Spring 8-3-2018

EXPLORING CLINICALLY USEFUL DEFINITIONS OF TREATMENT SUCCESS FOR INDIVIDUALS WITH ALCOHOL USE DISORDER

Megan Kirouac
University of New Mexico - Main Campus

Megan Kirouac_____
Candidate

Psychology_____
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Katie Witkiewitz, PhD, Chairperson___

Theresa B. Moyers, PhD_____

Matthew R. Person, PhD_____

Dennis M. Donovan, PhD_____

EXPLORING CLINICALLY USEFUL DEFINITIONS OF TREATMENT SUCCESS FOR INDIVIDUALS WITH ALCOHOL USE DISORDER

by

MEGAN KIROUAC

B.S., Psychology, University of Washington, 2010

M.S., Psychology, University of New Mexico, 2014

DISSERTATION

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctorate of Philosophy in Psychology

The University of New Mexico

Albuquerque, New Mexico

July, 2018

ACKNOWLEDGMENTS

I would like to thank Dr. Katie Witkiewitz, my advisor and dissertation committee chair, for her support, advice, and encouragement throughout the process of conducting my dissertation. I would also like to thank Drs. Moyers, Pearson, and Donovan for their guidance as members on my dissertation committee as well as the National Institute on Alcohol Abuse and Alcoholism for supporting this work via an F31 dissertation award. Additionally, I would like to thank members of my lab for their assistance with preparing the data that were used for my dissertation as well as their encouragement and peer mentorship. Finally, I would like to thank my family and friends for their continued support over the years.

EXPLORING CLINICALLY USEFUL DEFINITIONS OF TREATMENT SUCCESS FOR INDIVIDUALS WITH ALCOHOL USE DISORDER

by

MEGAN KIROUAC

B.S., Psychology, University of Washington, 2010

M.S., Psychology, University of New Mexico, 2014

Ph.D., Clinical Psychology, University of New Mexico, 2018

ABSTRACT

Given the widespread costs associated with alcohol use disorder (AUD; World Health Organization, 2011), it is unsurprising that many treatments exist for AUD. Moreover, many treatments have been rigorously studied via experimental research designs. In such research, treatment success has been defined predominantly as abstinence from alcohol or, more recently, no heavy drinking days. Consumption-based definitions of treatment success, rather than alternative non-consumption based definitions, have dominated in the field for at least two reasons. First, there are multiple measures of similar non-consumption constructs (e.g., quality of life, psychosocial functioning), and very little research has been conducted to direct researchers toward the best non-consumption measures to use among AUD populations. Second, it is assumed that non-consumption measures are insensitive and, therefore, consumption must be used as a surrogate measure for more clinically meaningful non-consumption measures. The present research study

empirically addressed these two barriers that have thwarted attempts to shift toward including non-consumption variables in our definitions of treatment success. Using secondary data analysis of data collected from the COMBINE Study (Anton et al., 2006) and Project MATCH (Project MATCH Research Group, 1997), the present study conducted several tests of measurement stability, reliability, validity, sensitivity, and specificity. To test measurement stability the current study examined effect sizes and measurement invariance across time to test if non-consumption measures may be viable options for comparing pre- and post-treatment scores on these measures. The present study also conducted analyses on psychometric properties of extant measures: internal consistency reliability, construct validity, convergent validity. Finally, receiver operating characteristic curve analyses were conducted of total scale scores, subscales, and individual items when available and appropriate to test the sensitivity and specificity of non-consumption measures in detecting post-treatment and 12-month outcomes. The Brief Symptom Inventory (BSI), Beck Depression Inventory (BDI), and the brief World Health Organization Quality of Life measure (WHOQOL-BREF) were invariant across time and performed the best overall across all psychometric and sensitivity/specificity analyses conducted in the present manuscript. All other measures examined in the current study had at least some promising results, with the sole exception of the Addiction Severity Index (ASI), which had weak findings across all analyses. Moreover, some non-consumption measures (e.g., Drinker Inventory of Consequences, Obsessive-Compulsive Drinking Scale) had baseline to post-treatment effect sizes as large as some consumption-based outcome effect sizes. The results of the present study have identified gold standard measures for assessing mental health and quality of life. Future research should use the

BSI, BDI, and WHOQOL-BREF to examine clinically-relevant changes beyond consumption outcomes. The present findings also indicate that consumption measures may not be needed to serve as surrogates for these clinically relevant constructs. These findings represent the possibility of a paradigm shift in the field of AUD treatment research evaluation to incorporate non-consumption outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

Background

Alcohol misuse causes significant problems worldwide and affects the lives of millions of people. Recent estimates suggest excessive alcohol consumption and alcohol-related problems comprise the third highest risk for disease and disability worldwide (World Health Organization, 2011). The prevalence of current (i.e., past twelve months) alcohol use disorder (AUD) in the United States has been estimated recently to be 13.9% (Grant et al., 2015). Given the prevalence of AUD and related public health concerns, research has focused on the development and evaluation of psychosocial and pharmacologic treatments for AUD.

Historically, success in AUD treatment trials has been defined primarily by abstinence (binary outcome) or percentage of days abstinent (PDA, continuous outcome; Food and Drug Administration (FDA), 2006). The FDA (2015) recently proposed percent subjects with no heavy drinking days (PSNHDD), in addition to abstinence, as primary endpoints for evaluating AUD treatment trials, with "no heavy drinking days" defined as no days with 4/5 or more standard drinks for women/men (FDA, 2015). Other AUD treatment outcome definitions that are commonly used include: number of drinks per drinking day (DDD; e.g., Greenfield, 2000), drinks per day (DPD; e.g., Morgenstern et al., 2007); and frequency of heavy drinking (percent heavy drinking days, PHDD; e.g., Anton et al., 2006). Importantly all the outcomes commonly used to define treatment success (PDA, PSNHDD, DDD, DPD, PHDD) are consumption-based measures; thus, treatment success is often solely defined by whether a client is still drinking and/or how much alcohol an individual drinks.

Alcohol consumption, variously defined, has been the dominant outcome variable in AUD treatment research for a number of reasons. First, consumption variables are easily quantified using standard drinks, and a variety of consumption variables can be examined, including dichotomous (e.g., abstinent or not) and continuous variables (e.g., PDA, DDD). Second, alcohol consumption is tied inherently to the development of AUD—without consuming alcohol, it is impossible for one to develop AUD. Third, it has been argued that alcohol consumption is likely a surrogate measure for how an individual is functioning (FDA, 2015). However, relying upon consumption-based variables as the sole markers of treatment success is limited in substantial ways.

Consumption-based outcome variables often fail to acknowledge the complex processes underlying the development, maintenance, and recovery from AUD. Recovering from addiction is more complex than simply abstaining from substance use, and defining treatment success purely by consumption often fails to adequately portray the complex, multifaceted recovery process (e.g., Donovan et al., 2012; Substance Abuse and Mental Health Services Administration, 2012; Tiffany et al., 2012). Cisler and Zweben (1999, 2003) attempted to address the need for a more complex representation of outcomes and created a composite measure to reconcile consumption with alcohol-related problems. This "composite clinical outcome" measure had four levels: (1) abstinence; (2) moderate drinking (<4 drinks for females; <6 drinks for males) without problems (drinking consequences occurred never or only once or twice); (3) heavy drinking (3+ occasions of 4+ drinks for females and 6+ drinks for males) or problems (recurrent drinking consequences occurring 3+ times); and (4) heavy drinking with problems.

Despite studies validating the composite clinical outcome measure (Cisler & Zweben, 1999, 2003), this measure has never been widely adopted in the field. Recently, Kaskutas and colleagues (2014) developed a measure to identify specific components that were most important to how clients and their loved ones evaluated whether or not a client's AUD had improved. Initial item testing and factor analysis indicated clients and their loved ones viewed a variety of non-consumption variables as important, including functioning and consequences (Kaskutas et al., 2014). Similarly, Neale and colleagues (2014) recently collected qualitative data to examine how treatment providers define treatment success. Findings suggest a broad range of outcomes are meaningful to treatment providers, including psychological and physical health, social functioning, and well-being (Neale et al., 2014). Thus, defining treatment success solely by consumption is an inaccurate definition of recovery from multiple clinically important perspectives.

Defining treatment success by non-consumption outcome variables may also be more consistent with the variety of theoretical models of addiction that underlie AUD treatment development. Various theories (e.g., cognitive theory versus behavioral theory) posit different key outcomes (e.g., changes in thoughts versus behavior) and examining all AUD treatments primarily by consumption (e.g., PDA, PSNHDD), regardless of underlying theory or hypothesized mechanisms of change is inconsistent with the myriad of addiction models. Moos and Finney (1983) criticized this inconsistency and noted that theory should be used to guide treatment evaluation. Similarly, addictions researchers have called for the evaluation of theory-specific outcomes in substance use treatment research rather than a single, one-size-fit all outcome variable (Del Boca & Darkes, 2012; Donovan et al., 2012). Further, Moos and Finney (1983) called for a greater

acknowledgment of the complexities associated with AUD in AUD treatment research (i.e., AUD is not simply a phenomenon of using too much alcohol but rather one of consequences incurred by such alcohol use). In sum, the AUD treatment research community has recognized that AUD is complex and that research must account for such complexities to be consistent with theory in evaluating treatment. This recognition means moving beyond the singular approach of using consumption-based definitions of treatment success.

Despite these arguments for shifting away from consumption as the sole index of AUD treatment success to more clinically and theoretically useful non-consumption measures, consumption outcomes have remained dominant in AUD treatment research. Efforts to incorporate non-consumption outcome measures into AUD treatment research have been stymied for at least two reasons. First, there are multiple measures of similar non-consumption constructs (e.g., quality of life), and research is needed to direct researchers toward the "gold standard" (i.e., psychometrically sound) measures that are viable for use among AUD populations (e.g., Del Boca & Darkes, 2012). Second, it is assumed that non-consumption measures are insensitive and, therefore, consumption must be used as a "surrogate" measure for more clinically meaningful non-consumption measures (FDA, 2015, p. 2). However, this assumption has not been subjected to empirical testing and some research has found non-consumption measures (e.g., temptation) to better predict AUD treatment outcomes (quantity, frequency, and alcohol-related problems) than consumption-based measures (Witkiewitz, 2013).

Present Study

The present study consisted of extensive secondary data analyses to evaluate and compare psychometric properties and the sensitivity/specificity of clinically meaningful non-consumption outcome variables (e.g., quality of life) for evaluating AUD treatment. To this end, the present study had two primary aims. Aim 1 was to examine the psychometric properties of several non-consumption self-report measures in order to explore the viability of these measures as potential "gold standard" measures to compare pre- and post-AUD treatment changes in these constructs. Accordingly, the present study conducted several tests of measurement stability, validity, and reliability. To test measurement stability the current study examined effect sizes and measurement invariance across time to test if non-consumption measures may be viable options for comparing pre- and post-treatment scores on these measures. Construct validity was examined via confirmatory factor analyses; convergent validity was examined via bivariate correlations with measures hypothesized to be related. These results informed further measure psychometric evaluation via examination of internal consistency total scale scores and sub-scale scores upheld via CFA and invariance testing. Aim 2 was to further test the viability of these non-consumption measures by examining total scale, subscale, and individual item sensitivity/specificity in an incremental approach based on CFA and invariance testing results and levels of sensitivity/specificity for each higher-level score (i.e., individual items were only examined if subscales had adequate sensitivity/specificity). Together, results from Aims 1 and 2 highlighted non-consumption outcome measures that may be most appropriate for use in AUD treatment research contexts to define treatment success in clinically meaningful ways.

Methods

Data

The present study used data collected from the COMBINE Study (Anton et al., 2006) and Project MATCH (Project MATCH Research Group, 1997). Table 1 summarizes the participant demographics, design, and exclusion criteria used in these two studies. Psychometric, measurement invariance, and sensitivity/specificity analyses were conducted using measures that were consistent with the variables previously identified as important by researchers, clients and their loved ones, and treatment providers (Donovan, et al., 2012; Kaskutas et al., 2014; Neale et al., 2014) and included: 1) drinking consequences/severity, 2) mental health, 3) craving/temptation, 4) quality of life/functioning. Table 2 details the measures used in these analyses and Figure 1 summarizes the analyses conducted. These non-consumption variables were comprised of full-measure information (e.g., Drinker Inventory of Consequences (DrInC) total summary score; Miller, Tonigan, & Longabaugh, 1995) as well as sub-scale data when available (e.g., factor-analytically supported subscales of the DrInC) and individual-item analyses where applicable (e.g., individual items of the DrInC if DrInC subscales performed adequately).

COMBINE. COMBINE (N = 1383) was a large, multisite, randomized controlled trial of 9 treatment combinations of psychosocial interventions (Combined Behavioral Intervention (CBI) or Medication Management (MM)) and medication (acamprosate, naltrexone, or placebo). Assessments were conducted at baseline (i.e., pre-treatment), during treatment, and post-treatment follow-ups at 10-weeks (immediately post-treatment = "week 16" post-baseline), 9-months, and 12-months. Participants were all seeking treatment and were recruited from 11 research sites across the United States. Although

there were relatively few exclusion criteria (presented in Table 1: history of other substance use disorder except cannabis, psychiatric diagnoses requiring medication, unstable medical conditions), there were strict inclusion criteria in the COMBINE Study (Anton et al., 2006). All eligible participants must have 1) met criteria for Alcohol Dependence per the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994), 2) have had between 4 and 21 days of abstinence prior to their baseline assessment session, and 3) have consumed > 14/21 drinks per week (for women/men) with at least 2 heavy drinking days ($\geq$ 4/5 drinks for women/men) within a consecutive 30 day period in the 90 days preceding their baseline assessment. These specific inclusion criteria resulted in greater homogeneity of alcohol consumption and problem severity in the COMBINE Study than those in Project MATCH. The sample homogeneity in COMBINE was intentional with the study design since the COMBINE Study was a pharmacotherapy study and selective recruitment was necessary to reduce medication complications.

Project MATCH. Project MATCH (N = 1726) was a large, multisite, randomized controlled trial of psychosocial treatments (Cognitive Behavioral Therapy (CBT), Motivational Enhancement Therapy (MET), or Twelve-Step Facilitation (TSF)). Participants were all seeking treatment and were provided 12 weeks of treatment. Assessments were conducted at baseline (i.e., pre-treatment), during treatment, and every three months post-treatment for up to 12 months. Participants were recruited from 9 research sites across the United States. Participants received these treatments in either an aftercare arm (after release from an inpatient treatment) or an outpatient treatment arm. Participants were substantially different between these two treatment arm settings and

had significantly different outcomes post-treatment (Project MATCH Research Group, 1998). Sample heterogeneity in Project MATCH was deliberate in the study design because the primary aim of MATCH was to find better treatment approaches to target client heterogeneity. A primary difference that resulted in differences in participant homogeneity between the COMBINE Study and Project MATCH was that Project MATCH inclusion criteria were DSM-III-R (American Psychiatric Association, 1987) diagnosis of alcohol abuse *or* dependence, whereas the COMBINE Study limited their sample to DSM-IV diagnoses of alcohol dependence only and had additional alcohol consumption inclusion criteria that were not paralleled in Project MATCH (Project MATCH Research Group, 1997).

Measures

Alcohol-Related Variables. Both COMBINE and MATCH employed the Form 90 (Miller, 1996) to collect 90-day assessment window information on daily drinking levels. From these data, multiple consumption outcome variables were computed: number of drinks per drinking day (DDD, including only days when alcohol was consumed), maximum number of drinks consumed in the 90-day window (MXD), drinks per day in the assessment window (DPD, averaging across drinking and abstinent days), binary heavy drinking (HD) data for each day, percent days abstinent (PDA), percent heavy drinking days (PHDD), World Health Organization risk levels (WHO risk levels; WHO, 2000), and composite clinical score (abstinent, abstinent or moderate drinking without problems, heavy drinking/problems, or heavy drinking and problems; Cisler & Zweben, 1999). WHO Risk levels included low risk (<20/40 grams of alcohol for women/men per day), medium risk (<40/60 grams of alcohol for women/men per day), high risk

(<60/<100 grams of alcohol for women/men per day) or very high risk (>60/100 grams of alcohol for women/men per day). Standard drinks were calculated using National Institute on Alcohol Abuse and Alcoholism guidelines of 14 grams. "Heavy drinking" was defined as 4/5 or more standard drinks for women/men (HD; NIAAA, 2004). The composite clinical score also used information collected in COMBINE and MATCH via the Drinker Inventory of Consequences (DrInC; Miller et al., 1995), a 45-item measure of alcohol-related consequences (plus 5 control-scale items not included in the present study analyses) on which higher scores indicated greater alcohol-consequence severity. The DrInC was initially conceptualized as containing 5 consequence factors: Interpersonal, Intrapersonal, Impulse Control, Physical, and Social Responsibility (Miller et al., 1995). It was according to these 5-factors that an abbreviated version of the DrInC was created: the Short Inventory of Problems (SIP; Feinn, Tennen, & Kranzler, 2003). Previously published findings have reported sufficient internal consistency reliability as well as convergent validity of the DrInC and the SIP in COMBINE and MATCH (α range from 0.61 to 0.87 with the DrInC generally having higher internal consistency than the SIP; Forcehimes, Tonigan, Miller, Kenna, & Baer, 2007). Moreover, Marra and colleagues (2014) found strict measurement invariance between Spanish and English speakers for the SIP. However, multiple publications have proposed alternative factor structures, including 3- and 1-factor models for the DrInC and the SIP, which may indicate instability of previously examined factor solutions or poor construct validity of the DrInC and SIP administrations (e.g., Alterman, Cacciola, Ivey, Habing, & Lynch, 2009; Feinn et al., 2003; Hagman et al., 2009; Kenna et al., 2005). Of particular importance to the current analyses, in Project MATCH the DrInC was non-uniformly administered to

individuals who reported 100% days of abstinence during the follow-up assessments. Specifically, the items on the follow-up version of the DrInC are worded such that items should be endorsed only in reference to consequences that occurred due to drinking during the assessment window and some assessors in MATCH did not administer the DrInC to some, but not all, of the individuals who were abstinent at follow-up. This inconsistent administration of the DrInC in Project MATCH may have important impacts on how well it performs psychometrically across COMBINE and MATCH.

In addition to alcohol-related consequences, alcohol dependence severity was assessed in MATCH via the Addiction Severity Index (ASI; McLellan, Luborsky, Woody, & O'Brien, 1980). The ASI was created with a conceptualization of having 6-7 factors: Medical Status, Employment/Social Support, Alcohol/Drug Use (sometimes conceptualized as separate factors), Legal Status, Family/Social, and Psychiatric Status (McLellan et al., 1992). However, other factor structures have also been published (e.g., Currie, El-Guebaly, Coulson, Hodings, & Mansley, 2004; Rogalski, 1987). Administrations of the ASI have found poor to good internal consistency of each of these factors (Currie et al., 2004) and other publications have cautioned against the use of the ASI as a research or diagnostic instrument (DeJong, Willems, Schippers, & Hendriks, 1995). Project MATCH did not administer the full ASI that has been factor analyzed in previous studies and only included a partial set of items.

Also included in COMBINE and MATCH were measures of alcohol temptation/craving. In COMBINE, temptation/craving was measured by the Obsessive-Compulsive Drinking Scale (Anton, 2000). The OCDS has been widely studied and conceptualized as a measure of alcohol craving (Anton, 2000) where higher scores

indicated greater alcohol craving, but various publications have found differing factor structures (e.g., Bohn, Barton, & Barron, 1996; Connor, Jack, Feeney, & Young, 2008; Connor, Feeney, Jack, & Young, 2010; Kranzler, Mulgrew, Modesto-Lowe, & Burleson, 1999; Roberts, Anton, Latham, & Moak, 1999). In addition to unclear construct validity regarding differing published factor analytic results, there is mixed evidence of the convergent validity of administrations of the OCDS (e.g., Anton, Moak, & Latham, 1996; Connor et al., 2008; Moak, Anton, & Latham, 1998). Similarly, various administrations of the OCDS have yielded variable internal consistency of the overall measure and its factor analyzed subscales (e.g., Bohn et al., 1996; Kranzler et al., 1999). A slightly less-studied measure that has been purported to measure temptation/craving is the Alcohol Abstinence Self-Efficacy Scale (AASE; DiClemente, Carbonari, Montgomery, & Hughes, 1994), which was administered in COMBINE and MATCH. The items in the AASE are worded to assess temptation/craving through the confidence an individual has in being able to avoid drinking in various circumstances. Therefore, lower AASE scores indicated higher temptation/craving to drink. Preliminary studies have identified the AASE as consisting of 4 factors related to situations in which individuals may be tempted to drink: Negative Affect, Social/Positive, Physical & Other Concern, Withdrawal or Urges (e.g., DiClemente et al., 1994; Hiller et al., 2000). Administrations of the AASE have demonstrated strong internal consistency reliability and modest convergent validity of total AASE score and each of the 4 subscales (DiClemente et al., 1994). In addition to the AASE, an individual item assessing overall temptation/craving was administered in MATCH.

A final alcohol-related measure used in MATCH was the Alcoholics Anonymous Involvement scale (Tonigan, Connors, & Miller 1996). The AAI assesses for attendance of AA meetings as well as involvement with each of the 12-steps of AA. This measure was examined in the present study as a means of examining convergent validity of items hypothesized to be negatively or positively correlated with AA involvement.

Mental Health Variables. Mental health was assessed via multiple assessment measures in COMBINE and MATCH. In COMBINE, mental health was assessed via the Brief Symptom Inventory (BSI; Derogatis & Melisaratos, 1983), which has been conceptualized as assessing 9 domains of mental health as well as overall global mental health problem severity; higher scores indicated greater mental health problem severity. These 9 domains have been upheld via numerous factor analyses, including analyses of the brief version of the BSI (BSI-18; e.g., Derogatis & Melisaratos, 1983; Long, Harring, Brekke, Test, & Greenberg, 2007; Recklitis et al., 2006; Wang et al., 2010). The BSI and BSI-18 have been administered in numerous mental health treatment studies and their psychometric properties (internal consistency, convergent validity) have been supported in multiple administrations of the measures.

A similarly well-studied measure of mental health is the Beck Depression Inventory (Beck, Steer, & Brown, 1996; Beck, Steer, & Garbin, 1988), which was administered in Project MATCH. The BDI and the second edition BDI-II measure depression symptoms, and higher levels indicated higher depression. The BDI and BDI-II have both been found to have either a 2- or a 3-factor structure and strong internal consistency and convergent validity in numerous measurement administrations (e.g.,

Arnau, Meagher, Norris, & Bramson, 2001; Beck et al., 1988; Visser, Leentjens, Marinus, Stiggelbout, & van Hilten, 2006).

In addition to depression, mental health as a construct was also measured in Project MATCH via a state-trait anger expression inventory: the Spielberger State-Trait Inventory (SSTI; Forgays, Forgays, & Spielberger, 1997). Only a subset of items of the SSTI were administered in Project MATCH. Though few studies have been published regarding the psychometric properties of the SSTI, extant literature suggests the SSTI items administered in Project MATCH consist of two factors: a Temperament and a Reaction factor (Forgays et al., 1997; Kroner & Reddon, 1992; van der Ploeg, 1988). Previous findings have also indicated administrations of the SSTI have had at least acceptable internal consistency and convergent validity (Forgays et al., 1997; Kroner & Reddon, 1992; van der Ploeg, 1988). The SSTI was only administered at the baseline timepoint in MATCH.

Quality of Life/Functioning Variables. Quality of life was assessed in COMBINE via the World Health Organization Quality of Life, brief measure (WHOQOL-BREF; WHOQOL Group, 1998) and the Health Survey (SF-12; Ware, Kosinski, & Keller, 1996). The WHOQOL-BREF is an abbreviated version of the 100-item WHOQOL where higher scores indicate better quality of life. Previous publications have identified the WHOQOL-BREF as comprised of a higher-order factor structure containing 4 lower-order factors (Physical Health, Psychological Health, Social Relationships, and Environment) and a higher-order Quality of Life factor (Skevington, Lofty, & O'Connell, 2004). One item of the WHOQOL-BREF assessing negative affect was erroneously omitted in administration of the WHOQOL-BREF in COMBINE.

Previous administrations of full version of the WHOQOL-BREF have been well-studied and have yielded good internal consistency, convergent validity, and measurement invariance across demographic groups (e.g., Jaracz, Kalfoss, Gorna, & Baczyk, 2006; Skevington et al., 2004; Yao & Wu, 2005).

The SF-12 is an abbreviated version of the SF-36, and has also been widely studied. However, unlike the WHOQOL-BREF there have been mixed findings regarding the psychometric properties of various SF-36 and SF-12 administrations (e.g., Hann & Reeves, 2008; Jakobsson, Westergren, Lindskov, & Hagell, 2012; Treanor & Donnelly, 2015). The SF-12 is most often conceptualized as consisting of 2 factors: Physical Health and Psychological Health. Importantly, many of the items are "double-barreled," meaning that a single item asks about both physical and mental health (e.g., "...how much of the time has your physical or emotional problems interfered with…"), which may mean participants are responding in different ways to the same item. The double-barreled nature of many of the SF-12 items may explain why construct validity of SF-12 administrations have been variable across studies (Miller et al., 2009). Other psychometric properties of SF-12 administrations have been more consistently strong. The convergent validity of the SF-12 has been supported in previous administrations (e.g., Salyers, Bosworth, Swanson, Lamb-Pagone, & Osher, 2000) as well as the internal consistency (e.g., Montazeri, Vahdaninia, Mousavi, & Omidvari, 2009). Both the WHOQOL-BREF and the SF-12 are purported to measure non-disease-specific quality of life; however, neither have been extensively examined in samples of individuals with AUD.

In Project MATCH, quality of life/psychosocial functioning was measured via the Psychosocial Functioning Inventory (PFI; Feragne, Longabaugh, & Stevenson, 1983). As described by Feragne and colleagues (1983), the PFI consists of 10 subscales and 2 composite scales and higher scores on the PFI reflect better psychosocial functioning. An abbreviated version of the PFI was administered in Project MATCH and was coded according to three subscales: Subjective Role Performance, Overall Social Role Performance, and Housemate/Roommate Role (Project MATCH Research Group, 1997). Psychometric properties of the PFI have not been well studied in either the full or abbreviated forms.

A final metric for functioning that was used in COMBINE and MATCH consisted of items that assessed employment status and income (ESI). These items were included in the present study to examine convergent validity of other assessment tools. It was hypothesized that individuals who were functioning less well would have poorer employment status and lower income than individuals who were functioning well. Only a single, categorical item was used for employment status in COMBINE and MATCH and income was assessed in COMBINE but not MATCH. Further, the employment status item had to be re-coded in COMBINE and MATCH to facilitate more meaningful categories for analyses. Specifically, the COMBINE and MATCH employment items were recoded to represent increasing levels of employment: unemployment or disabled = 0; homemaker, part-time employed, or retired = 1; and full-time employed = 2. COMBINE also included one item for income that was not paralleled in MATCH (< $15,000; $15,000 - $29,999; $30,000 - $59,000; $60,000 - $89,000; > $90,000).

