April 2012

Evaluation of the OTMA and ASUS-R among Colorado Offenders

Submitted by the Research Subcommittee of the Interagency Advisory Committee on Adult and Juvenile Correctional Treatment

Appointed Members
Interagency Advisory Committee on Adult and Juvenile Correctional Treatment

Marc Condojani
Director of Community Treatment and Recovery Programs, Colorado Department of Human Services, Division of Behavioral Health

Caren Leaf
Associate Director, Colorado Department of Human Services, Division of Youth Corrections

Jeaneene Miller*
Director, Colorado Department of Corrections, Division of Adult Parole, Community Corrections, and Youthful Offender System

Barry Pardus
Assistant Director, Colorado Department of Corrections, Finance and Administration

Eric Philp
Director, Colorado Division of Probation Services

Tom Quinn*
Director, Colorado Division of Probation Services

Charlie Smith*
Director, Colorado Department of Human Services, Division of Behavioral Health

Jeanne Smith
Director, Colorado Department of Public Safety, Division of Criminal Justice

*Former members

Report prepared by
Laura Engleman and Kasey R. Weber
Colorado Department of Corrections, Office of Planning & Analysis

In collaboration with
Kevin L. Ford
Division of Criminal Justice, Colorado Department of Public Safety
Office of Research and Statistics
William Dieterich, Markus Breitenbach, and William L. Oliver
Northpointe Institute for Public Management

with contributions from
Kelli J. Klebe
Department of Psychology, University of Colorado Colorado Springs

and Members of the Research Subcommittee
Interagency Advisory Committee on Adult and Juvenile Correctional Treatment

Laura Engleman
Colorado Department of Corrections, Office of Planning & Analysis

Peg Flick
Division of Criminal Justice, Colorado Department of Public Safety
Office of Research and Statistics

Kevin L. Ford
Division of Criminal Justice, Colorado Department of Public Safety
Office of Research and Statistics

Sherri Hufford
Division of Probation Services, Colorado Judicial Department

Ann Jones
Division of Behavioral Health, Colorado Department of Human Services

Bennie Lombard
Division of Behavioral Health, Colorado Department of Human Services

Glenn Tapia
Division of Criminal Justice, Colorado Department of Public Safety
Office of Community Corrections

Alexandra Walker
Division of Criminal Justice, Colorado Department of Public Safety
Office of Community Corrections

Kasey R. Weber
Colorado Department of Corrections, Office of Planning & Analysis

Dana Wilks
Division of Probation Services, Colorado Judicial Department

Table of Contents

Executive Summary

Introduction

As the relationship between substance abuse and criminal activity is undisputed, the state of Colorado intensified its commitment to substance abuse treatment across its criminal justice system with the passage of Colorado Revised Statute (C.R.S.) 16-11.5 in 1991. This legislation mandated standardized substance abuse assessment and treatment as well as punitive sanctions for offenders who fail to remain drug free. The Colorado Department of Corrections (CDOC), the State Judicial Department (Probation), the Division of Criminal Justice (DCJ), and the Department of Human Services have been collaborating on the Interagency Advisory Committee on Adult and Juvenile Correctional Treatment (IACAJCT) to fulfill these legislative mandates.

The current Standardized Offender Assessment-Revised (SOA-R) consists of four instruments:

- Level of Supervision Inventory-Revised (LSI-R) as a recidivism risk measure
- Simple Screening Instrument-Revised (SSI-R) as a substance abuse screening measure
- Adult Substance Use Survey-Revised (ASUS-R) as an addiction severity measure
- Treatment Recommendation Worksheet (TxRW) to derive a treatment level

A 2006 evaluation of the SOA-R recommended that the battery be replaced by an evidence-based, standardized, treatment-matching measure. In response, the Offender Treatment Matching Algorithm (OTMA) was developed in 2009 as a possible replacement (Melnick, Schoeneberger, & Sacks, 2009).

The current study sought to

- conduct a thorough literature review to explore assessment approaches and quality (including the issues of clinical judgment vs. statistical prediction, the accuracy of self-report data, and treatment matching);
- evaluate the measurement characteristics of the ASUS-R and OTMA scales that assess substance use/abuse and substance dependence;
- select the best substance abuse severity measure, considering psychometric properties as well as administration time, ease of use, cost, training, and staffing.

Literature Review

Debates over the best approach to assessment have revolved around whether statistical prediction outperforms clinical judgment. Although numerous studies across more than five decades have supported statistical prediction, many clinicians continue to rely primarily or exclusively on their professional expertise (Dawes, Faust, & Meehl, 1989; Taleff, 2006). Further complicating the argument is the issue of whether assessments based on self-report data are accurate. The research is not clear cut, and the best conclusion one can draw is that self-report data are valuable as long as the information is not relied upon exclusively.

Colorado has been committed to incorporating both criminal risk and addiction severity into its assessment approach, using the Risk-Need-Responsivity (RNR) model. That model identifies an offender's level of recidivism risk and criminogenic needs (factors predicting criminal behavior) so that the most effective treatment techniques may be utilized (responsivity; Andrews & Dowden, 2007). The RNR model suggests that matching offenders to the appropriate level and type of treatment may improve outcomes (Melnick, De Leon, Thomas, & Kressel, 2001; Simpson, Joe, Fletcher, Hubbard, & Anglin, 1999; Thanner & Taxman, 2003). Although the results of treatment-matching studies have not always been convincing, those findings perhaps can be attributed to the sheer complexity of the process, particularly when biopsychosocial factors beyond addiction severity are taken into account.

Method

- Participants were Colorado male and female adult inmates, parolees, and probationers assessed for substance abuse treatment needs from February to May 2011.
- Participating agencies were CDOC (Denver Reception and Diagnostic Center and Treatment Accountability for Safer Communities), probation, and community corrections.
- The total sample size was 238 (130 DOC inmates and parolees, 57 probationers, and 51 enrolled in community corrections programs), including 181 males and 57 females aged 19 to 55 years old (M = 32.8, SD = 9.2).
- Researchers, case managers, and assessors collected data after obtaining consent from voluntary participants and advising them that completing the two research instruments would extend the intake process by 30 to 40 minutes.
- Assessments administered were the LSI-R and ASUS-R, as part of the routine intake process, and the OTMA and the Global Appraisal of Individual Needs-Quick Core (GAIN-Q Core), as part of the research study. The order of the instruments was counterbalanced.
- When the OTMA was administered, a second independent rater joined the primary rater to record offenders' responses to determine the instrument's interrater reliability.

Results and Recommendations

- Both the ASUS-R and the OTMA demonstrated favorable psychometric properties in the evaluated areas, the ASUS-R having only a slightly better performance overall.
- The internal consistency among items was very good for both assessments. Interrater reliability of the OTMA was good, as expected, due to the highly structured nature of the questions. The ASUS-R was not tested for interrater reliability because it is a self-administered assessment.
- Validity tests also revealed high convergent validity, indicating the OTMA and the ASUS-R were measuring the same concepts as the GAIN-Q (i.e., substance abuse and dependence).
- Tests of discriminant validity demonstrated low correlations, confirming the assessments were *not* measuring concepts related to specific (non-substance abuse-related) LSI-R

subscales. It should be noted, however, that the OTMA had higher correlations in this area compared to the ASUS-R, indicating discriminant validity was an area of weakness.

- The depth of the ASUS-R and its sound psychometric properties, particularly in regard to validity, support its continued use in the SOA-R process, compared to the OTMA. That recommendation is strengthened by feedback from staff in the field, where the majority felt the ASUS-R was superior, and the fact that the ASUS-R is an existing measure requiring no additional training. Thus, the equivocal findings offer weak justification for ousting the ASUS-R and replacing it with the OTMA.
- In addition, the OTMA requires more work to improve the layout, flow, and scoring issues, all of which would necessitate additional reliability and validity testing. Furthermore, in its current form, the OTMA's treatment matching scheme does not incorporate criminal risk—considered a critical component—in the treatment decision.
- Among the study's limitations were difficulty in recruiting participants, alterations in how the OTMA was administered (due to the study design), and a floor effect on both the use and dependence subscales of the GAIN-Q (i.e., scores clustered near the bottom of the range).
- Further validation in future studies will consider the issue of treatment matching and whether that process is correlated with reductions in substance abuse and recidivism.

Introduction

The relationship between substance abuse and criminal activity is undisputed. According to the Bureau of Justice Statistics, nearly one-third of federal and state inmates in 2004 committed their crime while under the influence of drugs; obtaining drug money was a motive for 17% of state and 18% of federal offenders. In addition, 45% of federal and 53% of state offenders met the criteria for drug abuse or dependence (Mumola & Karberg, 2006) according to the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*; American Psychiatric Association, 2000).

The state of Colorado intensified its commitment to substance abuse treatment across its criminal justice system with the passage of Colorado Revised Statute (C.R.S.) 16-11.5 in 1991. This legislation mandated (a) a standardized procedure for assessment of substance abuse, including chemical testing, (b) a system of education and treatment programs for substance abusers, and (c) a system of punitive sanctions for offenders who test positive after the initial test. The Colorado Department of Corrections (CDOC), the State Judicial Department (Probation), the Division of Criminal Justice (DCJ), and the Department of Human Services have been collaborating on the Interagency Advisory Committee on Adult and Juvenile Correctional Treatment (IACAJCT) to fulfill these legislative mandates.

Assessment and Treatment in Colorado

Colorado has taken on the unique task of incorporating mechanisms for determining both criminal risk and addiction severity in its assessment approach. Although risk is not specifically mentioned in C.R.S. 16-11.5, state criminal justice agencies have been keenly aware of the potential impact of substance abuse treatment on criminal behavior (IACAJCT, 2009). Thus, IACAJCT's mission has been to identify the most accurate instruments to assess substance abuse treatment needs and develop a treatment system to address them. As this interagency work has progressed over the past decade, many factors based on the extensive research in this field have been taken into consideration. Following is a discussion of these factors as well as an overview of the treatment models and assessment measures currently used in Colorado and the evaluation of this process to date.

Assessment Approaches

Debates over the best approach to assessment have been raging for more than 50 years, and Colorado is no exception. In his seminal work, *Clinical vs. Statistical Prediction* (1954), Meehl can be credited with lighting the fire under the controversy with his conclusion that statistical prediction, also known as the actuarial method, consistently outperforms clinical judgment (Grove & Lloyd, 2006). In the actuarial method, conclusions are based on empirical relationships between certain predictor and outcome variables (Taleff, 2006). For example, criminogenic factors that predict future criminal behavior (outcome) would be used in a risk assessment tool in the same way a life insurance agent uses lifestyle characteristics in formulas that predict life expectancy. Clinical approaches, on the

other hand, are based largely on professional expertise, although *structured professional judgment* may rely on certain theoretical and empirical guidelines (Guy, 2008).

Numerous studies have supported statistical prediction, reaching as far back as a 1928 study of parole outcomes for 3,000 offenders, in which a fairly crude algorithmic method was a more accurate predictor than the judgments of three prison psychiatrists (Grove & Meehl, 1996). Grove and Meehl (1996) cited 136 studies, with a wide range of predictive criteria, showing that the mechanical method is equal or superior to the clinician's conclusion. Statistical approaches have resulted in more accurate diagnoses, predictions of human behavior, and treatment recommendations (Dawes, Faust, & Meehl, 1989). The data notwithstanding, it appears that many clinicians disregard the research and continue to rely primarily or exclusively on their professional expertise (Dawes et al., 1989; Taleff, 2006).

While practitioners may resist the actuarial approach, few empirical studies offer evidence in favor of clinical prediction (Grove & Meehl, 1996). Hanson's (2005) review of violence risk assessment tools, including those using structured professional judgment and others employing actuarial methods, found the predictive validity of the instruments to be similar (moderate accuracy). More recently, Guy's (2008) meta-analysis of 113 studies concluded that both structured professional judgment and actuarial approaches had equal merit for assessing violence risk in the 44 cases for which direct comparisons were possible. Silver and Miller (2002) also cautioned against using actuarial risk assessment tools for social control, marginalizing less fortunate populations in the name of efficient institutional management.

Further complicating the clinical versus actuarial debate is the issue of whether or not self-report data are accurate. Particularly in the criminal justice environment, it might be assumed that substance abusers would be less than forthcoming about their illegal behavior. However, the research is not clear cut. Early studies largely supported the validity of self-report data. A National Institute of Mental Health study found that addicts' responses generally were reliable (Amsel, Mandell, Matthias, Mason, & Hocherman, 1976). Comparing urinalysis to self-reported drug use, Mieczkowski's (1990) data supported previous studies showing concordance among all substance types and at each level of use.

More recently, the development of better biological markers (e.g., improved urinalysis techniques and hair analysis) began to cast doubt upon the accuracy of drug use self-report (Harrison, 1995). Preliminary data from the Arrestee Drug Abuse Monitoring (ADAM) program showed that, based on urinalyses, male arrestees underestimated their drug use by 40% to 60% (Taylor, Fitzgerald, Hunt, Reardon, & Brownstein, 2001). Similarly, based on a sample of 2,327 from the ADAM study, Lu, Taylor, and Riley (2001) found that only about half of those who tested positive for drugs were truthful. In her review of studies on self-report data, Harrison (1995) concluded that valid responses depend on the recency of drug use (most recent use is less reliable), the social desirability of the drug (cocaine use was least often reported by arrestees), and variations in how the data were collected (responding aloud increases deception). Statistics also differ depending on whether both positive and

negative urinalyses are considered and the time frame in question—whether the urinalysis and self-report cover the same period. After conducting a meta-analysis of 24 studies examining drug use in high-risk populations, Magura and Kang (1996) found the validity of self-report data to be so variable and unpredictable that assessments of prevalence and treatment outcomes should be viewed with caution. The greatest self-reported accuracy appeared to occur at intake, compared to posttreatment.

Among other studies involving a criminal justice population, mixed findings on self-report data are swinging the pendulum of research and opinion back toward the middle. A study funded by the National Institute on Drug Abuse (NIDA) found that results of a self-administered assessment in a probation-based drug treatment program were consistent with interview-based intake assessments (Broome, Knight, Joe, & Simpson, 1996). In another study by three of the same authors, self-reported cocaine use 6 months after prison release was evaluated among drug treatment graduates and a control group of parolees (Knight, Hiller, Simpson, & Broome, 1998). Compared to hair analysis, cocaine use was underreported but less so among the treatment group. Utilizing data from the Drug Use Forecasting project (precursor to the ADAM program), Rosay, Najaka, and Herz (2007) showed that offenders had a greater tendency to underreport crack/cocaine use and overreport marijuana use. Finally, Crawley's (2009) data on a random sample of 424 recently incarcerated male inmates in the Nebraska prison system found that self-report screening instruments generate high levels of confidence (i.e., criterion validity, construct validity, and efficacy) compared to clinical psychological assessments. Perhaps the best conclusion one can draw is that self-report data are valuable as long as the information is not relied upon exclusively. Noted Magura and Kang (1996): "Self-reports are still the best way of attempting to measure the duration, frequency, intensity, and other patterns of drug use, as well as the routes of administration and social context of use" (p. 1148).

