HHS Public Access

Author manuscript

Clin Psychol Sci. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

Clin Psychol Sci. 2015 November 1; 3(6): 819–835. doi:10.1177/2167702614553026.

Limits of Current Approaches to Diagnosis Severity Based on Criterion Counts: An Example with DSM-5 Alcohol Use Disorder

Sean P. Lane and **Kenneth J. Sher**

University of Missouri-Columbia and the Midwest Alcoholism Research Center, Columbia, MO, USA

Abstract

Within DSM-5, some diagnoses are now associated with a severity gradient based on the number of diagnostic criteria satisfied. Reasons for questioning the validity of this approach include the implicit assumptions of equal criterion severity and strict additivity of criteria combinations. To assess the implications of heterogeneity of criterion configurations on severity grading, we examined the association between all observed combinations of DSM-5 alcohol use disorder criteria endorsement, at each level of number of criteria endorsed, and multiple validity measures among 22,177 past-year drinkers from Wave 2 of the NESARC. Substantial variability of implied severity across criteria combinations was observed at each level of endorsement, with nontrivial overlap in implied severity across criterion counts. Findings suggest severity indices are at best imprecise and, potentially, misleading. These problems are likely inherent in traditional polythetic approaches to diagnosis and almost certainly applicable to other disorders. Approaches for improving severity grading are proposed.

Keywords

alcohol use disorder; clinical diagnosis; DSM-5; NESARC; severity; validity

The predominant approach to the assessment and diagnosis of mental disorders for the past six decades, both within the United States (e.g. DSM-5; APA, 2013) and internationally (e.g. ICD-10; WHO, 1992), has been largely categorical, based on the satisfaction of a necessary and/or sufficient number of established criteria. In many cases, this involves simple criterion counts in which the various criteria within a given criteria set are each given equal weight in determining the presence and/or severity of a disorder. However, for many, if not most disorders, diagnostic criteria inherently vary in their severity (Chung & Martin, 2005; Cooper & Balsis, 2009; Galatzer-Levy & Bryant, 2013; Martin et al., 2006). This has important implications for any polythetic approach to diagnosis where part of the diagnostic algorithm involves criterion counts and no single criterion is necessary or sufficient. For example, for a given disorder that has a threshold for diagnosis of four endorsed criteria, endorsing the three most severe criteria and falling below threshold is considered less

Corresponding author: Correspondence concerning this article should be addressed to Sean P. Lane, Department of Psychological Sciences, Psychology Building, 200 South Seventh Street, Columbia, MO 65211. *Phone*: +1 573 8841485. *Fax*: +1 573 884 5588. lanesp@missouri.edu.

adverse than endorsing the four least severe criteria and obtaining a diagnosis (see discussion on *diagnostic orphans* below).

More sophisticated approaches such as item response theory (IRT) have been employed across various disorders' criterion sets to assess the likely impact of and perhaps provide a solution to the equal weighting issue. Indeed, such methods are gaining popularity and have been applied to criterion sets of various personality (e.g. Cooper & Balsis, 2009; Balsis, Gleason, Woods, Oltmanns, 2007), mood (e.g. Uebelacker, Strong, Weinstock, & Miller, 2009), and substance use disorders (e.g. Langenbucher, Labouvie, Martin, Sanjuan, Bavly, Kirisci, & Chung, 2004). However, even IRT methods, with a particular focus on the unidimensional applications favored within the clinical diagnosis literature, are limited in that they assume an additive model to the underlying latent construct and ignore unique configural information that may be gathered from specific combinations of endorsed criteria beyond the information implied by a linear IRT-derived factor score. That is, particular subsets of criteria may confer super-additive or sub-additive information relating to the underlying construct. In the current investigation we seek to evaluate the validity of an additive approach to diagnosis, both in practice and for basic research, using as an example DSM-5 alcohol use disorder (AUD).