Aim 1 Analyses: Psychometric Properties

Extensive psychometric evaluation was conducted to help identify "gold-standard" measures for non-consumption outcomes. All psychometric analyses were conducted in SPSS version 23 (IBM Corp, 2015) and Mplus version 7.3 (Muthén & Muthén, 2012). Missing data were handled with maximum likelihood estimation, multiple imputation, or mean-and-variance adjusted weighted least squares (WLSMV) estimation as recommended by Kline (2011) and described in detail below. Descriptive statistics (mean, standard deviation, frequencies) were computed in SPSS and then effect sizes (Cohen's d; Cohen, 1988) were calculated per Lenhard and Lenhard (2016) to adjust for differences in sample sizes caused by attrition. *A priori* cutoffs for effect sizes were: large effect sizes $d > 0.8$, medium effect sizes $0.8 > d > 0.2$, small effect sizes $d < 0.2$ (Cohen, 1988).

Although work has already been done to examine various psychometric properties of some of these measures (e.g., Forcehimes et al., 2007), some non-consumption measures have not been evaluated in AUD-specific samples (e.g., the construct validity of the WHOQOL-BREF) and the present analyses were more comprehensive than previous studies. The present analyses included examinations of the following for every non-consumption measure specified in Figure 1: effect sizes, internal consistency reliability, convergent validity, construct validity (via confirmatory factor analyses), and measurement invariance across time. Few studies have examined the measurement invariance of non-consumption measures across time. Identifying measures that are invariant over time is critical for advancing non-consumption outcome measures that may be used to evaluate AUD treatment outcomes, assuming the changes from baseline to end

of treatment reflect true changes in the construct of interest and not changes in the measurement over time.

Construct Validity and Measurement invariance. In the present study, the construct validity of non-consumption measures was examined via confirmatory factor analysis (CFA) in the COMBINE Study and Project MATCH using baseline data to maximize sample size. CFA analyses were guided initially by factor structures that have been previously examined in prior studies. Data screening was conducted via SPSS version 23 (IBM Corp, 2015) to examine potential problems with the data prior to all analyses (e.g., nonnormality and outliers; Jackson, Gillaspy, & Purc-Stephenson, 2009). Specifically, measures with ordered categorical response options may yield data with rarely endorsed response categories and needed to be identified in data screening prior to CFA analyses.

Although some have argued that factor analyses should maximize the ratio of participants to parameters to assure model stability (e.g., Gorsuch, 1983; Streiner, 1994), others have recommended the use of random split-half designs to test and replicate factor structures (Floyd & Widaman, 1995). Accordingly, CFAs were conducted using randomly split-half samples in Project MATCH and the COMBINE Study. The first half of the sample was used to find a model with acceptable model fit (defined below); the second half was used to replicate the model in an independent sample. Data were split randomly via SPSS version 23 (IBM Corp, 2015). Moreover, demographic differences by treatment site in MATCH and COMBINE were accounted for via clustering by treatment site in all CFA and measurement invariance analyses as recommended by Heck and Thomas (2009). Treatment site was accounted for using a sandwich estimator to calculate

the standard errors as recommended by Muthén and Muthén (2012) for handling complex survey data. The use of sandwich estimators to calculate standard errors is an alternative to multilevel modeling approaches for accounting for treatment site effects in complex survey data (Muthén & Muthén, 2012).

Hu and Bentler (1998) recommended evaluating CFA fit based on indices that have different properties such as incremental fit and residual-based fit. In the present study, model fit was examined via the comparative fit index (CFI), the Tucker-Lewis Index (TLI), and the root-mean-square error of approximation (RMSEA). The CFI and TLI are indices of incremental and relative fit whereas RMSEA is a residual-based fit index. Several researchers have recommended the CFI as an alternative to other fit indices such as the chi-square test of fit that are easily influenced by sample size (e.g., Floyd & Widaman, 1995). Marsh and colleagues (1988) recommended the TLI as a measure of relative fit that is robust to effects of large sample sizes based on the results of their Monte Carlo simulation study. Steiger and Lind (1980) provided justification for using RMSEA to evaluate model fit because it has a known distribution and is robust to problems associated with model complexity. Although some have advised against the use of "rules of thumb" for model fit (e.g., Marsh, Hau, & Wen, 2004; Yuan, 2005), others have argued that *a priori* fit indices cutoffs are important to retain objectivity in model evaluation (Jackson et al., 2009). Accordingly, *a priori* cutoffs for the above fit indices were used, as informed by Hu & Bentler (1999) and Browne & Cudeck (1993) in order to minimize Type I and Type II error rates and reflect good model fit: CFI $\geq$ 0.95; TLI $\geq$ 0.95; RMSEA $\leq$ 0.06. Acceptable model fit *a priori* cutoffs were CFI $\geq$ 0.90; TLI $\geq$ 0.90; RMSEA $\leq$ 0.08. Fit indices outside of these cutoffs were deemed inadequate.

Inadequate model fit statistics in the CFAs prompted exploration of alternative factor structures. First, individual items within each measure were examined to identify common themes of question items. If face-valid themes were identifiable in this method, alternative factor structures were tested using CFA and the methodology described above regarding split-half and *a priori* fit cutoffs. If these models failed to provide adequate fit or there were no easily identifiable themes across items, exploratory factor analyses (EFA) were employed to examine alternative factor solutions. As recommended by Floyd and Widaman (1995), EFA using principal factor analysis (PFA) was used to explore the relationships among observed variables relative to underlying latent variables. We anticipated non-zero correlations among the latent factors for the analyzed measures, thus oblique rotation methods were used to allow for correlations between factors (e.g., geomin rotation; Floyd & Widaman, 1995).

The number of factors to be tested via CFA, based on EFA results, was determined by parallel analysis and the scree plot. As recommended by several publications (e.g., Floyd & Widaman, 1995; Gorsuch, 1983; Zwick & Velicer, 1986), the elbow in the scree plot of the eigenvalues was used to indicate the number of factors. If there were multiple points that could constitute an "elbow," alternative factor solutions were tested at each of these potential elbows. Notably, visual inspections of the scree plot of EFA results have demonstrated satisfactory performance and are generally less biased than reliance upon the Kaiser-Guttman rule of retaining all factors with eigenvalue > 1, especially when combined with a parallel-analysis of the eigenvalues against what might be expected by chance alone (Montanelli & Humphreys, 1976). Accordingly, scree plot and parallel analysis guided factor enumeration for EFAs. These factor solutions were

then tested via CFA as described above. Since items are allowed to be explained by multiple factors in EFA, item-factor assignment was based on the factor on which the item loaded strongest or where the item made most conceptual sense. However, if an item failed to load > 0.40 on any factor, that item was omitted from further analyses given research on the instability of factor solutions derived using items that load with factor loadings < 0.40 (Guadagnoli & Velicer, 1988). When models informed thusly by EFA results failed to provide adequate fit in CFA as defined by *a priori* cutoffs above, factor analyses were ceased for that measure and invariance testing was not pursued.

When an adequately fitting factor-solution was found and replicated in independent split-half sub-samples, measurement invariance across time was tested by examining nested models between baseline and post-treatment datasets. Measurement invariance over time was tested for possible non-equivalence of measurement parameters (e.g., item intercepts, item loadings) over time (Widaman et al., 2010). Specific procedures to test longitudinal measurement invariance followed the recommendations of Vandenberg and Lance (2000) based on the results of their literature review. First, an omnibus test of the equality of covariance matrices across time was tested. Next, configural invariance was tested wherein the overall factor structure is tested as equivalent across time (Horn & McArdle, 1992). Then, metric invariance was tested by constraining the factor loadings to be equivalent across time (Horn & McArdle, 1992). Next, thresholds were constrained to equality across time to establish scalar invariance (i.e., "strong invariance"). Since Widaman and colleagues (2010) argued that strong factorial invariance must be held across time to identify a consistent latent construct, analyses attempted to test at least partial scalar invariance by allowing some of the

constraints to be freed. Decisions to free constraints were based on a combination of what made conceptual sense to free based on item question content and modification indices (Vandenberg & Lance, 2000). Residual invariance (i.e., "strict invariance;" Widaman et al., 2010) was not tested in the current study because all measures had categorical items (mostly Likert-type scales) and residuals were necessarily constrained to 1 for model identification. Thus, additional equality constraints could not be examined.

In the cases where a measure had a higher-order factor structure, analyses followed the procedures for testing measurement invariance as specified by Chen and colleagues (2005). In this procedure, configural invariance across time of the full model was tested. Next, invariance of factor loadings in only the lower-order factor level was tested. Then, invariance of factor loadings in both the lower- and higher-order factor levels were analyzed. Once configural invariance was established, analyses included the additional constraint of equal intercepts of the observed variables across time. Next, analyses included the additional constraint of the intercepts of the lower-level factors to equivalence and then constrained the disturbances of the lower-level factors to be equivalent across time.

To determine if a more stringent level of invariance fit significantly worse than a less stringent level of invariance (i.e., to determine the level of invariance or non-invariance), analyses used the recommendations of Widaman and colleagues (2010) and synthesized the information provided by the fit indices (specifically, the CFI, TLI, and RMSEA). As noted by Cheung and Rensvold (2002) among others, chi-square difference testing is influenced by large sample sizes and may be too sensitive for measurement invariance testing with large samples like those in COMBINE and MATCH. Thus, chi-

square difference testing was not used in the present study. Instead, measurement invariance results were evaluated based on changes in CFI, TLI, and RMSEA or inadequate CFI, TLI, and/or RMSEA fit indices as indicators of poorer model fit across time (i.e., longitudinal measurement non-invariance). Many recommend against using "rules of thumbs" for fit statistics and overall change in fit across CFI, TLI, and RMSEA together was considered for determining measurement invariance or non-invariance (e.g., Vandenberg & Lance, 2000; Widaman et al., 2010). Some, however, suggest that a change (decrease) in CFI and TLI of greater than .01 and .0.5, respectively, from one level of invariance to another indicates the factor structure may be non-invariant (Cheung & Rensvold, 2002). Accordingly, these rules of thumbs were considered in evaluating change in fit statistics, but were not held as the sole determinants of non-invariance.

Reliability. Internal consistency reliability was examined via Cronbach's alpha of both total scale scores (e.g., DrInC summary score) and subscale scores (e.g., subscales of the DrInC) when available and applicable (i.e., when subscales were verified via factor analyses). Cronbach's alpha values closer to 1 indicated better internal consistency.

Convergent Validity. Convergent validity was examined via bivariate correlations of a given measure with measures that purport to measure similar and opposite constructs. See Table 3 for measures that were hypothesized to possess convergent validity (i.e., predicted significant positive or negative correlations with conceptually similar or dissimilar constructs).

Aim 2 Analyses: Sensitivity/Specificity

Receiver Operating Characteristic (ROC) curve analyses. Secondary data analyses of the COMBINE and MATCH data were conducted to examine the sensitivity

and specificity of the non-consumption outcome variables using ROC curve analyses (Hanley & McNeil, 1982). ROC curve analyses stem from signal detection theory where "sensitivity" is the ability of a measure to detect a signal (i.e., outcome) and "specificity" refers to the ability of a measure to discriminate between the target signal and other signals or noise. ROC curve analyses have been used extensively in other literatures, especially in the medical field for diagnostic testing (e.g., radiology; Hanley & McNeil, 1982). The ROC curve results were evaluated using the area under the curve (AUC) where measures with AUC = 1 are considered perfectly sensitive/specific to detection and discrimination of the target outcome variable and AUC < 0.50 are considered poor (Bradley & Longstaff, 2004). Generally, AUC values $\geq$ 0.65 are considered adequately sensitive/specific (Egger & Borg, 2016). Although AUC reflects an ability to both accurately detect and discriminate a target outcome variable, for parsimony of language, AUC results will be described using "detection" language throughout the manuscript.

All ROC curve analyses were conducted using non-consumption outcomes assessed at the assessment timepoint that immediately followed treatment in each study (4-month (i.e., "week-16") follow-up in COMBINE and 3-month follow-up in MATCH). These analyses examined how sensitive/specific each variable is at detecting binary outcomes at two timepoints: 4- or 3-months post-treatment and 12-months post-treatment for: 1) abstinence versus any drinking, 2) no heavy drinking days versus any heavy drinking days (Falk et al., 2010), 3) the World Health Organization risky drinking levels [European Medicines Agency, 2010; with three cutoffs: (a) low risk (<20/40g alcohol for women/men per day), (b) medium risk (<40/60g alcohol for women/men), and c) high risk or very high risk (>41/61g alcohol for women/men; English et al., 1995); all risk

levels were calculated via DDD, MXD, and DPD], and 4) scores on a composite clinical outcome measure of alcohol-related problems and consumption [Cisler & Zweben, 1999; with four cutoffs: a) abstinent, b) abstinent or moderate drinking without problems, c) heavy drinking/problems, and d) heavy drinking and problems]. As a further test of WHO risk levels, ROC curve analyses were also conducted for non-consumption variables' sensitivity/specificity for changes in WHO risk level between baseline and post-treatment (1+ and 2+ risk level changes, calculated via DDD and DPD). The results from the non-consumption variables ROC curve analyses were compared to those of the most widely used consumption-based measures (PDA and PHD) to understand if non-consumption outcomes were substantially less sensitive compared to consumption outcomes. Analyses were conducted separately in COMBINE and MATCH to examine the cross validation of findings whenever possible.

For measures with subscale factor structures that were upheld via the CFA results, ROC curve analyses were conducted for each subscale. When AUC $\geq$ 0.65 for subscales on at least one of the outcomes tested, individual item ROC curves were analyzed to try to identify individual items most sensitive/specific to the evaluated outcomes. Given the extensive number of items, we report the results from item level ROC curve analyses in Appendix A.

Summary of Analyses

In order to synthesize the myriad results of the present study, results were distilled and summarized using a 2-point system. Measures that had poor sensitivity/specificity, psychometric properties, or measurement invariance were allocated 0 points; those with mixed or modest properties were allocated 1 point. Measures with acceptable to excellent

sensitivity/specificity, psychometric properties, or measurement invariance were allocated 2 points. Sensitivity/specificity scores of 0 indicated area under the curve (AUC) < 0.650 across all outcomes; 1 point indicated AUC > 0.650 and < 0.700 or mixed results across studies or across consumption outcomes; 2 points indicated AUC > 0.700 in both COMBINE and MATCH or for most outcomes. Internal consistency reliability scores of 0 indicated $\alpha < 0.70$; 1 point indicated $\alpha > 0.70$ and < 0.80 or mixed results across studies; 2 points indicated $\alpha > 0.80$ in both COMBINE and MATCH. Convergent validity results with scores of 0 indicated non-significant ($p > 0.05$) or at least one correlation in the opposite direction than was expected; 1 point indicated significant correlations with some but not all the expected measures or mixed results across studies; 2 points indicated significant correlations in the expected direction for all measures in both COMBINE and MATCH. Confirmatory factor analysis (CFA) results with scores of 0 indicated RMSEA > 0.08 or CFI or TLI < 0.90; 1 point indicated RMSEA < 0.08 and > 0.06 and/or CFI or TLI > 0.90 and < 0.95 or mixed results across studies; 2 points indicated RMSEA < 0.06 and CFI or TLI > 0.95 in both COMBINE and MATCH. Measurement invariance results with scores of 0 indicated non-invariance at the configural level or did not proceed to invariance testing due to poor model fit; 1 point indicated at least adequate model fit through the metric invariance testing (constraint of the factor loadings for equivalence) or mixed results across both studies; 2 points indicated good model fit through strong invariance testing (highest possible level of invariance for categorical data) in both COMBINE and MATCH.

Results

Descriptive Results

Descriptive analyses of the data are presented in Tables 4 and 5. As depicted in Table 4, consumption outcome descriptives were largely similar between COMBINE and MATCH datasets. The percent days abstinent (PDA) were slightly higher at all timepoints in MATCH (baseline mean = 30.90% (SD =29.96%), *N* = 1725); post-treatment mean = 83.17% (SD = 28.51%), *N* = 1657; 12-month follow-up mean = 76.69% (SD = 33.55%), *N* = 1594) compared to COMBINE (baseline mean = 21.41% (SD = 22.50%), N = 1383); post-treatment mean = 72.66% (SD = 33.49%), N = 1288; 12-month follow-up mean = 62.63% (SD = 39.12%), N = 1099). Similarly, the percent heavy drinking days (PHDD) was higher in COMBINE (baseline mean = 70.52% (SD = 26.57%), N = 1383); post-treatment mean = 17.54% (SD = 28.69%), N = 1288; 12-month follow-up mean = 26.20% (SD = 34.27%), N = 1171) at all timepoints compared to MATCH (baseline mean = 63.18% (SD = 31.43%), N = 1725); post-treatment mean = 12.46% (SD = 25.09%), N = 1657; 12-month follow-up mean = 16.71% (SD = 29.17%), N = 1594). These consumption variables were also consistent across both COMBINE and MATCH in that the largest differences between timepoints occurred from baseline to post-treatment and from baseline to 12-month follow-up, as evidenced by very large effect sizes ($d > 1.0$) for difference from baseline and effect sizes around 0.2 in both COMBINE ($d = 0.277$ for PDA and $d = 0.275$ for PHDD) and MATCH (d = 0.208 for PDA and d = 0.156 for PHDD) for post-treatment to 12-month follow-up.

The pattern of change observed with consumption outcomes PDA and PHDD, whereby the smallest changes occurred between post-treatment and 12-month follow-up and the largest changes occurred between baseline and post-treatment and also between baseline and 12-month follow-up, was consistent with change patterns in the non-

consumption measures (see Table 4). However, there were differences in descriptive statistics for measures used in COMBINE versus MATCH. Similar to the different rates of abstinence and heavy drinking days, COMBINE and MATCH differed slightly with regards to their overall sample's endorsement of alcohol-related consequences on the Drinker Inventory of Consequences (DrInC) at each timepoint and for each of the 5 commonly used subscales. In COMBINE, the overall DrInC average summary score at baseline was 47.61 (SD = 20.42; *N* = 1381; baseline to post-treatment *d* = 1.735), at post-treatment was 13.36 (SD = 18.85; *N* = 1098; post-treatment to 12-month *d* = 0.322), and at 12-month follow-up was 19.89 (SD = 21.81; *N* = 965; baseline to 12-month *d* = 1.320). In contrast, for MATCH the overall DrInC average summary score at baseline was 52.63 (SD = 23.32; *N* = 1703; baseline to post-treatment *d* = 0.680), at post-treatment was 35.86 (SD = 26.78; *N* = 985; post-treatment to 12-month *d* = 0.323), and at 12-month follow-up was 27.50 (SD = 24.70; *N* = 789; baseline to 12-month *d* = 1.057). Higher DrInC scores in Project MATCH were likely due to the different administration procedures for COMBINE and MATCH with the DrInC, whereby the DrInC was administered to all abstainers in COMBINE and only some of the abstainers in MATCH. Similar patterns are observed in the commonly used subscales (physical health consequences, interpersonal consequences, intrapersonal consequences, impulse control, and social responsibility) of the DrInC for COMBINE and MATCH and effect sizes were generally higher in COMBINE overall. Effect sizes differed between COMBINE and MATCH in that the greatest changes in subscale scores occurred from baseline to post-treatment in COMBINE whereas the largest effect sizes in MATCH occurred baseline to 12-month follow-up. These changes may reflect the overall sample differences between COMBINE

and MATCH and the fact that overall sample was used for descriptive analyses rather than sub-samples (e.g., treatment arms in MATCH were not examined separately).

Other noteworthy findings from descriptive analyses were that effect sizes were found to be very large for several of the non-consumption measures and that these effects were on-par with those found for the primary consumption outcome variables of PDA and PHDD. For instance, the baseline to post-treatment effect size of the Obsessive-Compulsive Drinking Scale (OCDS) in COMBINE was $d = 1.762$, which is larger than the congruent effect size of the DrInC in COMBINE. Effect sizes of the Alcohol Abstinence Self-Efficacy Scale in COMBINE and MATCH were small for the total AASE; however, effect sizes were much larger when examining the Confidence and Temptation subscales independently. In COMBINE, the baseline to post-treatment effect sizes were $d = 1.078$ and $d = 1.022$ for the Confidence and Temptation subscales, respectively. These scores were noticeably smaller in MATCH: $d = 0.469$ and $d = 0.674$, respectively. The remaining measures had smaller effect sizes, although many effect sizes were still notable and were in the medium range ($0.2 < d < 0.8$; Cohen, 1988). Importantly, analysis of the descriptive statistics highlighted differences between COMBINE and MATCH that may reflect the overall sample differences between COMBINE and MATCH and the fact that overall sample was used for descriptive analyses rather than sub-samples (e.g., treatment arm).

Descriptive analyses also indicated the potential for problems with factor analyses given large standard deviations in some of the measures. Most notably, the Addiction Severity Index yielded very large standard deviations, especially for the family history (mean = 2.65, SD = 48.04, $N = 1726$) and legal status (mean = 141.90, SD = 384.71, $N =$

1726) subscales. Moreover, since the Psychiatric status mean scores were < 1 at all three timepoints at which this subscale was uniquely administered, descriptive analyses suggested combining psychiatric, family history, and legal status questions into one ASI summary score may be problematic given the range in responses possible for each subscale. These descriptive statistics and effect sizes provide a potentially useful overview of the performances of each measure in COMBINE and MATCH.

Descriptive results not presented in Table 4 are those for employment status and income (ESI) items used in COMBINE and MATCH. Employment status descriptives were: unemployment or disabled = 0 (*n* = 225 in COMBINE, 478 in MATCH), homemaker, part-time employed, or retired = 1 (*n* = 253 in COMBINE, 282 in MATCH), full-time employed = 2 (*n* = 838 in COMBINE, 847 in MATCH). COMBINE also included one item for income that was not paralleled in MATCH (< $15,000, *n* = 139; $15,000 - $29,999, *n* = 219; $30,000 - $59,000, *n* = 408; $60,000 - $89,000, *n* = 266; > $90,000, *n* = 330). These items indicate fairly even distribution of responses and again highlight important demographic differences between COMBINE and MATCH.

A final descriptive overview of the variables examined in the present study is provided by Table 5, which depicts the frequencies of the binary consumption outcome variables that were examined in the Receiver Operating Characteristic (ROC) curve analyses. Although base rate is mathematically unrelated to sensitivity and specificity (Pepe, 2003), the rates of each consumption outcome are interesting to consider in comparison to one another. Perhaps most notably, the rates of each of the World Health Organization Risk levels vary depending on how the risk levels were calculated (i.e., via Drinks per Drinking Day (DDD), Maximum number of drinks consumed in the 90-day

window (MXD), or via drinks per day in the assessment window (averaged across drinking and abstinent days; DPD)). For instance, the number of participants categorized as moderate or lower risk in COMBINE at post-treatment was: n=744, n=647, or n=1089 depending on if that risk level was calculated via DDD, MXD, or DPD, respectively.

Summary of Analyses

The overall results from the core analyses of the present study are depicted in Table 6 and highlight that no measure performed excellently across all examined domains. The Brief Symptom Inventory and Beck Depression Inventory performed best, with 2 points allocated for all examined properties except sensitivity/specificity, for which only 1 point was allocated due to mixed results. Similarly, the WHOQOL-BREF performed well and had 2 points allocated for all except CFA results, which fit adequately, and sensitivity/specificity, which were unable to be compared to other measures since WHOQOL-BREF administration occurred after the post-treatment timepoint in COMBINE. The remaining measures all had at least some promising qualities and are described below in order of how well they performed (best performance to poorest performance). The only measure examined in the present study that received 0 points for "poor" properties across all analyses was the Addiction Severity Index. These results were consistent with the fact that the full ASI was not used in MATCH and, more importantly, that the measure may be most helpful as an inventory of historical events (e.g., number of times incarcerated for various offenses, number of family members with histories of alcohol problems) rather than a measure that may hold utility for comparing scores pre- and post-treatment.

Strongest Results

Brief Symptom Inventory. The Brief Symptom Inventory was originally conceptualized as containing 9-factors with higher scores indicating more severe psychological symptoms: Somatization, Obsessive-Compulsive, Depression, Interpersonal Sensitivity, Hostility, Anxiety, Psychoticism, Phobic Anxiety, and Paranoid Ideation (Derogatis & Melisaratos, 1983). This 9-factor structure was upheld in the COMBINE study via CFA (replication half fit indices: RMSEA = 0.022 (90% CI: 0.019, 0.025); CFI = 0.975; TLI = 0.974; presented in Table 7). Moreover, this factor structure was invariant across time between baseline and post-treatment (week 16) timepoints in COMBINE. The configural model of invariance testing fit very well (RMSEA = 0.011 (90% CI: 0.010, 0.012); CFI = 0.981; TLI = 0.980), as did tests of metric invariance (RMSEA = 0.011 (90% CI: 0.009, 0.012); CFI = 0.982; TLI = 0.981), and strong invariance (RMSEA = 0.012 (90% CI: 0.011, 0.013); CFI = 0.977; TLI = 0.977). Moreover, as presented in Table 8, internal consistency reliability of the BSI and 9 factor subscales varied from good to excellent, with the sole exception of the Interpersonal Sensitivity subscale: total BSI $\alpha = 0.965$, Somatization subscale $\alpha = 0.798$, Obsessive-Compulsive subscale $\alpha = 0.862$, Depression subscale $\alpha = 0.882$, Interpersonal Sensitivity subscale $\alpha = 0.643$, Hostility subscale $\alpha = 0.790$, Anxiety subscale $\alpha = 0.824$, Psychoticism subscale $\alpha = 0.864$, Phobic Anxiety subscale $\alpha = 0.786$, and Paranoid Ideation subscale $\alpha = 0.836$. The BSI also had good convergent validity and all bivariate correlations were significant ($p < 0.01$) in the direction predicted. Specifically, the BSI was significantly, negatively correlated with quality of life (WHOQOL-BREF: $r = -0.698$, $p < 0.001$; SF-12: $r = -0.688$, $p < 0.001$), recoded employment status (ESI

employment item: r = -0.189, $p < 0.001$), and income (ESI income item: r = -0.211, $p < 0.01$).

The BSI also had modest sensitivity/specificity, as indicated by ROC curve results, as detailed in Table 9. The post-treatment BSI total summary score adequately detected 9 of 15 post-treatment consumption outcomes (AUC $\geq$ 0.650), however only detected 1 of 11 12-month follow-up consumption outcomes (AUC $\geq$ 0.650). The total BSI summary score had the highest AUC when detecting post-treatment composite clinical outcome of heavy or lower risk (AUC = 0.833) and had the lowest AUC when detecting 2+ level change in WHO risk level since baseline (calculated via drinks per drinking day (DDD); AUC = 0.511). For 12-month follow-up consumption outcomes, the BSI total score had the highest AUC when detecting the composite clinical outcome of heavy or lower risk (AUC = 0.708) and the lowest AUC when detecting 12-month abstinence (AUC = 0.545).

