanic, and 2% others; 2% were between the ages of 18 and 20 years, 17% between 21 and 29 years, 47% between 30 and 39 years, 28% between 40 and 49 years, and 5% were 50 years or older. All met criteria for lifetime dependence at the time of intake, including 7% for alcohol only, 20% for cocaine and alcohol, 29% for cocaine only, 8% for cocaine and opioids, 14% for opioids only, and 17% for other patterns of use. Seventy-seven percent reported additional co-occurring mental health problems. Over 26% reported health problems that bothered them daily or interfered with their responsibilities weekly, and 25% of the females reported being pregnant in the past year.

2.3. Psychometrics

The participant characteristics and primary outcomes were measured with the Global Appraisal of Individual Needs (GAIN) (Dennis, 1999; Dennis et al., 2003a,b), which is a comprehensive, semi-structured interview comprised of eight main sections (background, substance use, physical health, risk behaviors,

mental health, environment, legal, and vocational). The GAIN contains over 100 scales, with the main ones in this data set having alphas over .9 and the subscales generally having alphas over .7. Diagnoses based on the GAIN have been shown to have good test–retest reliability for substance use disorders (kappa = .55; Dennis et al., 2003a,b) and accurately predict independent and blind staff psychiatric diagnoses of co-occurring psychiatric disorders including ADHD (kappa = 1.00), mood disorders (kappa = .85), Conduct Disorder/Oppositional Defiant Disorder (kappa = .82), Adjustment Disorder (kappa = .69), or the absence of a non-substance use diagnosis (kappa = .91) (Shane et al., 2003). Relative to substance use based on self-report, urine, or saliva, each method was largely (but not perfectly) consistent with the combined estimate of any use (kappa of .59 for self-report, .69 for urine, and .56 for saliva) (Dennis et al., 2003a,b). Self-reported data on the days of treatment from the GAIN were largely consistent with agency treatment records ($r = .78$) (Godley et al., 2002). Table 1 describes the various self-reported measures of substance use from the GAIN used for the validity analysis (see Dennis et al., 2003a,b for a detailed description). This includes measures of the internal consistency and/or test–retest reliability on a subsample of 75 people over 1–3 days.

2.4. Biometrics

Urine and saliva samples were collected and tested as biological markers of recent substance use. Urine samples were checked for color and temperature, refrigerated, then shipped overnight to a SAMHSA NLCP-certified laboratory (MedTox, retrieved from www.medtox.com on 20 January 2005). The laboratory conducted screenings using kinetic interaction of microparticles in solution (KIMS) at the SAMHSA standard cut-off levels for a panel of five drugs: cannabinoids (marijuana/THC; 50 ng/ml), cocaine (300 ng/ml), amphetamines (1000 ng/ml), opiates (2000 ng/ml), and phencyclidine (PCP; 25 ng/ml). The laboratory also tested for adulteration by checking creatinine levels (less than 20 ng/ml suggests adulteration or high levels of kidney hydration) and, if below the threshold, the specific gravity (less than 1.003 suggests dilution). Two cases were discarded because of validity check problems.

Saliva tests (as well as a validity check) were done on-site using ORAL•Screen™4 tests for THC, opiates, cocaine/crack, and methadone using a lateral-membrane immunoassay technique at the SAMHSA recommended cut-offs. While a dozen on-site tests had to be redone, the final test was always acceptable. Both approaches replicate well relative to gas chromatography/mass spectrometry and will detect over 90% of use in the past 2 days as well as some earlier use during the past 1–4 weeks.

2.5. Analysis

Using AMOS 4.0 (Arbuckle, 1999), we conducted four separate confirmatory factor analyses related to marijuana use, cocaine use, opioid use, and any drug use. For each analysis we used maximum likelihood to estimate the combined measure-