Assessment Quality

Given the studies cited, there is no doubt that the accuracy of assessment data can be variable. One way to defuse subjectivity is to use standardized measures based on sound psychometric principles (Nunnally & Bernstein, 1994). These principles include both reliability, an instrument's ability to produce consistent results under varying conditions, and validity, an instrument's ability to measure the intended construct. A reliable and valid assessment is listed as the first priority in the implementation of evidence-based practices in corrections (Clawson, Bogue, & Joplin, 2005). Types of reliability commonly used in research include internal consistency, which measures how well items on an instrument relate to each other; test-retest reliability, which compares results on the same test given at different time periods; and interrater reliability, which measures the consistency of scores between two raters. Validity may be parsed into a number of categories, including face validity, content validity, construct validity, and criterion validity. Face validity is present when the items in a measure simply appear appropriate on a logical basis, while content validity is present when the items on an instrument, based on accepted theory, represent the domain of interest, such as substance abuse. Construct validity—whether one's "construction" of a measure accurately reflects its purpose-- may be subdivided into convergent validity, whether the instrument is similar to others measuring the same construct, and discriminant validity, whether the instrument is dissimilar from

measures of unrelated constructs (Trochim, 2006). One form of criterion validity is predictive validity, how well a measure predicts behaviors believed to be associated with the construct. The critical importance of these principles is well-known, but identifying psychometrically sound measures that can be administered with fidelity in the correctional environment is challenging.

Practical Considerations

Psychometric issues surrounding substance abuse assessments may, at times, be overshadowed by practical matters, including cost, length/complexity, scope, and treatment options (Wells, 2003). Instruments may be free (in the public domain) or available for a fee; brief, self-administered screening measures or lengthy interviews; or cover a long time frame (years) or a short period (30 days). More complex diagnostic instruments may require credentialed clinicians and extensive training, increasing the burden on short-staffed facilities (Knight, Simpson, & Hiller, 2002). Another consideration is whether a facility can provide a wide range of treatment options (determined by a more extensive assessment) (Wells, 2003). Instruments based on the *DSM-IV* criteria for substance *dependence* will be more likely to identify offenders with serious drug problems who would be referred to intensive treatment. Institutions that offer a range of treatment options, including drug education at the lower end of the spectrum, may want to choose a measure that meets the criteria for substance *abuse* (Knight et al., 2002). Other factors that influence an organization's approach to assessment include its own readiness for change, resources, climate, and staff attributes (Simpson, 2002).

Criminal Risk Assessment

The Risk-Need-Responsivity (RNR) model describes an approach that identifies an offender's level of recidivism risk and criminogenic needs (factors predicting criminal behavior) so that the most effective treatment techniques may be utilized (responsivity; Andrews & Dowden, 2007). The RNR model has been shown to reduce recidivism by up to 35% (Andrews & Bonta, 2010). In the RNR model, the *risk principle* asserts that the level of treatment must match the offender's risk to reoffend; the *need principle* requires that the needs of the offender, particularly those most strongly linked to their criminal behavior (e.g., education, alcohol, and drug use), are identified and subsequently addressed in treatment; and the *responsivity principle* states that the most effective evidence-based treatment techniques for offenders should be utilized in the treatment process (Andrews & Bonta, 2006). The idea that certain dynamic (changing) factors can predict risk of reoffending is an underlying assumption that drives the RNR model. The eight major risk/need factors include the "big four"-- history of antisocial behavior, antisocial personality patterns, antisocial cognition, and antisocial associates—as well as family/marital problems, school/work problems, unsatisfactory leisure/recreation activities, and substance abuse. Based on this model, effective programs would be designed to match high-risk offenders with more intensive treatment and with their individual learning abilities, motivational level, personality, and strengths (Andrews & Dowden, 2007).

Application of the RNR model requires an instrument that can effectively identify offenders who are more likely to recidivate. Developed by Andrews and Bonta in the 1970s, the Level of Supervision[1] Inventory-Revised (LSI-R) measures both static and dynamic risk factors during a semistructured interview (Andrews & Bonta, 2003). The assessment consists of 54 items across 10 subscales (criminal history, education/employment, financial, accommodation, family/marital, leisure/recreation, companions, alcohol/drug problems, emotional/personal and attitude/orientation). Studies show that the LSI-R demonstrates adequate internal consistency (Andrews & Bonta, 2003; Arens, Durham, O'Keefe, Klebe, & Olene, 1996) and acceptable levels of interrater reliability (Andrews, 1982; Austin, Coleman, Peyton, & Johnson, 2003; Lowenkamp, Holsinger, Brusman-Lovins, & Latessa, 2004). Most research on the LSI-R in Canada and the U.S. also supports its predictive validity—how accurately the instrument predicts behaviors believed to be associated with criminal risk (Andrews, 1982; Andrews & Bonta, 2003; O'Keefe, Klebe, & Hromas, 1998). However, results seem to vary depending on the population (O'Keefe et al., 1998). A 2002 study (Dowdy, Lacy, & Unnithan) of 140 male offenders found that the LSI-R was unable to predict outcomes for the community corrections population, but Flores, Lowenkamp, Holsinger, and Latessa (2006) concluded that the LSI-R *was* a significant predictor of reincarceration for a sample of federal probationers. Debate also continues about how well the measure performs with minority populations and women, although there is adequate evidence that the LSI-R outperforms other measures of risk assessment and functions equally well among all offenders (Weber, DeLaCerda, & O'Keefe, 2010).

Addiction Severity Assessment

In 2006, NIDA released a report titled *Principles of Drug Abuse Treatment for Criminal Justice Populations,* which recognized that drug addiction is a brain disease requiring effective long-term treatment, management, and monitoring. The Office of National Drug Control Policy estimated that, in 2002, the societal cost of drug abuse was $180.9 billion, $107.8 billion of which was associated with drug-related crime. The cost of treating drug abuse was estimated to be a fraction of the toll on society: $15.8 billion (NIDA, 2006). Comprehensive assessment was listed as the first step, to determine the nature and extent of the problem, as well as co-existing issues including mental disorders, and to formulate an appropriate treatment plan.

The first national survey of its kind, the National Criminal Justice Treatment Practices Survey (NCJTPS) was conducted to examine substance abuse programs and services in all correctional settings (prisons, jails, probation and parole offices, and other local community correctional agencies; Taxman, 2007). The survey was part of NIDA's Criminal Justice-Drug Abuse Treatment Studies (CJ-DATS), a 10-center network studying substance abuse treatment services for adult and juvenile offenders. Less than half of the administrators in the survey reported using a standardized substance abuse screening tool or an actuarial risk tool. Among a stratified sample (from the NCJTPS data) of 289 administrators at adult facilities (prisons, jails, and community correctional agencies),

[1] Originally called the Level of Supervision Inventory, the test authors sold the tool to Multi-Health Systems, who renamed the tool the Level of Service Inventory (both tools have since been revised). Currently the CDOC uses the Level of Supervision Inventory-Revised, as permissions for use were purchased (and subsequently renewed) directly from the authors, not through Multi-Health Systems.

58.2% used a standardized substance abuse screening tool and 34.2% used an actuarial risk tool (Taxman, Cropsey, Young, & Wexler, 2007). The use of a standardized substance abuse screening tool was associated with the availability of higher intensity treatment programs and community referral services and the use of standardized risk tools. The most frequently used screening and assessment tools (with facilities able to report more than one) were the Addiction Severity Index (ASI, 46.4%), the Substance Abuse Subtle Screening Inventory (SASSI, 42.3%), the Texas Christian University Drug Screen (TCUDS-II, 22%), the Michigan Alcohol Screening Tool (MAST, 20.8%), the Drug Abuse Screening Tool (DAST, 17.9%), and the Alcohol Dependence Scale (ADS, 14.9%); 29.8% used some other screening tool.

In Colorado, the Adult Substance Use Survey (ASUS) was created in 1992 for use within the state as a measure of addiction severity (Arens et al., 1996). The 96-question self-report instrument, now the ASUS-Revised (ASUS-R), consists of 15 basic scales and three supplemental scales. Scales measure patterns of substance use, alcohol and other drug (AOD) benefits, social and legal nonconformity, mental health concerns, defensiveness, motivation to change, and perceived strengths. An evaluator summary also compares the assessor's perception of drug use (from minimal to high) with that of the offender (Wanberg, 2009). Extensive research on the ASUS and ASUS-R among more than 40,000 respondents has demonstrated that the measure has good construct validity (criterion, concurrent, and predictive), optimal internal consistency, and scale independence (Wanberg, 2006, 2009). In addition, ASUS-R scales and external criterion variables are strongly correlated, and the measure is able to distinguish various levels of substance abuse severity and mental health and antisocial problems (Wanberg, 2009). The ASUS disruption subcomponent alone likewise has shown strong internal consistency based on Cronbach's alpha (.94) and convergent validity based on significant correlations with both the ADS and the DAST (Arens et al., 1996).

Colorado also uses the Simple Screening Instrument-Revised (SSI-R; Winters & Zenilman, 1994) as a substance use screening measure. The SSI-R spans a wide range of signs and symptoms of substance abuse based on the biopsychosocial view endorsed by the World Health Organization and the American Psychiatric Association. It consists of 16 items, most taken from existing screening tools, and measures five domains: substance consumption, preoccupation and loss of control, adverse consequences, problem recognition, and tolerance and withdrawal. Fourteen of the items are scored, and a score of 4 or higher is considered the cutoff point for a full substance abuse assessment. The SSI-R can be self-administered or used in an interview format. Research supports the sensitivity and test-retest reliability of the SSI among populations including inmates (Peters et al., 2000) and adolescent medical patients (Knight, Goodman, Pulerwitz, & DuRant, 2000), although it was not as effective as a screening instrument among college students (Kills Small, Simons, & Stricherz, 2007). The Peters et al. study (2000) found that the SSI demonstrated high sensitivity (92.6%) in detecting alcohol or drug dependence and high test-retest reliability (.97), and Knight et al. (2000) reported good internal consistency (.83) as well as excellent test-retest reliability (ranging from .82 to .90).

Treatment Matching

In addition to assessment, C.R.S. 16-11.5-102 mandates education and treatment for substance-abusing offenders. The RNR model suggests that matching offenders to the appropriate level and type of treatment may improve outcomes (Melnick, De Leon, Thomas, & Kressel, 2001; Simpson, Joe, Fletcher, Hubbard, & Anglin, 1999; Thanner & Taxman, 2003). Despite the intuitive appeal of this strategy, however, studies have not always been convincing (Kazdin, 2008; McKay, Cacciola, McLellan, Alterman, & Wirtz, 1997; McLellan & Alterman, 1991; Project MATCH Research Group, 1998a, 1998b).

At least 25 years before matching schemas gained a strong foothold among addiction treatment providers, Paul (1967) summed up the issue succinctly: "*What* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of circumstances?" (p. 111). Driven by a health care climate that demanded more efficient, cost-effective treatment methods, agencies in the late 1980s began to move away from the historical one-size-fits-all Minnesota Model: the 28-day inpatient program based largely on the 12-step Alcoholics Anonymous approach (Belenko & Peugh, 2005; Gartner & Mee-Lee, 1995; Gastfriend & McLellan, 1997). The goal was to place patients in the least expensive, least restrictive treatment environment most likely to produce a positive result. In 1991, the American Society of Addiction Medicine (ASAM) introduced the first edition of its Patient Placement Criteria (PPC), based on the Cleveland Criteria and input from the National Association of Addiction Treatment Providers (NAATP; Belenko & Peugh, 2005). As later modified, the ASAM criteria now outline five levels of service (early intervention, outpatient treatment, intensive outpatient and partial hospitalization, residential/inpatient treatment, and medically managed intensive inpatient treatment) determined by six dimensions of biopsychosocial needs (acute intoxication and/or withdrawal potential; biomedical conditions and complications; emotional, behavioral, or cognitive conditions and complications; readiness to change; relapse/continued use, continued problem potential; and recovery environment; Mee-Lee, 2005). Unfortunately, like the guidelines upon which they were based (Cleveland Criteria and NAATP), the ASAM criteria were not empirically validated before they were published and distributed (Gastfriend & McLellan, 1997). In an evaluation of the predictive validity of the psychosocial dimensions of the ASAM criteria, researchers found that cocaine- and alcohol-dependent patients correctly matched to inpatient or intensive outpatient programs did not show significantly more improvement than mismatched patients, although the authors cautioned against generalizing their results (McKay et al., 1997). However, Gastfriend, Lu, and Sharon (2000) cited more encouraging studies in support of placement matching based on the ASAM criteria while noting some of the challenges (e.g., patient acceptance of treatment recommendations, availability of services, and the needs of special populations). The authors distinguished *placement matching* (setting and resource intensity, such as long-term residential treatment) from *modality matching* (the theoretical model, such as cognitive-behavioral therapy), with initial studies providing limited justification for the latter (Project MATCH Research Group, 1998a, 1998b; see discussion below). Further, Magura et al. (2003) compared actual care received by alcoholic patients to the level of care recommended by a computerized algorithm and clinicians' evaluations based on the ASAM PPC. Results showed that certain patients who were

undertreated (the actual treatment level was lower than the recommended level) had poorer outcomes and overtreatment (the actual treatment was higher than the recommended level) did not result in improvement, thus lending support for using the criteria as a matching schema. Mismatching according to the ASAM PPC also promoted treatment no-shows in psychiatric patients with comorbid substance use disorders (Angarita et al., 2007). To help interviewers navigate the complex criteria, Turner, Turner, Reif, Gutowski, and Gastfriend (1999) developed a computerized version of the PPC as a means of addressing issues of interrater reliability.

Beyond the ASAM criteria, the largest and most notable early study to test the matching hypothesis was Project MATCH, a multisite study funded by the National Institute on Alcohol Abuse and Alcoholism. Patients were matched to 12-step therapy, motivational enhancement therapy, or cognitive-behavioral therapy based on more than 20 attributes. Disappointingly, only four client characteristics produced statistically significant matches: psychiatric severity, anger and support for drinking (among one's social network) in outpatients, and alcohol dependence in aftercare patients. Researchers were forced to conclude that "matching as an overall treatment strategy did not substantially enhance outcomes" (p. 589; Project MATCH Research Group, 1998b). Commentators followed up by suggesting, however, that unexpected findings in matching studies like Project MATCH and others were due to methodological flaws (Bühringer, 2006; Moyer, Finney, Elworth, & Kraemer, 2001). A later re-examination of Project MATCH data also tempered the earlier conclusion by providing insight on factors affecting treatment matching: Outpatient clients with low motivation, males with less severe alcohol dependence, and females in aftercare showed more improvement from motivational enhancement therapy compared to cognitive-behavioral approaches (Witkiewitz, Hartzler, & Donovan, 2010). Others have commented that Project MATCH incorporated treatment approaches not specifically targeted to clients' needs and characteristics and, therefore, results were unsurprising (Heather, 1999; Jolley & Kerbs, 2010).