All post-treatment BSI subscales, representing the 9 factors of the BSI, adequately detected at least 1 of 15 post-treatment consumption outcomes. The Depression factor adequately detected 8 out of 15 post-treatment outcomes and 1 of 11 12-month follow-up outcomes. The Interpersonal Sensitivity factor adequately detected 8 out of 15 post-treatment outcomes but 0 of 11 12-month follow-up outcomes. The Anxiety factor adequately detected 7 out of 15 post-treatment outcomes and 1 of 11 12-month follow-up outcomes. The Obsessive-Compulsive factor adequately detected 5 of 15 post-treatment outcomes and 1 of 11 12-month follow-up outcomes. The Psychoticism factor adequately detected 4 of 15 post-treatment outcomes and 1 of 11 12-month follow-up outcomes. The Somatization factor adequately detected 4 of 15 post-treatment outcomes but 0 of 11 12-

month follow-up outcomes (AUC's $\geq$ 0.650). The Hostility factor adequately detected 3 of 15 post-treatment outcomes and 1 of 11 12-month follow-up outcomes. The Paranoia symptom subscale adequately detected 1 of 15 post-treatment outcomes and 1 of 11 12-month follow-up outcomes. Finally, the Phobic Anxiety factor adequately detected only 1 of 15 post-treatment outcomes and none of the 12-month follow-up outcomes. Every BSI subscale had the highest AUC when detecting post-treatment and 12-month follow-up composite clinical outcome of heavy or lower risk (AUC's = 0.713 to 0.833 post-treatment; AUC's = 0.638 to 0.706 12-month follow-up). Further, six of the nine subscales had the lowest AUC's when detecting 1+ level change in WHO risk level since baseline (calculated via drinks per day (DPD); AUC's = 0.529 to 0.568). For 12-month outcomes, eight of the nine subscales had the lowest AUC's when detecting abstinence (AUC's = 0.508 to 0.557); only the Psychoticism symptoms subscale had the lowest AUC when detecting 12-month WHO risk level of moderate or lower (calculated via drinks per drinking day (DDD); AUC = 0.523). As reported in Appendix A, many of the items on the BSI yielded AUC's $\geq$ 0.650 at post-treatment and the 12-month follow-up.

Beck Depression Inventory. The majority of published studies report that the BDI is comprised of 2 or 3 factors; therefore, it was unsurprising that both 2- and 3-factor models fit well in the MATCH data (2-factor model: RMSEA = 0.030 (90% CI: 0.025-0.035); CFI = 0.978; TLI = 0.975; 3-factor model: RMSEA = 0.027 (90% CI: 0.021-0.032); CFI = 0.982; TLI = 0.980). Further, both the 2- and the 3-factor models sustained comparable levels of fit through strong invariance testing between baseline and post-treatment timepoints. The 2-factor model yielded good fit through constraints of thresholds to equivalence between timepoints (RMSEA = 0.019 (90% CI: 0.017-0.021);

CFI = 0.968; TLI = 0.969) as did the 3-factor model (RMSEA = 0.019 (90% CI: 0.017-0.021); CFI = 0.970; TLI = 0.971; see Table 7). Similarly, internal consistency was good for the overall BDI (α = 0.889) as well as each factor, with the sole exception of the Somatic factor in the 3-factor model (see Table 8): 2-factor Cognitive-Affective factor α = 0.848, 2-factor Somatic factor α = 0.771, 3-factor Negative Attitudes factor α = 0.859, 3-factor Performance Impairment factor α = 0.739, 3-factor Somatic factor α = 0.478. The convergent validity results for the total BDI were also good; the BDI was significantly negatively correlated with psychosocial functioning (PFI; r = -0.380, p < 0.001) and employment status (ESI employment item; r = -0.150, p < 0.001).

As detailed in Table 10, ROC curve analyses indicated the BDI more strongly detected 12-month follow-up outcomes (4 of 11 outcomes were detected at AUC $\geq$ 0.650) than post-treatment outcomes (2 of 15 outcomes were detected at AUC $\geq$ 0.650). The total BDI summary score was had the highest AUC values detecting post-treatment and 12-month composite clinical outcome of heavy or lower risk (AUC = 0.658; AUC = 0.681) and the lowest AUCs when detecting 2+ change in WHO risk level since baseline (computed via DPD; AUC = 0.555) and 12-month abstinence (AUC = 0.596). The same patterns were observed for all factors upheld via CFA and invariance testing (2-factor Cognitive-Affective factor, 2-factor Somatic factor; 3-factor Negative Attitudes factor, 3-factor Performance Impairment factor, 3-factor Somatic factor). For the 2-factor solution, the Cognitive-Affective factor and Somatic factor each adequately detected 3 of 15 post-treatment outcomes, and 0 of 11 12-month outcomes. The 3-factor solution, the Negative Affect factor adequately detected 8 of 15 post-treatment outcomes and 0 of 11 12-month follow-up outcomes; the Performance Impairment factor and the Somatic factor each

adequately detected 2 of 15 post-treatment outcomes and 0 of 11 12-month follow-up outcomes. As reported in Appendix A, many of the items on the BDI yielded AUC's $\geq$ 0.650 at post-treatment and the 12-month follow-up.

World Health Organization Quality of Life. The WHOQOL-BREF has been found to have a higher-order factor structure in previous publications (Skevington et al., 2004). Although one item was omitted from the COMBINE Study administration, this higher-order factor structure was acceptably upheld in COMBINE via CFA (replication half fit indices: RMSEA = 0.050 (90% CI: 0.045, 0.055); CFI = 0.942; TLI = 0.932). Although other factor structures have been published in other datasets (e.g., Yao & Wu, 2005), several previously published factor structures were tested in COMBINE with no one factor structure fitting substantially better than others. Accordingly, invariance testing was continued with the most widely-cited factor structure that made the most conceptual sense: the higher-order factor structure described by Skevington et al. (2004). Higher-order factor invariance testing was completed as described above with very little change to any of the fit indices. The least constrained level of invariance testing fit acceptably well across baseline and week 26 timepoints (RMSEA = 0.037 (90% CI: 0.035, 0.038); CFI = 0.921; TLI = 0.916) as did the highest level of constraints possible for this measure (strong invariance: RMSEA = 0.037 (90% CI: 0.036, 0.039); CFI = 0.926; TLI = 0.930). Further, the overall WHOQOL-BREF as well as the subscale factors had good to excellent internal consistency reliability, as indicated in Table 8 ($\alpha = 0.901$ for the overall measure; Physical Health $\alpha = 0.798$, Psychological Health $\alpha = 0.770$, Social Relationships $\alpha = 0.718$, Environment $\alpha = 0.812$). Similarly, the WHOQOL-BREF had good convergent validity and was significantly (*p's* < 0.001), negatively correlated to

both alcohol-related consequences (DrInC; r = -0.456) and psychological symptoms (BSI; r = -0.698) as hypothesized.

Although not accounted for in the overall results summary (Table 6) because WHOQOL-BREF data were collected after the week 16 post-treatment timepoint, ROC curve analyses indicated that the week 26 WHOQOL-BREF data were adequately able to detect 12-month follow-up consumption outcomes (presented in Table 11). The total WHOQOL-BREF score adequately detected 4 of 11 consumption outcomes. The total WHOQOL-BREF summary score had the highest AUC when detecting 12-month composite clinical outcome of heavy or lower risk (AUC = 0.715) and had the lowest AUC when detecting 12-month abstinence (AUC = 0.508). The same pattern was observed for each of the 4 subscales of the WHOQOL-BREF that were upheld via CFA and invariance testing: 12-month composite clinical outcome of heavy or lower risk was the outcome with the highest AUC and 12-month abstinence was the outcome with the lowest AUC. The Physical Health, Social, and Environment subscales each only adequately detected 12-month composite clinical outcome of heavy or lower risk (AUC's = 0.698, 0.691, 0.653). However, the Psychological Health subscale adequately detected 12-month WHO risk moderate or lower risk (calculated via DDD and DPD; AUC's = 0.652, 0.659), 12-month composite clinical outcome of moderate or lower risk (AUC = 0.665), and 12-month composite clinical outcome of heavy or lower risk (AUC = 0.704). As reported in Appendix A, several individual items of the WHOQOL-BREF adequately detected 12-month outcomes.

Measures with Mixed Results

Obsessive-Compulsive Drinking Scale. The Obsessive-Compulsive Drinking Scale (OCDS) has been conceptualized as a measure of alcohol craving, but publications have differed in their conceptualizations of factors comprising the OCDS. Unsurprisingly, multiple steps were taken to test various factor models. First, a 2-factor solution was tested based on the original conceptualization of the OCDS, which has been replicated in other studies (Ansseau et al., 2000; Anton 2000; Cordero, Solis, Cordero, Torruco, & Cruz-Fuentes, 2009). However, model fit was poor in the COMBINE data, and the model was not tested in the replication half of the sample (RMSEA = 0.109 (90% CI: 0.102-0.117); CFI = 0.852; TLI = 0.823). Model fit was similarly poor for alternative 2-, 3-, and 4-factor models that were based on other published findings (see Table 7). Consequently, EFA was combined with conceptual interpretation of the items to generate a new 4-factor model that differs slightly from other studies. These 4-factors are consistent with a conceptualization of the OCDS as measuring: frequency of craving thoughts ("Factor 1"), craving interference with activities ("Factor 2"), distress of the craving ("Factor 3"), and controllability of craving ("Factor 4"). Notably, the two items of the OCDS that assess for alcohol consumption directly were omitted from the final factor model due to their poor contribution to the factor structures tested and per previous findings (Anton, 2000). This 4-factor model provided acceptable fit to baseline data in COMBINE (replication half sample: RMSEA = 0.072 (90% CI: 0.062-0.082); CFI = 0.968; TLI = 0.596). Measurement invariance testing of this new 4-factor model yielded adequate fit statistics through strong invariance (see Table 7; RMSEA = 0.076 (90% CI: 0.062, 0.082); CFI = 0.910; TLI = 0.912). However, fit substantially worsened between metric and strong invariance (i.e., when adding the constraint of equality of thresholds),

which may reflect measurement non-invariance at the threshold level. Attempts to establish partial strong invariance were unsuccessful and yielded consistently inadequate model fit.

Despite the potential non-invariance of the OCDS at the strong invariance level, the OCDS did have other psychometric strengths in COMBINE. The internal consistency was strong for the total OCDS score ($\alpha = 0.852$) and convergent validity results yielded significant bivariate correlations in all Form 90 consumption variables, as predicted: PDA ($r = -0.103$, $p < 0.001$), PHDD ($r = 0.202$, $p < 0.001$), change in WHO risk level since baseline (as calculated with DDD: ($r = 0.288$, $p < 0.001$)), DPD ($r = 0.305$, $p < 0.001$), DDD ($r = 0.360$, $p < 0.001$ and MXD ($r = 0.320$, $p < 0.001$).

Moreover, the total summary score of the OCDS strongly detected all post-treatment and 12-month follow-up consumption outcomes (presented in Table 12). Area under the curve (AUC) values for detecting post-treatment outcomes ranged from 0.686 to 0.934; those for 12-month outcomes ranged from 0.690 and 0.756. For post-treatment outcomes, the OCDS total summary score yielded the highest AUC for post-treatment composite clinical outcome of heavy or lower risk and was lowest AUC for 1+ level change in WHO risk level since baseline (computed via DDD). For 12-month follow-up outcomes, the OCDS total summary score yielded highest AUC for composite clinical outcome of moderate or lower risk and lowest AUC for WHO moderate or lower risk (calculated via DDD).

The 4 factors that were partially supported per adequate CFA results and partial invariance testing yielded several promising ROC curve results (AUC's ≥ 0.650). Factors 1 and 2 both adequately detected 11 of 15 post-treatment outcomes (all but 1+ and 2+

risk level changes in WHO risk levels since baseline as computed via DDD and DPD): AUC's ranged from 0.589 to 0.782 for Factor 1 and 0.614 to 0.871 for Factor 2. Factors 1 and 2 also adequately detected 12-month composite clinical outcome of heavy or lower risk (AUC's = 0.668 and 0.672) and had lowest AUCs when detecting 12-month abstinence (AUC's = 0.604 and 0.566). Factors 3 and 4 performed even better than Factors 1 and 2. Factor 3 adequately detected 13 out of 15 post-treatment outcomes; AUC's ranged between 0.631 (2+ change in WHO risk level since baseline, calculated via DPD) and 0.877 (composite clinical outcome of heavy or lower risk). Factor 3 also adequately detected 7 of 11 12-month outcomes; AUC's ranged from 0.641 (WHO moderate or lower risk, calculated via max drinks (MXD) to 0.716 (composite clinical outcome of heavy or lower risk). Factor 4 performed superior and adequately detected 15 of 15 and 11 of 11 post-treatment (AUC's ranged from 0.676 to 0.912) and 12-month outcomes (AUC's ranged from 0.682 to 0.742). For post-treatment outcomes, Factor 4 had highest AUC for composite clinical outcome of heavy or lower risk and lowest AUC for 1+ change in WHO risk level since baseline (as calculated via DDD). For 12-month follow-up outcomes, Factor 4 had highest AUC for composite clinical outcome of moderate or lower risk and lowest AUC for WHO moderate or lower risk level (calculated via DDD). As reported in Appendix A, many of the items on the OCDS yielded AUC's $\geq$ 0.650 at post-treatment and the 12-month follow-up.

Drinker Inventory of Consequences. Similar to the OCDS, the factor structure of the DrInC has been explored in numerous prior studies, including several publications on the 15-item abbreviated version, the Short Inventory of Problems (SIP; e.g., Forcehimes, Tonigan, Miller, Kenna, & Baer, 2007; Kiluk, Dreifuss, Weiss,

Morgenstern, & Carroll, 2012). Accordingly, a number of factor structures were tested in COMBINE and MATCH via CFA for the DrInC as well as the SIP (see Table 7 for fit statistics). The only factor structure tested in the present study that yielded adequate fit in both COMBINE and MATCH and was invariant beyond configural invariance was a 3-factor solution created in the present study based on a conceptualization of the DrInC as consisting of alcohol related consequences that occur at different frequencies (e.g., consequences that occur commonly, such as hangovers; consequences that occur moderately commonly; and rare consequences). As detailed in Table 7, this 3-factor solution fit adequately at baseline in both COMBINE and MATCH (COMBINE: RMSEA = 0.041 (90% CI: 0.038, 0.043); CFI = 0.920; TLI = 0.916; MATCH: RMSEA = 0.040 (90% CI: 0.038, 0.042); CFI = 0.908; TLI = 0.904) and fit improved as additional constraints were added through constraining thresholds to equivalence across time (COMBINE: RMSEA = 0.024 (90% CI: 0.023, 0.025); CFI = 0.951; TLI = 0.952; MATCH: RMSEA = 0.018 (90% CI: 0.017, 0.019); CFI = 0.941; TLI = 0.942). Similarly, the total DrInC and the 3 factors all had strong internal consistency in COMBINE and MATCH. As described in Table 8, total DrInC in COMBINE and MATCH had $\alpha = 0.937$ and $\alpha = 0.938$, respectively. The 3-factor subscales in COMBINE and MATCH were: Common Consequences $\alpha = 0.855$, $\alpha = 0.833$; Moderately Common Consequences $\alpha = 0.905$, $\alpha = 0.905$; and Rare Consequences $\alpha = 0.808$, $\alpha = 0.830$.

Bivariate correlations were significant and in the expected direction for Form 90 variables ($p < .01$ in both COMBINE and MATCH). Other variables were also significantly correlated in the hypothesized directions in COMBINE: OCDS $r = 0.519$ ($p < 0.001$), AASE $r = 0.153$ ($p < 0.001$), WHOQOL-BREF $r = -0.456$ ($p < 0.001$), SF-12 r

= -0.486 ($p < 0.001$), recoded employment status r = -0.225 ($p < 0.001$) and income r = -0.233 ($p < 0.001$). In MATCH, the DrInC also significantly correlated with other hypothesized measures as predicted: PFI r = -0.479 ($p < 0.001$) and recoded employment status r = -0.164 ($p < 0.001$). However, the DrInC had poor convergent validity in MATCH with temptation to drink (r = -0.060, $p < 0.05$) and AA involvement (AAI; r = 0.336, $p < .001$). These inconsistent psychometrics across COMBINE and MATCH may be due to the inconsistent administration in MATCH to individuals who had been 100% abstinent during the follow-up window.

For the DrInC summary ROC curve analyses, all DrInC summary AUC values $\geq$ 0.650 in COMBINE for both timepoints (post-treatment AUC's = 0.677-0.944; 12-month follow-up AUC's = 0.684-0.736; see Table 13). Further, the first two factors adequately detected all post-treatment and 12-month follow-up outcomes: Common Consequence Factor post-treatment AUC's = 0.677-0.939, 12-month AUC's = 0.671-0.719; Moderately Common Consequences Factor post-treatment AUC's = 0.663-0.939, 12-month AUC's = 0.659-0.715. The Rare Consequences Factor adequately detected 14 of 15 post-treatment outcomes (AUC's = 0.642-0.917) and 10 of 11 12-month AUC's = 0.645-0.699). The DrInC total score and each of the 3 factors had the highest AUC's for composite clinical outcome of heavy or lower risk and the lowest AUC's for 2+ level change in WHO risk level since baseline (computed via DPD). For 12-month follow-up outcomes, DrInC total score and each of the 3 factors all had the highest AUC's for composite clinical outcome of moderate or lower risk and the lowest AUC's for abstinence. As reported in Appendix A, many of the items on the DrInC yielded AUC's $\geq$ 0.650 at post-treatment and the 12-month follow-up.

In MATCH, the total DrInC summary adequately detected 9 of 15 post-treatment outcomes (AUC's = 0.493-0.886) but failed to adequately detect any 12-month outcomes (AUC's = 0.424-0.590; see Table 14). Similarly, each of the 3-factors adequately detected 8 of 15 post-treatment outcomes but failed to detect any 12-month follow-up outcomes: Common Consequence Factor post-treatment outcomes AUC's = 0.508-0.850; Moderately Common Consequences Factor post-treatment outcomes AUC's = 0.495-0.879; Rare Consequences Factor post-treatment outcomes AUC's = 0.513-0.876. Similar with the COMBINE study results, the DrInC total score and each of the 3 factors all had the highest AUC's for composite clinical outcome of moderate or lower risk and the lowest AUC's for 2+ level change in WHO risk level since baseline (computed via DPD). For 12-month follow-up outcomes, DrInC total score and each of the 3 factors all had the highest AUC's for composite clinical outcome of moderate or lower risk and the lowest AUC's for abstinence. As reported in Appendix A, many of the items on the DrInC yielded AUC's $\geq$ 0.650 at post-treatment and the 12-month follow-up in Project MATCH.

Measures with Poorer Results

Spielberger State-Trait Inventory. The two factors (Temperament and Reaction) that have been previously described in the literature (Forgays et al., 1997) provided poor fit to the data in Project MATCH via CFA (split half 1 sample: RMSEA = 0.116 (90% CI: 0.109-0.122); CFI – 0.902; TLI = 0.884). However, Forgays and colleagues (1997) also described the SSTI in terms of 7 subscales. The items administered in MATCH consisted of 4 of those 7 subscales; this 4-factor model provided excellent fit via CFA in MATCH data (replication half sample: RMSEA =

0.056 (90% CI: 0.048-0.064); CFI – 0.976; TLI = 0.969). Because the SSTI was only administered at baseline in Project MATCH, measurement invariance across time was not tested in the present study for this 4-factor solution. Factor analysis results are detailed in Table 7. Further, the overall SSTI scale had excellent internal consistency ($\alpha = 0.887$) and all but one of these 4 factors had good internal consistency: "Factor 1" $\alpha = 0.746$, "Factor 2" $\alpha = 0.865$, "Factor 4" $\alpha = 0.496$, and "Factor 6" $\alpha = 0.781$ (as detailed in Table 8).

The SSTI performed inadequately in terms of both convergent validity and sensitivity/specificity. The SSTI was significantly correlated with the PFI as hypothesized ($r = -0.200$; $p < 0.001$) and the SSTI summary score using only the items in the CFA ($r = -0.053$, $p < 0.05$) but the full SSTI summary score was not significantly correlated with recoded employment status in MATCH ($r = -0.051$, $p > 0.05$). A bigger weakness of the SSTI was its poor sensitivity and specificity for detecting consumption outcomes. For the full SSTI and each of the 4 factors tested via CFA, AUC values were below the 0.650 cutoff for all consumption outcomes at both post-treatment and 12-month follow-up (see Table 15). The total SSTI score (including all items administered in MATCH) had AUC values ranging from 0.503-0.577 and 0.499-0.559 for post-treatment and 12-month follow-up outcomes. Similarly, AUC values for the total SSTI score (including only the items from CFA) ranged from 0.504-0.575 and 0.501-0.561 for post-treatment and 12-month follow-up outcomes. Both of these forms of the total SSTI score had highest AUC for post-treatment WHO moderate or lower risk (calculated via DPD) and 12-month composite clinical outcome of heavy or lower risk; the lowest AUC's for both total SSTI scores were detecting 2+ level change in WHO risk level since baseline (calculated via DPD) and 12-month abstinence. No consistent pattern emerged for which

post-treatment and 12-month follow-up outcomes had consistently the highest or lowest AUC for the 4 factors that were upheld via CFA, but none of the 4 factors were able to adequately detect post-treatment or 12-month follow-up outcomes (AUC's = 0.490-0.576).

Health Survey (SF-12). The SF-12 yielded mixed and modest findings for construct validity, measurement invariance testing, and ROC curve analyses results but had promising internal consistency and convergent validity in COMBINE. Despite the substantial body of literature and conceptual sense supporting the SF-12 as comprised of 2 factors (Physical Health and Psychological Health), CFA analyses indicated only an adequate fit of this model (RMSEA = 0.080 (90% CI: 0.071, 0.090); CFI = 0.951; TLI = 0.939). Further, configural invariance was not upheld per fit indices that were outside *a priori* cutoff values and the configural invariance mode fit substantially poorer than the baseline CFA (RMSEA = 0.075 (90% CI: 0.072, 0.078); CFI = 0.854; TLI = 0.839). Additional levels of invariance were not tested per the above evidence of configural non-invariance.

The SF-12 as administered in COMBINE had good internal consistency and convergent validity. Internal consistency was good for total SF-12 ($\alpha = 0.874$) as well as both Physical ($\alpha = 0.805$) and Psychological Health ($\alpha = 0.861$) factors. Convergent validity was particularly strong. The total SF-12 score was highly, negatively correlated with both alcohol-related consequences (DrInC; $r = -0.486$, $p < 0.001$) and total psychiatry symptom severity (BSI; $r = -0.688$, $p < 0.001$).

ROC curve analyses results were mixed, as detailed in Table 16. The total SF-12 summary score performed well for detecting consumption outcomes at post-treatment (8

of 15 outcomes were adequately detected, AUC's = 0.577-0.836), performed poorly at detecting 12-month follow-up consumption outcomes (1 of 11 outcomes was adequately detected, AUC's = 0.548-0.681). Similarly, the Physical Health and Psychological Health factors detected post-treatment consumption outcomes fairly well (4 of 15 and 9 of 15, respectively; AUC's = 0.522-0.766 and 0.600-0.840) whereas the Physical Health factor failed to adequately detect any 12-month follow-up consumption outcomes (AUC's = 0.518-0.627) and the Psychological Health factor only adequately detected 1 of 11 consumption outcomes for the 12-month follow-up (AUC's = 0.561-0.693). The total summary score and each of the 2 factors all had highest AUC's for post-treatment and 12-month follow-up composite clinical outcome of heavy or lower risk. The total SF-12 score and the Physical Health factor both had lowest AUC's for 1+ level change in WHO risk level since baseline (calculated via DPD); the Psychological Health factor had lowest AUC for 2+ level change in WHO risk level since baseline (calculated via DPD). The 12-month abstinence was the lowest AUC for total SF-12 and both factors. As reported in Appendix A, many of the items on the SF-12 yielded AUC's $\geq$ 0.650 at post-treatment. Only one item on the SF-12 (item 6A: "felt calm or peaceful") yielded AUC's $\geq$ 0.650 in predicting the composite clinical outcome of heavy or lower risk at the 12-month follow-up.

Psychosocial Functioning Inventory. The items of the PFI administered in MATCH are consistent with three of the factors in the original conceptualization of the PFI's construction (Feragne et al., 1983). Those three factors are: Subjective Role Performance, Overall Social Role Performance, and Housemate/Roommate Role. In CFA, results indicated this 3-factor model fit acceptably well and replicated in the second

split half sample (RMSEA = 0.052 (90% CI: 0.047, 0.057); CFI = 0.933; TLI = 0.923). However, model fit was substantially poorer in the configural invariance model (RMSEA = 0.042 (90% CI: 0.041, 0.044); CFI = 0.822; TLI = 0.811). Because CFI and TLI fit indices decreased below *a priori* cutoffs for adequate model fit, configural invariance fit was determined to be too inadequate to warrant further invariance testing. Accordingly, the 3-factor model of the items of the PFI administered in MATCH was non-invariant over baseline and post-treatment timepoints. These poor findings may be due to the fact that MATCH administered an abbreviated version of the PFI.

Despite these findings, the PFI performed reasonably well on internal consistency and convergent validity analyses. Except for the Housemate/Roommate role factor (α = 0.531), internal consistency was good: total PFI $\alpha = 0.867$, Subjective Role Performance factor $\alpha = 0.817$, Overall Social Role Performance factor $\alpha = 0.818$ (see Table 8). Further, the convergent validity results were all significant ($p < 0.001$) and in the expected direction. The PFI was significantly, negatively correlated with SSTI (r = -0.200) and BDI (r = -0.380).

The PFI performed poorly with respect to sensitivity/specificity and adequately detected only some consumption outcomes, as detailed in Table 17. Total PFI summary scores only adequately detected 4 of 15 post-treatment outcomes (AUC's = 0.538-0.700) and failed to adequately detect any consumption outcomes at 12-month follow-up (AUC's = 0.533-0.600). Since CFA results indicated an adequately-fitting 3-factor structure, these 3 factor summary scores were examined via ROC curve analyses, yielding very few positive results. The Subjective Role Performance factor failed to adequately detect any post-treatment (AUC's = 0.522-0.644) or 12-month follow-up

outcomes (AUC's = 0.517-0.582). The Overall Social Role Performance factor adequately detected only 1 of 15 post-treatment consumption outcomes (AUC's = 0.520-0.663) and the Housemate/Roommate Role factor only adequately detected 2 of 15 post-treatment outcomes (AUC's = 0.537-0.660). Neither the Overall Social Role Performance factor nor the Housemate/Roommate Role factor adequately detected any 12-month follow-up outcomes (AUC's = 0.517-0.582, 0.531-0.585). There was no consistent pattern for which post-treatment or 12-month follow-up outcomes yielded highest AUC's when detected by the PFI total summary score or any of the factors, but 2+ level change in WHO risk level since baseline (calculated via DPD) and 12-month abstinence were the lowest AUC's for all 4 PFI variables. As reported in Appendix A, only 2 items adequately detected any consumption outcomes: Item 11 (spousal/mate overall role performance; Social Role Performance factor) and 19 (housemate/roommate overall role performance).