Table 1
Self-report: summary of key measures from GAIN^a

Self-reported measures of substance use	
Past-month use (test–retest kappa). Yes (1) or no (0) for self-reported use in the past month of marijuana (.93), cocaine (.94), opioids (.85), or any (including alcohol and other drugs) substance (.87) (recoded from recency below)	
Recency of use (test–retest rho). Recency of last use of marijuana (.94), cocaine (.95), opioids (.84), or any substance (.88) rated as 4: past 48 h, 3: 3–7 days ago, 2: 2–4 weeks ago, 1: 2–3 months ago, 0: more than 3 months ago/never	
Peak quantity of use (test–retest rho). During the past 90 days, the peak amount consumed of marijuana (standard joint, calculated as ounce = 25–30 joints; dime = 4–5 joints; nickel = 2–3 joints; 1 blunt = 2–6 joints; 1 g = 1–2 joints; 1 bowl = 1 joint; 10 one-hit pipes = 1 joint; capped 20 joints; rho = .97), cocaine (in standard rocks, calculated 8 ball = 32 rocks; teen = 16 rocks; gram = 10 rocks; quarter gram = 2.5 rocks; dime = 1 rock; nickel = 1 hit = 1/2 rock; capped at 20 rocks; rho = .94), opioids (in standard dime bags, calculated as 1 g = 10 dime bags; rho = .94), or any drug (using maximum of above or standard drinks, calculated as 1 standard drink = 1 beer = 1 glass wine = 1 mixed drink = 1 shot; 40 oz beer = 3 drinks; fifth = up to 26 drinks; capped at 20; rho = .89)	
Frequency of use (test–retest rho). During the past 90 days, how many days were reported using marijuana (.95), cocaine (.96), opioids (.95), or any (including alcohol and other drugs) substance (.94)	
Substance Frequency Scale (SFS; alpha = .85; test–retest rho = .94). The GAIN SFS is a multiple-item measure that averages percent of days reported of any AOD use, days of heavy AOD use, days of problem from AOD use, days of alcohol, marijuana, crack/cocaine, and heroin/opioid use	
Self-reported measures for validation	
Substance Problems Scale (SPS; alpha = .93; test–retest rho = .81). This is a count of past-month symptoms of substance abuse, dependence, or substance induced disorders and is based on DSM-IV (APA, 1994, 2001)	
Current Withdrawal Scale (CWS; alpha = .95; test–retest rho = .59). The GAIN CWS is a count of past-week psychological (tired and anxious) and physiological (e.g., diarrhea and fever) symptoms of withdrawal and is based on DSM-IV (APA, 1994, 2001)	
Recovery Environmental Risk Index (RERI; alpha = n.a.; test–retest rho = .75). This is an average of items (divided by their range) for the days (during the past 90) of alcohol in the home, drug use in the home, fighting, victimization, homelessness, and structured activities that involved substance use and the inverse (90 minus answer) percent of days going to self-help meetings and involvement in structured substance-free activities	
Illegal Activities Scale (IAS; alpha = .71; test–retest rho = .48). This is an average of items (divided by their range) for the recency of illegal activity and number of days (during the past 90) of any illegal activity and supporting oneself financially with illegal activity	
Emotional Problems Scale (EPS; alpha = .90; test–retest rho = .56). This is an average of items (divided by their range) for recency of mental health problems, memory problems, and behavioral problems and the number of days (during the past 90) of being bothered by mental problems, memory problems, and behavioral problems, and the number of days the problems kept participant from responsibilities	
Employment Activity Scale (EAS; alpha = .96; test–retest rho = .81). This is an average of items (divided by their range) for the recency of working and days (of 90) working, working full-time, and the inverse (90 minus answer) of the days in trouble, days missed, and days suspended	

n.a.: not applicable because this is a formative scale.

^a From the GAIN (Dennis et al., 2003b).

ment model for indicators of positive urine test, positive saliva test, self-reported recency of use, self-reported peak quantity of use, and self-reported frequency of use. For each substance the model was evaluated with multiple goodness-of-fit statistics (see Bollen, 1989; Byrne, 2001 for review), including the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), and the Root Mean Square of Approximation (RMSEA). Bentler and Bonnet (1980) replaced the earlier Normed Fit Index (NFI) with the CFI to adjust for sample size; CFI should be over .9, with .95 or higher being considered a very good fit. The TLI scales chi-square to a range of approximately 0–1 (null to perfect fit) and is one of the mostly commonly used fit indices that is not affected by sample size (Marsh et al., 1988). TLI should be over .8 (moderate fit), with .9 being a good fit and .95 or higher being considered a very good fit (Bentler and Bonnet, 1980; Tucker and Lewis, 1973). RMSEA should be less than .1, with values less than .08 being a moderate fit and less than .06 being a very good fit (Hu and Bentler, 1999). While good for complex models, RMSEA is very sensitive to the ratio of parameters to degrees of freedom and can be overly conservative in a simple model (as used here) with a single factor (Byrne, 2001).

For the analysis of construct validity, we used SPSS Version 11.0 to correlate various measures of substance use with the self-reported GAIN indices of substance (abuse/dependence) problems, withdrawal, recovery environment risk, illegal activity, emotional problems, and employment activity.