The problem with polythetic diagnosis was clearly demonstrated by Cooper and Balsis (2009) in their analysis of the psychometric properties of diagnostic criteria for schizoid personality disorder using data from the National Epidemiological Survery on Alcoholism and Related Conditions (NESARC; Grant, Moore, & Kaplan, 2003; Grant, Stinson, Dawson, Chou, & Ruan, 2005). They showed that some configurations of subthreshold symptomatology implied greater severity than some configurations at diagnostic threshold. Examining the distribution of implied latent severity of different suprathreshold configurations revealed that there is overlap in latent severity associated with different symptom configurations across all adjacent symptom counts except for the most severe configuration where all items are endorsed (Cooper and Balsis, 2009, their Figure 2). Given the distribution of criteria thresholds from various IRT analyses of AUD criteria, the same basic problem can be directly inferred with respect to AUD (e.g. Dawson, Saha, & Grant, 2010; Keyes et al.,2011; Langenbucher et al., 2004; Mewton et al., 2011a. The Cooper and Balsis (2009) analysis, as well as other recent analyses (e.g. posttraumatic stress disorder; Galatzer-Levy & Bryant, 2013), highlight that any categorical diagnosis based on a criteria set can be conceptualized on an underlying severity metric and that variability in criteria severity can lead to ambiguity with respect to the meaning of the diagnostic threshold in categorical diagnosis. This is particularly problematic if that diagnostic boundary has implications for research on diagnostic groups and for clinical decision-making surrounding the need to treat. To that end, more explicit severity gradients to diagnosis have been adopted for various mood and substance use disorders, and similar suggestions have been made with respect to personality disorders (e.g. Section III DSM-5, APA, 2013; Widiger & Trull, 2007). However, to the extent that severity gradients still form cutoffs in terms of criterion endorsement across various levels the same problems inherent in the categorical approach may still apply. Currently, there are three different, albeit related, approaches to grading severity within DSM-5: 1) a simple count of endorsed criteria (e.g. substance use

disorders [SUDs]; p. 491; APA, 2013), 2) a hierarchical approach where criteria count is used with severity being graded contingent upon the endorsement of a necessary symptom (e.g. unspecified depressive disorder; p. 184; APA, 2013), and 3) a more vague algorithm combining a threshold for criteria count with an individualized grading of given criteria (e.g. bipolar I disorder; p. 154; APA, 2013). We chose SUDs, focusing on AUDs, because they have the most concrete definition by which to assess the effect of non-additivity (i.e. a simple count), and because AUDs generally have the highest prevalence rate among SUDs and DSM conditions overall (Grant et al., 2006; Kessler et al., 2005; Regier et al., 1990).

The notion that some symptom configurations of AUD criteria that are suprathreshold can imply less pathology than other configuration that are subthreshold has been discussed in several contexts. Martin and colleagues (1999, 2002, 2006) have argued that in the DSM-IV, individuals who fall below the alcohol dependence threshold of three criteria, only endorsing one or two, and do not qualify for abuse ("*diagnostic orphans*"), are actually more similar to those who qualify for abuse, and more dissimilar from others who endorse no criteria. This, in part, motivated the shift in diagnosis from DSM-IV to DSM-5 in which individuals who endorse just two dependence criteria would not qualify for dependence using DSM-IV, but would qualify as having a mild AUD using the new unidimensional DSM-5 grading (see Dawson, Goldstein, & Grant, 2013; Compton, Dawson, Goldstein, & Grant, 2013). Similarly, with regard to alcohol abuse, which had a one-criterion threshold in the DSM-IV guidelines, some adolescents and young adults with relatively low levels of alcohol use and correspondingly low endorsement of dependence criteria, may qualify for abuse through endorsement of role interference or interpersonal problems as a result of conflict with parents with strict parenting practices (Martin, Chung, & Langenbucher, 2008). These individuals have been referred to as "*diagnostic impostors*" (Langenbucher, Martin, Hasin, & Helzer, 1996). While the DSM-IV would qualify these individuals for abuse due to their endorsement of a single criterion, under DSM-5 guidelines they would not qualify for any AUD. Although not likely eliminating the problem (Martin, Steinley, Verges, & Sher, 2011), the requirement of at least two criteria should help somewhat with both the diagnostic impostor and diagnostic orphan problems noted for the DSM-IV.

Relatedly, others have argued that some symptoms are so severe as to be pathognomic (i.e., sufficient) for a diagnosis of alcohol dependence. Langenbucher and colleagues (2000) have argued this should be the case for withdrawal. Although in the ICD-10 (World Health Organization [WHO], 1992) withdrawal is not viewed as pathognomic of dependence, it is accorded a special status in grading severity. Like the DSM-5, ICD-10 grades severity of dependence based on symptom counts, but the presence of withdrawal automatically warrants a severity grading of moderate severity (or higher). That is, dependence characterized by withdrawal *ipso facto* implies moderate or severe dependence. This convention represents an implicit weighting of withdrawal as a qualitatively more severe criterion.