Alcohol-Abstinence Self-Efficacy Scale. The Alcohol-Abstinence Self-Efficacy Scale (AASE; DiClemente et al., 1994) was tested in both COMBINE and MATCH for the present study. Since the bulk of previous studies have found a 4-factor model in previous administrations of the AASE, this model was tested via CFA. This model provided an adequate fit in COMBINE (replication half fit indices: RMSEA = 0.050 (90% CI: 0.048, 0.053); CFI = 0.919; TLI = 0.914), but failed to provide an adequate fit in the overall MATCH dataset (split half 1 sub-sample fit indices: RMSEA = 0.060 (90% CI: 0.058, 0.062); CFI = 0.866; TLI = 0.857). An additional CFA was conducted to examine the fit of this 4-factor model in each treatment arm (aftercare and outpatient) for three reasons: 1) known differences in participant characteristics in the treatment arms of

Project MATCH, 2) the fact that the previously published 4-factor model of the AASE fit acceptably well in COMBINE, and 3) that the present study aimed for measurement preservation rather than data-driven methodology that could limit generalizability of findings. As shown in Table 7, the model fit adequately in the outpatient treatment arm of MATCH (replication half fit indices: RMSEA = 0.050 (90% CI: 0.047, 0.053); CFI = 0.931; TLI = 0.926). Moreover, this 4-factor model was invariant across time through strong invariance in both COMBINE and MATCH: COMBINE strong invariance RMSEA = 0.033 (90% CI: 0.032, 0.034); CFI = 0.953; TLI = 0.954; MATCH outpatient arm strong invariance RMSEA = 0.027 (90% CI: 0.026, 0.029); CFI = 0.917; TLI = 0.914.

Despite the strong invariance across baseline and post-treatment timepoints that replicated in both COMBINE and MATCH, however, other properties of the AASE were less promising. As described in Table 8, internal consistency of the overall AASE was good ($\alpha = 0.752$ in COMBINE; $\alpha = 0.841$ in MATCH) but internal consistency of each of the 4 factors was sub-optimal in both COMBINE and MATCH (COMBINE, MATCH: Negative Affect factor $\alpha = 0.356, 0.557$; Social/Positive factor $\alpha = 0.341, 0.460$; Physical & Other Concern factor $\alpha = 0.162, 0.546$; and Withdrawal & Urges factor $\alpha = 0.279, 0.458$). However, internal consistency is greatly influenced by the number of items contained in a measure or factor, so these low internal consistency reliability values may be because the AASE is a brief questionnaire.

More importantly, there were mixed findings for the convergent validity of the AASE. In COMBINE and MATCH, the AASE and its 4 factors were only significantly correlated with some of the Form 90 consumption variables and since the AASE purports

to measure craving and temptation to drink, one would expect all Form 90 consumption variables to be significantly correlated with the AASE. In COMBINE and MATCH, the total AASE correlations were: PDA r = -0.012 (p > 0.05) and r = -0.063 (p < 0.01); PHDD r = 0.011 (p > 0.05) and r = 0.062 (p < 0.05); DDD r = 0.079 (p < 0.01) and r = 0.032 (p > 0.05); MXD r = 0.104 (p < 0.001) and r = 0.061 (p < 0.05); and DPD r = 0.070 (p > 0.05) and r = 0.067 (p < 0.01). Bivariate correlations were similarly mixed for each of the 4 factors in COMBINE and MATCH.

Sensitivity/specificity of the AASE and the 4 subscales were all below the AUC $\geq$ 0.650 cutoff for all post-treatment and 12-month follow-up outcomes in both COMBINE and MATCH (detailed in Tables 18 and 19). This suggests the total AASE summary score and 4 factors are markedly poor at detecting concurrent and future consumption outcomes.

One potential exception to the poor sensitivity/specificity of the AASE in COMBINE and MATCH was the notably strong AUC values for the Temptation and Confidence subscales of the AASE. The summary scores of items in each of these 2 categories of items (Confidence and Temptation) were examined in terms of their sensitivity/specificity (presented in Tables 18 and 19). In the COMBINE Study, the summary of Confidence items and summary of Temptation items each adequately detected all concurrent (AUC's = 0.674-0.879, 0.669-0.858) and 12-month follow-up consumption outcomes (AUC's = 0.664-0.732, 0.690-0.734). In Project MATCH, the summary of Confidence items adequately detected 13 of 15 post-treatment (AUC's = 0.611-0.733) and 10 of 11 12-month follow-up consumption outcomes (AUC's = 0.649-0.690). Similarly, the summary of Temptation items in MATCH adequately detected 11

of 15 post-treatment (AUC's = 0.537-0.655) and all 11 12-month follow-up consumption outcomes (AUC's = 0.655-0.712). No consistent pattern emerged for which post-treatment or 12-month consumption outcomes yielded highest AUC's for the AASE Confidence or Temptation subscales across COMBINE and MATCH, except that 12-month composite clinical outcome of moderate or lower risk had the highest AUC's for both Confidence and Temptation subscales in both COMBINE and MATCH.

Consideration was given to these two summary scores via preliminary CFA testing of the AASE as a 2-factor solution; however, model fit was inadequate and not explored further (COMBINE split half 1 fit indices: RMSEA = 0.068 (90% CI: 0.065, 0.070); CFI = 0.897; TLI = 0.889; MATCH full sample split half 1 fit indices: RMSEA = 0.053 (90% CI: 0.050, 0.055); CFI = 0.856; TLI = 0.848). Future research should investigate these two summary scores further given their strong sensitivity/specificity for detecting consumption outcomes.

The Addiction Severity Index. Because the original ASI was not administered in Project MATCH, numerous strategies were employed to identify a factor model to test via CFA (detailed in Table 7). First, a 3-factor model based on the structure described by McLellan and colleagues (1992) was tested and failed to converge. Then, that model was altered to account for the restricted distributions of data in MATCH and the high intercorrelations between items; this model also failed to converge. Next, item question content and data distributions were explored closely to identify overlapping question content and to identify response categories that could be collapsed to mitigate potential data sparseness. No alternative factor structures emerged successfully from these procedures.

The total ASI score, as administered in Project MATCH, also performed poorly from other psychometric perspectives. The total ASI internal consistency was poor (α = 0.327), which is unsurprising given the content of the items differed substantially from one category (e.g., legal problems) to the next (e.g., psychological problems). Convergent validity results yielded mixed findings whereby the total ASI score was significantly correlated with some but not all Form 90 consumption variables. Specifically, the ASI was significantly correlated with DPD and MXD (r = 0.068, $p < 0.01$; r = 0.064, $p <$ 0.01), but these low correlation values may have been significant due to the large sample size of Project MATCH. All other convergent validity results were non-significant, including other consumption variables (e.g., PDA, PHDD), temptation items, PFI, ESI, and Alcoholic Anonymous Involvement (p's < 0.05). Finally, the ASI failed to detect any post-treatment or 12-month follow-up consumption outcomes over the AUC $\geq$ 0.650 level. Post-treatment consumption outcomes were detected at AUC values between 0.494 and 0.612; 12-month follow-up consumption outcomes were detected at AUC values between 0.492 and 0.544 (detailed in Table 20). Accordingly, the ASI as administered in Project MATCH had poor construct validity, internal consistency, convergent validity, and sensitivity/specificity to detect consumption outcomes.

ROC Curve Analyses of Consumption Variables

Results indicated that both PDA and PHDD were largely adequately sensitive/specific to detect consumption-based outcomes per the AUC $\geq$ 0.65 *a priori* cutoff (presented in Table 21). In the COMBINE Study, post-treatment outcome AUC values for PDA and PHDD ranged from 0.670 to 0.988. AUC values for both PDA and PHDD were largest for detecting specific WHO risk levels and composite clinical scores

and lowest for detecting changes in WHO risk level since baseline timepoint. These overall patterns were replicated in the post-treatment MATCH data (month 3) where AUC values ranged from 0.610 to 0.994 with specific WHO risk levels and composite clinical scores had highest AUC's and changes in WHO risk had lowest AUC's.

For ROC curve analyses of post-treatment PDA and PHDD detecting 12-month follow-up consumption outcomes, all AUC values were lower than for those of congruent timepoints in COMBINE and MATCH (described above). AUC values for PDA and PHDD detecting 12-month follow-up consumption outcomes ranged from 0.682 to 0.798 in COMBINE data and 0.668 to 0.754 in MATCH. Given the small range in AUC values, no clear pattern emerged as to which 12-month follow-up outcomes were most or least easily detected via PDA or PHDD in COMBINE or MATCH.

Discussion

The current study was a secondary analysis of data from the two largest randomized clinical trials for alcohol use disorder (AUD) ever conducted (the COMBINE Study (Anton et al., 2006) and Project MATCH (Project MATCH Research Group, 1996)) to examine the measurement stability, internal consistency reliability, construct and convergent validity, sensitivity, and specificity of numerous non-consumption outcome measures. These analyses constitute some of the most broad and rigorous analyses ever conducted with these measures of non-consumption constructs. Such extensive analyses of measures administered in the COMBINE Study and Project MATCH have highlighted several promising measures for use in future research. Although no one construct performed consistency well across all measures tested, individual measures stood out as viable options for measuring a variety of constructs. The

Brief Symptom Inventory (BSI), Beck Depression Inventory (BDI), and the brief World Health Organization Quality of Life measure (WHOQOL-BREF) as administered in COMBINE or MATCH demonstrated strong viability for future use as measures of psychological health and quality of life. Importantly, these three measures had good construct validity and were invariant across time, which supports their use for examining pre- to post-treatment changes in these constructs. Moreover, these three measures adequately detected at least some consumption outcomes, as illustrated via ROC curve analyses. These administrations of the BSI, BDI, and WHOQOL-BREF in COMBINE and MATCH also had good convergent validity and internal consistency reliability.

Other measures administered in COMBINE and MATCH yielded mixed findings. Measures that showed the most promise for use in AUD treatment research were the Obsessive-Compulsive Drinking Scale (OCDS) and the Drinker Inventory of Consequences (DrInC) due to their having evidence for at least partial measurement invariance across time and at least some evidence for their sensitivity/specificity to detect consumption outcomes. Since the main goal of the present study was to identify measures that may be viable for use in examining pre- and post-treatment changes for AUD treatment research, more credence was given to these results (i.e., invariance, sensitivity/specificity) as more directly related to the research question. As such, the OCDS and DrInC were shown to have the most promise and minor refinement may further improve their suitability for use in AUD treatment research.

Findings from the present study also highlighted a few measures that warrant further investigation and refinement for use in AUD treatment research. The Spielberger State-Trait Inventory (SSTI) had strong internal consistency and construct validity, but

more research is needed to improve the convergent validity and sensitivity/specificity. The Short Form Health Survey (SF-12) and the Psychosocial Functioning Inventory (PFI) each had acceptable convergent validity, strong internal consistency, strong convergent validity, and had acceptable sensitivity/specificity but more research is needed to establish more robust factor structures that will be invariant across time. The Alcohol Abstinence Self-Efficacy Scale (AASE) had acceptable construct validity and was invariant across time in both COMBINE and MATCH, had acceptable internal consistency, and had acceptable convergent validity. However, the AASE had poor sensitivity/specificity in both COMBINE and MATCH for detecting all post-treatment and 12-month follow-up consumption outcomes. However, ROC curve results from the present study highlighted the good sensitivity/specificity of the AASE Confidence and Temptation subscales and future research may be able to use these two subscales rather than the total AASE score to examine pre- and post-treatment differences in self-efficacy and temptation to drink.

Only the Addiction Severity Index (ASI), as administered in MATCH, failed to perform adequately across all psychometric analyses. Data distributions appeared to be driving the poor performance of the ASI, which likely indicates that aggregating data from the ASI items is counter to its utility. The content of individual items in the ASI may still be useful in clinical settings and in research from an individual participant perspective. For instance, it may be interesting and informative clinically and in research to know the number of times an individual client or participant has experienced each of the psychiatric symptoms or engaged in specific illegal activities that are each assessed via the ASI. However, summarizing items and averaging scores across clients based on

the published ASI factor structure does not appear to be advisable based on the present findings. These findings are consistent with previously published cautions against the ASI (DeJong et al., 1995).

The limitations highlighted by the ASI and the mixed results for many other measures administered in the COMBINE Study and Project MATCH highlight an important gap in current AUD treatment research: a need for more rigorously developed assessment tools. Although the COMBINE Study and Project MATCH utilized the state of the science assessment tools, many of the measures that were used in these studies continue to be used in current research despite a lack of clear psychometric strength. The present results highlight that many non-consumption measures are viable or promising for use in future research; however, non-consumption measures could be even stronger psychometrically. First, more stringent, empirically-driven methodologies should be used to develop measures whereas many currently used measures were developed solely by researchers. Ideally, measurement development is a multiphase process where a measure is developed iteratively based on data from multiple sources, including experts, researchers, and study populations themselves (e.g., the Delphi process; Polit & Hungler, 1999). Second, more extensive and appropriate psychometric analyses should be conducted for measures before they become widely used. Many of the previously published studies of measures' psychometric appropriateness of use in AUD treatment research stemmed from internal consistency and principal components analyses (PCA). However, many psychometricians would argue that internal consistency is not the sole, nor best, metric for evaluating reliability. Further, PCA is ill-suited for examining factor structure and construct validity (e.g., Floyd & Widaman, 1995). Instead, other forms of

reliability and validity testing could be used in initial measurement development (e.g., test-retest reliability, confirmatory factor analyses; DeVellis, 2012). Unsurprisingly, the measures that performed the best in the present study (the Brief Symptom Inventory, Beck Depression Inventory, and the World Health Organization Quality of Life, Brief version) were three of the more rigorously developed and evaluated measures utilized in the COMBINE Study and Project MATCH. With more rigorously developed measures, it is entirely possible that non-consumption variables would yield consistently strong results and AUD researchers would no longer need to use consumption variables as "surrogates" for these more clinically meaningful, non-consumption constructs.

With respect to consumption outcomes being necessary as "surrogates" for non-consumption outcomes, it is important to note that the present study found evidence to contradict this assumption. Many non-consumption variables were not exceptionally bad at detecting consumption outcomes and numerous non-consumption measures had large effect sizes, even when aggregated across the overall samples in the COMBINE Study and Project MATCH. For example, the Drinker Inventory of Consequences (DrInC) and the 3 factors identified in the present study, the Obsessive-Compulsive Drinking Scale (OCDS) and the 4 factors used in the present study, and the Confidence and Temptation subscales of the Alcohol-Abstinence Self-Efficacy (AASE) all had good sensitivity/specificity for detecting consumption outcomes and demonstrated large effect sizes, especially in the COMBINE Study. These findings are despite the long-standing claim that consumption variables must be used as a "surrogate" for more clinically meaningful non-consumption outcomes due to the inability of non-consumption outcomes to be sensitive to change (e.g., FDA, 2015, p. 2). Although area under the curve

(AUC) values were generally higher for consumption variables detecting consumption outcomes than those for non-consumption variables, several non-consumption variables performed adequately and several non-consumption variables had AUC values comparable to those for consumption variables. For example, the 3-factor model of the DrInC described in the present study yielded AUC values similar to those of consumption variables, especially in the COMBINE Study. Some AUC values for the DrInC subscales in COMBINE were as high as 0.939, which corresponds to an almost perfect detection/discrimination ability. The DrInC corresponds to alcohol-related consequences that are far more aligned with diagnostic criteria for AUD than consumption variables, which have never been part of diagnostic criteria used by the American Psychiatric Association.

Another interesting finding was a general pattern of which consumption outcomes were adequately detected or not detected by non-consumption variables. The majority of non-consumption variables had their highest AUC values for composite clinical outcome scores (especially the heavy drinking/problems or lower risk level) for both post-treatment and 12-month follow-up timelines. The post-treatment outcome that yielded the lowest AUC's by the majority of non-consumption variables was the 2+ levels of change in WHO risk level from baseline. The 12-month consumption outcome that yielded the lowest AUC's by many non-consumption variables was 12-month abstinence. These patterns were largely consistent with consumption variable ROC curve results (PDA and PHDD ROC curve analyses), suggesting that post-treatment WHO Risk Level changes from baseline and 12-month abstinence may not be the best outcomes to use in evaluating treatment efficacy given the difficulty of accurately detecting these outcomes. Instead,

the present findings suggest future research evaluate treatment efficacy based on the most readily detected outcome: composite clinical score. The composite clinical scores that systematically yielded the highest AUC's for both consumption and non-consumption variables were moderate or lower risk and heavy drinking/problems or lower risk.

Recent studies have suggested the shift in WHO risk level, as a consumption based outcome, may provide a viable alternative to abstinence and percent subjects with no heavy drinking days as endpoints for alcohol clinical trials (Hasin et al., in press; Witkiewitz et al., 2017). In the current study, the method of calculation of the WHO risk level variables impacted how those consumption outcomes were detected by both non-consumption variables as well as PDA and PHDD. Specifically, WHO risk levels calculated via drinks per drinking day (DDD), maximum number of drinks consumed in the 90-day window (MXD) and drinks per day in the assessment window (averaged across drinking and abstinent days; DPD) yielded inconsistent AUC values across variables. For some variables, AUC values were relatively equitable across DDD, MXD, and DPD calculations. However, for some variables such as the Beck Depression Inventory scores and Drinker Inventory of Consequences (in Project MATCH especially), AUC values varied substantively between DDD, MXD, and DPD calculations of WHO risk levels. These findings suggest more research is needed to establish which calculation method (DDD, MXD, or DPD) is most stable across various conditions. Further, the fact that sensitivity/specificity can vary so meaningfully in consumption outcomes is another indication that consumption variables may not be as dependable as previously assumed by AUD treatment researchers.

Limitations

The primary limitation of the present study was that findings are specific to administration of the measures in the COMBINE Study and Project MATCH. Many of the measures examined in the present study were abbreviated versions of the full measures. For example, the WHOQOL-BREF as administered in COMBINE was one item short of the full WHOQOL-BREF. It is unclear why one item was omitted in the COMBINE Study, though it seems likely the item was omitted erroneously since it is the last item of the measure. However, the WHOQOL-BREF as administered in COMBINE performed well despite this missing item. Other abbreviated measures did not perform so well. Notably, the ASI, SSTI, and PFI were all abbreviated versions of these measures, which may or may not account for the poor results found for these measures in MATCH compared to previous studies (e.g., Feragne, 1983; McLellan et al., 1980). The results from the present study provide evidence that abbreviated versions of measures should be thoroughly explored psychometrically prior to their use in AUD treatment research.

The present findings are also limited by the availability of certain measures at different assessment periods. The post-treatment WHOQOL-BREF in COMBINE was administered at week 26, which fails to map on to the week 16 data that were examined for other measures. This inconsistency restricted the examination of post-treatment ROC curves for WHOQOL-BREF. Moreover, baseline data were only available for the SSTI, which prevented the examination of measurement invariance of the 4-factor model that fit excellently in MATCH data.

Another limitation of the present study was that the administration of identical measures may have varied between COMBINE and MATCH. For instance, the ROC curve analyses of the DrInC differed between COMBINE and MATCH, which may be

the result of how this measure was administered. The DrInC was inconsistently administered to individuals who reported total abstinence during the assessment window in Project MATCH and was administered to all individuals in the COMBINE Study, regardless of drinking status. These differences in assessment administration could explain why ROC curve results were inconsistent between the two studies. Further, the bivariate correlations had mixed results than were predicted for MATCH (but not COMBINE), which may indicate the DrInC as administered in MATCH was unstable psychometrically. Future administrations of the DrInC in other longitudinal studies may help elucidate these nuances to identify if the DrInC is poorly suited for longitudinal research or if methodology employed in MATCH was ill-suited for the DrInC.

Similarly, two of the three measures that performed most strongly in the present analyses (the BSI and WHOQOL-BREF) were only administered in the COMBINE Study. COMBINE was a smaller and more homogeneous sample than Project MATCH (primarily due to more strict alcohol consumption and diagnostic inclusion criteria in COMBINE). It is unclear how these two measures may perform in more heterogeneous samples. The current findings may not be generalizable to other studies and replication of the present findings is warranted.

Finally, findings from the present study may be limited by the fact that the full samples in COMBINE and MATCH were used. Analyses from Project MATCH included both aftercare and outpatient arms, which had different demographic characteristics and different levels of AUD severity (the aftercare arm had greater baseline drinking and severity than the outpatient arm). These differences may be one factor that could have

hurt the psychometrics of examined measures. Findings from the present study should be examined in other studies to attempt replication.

Future Directions

Findings from the present study highlight strengths in several non-consumption measures for use in AUD treatment research. However, the present study also demonstrates that not all measures are created equally. For instance, although both the WHOQOL-BREF and the SF-12 are purported to measure quality of life, the SF-12 was non-invariant across time. Examining the items reveals that many items are "double-barreled," meaning they essentially ask more than one question per item (e.g., "…how much of the time has your physical health or emotional problems interfered with your social activities…"), which is particularly problematic when a measure was conceptualized as measuring two separate factors: physical and psychological health. These "double-barreled" items may be problematic for invariance of a measure across time because participants may respond predominantly to one aspect of the question item (e.g., physical health interference in social activities) at one timepoint and to another aspect of the question (e.g., emotional problem interference in social activities) at another timepoint. Careful development of question items comprising measures to avoid "double-barreled" items and other pitfalls of measurement development and refinement are essential in future research (see recommendations by DeVellis, 2012; Miller et al., 2009). As the measures currently exist, it is recommended that future research use the WHOQOL-BREF rather than the SF-12 to assess changes in quality of life over time.

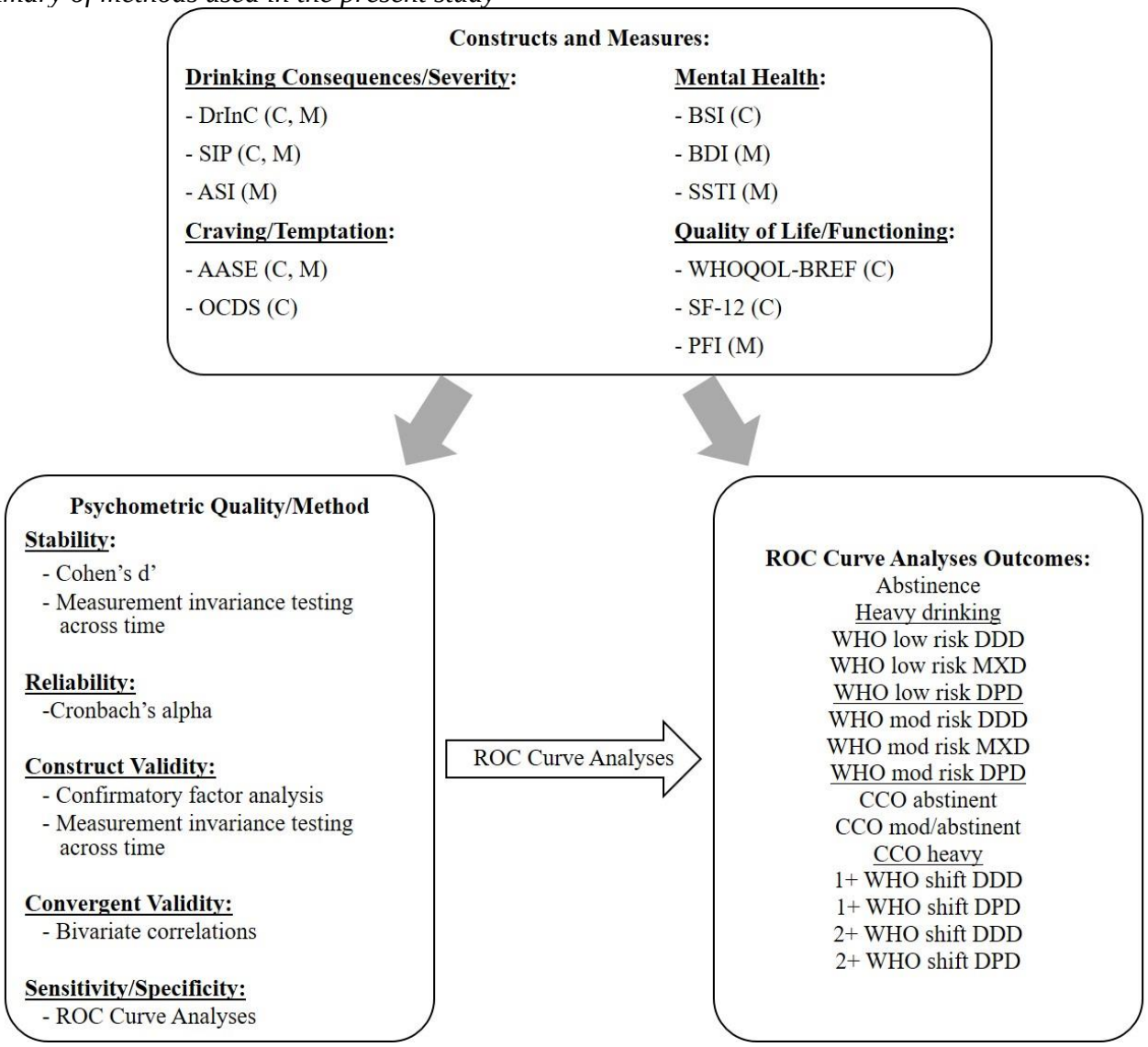
In addition to highlighting the need for future research to refine existing and develop new measures of non-consumption constructs, future research should prioritize

consistent administration of assessment tools. Assessments should be administered at multiple timepoints whenever possible, so future research can expand the extant knowledge of measurement invariance of commonly used assessment tools. The measurement non-invariance of several measures examined in the present study highlights the need for future research on testing for measurement invariance across time as well as across other domains such as gender, race, treatment sample, and other demographic characteristics.

Most importantly, the present study highlights the promise of several non-consumption measures as viable means of examining clinically meaningful outcomes in AUD treatment research beyond changes in alcohol consumption alone. Accordingly, future researchers are supported in expanding their definitions of treatment success to include at least psychological health (via the BSI and BDI) and quality of life (via the WHOQOL-BREF). Additional research is needed to continue this vein of research to discover and develop other potentially viable measures that map onto definitions of treatment success used by clinicians as well as clients and their loved ones (Kaskutas et al., 2014; Neale et al., 2014). Future research need not rely solely upon consumption-based outcome variables due to a lack of information on which measures are appropriate to use in AUD treatment research. Several non-consumption measures showed promise and more measures could be refined and developed to allow researchers to further enhance their ability to examine treatment efficacy using non-consumption outcomes.

Figure 1

Summary of methods used in the present study



Note. (C) = COMBINE Study; (M) = Project MATCH. Abbreviated measure names are: Drinker Inventory of Consequences (DrInC), Short Inventory of Problems (SIP), Addiction Severity Index (ASI), Alcohol Abstinence Self-Efficacy Scale (AASE), Obsessive-Compulsive Drinking Scale (OCDS), Brief Symptom Inventory (BSI), Beck Depression Inventory (BDI), Spielberger State-Trait Inventory (SSTI), World Health Organization Quality of Life brief version (WHOQOL-BREF), Health Survey (SF-12), Psychosocial Functioning Inventory (PFI). Abbreviated terms are: WHO = World Health Organization; DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days), CCO = Composite Clinical Outcome, mod = moderate.