3. Results

3.1. Comparing the measures

For each substance (marijuana, cocaine, opioid, and any drug) we conducted a confirmatory factor analysis to see if they were being driven by the same underlying factor (i.e., unobserved substance use). Table 2 presents the standardized loadings (all

Table 2
Standardized factor loadings by substance use factor

	Standardized loadings			
	Marijuana	Cocaine	Opioids	Any drug
Measure				
Urine positive/negative	.47	.71	.76	.65
Saliva positive/negative	.36	.73	.76	.68
Self-report recency	.71	.89	.91	.89
Self-report quantity peak	.83	.52	.63	.46
Self-report frequency	.76	.61	.75	.60
Goodness-of-fit				
Degrees of freedom	5	5	5	5
Chi-square ^a	34.91	79.31	129.50	37.24
Comparative Fit Index	.96	.95	.90	.98
Tucker–Lewis Index	.87	.84	.71	.94
Root mean square error	.13	.21	.27	.14

^a All the chi-squares for goodness-of-fit had $p < .001$.

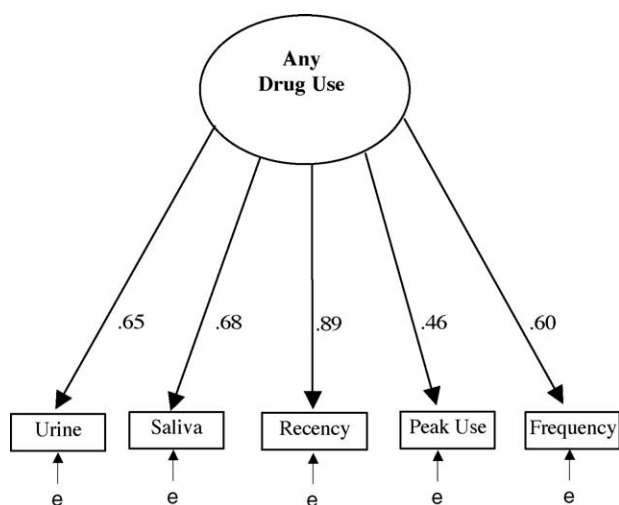


Fig. 1. Any substance use model.

significant) for each item in each of the four models along with the goodness-of-fit statistics for each model. Factor loadings that are significantly different than 0 (all in this study) are “reliably” measured, with higher loadings more closely reflecting variation in the unobserved or latent construct (in this case substance use). Loadings over .7 are particularly strong, suggesting that half or more of the variance in the observed measure is being driven by the latent construct. Fig. 1 illustrates the confirmatory factor analysis model for any drug use with its standardized factor loadings. Below is a short summary of the findings by substance:

- For marijuana use, all loadings were significantly different than 0, with higher (>.70) loadings occurring for self-reported peak use (.83), frequency of use (.76), and recency of use (.71). For the biometrics, urine had a higher loading than saliva (.47 versus .36). The goodness-of-fit was acceptable (CFI = .96; TLI = .87), though the RMSEA was high (.13).
- For cocaine use, all loadings were significantly different than 0, with higher loadings occurring for self-reported recency of use (.89), saliva (.73), and urine (.71); the goodness-of-fit was acceptable (CFI = .95; TLI = .81), though the RMSEA was high (.21).
- For opioid use, all loadings were significantly different than 0, with higher loadings occurring for self-reported recency of use (.91), urine (.76), saliva (.76), and self-reported frequency of use (.75); the goodness-of-fit was marginally acceptable (CFI = .90; TLI = .71), and the RMSEA was high (.21).
- For any substance use, all loadings were significantly different than 0, but only self-reported recency of use (.89) was high; the goodness-of-fit was very good (CFI = .98; TLI = .94), though the RMSEA was still high (.14).

On the whole, the loadings are respectable and indicative of a high degree of convergence among the five distinct measures. While the RMSEA all exceed .10 (indicating a modest fit to the data and the possibility of multiple factors or correlated errors), the residual covariances are normally distributed around 0, and allowing for a correlation among the error terms for the

various measures, did not bring the RMSEA below .10. The RMSEA is important because it suggests the lack of significant methods effects (which would have produced correlated errors within the self-report or biometric measures). While CFA often compares models, it is not always the case (see Bollen, 1989), especially when the implicit model is that the “Null model” is itself theoretically meaningful, as it is here. A comparison of the two-factor solutions for biometric and psychometric measures did not improve the model fit.