However, there is little in the IRT studies of AUD criteria that suggests that withdrawal should warrant a unique status with respect to pathognomicity or severity grading. Indeed, in most published studies other items appear to be more severe, including giving up important activities, failure to fulfill role obligations, and continued use despite interpersonal problems

(e.g. Dawson, Saha, & Grant, 2010; Saha, Chou, & Grant, 2006); although there is considerable variability in findings with withdrawal sometimes found to be a severe criterion (e.g. Gilder, Gizer, & Ehlers, 2011; McCutcheon et al., 2011). These between-study differences highlight how IRT studies can be misleading since all IRT studies of AUD are dependent upon the specific operationalization of each criterion; how criteria, such as withdrawal, are assessed varies across diagnostic instruments employed in different studies. Despite this potentially serious limitation, IRT studies of latent severity provide a psychometric foundation for viewing severity grading of mental disorders in AUDs based upon equal weighting with some degree of skepticism (e.g. Dawson, Saha, & Grant, 2010; Kahler & Strong, 2006; Krueger, Nichol, Hicks, Markon, Patrick, & McGue, 2004; Langenbucher et al., 2004; Martin, Chung, Kirisci, & Langenbucher, 2006; Saha, Chou, & Grant, 2006).

Dawson, Saha, and Grant (2010), using the DSM-IV AUD criteria set in Wave 1 of NESARC, found that the simple criterion counts were highly correlated with latent trait scores derived from IRT analysis, and they concluded that, from a practical perspective, accounting for the severity of individual criteria provides little additional information that would strongly support the use of weighted criterion counts (essentially factor scores). They note that “severity weighting would be useful if it could distinguish individuals with just a few positive criteria, all of which were mild, from those with just a few criteria, all of which were severe; however, our data clearly indicated that among individuals with just a few positive criteria, there was a very low likelihood that those criteria would be severe ones” (p. 37; Dawson, Saha, & Grant, 2010). That is, they argue that with the DSM-5 AUD criteria set as it is currently constructed, there is unlikely to be much advantage of an IRT-based severity rating over a simple criteria count.

We view the Dawson et al. (2010) analyses as providing an important “first step” in evaluating the reasonableness of severity grading of AUD (and other disorders) in DSM-5. However, we view their analysis as incomplete in that it fails to fully consider the range of criteria configurations that can occur at a given criteria count in a polythetic diagnostic framework. Just considering those individuals who are at threshold for diagnosis of AUD (i.e., meeting exactly two of 11 criteria), there are 55 possible combinations (although not all might be observed in real data). Because both IRT-based factor scores and simple criterion counts are assuming additive, linear models, important configural information in the patterning of criteria is not resolved. To the extent that a simple additive model represents a poor representation of the underlying data, it is possible that both simple criterion counts and IRT-based factor scores fail to adequately estimate likely severity.

We investigated the extent to which there is variation in severity across different configurations of criteria for a given symptom count, both with respect to an internal, theoretically-based AUD severity indicator (i.e., latent IRT-derived AUD factor scores) and multiple external, empirically driven and clinically relevant AUD severity indicators (i.e., behavioral correlates; Burns & Teesson, 2002; Dawson & Grant, 2010; Dawson, Saha, & Grant, 2010; Pollock & Martin, 1999; Saha, Stinson, & Grant, 2007; WHO, 2011). Our goal was to first cross-validate our external severity measures with results from past IRT analyses and our own IRT analysis of the current data. In doing so, we identify/corroborate that

particular criteria are associated with both greater latent and manifest measures of severity. We then estimate the degree of variation in latent (i.e., IRT-based) and empirical (i.e., external criterion-based) severity indicators across the different specific criteria configurations *within* each level of criteria endorsement (from 0 to 11). To the extent that there is relatively high variation in severity *within* a criteria count that then directly leads to substantial overlap in implied severity *between* adjacent symptom counts, the utility of the DSM-5 approach to grading severity is compromised. Such a result suggests that the configuration of criterion endorsement is an important but severely neglected factor in determining clinical diagnosis, and while theoretically clusters of criteria may be expected to co-occur more/less often within a given disorder, both count and IRT algorithms implicitly ignore this despite the fact that such clusters may be serve theoretical and practical utility.

Method

Sample

The NESARC is a nationally representative sample of non-institutionalized United States civilians 18 years and older. The survey oversampled minority ethnicities (Blacks and Hispanics) and young adults between the ages of 18 and 24. The initial wave, which was administered using face-to-face interviews between 2001–2002 contained 43,093 respondents (Grant, Moore, & Kaplan, 2003). A second follow-up wave of assessment was conducted during 2004–2005 and contained 34,653 of the same respondents (Grant & Kaplan, 2005). In the current analyses we limit our sample to the 22,177 individuals (male = 10,395 [47%], female = 11,782 [53%]) in Wave 2 who had consumed at least one alcoholic beverage in the past year. We did so because individuals who abstained from drinking would contribute little to no variance to both AUD criteria endorsement and the drinking-related validity measures. Similarly, we focused on the Wave 2 assessment because it included items corresponding to the DSM-5 AUD craving criterion whereas Wave 1 did not. Individuals ranged in age from 20–90 years (M = 45.9, SD = 15.9), with a majority White (63.3%) and the remainder Hispanic (17.4%), Black (15.4%), Asian (2.2%), and Native American (1.6%). Analyses weighted participants to approximately represent the overall population of the United States.