Table 1

Demographic, design, and exclusion criteria for COMBINE and MATCH

	COMBINE	MATCH
Demographic characteristic		
Sample size	1383	1726
Gender -- % Male	69.1%	75.7%
Age – Mean (SD)	44.4 (10.2)	40.2 (10.9)
Ethnicity -- % White	76.8%	80.0%
Marital status -- % Married, in relationship	46.3%	41.4%
Employment status -- % Full or part-time	71.4%	82.1%
Higher education or equivalent	70.6%	53.4%
Design		
Randomization to treatment	9 groups	3 groups
Length of treatment	16 weeks	12 weeks
Follow-up assessments	12 months	12 months
Exclusion criteria		
Age	18+	18+
Meet criteria for abuse/dependence	Past year	Past year
Reading level	Literate	6^{th} grade
Comorbid psychiatric diagnoses	X	X
Unable to identify collateral informant		X
Severe cognitive impairment		X
Residential instability		X
Other illicit drug dependence	X	X

Table 2

Outcome variables and their corresponding measures and timepoints available in COMBINE and MATCH

Variable	Measure in COMBINE (C) or MATCH (M)	Timepoint(s) assessed in COMBINE (C) or MATCH (M)
Alcohol/ Drug use	Form-90 (C, M)	Pre-treatment (C, M), During-treatment (C, M), Post-treatment (C, M)
Drinking Consequences /Severity	Drinker Inventory of Consequences/ Short Inventory of Problems (DrInC/SIP; C, M)	Pre-treatment (C, M), During-treatment (C, M), Post-treatment (C, M)
	Addiction Severity Index (ASI; M)	Pre-treatment (M), During-treatment (M), Post-treatment (M)
Mental health	Brief Symptom Inventory (BSI; C)	Pre-treatment (C), During-treatment (C), Post-treatment (C)
	Beck Depression Inventory (BDI; M)	Pre-treatment (M), During-treatment (M), Post-treatment (M)
	Spielberger State-Trait Inventory (SSTI; M)	Pre-treatment (M), During-treatment (M), Post-treatment (M)
Craving/ Temptation	Alcohol Abstinence Self-Efficacy (AASE; C, M)	Pre-treatment (C, M), During-treatment (C, M), Post-treatment (C, M)
	Obsessive-Compulsive Drinking Scale (OCDS; C)	Pre-treatment (C), During-treatment (C), Post-treatment (C)
	Temptation/craving items during treatment (M)	Pre-treatment (M), During-treatment (M), Post-treatment (M)

Quality of life/ Functioning	World Health Organization Quality of Life (WHOQOL-BREF; C)	Pre-treatment (C), During-treatment (C), Post-treatment (C)
	Health Survey (SF-12; C)	Pre-treatment (C), During-treatment (C), Post-treatment (C)
	Psychosocial Functioning Inventory (PFI; M)	Pre-treatment (M), During-treatment (M), Post-treatment (M)

Table 3

Outcome variables and the measures hypothesized to have convergent (+, significant positive correlation; -, significant negative correlation) validity in COMBINE and MATCH

	Form-90	DrInC, ASI	BSI, SSTI, BDI	OCDS, Temptation/ craving items	PFI, WHOQOL-BREF	SF-12	ESI	AASE	AAI
Form-90		+		+				-	-
DrInC, ASI	+			+	-	-	-		-
BSI, SSTI, BDI					-	-	-		
OCDS, Temptation/ craving	+	+							
PFI, WHOQOL-BREF		-	-						
SF-12		-	-						
ESI		-	-						
AASE	-								
AAI	-	-							

Note. Abbreviated measure names are: Drinker Inventory of Consequences (DrInC), Addiction Severity Index (ASI), Brief Symptom Inventory (BSI), Spielberger State-Trait Inventory (SSTI), Beck Depression Inventory (BDI), Obsessive-Compulsive Drinking Scale (OCDS), Psychosocial Functioning Inventory (PFI), World Health Organization Quality of Life brief version (WHOQOL-BREF), Health Survey (SF-12), Employment Status and Income (ESI), Alcohol Abstinence Self-Efficacy Scale (AASE), and Alcoholics Anonymous Involvement (AAI).

Table 4

Descriptive statistics and Cohen's d effect sizes for measures used in the present study at baseline, post-treatment (post-tx), and 12-month follow-up (12mo)

			N	M (SD)	Cohen's d
Percent Days Abstinent: COMBINE Study		Baseline:	1383	21.41 (22.50)	Baseline to Post-tx: 1.809
		Post-Tx:	1288	72.66 (33.49)	Post-tx to 12mo: 0.277
		12mo:	1099	62.63 (39.12)	Baseline to 12mo: 1.331
Percent Days Abstinent: Project MATCH		Baseline:	1725	30.90 (29.96)	Baseline to Post-tx: 1.786
		Post-Tx:	1657	83.17 (28.51)	Post-tx to 12-moth: 0.208
		12mo:	1594	76.69 (33.55)	Baseline to 12mo: 1.443
Percent Heavy Drinking Days: COMBINE Study		Baseline:	1383	70.52 (26.57)	Baseline to Post-tx: 1.919
		Post-Tx:	1288	17.54 (28.69)	Post-tx to 12mo: 0.275
		12mo:	1171	26.20 (34.27)	Baseline to 12mo: 1.461
Percent Heavy Drinking Days: Project MATCH		Baseline:	1725	63.18 (31.43)	Baseline to Post-tx: 1.780
		Post-Tx:	1657	12.46 (25.09)	Post-tx to 12mo: 0.156
		12mo:	1594	16.71 (29.17)	Baseline to 12mo: 1.530
DrInC: COMBINE Study		Baseline:	1381	47.61 (20.42)	Baseline to Post-tx: 1.735
		Post-Tx:	1098	13.36 (18.85)	Post-tx to 12mo: 0.322
		12mo:	965	19.89 (21.81)	Baseline to 12mo: 1.320
	Physical Health Subscale	Baseline:	1381	9.28 (4.36)	Baseline to Post-tx: 1.607
		Post-Tx:	1098	2.61 (3.87)	Post-tx to 12mo: 0.320
		12mo:	965	3.95 (4.53)	Baseline to 12mo: 1.203
	Interpersonal Subscale	Baseline:	1381	10.06 (6.01)	Baseline to Post-tx: 1.389
		Post-Tx:	1098	2.60 (4.44)	Post-tx to 12mo: 0.302
		12mo:	965	4.06 (5.25)	Baseline to 12mo: 1.051

	Intrapersonal Subscale	Baseline:	1381	14.44 (5.66)	Baseline to Post-tx: 1.749
		Post-Tx:	1098	4.45 (5.78)	Post-tx to 12mo: 0.292
		12mo:	965	6.26 (6.66)	Baseline to 12mo: 1.343
	Impulse Subscale	Baseline:	1381	7.56 (4.25)	Baseline to Post-tx: 1.414
		Post-Tx:	1098	2.11 (3.29)	Post-tx to 12mo: 0.318
		12mo:	965	3.24 (3.83)	Baseline to 12mo: 1.058
	Social Responsibility Subscale	Baseline:	1381	6.27 (4.15)	Baseline to Post-tx: 1.260
		Post-Tx:	1098	1.60 (3.04)	Post-tx to 12mo: 0.242
		12mo:	965	2.39 (3.46)	Baseline to 12mo: 1.000
DrInC: Project MATCH		Baseline:	1703	52.63 (23.32)	Baseline to Post-tx: 0.680
		Post-Tx:	985	35.86 (26.78)	Post-tx to 12mo: 0.323
		12mo:	789	27.50 (24.70)	Baseline to 12mo: 1.057
	Physical Health Subscale	Baseline:	1626	9.48 (4.94)	Baseline to Post-tx: 0.666
		Post-Tx:	966	6.14 (5.13)	Post-tx to 12moFU: 0.204
		12mo:	818	5.12 (4.85)	Baseline to 12mo: 0.945
	Interpersonal Subscale	Baseline:	1558	12.21 (6.98)	Baseline to Post-tx: 0.568
		Post-Tx:	942	8.17 (7.34)	Post-tx to 12mo: 0.295
		12mo:	807	6.09 (6.72)	Baseline to 12mo: 0.888
	Intrapersonal Subscale	Baseline:	1626	14.51 (6.02)	Baseline to Post-tx: 0.653
		Post-Tx:	964	10.38 (6.81)	Post-tx to 12mo: 0.297
		12mo:	819	8.31 (7.16)	Baseline to 12mo: 0.965
	Impulse Subscale	Baseline:	1572	8.69 (5.10)	Baseline to Post-tx: 0.503

		Post-Tx:	967	6.06 (5.43)	Post-tx to 12mo: 0.540
		12mo:	820	3.41 (4.21)	Baseline to 12mo: 1.097
	Social Responsibility Subscale	Baseline:	1598	7.49 (4.71)	Baseline to Post-tx: 0.591
		Post-Tx:	964	4.71 (4.69)	Post-tx to 12mo: 0.004
		12mo:	822	4.73 (4.56)	Baseline to 12mo: 0.592
ASI Psychiatric Severity: Project MATCH		Baseline:	1714	0.21 (0.20)	Baseline to Post-tx: 0.358
		Post-Tx:	1566	0.14 (0.19)	Post-tx to 12mo: 0.053
		12mo:	1554	0.13 (0.19)	Baseline to 12mo: 0.410
ASI Family History: Project MATCH		Baseline:	1726	2.65 (48.04)	-
ASI Legal Status: Project MATCH		Baseline:	1726	141.90 (384.71)	-
OCDS: COMBINE Study		Baseline:	1383	26.60 (8.20)	Baseline to Post-tx: 1.762
		Post-Tx:	1101	11.25 (9.32)	-
		12mo:	2	22.50 (6.36)	-
AASE: COMBINE Study		Baseline:	1382	113.26 (15.39)	Baseline to Post-tx: 0.101
		Post-Tx:	1103	114.78 (14.71)	-
	Confidence Subscale	Baseline:	1377	2.61 (0.74)	Baseline to Post-tx: 1.078
		Post-Tx:	1100	3.50 (0.92)	-
	Temptation Subscale	Baseline:	1374	3.11 (0.78)	Baseline to Post-tx: 1.022
		Post-Tx:	1093	2.28 (0.85)	-
AASE: Project MATCH		Baseline:	1700	117.37 (21.48)	Baseline to Post-tx: 0.130
		Post-Tx:	1557	114.56 (21.87)	-
	Confidence Subscale	Baseline:	1662	3.06 (0.92)	Baseline to Post-tx: 0.469
		Post-Tx:	1528	3.51 (1.00)	-
	Temptation Subscale	Baseline:	1688	2.91 (0.90)	Baseline to Post-tx: 0.674

		Post-Tx:	1545	2.31 (0.88)	-
Temptation Item: Project MATCH		Baseline:	1531	2.66 (1.34)	-
AAI: Project MATCH		Baseline:	1624	4.29 (2.60)	-
BSI: COMBINE Study		Baseline:	1356	60.29 (10.91)	Baseline to Post-tx: 0.695
		Post-Tx:	1101	52.32 (12.11)	Post-tx to 12mo: 0.031
		12mo:	959	51.93 (13.18)	Baseline to 12mo: 0.702
	Somatic Subscale	Baseline:	1356	54.50 (10.37)	Baseline to Post-tx: 0.467
		Post-Tx:	1101	49.89 (9.24)	Post-tx to 12mo: 0.123
		12mo:	959	51.10 (10.14)	Baseline to 12mo: 0.331
	Obsessive Compulsive Subscale	Baseline:	1356	58.69 (10.75)	Baseline to Post-tx: 0.560
		Post-Tx:	1101	52.66 (10.78)	Post-tx to 12mo: 0.014
		12mo:	959	52.81 (11.38)	Baseline to 12mo: 0.534
	Interpersonal Sensitivity Subscale	Baseline:	1356	56.75 (10.96)	Baseline to Post-tx: 0.405
		Post-Tx:	1101	52.44 (10.24)	Post-tx to 12mo: 0.061
		12mo:	959	51.81 (10.45)	Baseline to 12mo: 0.459
	Depression Subscale	Baseline:	1356	61.55 (10.60)	Baseline to Post-tx: 0.603
		Post-Tx:	1101	55.15 (10.64)	Post-tx to 12mo: 0.001
		12mo:	959	55.16 (11.15)	Baseline to 12mo: 0.590
	Anxiety Subscale	Baseline:	1356	58.37 (11.48)	Baseline to Post-tx: 0.628
		Post-Tx:	1101	51.26 (11.11)	Post-tx to 12mo: 0.012
		12mo:	959	51.12 (11.44)	Baseline to 12mo: 0.632
	Hostility Subscale	Baseline:	1356	55.68 (9.65)	Baseline to Post-tx: 0.584
		Post-Tx:	1101	50.04 (9.67)	Post-tx to 12mo:

					0.103
		12mo:	959	49.32 (9.93)	Baseline to 12mo: 0.651
	Phobic Anxiety Subscale	Baseline:	1356	53.43 (9.45)	Baseline to Post-tx: 0.262
		Post-Tx:	1101	51.10 (8.13)	Post-tx to 12mo: 0.025
		12mo:	959	50.89 (8.40)	0.281
	Paranoia Subscale	Baseline:	1356	54.53 (10.45)	Baseline to Post-tx: 0.304
		Post-Tx:	1101	51.47 (9.61)	Post-tx to 12mo: 0.103
		12mo:	959	50.46 (9.97)	Baseline to 12mo: 0.397
	Psychoticism Subscale	Baseline:	1356	63.29 (10.09)	Baseline to Post-tx: 0.633
		Post-Tx:	1101	56.73 (10.69)	Post-tx to 12mo: 0.006
		12mo:	959	56.67 (11.00)	Baseline to 12mo: 0.632
BDI: Project MATCH		Baseline:	1618	10.17 (8.24)	Baseline to Post-tx: 0.322
		Post-Tx:	1532	7.57 (7.87)	Post-tx to 12mo: 0.045
		12mo:	1505	7.94 (8.40)	Baseline to 12mo: 0.268
SSTI: Project MATCH (full SSTI)		Baseline:	1553	27.70 (7.14)	-
SSTI: Project MATCH (items used in factor analyses)		Baseline:	1553	25.79 (6.67)	-
WHOQOL-BREF: COMBINE Study		Baseline:	1351	87.94 (13.44)	Baseline to Post-tx: 0.679
		Post-Tx:	1062	97.67 (15.38)	Post-tx to 12mo: 0.257
		12mo:	954	93.88 (14.06)	Baseline to 12mo: 0.434
	Physical Health Subscale	Baseline:	1351	27.29 (4.30)	Baseline to Post-tx: 0.420
		Post-Tx:	1060	29.11 (4.38)	Post-tx to 12mo: 0.037
		12mo:	954	28.95 (4.36)	Baseline to 12mo:

					0.384
	Psychological Health Subscale	Baseline:	1351	21.04 (3.97)	Baseline to Post-tx: 0.417
		Post-Tx:	1060	22.75 (4.27)	Post-tx to 12mo: 0.022
		12mo:	954	22.84 (4.03)	Baseline to 12mo: 0.451
	Social Subscale	Baseline:	1351	9.84 (2.63)	Baseline to Post-tx: 0.373
		Post-Tx:	1060	10.82 (2.62)	Post-tx to 12mo: 0.008
		12mo:	953	10.84 (2.52)	Baseline to 12mo: 0.389
	Environment Subscale	Baseline:	1351	29.77 (5.44)	Baseline to Post-tx: 0.280
		Post-Tx:	1060	31.31 (5.57)	Post-tx to 12mo: 0.007
		12mo:	954	31.27 (5.49)	Baseline to 12mo: 0.275
SF-12: COMBINE Study		Baseline:	1357	42.28 (7.14)	Baseline to Post-tx: 0.710
		Post-Tx:	1102	47.13 (6.43)	Post-tx to 12mo: 0.275
		12mo:	951	45.26 (7.18)	Baseline to 12mo: 0.416
	Physical Health Subscale	Baseline:	1346	0.27 (0.83)	Baseline to Post-tx: 0.129
		Post-Tx:	1099	0.37 (0.70)	Post-tx to 12mo: 0.244
		12mo:	948	0.18 (0.86)	Baseline to 12mo: 0.107
	Psychological Health Subscale	Baseline:	1346	-0.86 (1.10)	Baseline to Post-tx: 0.760
		Post-Tx:	1099	-0.07 (0.96)	Post-tx to 12mo: 0.130
		12mo:	948	-0.20 (1.04)	Baseline to 12mo: 0.614
PFI: Project MATCH		Baseline:	1695	50.91 (11.42)	Baseline to Post-tx: 0.532
		Post-Tx:	1556	56.88 (10.98)	-

Note. Abbreviated measure names are: Drinker Inventory of Consequences (DrInC), Addiction Severity Index (ASI), Brief Symptom Inventory (BSI), Spielberger State-Trait Inventory (SSTI), Beck Depression Inventory (BDI), Obsessive-Compulsive Drinking Scale (OCDS), Psychosocial Functioning Inventory (PFI), World Health Organization Quality of Life brief version

(WHOQOL-BREF), Health Survey (SF-12), Employment Status and Income (ESI), Alcohol Abstinence Self-Efficacy Scale (AASE), and Alcoholics Anonymous Involvement (AAI).

Table 5

Frequencies for dichotomous consumption outcome variables used in Receiver Operating Characteristic (ROC) Curve analyses at post-treatment (post-tx) and 12-month follow-up (12mo)

	COMBINE Study	Project MATCH
Post-tx Abstinence	0 = 829 1 = 459 *N* = 1288	0 = 807 1 = 850 *N* = 1657
Post-tx Heavy Drinking	0 = 858 1 = 630 *N* = 1288	0 = 1017 1 = 640 *N* = 1657
Post-tx WHO risk: low risk or lower risk computed via DDD	0 = 699 1 = 589 *N* = 1288	0 = 692 1 = 965 *N* = 1657
Post-tx WHO risk: low risk or lower risk computed via MXD	0 = 751 1 = 537 *N* = 1288	0 = 720 1 = 937 *N* = 1657
Post-tx WHO risk: low risk or lower risk computed via DPD	0 = 325 1 = 963 *N* = 1288	0 = 304 1 = 1353 *N* = 1657
Post-tx WHO risk: moderate risk or lower risk computed via DDD	0 = 544 1 = 744 *N* = 1288	0 = 592 1 = 1065 *N* = 1657
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0 = 641 1 = 647 *N* = 1288	0 = 645 1 = 1012 *N* = 1657
Post-tx WHO risk: moderate risk or lower risk computed via DPD	0 = 199 1 = 1089 *N* = 1288	0 = 219 1 = 1438 *N* = 1657
Post-tx Composite clinical Outcome: Abstinent	0 = 666 1 = 471 *N* = 1137	0 = 1057 1 = 599 *N* = 1656
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0 = 348 1 = 789 *N* = 1137	0 = 850 1 = 806 *N* = 1656
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0 = 98 1 = 1039 *N* = 1137	0 = 531 1 = 1125 *N* = 1656
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0 = 473 1 = 763 *N* = 1236	0 = 571 1 = 1069 *N* = 1640

1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0 = 261 1 = 975 *N* = 1236	0 = 335 1 = 1305 *N* = 1640
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0 = 686 1 = 550 *N* = 1236	0 = 784 1 = 856 *N* = 1640
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0 = 493 1 = 743 *N* = 1236	0 = 624 1 = 1016 *N* = 1640
12mo Abstinence	0 = 817 1 = 355 *N* = 1172	0 = 847 1 = 747 *N* = 1594
12mo Heavy Drinking	0 = 539 1 = 633 *N* = 1172	0 = 923 1 = 671 *N* = 1594
12mo WHO risk: low risk or lower risk computed via DDD	0 = 719 1 = 453 *N* = 1172	0 = 729 1 = 865 *N* = 1594
12mo WHO risk: low risk or lower risk computed via MXD	0 = 756 1 = 416 *N* = 1172	0 = 763 1 = 831 *N* = 1594
12mo WHO risk: low risk or lower risk computed via DPD	0 = 448 1 = 724 *N* = 1172	0 = 384 1 = 1210 *N* = 1594
12mo WHO risk: moderate risk or lower risk computed via DDD	0 = 581 1 = 591 *N* = 1172	0 = 615 1 = 979 *N* = 1594
12mo WHO risk: moderate risk or lower risk computed via MXD	0 = 646 1 = 526 *N* = 1172	0 = 682 1 = 912 *N* = 1594
12mo WHO risk: moderate risk or lower risk computed via DPD	0 = 323 1 = 849 *N* = 1172	0 = 255 1 = 1339 *N* = 1594
12mo Composite clinical Outcome: Abstinent	0 = 750 1 = 284 *N* = 1034	0 = 1024 1 = 581 *N* = 1605
12mo Composite clinical Outcome: Moderate drinking or abstinent	0 = 554 1 = 480 *N* = 1034	0 = 827 1 = 778 *N* = 1605

12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0 = 234 1 = 800 *N* = 1034	0 = 488 1 = 1117 *N* = 1605

Table 6

Summary of Psychometric Findings across Measures and Methods (0 = Poor Psychometric Qualities; 1 = Mixed or Modest Psychometric Qualities; and 2 = Acceptable to Excellent Psychometric Qualities)

Measure (Study: COMBINE (C) or MATCH (M))	ROC	Reliability	Convergent Validity	CFA	Invariance	Total Points/ Possible Points
Alcohol Abstinence Self-Efficacy Scale (AASE) (C; M)	0	1	1	1	2	5/10
AASE-Confidence (C; M)	2	N/A	N/A	N/A	N/A	2/2
AASE-Temptation (C; M)	2	N/A	N/A	N/A	N/A	2/2
Addiction Severity Index (M)	0	0	0	0	0	0/10
Beck Depression Inventory (M)	1	2	2	2	2	9/10
Brief Symptom Inventory (C)	1	2	2	2	2	9/10
Drinker Inventory of Consequences (C; M)	1	2	1	1	2	7/10
Obsessive-Compulsive Drinking Scale (C)	2	2	2	1	1	8/10
Psychosocial Functioning Inventory (M)	1	2	2	1	0	6/10
Short Form Health Survey-12 (C)	1	2	2	1	0	6/10
Spielberger State-Trait Inventory (M)	0	2	1	2	N/A	5/8
WHO Quality of Life Scale - Brief (C)	N/A	2	2	1	2	7/8

Note. **ROC.** Sensitivity/specificity scores of 0 indicated area under the curve (AUC) < 0.650 across all outcomes; 1 point indicated AUC > 0.650 and < 0.700 or mixed results across studies or across consumption outcomes; 2 points indicated AUC > 0.700 in both COMBINE and MATCH or for most outcomes. **Internal consistency reliability** scores of 0 indicated $\alpha < 0.70$; 1 point indicated $\alpha > 0.70$ and < 0.80 or mixed results across studies; 2 points indicated $\alpha > 0.80$ in both COMBINE and MATCH. **Convergent validity** results with scores of 0 indicated non-significant ($p > 0.05$) or at least one correlation in the opposite direction than was expected; 1 point indicated significant correlations with some but not all the expected measures or mixed results across studies; 2 points indicated significant correlations in the expected direction for all measures in both COMBINE and MATCH. **Confirmatory factor analysis** (CFA) results with scores of 0 indicated RMSEA > 0.08 or CFI or TLI < 0.90; 1 point indicated RMSEA < 0.08 and > 0.06 and/or CFI or TLI > 0.90 and < 0.95 or mixed results across studies; 2 points indicated RMSEA < 0.06 and CFI or TLI > 0.95 in both COMBINE and MATCH. **Measurement invariance** results with scores of 0 indicated non-invariance at the configural level or did not proceed to invariance testing due to poor model fit; 1 point indicated at least adequate model fit through the metric invariance testing (constraint

of the factor loadings for equivalence) or mixed results across both studies; 2 points indicated good model fit through strong invariance testing (highest possible level of invariance for categorical data) in both COMBINE and MATCH.