To paraphrase Gastfriend and McLellan (1997) with an old cliché, "different strokes for different folks:" no program works for everyone all the time, but particular programs may work for particular clients at particular times. In asking why drug abuse treatment matching has been so difficult, Ball (1994) suggested that it may be a changing process influenced by innumerable variables rather than a static, one-time event. Findings from a number of studies generally indicate that clients with higher levels of drug-use severity benefit from longer, more intensive, and more structured treatment (Budde, Rounsaville, & Bryant, 1992; McKay et al., 2002; Rychtarik et al., 2000; Simpson et al., 1999; Thanner & Taxman, 2003; Thornton, Gottheil, Weinstein, & Kerachsky, 1998; Tiet, Ilgen, Byrnes, Harris, & Finney, 2007). Such a statement may be an oversimplification of the treatment matching issue, however, and studies have sometimes been small or not entirely consistent (Budde et al., 1992; McKay et al., 2002; Thornton et al., 1998). Multiple client characteristics and needs must be matched with a wide range of services and programs in a multidimensional approach (Gastfriend & McLellan, 1997). Mattson et al. (1994) found 31 empirical studies that demonstrated interactions between treatment options and four categories of client variables: demographic factors, alcohol-specific characteristics (e.g., amount and duration of alcohol consumption), intrapersonal characteristics (e.g., cognitive status and emotional states), and interpersonal functioning (e.g., social

support). In a review of substance abuse treatment studies with 2-year follow-ups or more, McKay and Weiss (2001) found that psychiatric severity was significantly related to outcomes. In addition, positive long-term outcomes were associated with relapse-prevention behaviors such as better performance during treatment; more self-help participation; lower social/family and legal problems; and fewer treatment readmissions, less life stress, and positive coping responses after treatment. In a study exploring the matching of clients' needs with drug treatment, requested services providing vocational training, child care, transportation, and housing were associated with longer treatment retention (Hser, Polinsky, Maglione, & Anglin, 1999). Smith and Marsh (2002) obtained similar results when considering client-service matching for women with children: The provision of counseling services was related to less substance abuse, and matched housing, job training, and legal services predicted treatment satisfaction. Other research considering client characteristics such as age, gender, substance abuse profile, psychological status, and motivation for change lends further support to the matching hypothesis (Brown, Seraganian, Tremblay, & Annis, 2002; Greenfield et al., 2007; Magura et al., 2005). Jolley and Kerbs (2010) urged correctional agencies to emphasize "demand-side programming" based on offenders' needs over "supply-side programming" driven by organizational factors (p. 291).

The logic of finding the best client-treatment fit to maximize benefits within the constraints of available resources continues to be appealing, if elusive. In its *Principles of Drug Addiction Treatment: A Research-Based Guide*, NIDA (2009) listed the need to match individuals to specific settings, interventions, and services as well as factors beyond drug abuse (medical, psychological, social, vocational, legal, age, gender, ethnicity, and culture). Thus, researchers have begun to move beyond assessment of substance-use severity to develop more comprehensive biopsychosocial treatment-matching frameworks. As Latessa, Cullen, and Gendreau (2002) observed, without adequate risk and needs assessments, "offenders in effect enter a treatment lottery in which their access to effective intervention is a chancy proposition" (p. 48). Among emerging instruments, De Leon, Melnick, and Cleland's (2008) Client Matching Protocol (CMP) employed a decision tree based on four domains (pattern of use, previous long abstinence, social factors, and habilitation) and a severity-intensity paradigm. That paradigm supports placing offenders in the least-intensive programs needed to achieve positive outcomes, reflecting studies that show undertreatment (lower intensity) is detrimental to high-risk individuals but overtreatment (higher intensity) is not harmful to those in low-risk categories (De Leon et al., 2008). Validity studies on the CMP, using data from the Drug Abuse Treatment Outcome Studies (DATOS), indicated more favorable results for matched and overtreated clients compared to undertreated participants; however, the fact that clients were allowed to self-match (choose their own treatment regardless of the assessed level) limited conclusions (De Leon et al., 2008; De Leon, Melnick, & Cleland, 2010). Belenko and Peugh (2005) have developed treatment-matching guidelines by analyzing data from the 1997 Survey of Inmates in State Correctional Facilities (14,284 inmates from 275 state prisons), including recency and severity of drug use, drug-related behavioral consequences, and other social and health problems. The researchers estimated that, defined broadly, 82% of state prison inmates are involved with drugs or alcohol but correctional agencies cannot meet the need for treatment, particularly for residential programs (needed by one-third of male inmates and more than half of female inmates). Belenko and

Peugh (2005) classified treatment needs into four levels: no treatment, short-term intervention (e.g., drug education), outpatient treatment (e.g., group counseling), and residential treatment (e.g., long-term treatment in separate housing). These levels are based on four assumptions: (a) "any inmate in the most severe drug use category… should receive residential treatment," (b) "any inmate who has ever used non-marijuana illegal drugs should, at a minimum, receive outpatient treatment," (c) "having multiple other social or health problems implies a need for more intensive treatment than would otherwise be suggested by drug use pattern alone," and (d) "having three or more drug-related consequences generally should move an inmate up one level of treatment intensity" (p. 273). In a later study of released inmates, Belenko (2006) advocated new, psychometrically sound assessment instruments that span multiple dimensions, including public safety risk; are easy to administer and score; are adaptable to real-world settings; assess both static (stable) and dynamic (changing) factors; and improve links to services.

Standardized Offender Assessment

Belenko's (2006) recommendation encapsulates the goal of the Standardized Offender Assessment (SOA) in Colorado. The original SOA yielded a treatment recommendation based on the Substance Use History Matrix (SUHM), a risk/needs algorithm that measured risk with the LSI total score and addiction severity with the disruption subscale of the ASUS (Tapia, 2010). The SUHM also included substance use history information such as age at first use, duration of use, frequency, intensity, method of ingestion, social context and setting, and longest reported abstinence. Problems with the SOA included false positives, false negatives, and frequent overrides of the assessed treatment level (more than 60% of the time). There were at least 18 clinical reasons for overrides, ranging from acute intoxication, significant denial, and tolerance to lack of a substance-free social network, long-term abstinence, and successful treatment completion. Overrides were the rule rather than the exception. In addition, the former Alcohol and Drug Abuse Division (ADAD; now part of the Division of Behavioral Health, Colorado Department of Human Services) required treatment providers to use the ASAM PPC (modified for offenders) to admit clients into treatment. Of these criteria—biomedical condition, emotional/behavioral or cognitive condition, relapse/recidivism risk, readiness to change, and recovery environment—only one (relapse/recidivism risk) was congruent with the SOA (Mee-Lee, 2005; Tapia, 2010). To address these issues, changes were made to the SOA. The goals of the SOA-Revised (SOA-R), introduced in 2006, were to incorporate other ASUS scales and offender information into the treatment referral decision, link to the new ADAD treatment level (enhanced outpatient), achieve congruence with ASAM domains, address false positives and negatives, and mitigate overrides of the assessed treatment level. The SOA-R kept the LSI-R and a revised ASUS, modified and incorporated the ASAM criteria, and replaced the SUHM with the Treatment Recommendation Worksheet (TxRW; Tapia, 2010).

The current SOA-R consists of four instruments: the LSI-R as a recidivism risk measure, the SSI-R as a substance abuse screening measure, the ASUS-R as an addiction severity measure, and the TxRW to derive a treatment level. The TxRW combines the LSI-R total score, ASUS-R involvement and disruption scores, substance use patterns, previous treatment and supervision performance, and the modified ASAM dimensions (mentioned above). The TxRW requires the use of structured

clinical judgment, by gathering information from the actuarial assessments to reach a treatment decision, and asks assessors to document reasons why a treatment referral might differ from the actual treatment recommendation.

In conjunction with the SOA-R, an eight-level treatment system was formulated and tied to the assessment battery. Scores on the standardized assessment drive placement into one of the treatment levels, all of which provide education and therapy of varying intensity and duration. (See the Appendix for a complete description.)

SOA-R Evaluation

The IACAJCT requested an evaluation of the SOA-R in 2006, specifically targeting the treatment recommendation produced by the TxRW. The year-long study examined the interrater reliability of the SOA-R, or how often different assessors recommended similar treatment for the same offenders (Colorado Department of Human Services, 2007). The evaluation focused on eight fictitious offender profiles (three of which included videotaped interviews) rated by 173 assessors from three agencies (DOC, the Division of Criminal Justice, and the Department of Probation). All assessors completed SOA-R training but did not represent a random sample. The study's author justified the use of a large number of raters as providing greater ecological (real-world) validity, while noting that the limited number of cases was a limitation (Colorado Department of Human Services, 2007).

The resulting report concluded that the SOA-R had unacceptable interrater reliability, included too many factorial combinations for assessors to consider (900 trillion), failed to meet accepted criteria for standardization, and used clinical judgment rather than an algorithm to reach a treatment decision (Colorado Department of Human Services, 2007; Tapia, 2010). The evaluation recommended that the SOA-R be replaced by an evidence-based, standardized, treatment-matching measure. In response, the Offender Treatment Matching Algorithm (OTMA) was developed in 2009 as a possible replacement (Melnick, Schoeneberger, & Sacks, 2009).

A modified version of the CMP (De Leon et al., 2008), the OTMA utilizes a decision tree with five assessment points based on the severity of current substance abuse, determined from the degree of alcohol and/or substance abuse during the previous 6 months, as well as the consequences; social factors related to relapse and recidivism risk; and habilitation (the ability to earn a living; Melnick et al., 2009). A previous study showed that clients matched to long-term residential treatment with the CMP algorithm had successful 1-year outcomes, based on reduced recidivism and substance abuse, compared to undertreated clients in outpatient treatment (De Leon et al., 2008).

The OTMA is structured so that clients with low alcohol/drug use frequency and consequences scores are matched to no treatment, alcohol/drug education, or weekly outpatient treatment. Clients with higher scores in those areas move on to the next level of assessment, which includes social factors, habilitation, and an LSI-R score; they are matched to enhanced outpatient therapy (EOP) or intensive outpatient therapy (IOP), intensive residential treatment (IRT), or a therapeutic community (TC). The OTMA also includes mitigating circumstances, such as previous treatment, responsibilities

to dependents, and criminal history limitations, which may factor into a treatment referral. The OTMA takes about 15 minutes to administer, compared to about 90 minutes for the entire SOA-R. (Note that if the OTMA were to replace the ASUS-R and the TxRW, the LSI-R would still be administered, adding another 45 to 60 minutes.)

A pilot study was conducted as a preliminary test of the OTMA's validity (Melnick et al., 2009). Forty-four staff members from four treatment agencies administered the OTMA and the SOA-R to 185 new admissions. Findings showed that the OTMA was more consistent with the results of the SSI-R than the SOA-R: The OTMA identified 21.3% of offenders as having a substance use problem, compared to 27% by the SSI-R and 1.1% by the SOA-R. The OTMA also was compared to actual 2007 Colorado treatment referrals as well as treatment levels assessed by the SOA-R. A simplified matching schema condensed to only four treatment levels (drug education, outpatient, intensive outpatient, and residential treatment) was used for the comparison. Data revealed that the OTMA's treatment referrals more closely matched actual DOC referrals rather than the SOA-R assessments: The average OTMA deviation was 6.8% versus 24.9% for the SOA-R. Although both the OTMA and the SOA-R underreferred offenders to higher treatment levels, the OTMA was considered more accurate--more likely to make referrals to adjacent treatment levels (enhanced or intensive outpatient) while the SOA-R's deviation involved two different levels (weekly and intensive outpatient; Melnick et al., 2009).

Based on the findings of the pilot study, the OTMA was revised to increase the likelihood of referring clients to enhanced outpatient treatment, intensive outpatient treatment, or a therapeutic community, increased the weighting of criminogenic influences in social factors, and added the risk of HIV and violence to the consequences of substance use. Developers also reduced the complexity of the mitigating circumstances and removed some items to make the instrument less confusing (Melnick et al., 2009). A second pilot study using the revised OTMA, which produced 74 ratings, provided preliminary indications of the tool's interrater reliability, with an average agreement rate of 80% (Melnick et al., 2009).

Current Study

The current study sought to further test the reliability of the OTMA and the ASUS-R and to assess the convergent and discriminant validity of both the OTMA and the ASUS-R as part of a second revision of the SOA-R. The LSI-R was not reevaluated, as its psychometric properties have been validated elsewhere (Andrews & Bonta, 2003). Other issues considered in selecting an instrument were administration time, ease of use, cost, training, and staffing. Further validation in future studies will consider the issue of treatment matching and whether that process is correlated with reductions in substance abuse and recidivism.

Method

Participants

Colorado male and female adult inmates, parolees, and probationers assessed for substance abuse treatment needs were the target population for this study. Offenders were recruited for participation during the intake process to the CDOC (Denver Reception and Diagnostic Center; DRDC), probation, community corrections, and parole (Treatment Accountability for Safer Communities; TASC) in the Denver/Colorado Springs area. Data for some of the assessments were missing for 30 participants, reducing the initial sample of 268 to 238. Of the total participants, 130 were DOC inmates and parolees, 57 were probationers, and 51 were enrolled in community corrections programs. There were 181 males and 57 females in the sample. Ages ranged from 19 to 55 years old (M = 32.8, SD = 9.2); 36% were Caucasian/Hispanic, 47% Caucasian/Non-Hispanic, 12% African American, and 5% other (e.g., Native American and Asian). Nine participants did not list race. The average LSI-R total score was 29.4 (SD = 8.3).

Materials

Data were gathered using self-administered and structured interview assessments; a coversheet was also completed detailing the participant's demographic, criminal history, and criminal risk information. The LSI-R and the ASUS-R are included in the SOA-R during the routine intake process. As part of the research protocol, the OTMA and the Global Appraisal of Individual Needs-Quick Core (GAIN-Q Core) also were administered. Although most of these instruments have been described earlier, a brief summary is provided below:

LSI-R. The LSI-R measures an offender's level of recidivism risk by assessing both static and dynamic risk factors during a semistructured interview. The assessment consists of 54 items across 10 subscales (criminal history, education/employment, financial, accommodation, family/marital, leisure/recreation, companions, alcohol/drug problems, emotional/person, and attitude/orientation), each of which has items that are scored as a 1 or 0 to indicate the presence or absence of the risk factor. With a total score range of 0 to 54, higher scores represent a greater likelihood of recidivating. The total score is divided into three bands of low-, medium-, and high-risk categories, although with slightly different score ranges set by different Colorado agencies. Time frames also influence how an item is scored, and 13 rater boxes allow assessors to provide additional information about the severity of a risk factor (Weber, DeLaCerda, & O'Keefe, 2010).

ASUS-R. The ASUS-R is a 96-question self-administered measure of addiction severity. It consists of 15 basic scales and three supplemental scales. Scales measure patterns of substance use, alcohol and other drug benefits, social and legal nonconformity, mental health concerns, defensiveness, motivation to change, and perceived strengths. An evaluator summary also compares the assessor's perception of drug use (from minimal to high) with that of the offender (ASUS-R, 2009).

OTMA. The OTMA is an interview-based treatment-matching measure that utilizes a decision tree with question sequences determined by conditional branching (i.e., the answer to a particular question determines whether the next question is asked or whether the interviewer skips ahead to subsequent sections). There are five assessment points and seven scales based on the severity of current substance abuse, determined from the degree of alcohol and/or substance abuse during the previous 6 months, as well as the consequences; social factors related to relapse and recidivism risk; and habilitation (the ability to earn a living; Melnick et al., 2009). Scales A, B, and C sum to produce an overall alcohol and drug frequency (use) score in scale D; scale E provides a consequences score; and scales F and G result in social indicators and habilitation scores, respectively. The OTMA is structured so that clients with low alcohol/drug use frequency and consequences scores are matched to no treatment, alcohol/drug education, or weekly outpatient treatment. Clients with higher scores in those areas move on to the next level of assessment, which includes social factors, habilitation, and an LSI-R score; they are matched to enhanced outpatient therapy (EOP) or intensive outpatient therapy (IOP), intensive residential treatment (IRT), or a therapeutic community (TC). The OTMA also includes mitigating circumstances, such as previous treatment, responsibilities to dependents, and criminal history limitations, which may factor into a treatment referral.