Measures

AUD Criteria—The 11 criteria used to diagnose the presence and severity of AUD include 1) drinking larger amounts or for longer periods of time than intended, 2) attempts to cut down or control use, 3) spending a lot of time obtaining, using, or recovering, 4) craving or strong desires/urges to use, 5) experiencing withdrawal symptoms, 6) increased tolerance to alcohol's effects, 7) using alcohol in physically hazardous situations, 8) giving up important social, occupational, or recreational activities, 9) continued use despite physical or psychological problems, 10), failure to fulfill role obligations, and 11) continued use despite social and interpersonal problems. Using DSM-5 guidelines (American Psychiatric Association, 2013) for criterion endorsement each criterion was coded as present ('1') or absent ('0') based on individuals' responses to the particular criterion items using the Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV;

Grant, Dawson, Stinson, Chou, Kay, & Pickering, 2003). Population-weighted prevalence rates and associated IRT parameters for the current sample are shown in Table 1.

Drinking Behavior—To create an overall index of the severity of respondents' self-reported drinking behavior we standardized and then averaged 7 different questions that asked about various aspects of their drinking behavior over the past year (Dawson, Grant, Stinson, & Zhou, 2005). These included 1) the maximum number of standard drinks that a respondent consumed in a single day in the last year, 2) how often a person drank his/her largest number of drinks over the past year, 3) the usual number of drinks a person reported drinking on days when he/she drank over the past year, 4) the number of days in the past year that a respondent reported drinking *any* alcoholic beverage, 5) the frequency with which an individual reported drinking enough to feel intoxicated over the past year 6) the number of times in the past 12 months that a person engaged in binge drinking behavior (>5 drinks in a two hour window for men, >4 for women), and 7) the number of times an individual exceeded the recommended physicians' limits for risky drinking (>14 drinks per week or >4 drinks on a given day for men, and >7 drinks per week or >4 drinks on a given day for women; NIAAA, 1995). The standardized behavioral drinking measure had good reliability (α = .84; see Supplementary Material for more detailed information regarding each drinking measure).

General Physical and Mental Functioning—A measure of overall health and well-being was constructed from the SF12-V2 physical and mental functioning scales (Ware, Kosinski, Turner-Bowker, & Gandek, 2002). There are 10 subscales, all of which were positively correlated and loaded onto a single factor, so they were averaged into an overall index (α = .90 across the 10 subscales). The scales were originally constructed to have individual means of 50 and standard deviations of 10. To retain this metric but facilitate comparison with the other validity measures the overall health index was reverse scored based on an assumed mean of 50. The resulting health index had a mean of 48.5 (*SD* = 7.1) with higher scores indicating worse health.

Lifetime Axis I and Axis II Diagnoses—A sum score of the total number of other (i.e. non-AUD/SUD) lifetime Axis I and Axis II diagnoses was created using responses from Wave 1 and Wave 2. The same eleven Axis I disorders were assessed in both the first and second waves of the NESARC (Major depressive episode, Manic episode, Dysthymic episode, Hypomanic episode, Panic disorder with agoraphobia, Panic disorder without agoraphobia, Agoraphobia without panic disorder, Social phobia, Specific phobia, Generalized anxiety disorder, Posttraumatic stress disorder). Seven Axis II personality disorders (PDs) were assessed at Wave 1 (Avoidant, Dependent, Paranoid, Obsessive-Compulsive, Schizoid, Histrionic, and Antisocial). In contrast, only three PDs were assessed at Wave 2 (Borderline, Schizotypal, Narcissistic). For Antisocial PD, Wave 1 included the assessment of Conduct Disorder before age 15 and Adult Antisocial Behavior at or after age 15, whereas Wave 2 included only an assessment of Adult Antisocial Behavior since the last interview.