Table 7

Model results for CFA and measurement invariance testing

Measure (Dataset)	CFA Model	Invariance Testing Model	RMSEA (90% CI)	CFI	TLI
AASE (COMBINE)	4-factors based on DiClemente et al., 1994 structure		0.050 (0.048, 0.053)	0.919	0.914
		Configural: Baseline to Post-Treatment	0.030 (0.029, 0.031)	0.964	0.963
		Loadings Constrained: Baseline to Post-Treatment	0.029 (0.028, 0.030)	0.964	0.964
		Thresholds Constrained: Baseline to Post-Treatment	0.033 (0.032, 0.034)	0.953	0.954
AASE (MATCH)	4-factors based on DiClemente et al., 1994 structure*		0.060 (0.058, 0.062)	0.866	0.857
AASE (MATCH aftercare arm only)	4-factors based on DiClemente et al., 1994 structure		0.081 (0.078, 0.084)	0.879	0.872
AASE (MATCH outpatient arm only)	4-factors based on DiClemente et al., 1994 structure		0.050 (0.047, 0.053)	0.931	0.926
		Configural: Baseline to Post-Treatment	0.027 (0.026, 0.029)	0.917	0.914
		Loadings Constrained: Baseline to Post-Treatment	0.027 (0.026, 0.029)	0.915	0.913
		Thresholds Constrained: Baseline to Post-Treatment	0.028 (0.026, 0.029)	0.912	0.912
ASI (MATCH)	3-Factor Solution based on McLellan et al., 1992 structure*		N/A (failed convergence)	N/A	N/A
	Modified 3-Factor Solution based on McLellan et al.,		N/A (failed convergence)	N/A	N/A

	1992 structure with items ASIAF1, ASIAF2, ASIAF3, ASIAF6, ASIAF7, ASIAF8 specified as categorical per limited distributions and with item ASIAL13 removed due to high correlations with other variables and sparseness in item endorsement.				
BDI (MATCH)	2-factors (cognitive-affective and somatic factors)		0.030 (0.025, 0.035)	0.978	0.975
		Configural: Baseline to Post-Treatment	0.019 (0.017, 0.021)	0.971	0.970
		Loadings Constrained: Baseline to Post-Treatment	0.018 (0.016, 0.020)	0.973	0.972
		Thresholds Constrained: Baseline to Post-Treatment	0.019 (0.017, 0.021)	0.968	0.969
	3-factors (negative attitude, performance impairment, and somatic factors)		0.027 (0.021, 0.032)	0.982	0.980
		Configural: Baseline to Post-Treatment	0.019 (0.017, 0.020)	0.973	0.971
		Loadings Constrained: Baseline to Post-Treatment	0.018 (0.016, 0.020)	0.975	0.974
		Thresholds Constrained: Baseline to Post-Treatment	0.019 (0.017, 0.021)	0.970	0.971
BSI (COMBINE)	9-factors based on Derogatis & Melisaratos, 1983 structure		0.022 (0.019, 0.025)	0.975	0.974
		Configural: Baseline to	0.011 (0.010,	0.981	0.980

		Post-Treatment	0.012)		
		Loadings Constrained: Baseline to Post-Treatment	0.011 (0.009, 0.012)	0.982	0.981
		Thresholds Constrained: Baseline to Post-Treatment	0.012 (0.011, 0.013)	0.977	0.977
DrInC (COMBINE)	5-factors based on original conceptualization*		0.044 (0.041, 0.046)	0.900	0.894
	1-factor based on previously published models*		0.051 (0.049, 0.054)	0.861	0.854
	3-factor solution based on my conceptualization of the DrInC as comprised of consequences that occur commonly, moderately, and rarely		0.041 (0.038, 0.043)	0.920	0.916
		Configural: Baseline to Post-Treatment	0.017 (0.016, 0.019)	0.975	0.974
		Loadings Constrained: Baseline to Post-Treatment	0.019 (0.018, 0.020)	0.969	0.968
		Thresholds Constrained: Baseline to Post-Treatment	0.024 (0.023, 0.025)	0.951	0.952
DrInC (MATCH)	3-factor solution based on my conceptualization of the DrInC as comprised of consequences that occur commonly, moderately, and rarely		0.040 (0.038, 0.042)	0.908	0.904
		Configural: Baseline to Post-Treatment	0.018 (0.017, 0.019)	0.945	0.944
		Loadings Constrained: Baseline to Post-Treatment	0.018 (0.017, 0.018)	0.946	0.946
		Thresholds Constrained:	0.018 (0.017,	0.941	0.942

		Baseline to Post-Treatment	0.019)		
SIP (COMBINE)	5-factors based on original conceptualization		0.061 (0.053, 0.069)	0.969	0.960
		Configural: Baseline to Post-Treatment	0.086 (0.084, 0.086)	0.894	0.883
		Loadings Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
		Thresholds Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
	1-factor based on previously published models*		0.109 (0.102, 0.116)	0.890	0.871
SIP (MATCH)	5-factors based on original conceptualization		0.077 (0.070, 0.084)	0.949	0.933
		Configural: Baseline to Post-Treatment	0.059 (0.057, 0.061)	0.894	0.883
		Loadings Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
		Thresholds Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
OCDS (COMBINE)					
	2-factor model based on Ansseau et al., 2000; Anton 2000; Cordero et al., 2009*		0.109 (0.102, 0.117)	0.852	0.823
	2-factor model based on Ansseau et al., 2000; Anton 2000; Cordero et al., 2009 with consumption items		0.123 (0.114, 0.131)	0.869	0.837

	removed*				
	3-factor model based on Roberts et al., 1999*		0.096 (0.089, 0.104)	0.889	0.863
	3-factor model based on Kranzler et al., 1999*		0.104 (0.096, 0.111)	0.881	0.841
	4-factor model based on Bohn et al., 1996*		0.100 (0.093, 0.108)	0.885	0.852
	4-factor model based on Connor et al., 2008*		0.084 (0.076, 0.091)	0.920	0.897
	4-factor model based on previous published structures, EFA results, and conceptualization of items		0.072 (0.062, 0.082)	0.968	0.956
		Configural: Baseline to Post-Treatment	0.032 (0.029, 0.036)	0.987	0.984
		Loadings Constrained: Baseline to Post-Treatment	0.035 (0.032, 0.038)	0.984	0.981
		Thresholds Constrained: Baseline to Post-Treatment	0.076 (0.073, 0.079)	0.910	0.912
PFI (MATCH)	3-factor model based on original conceptualization of factors available in MATCH abbreviated version of the PFI		0.052 (0.047, 0.057)	0.933	0.923
		Configural: Baseline to Post-Treatment	0.042 (0.041, 0.044)	0.822	0.811
		Loadings Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
		Thresholds Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A

SF-12 (COMBINE)	2-factor model based on previously published models		0.080 (0.071, 0.090)	0.951	0.939
		Configural: Baseline to Post-Treatment	0.075 (0.072, 0.078)	0.854	0.839
		Loadings Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
		Thresholds Constrained: Baseline to Post-Treatment	N/A (Not tested due to failed configural invariance)	N/A	N/A
SSTI (MATCH) *only administered at baseline*	2-factor model based on previously published models*		0.116 (0.109, 0.122)	0.902	0.884
	4-factor model based on original conceptualization of 7 factors (items for 4 of the 7 factors were available in MATCH)		0.056 (0.048, 0.064)	0.976	0.969
WHOQOL-BREF (COMBINE)	4-factor, higher order model based on Skevington et al., 2004 structure		0.050 (0.045, 0.055)	0.942	0.935
		Model 1: Baseline to Week 26 (*N*=1381)	0.037 (0.035-0.038)	0.921	0.916
		Model 2: Baseline to Week 26	0.035 (0.034-0.037)	0.926	0.923
		Model 3: Baseline to Week 26	0.035 (0.033-0.037)	0.927	0.924
		Model 4: Baseline to Week 26	0.033 (0.032-0.035)	0.927	0.931
		Model 5: Baseline to Week 26	0.035 (0.033-0.036)	0.920	0.925

	Model 1: Week 26 to Week 52	0.042 (0.040-0.044)	0.917	0.912
	Model 2: Week 26 to Week 52	0.040 (0.038-0.041)	0.924	0.921
	Model 3: Week 26 to Week 52	0.039 (0.038-0.041)	0.925	0.922
	Model 4: Week 26 to Week 52	0.039 (0.036-0.039)	0.925	0.929
	Model 5: Week 26 to Week 52	0.037 (0.036-0.039)	0.926	0.930
4-factor model based on Jaracz et al., 2006 structure		0.053 (0.048-0.058)	0.944	0.936
4-factor model based on Trompenaars et al., 2005 structure		0.053 (0.048-0.058)	0.938	0.930
4-factor model based on Yao & Wu, 2002 structure		Non-positive definite matrix	N/A	N/A

Note. CFA results are for the replication split half sample unless specified as the first split half via *

Table 8

Baseline measure internal consistency reliability

Measure	Subscale (from CFA)	Cronbach's Alpha
AASE (COMBINE)		0.752*
	Negative Affect	0.356
	Social/Positive	0.341
	Physical & Other Concern	0.162
	Withdrawal & Urges	0.279
AASE (MATCH)		0.841**
	Negative Affect	0.557
	Social/Positive	0.460
	Physical & Other Concern	0.546
	Withdrawal & Urges	0.458
ASI		0.327
BDI		0.889**
	2-Factor Model: Cognitive-Affective	0.848**
	2-Factor Model: Somatic	0.771*
	3-Factor Model: Negative Attitudes	0.859**
	3-Factor Model: Performance Impairment	0.739*
	3-Factor Model: Somatic	0.478
BSI		0.965**
	Somatization	0.798*
	Obsessive-Compulsive	0.862**
	Depression	0.882**
	Interpersonal Sensitivity	0.643
	Hostility	0.790*
	Anxiety	0.824**
	Psychoticism	0.864**
	Phobic Anxiety	0.786*
	Paranoid Ideation	0.836**
DrInC (COMBINE)		0.937**
	Common Consequences	0.855**

	Moderately Common Consequences	0.905**
	Rare Consequences	0.808**
DrInC (MATCH)		0.938**
	Common Consequences	0.833**
	Moderately Common Consequences	0.905**
	Rare Consequences	0.830**
OCDS		0.852**
PFI		0.867**
	Subjective Role Performance	0.817**
	Overall Social Role Performance	0.818**
	Housemate/Roommate Role	0.531
SF-12		0.874**
	Physical Health	0.805**
	Psychological Health	0.861**
SSTI		0.887**
	"Factor 1" by Forgays et al., 1997	0.746*
	"Factor 2" by Forgays et al., 1997	0.865**
	"Factor 4" by Forgays et al., 1997	0.496
	"Factor 6" by Forgays et al., 1997	0.781*
WHOQOL-BREF		0.901**
	Physical Health	0.798*
	Psychological Health	0.770*
	Social Relationships	0.718*
	Environment	0.812**

Note. ***** indicates good internal consistency of $\alpha \geq 0.70$ and < 0.80; ****** indicates excellent internal consistency of $\alpha \geq 0.80$.

Table 9

Receiver operating characteristic curve area under the curve (AUC) results for the Brief Symptom Inventory (BSI) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	BSI total score	SOM Factor	OC Factor	DEP Factor	IS Factor	HOS Factor	ANX Factor	PSY Factor	PHOB Factor	PARA Factor
Post-tx Abstinence	0.628	0.613	0.617	0.626	0.622	0.602	0.606	0.600	0.566	0.575
Post-tx Heavy Drinking	**0.676**	0.629	0.644	**0.672**	**0.667**	0.636	**0.652**	0.648	0.610	0.619
Post-tx WHO risk: low risk or lower risk computed via DDD	**0.658**	0.622	0.634	**0.661**	**0.650**	0.619	0.645	0.634	0.594	0.603
Post-tx WHO risk: low risk or lower risk computed via MXD	**0.650**	0.617	0.627	**0.652**	0.642	0.624	0.631	0.620	0.589	0.598
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.690**	**0.674**	**0.691**	**0.695**	**0.672**	0.627	**0.693**	**0.650**	0.608	0.609
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.675**	0.641	**0.650**	**0.670**	**0.662**	0.627	**0.656**	0.647	0.613	0.619
Post-tx WHO risk: moderate risk or lower risk computed via MXD	**0.670**	0.634	0.645	**0.667**	**0.662**	0.633	**0.650**	0.641	0.606	0.614
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.724**	**0.683**	**0.714**	**0.718**	**0.695**	**0.666**	**0.731**	**0.655**	0.624	0.623
Post-tx Composite clinical Outcome: Abstinent	0.638	0.624	0.619	0.639	0.636	0.612	0.622	0.610	0.575	0.583
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	**0.746**	**0.698**	**0.710**	**0.744**	**0.723**	**0.677**	**0.735**	**0.704**	0.639	0.646
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.833**	**0.767**	**0.796**	**0.833**	**0.809**	**0.754**	**0.799**	**0.774**	**0.713**	**0.721**

1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.603	0.569	0.587	0.606	0.595	0.567	0.605	0.585	0.569	0.573
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.591	0.529	0.568	0.601	0.564	0.581	0.589	0.565	0.535	0.539
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.511	0.581	0.588	0.616	0.592	0.581	0.599	0.593	0.568	0.574
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.585	0.550	0.574	0.584	0.570	0.567	0.580	0.566	0.550	0.554
12mo Abstinence	0.545	0.549	0.543	0.555	0.543	0.557	0.541	0.539	0.508	0.512
12mo Heavy Drinking	0.603	0.569	0.578	0.608	0.580	0.600	0.585	0.601	0.557	0.562
12mo WHO risk: low risk or lower risk computed via DDD	0.599	0.581	0.580	0.610	0.582	0.593	0.592	0.596	0.556	0.560
12mo WHO risk: low risk or lower risk computed via MXD	0.589	0.573	0.576	0.596	0.579	0.588	0.577	0.583	0.546	0.550
12mo WHO risk: low risk or lower risk computed via DPD	0.602	0.568	0.604	0.619	0.578	0.589	0.600	0.597	0.551	0.551
12mo WHO risk: moderate risk or lower risk computed via DDD	0.618	0.580	0.587	0.627	0.597	0.599	0.605	0.523	0.574	0.578
12mo WHO risk: moderate risk or lower risk computed via MXD	0.603	0.578	0.579	0.611	0.583	0.597	0.588	0.599	0.562	0.566
12mo WHO risk: moderate risk or lower risk computed via DPD	0.613	0.572	0.594	0.627	0.590	0.586	0.607	0.610	0.562	0.568
12mo Composite clinical Outcome: Abstinent	0.550	0.556	0.545	0.562	0.547	0.557	0.543	0.552	0.522	0.523
12mo Composite clinical	0.625	0.586	0.604	0.639	0.612	0.619	0.606	0.620	0.571	0.574

Outcome: Moderate drinking or abstinent										
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.708**	0.638	**0.686**	**0.706**	0.648	**0.691**	**0.677**	**0.686**	0.647	**0.650**

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). BSI sub-scale factors are: Somatic symptoms (SOM), Obsessive-Compulsive symptoms (OC), Depressive symptoms (DEP), Interpersonal Sensitivity (IS), Hostility (HOS), Anxiety symptoms (ANX), Psychoticism symptoms (PSY), Phobic Anxiety symptoms (PHOB), and Paranoia symptoms (PARA). AUC ≥ 0.650 have been bolded and underlined for improved readability.

Table 10

Receiver operating characteristic curve area under the curve (AUC) results for the Beck Depression Inventory (BDI) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	BDI total score	2-Factor Solution: Cognitive Factor	2-Factor Solution: Somatic Factor	3-Factor Solution: Negative Affect Factor	3-Factor Solution: Performance Impairment Factor	3-Factor Solution: Somatic Factor
Post-tx Abstinence	0.582	0.606	0.598	0.626	0.585	0.558
Post-tx Heavy Drinking	0.597	0.636	0.616	**0.654**	0.604	0.571
Post-tx WHO risk: low risk or lower risk computed via DDD	0.598	0.633	0.619	**0.653**	0.605	0.575
Post-tx WHO risk: low risk or lower risk computed via MXD	0.589	0.625	0.612	0.645	0.602	0.569
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.653**	**0.689**	**0.706**	**0.701**	**0.680**	**0.658**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	0.600	0.641	0.624	**0.659**	0.611	0.577
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0.595	0.634	0.615	**0.651**	0.603	0.569
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.658**	**0.705**	**0.710**	**0.714**	**0.685**	**0.668**
Post-tx Composite clinical Outcome: Abstinent	0.588	0.604	0.594	0.625	0.580	0.554
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0.613	0.640	0.639	**0.665**	0.615	0.585
Post-tx Composite clinical Outcome: Heavy drinking OR	0.624	**0.656**	**0.655**	**0.678**	0.633	0.596

problems, moderate drinking, or abstinent						
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.630	0.608	0.593	0.626	0.579	0.563
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.586	0.584	0.570	0.592	0.573	0.537
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.604	0.587	0.570	0.606	0.562	0.541
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.555	0.538	0.530	0.551	0.531	0.508
12mo Abstinence	0.596	0.558	0.556	0.574	0.553	0.523
12mo Heavy Drinking	0.623	0.574	0.579	0.591	0.571	0.539
12mo WHO risk: low risk or lower risk computed via DDD	0.619	0.577	0.576	0.593	0.570	0.537
12mo WHO risk: low risk or lower risk computed via MXD	0.610	0.564	0.574	0.580	0.566	0.530
12mo WHO risk: low risk or lower risk computed via DPD	**0.674**	0.597	0.607	0.615	0.592	0.572
12mo WHO risk: moderate risk or lower risk computed via DDD	0.631	0.586	0.587	0.604	0.583	0.536
12mo WHO risk: moderate risk or lower risk computed via MXD	0.624	0.574	0.582	0.593	0.573	0.539
12mo WHO risk: moderate risk or lower risk computed via DPD	**0.669**	0.586	0.606	0.603	0.599	0.569
12mo Composite clinical	0.623	0.565	0.566	0.582	0.564	0.533

Outcome: Abstinent						
12mo Composite clinical Outcome: Moderate drinking or abstinent	**<u>0.667</u>**	0.595	0.604	0.617	0.593	0.558
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**<u>0.681</u>**	0.621	0.630	0.635	0.622	0.572

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 11

Receiver operating characteristic curve area under the curve (AUC) results for the World Health Organization Quality of Life Brief scale (WHOQOL-BREF) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	WHOQOL-BREF total score	Physical Factor	Psychological Factor	Social Factor	Environment Factor
12mo Abstinence	0.508	0.555	0.569	0.583	0.556
12mo Heavy Drinking	0.643	0.620	0.630	0.627	0.609
12mo WHO risk: low risk or lower risk computed via DDD	0.631	0.604	0.615	0.620	0.603
12mo WHO risk: low risk or lower risk computed via MXD	0.623	0.595	0.604	0.616	0.595
12mo WHO risk: low risk or lower risk computed via DPD	0.632	0.616	0.635	0.612	0.579
12mo WHO risk: moderate risk or lower risk computed via DDD	**0.664**	0.638	**0.652**	0.642	0.629
12mo WHO risk: moderate risk or lower risk computed via MXD	0.644	0.618	0.630	0.632	0.610
12mo WHO risk: moderate risk or lower risk computed via DPD	**0.651**	0.633	**0.659**	0.627	0.591
12mo Composite clinical Outcome: Abstinent	0.610	0.589	0.609	0.597	0.581
12mo Composite clinical Outcome: Moderate drinking or abstinent	**0.668**	0.644	**0.665**	0.638	0.624
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.715**	**0.698**	**0.704**	**0.691**	**0.653**

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 12

Receiver operating characteristic curve area under the curve (AUC) results for the Obsessive-Compulsive Drinking Scale (OCDS) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	OCDS total score	Factor 1	Factor 2	Factor 3	Factor 4
Post-tx Abstinence	**0.864**	**0.659**	**0.657**	**0.773**	**0.832**
Post-tx Heavy Drinking	**0.846**	**0.671**	**0.697**	**0.775**	**0.826**
Post-tx WHO risk: low risk or lower risk computed via DDD	**0.846**	**0.668**	**0.683**	**0.766**	**0.825**
Post-tx WHO risk: low risk or lower risk computed via MXD	**0.863**	**0.668**	**0.674**	**0.780**	**0.838**
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.868**	**0.669**	**0.708**	**0.751**	**0.851**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.819**	**0.667**	**0.710**	**0.763**	**0.806**
Post-tx WHO risk: moderate risk or lower risk computed via MXD	**0.844**	**0.667**	**0.689**	**0.768**	**0.823**
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.876**	**0.684**	**0.758**	**0.763**	**0.851**
Post-tx Composite clinical Outcome: Abstinent	**0.857**	**0.673**	**0.666**	**0.766**	**0.820**
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	**0.919**	**0.726**	**0.771**	**0.821**	**0.888**
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.934**	**0.782**	**0.871**	**0.877**	**0.912**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	**0.686**	0.610	0.629	0.646	**0.676**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	**0.709**	0.618	0.614	**0.658**	**0.701**
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	**0.732**	0.620	0.627	**0.675**	**0.719**
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	**0.696**	0.589	0.627	0.631	**0.686**

12mo Abstinence	**0.709**	0.604	0.566	0.643	**0.695**
12mo Heavy Drinking	**0.695**	0.606	0.603	0.644	**0.684**
12mo WHO risk: low risk or lower risk computed via DDD	**0.717**	0.626	0.599	**0.661**	**0.703**
12mo WHO risk: low risk or lower risk computed via MXD	**0.730**	0.627	0.599	**0.667**	**0.716**
12mo WHO risk: low risk or lower risk computed via DPD	**0.729**	0.616	0.602	**0.659**	**0.718**
12mo WHO risk: moderate risk or lower risk computed via DDD	**0.690**	0.616	0.620	0.647	**0.682**
12mo WHO risk: moderate risk or lower risk computed via MXD	**0.693**	0.603	0.599	0.641	**0.684**
12mo WHO risk: moderate risk or lower risk computed via DPD	**0.728**	0.630	0.623	**0.666**	**0.718**
12mo Composite clinical Outcome: Abstinent	**0.734**	0.614	0.587	**0.681**	**0.721**
12mo Composite clinical Outcome: Moderate drinking or abstinent	**0.756**	0.647	0.629	**0.694**	**0.742**
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.737**	**0.668**	**0.672**	**0.716**	**0.722**

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 13

Receiver operating characteristic curve area under the curve (AUC) results for the Drinker Inventory of Consequences (DrInC) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes in the COMBINE Study

	COMBINE DrInC total score	Factor 1 (Common Consequences)	Factor 2 (Moderately Common Consequences)	Factor 3 (Rare Consequences)
Post-tx Abstinence	**0.845**	**0.833**	**0.803**	**0.780**
Post-tx Heavy Drinking	**0.845**	**0.840**	**0.824**	**0.782**
Post-tx WHO risk: low risk or lower risk computed via DDD	**0.841**	**0.835**	**0.810**	**0.775**
Post-tx WHO risk: low risk or lower risk computed via MXD	**0.843**	**0.834**	**0.807**	**0.774**
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.825**	**0.825**	**0.716**	**0.771**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.831**	**0.821**	**0.814**	**0.781**
Post-tx WHO risk: moderate risk or lower risk computed via MXD	**0.841**	**0.835**	**0.819**	**0.780**
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.859**	**0.842**	**0.853**	**0.821**
Post-tx Composite clinical Outcome: Abstinent	**0.846**	**0.834**	**0.801**	**0.784**
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	**0.921**	**0.909**	**0.914**	**0.870**
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.944**	**0.939**	**0.939**	**0.917**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	**0.701**	**0.692**	**0.687**	**0.673**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	**0.691**	**0.687**	**0.677**	**0.655**
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	**0.731**	**0.722**	**0.713**	**0.688**

2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	**0.677**	**0.677**	**0.663**	0.642
12mo Abstinence	**0.684**	**0.671**	**0.659**	0.645
12mo Heavy Drinking	**0.702**	**0.685**	**0.683**	**0.674**
12mo WHO risk: low risk or lower risk computed via DDD	**0.705**	**0.696**	**0.681**	**0.663**
12mo WHO risk: low risk or lower risk computed via MXD	**0.712**	**0.702**	**0.685**	**0.669**
12mo WHO risk: low risk or lower risk computed via DPD	**0.710**	**0.711**	**0.688**	**0.662**
12mo WHO risk: moderate risk or lower risk computed via DDD	**0.699**	**0.688**	**0.683**	**0.668**
12mo WHO risk: moderate risk or lower risk computed via MXD	**0.700**	**0.686**	**0.681**	**0.670**
12mo WHO risk: moderate risk or lower risk computed via DPD	**0.712**	**0.709**	**0.692**	**0.677**
12mo Composite clinical Outcome: Abstinent	**0.705**	**0.694**	**0.689**	**0.663**
12mo Composite clinical Outcome: Moderate drinking or abstinent	**0.736**	**0.719**	**0.715**	**0.699**
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.755**	**0.748**	**0.747**	**0.727**

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 14

Receiver operating characteristic curve area under the curve (AUC) results for the Drinker Inventory of Consequences (DrInC) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes in Project MATCH

	MATCH DrInC total score	Factor 1 (Common Consequences)	Factor 2 (Moderately Common Consequences)	Factor 3 (Rare Consequences)
Post-tx Abstinence	0.583	0.585	0.573	0.586
Post-tx Heavy Drinking	**0.679**	**0.672**	**0.673**	**0.671**
Post-tx WHO risk: low risk or lower risk computed via DDD	**0.658**	**0.650**	**0.651**	**0.654**
Post-tx WHO risk: low risk or lower risk computed via MXD	0.642	0.637	0.635	0.639
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.675**	**0.747**	**0.756**	**0.718**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.688**	**0.678**	**0.683**	**0.680**
Post-tx WHO risk: moderate risk or lower risk computed via MXD	**0.677**	**0.669**	**0.671**	**0.671**
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.784**	**0.778**	**0.787**	**0.737**
Post-tx Composite clinical Outcome: Abstinent	**0.652**	0.649	0.647	0.631
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	**0.886**	**0.850**	**0.879**	**0.876**
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.836**	**0.812**	**0.835**	**0.809**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.579	0.590	0.589	0.597
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.549	0.541	0.550	0.568
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.550	0.558	0.554	0.569

2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.493	0.508	0.495	0.513
12mo Abstinence	0.424	0.432	0.425	0.436
12mo Heavy Drinking	0.511	0.515	0.511	0.509
12mo WHO risk: low risk or lower risk computed via DDD	0.499	0.505	0.496	0.504
12mo WHO risk: low risk or lower risk computed via MXD	0.470	0.473	0.468	0.478
12mo WHO risk: low risk or lower risk computed via DPD	0.557	0.569	0.561	0.535
12mo WHO risk: moderate risk or lower risk computed via DDD	0.547	0.553	0.543	0.546
12mo WHO risk: moderate risk or lower risk computed via MXD	0.507	0.517	0.506	0.508
12mo WHO risk: moderate risk or lower risk computed via DPD	0.575	0.581	0.582	0.547
12mo Composite clinical Outcome: Abstinent	0.453	0.465	0.455	0.462
12mo Composite clinical Outcome: Moderate drinking or abstinent	0.558	0.566	0.558	0.549
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.590	0.598	0.589	0.573

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 15

Receiver operating characteristic curve area under the curve (AUC) results for the baseline Spielberger State-Trait Inventory (SSTI) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	SSTI total score	SSTI total score (CFA items only)	"Factor 1" by Forgays et al., 1997	"Factor 2" by Forgays et al., 1997	"Factor 4" by Forgays et al., 1997	"Factor 6" by Forgays et al., 1997
Post-tx Abstinence	0.524	0.526	0.514	0.529	0.524	0.500
Post-tx Heavy Drinking	0.549	0.548	0.538	0.544	0.546	0.518
Post-tx WHO risk: low risk or lower risk computed via DDD	0.540	0.540	0.536	0.539	0.539	0.508
Post-tx WHO risk: low risk or lower risk computed via MXD	0.530	0.530	0.523	0.533	0.530	0.502
Post-tx WHO risk: low risk or lower risk computed via DPD	0.565	0.564	0.564	0.539	0.568	0.531
Post-tx WHO risk: moderate risk or lower risk computed via DDD	0.555	0.554	0.550	0.546	0.541	0.523
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0.548	0.547	0.535	0.543	0.543	0.520
Post-tx WHO risk: moderate risk or lower risk computed via DPD	0.577	0.575	0.560	0.543	0.568	0.554
Post-tx Composite clinical Outcome: Abstinent	0.534	0.534	0.511	0.539	0.530	0.516
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0.563	0.562	0.538	0.553	0.552	0.543
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.569	0.567	0.553	0.534	0.576	0.550
1 or more risk level change in	0.562	0.561	0.564	0.541	0.548	0.540

WHO risk level baseline to post-tx (computed via DDD)						
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.546	0.548	0.547	0.555	0.523	0.524
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.536	0.537	0.539	0.532	0.533	0.515
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.503	0.504	0.510	0.516	0.492	0.491
12mo Abstinence	0.499	0.501	0.501	0.501	0.507	0.490
12mo Heavy Drinking	0.518	0.521	0.530	0.513	0.528	0.495
12mo WHO risk: low risk or lower risk computed via DDD	0.511	0.513	0.517	0.510	0.518	0.493
12mo WHO risk: low risk or lower risk computed via MXD	0.509	0.511	0.512	0.510	0.514	0.496
12mo WHO risk: low risk or lower risk computed via DPD	0.530	0.532	0.545	0.524	0.534	0.501
12mo WHO risk: moderate risk or lower risk computed via DDD	0.532	0.533	0.540	0.527	0.529	0.503
12mo WHO risk: moderate risk or lower risk computed via MXD	0.520	0.522	0.528	0.517	0.528	0.495
12mo WHO risk: moderate risk or lower risk computed via DPD	0.547	0.548	0.545	0.537	0.548	0.514
12mo Composite clinical Outcome: Abstinent	0.506	0.509	0.512	0.509	0.505	0.497
12mo Composite clinical Outcome: Moderate drinking or abstinent	0.534	0.535	0.536	0.525	0.541	0.512
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.559	0.561	0.552	0.532	0.563	0.539

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 16

Receiver operating characteristic curve area under the curve (AUC) results for the Health Survey (SF-12) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	SF-12 total score	Physical Factor	Psychological Factor
Post-tx Abstinence	0.624	0.578	0.637
Post-tx Heavy Drinking	**0.671**	0.610	**0.686**
Post-tx WHO risk: low risk or lower risk computed via DDD	**0.654**	0.598	**0.667**
Post-tx WHO risk: low risk or lower risk computed via MXD	0.646	0.586	**0.663**
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.701**	**0.652**	**0.695**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.677**	0.632	**0.681**
Post-tx WHO risk: moderate risk or lower risk computed via MXD	**0.672**	0.615	**0.684**
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.724**	**0.672**	**0.724**
Post-tx Composite clinical Outcome: Abstinent	0.633	0.593	0.641
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	**0.754**	**0.707**	**0.745**
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.836**	**0.766**	**0.840**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.601	0.574	0.611
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.577	0.522	0.603
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.601	0.560	0.616
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.578	0.529	0.600
12mo Abstinence	0.548	0.518	0.561