GAIN-Q Core. The GAIN-Q Core is a 45-question biopsychosocial assessment with eight sections (background, general factors, sources of stress, physical health, emotional health, behavioral health, substance-related issues, and service utilization; Chestnut Health Systems, 2002). Most questions are answered with yes/no responses, with the exception of those asking for time frames. Although the GAIN-Q can be given by an interviewer, it was self-administered for purposes of the current study.

The GAIN-Q was chosen as the comparison measure to test the convergent validity of the OTMA and ASUS-R. Research has shown good test/retest reliability and diagnostic consistency for the GAIN-Q among adults and adolescents. In addition to its sound psychometrics, the GAIN-Q was selected because of its compatibility with the *DSM-IV* and the ASAM PPC-2R and its widespread use by more than 700 agencies in the U.S., Canada, Mexico, England, and other countries (Chestnut Health Systems, 2002).

Procedure

Data were collected from February through the first week in May 2011. The research protocol was approved by the institutional review board at Ethical and Independent Review Services, Corte Madera, California.

At the start of the intake process at each location, a case manager, assessor, or research staff member gave the offender a copy of the informed consent document to read and verbally provided an explanation of the study as well. If there was any concern regarding the offender's literacy skills, the consent form was read aloud. Staff also discussed mandatory reporting requirements (intent to harm self or others, child/elder abuse, or other undisclosed criminal activity), the voluntary nature of

participation, and options for withdrawal. All assessments were coded with a participant identification number to ensure confidentiality.

The data collection procedure was similar for each agency. Upon arrival at DRDC, a community corrections center, probation office, or parole agency (TASC), offenders completed routine intake paperwork. During this initial intake meeting, the offender was provided with a description of the research study and given an opportunity to participate. If the offender declined to participate, he/she continued with the intake process and was not given any of the additional study assessments.

If the offender was willing to participate, he/she was asked to sign the informed consent document. Participants were advised that each of the research instruments would require 15 to 20 minutes to complete, adding another 30 to 40 minutes to the standard intake process. Then the study assessments (OTMA and GAIN-Q together) were administered either before or after the standard intake process was complete, in alternating order. The order of the OTMA and GAIN-Q also was rotated to ensure that some participants completed the GAIN-Q first and others the OTMA first. Assessors informed participants when the research assessments were beginning and ending to avoid confusion with the standard intake process. When the OTMA was administered to offenders, a second independent rater joined the primary rater at each agency to record offenders' responses to determine the interrater reliability of the OTMA. Also, for data analysis purposes, rather than following the conditional branching in the measure, the primary assessor asked all questions and both assessors recorded all answers on the OTMA. For example, under typical circumstances, if an offender answered that he/she never drank beer and/or wine during the previous 6 months (question 5), the assessor would skip to question 9 related to hard liquor. In the current study, the assessor asked questions 6, 7, and 8 concerning frequency of beer and/or wine consumption, regardless of the initial answer. Additionally, assessors did not score the assessment subsections to derive a treatment recommendation because the issue of treatment matching will be evaluated in a later study. The scoring for both the OTMA and the GAIN-Q and the derivation of the OTMA treatment recommendation were completed at the time of analysis.

Following administration of the assessments and the intake process, the assessor completed the data collection cover sheet, including the participant's gender, age, age at first arrest, race/ethnicity, LSI-R subscale and total score, SSI-R total score, and assessment treatment level to ensure a representative sample. Responses were coded and aggregated to protect the identity of the participants. A research coordinator at each agency (CDOC/parole, probation, and DCJ-community corrections) was responsible for collecting the informed consent forms and assessment data packages. Upon receipt of the data, each research coordinator entered participant names and code numbers into a master list database, secured on a password-protected computer and accessible only by project researchers. Data were then forwarded to the data analysis firm.

Data analysis plan. In this study, the interrater item-level reliability of the OTMA, the internal consistency of the OTMA and the ASUS-R, and the convergent/discriminant validity of the OTMA and the ASUS-R were examined. The responses for the OTMA were concurrently collected by two

interviewers. The responses for the ASUS-R were provided directly by participants on a paper-and-pencil form. Hence, interrater reliability was examined only for the OTMA but internal consistency was examined for all scales on the OTMA and the ASUS-R.

OTMA interrater reliability. An item-level interrater reliability analysis of the OTMA was conducted using the intraclass correlation coefficient (ICC) from the one-way ANOVA model (Shrout & Fleiss, 1979). The ICC is a measure of agreement between different raters and can be used to estimate reliability on item-level data (Dunn, 2004; Fleiss, 1975). Intraclass correlations were computed for all the scale items in the OTMA. Missing data were omitted using a listwise deletion strategy (i.e., only cases with complete data were analyzed). The criterion for good item-level reliability between raters was set at .80 (Anastasi & Urbina, 1998).

Power and sample size calculations were conducted to examine the precision of the ICC estimates in the study, given the current design. Although the sample was smaller than anticipated, it was large enough to obtain ICC estimates with good precision. For example, for an ICC value of .70, the 95% confidence interval would have a precision within 10% (.63, .77).

Internal consistency of the OTMA and ASUS-R scales. The internal consistency measure of reliability refers to the degree to which items within each scale correlate with one another. High internal consistency indicates that the scale is reliable. Cronbach's alpha was calculated for each scale, using an alpha of .70 or greater (on a scale of 0 to 1) to indicate a high level of internal consistency. An alpha above .95 suggests colinearity or a high degree of redundancy in the items. Low alphas indicate the scale has too few items or the items do not have much in common (Nunnally & Bernstein, 1994). Generally, if the items on a scale are highly correlated (internally consistent), then the summated scale will be reliable.

Scale validity. To assess validity, the multitrait-multimethod matrix (Campbell & Fiske, 1959) was used to examine the pattern of correlations between the ASUS-R involvement and disruption scales and the OTMA use and consequences scales. Those scales were chosen because of their correspondence to the *DSM-IV*-specified substance use/abuse and substance dependence criteria, which provided the theoretical foundation for the validity aspect of the study. ASUS-R involvement and OTMA use were equated with substance use/abuse and ASUS-R disruption and OTMA consequences reflected substance dependence. The use and dependence scales from the GAIN-Q were used for comparison to assess convergent validity. The attitude and accommodation scales from the LSI-R were included to assess discriminant validity, as they were assumed to be unrelated to substance use/abuse and dependence. When computing correlations between ordinal and other scales, the appropriate correlation coefficients (polyserial, polychoric, etc.) were used to approximate the Pearson correlations that would be obtained if the data were continuous.

Convergent validity. The convergent validity of the ASUS-R and OTMA scales were assessed by examining the degree to which each scale correlated with the corresponding GAIN-Q scales. These correlations are referred to as convergent validity coefficients. Following recommendations of Hemphill (2003), correlations less than .20 were considered low, correlations between .20 and .30 were considered medium, and correlations above .30 were considered high. Large convergent validity coefficients indicate convergence onto the same traits measured by the GAIN-Q scales.

Discriminant validity. The discriminant validity of the ASUS-R and OTMA scales was assessed by examining their correlations with the LSI-R scales. Large correlations indicate that the scales are unable to discriminate unrelated traits, hence, values less than .20 were considered to be ideal (Hemphill, 2003). Other patterns in the correlation matrix also provide evidence of discriminant validity. These patterns are described in the Results section and involved testing whether there was a significant difference between specific pairs of correlations in the matrix. These tests of differences between correlation coefficients were based on an approach developed by Steiger (1980) for comparing two dependent correlations. Two correlations being tested are considered "dependent" when both are based on scores from individuals in the same sample.

Results

The main goal of the project was to evaluate the reliability and validity of the scales from two instruments (ASUS-R and OTMA) that assess two traits: substance use/abuse and substance dependence. Reliability is a reflection of consistency across responses on items or by raters. Validity is whether a scale measures what it is intended to measure. One should note that the degree of reliability determines the potential for validity. Another goal of the project was to examine all the patterns of relationships across all scales to determine whether these patterns were as one would expect. The following provides a summary of the technical results section that may be found in Appendix B.

Reliability

Reliability is an essential element of measurement. If a scale is unreliable, the scale scores are prone to error, such that an individual's score may not accurately reflect his or her true level of a given trait, preventing comparisons with individuals who score high versus low on the trait. Scale reliability (internal consistency) is commonly assessed with a statistic called Cronbach's alpha. The Cronbach's alphas for the ASUS-R scales that measured the traits use/abuse and dependence were .84 and .94, respectively. The corresponding Cronbach's alphas for the OTMA scales were .75 and .85. Hence, the reliabilities were very good.

Interrater reliability, or the degree to which responses on an instrument are consistent between two assessors, also was tested for the OTMA using intraclass correlation coefficients (ICCs). (The ASUS-R was not tested for interrater reliability because it is a self-administered assessment.) Overall, OTMA items demonstrated excellent interrater reliability, with ICCs ranging from .82 to 1.00, a result that was expected given the instrument's highly structured design. In addition, computer-calculated OTMA treatment level assignments indicated good agreement between the raters, with only 10% discordance.

Validity

Having established that the scales were reliable, analyses were conducted to determine whether the instruments measured the intended traits. One aspect of validity focuses on the accuracy in the measurement of the theoretical concepts of interest (substance use/abuse and dependence). This is described as the construct validity of the scale. This analysis attempted to assess two aspects of construct validity: *convergent* and *discriminant* validity.

Convergent validity. If a scale correlates highly with other scales that are believed to measure the same trait, then it *converges*, or has high convergent validity. Data were collected on two subscales from the GAIN-Q instruments that were believed to be good measures of the substance use/abuse and dependence traits. For example, a person who reported drinking a great deal each day would have a high score on the GAIN-Q use scale. If that person reported that his daily life was greatly

disrupted as a consequence of his drinking, he would have a high score on the GAIN-Q dependence scale. Convergent validity of the ASUS-R and OTMA scales was assessed by determining if the subscales on these measures correlated with the corresponding GAIN-Q subscales.

The correlations of the two ASUS-R subscales (involvement and disruption) with the two GAIN-Q subscales were .39 and .48 for the use/abuse and dependence traits, respectively. Likewise, the corresponding correlations for the two OTMA scales (use and consequences) with the GAIN-Q scales were .44 and .57. These correlations reflect high convergent validity.

Discriminant validity. A valid scale should *not* correlate with traits that are unrelated to the trait being measured. When a scale correlates with unrelated traits, it fails to *discriminate*, and the resulting measurements cannot be unambiguously associated with the trait that one is trying to measure. To assess discriminant validity, one looks at the correlation between one scale and other scales that are believed to measure unrelated traits.

Two scales from the LSI-R—attitude and accommodation—were collected to assess the discriminant validity of the ASUS-R and OTMA scales. The LSI-R scales were presumed to measure traits unrelated to the use/abuse and dependence traits.

The correlations between the two ASUS-R subscales and the two LSI-R subscales ranged in value from .08 to .20. The correlations between the OTMA subscales and the same LSI-R subscales ranged from .14 to .29. Several values in these sets of correlations were undesirably large, especially for the OTMA scale. Other patterns in the correlations between the scales are also of interest when assessing discriminant validity. However, these were more difficult to interpret, as high correlations between the substance use and dependence scales suggested those traits may not be distinct. (See Appendix B for a more complete discussion.)

Conclusion

Results for the ASUS-R and the OTMA were similar, with the ASUS-R performing slightly better than the OTMA. Regarding relative performance on the characteristic of scale reliability, the ASUS-R slightly exceeded the OTMA. The answers to items for the OTMA scale were collected by interviewers, whereas the items for the ASUS-R were directly answered by offenders on a paper form. The use of interviewers theoretically could have reduced the reliability of the OTMA scales due to data recording errors or differences in the interpretation of the offenders' answers. However, item-level data were consistent across the offender interviews by the two different raters in the different agency settings. This result indicated that errors made by the interviewers did not contribute to inconsistencies in the measures.

On convergent validity, the OTMA's correlations were better, but not significantly so. Comparisons with two LSI-R subscales provided evidence of discriminant validity for both measures, although the ASUS-R was slightly better.

Discussion

As measures of substance abuse severity, both the ASUS-R and the OTMA were found to have favorable psychometric properties in the evaluated areas, the ASUS-R having only a slightly better performance overall. The internal consistency among items was very good for both assessments. Interrater reliability of the OTMA was good, as expected, due to the highly structured nature of the questions. The ASUS-R was not tested for interrater reliability because it is a self-administered assessment. Validity tests also revealed high convergent validity, indicating the OTMA and the ASUS-R were measuring the same concepts as the GAIN-Q (i.e., substance abuse and dependence). Equally important, tests of discriminant validity demonstrated low correlations, confirming the assessments were *not* measuring concepts related to specific (non-substance abuse-related) LSI-R subscales. It should be noted, however, that the OTMA had higher correlations in this area compared to the ASUS-R, indicating discriminant validity was an area of weakness. Analyses of other discriminant validity comparisons were difficult to interpret, as the substance use and dependence traits on all measures were highly correlated.

Limitations

While the results are informative, the study's limitations suggest cautious interpretation of the data. One challenge of the study was the recruitment of participants. The low participation numbers may have been partly due to the lack of incentives for offenders' participation. It was reported in the field that the additional time required to complete the research assessments was often a deterrent, particularly for those offenders in the community who did not want or have the ability to spend an additional 40 minutes at the agency. Offering compensation might have encouraged certain offenders to participate, boosting the recruitment levels at all sites.

To test interrater reliability, data collection procedures required the assessor to ask all the questions on the OTMA, regardless of the offender's response, thus ignoring the branching nature of the measure. This was done to give both raters an equal and independent opportunity to score each item. However, the instrument is designed to allow the assessor to skip questions based on the offender's answer (e.g., if the offender reports he does not drink beer or wine, skip the follow-up question asking how often he drinks beer or wine). Therefore, under the study design, the assessment process was somewhat artificial. For this reason, the assessors were instructed not to score the contingency tables (informing them to stop or move on to the next section) or to score the final table indicating a recommended treatment level. Because neither of these sections of the OTMA was completed, interrater reliability was determined at the item level.

Finally, results indicated that scores on both the use and dependence subscales of the GAIN-Q displayed a floor effect; that is, respondent scores on these subscales clustered near the bottom of the potential range of scores. This is problematic because these two GAIN-Q measures served as the basis for the convergent validity analysis and to establish the other interconnected patterns of relationships interpreted from the multitrait-multimethod matrix. It is likely that any correlation with

either of these GAIN-Q measures limited the potential size of the correlation coefficients and, therefore, may underestimate the convergent validity and other expected interrelationships of the scales on the ASUS-R and the OTMA.

User Feedback

One of the primary goals of the current evaluation was to gather information about the OTMA and the ASUS-R assessments to determine which instrument would be the best addiction severity measure for use in the SOA-R process. In making this decision, other factors were considered in addition to the psychometric qualities of the tools, including ease of administration (for offender and staff), time to complete, level and utility of information provided, and time and resources required for training.

These factors were explored with the staff members who administered the assessments during data collection. User feedback was collected through focus group and survey methods from the two research probation staff, eight community corrections staff, and nine CDOC staff. All methods of gathering user feedback explored the staff's experience with the ASUS-R and OTMA, perceived strengths and weaknesses, and preferences. This feedback was then organized and reviewed, and common themes and categories were identified. Although somewhat cursory, this information provided useful insight about the field staff's perspectives on what the instruments had to offer.