3.2. Construct validity

Next, we compared each individual measure, simple combinations of these measures, and complex combinations (structural equation modeling, GAIN Substance Frequency Scale) in terms of their ability to predict external criteria that are known to be correlated with actual substance use. Table 3 presents the zero-order correlations between the individual measures of substance use (and their various combinations) with the following problems related to substance use: withdrawal, substance abuse/dependence problems, health problems, emotional problems, recovery environment, illegal activity, and employment activity. Note that the percentage of variance explained for each of these problems by any given substance use measure is simply the square of the r in the table. The latent variable measure is a linear composite of the five measures (standardized to z -scores and weighted by their respective factor loading from Table 2). All the substance measures are stipulated as methods of establishing positive drug use status, with higher numbers indicating more recent or more use. Correlations in bold text are statistically significant at the .05 level. Below is a summary by substance:

- For marijuana use the biomarkers were generally uncorrelated with other problems while the psychometric measures were correlated. The latent variable did as well as the simple frequency of substance use. The one exception was the emotional problem scale, where the “peak” use did much better than any of the other measures.
- For cocaine use both the biometric and psychometric measures were correlated with these other problems, though again the psychometric measures were always higher (sometimes two-fold). The latent variable did as well or slightly worse than the simple self-reported frequency of use (but better than all other variables or their combination).
- For opioid use there was a similar pattern, with both the biometric and psychometric measures being correlated with these other problems. However, there was also more diversity in what worked best. For withdrawal and illegal activity, frequency and recency of use worked better than the other measures (including the latent variable). For substance problems ($r = .62$) and the other measures, the latent variable was as good or better than simple frequency of opioid use.
- For any substance use (including alcohol and other drugs), both biometric and psychometric measures were correlated with these problems. The rates are much higher in part because participants used a wide range of different drugs (frequently multiple) that have similar effects on these other problem

Table 3
Correlations between self-report, biometric, and combined measures of drug use and substance use-related problems

Method of determining positive drug use status	Substance Problems Scale	Withdrawal Problems Scale	Recovery Environment Risk Index	Illegal Activity Scale	Emotional Problems Scale	Employment Activity Scale
Marijuana use (n = 332)						
Urine only	.01	.02	.06	−.02	.02	.06
Saliva only	.06	.07	.05	.08	−.12	−.02
Self-report past month only	.19	.20	.20	.09	.04	−.08
Self-report recency only	.19	.18	.21	.10	.06	−.05
Self-report peak use only	.18	.18	.24	.14	.22	−.11
Self-report frequency only	.16	.23	.24	.20	.14	−.03
Urine and saliva positive	.11	.16	.13	.04	−.04	−.03
Urine and past month self-report	.06	.05	.10	.00	.02	.02
Saliva and past month self-report	.10	.12	.15	.13	−.07	−.04
All three methods	.11	.16	.13	.04	−.04	−.03
Any of the three methods	.13	.15	.11	.05	.00	−.03
Latent variable	.19	.21	.25	.15	.12	−.05
Cocaine use (n = 336)						
Urine only	.37	.14	.21	.10	.07	−.14
Saliva only	.37	.17	.19	.09	.09	−.09
Self-report past month only	.58	.34	.34	.17	.21	−.18
Self-report recency only	.59	.29	.35	.19	.20	−.19
Self-report peak use only	.41	.21	.32	.17	.16	−.09
Self-report frequency only	.64	.22	.48	.31	.24	−.21
Urine and saliva positive	.37	.14	.17	.07	.07	−.07
Urine and self-report past month	.54	.26	.28	.15	.14	−.19
Saliva and self-report past month	.46	.24	.22	.13	.13	−.12
All three methods	.18	.12	.10	.10	−.04	−.03
Any of the three methods	.40	.20	.27	.13	.11	−.17
Latent variable	.62	.27	.39	.22	.20	−.19
Opioids use (n = 335)						
Urine only	.20	.20	.10	.18	.07	−.08
Saliva only	.21	.17	.17	.17	.06	−.08
Self-report past month only	.34	.26	.22	.22	.18	−.05
Self-report recency only	.35	.27	.20	.25	.16	−.06
Self-report peak use only	.28	.32	.25	.21	.15	−.09
Self-report frequency only	.40	.30	.27	.32	.17	−.08
Urine and saliva positive	.20	.16	.12	.19	.06	−.06
Urine and past-month self-report	.27	.19	.13	.22	.10	−.04
Saliva and past-month self-report	.30	.21	.19	.21	.11	−.04
All three methods	.16	.08	.11	.08	.01	−.04
Any of the three methods	.24	.23	.15	.20	.07	−.08
Latent variable	.62	.26	.39	.22	.20	−.20
Any substance use (n = 337)						
Urine only	.32	.21	.22	.14	.09	−.14
Saliva only	.32	.22	.21	.16	.05	−.09
Self-report past month only	.52	.38	.38	.19	.16	−.21
Self-report recency only	.54	.34	.37	.22	.15	−.18
Self-report peak use only	.41	.31	.35	.14	.20	−.13
Self-report frequency only	.69	.42	.54	.41	.20	−.23
Urine and saliva positive	.37	.25	.23	.16	.09	−.07
Urine and past-month self-report	.46	.32	.30	.20	.14	−.16
Saliva and past-month self-report	.41	.29	.27	.19	.09	−.09
All three methods	.43	.30	.26	.18	.12	−.08
Any of the three methods	.38	.26	.30	.14	.10	−.20
Latent variable	.63	.41	.47	.29	.19	−.21
Maximum correlation above	.69	.42	.54	.41	.24	−.23
Substance Frequency Scale	.71	.40	.55	.48	.28	−.24