12mo Heavy Drinking	0.595	0.562	0.604
12mo WHO risk: low risk or lower risk computed via DDD	0.599	0.562	0.608
12mo WHO risk: low risk or lower risk computed via MXD	0.592	0.557	0.601
12mo WHO risk: low risk or lower risk computed via DPD	0.630	0.586	0.637
12mo WHO risk: moderate risk or lower risk computed via DDD	0.609	0.580	0.614
12mo WHO risk: moderate risk or lower risk computed via MXD	0.603	0.570	0.608
12mo WHO risk: moderate risk or lower risk computed via DPD	0.632	0.593	0.635
12mo Composite clinical Outcome: Abstinent	0.561	0.540	0.565
12mo Composite clinical Outcome: Moderate drinking or abstinent	0.632	0.590	0.636
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.681**	0.627	**0.693**

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 17

Receiver operating characteristic curve area under the curve (AUC) results for the Psychosocial Functioning Inventory (PFI) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	PFI total score	Subjective Role Performance Factor	Overall Social Role Performance Factor	Housemate/ Roommate Factor
Post-tx Abstinence	0.608	0.545	0.575	0.575
Post-tx Heavy Drinking	0.646	0.599	0.617	0.604
Post-tx WHO risk: low risk or lower risk computed via DDD	0.642	0.588	0.608	0.613
Post-tx WHO risk: low risk or lower risk computed via MXD	0.631	0.571	0.593	0.601
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.671**	0.630	0.625	**0.660**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.651**	0.610	0.626	0.611
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0.644	0.596	0.614	0.602
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.700**	0.644	**0.663**	**0.658**
Post-tx Composite clinical Outcome: Abstinent	0.611	0.559	0.579	0.573
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0.648	0.611	0.599	0.614
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.665**	0.626	0.637	0.613
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.595	0.598	0.580	0.587
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.574	0.556	0.552	0.552
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.584	0.577	0.571	0.571

2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.538	0.522	0.520	0.537
12mo Abstinence	0.553	0.517	0.531	0.538
12mo Heavy Drinking	0.591	0.552	0.551	0.577
12mo WHO risk: low risk or lower risk computed via DDD	0.586	0.546	0.549	0.568
12mo WHO risk: low risk or lower risk computed via MXD	0.574	0.532	0.537	0.554
12mo WHO risk: low risk or lower risk computed via DPD	0.581	0.554	0.540	0.569
12mo WHO risk: moderate risk or lower risk computed via DDD	0.596	0.568	0.559	0.579
12mo WHO risk: moderate risk or lower risk computed via MXD	0.589	0.554	0.548	0.572
12mo WHO risk: moderate risk or lower risk computed via DPD	0.574	0.576	0.542	0.563
12mo Composite clinical Outcome: Abstinent	0.565	0.523	0.538	0.545
12mo Composite clinical Outcome: Moderate drinking or abstinent	0.600	0.573	0.583	0.571
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.597	0.582	0.585	0.564

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 18

Receiver operating characteristic curve area under the curve (AUC) results for the Alcohol Abstinence Self-Efficacy (AASE) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes in the COMBINE Study

	COMBINE AASE total score	AASE Confidence Subscale	AASE Temptation Subscale	Negative Affect Factor	Social Factor	Physical Factor	Urge Factor
Post-tx Abstinence	0.583	**0.793**	**0.790**	0.561	0.507	0.592	0.574
Post-tx Heavy Drinking	0.592	**0.804**	**0.790**	0.569	0.531	0.599	0.576
Post-tx WHO risk: low risk or lower risk computed via DDD	0.589	**0.796**	**0.788**	0.568	0.521	0.588	0.578
Post-tx WHO risk: low risk or lower risk computed via MXD	0.593	**0.808**	**0.804**	0.566	0.513	0.592	0.587
Post-tx WHO risk: low risk or lower risk computed via DPD	0.593	**0.814**	**0.799**	0.571	0.554	0.585	0.580
Post-tx WHO risk: moderate risk or lower risk computed via DDD	0.601	**0.790**	**0.766**	0.574	0.551	0.600	0.580
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0.590	**0.796**	**0.784**	0.566	0.521	0.597	0.577
Post-tx WHO risk: moderate risk or lower risk computed via DPD	0.625	**0.845**	**0.810**	0.615	0.580	0.590	0.601
Post-tx Composite clinical Outcome: Abstinent	0.579	**0.787**	**0.790**	0.570	0.498	0.587	0.568
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0.593	**0.867**	**0.858**	0.569	0.554	0.592	0.572
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.620	**0.879**	**0.852**	0.618	0.564	0.621	0.578

1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.554	**0.674**	**0.669**	0.545	0.530	0.548	0.547
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.547	**0.699**	**0.698**	0.535	0.535	0.537	0.547
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.546	**0.711**	**0.719**	0.527	0.504	0.553	0.537
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.544	**0.691**	**0.688**	0.527	0.509	0.550	0.542
12mo Abstinence	0.516	**0.664**	**0.690**	0.509	0.459	0.532	0.521
12mo Heavy Drinking	0.542	**0.687**	**0.701**	0.529	0.478	0.543	0.541
12mo WHO risk: low risk or lower risk computed via DDD	0.532	**0.685**	**0.709**	0.510	0.474	0.538	0.538
12mo WHO risk: low risk or lower risk computed via MXD	0.523	**0.692**	**0.717**	0.505	0.465	0.534	0.531
12mo WHO risk: low risk or lower risk computed via DPD	0.557	**0.698**	**0.707**	0.531	0.509	0.543	0.554
12mo WHO risk: moderate risk or lower risk computed via DDD	0.548	**0.683**	**0.692**	0.523	0.500	0.543	0.547
12mo WHO risk: moderate risk or lower risk computed via MXD	0.537	**0.683**	**0.700**	0.515	0.476	0.541	0.540
12mo WHO risk: moderate risk or lower risk computed via DPD	0.549	**0.715**	**0.724**	0.526	0.496	0.540	0.547
12mo Composite clinical Outcome: Abstinent	0.550	**0.698**	**0.708**	0.535	0.480	0.562	0.555
12mo Composite clinical	0.577	**0.732**	**0.734**	0.558	0.507	0.568	0.576

Outcome: Moderate drinking or abstinent							
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.597	**<u>0.729</u>**	**<u>0.723</u>**	0.572	0.541	0.588	0.580

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 19

Receiver operating characteristic curve area under the curve (AUC) results for the Alcohol Abstinence Self-Efficacy (AASE) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes in Project MATCH

	MATCH AASE total score	AASE Confidence Subscale	AASE Temptation Subscale	Negative Affect Factor	Social Factor	Physical Factor	Urge Factor
Post-tx Abstinence	0.573	**0.720**	**0.703**	0.527	0.533	0.576	0.578
Post-tx Heavy Drinking	0.554	**0.726**	**0.735**	0.501	0.519	0.554	0.554
Post-tx WHO risk: low risk or lower risk computed via DDD	0.563	**0.726**	**0.722**	0.507	0.529	0.564	0.564
Post-tx WHO risk: low risk or lower risk computed via MXD	0.562	**0.725**	**0.721**	0.512	0.523	0.566	0.567
Post-tx WHO risk: low risk or lower risk computed via DPD	0.532	**0.730**	**0.758**	0.471	0.528	0.538	0.544
Post-tx WHO risk: moderate risk or lower risk computed via DDD	0.539	**0.716**	**0.735**	0.485	0.512	0.543	0.541
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0.553	**0.726**	**0.735**	0.502	0.517	0.553	0.554
Post-tx WHO risk: moderate risk or lower risk computed via DPD	0.523	**0.733**	**0.767**	0.475	0.519	0.531	0.533
Post-tx Composite clinical Outcome: Abstinent	0.550	**0.701**	**0.702**	0.511	0.524	0.560	0.550
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0.542	**0.720**	**0.735**	0.492	0.516	0.563	0.542
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.540	**0.717**	**0.735**	0.483	0.514	0.550	0.542

1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.534	**0.668**	0.626	0.502	0.517	0.537	0.534
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.545	0.628	0.565	0.515	0.529	0.555	0.534
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.543	**0.685**	0.537	0.503	0.515	0.552	0.544
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.536	0.611	0.572	0.520	0.512	0.547	0.529
12mo Abstinence	0.555	**0.666**	**0.655**	0.524	0.512	0.568	0.556
12mo Heavy Drinking	0.536	**0.667**	**0.674**	0.499	0.511	0.545	0.534
12mo WHO risk: low risk or lower risk computed via DDD	0.543	**0.681**	**0.681**	0.505	0.509	0.552	0.542
12mo WHO risk: low risk or lower risk computed via MXD	0.544	**0.672**	**0.670**	0.510	0.508	0.556	0.546
12mo WHO risk: low risk or lower risk computed via DPD	0.540	**0.689**	**0.700**	0.497	0.517	0.551	0.544
12mo WHO risk: moderate risk or lower risk computed via DDD	0.534	**0.682**	**0.694**	0.494	0.516	0.544	0.528
12mo WHO risk: moderate risk or lower risk computed via MXD	0.538	**0.675**	**0.681**	0.502	0.511	0.547	0.535
12mo WHO risk: moderate risk or lower risk computed via DPD	0.522	**0.672**	**0.700**	0.504	0.514	0.524	0.513
12mo Composite clinical Outcome: Abstinent	0.556	**0.689**	**0.679**	0.520	0.519	0.563	0.550
12mo Composite clinical	0.537	**0.690**	**0.712**	0.493	0.504	0.553	0.531

Outcome: Moderate drinking or abstinent							
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.503	0.649	**<u>0.696</u>**	0.465	0.501	0.516	0.500

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 20

Receiver operating characteristic curve area under the curve (AUC) results for the Addiction Severity Index (ASI) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes

	ASI total score
Post-tx Abstinence	0.551
Post-tx Heavy Drinking	0.568
Post-tx WHO risk: low risk or lower risk computed via DDD	0.569
Post-tx WHO risk: low risk or lower risk computed via MXD	0.559
Post-tx WHO risk: low risk or lower risk computed via DPD	0.599
Post-tx WHO risk: moderate risk or lower risk computed via DDD	0.583
Post-tx WHO risk: moderate risk or lower risk computed via MXD	0.568
Post-tx WHO risk: moderate risk or lower risk computed via DPD	0.612
Post-tx Composite clinical Outcome: Abstinent	0.538
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	0.562
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.573
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.528
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.514
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	0.510
2 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	0.494

12mo Abstinence	0.504
12mo Heavy Drinking	0.515
12mo WHO risk: low risk or lower risk computed via DDD	0.514
12mo WHO risk: low risk or lower risk computed via MXD	0.508
12mo WHO risk: low risk or lower risk computed via DPD	0.523
12mo WHO risk: moderate risk or lower risk computed via DDD	0.525
12mo WHO risk: moderate risk or lower risk computed via MXD	0.515
12mo WHO risk: moderate risk or lower risk computed via DPD	0.492
12mo Composite clinical Outcome: Abstinent	0.511
12mo Composite clinical Outcome: Moderate drinking or abstinent	0.535
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	0.544

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Table 21

Receiver operating characteristic curve area under the curve (AUC) results for percent days abstinent (PDA) and percent heavy drinking days (PHDD) for detecting/discriminating post-treatment (post-tx) and 12-month follow-up (12mo) consumption outcomes in the COMBINE Study and Project MATCH

	COMBINE PDA	COMBINE PHDD	MATCH PDA	MATCH PHDD
Post-tx Abstinence	-	**0.880**	-	**0.897**
Post-tx Heavy Drinking	**0.900**	-	**0.955**	-
Post-tx WHO risk: low risk or lower risk computed via DDD	**0.922**	**0.937**	**0.962**	**0.945**
Post-tx WHO risk: low risk or lower risk computed via MXD	**0.962**	**0.919**	**0.979**	**0.944**
Post-tx WHO risk: low risk or lower risk computed via DPD	**0.960**	**0.955**	**0.971**	**0.983**
Post-tx WHO risk: moderate risk or lower risk computed via DDD	**0.849**	**0.970**	**0.929**	**0.987**
Post-tx WHO risk: moderate risk or lower risk computed via MXD	**0.909**	**0.986**	**0.958**	**0.994**
Post-tx WHO risk: moderate risk or lower risk computed via DPD	**0.954**	**0.988**	**0.970**	**0.990**
Post-tx Composite clinical Outcome: Abstinent	**0.975**	**0.846**	**0.882**	**0.803**
Post-tx Composite clinical Outcome: Moderate drinking or abstinent	**0.946**	**0.960**	**0.872**	**0.844**
Post-tx Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.876**	**0.902**	**0.834**	**0.849**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	**0.670**	**0.726**	**0.698**	**0.672**
1 or more risk level change in WHO risk level baseline to post-tx (computed via DPD)	**0.724**	**0.731**	**0.682**	0.611
2 or more risk level change in WHO risk level baseline to post-tx (computed via DDD)	**0.757**	**0.753**	**0.745**	**0.679**
2 or more risk level change in WHO risk level baseline	**0.745**	**0.732**	**0.666**	0.610

to post-tx (computed via DPD)				
12mo Abstinence	**0.806**	**0.682**	**0.724**	**0.668**
12mo Heavy Drinking	**0.736**	**0.740**	**0.721**	**0.703**
12mo WHO risk: low risk or lower risk computed via DDD	**0.753**	**0.719**	**0.726**	**0.695**
12mo WHO risk: low risk or lower risk computed via MXD	**0.773**	**0.709**	**0.728**	**0.686**
12mo WHO risk: low risk or lower risk computed via DPD	**0.798**	**0.768**	**0.748**	**0.724**
12mo WHO risk: moderate risk or lower risk computed via DDD	**0.719**	**0.742**	**0.711**	**0.709**
12mo WHO risk: moderate risk or lower risk computed via MXD	**0.737**	**0.734**	**0.723**	**0.700**
12mo WHO risk: moderate risk or lower risk computed via DPD	**0.773**	**0.778**	**0.733**	**0.727**
12mo Composite clinical Outcome: Abstinent	**0.773**	**0.686**	**0.743**	**0.688**
12mo Composite clinical Outcome: Moderate drinking or abstinent	**0.787**	**0.756**	**0.754**	**0.724**
12mo Composite clinical Outcome: Heavy drinking OR problems, moderate drinking, or abstinent	**0.716**	**0.723**	**0.694**	**0.683**

Note: DDD = drinks per drinking day; MXD = maximum number of drinks consumed in the 90-day window; DPD = drinks per day in the assessment window (averaged across drinking and abstinent days). AUC $\geq$ 0.650 have been bolded and underlined for improved readability.

Appendix A

Individual Item Receiver Operating Characteristic Curve Results

Brief Symptom Inventory. Since ROC curve analyses for the BSI total and subscale score yielded AUC's $\geq$ 0.650, ROC curve analyses were conducted for individual items. Items 1 (nervousness or shakiness; Anxiety subscale item), 15 (feeling blocked in getting things done; Obsessive-Compulsive subscale item), and 17 (feeling blue; Depression subscale item) adequately detected 4 of 15 post-treatment outcomes: WHO low or lower risk level (calculated via DPD; AUC = 0.687, 0.687, 0.674), WHO moderate or lower risk level (calculated via DPD; AUC = 0.720, 0.684, 0.704), composite clinical outcome of moderate or lower risk (AUC = 0.717, 0.686, 0.706), and composite clinical outcome of heavy or lower risk (AUC = 0.772, 0.767, 0.786). Items 6 (annoyed or irritated; Hostility subscale item), 18 (lack of interest; Depression subscale item), 19 (feeling fearful; Anxiety subscale item), 36 (trouble concentrating; Obsessive-Compulsive subscale item), and 38 (tense/keyed up; Anxiety subscale item) all adequately detected 3 of 15 post-treatment outcomes: WHO moderate or lower risk level (calculated via DPD; AUC = 0.683, 0.660, 0.656, 0.672, 0.669), composite clinical outcome of moderate or lower risk (AUC = 0.677, 0.678, 0.650, 0.670, 0.670), and composite clinical outcome of heavy or lower risk (AUC = 0.768, 0.760, 0.697, 0.751, 0.738). In addition to the above items, post-treatment composite clinical outcome of moderate or lower risk and composite clinical outcome of heavy or lower risk were also both adequately detected by item 5 (trouble remembering; Obsessive-Compulsive subscale item; AUC's = 0.667, 0.678), item14 (feeling lonely; Psychoticism subscale item; AUC's = 0.653, 0.710), item 16 (feeling lonely; Depression subscale item; AUC's = 0.686, 0.763), item 35 (hopeless; Depression subscale item; AUC's = 0.674, 0.758), item 50 (worthlessness; Depression subscale item; AUC's = 0.659, 0.724), and

items 52 (guilt; Interpersonal Sensitivity subscale item; AUC's = 0.683, 0.737) and 53 (ideas something is wrong with you; Psychoticism subscale item; AUC's = 0.677, 0.747). Post treatment composite clinical outcome of heavy or lower risk was also adequately detected by 15 other items, which made it the most readily detected post-treatment consumption outcome. Moreover, 12-month follow-up composite clinical outcome of heavy or lower risk was the only 12-month follow-up consumption outcome that was adequately detected by any BSI individual items and it was adequately detected by 7 items: 6 (annoyed or irritated; Hostility subscale item), 15 (feeling blocked in getting things done; Obsessive-Compulsive subscale item), 16 (feeling lonely; Depression subscale item), 17 (feeling blue; Depression subscale item), 18 (lack of interest; Depression subscale item), 52 (guilt; Interpersonal Sensitivity subscale item), and 53 (ideas something is wrong with you; Psychoticism subscale item).

World Health Organization Quality of Life, Brief Version. Since ROC curve analyses for the week 26 WHOQOL-BREF subscales yielded adequate detection of 12-month consumption outcomes (AUC's $\geq$ 0.650), analyses were conducted for individual items of the week 26 WHOQOL-BREF for detecting 12-month consumption outcomes. The 12-month composite clinical outcome of heavy or lower risk was adequately detected by 8 of the individual items: item 5 (enjoy life; Psychological Health subscale item; AUC = 0.675), item 6 (life is meaningful; Psychological Health subscale item; AUC = 0.665), item 7 (able to concentrate; Psychological Health subscale item; AUC = 0.655), item 16 (sleep satisfaction; Physical Health subscale item; AUC = 0.661), item 17 (daily activities; Physical Health subscale item; AUC = 0.696), item 18 (capacity for work; Physical Health subscale item; AUC = 0.650), item 20 (personal relationships; Social subscale item; AUC = 0.673), and item 22 (friend support; Social subscale item; AUC = 0.654).

Beck Depression Inventory. Since the BDI subscales adequately detected some post-treatment consumption outcomes, ROC curve analyses were conducted for individual items' ability to detect post-treatment outcomes. The outcomes that were adequately detected were WHO low or lower risk level (calculated via DPD; WLLP) and WHO moderate or lower risk level (calculated via DPD; WMLP). Items 4 (satisfaction in activities) and 7 (self-dislike) were able to adequately detect both of these outcomes: WLLP (AUC's = 0.670, 0.671) and WMLP (AUC's = 0.674, 0.683). Item 15 (work ability) was only able to adequately detect post-treatment WLLP (AUC = 0.650). Items 3 (personal failure), 5 (guilt), and 16 (sleep disturbance) were all able to adequately detect post-treatment WMLP (AUC's = 0.664, 0.661, 0.660).

Obsessive-Compulsive Drinking Scale. Given how well each sub-factor of the OCDS did in ROC curve analyses, it is unsurprising individual items also performed well. Items 13 (drive to consume), and 14 (control over drinking) all adequately detected 15 of 15 post-treatment outcomes: AUC's = 0.650-0.880, and 0.664-0.894. Item 12 (effort to resist drinking) adequately detected 13 of 15 post-treatment outcomes: AUC's = 0.624-0.845. Items 5 (effort to resist thoughts), 6 (success in stopping thoughts), and 11 (anxiety over being prevented from drinking) each adequately detected 12 of 15 post-treatment outcomes: AUC's = 0.610-0.848, 0.635-0.848, 0.640-0.867). Items 1 (time thinking) and 4 (distress of thoughts) each adequately detected 11 of 15 whereas item 10 (social functioning interference) adequately detected 10 of 15 post-treatment outcomes: AUC's = 0.590-0.787, 0.574-0.806 and 0.598-0.853. Item 9 (work functioning interference) adequately detected 8 of 15 post-treatment outcomes (AUC's = 0.584-0.811) and items 2 (thought frequency) and 3 (thought interference with social or work functioning) each adequately detected 3 of 15 post-treatment outcomes: AUC's = 0.576-0.718, 0.556-0.789. For 12-month follow-up outcomes, items 7 and 8 adequately detected 11 of 11

outcomes (AUC's = 0.672-0.737, 0.661-0.740); items 13 and 14 adequately detected 9 of 11 outcomes (AUC's = 0.644-0.696, 0.647-0.704). Items 6, 11, and 12 each adequately detected 8 of 11 outcomes: AUC's = 0.632-0.690, 0.642-0.698, and 0.639-0.690. Item 5 adequately detected 2 of 11 12-month outcomes (AUC's = 0.608-0.678) and items 1 and 4 adequately detected 1 of 11 12-month outcomes: AUC's = 0.602-0.668, 0.573-0.667. Post-treatment outcomes of composite clinical outcome of heavy or lower risk and 2+ change in WHO risk level since baseline (calculated via DPD) were the most and least detectable consumption outcomes for 9 of the 14 of individual items. For items that were able to adequately detect at least one 12-month outcome (AUC $\geq$ 0.650), composite clinical outcome of moderate or lower risk and composite clinical outcome of heavy or lower risk yielded the highest AUC's; abstinence and WHO moderate or lower risk (calculated via MXD) yielded the lowest AUC's.

Drinker Inventory of Consequences. The 3 factors that were upheld via CFA and measurement invariance testing for the DrInC in both COMBINE and MATCH yielded high AUC values in ROC curve analyses. Therefore, individual item ROC curve analyses were conducted for COMBINE and MATCH DrInC data.

In COMBINE, several individual items were able to adequately detect post-treatment outcomes. Items 1 and 2 (hangover, felt bad about self) adequately detected all post-treatment outcomes (AUC's = 0.664-0.880, 0.651-0.870); item 12 (unhappy due to drinking) adequately detected 14 of 15 post-treatment outcomes (AUC's = 0.640-0.898). Item 16 (guilt/ashamed; AUC's = 0.630-0.881) adequately detected 13 of 15 post-treatment outcomes. Several individual items were able to adequately detect 11 of 15 post-treatment outcomes: item 4 (family/friends worried/complained; AUC's = 0.602-0.805), item 8 (sleep disturbances; AUC's = 0.578-0.833), item 13 (eating disturbances; AUC's = 0.597-0.888), item 18 (personality worsened; AUC's =

0.587-0.794), item 37 (undesired life; AUC's = 0.607-0.883), and item 38 (personal growth interference; AUC's = 0.619-0.895). Item 28 (smoked more tobacco) adequately detected 10 of 15 post-treatment outcomes (AUC's = 0.589-0.742); item 17 (said/done embarrassing things) adequately detected 9 of 15 post-treatment outcomes (AUC's = 0.591-0.804). Item 40 (spent too much money) adequately detected 8 out of 15 post-treatment outcomes (AUC's = 0.576-0.821). Many items adequately detected 7 out of 15 post-treatment outcomes: item 14 (failed expectations; AUC's = 0.555-0.831), item 24 (physical health harmed; AUC's = .593-0.838), item 30 (hurt family; AUC's = 0.560-0.784), item 34 (lost interest; AUC's = 0.576-0.823), and item 36 (spiritual/moral life harmed; AUC's = 0.583-0.794). Items 22 (impulsivity) and 39 (damaged social life) adequately detected 5 of 15 post-treatment outcomes (AUC's = 0.569-0.818, 0.555-0.794). Four items adequately detected 4 of 15 post-treatment outcomes: item 6 (work quality suffered; AUC's = 0.582-0.801), item 9 (driven after 3+ drinks; AUC's = 0.594-0.744), item 29 (physical appearance harmed; AUC's = 0.567-0.842), and item 33 (sex life suffered; AUC's = 0.569-0.776). Six items each adequately detected 3 of 15 post-treatment outcomes: item 19 (foolish risks; AUC's = 0.555-0.777), item 21 (said cruel things; AUC's = 0.562-0.710), item 26 (money problems; AUC's = 0.564-0.774), item 27 (marriage/love relationship harmed; AUC's = 0.559-0.717), item 31 (friendship damaged; AUC's = 0.557-0.751), and item 32 (overweight; AUC's = 0.583-0.708). Item 20 (trouble; AUC's = 0.527-0.677) adequately detected 2 of 15 post-treatment outcomes; items 3 (missed school/work), item 7 (parenting ability), and 11 (vomited) each adequately detected 1 of 15 post-treatment outcomes (AUC's = 0.535-0.716, 0.535-0.684, 0.522-0.676). Items 10 (other drug use), 23 (physical fight), and 41 through 50 (arrested for DWI, trouble with the law, lost marriage/love relationship, suspended/fired, lost a friend, had an accident, been physically hurt, injured someone else, and

broken things) all failed to detect any post-treatment outcomes (0 out of 15 outcomes; AUC's < 0.650).

In COMBINE, some individual items of the DrInC also adequately detected 12-month follow-up outcomes. Item 1 (hangover) adequately detected all 11 out of 11 12-month outcomes (AUC's = 0.663-0.698); item 2 (felt bad about self) adequately detected 10 of 11 12-month outcomes (AUC's = 0.639-0.713). Item 16 (guilt/ashamed) adequately detected 9 of 11 12-month outcomes (AUC's = 0.638-0.714); item 12 (unhappy due to drinking) adequately detected 6 of 11 12-month outcomes (AUC's = 0.629-0.707). These findings are consistent with the ability of these items to detect the majority of post-treatment outcomes. Additionally, item 38 (personal growth interference) adequately detected 2 of 11 12-month outcomes (AUC's = 0.559-0.706) and several items adequately detected 1 of 11 12-month outcomes: item 8 (sleep disturbances), 13 (eating disturbances), 18 (personality worsened), 24 (physical health harmed), 29 (physical appearance harmed), 30 (hurt family), 34 (lost interest), 36 (spiritual/moral life harmed), 37 (undesired life), and 40 (spent too much money).