Comments on the ASUS-R generally were favorable. Staff found it easy to administer, largely because it is self-report. This seemed to be an especially important aspect for staff with large caseloads. In addition, offenders benefitted from having the privacy and time to answer the questions at their own pace. It was acknowledged, however, that administering the ASUS-R required more time up front to explain the instructions, as well as to clarify the questions or read them aloud, depending on the offender's literacy level.

The ASUS-R was praised for having a visual component (a bar graph) that allowed staff to see the client's growth and progress. Staff reported the level of detail available offered a wide spectrum of information for case planning, beyond what was necessary for a treatment recommendation. A key part of this was the offender's substance abuse history, providing greater overall understanding of the offender's addiction severity. A few staff voiced a concern, however, that the accuracy of this information may be questionable because offenders have been overly exposed to the assessment.

User feedback on the OTMA was mixed among the agencies. Community corrections and probation staff overwhelmingly endorsed the ASUS-R compared to the OTMA. However, feedback from CDOC staff varied. An identified benefit of the OTMA was the minimal skill and training required to administer the assessment, rendering it accessible to a wider range of staff. Users also reported they felt the OTMA was easier for the offender, as there is no reading required and the limited number of questions decreased the time to complete. The OTMA's interview format also allowed the opportunity to open a dialogue with the offender. Unfortunately, staff also felt the OTMA's

questions alone did not elicit the necessary depth of information. Rather, in order to understand the offender's past and current substance use, the assessor would need to ask further follow-up questions not included on the instrument.

The layout and organization of the OTMA was another reported concern. Several users felt it was poorly constructed, jumping from benign to very intrusive and sensitive final questions, thought to be a poor way to conclude an assessment. The scoring of certain items was deemed difficult, confusing, and inaccurate. This feedback was captured at the time of data collection and also during the follow-up with the staff. For example, for the question that asked how many times in the past 6 months the offender drank beer or wine, the only available answers were a) never, b) 1-2 times a day, or c) 3 times a day. Yet in several cases the offender reported drinking only a few times a *month* or a few times over the past *6 months*. In these cases, the scoring instructions required the assessor to score the item as *"1-2 times a day,"* even though this was substantially more than the offender reported and likely painted an inaccurate picture of an individual's alcohol use.

Recommendations

The goal of the SOA-R process is to identify the appropriate level of treatment for an offender that will result in a successful outcome without exceeding the level of care necessary to achieve that outcome. Although the OTMA provides a mechanism for treatment matching and the ASUS-R does not, the depth of the ASUS-R and its sound psychometric properties, particularly in regard to validity, support its continued use in the SOA-R process, compared to the OTMA. That recommendation is strengthened by feedback from staff in the field, where the majority felt the ASUS-R was superior, and the fact that the ASUS-R is an existing measure requiring no additional training. Thus, the equivocal findings offer weak justification for ousting the ASUS-R and replacing it with the OTMA. Moreover, the OTMA requires modifications to improve the layout, flow, and scoring issues, all of which would necessitate further reliability and validity testing. Additionally, in its current form, the OTMA's treatment matching scheme does not incorporate criminal risk—considered a critical component—in the treatment decision.

Notwithstanding the ASUS-R's strengths as a substance abuse severity measure, the original challenge remains: how best to make an informed treatment decision. A pivotal question is how risk factors into the equation; i.e., is criminal risk important because of its role in allocating funds and services (in the name of public safety) or is risk an element in the treatment of a substance-addicted offender? Recall that a criticism of the treatment recommendation algorithm in the original SOA was that it had a high rate of false positives into higher treatment levels. Specifically, it was common for high-risk offenders with low addiction severity indications to be assessed into intensive outpatient and inpatient programs, simply due to their risk of recidivism. This resulted in a very high rate (nearly 67% in some agencies) of judgmental overrides of the assessed treatment level without any structured clinical rationale. Thus, the SOA process overemphasized criminal risk in the treatment-matching decision. This committee wishes to avoid retreading old paths. At the same time, the

complexity of treatment matching and the difficulty of isolating variables that contribute to successful outcomes are daunting goals still to be addressed.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Amsel, Z., Mandell, W., Matthias, L., Mason, C., & Hocherman, I. (1976). Reliability and validity of self-reported illegal activities and drug use collected from narcotics addicts. *International Journal of the Addictions, 11,* 325-336.

Anastasi, A., & Urbina, S. (1998). *Psychological testing* (6th ed.). New York, NY: Macmillan.

Andrews, D. A. (1982). *The Level of Supervisory Inventory (LSI): 1. First follow-up.* Ottawa, Ontario: Carleton University.

Andrews, D. A., & Bonta, J. (2003). *LSI-R Level of Service Inventory Revised: U.S. norms manual supplement.* North Tonawanda, NY: Multi-Health Systems, Inc.

Andrews, D. A., & Bonta, J. (2006). *The psychology of criminal conduct* (4th ed.). Newark, NJ: LexisNexis.

Andrews, D. A., & Bonta, J. (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law, 16,* 39-55.

Andrews, D. A., & Dowden, C. (2007). The Risk-Need-Responsivity model of assessment and human service in prevention and corrections: Crime-prevention jurisprudence. *Canadian Journal of Criminology and Criminal Justice/LaRevue Canadienne de Criminologie et de Justice Pénale, 49*(4), 439-464. doi:10.3138/cjccj.49.4.439

Angarita, G. A., Reif, S., Pirard, S., Lee, S., Sharon, E., & Gastfriend, D. R. (2007). No-show for treatment in substance abuse patients with comorbid symptomatology: Validity results from a controlled trial of the ASAM Patient Placement Criteria. *Journal of Addiction Medicine, 1,* 79-87.

Arens, S. A., Durham, B., O'Keefe, M., Klebe, K., & Olene, S. (1996). *Psychometric properties of Colorado substance abuse assessment instruments.* Colorado Springs, CO: Department of Corrections.

Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument.* Washington, DC: Institute on Crime, Justice, and Corrections at the George Washington University.

Ball, J. C. (1994). What I would most like to know: Why has it proved so difficult to match drug abuse patients to appropriate treatment? *Addiction, 89,* 263-265.

Belenko, S. (2006). Assessing released inmates for substance-abuse-related service needs. *Crime & Delinquency, 52,* 94-113. doi:10.1177/0011128705281755

Belenko, S., & Peugh, J. (2005). Estimating drug treatment needs among state prison inmates. *Drug and Alcohol Dependence, 77,* 269-281.

Broome, K. M., Knight, K., Joe, G. W., & Simpson, D. D. (1996). Evaluating the drug-abusing probationer: Clinical interview versus self-administered assessment. *Criminal Justice and Behavior, 23,* 593-606. doi:10.1177/0093854896023004005

Brown, T. G., Seraganian, P., Tremblay, J., & Annis, H. (2002). Matching substance abuse aftercare treatments to client characteristics. *Addictive Behaviors, 27,* 585-604.

Budde, D., Rounsaville, B., & Bryant, K. (1992). Inpatient and outpatient cocaine abusers: Clinical comparisons at intake and one-year follow-up. *Journal of Substance Abuse Treatment, 9,* 337-342.

Bühringer, G. (2006). Allocating treatment options to patient profiles: Clinical art or science? *Addiction, 101,* 646-652.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Chestnut Health Systems. (2002). *Global Appraisal of Individual Needs (GAIN) overview.* Retrieved from http://www.chestnut.org/li/gain/GAIN_Overview.pdf

Clawson, E., Bogue, B., & Joplin, L. (2005). Implementing evidence-based practices in corrections; using an integrated model to implement evidence-based practices in corrections. Washington, DC: National Institute of Corrections. Retrieved from http://nicic.gov/library/020174

Colorado Department of Human Services. (2007). SOA-R interrater reliability evaluation: A technical report from the Colorado Division of Mental Health. Denver: Colorado Department of Human Services, Division of Mental Health.

Crawley, W. R. (2009). *Self-reporting and the diagnostic screening of substance abusing prisoners.* El Paso, TX: LFB Scholarly Publishing.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668-1674.

De Leon, G., Melnick, G., & Cleland, C. (2008). Client matching: A severity-treatment intensity paradigm. *Journal of Addictive Diseases, 27,* 99-113. doi:10.1080/10550880802122703

De Leon, G., Melnick, G., & Cleland, C. (2010). Matching to sufficient treatment: Some characteristics of undertreated (mismatched) clients. *Journal of Addictive Diseases, 29,* 59-67. doi:10.1080/10550880903436051

Dowdy, E. R., Lacy, M. G., & Unnithan, N. P. (2002). Correctional prediction and the Level of Supervision Inventory. *Journal of Criminal Justice, 30,* 29-39.

Dunn, G. (2004). *Statistical evaluation of measurement errors* (2nd ed.). New York, NY: Arnold Publishers.

Engleman, L., & Weber, K. R. (2011). *Overview of substance abuse treatment services: Fiscal year 2010.* Colorado Springs, CO: Department of Corrections.

Fleiss, J. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651-659.

Flores, A. W., Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2006). Predicting outcome with the Level of Service Inventory-Revised: The importance of implementation integrity. *Journal of Criminal Justice, 34,* 523-529. doi:10.1016/j.jcrimjus.2006.09.007

Gartner, L., & Mee-Lee, D. (1995). *The role and current status of patient placement criteria in the treatment of substance abuse disorders* (Treatment Improvement Protocol No. 13). Rockville, MD: Center for Substance Abuse Treatment, Substance Abuse & Mental Health Services Administration. Retrieved from http://radar.boisestate.edu/pdfs.TIP13.pdf

Gastfriend, D. R., Lu, S., & Sharon, E. (2000). Placement matching: Challenges and technical progress. *Substance Use & Misuse, 35,* 2191-2213.

Gastfriend, D. R., & McLellan, T. (1997). Treatment matching: Theoretic basis and practical implications. *Medical Clinics of North America, 81,* 945-966.

Greenfield, S. F., Brooks, A. J., Gordon, S. M., Green, C. A., Kropp, F., McHugh, R. K., Lincoln, M., …Miele, G. M. (2007). Substance abuse treatment entry, retention, and outcome in women: A review of the literature. *Drug and Alcohol Dependence, 86,* 1-21. doi:10.1016/j.drugalcdep.2006.05.012

Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. *Journal of Abnormal Psychology, 115,* 192-194. doi:10.1037/0021-843X.115.2.192

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293-323.

Guy, L. S. (2008). *Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey* (Doctoral dissertation). Retrieved from http://ir.lib.sfu.ca/handle/1892/10581

Hanson, R. K. (2005). Twenty years of progress in violence risk assessment. *Journal of Interpersonal Violence, 20,* 212-217. doi:10.1177/0886260504267740

Harrison, L. (1995). The validity of self-reported data on drug use. *Journal of Drug Issues, 25,* 91-111.

Heather, N. (1999). Some common methodological criticisms of Project MATCH: Are they justified? Comments on Project MATCH: Matching alcohol treatments to client heterogeneity. *Addiction, 94,* 31-69.

Hemphill, J. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58,* 78-79.

Hser, Y.-I., Polinsky, M. L., Maglione, M., & Anglin, M. D. (1999). Matching clients' needs with drug treatment services. *Journal of Substance Abuse Treatment, 16,* 299-305.

Interagency Advisory Committee on Adult & Juvenile Correctional Treatment. (2009, November 30). *Meeting of the Interagency Advisory Committee on Adult & Juvenile Correctional Treatment: Special meeting—Standardized Offender Assessment.*

Jolley, J. M., & Kerbs, J. J. (2010). Risk, need, and responsivity: Unrealized potential for the international delivery of substance abuse treatment in prison. *International Criminal Justice Review, 20,* 280-301. doi:10.1177/1057567710373115

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63,* 146-159. doi:10.1037/0003-066X.63.3.146

Kills Small, N. J., Simons, J. S., & Stricherz, M. (2007). Assessing criterion validity of the Simple Screening Instrument for Alcohol and Other Drug Abuse (SSI-AOD) in a college population. *Addictive Behaviors, 32,* 2425-2431. doi:10.1016/j.addbeh.2007.04.003

Knight, J. R., Goodman, E., Pulerwitz, T., & DuRant, R. H. (2000). Reliabilities of short substance abuse screening tests among adolescent medical patients. *Pediatrics, 105,* 948-953.

Knight, K., Hiller, M. L., Simpson, D. D., & Broome, K. M. (1998). The validity of self-reported cocaine use in a criminal justice treatment sample. *American Journal of Drug and Alcohol Abuse, 24,* 647-660.

Knight, K., Simpson, D. D., & Hiller, M. L. (2002). Screening and referral for substance-abuse treatment in the criminal justice system. In C. G. Leukefeld, F. Tims, & D. Farabee (Eds.), *Treatment of drug offenders: Policies and issues* (pp. 259-272). New York: Springer.

Latessa, E. J., Cullen, F. T., & Gendreau, P. (2002). Beyond professional quackery-professionalism and the possibility of effective treatment. *Federal Probation, 66,* 43-49. doi:10.1177/1057567710373115

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E. J. (2004). Assessing the inter-rater agreement of the Level of Service Inventory-Revised. *Federal Probation, 68*(3), 34-38.

Lu, N. T., Taylor, B. G., & Riley, K. J. (2001). The validity of adult arrestee self-reports of crack cocaine use. *American Journal of Drug and Alcohol Abuse, 27,* 399-419.

Magura, S., Fong, C., Staines, G. L., Cleland, C., Foote, J., Rosenblum, A., Kosanke, N., & DeLuca, A. (2005). The combined effects of treatment intensity, self-help groups and patient attributes on drinking outcomes. *Journal of Psychoactive Drugs, 37,* 85-92.

Magura, S., & Kang, S.-Y. (1996). Validity of self-reported drug use in high risk populations: A meta-analytical review. *Substance Use & Misuse, 31,* 1131-1153.

Magura, S., Staines, G., Kosanke, N., Rosenblum, A., Foote, J., DeLuca, A., & Bali, P. (2003). Predictive validity of the ASAM patient placement criteria for naturalistically matched vs. mismatched alcoholism patients. *American Journal on Addictions, 12,* 386-397. doi:10.1080/10550490390240765

Mattson, M.E., Allen, J. P., Longabaugh, R., Nickless, C., Connors, G. J., & Kadden, R. M. (1994). A chronological review of empirical studies matching alcoholic clients to treatment. *Journal of Studies on Alcohol (Supplement), 12,* 16-29.

McKay, J. R., Cacciola, J. S., McLellan, A.T., Alterman, A. I., & Wirtz, P. W. (1997). An initial evaluation of the psychosocial dimensions of the American Society of Addiction Medicine criteria for inpatient versus intensive outpatient substance abuse rehabilitation. *Journal of Studies on Alcohol, 58,* 239-252.

McKay, J. R., Donovan, D. M., McLellan, T., Krupski, A., Hansten, M., Stark, K. D., Geary, K., & Cecere, J. (2002). Evaluation of full vs. partial continuum of care in the treatment of publicly funded substance abusers in Washington state. *American Journal of Drug and Alcohol Abuse, 28,* 307-338.

McKay, J. R., & Weiss, R. V. (2001). A review of temporal effects and outcome predictors in substance abuse treatment studies with long-term follow-ups: Preliminary results and methodological issues. *Evaluation Review, 25,* 113-161.

McLellan, A. T., & Alterman, A. I. (1991). Patient treatment matching: A conceptual and methodological review with suggestions for future research. In R. W. Pickens, C. G. Leukefeld, & C. R. Schuster (Eds.), *Improving drug abuse treatment* (pp. 114-135) (Research Monograph 106). Rockville, MD: National Institute on Drug Abuse.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota Press.