Note: Bold indicates $p < .05$.

areas. Again, the best two measures were consistently either the self-reported frequency of use or the latent variable.

After “any drug use,” we have also included the maximum correlation of any individual substance use measure with each of these scales and the GAIN Substance Frequency Scale (SFS). This scale combines the self-reported measures of use without collapsing them to dichotomies and is designed to provide a more robust measure of self-reported use. While the specific measure that correlated best with the substance-related problem was different for different problems and for different substances, the SFS correlation with each respective problem is about the same or higher—particularly in terms of substance problems ($r = .71$), recovery environment risk ($r = .55$), illegal activity ($r = .48$), and employment activity ($r = -.24$).

While early research often focused on the advantages of biometrics for increasing “detection,” they were not as useful for predicting other measures, in part because they are limited to dichotomous responses. Variability by drug and problem area, in terms of which measure was the most correlated with this list of substance-related problems, is further evidence of the need for a composite measure. The latent measure consistently did better than the various logical combinations of biometrics and self-reported past-month use, and it is worth noting that it did much better than the “any method” approach that anecdotally appears to be the most common in practice. Moreover, the SFS (based only on self-reports) did as well or better than the latent variable with biometrics included and appears to be a useful alternative when biometrics are not available. It should also be noted that while frequency or days of use (again one of the more widely used measures) were not always the best, they actually did fairly well.

4. Discussion

The results of this analysis support our contention that psychometrics and biometrics each primarily measures a single underlying latent variable (substance use) and they *each* do so imperfectly due to a combination of measurement error and differences in how they are operationalized. The results of the confirmatory factor analysis were mixed; however, the model fits sufficiently well to support the general strategy of combining biometric and psychometric in a single composite measure. The correlational analysis also showed support for using a composite measure, since the measure that worked best varied by substance and problem. The latent variable approach and GAIN SFS both clearly worked better than the common approach of combining self-report and urine data by collapsing the former into a dichotomy. The success of the SFS is particularly useful, since it is not always feasible to collect biometric data.

We do, however, need to acknowledge the limitations of this work. First, the RMSEAs were higher than desired, suggesting that with more measures we might have been able to improve accuracy by specifying a more complex model (e.g., with separate factors for recency and frequency of use), or other variables that influence the answers (e.g., use of other drugs and interview context). This was particularly true for opioids, where the

RMSEA and TLI were lower than one would typically want to see. Second, it was also very surprising that days of use (an individual simple measure) did as well or better than several of the composite measures on many correlations. That such a simple measure would do so well is good news for many, though obviously these findings need to be replicated. Third, it is possible that the correlation of the self-reported measures (including SFS) with the self-reported problem scales is partially due to a methods artifact. Ideally, this analysis should be replicated using collateral or observational data to reduce this potential bias. Finally, it is important to note that our analysis focused on the accurate quantification of substance use. If we had relied on (the much more expensive) quantified urine or saliva results, they might have been more useful. Conversely, if we had focused on detection of any use (as a 0/1 measure), it is possible that the biometrics might have been more useful, particularly in a context like a jail or detention center or as part of a protocol designed to reduce underreporting.