In MATCH, several individual items were able to detect any post-treatment outcomes. Item 1 (hangovers) adequately detected 9 out of the 11 post-treatment outcome tested (changes in WHO risk since baseline were not analyzed due to the null findings for total DrInC and individual factors for detecting those outcomes adequately; AUC's = 0.596-0.733). Item 13 (eating disturbances) adequately detected 7 of 11 post-treatment outcomes (AUC's = 0.584-0.768). Items 12 (unhappy due to drinking) and 17(said/done embarrassing things) adequately detected 6 of 11 post-treatment outcomes (AUC's = 0.598-0.774, 0.575-0.771). Several items were able to adequately detect 4 of the 11 post-treatment outcomes examined: item 2 (felt bad about self ; AUC's = 0.572-0.746), item 4 (family/friends worried/complained; AUC's = 0.566-

0.757), item 6 (work quality suffered; AUC's = 0.548-0.685), item 8 (sleep disturbances; AUC's = 0.521-0.719), item 14 (failed expectations; AUC's = 0.550-0.757), item 16 (guilt/ashamed; AUC's = 0.568-0.740), item 18 (personality worsened; AUC's = 0.540-0.784), item 21 (said cruel things; AUC's = 0.571-0.745), item 24 (physical health harmed; AUC's = .563-0.755), item 26 (money problems; AUC's = 0.551-0.733), item 29 (physical appearance harmed; AUC's = 0.546-0.771), item 30 (hurt family; AUC's = 0.551-0.772), item 34 (lost interest; AUC's = 0.535-0.753), item 36 (spiritual/moral life harmed; AUC's = 0.518-0.719), item 37 (undesired life; AUC's = 0.561-0.815), and item 38 (personal growth interference; AUC's = 0.563-0.817), and item 40 (spent too much money; AUC's = 0.581-0.778). Five individual items each adequately detected 3 of 11 post-treatment outcomes that were analyzed: item 9 (driven after 3+ drinks; AUC's = 0.582-0.701), item 19 (foolish risks; AUC's = 0.548-0.738), item 27 (marriage/love relationship harmed; AUC's = 0.542-0.740), item 31 (friendship damaged; AUC's = 0.539-0.723), and item 39 (damaged social life; AUC's = 0.509-0.752). Five items also detected only 2 of the 11 analyzed post-treatment items: item 20 (trouble; AUC's = 0.540-0.675), item 22 (impulsivity; AUC's = 0.540-0.727), item 28 (smoked more tobacco; AUC's = 0.540-0.721), and items 32 and 33 (overweight, AUC's = 0.566-0.663; sex life suffered, AUC's = 0.513-0.687). Item 7 (parenting ability) was able to adequately detect 1 of the 11 examined post-treatment outcomes (AUC's = 0.537-0.657). All other items failed to adequately detect (AUC's < 0.650) any post-treatment consumption outcomes. In MATCH, several individual items on the DrInC were most able to detect post-treatment composite clinical outcome of moderate or lower risk and were poorest at detecting post-treatment abstinence.

Health Survey (SF-12). Since some of the post-treatment outcomes were adequately detected by each of the factors, select ROC curve analyses were conducted for individual items

based on which consumption outcomes were adequately detected by each item's respective factor. Item 6A (felt calm or peaceful) had the highest AUC values and adequately detected 8 out of 9 post-treatment outcomes that were examined (AUC's = 0.647-0.769). Item 6A was also the only examined item that was able to adequately detect any 12-month follow-up outcome: AUC = 0.650 for 12-month follow-up composite clinical outcome of heavy or lower risk. Items 4B (emotional interference with work/activities) and 6B (lots of energy) were able to adequately detect 4 post-treatment outcomes with positive AUC's = 0.651-0.763 and 0.650-0.775. Items 1 (general health), 4A (emotional interference with accomplishments)m and 6C (downhearted/depressed) adequately detected 3 post-treatment outcomes: positive AUC's = 0.656-0.773, 0.669-0.775, 0.654-0.715. Items 3A (physical health interference with accomplishments) and 7 (physical/emotional interference with social activities) adequately detected 2 post-treatment outcomes: positive AUC's = 0.670-0.684, 0.650-0.713. Items 2A (health limit moderate activities), 2B (health limit climbing stairs), 3B (physical health limits work/activities times), and 5 (pain interference with work) were unable to detect any of the post-treatment outcomes that were examined in the present study.

Psychosocial Functioning Inventory. Only the Social Role Performance and Housemate/Roommate Role factors yielded some adequate AUC values in ROC curve analyses. Accordingly, individual item ROC curve analyses were conducted only for items from these two factors for consumption outcomes that were adequately detected from the larger factor subscale scores (i.e., only certain post-treatment consumption outcomes, and none of the 12-month follow-up consumption outcomes since AUC's were all < 0.650 for every factor subscale score). From individual item ROC curve analyses conducted based on positive Social Role Performance and Housemate/Roommate Role factors ROC curve analyses, only 2 items adequately detected

any consumption outcomes. Item 11 (spousal/mate overall role performance; Social Role Performance factor) adequately detected WHO moderate or lower risk level (calculated via DPD; AUC = 0.677) as did item 19 (housemate/roommate overall role performance, AUC = 0.658). These two items were the only examined items that detected any post-treatment or 12-month follow-up outcomes and each only adequately detected one consumption outcome (WHO moderate or lower risk level (calculated via DPD).

References

Alterman, A. I., Cacciola, J. S., Ivey, M. A., Habing, B., & Lynch, K. G. (2009). Reliability and validity of the alcohol short index of problems and a newly constructed drug Short Index of Problems. *Journal of Studies on Alcohol and Drugs, 70*(2), 304. PMCID: PMC2653616

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., revised). Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: Author.

Ansseau, M., Besson, J., Lejoyeux, M., Pinto, E., Landry, U., Cornes, M., … Ades, J. (2000). A French translation of the obsessive-compulsive drinking scale for craving in alcohol-dependent patients: a validation study in Belgium, France, and Switzerland. *European Addiction Research, 6*(2), 51–6. doi: 10.1159/000019010

Anton, R. F. (2000). Obsessive-compulsive aspects of craving: development of the Obsessive Compulsive Drinking Scale. *Addiction, 95 Supplement 2*(January), S211–7. doi: 10.1080/09652140050111771

Anton, R. F., Moak, D. H., & Latham, P. K. (1996). The Obsessive Compulsive Drinking Scale. A new method of assessing outcome in alcoholism treatment studies. *Archives of General Psychiatry, 53*, 225-231.

Anton, R. F., O'Malley, S. S., Ciraulo, D. A., Cisler, R. A., Couper, D., Donovan, D. M., … COMBINE Study Research Group (2006). Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence: The COMBINE Study: A Randomized Controlled

Trial. *The Journal of the American Medical Association, 295*(17), 2003-2017. doi: 10.1001/jama.295.17.2003

Arnau, R. C., Meagher, M. W., Norris, M. P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. Health Psychology: Official Journal of the Division of Health Psychology, *American Psychological Association, 20*(2), 112–119. doi: 10.1037//1040-3590.10.2.83

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory—Second Edition manual*. San Antonio, TX: The Psychological Corporation.

Beck, A. T., Steer, R., & Garbin, M. (1988). Psychometric properties of the Beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*, 77–100. doi: 10.1016/0272-7358(88)90050-5

Bohn, M. J., Barton, B. A., & Barron, K. E. (1996). Psychometric properties and validity of the Obsessive-Compulsive Drinking Scale. *Alcoholism: Clinical and Experimental Research, 20*(5), 817-823. doi: 10.1111/j.1530-0277.1996.tb05257.x

Bradley, A. P., & Longstaff, I. D. (2004). Sample size estimation using the receiver operating characteristic curve. *Proceedings of the 17th International Conference on Pattern Recognition, 4*, 428-431. doi: 10.1109/ICPR.2004.1333794

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (p. 136-162). Beverly Hills, CA: Sage.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471-492.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi: 10.1207/S15328007SEM0902_5

Cisler, R. A., & Zweben, A. (1999). Development of a composite measure for assessing alcohol treatment outcome: operationalization and validation. *Alcoholism: Clinical and Experimental Research, 23*(2), 263-271. doi: 10.1097/00000374-199902000-00011

Cisler, R. A., & Zweben, A. (2003). Clinical and methodological utility of a composite outcome measure for alcohol treatment research. *Alcoholism: Clinical and Experimental Research, 27*(10), 1680-1685. doi: 10.1097/01.ALC.0000091237.34225.D7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Connor, J. P., Feeney, G. F. X., Jack, A., & Young, R. M. (2010). The obsessive compulsive drinking scale is a valid measure of alcohol craving in young adults. *Alcoholism: Clinical and Experimental Research, 34*(12), 2155–2161. doi: 10.1111/j.1530-0277.2010.01312.x

Connor, J. P., Jack, A., Feeney, G. F. X., & Young, R. M. (2008). Validity of the Obsessive Compulsive Drinking Scale in a Heavy Drinking Population. *Alcoholism: Clinical and Experimental Research, 32*(6), 1067-1073. doi: 10.1111/j.1530-0277.2008.00668.x

Cordero, M., Solís, L., Cordero, R., Torruco, M., & Cruz-Fuentes, C. (2009). Factor structure and concurrent validity of the obsessive compulsive drinking scale in a group of alcohol-dependent subjects of Mexico City. *Alcoholism: Clinical and Experimental Research, 33*(7), 1145–1150. doi: 10.1111/j.1530-0277.2009.00937.x

Currie, S. R., El-Guebaly, N., Coulson, R., Hodings, D., & Mansley, C. (2004). Factor validation of the addiction severity index scale structure in persons with concurrent disorders. *Psychological Assessment, 16*(3), 326–329. doi: 10.1037/1040-3590.16.3.326

DeJong, C. A. J., Willems, J. C. E. W., Schippers, G. M., & Hendriks, V. M., (1995). The Addiction Severity Index: Reliability and validity in a Dutch alcoholic population. *The Internal Journal of the Addictions, 30*(5), 605-616. doi: 10.3109/10826089509048747

Del Boca, F. K., & Darkes, J. (2012). "Nothing is more practical than a good theory": Outcome measures in addictions treatment research. *Addiction, 107*, 719-726.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: an introductory report. *Psychological Medicine*. doi: 10.1017/S0033291700048017

DeVellis, R. F. (2012). *Scale Development: Theory and Applications* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

DiClemente, C. C., Carbonari, J. P., Montgomery, R. P., & Hughes, S. O. (1994). The Alcohol Abstinence Self-Efficacy scale. *Journal of Studies on Alcohol, 55*(2), 141–148. doi: 10.15288/jsa.1994.55.141

Donovan, D. M., Bigelow, G. E., Brigham, G. S., Carroll, K. M., Cohen, A. J., Gardin, J. G., … Wells, E. A. (2012). Primary outcome indices in illicit drug dependence treatment research: systematic approach to selection and measurement of drug use end-points in clinical trials. *Addiction, 107*(4), 694–708. doi: 10.1111/j.1360-0443.2011.03473.x

Egger, D., & Borg, J. S. (2016). Introduction to binary classification [online lecture]. Retrieved from Coursera Web Series by Duke University, *Mastering Data Analysis in Excel*: https://www.coursera.org/learn/analytics-excel/lecture/TUihw/introduction-to-binary-classification

English, D. R., Holman, C. D. J., Milne, E., et al. (1995). *The Quantification of Drug Caused Morbidity and Mortality in Australia 1995*. Canberra: Commonwealth Department of Human Services and Health.

European Medicines Agency. (2010). *Guideline on the development of medicinal products for the treatment of alcohol dependence*: 1-17. London: United Kingdom.

Falk, D., Wan, X. Q., Liu, L., Fertig, J., Mattson, M., Ryan, M., … Litten, R. Z. (2010). Percentage of subjects with no heavy drinking days: Evaluation as an efficacy endpoint for alcohol clinical trials. *Alcoholism: Clinical and Experimental Research, 34*(12), 2022-2034. doi: 10.1111/j.1530-0277.2010.01290.x

Feinn, R., Tennen, H., & Kranzler, H. R. (2003). Psychometric properties of the short index of problems as a measure of recent alcohol-related problems. *Alcoholism, Clinical and Experimental Research, 27*(9), 1436–41. doi: 10.1097/01.ALC.0000087582.44674.AF

Feragne, M. A., Longabaugh, R., & Stevenson, J. F. (1983). The Psychosocial Functioning Inventory. *Evaluation & The Health Professions, 6*(1), 25-48. doi: 10.1177/016327878300600102

Floyd, F. J. & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Food and Drug Administration. (2006). *Medical Review of Vivitrol*: 21-897. U.S. Government, Rockville, MD.

Food and Drug Administration. (2015). *Alcoholism: Developing Drugs for Treatment: Guidance for Industry*: 1-14. U.S. Government, Rockville, MD.

Forcehimes, A. A., Tonigan, J. S., Miller, W. R., Kenna, G. A., & Baer, J. S. (2007). Psychometrics of the Drinker Inventory of Consequences (DrInC). *Addictive Behaviors, 32*(8), 1699-1704. doi: 10.1016/j.addbeh.2006.11.009

Forgays, D. G., Forgays, D. K., & Spielberger, C. D. (1997). Factor Structure of the State-Trait Anger Expression Inventory. *Journal of Personality Assessment, 69*(3), 497–507. doi: 10.1207/s15327752jpa6903

Gorsuch, R. L., (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Grant, B. F., Goldstein, R. B., Saha, T. D., Chou, S. P., Jung, J., Zhang, H., … Hasin, D. S. (2015). Epidemiology of DSM-5 Alcohol Use Disorder: Results from the National Epidemiologic Survey on Alcohol and Related Conditions III. *Journal of the American Medical Association Psychiatry, Aug 72*(8), 757-766. doi: 10.1001/jamapsychiatry.2015.0584.

Greenfield, T. K. (2000). Ways of measuring drinking patterns and the difference they make: Experience with graduated frequencies. *Journal of Substance Abuse, 12*, 33-49. doi: 10.1016/S0899-3289(00)00039-0

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265-275. doi: 10.1037/0033-2909.103.2.265

Hagman, B. T., Kuerbis, A. N., Morgenstern, J., Bux, D. A., Parsons, J. T., & Heidinger, B. E. (2009). An Item Response Theory (IRT) analysis of the Short Inventory of Problems-Alcohol and Drugs (SIP-AD) among non-treatment seeking men-who-have-sex-with-men: Evidence for a shortened 10-item SIP-AD. *Addictive Behaviors, 34*(11), 948–954. doi: 10.1016/j.addbeh.2009.06.004

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Radiology, 143*, 29-36. doi: 10.1148/radiology.143.1.7063747

Hann, M., & Reeves, D. (2008). The SF-36 scales are not accurately summarised by independent physical and mental component scores. *Quality of Life Research, 17*(3), 413–423. doi: 10.1007/s11136-008-9310-0

Hasin, D., S., Wall, M., Witkiewitz, K., Kranzler, H. R., Falk, D. E., Litten, R. Z., Mann, K. F., O'Malley, S. S., Scodes, J., Robinson, R. K., Anton, R. F. (in press). Change in non-abstinent WHO risk drinking levels and alcohol dependence: A 3-year follow-up study in the United States general population. *Lancet Psychiatry.*

Heck R. H., & Thomas S. L. (2009). *An Introduction to Multilevel Modeling Techniques* (2nd ed.). New York, NY: Routledge.

Hiller, M. L., Broome, K. M., Knight, K., & Simpson, D. D. (2000). Measuring self-efficacy among drug-involved probationers. *Psychological Reports, 86*, 529–538.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144. doi: 10.1080/03610739208253916

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi: 10.1080/10705519909540118

IBM Corp. (2015). *IBM SPSS Statistics for Windows*. Version 23.0. Armonk, NY: IBM Corp.

Jackson, D. L., Gillaspy, J. R., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6-23. doi: 10.1037/a0014694

Jakobsson, U., Westergren, A., Lindskov, S., & Hagell, P. (2012). Construct validity of the SF-12 in three different samples. *Journal of Evaluation in Clinical Practice, 18*(3), 560–566. doi: 10.1111/j.1365-2753.2010.01623.x

Jaracz, K., Kalfoss, M., Górna, K., & Bączyk, G. (2006). Quality of life in Polish respondents: Psychometric properties of the Polish WHOQOL - Bref. Scandinavian Journal of Caring Sciences, 20(3), 251–260. doi: 10.1111/j.1471-6712.2006.00401.x

Kaskutas, L. A., Borkman, T. J., Laudet, A., Ritter, L. A., Witdbrodt, J., Subbaraman, M. S., … Bond, J. (2014). Elements that define recovery: The experiential perspective. *Journal of Studies on Alcohol and Drugs, 75,* 999-1010. doi: 10.15288/jsad.2014.75.999

Kenna, G. A., Longabaugh, R., Gogineni, A., Woolard, R. H., Nirenberg, T. D., Becker, B., … Karolczuk, K. (2005). Can the short index of problems (SIP) be improved? Validity and reliability of the three-month SIP in an emergency department sample. *Journal of Studies on Alcohol, 66*(3), 433–437. doi: 10.15288/jsa.2005.66.433

Kiluk, B. D., Dreifuss, J. a., Weiss, R. D., Morgenstern, J., & Carroll, K. M. (2012). The Short Inventory of Problems – Revised (SIP-R): Psychometric Properties Within a Large, Diverse Sample of Substance Use Disorder Treatment Seekers. *Psychology of Addictive Behaviors, 27*(1), 307–314. doi: 10.1037/a0028445

Kline, R. B. (2011). *Principles and practice of Structural Equation Modeling* (3rd ed.). New York: Guilford Press.

Kranzler, H. R., Mulgrew, C. L., Modesto-Lowe, V., & Burleson, J. A. (1999). Validity of the Obsessive-Compulsive Drinking Scale (OCDS): Does craving predict drinking behavior? *Alcoholism: Clinical and Experimental Research, 23*(1), 108-114. doi: 10.1111/j.1530-0277.1999.tb04030.x

Kroner, D. G., & Reddon, J. R. (1992). The anger expression scale and state-trait anger scale. *Criminal Justice & Behavior, 19*(4), 397–408. doi: 10.1177/0093854892019004004

Lenhard, W. & Lenhard, A. (2016). *Calculation of Effect Sizes*. available: https://www.psychometrica.de/effect_size.html. Dettelbach (Germany): Psychometrica. DOI: 10.13140/RG.2.1.3478.4245

Long, J. D., Harring, J. R., Brekke, J. S., Test, M. A., & Greenberg, J. (2007). Longitudinal construct validity of brief symptom inventory subscales in schizophrenia. *Psychological Assessment, 19*(3), 298–308. doi: 10.1037/1040-3590.19.3.298

Marra, L. B., Field, C. A., Caetano, R., & von Sternberg, K. (2014). Construct validity of the Short Inventory of Problems among Spanish speaking Hispanics. *Addictive Behaviors, 39*, 205-210. doi: 10.1016/j.addbeh.2013.09.02

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.

Marsh, H. W., Hua, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341. doi: 10.1207/s15328007sem1103_2

McLellan, A. T., Kushner, H., Metzger, D., Peters, R., Smith, I., Grissom, G., … Argeriou, M. (1992). The fifth edition of the Addiction Severity Index. *Journal of Substance Abuse Treatment, 9*(3), 199–213. doi: 10.1016/0740-5472(92)90062-S

McLellan, A. T., Luborsky, L., Woody, G. E., & O'Brien, C. P. (1980). An improved diagnostic evaluation instrument for substance abuse patients: The Addiction Severity Index. *Journal of Nervous Mental Disorders, 178,* 26-33. doi: 10.1097/00005053-198001000-00006

Miller, V. A., Reynolds, W. W., Ittenbach, R. F., Luce, M. F., Beauchamp, T. L., & Nelson, R. M. (2009). Challenges in measuring a new construct: Perception of voluntariness for research and treatment decision making. *Journal of Empirical Research & Human Research Ethics, 4*(3), 21-31. doi: 10.1525/jer.2009.4.3.21

Miller, W. R. (1996). Form 90: A structured assessment interview for drinking and related behaviors. *NIAAA Project MATCH Monograph Series, NIH Publication No. 96-4004,* Volume 5. Washington: Government Printing Office.

Miller, W. R., Tonigan, J. S., Longabaugh, R. (1995). *The Drinker Inventory of Consequences (DrInC): An Instrument for Assessing Adverse Consequences of Alcohol Abuse.* Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism.

Moak, D. H., Anton, R. F., & Latham, P. K. (1998). Further validation of the Obsessive-Compulsive Drinking Scale (OCDS). *American Journal of Addiction, 7*, 14-23. doi: 10.1111/j.1521-0391.1998.tb00463.x

Montanelli, R. G. & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41*, 341-348. doi: 10.1007/BF02293559

Montazeri, A., Vahdaninia, M., Mousavi, S. J., & Omidvari, S. (2009). The Iranian version of 12-item Short Form Health Survey (SF-12): factor structure, internal consistency and construct validity. *BMC Public Health, 9*, 341. doi: 10.1186/1471-2458-9-341

Moos, R.H., & Finney, J.W. (1983). The expanding scope of alcoholism treatment evaluation. *American Psychologist, 38*(10), 1036-1044. doi: 10.1037/0003-066X.38.10.1036

Morgenstern, J., Irwin, T. W., Wainberg, M. L., Parsons, J. T., Muench, F., Bux, D. A., … Schulz-Heik, J. (2007). A randomized controlled trial of goal choice interventions for alcohol use disorders among men who have sex with men. *Journal of Consulting and Clinical Psychology, 75*(1), 72-84. doi: 10.1037/0022-006X.75.1.72

Muthén, L. K., & Muthén, B. O. (2012). *Mplus users guide* (Version 7).

National Institute on Alcohol Abuse and Alcoholism. (2004). NIAAA Council approves definition of binge drinking. *NIAAA Newsletter*. Retrieved from: http://pubs.niaaa.nih.gov/publications/Newsletter/winter2004/Newsletter_Number3.pdf

Neale, J., Finch, E., Marsden, J., Mitcheson, L., Rose, D., Strang, J., … Wykes, T. (2014). How should we measure addiction recovery? Analysis of service provider perspectives using online Delphi groups. *Drugs: education, prevention and policy, 21*(4), 310-323. doi: 10.3109/09687637.2014.918089

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Wiley.

Polit, D. F., & Hungler, B. P. (1999). *Nursing Research: Principles and Methods*. Philadelphia, PHA: Lippincott.

Project MATCH Research Group (1997). Matching alcoholism treatments to client heterogeneity: project MATCH post-treatment drinking outcomes. *Journal of Studies on Alcohol, 58,* 7-29.

Project MATCH Research Group. (1998). Matching alcoholism treatments to client heterogeneity: Project MATCH three-year drinking outcomes. *Alcoholism: Clinical and Experimental Research, 22*(6), 1300-1311. doi: 10.1111/j.1530-0277.1998.tb03912.x

Recklitis, C. J., Parsons, S. K., Shih, M.-C., Mertens, A., Robison, L. L., & Zeltzer, L. (2006). Factor structure of the brief symptom inventory--18 in adult survivors of childhood cancer: results from the childhood cancer survivor study. *Psychological Assessment, 18*(1), 22–32. doi: 10.1037/1040-3590.18.1.22

Roberts, J. S., Anton, R. F., Latham, P. K., & Moak, D. H. (1999). Factor structure and predictive validity of the Obsessive Compulsive Drinking Scale. *Alcoholism: Clinical and Experimental Research, 23*(9), 1484–1491. doi: 10.1111/j.1530-0277.1999.tb04671.x

Rogalski, C. J. (1987). Factor structure of the Addiction Severity Index in an inpatient detoxification sample. *International Journal of Addiction, 22*(10), 981–992. doi: 10.3109/10826088709109693

Salyers, M. P., Bosworth, H. B., Swanson, J. W., Lamb-Pagone, J., & Osher, F. C. (2000). Reliability and validity of the SF-12 health survey among people with severe mental illness. *Medical Care, 38*(11), 1141–1150. doi: 10.1097/00005650-200011000-00008

Skevington, S. M., Lofty, M., & O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF Quality of life assessment: Psychometric properties and results of the international field trial A report from the WHOQOL Group. *Quality of Life Research, 13*(2), 299-310.

Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry, 39*, 135-140. doi: 10.1177/070674379403900303

Substance Abuse and Mental Health Administration. (2012). SAMHSA's working definition of recovery: 10 guiding principles of recovery. (SAMHSA Publication No. PEP12-RECDEF). Retrieved from http://store.samhsa.gov/shin/content//PEP12-RECDEF/PEP12-RECDEF.pdf

Tiffany, S.T., Friedman, L., Greenfield, S.F., Hasin, D.S., & Jackson, R. (2012). Beyond drug use: a systematic consideration of other outcomes in evaluations of treatments for substance use disorders. *Addiction, 107*, 709–718. doi: 10.1111/j.1360-0443.2011.03581.x

Tonigan, J. S., Connors, G. J., & Miller, W. R. (1996). Alcoholics Anonymous Involvement (AAI) scale: Reliability and norms. *Psychology of Addictive Behaviors, 10*, 75-80. doi: 10.1037/0893-164X.10.2.75

Treanor, C., & Donnelly, M. (2014). A methodological review of the Short Form Health Survey 36 (SF-36) and its derivatives among breast cancer survivors. *Quality of Life Research, 36*, 339–362. doi: 10.1007/s11136-014-0785-6

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods,* 3, 4–69. doi: 10.1177/109442810031002

Van Der Ploeg, H. M. (1988). The factor structure of the state-trait anger scale. *Psychological Reports, 63*, 978. doi: 10.2466/pr0.1988.63.3.978

Visser, M., Leentjens, A. F. G., Marinus, J., Stiggelbout, A. M., & van Hilten, J. J. (2006). Reliability and validity of the Beck Depression Inventory in patients with Parkinson's disease. *Movement Disorders, 21*(5), 668–672. doi: 10.1002/mds.20792

Wang, J., Kelly, B. C., Booth, B. M., Falck, R. S., Leukefeld, C., & Carlson, R. G. (2010). Examining factorial structure and measurement invariance of the Brief Symptom Inventory (BSI)-18 among drug users. *Addictive Behaviors, 35*(1), 23–29. doi: 10.1016/j.addbeh.2009.08.003

Ware, J. E. Jr., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*, 220-233. doi: 10.1097/00005650-199603000-00003

WHOQOL Group. (1998). Development of the World health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine, 28*, 551-558.

Widaman K. F., Ferrer E., & Conger R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child Development Perspectives, 4*, 10–18. doi: 10.1111/j.1750-8606.2009.00110.x

Witkiewitz, K. (2013). Temptation to drink as a predictor of drinking outcomes following psychosocial treatment for alcohol dependence. *Alcoholism: Clinical and Experimental Research, 37*(3), 529-537. doi: 10.1111/j.1530-0277.2012.01950

Witkiewitz, K., Hallgren, K. A., Kranzler, H. R., Mann, K. F., Hasin, D. S., Falk, D. E., Litten, R. Z., & Anton, R. F. (2017). Clinical validation of reduced alcohol consumption after

treatment for alcohol dependence: Using the World Health Organization risk drinking levels. *Alcoholism: Clinical and Experimental Research, 41* (1), 179-186.

World Health Organization. (2000). *International Guide for monitoring alcohol consumption and related harm*. retrieved from: whqlibdoc.who.int/HQ/2000/WHO_MSD_MSB_00.4.pdf

World Health Organization. (2011). *Global Status Report on Alcohol and Health*. Geneva, Switzerland.

Yao, G., & Wu, C. H. (2005). Factorial invariance of the WHOQOL-BREF among disease groups. *Quality of Life Research, 14*(8), 1881-1888. doi: 10.1007/s11136-005-3867-7

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*, 115–148. doi: 10.1207/s15327906mbr4001_5

Zwick, W. R., & Velicer, W. F. (1986). A comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-44. doi: 10.1037/0033-2909.99.3.432