Mee-Lee, D. (2005). *Overview of the ASAM patient placement criteria, second edition revised (ASAM PPC-2R)* [PowerPoint slides]. Retrieved from

http://coce.samhsa.gov/cod_resources/PDF/ASAMPatientPlacementCriteriaOverview5-05.pdf

Melnick, G., De Leon, G., Thomas, G., & Kressel, D. (2001). A client-treatment matching protocol for therapeutic communities: First report. *Journal of Substance Abuse Treatment, 21,* 119-128.

Melnick, G., Schoeneberger, M., & Sacks, S. (2009). *Final report: Updating the Colorado Standardized Offender Assessment-Revised (SOA-R).* New York: Center for the Integration of Research and Practice, National Development and Research Institutes.

Mieczkowski, T. (1990). The accuracy of self-reported drug use: An evaluation and analysis of new data. In R. Weisheit (Ed.), *Drugs, crime and the criminal justice system* (pp. 275-302). Cincinnati, OH: Anderson.

Moyer, A., Finney, J. W., Elworth, J. T., & Kraemer, H. C. (2001). Can methodological features account for patient-treatment matching findings in the alcohol field? *Journal of Studies on Alcohol, 62,* 62-73.

Mumola, C. J., & Karberg, J. C. (2006). *Drug use and dependence, state and federal prisoners, 2004* (NCJ 213530). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

National Institute on Drug Abuse. (2006). *Principles of drug abuse treatment for criminal justice populations: A research-based guide* (NIH Publication No. 06-5316). Bethesda, MD: National Institute on Drug Abuse. Retrieved from http://www.atforum.com/addiction-resources/documents/PODAT_CJ.pdf

National Institute on Drug Abuse. (2009). *Principles of drug addiction treatment: A research-based guide* (2nd ed.) (NIH Publication No. 09-4180). Bethesda, MD: National Institute on Drug Abuse. Retrieved from http://www.nida.nih.gov/podat/Principles.html

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

O'Keefe, M., Klebe, K., & Hromas, S. (1998). *Validation of the Level of Supervision Inventory (LSI) for community based offenders in Colorado: Phase II* (Technical Report). Colorado Springs, CO: Department of Corrections.

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology, 31,* 109-118.

Peters, R. H., Greenbaum, P. E., Steinberg, M. L., Carter, C. R., Ortiz, M. M., Fry, B. C., & Valle, S. K. (2000). Effectiveness of screening instruments in detecting substance abuse disorders among prisoners. *Journal of Substance Abuse Treatment, 18,* 349-358.

Project MATCH Research Group. (1998a). Matching alcoholism treatments to client heterogeneity: Project MATCH three-year drinking outcomes. *Alcoholism: Clinical and Experimental Research, 22,* 1300-1311.

Project MATCH Research Group. (1998b). Matching patients with alcohol disorders to treatments: Clinical implications from Project MATCH. *Journal of Mental Health, 7,* 589-602.

Rosay, A. B., Najaka, S. S., & Herz, D. C. (2007). Differences in the validity of self-reported drug use across five factors: Gender, race, age, type of drug, and offense seriousness. *Journal of Quantitative Criminology, 23,* 41-58. doi:10.1007/s10940-006-9018-7

Rychtarik, R. G., Connors, G. J., Whitney, R. B., McGillicuddy, N. B., Fitterling, J. M., & Wirtz, P. W. (2000). Treatment settings for persons with alcoholism: Evidence for matching clients to inpatient versus outpatient care. *Journal of Consulting and Clinical Psychology, 68,* 277-289. doi:10.1037//0022-006X.68.2.277

Shaffer, D., Schwab-Stone, M., Fisher, P. W., Cohen, P., Piacentini, J., Davies, M., Conners, C. K., & Regier, D. (1993). The Diagnostic Interview Schedule for Children—Revised version (DISC–R): I. Preparation, field testing, interrater reliability, and acceptability. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 643–650.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Silver, E., & Miller, L. L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime & Delinquency, 48,* 138-161. doi:10.1177/0011128702048001006

Simpson, D. D. (2002). A conceptual framework for transferring research to practice. *Journal of Substance Abuse Treatment, 22,* 171-182.

Simpson, D. D., Joe, G. W., Fletcher, B. W., Hubbard, R. L., & Anglin, M. D. (1999). A national evaluation of treatment outcomes for cocaine dependence. *Archives of General Psychiatry, 56*, 507-514.

Smith, B. D., & Marsh, J. C. (2002). Client-service matching in substance abuse treatment for women with children. *Journal of Substance Abuse Treatment, 22*, 161-168.

Steiger, J. H. (1980). Test for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

Taleff, M. (2006). The accuracy of clinical judgment. *Counselor*, *7*(4), 29-30.

Tapia, G. (2010, August). *Standardized offender assessment* [PowerPoint slides]. Presented to the Interagency Advisory Committee on Adult and Juvenile Correctional Treatment (IACAJCT), Denver.

Taylor, B. G., Fitzgerald, N., Hunt, D., Reardon, J. A., & Brownstein, H. H. (2001). *ADAM preliminary 2000 findings on drug use and drug markets—adult male arrestees* (NCJ 189101). Washington, DC: National Institute of Justice. Retrieved from http://www.ncjrs.gov/pdffiles1/nij/189101.pdf

Taxman, F. S. (2007, May). Findings from a national survey of correctional agencies on substance abuse treatment and health services: Who can get served? In T. P. Condon & R. K. Chandler (Co-chairs), *Drug abuse treatment within the criminal justice system: Addressing our nation's public health needs.* Symposium conducted at the meeting of the National Institute on Drug Abuse, National Institutes of Health, and the American Psychiatric Association, San Diego, CA.

Taxman, F. S., Cropsey, K. L., Young, D. W., & Wexler, H. (2007). Screening, assessment, and referral practices in adult correctional settings: A national perspective. *Criminal Justice and Behavior, 34,* 1216-1234. doi:10.1177/0093854807304431

Thanner, M. H., & Taxman, F. S. (2003). Responsivity: The value of providing intensive services to high-risk offenders. *Journal of Substance Abuse Treatment, 24,* 137-147. doi:10.1016/S0740-5472(02)00352-5

Thornton, C. C., Gottheil, E., Weinstein, S. P., & Kerachsky, R. S. (1998). Patient-treatment matching in substance abuse. *Journal of Substance Abuse Treatment, 15,* 505-511.

Tiet, Q. A., Ilgen, M. A., Byrnes, H. F., Harris, A. H. S., & Finney, J. W. (2007). Treatment setting and baseline substance use severity interact to predict patients' outcomes. *Addiction, 102,* 432-440. doi:10.1111/j.1360-0443.2006.01717.x

Trochim, W. M. K. (2006). Measurement validity types. *Research Methods Knowledge Base.* Retrieved from http://www.socialresearchmethods.net/kb/measval.php

Turner, W. M., Turner, K. H., Reif, S., Gutowski, W. E., & Gastfriend, D. R. (1999). Feasibility of multidimensional substance use treatment matching: Automating the ASAM Patient Placement Criteria. *Drug and Alcohol Dependence, 55,* 35-43.

Wanberg, K. W. (2006). *User's guide: Adult Substance Use Survey-Revised.* Arvada, CO: Center for Addictions Research and Evaluation.

Wanberg, K. W. (2009). *ASUS-R/Adult Substance Use Survey-Revised.* Retrieved from http://aodassess.com/assessment_tools/asus/

Weber, K., DeLaCerda, D., & O'Keefe, M. (2010). *An evaluation of the use of the LSI-R with Colorado inmates.* Colorado Springs, CO: Department of Corrections.

Wells, D. (2003). Drug-assessment instruments: Making wise choices. *Corrections Today, 65*(6), 28-30.

Winters, K. C., & Zenilman, J. M. (1994). *Simple screening instruments for outreach for alcohol and other drug abuse and infectious diseases. Treatment Improvement Protocol (TIP) Series 11* (DHHS Publication No. (SMA) 95-3058). Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Substance Abuse Treatment. Retrieved from http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=hssamhsatip&part=A32939

Witkiewitz, K., Hartzler, B., & Donovan, D. (2010). Matching motivation enhancement treatment to client motivation: Re-examining the Project MATCH motivation matching hypothesis. *Addiction, 105,* 1403-1413. doi:10.1111/j.1360-0443.2010.02954.x

Appendix A: Colorado's Standardized Treatment System

Level 1: No Treatment. Offenders assigned to the first level include those who require no substance abuse treatment. Level 1 individuals have no current clinical diagnosis of substance abuse or dependence.

Level 2: Drug and Alcohol Education and Increased Urinalyses. The second treatment level is prescribed for individuals with no clinical diagnosis of substance abuse or dependence and no withdrawal symptoms. Participants must be mentally, physically, and cognitively capable of participation. Level 2 education should be structured and curriculum-based; however, content may vary. These programs must incorporate urinalyses and substance abuse education that strives to increase awareness of the effects of drug and alcohol usage. The hours for this level are generally open to the facilities' own guidelines.

Level 3: Weekly Outpatient Treatment (WOP). WOP is structured for individuals who may have a clinical diagnosis of substance abuse or dependence. Level 3 programs are designed for individuals who experience up to two of the following consequences of substance abuse: employment difficulties, legal difficulties, or damaged personal relationships. Individuals may be appropriate for weekly outpatient treatment if they completed a more intensive program and need continuing support. It is recommended that participants at this level are voluntary, motivated, and have positive support systems outside of treatment. The general parameters of weekly outpatient treatment are 1 to 3 hours per week within a 3- to 12-month duration.

Level 4a: Enhanced Outpatient Treatment (EOP). EOP is also structured for individuals who may have a clinical diagnosis of substance abuse or dependence. Individuals may have an emotional, behavioral, or cognitive disorder requiring intensive monitoring to minimize distractions from treatment. Individuals also may have performed poorly in weekly outpatient therapy while under supervision and may have a poor support system, which necessitates more intensive treatment services. Individuals may be appropriate for enhanced outpatient treatment if they completed a more intensive program and need continuing support. It is recommended, although not required, that participants at this level are voluntary and motivated. Level 4a programs typically involve 3 to 8 hours of treatment services each week for a period of 4 to 12 weeks. Level 4a treatment should be followed by level 3 treatment.

Level 4b: Intensive Outpatient Treatment (IOP). IOP is designed for individuals who require more structured therapy than is provided through weekly or enhanced outpatient treatment. Individuals may be assigned to level 4b treatment if their histories indicate unsuccessful treatment attempts, prolonged intoxication or lengthy withdrawal, and serious behavioral changes (e.g., legal or family problems) due to substance abuse. Intensive outpatient programs are best suited to individuals who realize the seriousness of their substance abuse problems, but have negative or

nonexistent support systems and are a high risk for continued drug use. This level may be appropriate for individuals who have a dual diagnosis, present a risk to self or others, or have successfully completed an intensive residential program. Level 4b differs from levels 3 and 4a primarily in intensity; level 4b programs are typically 9 to 19 treatment hours per week for 4 to 6 weeks with continuing care and should be followed by level 4a or level 3 treatment.

Level 4c: Intensive Residential Treatment (IRT). IRT is designed for individuals presenting with serious substance abuse problems. Intensive residential programs typically last 30 days; however, some programs may be as short as 14 days or as long as 120 days. The purpose of level 4c treatment is to provide a brief, intense treatment intervention. The programs are structured to accommodate persons requiring medical attention, particularly for disorders related to prolonged substance abuse. Individuals who are diagnosed with a chronic psychiatric disorder or present a danger to self or others are appropriate for placement in a level 4c program. Additionally, level 4c programs treat individuals who lack a positive support system, experience substantial denial, or exhibit inability to sustain independent functioning outside of a controlled environment. Level 4c treatment should be followed by a progressive continuum of level 4b, 4a, and 3 treatments.

Level 4d: Therapeutic Community (TC). TCs are similar to level 4c in that they are residential in nature, but 4d programs have greater lengths of stay and are more structured than the lower levels. Typically, TCs have a 6- to 12-month length of stay. TCs are designed for individuals with extensive criminal histories, antisocial behavior, limited social support, and multiple unsuccessful treatment attempts. Level 4d programs use confrontation techniques and place high levels of responsibility on the individual participants for their treatment.

Level 5: Medical/Mental Health Referral. The final level of treatment is geared toward individuals with the most serious cognitive, emotional, and/or behavioral problems. Typically, individuals have a psychiatric disorder or a severe medical issue that warrants a suspension of substance abuse treatment until the issue can be addressed. Due to the extreme severity of their psychological or physical problems, individuals in this category may be considered inappropriate for treatment and should be referred for a medical and/or mental health evaluation by a licensed professional.

Sanctions. Another mandate under C.R.S. 16–11.5 is a system of punitive sanctions for continued substance use. The system of sanctions is related to the SOA-R and takes the following factors into consideration: (a) level of criminal risk, (b) severity of substance abuse, (c) compliance in supervision, and (d) treatment participation and compliance. Recognizing that each agency must develop its own policies and procedures to respond to offender substance use, specific sanctions (e.g., SOA-R reassessment, increased supervision, loss of privileges) are suggested depending on the four factors (Engleman & Weber, 2011).

Appendix B: Technical Results

Descriptive Statistics

Table 1 displays the means and standard deviations of the study scales for each of the participating agencies. In all cases, the means for the Department of Corrections (DOC) and community corrections (CC) agencies were higher than the means for the parole and probation agencies. It is likely that the variation is partly an artifact of how study participants from each agency were recruited.

Figure 1 displays a scatterplot matrix for the scales that measure the use/abuse and dependence traits. Histograms for each scale appear along the diagonal. It is apparent from the histograms that the distributions of the scales are positively skewed, especially the GAIN-Q scales. These departures from normality, with most scores falling at the low end of the scale, will tend to attenuate the bivariate correlations between the scales, thereby reducing the validity coefficients.

To test for differences in study scale scores across the agencies, a one-way analysis of variance was conducted on each scale. Significant agency differences were found for all scales except the GAIN-Q use and dependence scales. Post-hoc comparisons using Tukey's multiple comparison of means test indicated that the community corrections sample scored higher than the probation sample on the OTMA consequences and the LSI-R accommodations scales and higher than both the probation and DOC samples on the OTMA use and LSI-R attitude scales. The DOC sample scored higher than the probation sample on the OTMA use scale and higher than both the probation and community corrections samples on the LSI-R attitude scale. The community corrections sample also scored higher than the parole sample on the OTMA use and consequences scales and the LSI-R accommodations scale.