These results also provide some theoretical perspective on the use of summary scores in social research. Although most social scientists believe that multiple measures are better than single item measures, most operationalizations are based more on faith than on empirical verification or even specification. There appears to be three categories of measurement models in use in conventional social science: the opaque model, the translucent model, and the transparent model. The opaque model consists entirely of a computation procedure such as a simple sum or average without any accompanying logic for how the calculation approximates the latent construct. The combination of the items is not structurally verifiable and, therefore, resembles a black box that does not allow us to look at the operation of the computational procedures. The translucent model usually includes some psychometric logic for combining items, such as true score theory or a less formal linear composite logic. However, it stops short of providing a method validating the structural integrity of the computational procedures, although it often cites summary internal consistency measures as evidence of scalability, and occasionally, even with the models, it is inconsistent with the assumption that the items need to be highly intercorrelated with one another (Bollen and Lennox, 1991). Here again, the combination of the items is only partially verifiable by use of the summary statistics and only interpretable if the specific measurement model is correct. The transparent model not only offers a psychometric theory, but it also offers a formal specification of the measurement model that is testable and falsifiable. For example, if a scale is created using an explicit “effect-indicator model” (Bollen and Lennox, 1991), then the logic is ultimately testable and falsifiable. From a research practice perspective, our latent variable model may have some advantages over any of the other multi-item approaches simply because its structure is explicit and testable. We also need to examine other populations and explore a broader range of external criteria.

Future studies are needed to better understand the exact nature of the systematic bias in self-report measures. These results suggest that it is not enough to simply assert that self-reported measures are underreported—or that they are without problems. Conversely, the results suggest that the biometrics come with

their own sets of measurement limits that reduce their construct validity and utility for predicting other problems. There is clear convergence among the different approaches, suggesting that it is probably the most useful to report both individual measures/substances and composite measures that go across them. It would also be useful in the future to examine the ability of different measures and composite measure to predict outcomes over time.

Acknowledgements

This work was completed with support provided by the National Institute on Drug Abuse Grant No. DA 11323. The second author developed the GAIN, but no royalties are associated with the GAIN. The authors would like to thank Joan Unsicker and Tim Feeney for assistance in preparing the manuscript and the study staff and participants for their time and effort.