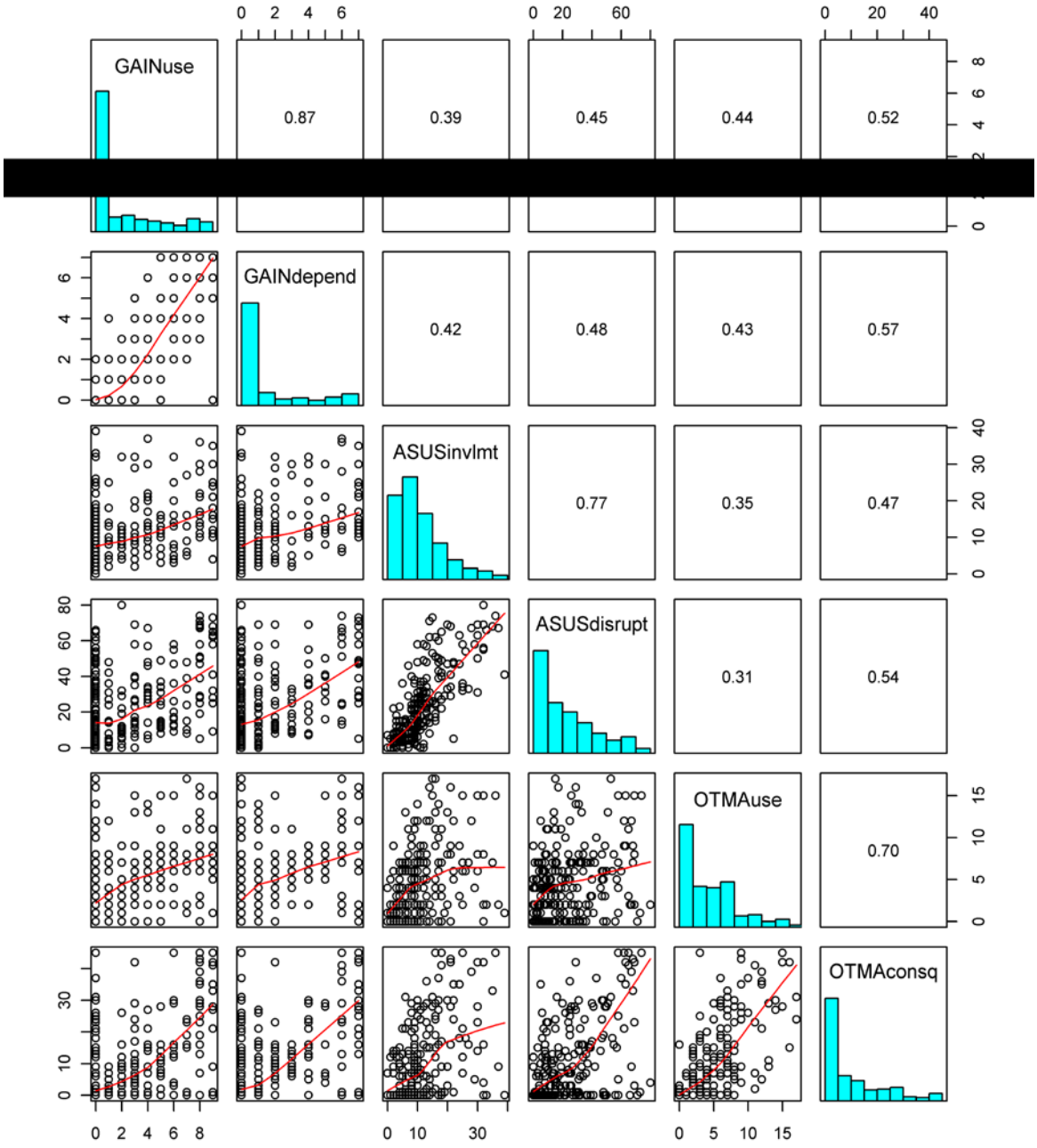
In summary, the descriptive statistics and pattern of mean differences between the agencies follow expected patterns for these kinds of scales. The positively skewed distributions for the scales are in line with what would be expected for summated scales, but the degree of skew is cause for concern, especially for the GAIN-Q scales that are reference scales for the assessment of construct validity.

Table 1. Descriptive Statistics for the Study Scales in the Four Agency Groups

	DOC	CC	PAROLE	PROBATION
	n = 105	*n* = 51	*n* = 25	*n* = 57
GAIN-Q Use/Abuse				
M	2.74	2.55	1.68	2.47
SD	2.92	3.15	2.64	3.08
GAIN-Q Dependence				
M	1.75	2.18	1.08	1.32
SD	2.31	2.69	1.87	2.13
ASUS-R Involvement				
M	$11.62_{a,b}$	13.84_{a}	7.92_{b}	$10.18_{a,b}$
SD	7.62	8.29	5.63	8.70
ASUS-R Disruption				
M	25.75_{a}	27.75_{a}	13.48_{b}	$18.00_{a,b}$
SD	19.04	21.05	14.35	21.52
OTMA USE				
M	$4.82_{a,b}$	6.55_{b}	$4.16_{a,b,c}$	2.93_{c}
SD	4.20	4.14	3.87	3.34
OTMA Consequences				
M	9.60_{a}	15.98_{b}	7.36_{a}	5.70_{a}
SD	11.84	13.41	11.34	10.12
LSI-R Attitude				
M	4.13_{a}	3.25_{b}	$2.60_{b,c}$	2.49_{c}
SD	1.16	1.56	1.38	1.28
LSI-R Accommodation				
M	$2.18_{a,b,c}$	2.57_{a}	1.96_{b}	1.89_{b}
SD	.98	1.14	.93	.77

Note. Means not sharing a subscript are significantly different from each other ($p < .05$, two-tailed).

Figure 1. Scatterplot Matrix for the Study Scales



Note. Correlations appear above the diagonal, with corresponding scatterplots below the diagonal. Histograms for the scales appear along the diagonal.

Internal Consistency Reliability

OTMA scales. Table 2 reports the alpha coefficients for each OTMA scale. The alpha coefficients ranged from .63 for Scale C (frequency drugs) to .85 for Scale E (consequences). Only Scale A (frequency beer/wine), Scale D (substance use), and Scale E (consequences) have alphas equal to or above .70. The alpha coefficients for the use (D) and consequences (E) scales are the same as those shown in Table 10's construct validity correlation matrix.

Table 2. OTMA Internal Consistency Reliability Coefficients

Scale	Alpha	N	Number of Items
A: frequency beer/wine	.70	238	4
B: frequency hard liquor	.63	238	4
C: frequency drugs	.63	238	4
D: substance use overall	.75	238	12
E: consequences	.85	238	12
F: social indicators	.66	238	5
G: habilitation	.63	238	3

Results from an item-level analysis (not shown) indicate that the low alphas on Scales B, C, F, and G stem primarily from too few items. The correlations of each item with the total scale score were reasonable in almost all cases. On Scale F (social factors), item 21A (previous arrest for violence) correlated at only .14 with the total score, indicating that it had little in common with the other items on the scale. Most of the short scales with weak alpha coefficients could be improved to a level above .70 by adding 50% more items of similar quality, based on the Spearman-Brown prophecy formula (Shrout & Fleiss, 1979).

ASUS-R scales. Table 3 reports the alpha coefficients for the ASUS-R scales. The ASUS-R scales had higher alpha coefficients compared to those obtained for the OTMA scales. The alpha coefficients for the ASUS-R scales ranged from .79 for social non-conforming to .96 for global AOD. The ASUS-R disruption and involvement scales demonstrated somewhat better internal consistency reliability than the OTMA consequences and use scales, respectively, and the difference between the coefficients was significant. In contrast to the OTMA scales, the ASUS-R scales have more items per scale. Alpha coefficients above .90 sometimes indicate narrow content sampling (e.g., items that are the same except for minor changes in wording) or possible problems due to response set bias. Although our analyses did not directly address these issues, there are certain features of the ASUS-R that make it vulnerable to response set bias. For example, the disruption items all have the same number of categories scored in the same direction, and it is possible for respondents to circle the same answer on all 20 items simultaneously. Note that the number of cases with complete data on each ASUS-R scale varied. The alpha coefficients for involvement and disruption are the same as those reported later in Table 10 for the construct validity analysis.

Table 3. ASUS-R Internal Consistency Reliability Coefficients

Scale	Alpha	N	Items
Involvement	.84	225	10
Disruption	.94	221	20
AOD 6 Mths	.95	194	30
AOD Use Benefits	.93	225	10
Social Non-Conforming	.79	213	12
Legal Non-Conforming	.84	218	13
Legal Non-Conforming 6 Mths	.83	195	13
Mood Adjustment	.92	217	10
Global AOD	.96	202	52
Defensive	.85	224	7
Motivation Change	.89	218	7
Strengths	.86	220	9

Interrater Reliability

Calculated treatment level assignment. Recall that interviewers only recorded answers to the items on the OTMA but did not complete the treatment assignment tables. The treatment assignment level was computer-calculated using the answers the interviewers provided. The results for raters A and B are reported in Tables 4 and Table 5, respectively (n = 238). Table 6 shows the cross-tabulation of the calculated treatment levels of raters A and B. Approximately 10% (n = 23) of the calculated treatment assignments were discordant, indicating a high level of agreement. Overall, approximately 17% of the participants were assigned to *No Treatment* (Level 1); 16% to *Basic Services* (Levels 2-3); and 68% to *Enhanced Services* (Levels 4a-4d). The calculation did not assign cases to Level 5 (medical/mental health referral) because that information was not available for the cases under consideration.

Table 4. Calculated OTMA Assigned Treatment Levels: Rater A

Calculated Level	Frequency	Percent
1: No treatment	41	17.2
2: Education	20	8.4
3: Weekly outpatient	14	5.9
4a: Enhanced outpatient	34	14.3
4b: Intensive outpatient	51	21.4
4c: Intensive residential	16	6.7
4d: Therapeutic community	62	26.1
Sum	238	100.0

Table 5. Calculated OTMA Assigned Treatment Levels: Rater B

Calculated Level	Frequency	Percent
1: No treatment	39	16.4
2: Education	22	9.2
3: Weekly outpatient	13	5.5
4a: Enhanced outpatient	34	14.3
4b: Intensive outpatient	55	23.1
4c: Intensive residential	14	5.9
4d: Therapeutic community	61	25.6
Sum	238	100.0

Table 6. Comparison of the Calculated OTMA Assigned Treatment Levels: Two Raters

Rater A	1	2	3	4a	4b	4c	4d	Sum
1: No treatment	36	5	0	0	0	0	0	41
2: Education	2	14	3	1	0	0	0	20
3: Weekly outpatient	1	2	10	1	0	0	0	14
4a: Enhanced outpatient	0	1	0	32	1	0	0	34
4b: Intensive outpatient	0	0	0	0	50	0	1	51
4c: Intensive residential	0	0	0	0	1	14	1	16
4d: Therapeutic community	0	0	0	0	3	0	59	62
Sum	39	22	13	34	55	14	61	238

Table 7 reports the percent of participants in the calculated OTMA-assigned treatment levels within each agency. More DOC participants (13.3%) were assigned to 4c (intensive residential treatment) compared to those in community corrections (0.0%), parole (0.0%), and probation (3.5%). The percent assigned to 4d (therapeutic community) was higher for participants in DOC (29.5%) and community corrections (45.1%) compared to parole (8.0%) and probation (10.5%). The proportion of participants with a calculated assignment to 4b (intensive outpatient) was very similar across agencies.

Table 7. Calculated OTMA Assigned Treatment Levels in Agencies

		Agency							
Tx Level		DOC		CC		Parole		Probation	
1:	No treatment	14	(13.3)	3	(5.9)	3	(12.0)	21	(36.8)
2:	Education	10	(9.5)	1	(2.0)	4	(16.0)	5	(8.8)
3:	Weekly outpatient	7	(6.7)	1	(2.0)	2	(8.0)	4	(7.0)
4a:	Enhanced outpatient	9	(8.6)	12	(23.5)	7	(28.0)	6	(10.5)
4b:	Intensive outpatient	20	(19.0)	11	(21.6)	7	(28.0)	13	(22.8)
4c:	Intensive residential	14	(13.3)	0	(0.0)	0	(0.0)	2	(3.5)
4d:	Therapeutic community	31	(29.5)	23	(45.1)	2	(8.0)	6	(10.5)
Sum		105	(100.0)	51	(100.0)	25	(100.0)	57	(100.0)

Intraclass correlation coefficients. One-way analysis of variance was conducted on each item to obtain intraclass correlation coefficients (ICC) at the item level (Shrout & Fleiss, 1979). The ICC is the correlation between interviewer A's answer on an item with interviewer B's answer on that item. Table 8 reports the ICC and 95% confidence interval for each item from the Suitability, Screening, and Alcohol /Drug Use sections of the OTMA. The ICCs are displayed in Table 9 along with the 95% confidence interval for each item for the Social, Habilitation, and Mitigating Factors sections of the OTMA. The percent of participants interviewed by interviewer A that endorsed higher than the lowest response category is also reported in each table as an indication of the most common respondent characteristics.

Table 8. ICC Coefficients for OTMA Questions 1-17 (n = 238)

OTMA Question	Description	% (A) Endorsed	ICC	Lower 95% CI	Upper 95% CI
Q1	Mental disorder	18.5	.93	.91	.95
Q2	Medical condition	3.8	.87	.84	.90
Q3	Contagious disease	5.5	.96	.95	.97
Q5	Alcohol often	47.5	.82	.77	.86
Q6	More alcohol	10.9	.98	.97	.98
Q7	Try cut back	9.7	.98	.97	.98
Q8	Five or more	29.4	.96	.95	.97
Q9	Hard liquor	42.9	.90	.87	.92
Q10	More hard liquor	8.4	1.00		
Q11	Try cut back	6.3	.97	.96	.97
Q12	Five or more	23.1	.99	.99	.99
Q13	Illegal drugs	54.6	1.00		
Q14	Frequency drugs	63.4	.98	.97	.98
Q15	More drugs	18.9	.99	.98	.99
Q16	Try cut back	10.9	.96	.95	.97
Q17A	Medical problems	18.9	.99	.98	.99
Q17B	Blacking out	14.3	1.00		
Q17C	Family relations	36.6	.99	.99	.99
Q17D	General attitude	34.5	.99	.98	.99
Q17E	Concentration	36.6	.97	.97	.98
Q17F	Going to work	23.1	.97	.96	.98
Q17G	Money	29.4	.99	.99	1.00
Q17H	Legal trouble	49.6	.97	.96	.98
Q17I	Shaking	13.4	1.00		
Q17J	Need drugs	28.2	.98	.98	.99
Q17K	Risky behavior	17.2	1.00	1.00	1.00
Q17L	Fighting	27.3	.96	.95	.97

Note: Question 4 was not answered.

Table 9. ICC Coefficients for OTMA Questions 18-29 (n = 238)

OTMA Question	Description	% (A) Endorsed	ICC	Lower 95% CI	Upper 95% CI
Q18	Friends drugs	64.3	.97	.97	.98
Q19	Friends criminals	31.1	.98	.98	.99
Q20	Live w drunks	31.1	.98	.98	.99
Q21	Live w criminals	19.7	.95	.93	.96
Q21A	Previous arrests	34.0	.97	.96	.98
Q22	GED	26.9	.96	.95	.97
Q23	Fulltime job	40.3	.94	.92	.95
Q24	Earn a living	20.2	.95	.93	.96
Q25A	Court mandated	35.3	.97	.97	.98
Q25B	Treatment mandated	21.8	.85	.81	.88
Q26A	Substance abuse tx	37.0	.91	.89	.93
Q26B	Treatment outpatient	20.2	.96	.95	.97
Q26C	Treatment attendance	28.6	.91	.88	.93
Q27A	Sole caretaker	21.8	.98	.97	.98
Q27B	Caretaking help	24.8	.92	.90	.94
Q28	Arrested sex assault	5.9	.96	.95	.97
Q29	Arrested arson	1.3	1.00		

The general consensus in the literature is that an ICC above .80 indicates good reliability (see, for example, Anastasi & Urbina, 1998). The ICCs in the interrater reliability study ranged from .82 to 1.00. More than 85% of the items had ICCs above .90 and more than 50% of the items had ICCs above .95. Given the highly structured design of the OTMA, large ICCs were expected. Large ICCs in the range found for the OTMA have been observed for highly structured instruments such as the Diagnostic Interview Schedule for Children (Shaffer et al., 1993).

The ICCs for items on the substance use scale (Q5 – Q16) were very similar to the ICCs on the consequences scale (Q17A – Q17L). For some of the items, the ICCs were equal to 1.00, indicating that the interviewers were in perfect agreement on these items. The two interviewers were in perfect agreement on Q10 (more hard liquor); Q13 (illegal drugs); Q17B (blacking out); Q17I (shaking); and Q29 (arrested arson). These items have in common low (or high) base rate (percent endorsed) and concrete underlying behaviors.

In summary, the OTMA items demonstrated excellent interrater reliability overall. In addition, the treatment level assignments calculated from the item answers of the two raters resulted in good agreement, with an error rate of 10%. This demonstrates that very high agreement at the item level can translate into slightly lower agreement in calculated treatment assignment.

Construct Validity

To evaluate the convergent and discriminant validity of the OTMA and ASUS-R, the general logic of the multitrait-multimethod matrix (MTMMM) of Campbell and Fiske (1959) was used. The present study included four instruments (OTMA, ASUS-R, GAIN-Q, and LSI-R) and attempted to validate the OTMA and the ASUS-R based on the *DSM-IV* substance abuse and dependence criteria.

The method cannot be used exactly as described by Campbell and Fiske (1959) due to the absence of multiple assessment models that match one-to-one on the traits measured. The MTMMM approach requires that several traits be measured by each of several methods. One aspect of the current design that does meet this requirement is that three instruments (GAIN-Q, OTMA, ASUS-R) measure the same two traits (use and dependence). Thus, this part of the correlation matrix of study measures can be interpreted closely following Campbell and Fiske's (1959) approach. However, the design also includes two traits from the LSI-R (attitude/orientation and accommodation) that do not match up with the other traits. The part of the matrix involving LSI-R traits will be interpreted separately and only in terms of discriminant validity. The correlation matrix adapted for the current study is displayed in Table 10.

Table 10. Correlations of ASUS-R and OTMA Scales with GAIN-Q Scales: Construct Validity

	GAIN-Q Use	GAIN-Q Depend.	ASUS-R Invlvmnt	ASUS-R Disrupt.	OTMA Use	OTMA Conseq.	LSI-R Attitude	LSI-R Accom.
GAIN-Q Use	(.90)							
GAIN-Q Depend.	*.87*	(.89)						
ASUS-R Invlvmnt	.39	.42	(.84)					
ASUS-R Disrupt.	.45	.48	*.77*	(.94)				
OTMA Use	.44	.43	.35	.31	(.75)			
OTMA Conseq.	.52	.57	.47	.54	*.70*	(.85)		
LSI-R Attitude	.09	.09	.11	.08	.14	.14	(NA)	
LSI-R Accom.	.20	.22	.20	.10	.29	.23	*.31*	(NA)

Note. With a sample size of 238, correlations greater than .13 are significantly different from zero ($p < .05$, two-tailed). Shading indicates heteromethod blocks; the different shades distinguish the different method pairs. Underline indicates monotrait-heteromethod correlations; italics indicate heterotrait-monomethod correlations; parentheses indicate reliability coefficients.

Prior to conducting the correlation analysis, a low Pearson correlation threshold was defined as $r = .20$. For the correlations between a continuous scale (GAIN-Q, ASUS-R, and OTMA scales) and a categorical scale (LSI-R accommodation or attitude scales), polyserial correlations were computed. For correlations between two categorical scales (LSI-R accommodation with LSI-R attitude), polychoric correlations were computed. The polyserial and polychoric coefficients are estimates of what the Pearson correlation would be if both scales were continuous. The interpretation of the pattern of correlations in the matrix, according to the five requirements of construct validity posited by Campbell and Fiske (1959), is as follows:

1. Coefficients in the reliability diagonal should consistently be the highest in the matrix. The reliability coefficients (internal consistency alpha) are along the main diagonal of the matrix in parentheses. Note that internal consistency for the LSI-R traits could not be estimated because the items were not available in the study. The reliability coefficients are primarily the highest values in the matrix, though one should note that several scales had reliability coefficients that were only marginally higher than their correlations with other scales. For example, the reliability coefficient for the GAIN use scale (.90) was only slightly higher than its correlation with the dependence scale (.87).

2. Correlations for convergent validity should be significantly different from zero and sufficiently large. These are correlations of measures of the same trait (monotrait) using different methods (heteromethod). Referring to Table 10, there are four convergent validity correlations of interest (underlined): ASUS-R Involvement with GAIN-Q substance use (.39); ASUS-R disruption with GAIN-Q dependence (.48); OTMA substance use with GAIN-Q substance use (.44); and OTMA consequences with GAIN-Q dependence (.57). It is evident that the convergent validity correlations are large (using the threshold of .30), and they are all significantly different from zero. Thus, the results indicated convergent validity for both the ASUS-R and the OTMA. Note that the convergent correlations for the OTMA subscales were not significantly larger than their counterparts for the ASUS-R.

There were also monotrait-heteromethod correlations between the OTMA substance use and the ASUS-R involvement scales (.35) and between the OTMA consequence and the ASUS-R disruption scales (.54) (i.e., correlations between the same traits using different methods). These latter correlations were set aside because they are the scales being tested and compared in the current study. In other words, the GAIN-Q was treated as the reference for determining the relative construct validity of the ASUS-R and the OTMA.

3. A validity value for a variable should be higher than the correlations obtained between that variable and any other variable having neither trait nor method (instrument) in common. To evaluate this requirement, the validity correlation for a scale (monotrait-heteromethod) is compared to the correlations of that scale with other scales that measure different traits and use different methods (heterotrait-heteromethod). The monotrait-heteromethod correlation (validity coefficient) should be higher than the heterotrait-heteromethod correlations in the same block. In Table 10, the heteromethod blocks are shaded; different shades distinguish different method pairs.

The first heteromethod block includes the validity coefficients for the ASUS-R involvement and disruption scales. To test if the validity coefficients were larger than the other correlation coefficients in the same block, Steiger's (1980) approach for testing the difference between two dependent correlations was used.

The convergent validity correlation for the ASUS-R involvement scale was .39, which was not significantly different from the heterotrait-heteromethod correlation of .42 between the ASUS-R involvement scale and GAIN-Q disruption scale (t = -0.79, p = .430), and was not significantly different from the heterotrait-heteromethod correlation of .45 between the ASUS-R disruption scale and GAIN-Q use scale (t = -1.77, p = .079). These results indicated a lack of discriminant validity.

The convergent validity correlation for the ASUS-R disruption scale was .48, and it was, as expected, significantly higher than the heterotrait-heteromethod correlation of .42 between the ASUS-R involvement scale and GAIN-Q dependence scale (t = 2.18, p = .03), but not

significantly higher than the heterotrait-heteromethod correlation of .45 between the ASUS-R disruption scale and GAIN-Q use scale (t = 1.19, p = .234). These results provided limited evidence of discriminant validity, a conclusion likely impacted by the significant overlap between the use and dependence traits on all scales.

The LSI-R attitude/orientation and accommodation scales were used for an ancillary block of heterotrait-heteromethod measures. From the ASUS-R, there were four discriminant validity coefficients involving the LSI-R scales: ASUS-R involvement with LSI-R attitude/orientation (.11); ASUS-R involvement with LSI-R accommodation (.20); ASUS-R disruption with LSI-R attitude/orientation (.08); and ASUS-R disruption with LSI-R accommodation (.10). Based on the predefined threshold of .20, it is evident that these coefficients are small. However, formal tests were conducted to determine whether the differences between these correlations and the validity correlations were significantly different from zero.
Tests indicated a significant difference between the ASUS-R involvement scale validity coefficient (.39) and the correlation of .11 between the ASUS-R involvement and LSI-R attitude/orientation scales (t = 3.53, p = .001). Tests also showed a significant difference between the ASUS-R involvement scale validity coefficient (.39) and the correlation of .20 between the ASUS-R involvement and LSI-R accommodation scales (t = 2.59, p = .010).

Tests indicated a significant difference between the ASUS-R disruption scale validity coefficient (.48) and the correlation of .08 between the ASUS-R involvement and LSI-R attitude/orientation scales (t = 5.19, p < .001). Also evident was a significant difference between the ASUS-R disruption scale validity coefficient (.48) and the correlation of .10 between the ASUS-R disruption and LSI-R accommodation scales (t = 5.26, p < .001). These results suggested that the ASUS-R scales measure a distinct construct, but there are no monomethod-heterotrait correlations within the LSI-R block against which the estimates can be evaluated.

The next heteromethod block in the matrix includes the validity coefficients for the OTMA use and consequence scales. The convergent validity correlation for the OTMA use scale was .44, but this was not significantly different from the heterotrait-heteromethod correlation of .43 between the OTMA use scale and GAIN-Q dependence scale (t = 0.30, p = .764), and it was significantly lower (not higher) than the heterotrait-heteromethod correlation of .52 between the OTMA consequence scale and GAIN-Q use scale (t = -2.76, p = .006). Similar to the results for the ASUS-R scales, these results indicated a lack of discriminant validity.

The convergent validity correlation for the OTMA consequence scale was .57, and, as expected, this was significantly higher than the heterotrait-heteromethod correlation of .43 between the OTMA use scale and GAIN-Q dependence scale (t = 4.99, p < .001), but this was not significantly higher than the heterotrait-heteromethod correlation of .52 between the OTMA consequence scale and GAIN-Q use scale (t = 1.74, p = .083). These results provided limited evidence of discriminant validity.

For the OTMA, there were four discriminant validity coefficients that involved the LSI-R scales: OTMA use with LSI-R attitude/orientation (.14); OTMA use with LSI-R accommodation (.29); OTMA consequence with LSI-R attitude/orientation (.14); and OTMA consequence with LSI-R accommodation (.23). Based on the predefined threshold of .20, only the correlations between the OTMA scales and the LSI-R attitude/orientation scale qualified as small.

Tests indicated a significant difference between the OTMA use scale validity coefficient (.44) and the correlation of .14 between the OTMA use and LSI-R attitude/orientation scales ($t = 3.74$, $p < .001$). In addition, there was a significant difference between the OTMA use scale validity coefficient (.44) and the correlation of .29 between the OTMA use and LSI-R accommodation scales ($t = 2.04$, $p = .043$).

There was a significant difference as well between the OTMA consequence scale validity coefficient (.57) and the correlation of .141 between the OTMA use and LSI-R attitude/orientation scales ($t = 5.76$, $p < .001$), plus a significant difference between the OTMA consequence scale validity coefficient (.57) and the correlation of .23 between the OTMA consequence and LSI-R accommodation scales ($t = 4.91$, $p < .001$). These results suggested that the OTMA scales measure distinct factors, but there are no monomethod-heterotrait correlations within the LSI-R block to which the estimates can be compared to tease out the methods effect. Importantly, the correlations between the OTMA consequences scale and the LSI-R scales are both above the study definition of a small correlation.

4. A variable should correlate higher with an independent effort to measure the same trait than with measures of a different trait that happen to employ the same method (instrument). To meet this requirement, the validity coefficient should be higher than the heterotrait-monomethod coefficient. The heterotrait-monomethod correlations are displayed in italics above in Table 10. For example, the correlation between ASUS-R disruption and GAIN-Q dependence (monotrait-heteromethod) should be higher than the correlation between ASUS-R disruption and ASUS-R involvement (heterotrait-monomethod). Meeting this requirement provides evidence of discriminant validity. This requirement was not met for any of the study scales. In all cases, the monotrait-heteromethod correlation was significantly lower than the heterotrait-monomethod correlation. If the correlation between different traits measured with the same method is higher than the correlation between the same trait measured with different methods, this indicates a methods factor. These results also suggested a lack of discriminant validity, although (as mentioned above), the strong relationship between substance use and dependence tempers that conclusion. Note that no monotrait-heteromethod correlations were available for the LSI-R scales in the study.

5. The same pattern of trait interrelationships should be demonstrated in all the heterotrait correlations (between different instruments) on both mono- and hetero-method combinations. Although a similar pattern of correlations was observed in the heteromethod blocks, either the pattern was in the wrong direction (heterotrait-heteromethod correlation > monotrait-

heteromethod correlation) or the pattern revealed small differences between the heterotrait-heteromethod and monotrait-heteromethod correlations.

Summary

The internal consistency reliability results for the ASUS-R involvement and disruption scales were robust and indicated that these traits were more highly correlated with themselves than with other traits in the matrix. The reliability coefficient for the OTMA use scale was only .75, but the results overall indicated that both the use and consequence scales were more highly correlated with themselves than with any other trait in the matrix. The ASUS-R involvement scale demonstrated somewhat better internal consistency reliability (significantly different from zero) than the OTMA use scale.

The ASUS-R involvement and disruption scales and the OTMA use and consequences scales demonstrated good convergent validity. The convergent validity correlations of the OTMA scales with the GAIN-Q scales (.44 and .57) were somewhat larger than the corresponding correlations between the ASUS-R and GAIN-Q scales (.39 and .48), but the differences between the respective convergent validity coefficients were not significantly different from zero. The two instruments performed equally well on this requirement.

Overall, there was limited evidence of discriminant validity for either the ASUS-R involvement and disruption scales or the OTMA use and consequences scales. The main evidence of discriminant validity for the ASUS-R scales was that their respective convergent validity values were significantly higher than their respective correlations with the LSI-R attitude/orientation and accommodation scales. The OTMA discriminant validity evidence was similar, except that the correlations of the OTMA use and consequences scales with the LSI-R accommodations scale were larger than the predefined threshold of .20.

The comparison of the convergent validity values with correlations in the same heteromethod block provided modest evidence of discriminant validity for some of the study scales. The convergent validity of the ASUS-R disruption scale was higher than the heterotrait-heteromethod correlation between the ASUS-R involvement and GAIN-Q dependence scales. Likewise, the convergent validity of the OTMA consequences scale was higher than only one of the heterotrait-heteromethod correlations in its block.

For all the study scales, the heterotrait-monomethod correlations (e.g., ASUS-R involvement with ASUS-R disruption) were much higher than the convergent validity correlations. These results suggested strong methods factors for all the scales.

The high correlations between the use and dependence traits on the GAIN-Q, ASUS-R, and OTMA indicated considerable shared variance among these traits, especially on the GAIN-Q (r = .87), which was used as the primary reference scale for convergent and discriminant validity comparisons.

It was difficult to demonstrate that a scale discriminated between use and dependence because these traits were highly correlated in the study data. In addition, a floor effect was evident for most of the substance abuse measures. However, neither recalculation of the data after removal of the lowest scores nor a square root transformation changed the basic pattern of the correlations.

The ASUS-R and GAIN-Q are both self-administered and the OTMA and LSI-R are both interview-based. The shared methods variance could be higher between the ASUS-R and GAIN-Q, and this could increase the convergent validity (monotrait-heteromethod) while increasing the heterotrait-heteromethod correlations. The shared methods variance may be higher between the OTMA and LSI-R, which could increase the convergent validity (monotrait-heteromethod) and increase the heterotrait-heteromethod correlations. There also is a possible response set bias specific to the ASUS and not the GAIN-Q, which could attenuate the convergent validity correlations involving these two instruments.

Colorado Department of Corrections
Office of Planning & Analysis

2862 South Circle Drive
Colorado Springs, CO 80906
http://www.doc.state.co.us

Cover designed by Camille S. Rhoades