References

- Amsel, Z., Mandell, W., Matthias, L., Mason, C., Hoeherman, I., 1976. Reliability and validity of self-reported illegal activities and drug use collected from narcotic addicts. *Int. J. Addict.* 11, 325–336.
- Arbuckle, J.L., 1999. AMOS 4.0 (Computer Software). Smallwaters, Chicago, IL.
- Bennett, G.A., Davies, E., Thomas, P., 2003. Is oral fluid analysis as accurate as urinalysis in detecting drug use in a treatment setting? *Drug Alcohol Depend.* 72, 265–269.
- Bentler, P.M., Bonnet, D.G., 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 588–606.
- Bollen, K., Lennox, R., 1991. Conventional wisdom on measurement: a structural equation perspective. *Psychol. Bull.* 110, 305–314.
- Bollen, K.A., 1989. *Structural Equations with Latent Variables*. Wiley, New York.
- Buchan, B.J., Dennis, M.L., Tims, F.M., Diamond, G.S., 2002. Cannabis use: consistency and validity of self report, on-site urine testing, and laboratory testing. *Addiction* 97, S98–S108.
- Byrne, B.M., 2001. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. Lawrence Erlbaum, Mahwah, NJ.
- Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multi-trait–multi-method matrix. *Psychol. Bull.* 56, 81–105.
- Carroll, K.M., 1995. Methodological issues and problems in the assessment of substance use. *Psychol. Assess.* 7, 349–358.
- Cone, E.J., Weddington Jr., W.W., 1989. Prolonged occurrence of cocaine in human saliva and urine after chronic use. *J. Anal. Toxicol.* 13, 65–68.
- Conway, K.P., Kane, R.J., Ball, S.A., Poling, J.C., Rounsaville, B.J., 2003. Personality, substance of choice, and polysubstance involvement among substance dependent patients. *Drug Alcohol Depend.* 71, 65–75.
- Cook, R.F., Bernstein, A.D., Arrington, T.L., Andrews, C.M., Marshall, G.A., 1995. Methods for assessing drug use prevalence in the workplace: a comparison of self-report, urinalysis, and hair analysis. *Int. J. Addict.* 30, 403–426.
- Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302.
- Darke, S., 1998. Self report among injecting drug users: a review. *Drug Alcohol Depend.* 51, 253–263.
- Del Boca, F.K., Darkes, J., 2003. The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction* 98 (S2), 1–12.
- Del Boca, F.K., Noll, J.A., 2000. Truth or consequences: the validity of self-report data in health services research on addictions. *Addiction* 95, S347–S360.
- Dennis, M.L., 1999. *Global Appraisal of Individual Needs (GAIN): Administration Guide for the GAIN and Related Measures (Version 1299)*. Chestnut Health Systems, Bloomington, IL.
- Dennis, M.L., Godley, S.H., Diamond, G., Tims, F.M., Babor, T., Donaldson, J., Liddle, H., Titus, J.C., Kaminer, Y., Webb, C., Hamilton, N., Funk, R., 2004. The Cannabis Youth Treatment (CYT) study: main findings from two randomized trials. *J. Subst. Abuse Treat.* 27, 197–213.
- Dennis, M.L., Scott, C.K., Funk, R., 2003a. An experimental evaluation of recovery management checkups (RMC) for people with chronic substance use disorders. *Eval. Program Plann.* 26, 339–352.
- Dennis, M.L., Titus, J.C., White, M., Unsicker, J., Hodgkins, D., 2003b. *Global Appraisal of Individual Needs (GAIN): Administration Guide for the GAIN and Related Measures*. Chestnut Health Systems, Bloomington, IL (retrieved from <http://www.chestnut.org/li/gain> on 20 January 2005).
- Feucht, T.E., Stephens, R.C., Walker, M.L., 1994. Drug use among juvenile arrestees: a comparison of self-report, urinalysis and hair assay. *J. Drug Issues* 24, 99–116.
- Godley, M.D., Godley, S.H., Dennis, M.L., Funk, R., Passetti, L., 2002. Preliminary outcomes from the assertive continuing care experiment for adolescents discharged from residential treatment. *J. Subst. Abuse Treat.* 23, 21–32.
- Goldstein, A., Brown, B.W., 2003. Urine testing in methadone maintenance treatment: applications and limitations. *J. Subst. Abuse Treat.* 25, 61–63.
- Harrison, L.D., 1995. The validity of self-reported data on drug use. *J. Drug Issues* 25, 91–111.
- Hasin, D.S., Schuckit, M.A., Martin, C.S., Grant, B.F., Bucholz, K.K., Helzer, J.E., 2003. The validity of DSM-IV alcohol dependence: what do we know and what do we need to know? *Alcohol. Clin. Exp. Res.* 27, 244–252.
- Hersh, D., Mulgrew, C.L., Van Kirk, J., Kranzler, H.R., 1999. The validity of self-reported cocaine use in two groups of cocaine abusers. *J. Consult. Clin. Psychol.* 67, 37–42.
- Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equation Model.* 6, 1–55.
- Katz, C.M., Webb, V.J., Decker, S.H., 2005. Using the Arrestee Drug Abuse Monitoring (ADAM) program to further understand the relationship between drug use and gang membership. *Justice Q.* 22, 58–88.
- Kranzler, H.R., Tennen, H., Babor, T.F., Kadden, R.M., Rounsaville, B.J., 1997. Validity of the longitudinal, expert, all data procedure for psychiatric diagnosis in patients with psychoactive substance use disorders. *Drug Alcohol Depend.* 45, 93–104.
- Landry, M., Brochu, S., Bergeron, J., 2003. Validity and relevance of self-report data provided by criminalized addicted persons in treatment. *Addict. Res. Theory* 11, 415–426.
- Lennox, R.D., Dennis, M.L., 1994. Measurement error issues in substance abuse services research: lessons from structural equation modeling and psychometric theory. *Eval. Program Plann.* 17, 399–407.
- Lennox, R.D., Steele, P., Zarkin, G., Bray, J., 1998. The differential effects of alcohol consumption and dependence on adverse alcohol-related consequences. *Drug Alcohol Depend.* 50, 211–220.
- Lennox, R.D., Zarkin, G.A., Bray, J.W., 1996. Latent variable models of alcohol-related constructs. *J. Subst. Abuse* 8, 241–250.
- Marsh, H.W., Balla, J.R., McDonald, R.P., 1988. Goodness of fit indices in confirmatory factor analysis: the effect of sample size. *Psychol. Bull.* 103, 391–410.
- Matt, G.E., Turingan, M.R., Dinh, Q.T., Felsch, J.A., Hovell, M.F., Gehrman, C., 2003. Improving self-reports of drug-use: numeric estimates as fuzzy sets. *Addiction* 98, 1239–1247.
- Matt, G.E., Wilson, S.J., 1994. Describing the frequency of marijuana use: fuzziness and context-dependent interpretation of frequency expressions. *Eval. Program Plann.* 17, 357–369.
- McNagny, S.E., Parker, R.M., 1992. High prevalence of recent cocaine use and the unreliability of patient self-report in an inner-city walk-in clinic. *JAMA* 267, 1106–1108.
- Messina, N.P., Wish, E.D., Nemes, S., Wraight, B., 2000. Correlates of under-reporting of post-discharge cocaine use among therapeutic community clients. *J. Drug Issues* 30, 119–132.

- Mieczkowski, T., 1990. The accuracy of self-reported drug use: an evaluation and analysis of new data. In: Weisheit, R. (Ed.), *Drugs, Crime and the Criminal Justice System*. Anderson Publishing, Cincinnati, OH, pp. 275–302.
- Mieczkowski, T., Barzelay, D., Gropper, B., Wish, E., 1991. Concordance of three measures of cocaine use in an arrestee population: hair, urine, and self-report. *J. Psychoactive Drugs* 23, 241–249.
- Nelson, D.B., Kotranski, L., Semaan, S., Collier, K., Lauby, J., Feighan, K., Halbert, J., 1998. The validity of self-reported opiate and cocaine use by out-of-treatment drug users. *J. Drug Issues* 28, 483–494.
- O'Brien, C.P., Lynch, K.G., 2003. Can we design and replicate clinical trials with a multiple drug focus? *Drug Alcohol Depend.* 70, 135–137.
- Preston, K.L., Silverman, K., Schuster, C.R., Cone, E.J., 1997. Comparison of self-report drug use with quantitative and qualitative urinalysis for assessment of drug use in treatment studies. In: Harrison, L., Hughes, A. (Eds.), *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*. Rockville, MD, NIDA Research Monograph, No. 167, pp. 130–145.
- Rounsaville, B.J., Petry, N.M., Carroll, K.M., 2003. Single versus multiple drug focus in substance abuse clinical trials research. *Drug Alcohol Depend.* 70, 117–125.
- Scott, C.K., Dennis, M.L., Foss, M.A., 2005. Recovery management checkups to shorten the cycle of relapse, treatment re-entry, and recovery. *Drug Alcohol Depend.* 78, 325–338.
- Shane, P., Jasiukaitis, P., Green, R.S., 2003. Treatment outcomes among adolescents with substance abuse problems: the relationship between comorbidities and post-treatment substance involvement. *Eval. Program Plann.* 26, 393–402.
- Skog, O.J., 1992. The validity of self-reported drug use. *Br. J. Addict.* 87, 539–548.
- Stephens, R., 1972. The truthfulness of addict responses in research projects. *Int. J. Addict.* 7, 549–558.
- Stephens, R.C., Feucht, T.E., 1993. Reliability of self-reported drug use and urinalysis in the drug use forecasting system. *Prison J.* 73, 279–289.
- Strain, E.C., 2003. Single versus multiple drug focus in substance abuse clinical trials research: the devil is in the details. *Drug Alcohol Depend.* 70, 131–134.
- Substance Abuse and Mental Health Services Administration, 2003a. Results from the 2002 National Survey on Drug Use and Health: National Findings. SAMHSA, Rockville, MD (Office of Applied Studies, NHSDA Series H-22, DHHS Publication No. SMA 03-3836).
- Substance Abuse and Mental Health Services Administration, 2003b. Treatment Episode Data Set (TEDS): 1992–2001. National Admissions to Substance Abuse Treatment Services, DASIS Series: S-20. SAMHSA, Rockville, MD (DHHS Publication No. (SMA) 03-3778).
- Tucker, L.R., Lewis, C., 1973. The reliability of coefficient for maximum likelihood factor analysis. *Psychometrika* 38, 1–10.
- Visher, C., McFadden, K., 1991. A Comparison of Urinalysis Technologies for Drug Testing in Criminal Justice. National Institute of Justice, Washington, DC.
- Weatherby, N.L., Needle, R., Cesari, H., Booth, R., McCoy, C., Watters, J., Williams, M., Chitwood, D., 1994. Validity of self-reported drug use among injection drug users and crack cocaine users recruited through street outreach. *Eval. Program Plann.* 17, 347–355.
- Wish, E.D., Hoffman, J.A., Nemes, S., 1997. The validity of self-reports of drug use at treatment admission and at follow-up: comparisons with urinalysis and hair assays. In: Harrison, L. (Ed.), *The Validity of Self-Reports: The Implications for Survey Research*. National Institute on Drug Abuse, Rockville, MD, pp. 200–225.