We used lifetime diagnoses to facilitate aggregation across Wave 1 and Wave 2 measures given the unbalanced design. This involved using lifetime measures of the eleven Axis I

disorders assessed across Waves 1 and 2, lifetime measures of the three Axis II disorders assessed at Wave 2, the lifetime measures of the six Axis II disorders assessed at Wave 1 (acknowledging the limitation that individuals could have developed these disorders after Wave 1), and an inclusive combination of individuals who qualified for either Wave 1 lifetime Antisocial PD or both Wave 2 Adult Antisocial Behavior and Wave 1 pre-age-15 Conduct Disorder. Presence/absence of each disorder was summed to create an indicator of total number of Axis I and Axis II disorders (M = 1.35, SD = 2.13).

Composite Measure—Because none of our external criteria can be thought of as a "gold standard" and each holds considerable "unique" variance, we also created a composite of the three external/manifest AUD severity indicators, the above summary measures for drinking behavior, general health, and lifetime Axis I and II diagnoses. These were individually standardized and then averaged to create a composite. (α = .35). Note that from our perspective, this composite is not meant to be unidimensional, nor are the constituent variables meant to be viewed as "effect indicators" (i.e., caused by the same latent variable; see Bollen & Lennox, 1991). Rather, we view the measure as a multifaceted composite of three broad domains of AUD-related life challenges and thus do not view the low coefficient alpha as problematic for present purposes. While we do not view the composite as a "gold standard" either, having a summary index that contains elements of heaviness of drinking pattern, general well being, and other psychopathology provides a convenient variable that encompasses a range of correlates of pathological alcohol involvement.

Results

Item Response Theory Analysis

In an effort to provide a point of comparison between the current analyses and past findings regarding criteria severity from the IRT literature, we first conducted an IRT analysis using the 11 DSM-5 AUD criteria for our sample of Wave 2 respondents (Table 1). We conducted the analysis using Mplus 6.0 (Muthén & Muthén, 2010) taking into account primary sampling unit, stratum, and sample weights. Rank-order correlations between threshold estimates, which are used to index criteria severity, for the 10 overlapping criteria from DSM-IV to DSM-5 for our analysis compared to other published NESARC IRT analyses were very high (ρ = .99 – Dawson, Saha, & Grant, 2010; ρ = .99 – Saha, Chou, & Grant, 2006; ρ = .99 – Saha, Stinson, & Grant, 2007).[1] Our results replicate past findings, at least those using the NESARC, in terms of the ordering of individual criterion severity. This analysis serves as a validity check for both our external severity measures and the IRT methodology as described in the next two sections.

Criteria Severity Validation using External Severity Indicators

Next, we performed a set of regression analyses with the 11 DSM-5 AUD criteria predicting the four external severity indicators (again using Mplus 6.0 to maintain consistent

[1]We note that although our findings were not fully consistent (ρ = .86) with those reported by Casey, Adamson, Shelvin, and McKinney (2012) despite both of us using DSM-5 and Wave 2 NESARC data, this is attributable to those authors reporting an IRT parameterization of the threshold estimates (which incorporates threshold and discrimination information) instead of standardized estimates. When transformed our threshold estimates and theirs correlate .99.

adjustment for the sampling weights and stratum in the calculation of standard errors across different analyses; Muthén & Muthén, 2010). These tested the bivariate association for each criterion with each of the four external severity indicators. Across all four external severity indicators, we consistently observe that giving up activities, role interference, and interpersonal problems are the three most "severe" items whereas trying to cut down, drinking larger quantities or for longer periods of time, withdrawal, and tolerance are consistently the least severe criteria (see Table S1 and Figure S1 in Supplementary Material). More generally, not only do the rank orderings of the individual criterion parameter estimates all correlate above .90 between the four external severity measures (showing consistency across observed measures), but the rank orderings of the criterion parameter estimates for each of the four measures also all correlate above .90 with the IRT criterion thresholds (showing consistency across methodologies). Importantly, the effects sizes for each criterion using our composite measure closely paralleled the analogous standardized thresholds of the IRT analysis (r = .96). In addition to demonstrating the nature and extent of differential severity across the 11 criteria, these results validate our constructed measures, particularly the composite measure, as indicators of AUD severity, and provide external criterion validity for the IRT methodology and results. These results allow us to more concretely compare IRT latent factor score indicators of severity and our observed severity measures when assessing within-category (i.e., symptom count) and between-category variability in the different combinations of criteria endorsement.

Illustrating the Severity Problem

To provide graphical evidence that individual criteria, and more importantly, different combinations of criteria, confer different degrees of AUD severity above and beyond a count of how many criteria are endorsed, we plotted the number of criteria endorsed against our severity indicators, but for each level of endorsement we partitioned individuals into the various unique combinations of criteria endorsement (see Table 2 and Figures 1–2). Table 2 shows the number of sample participants falling within each level of criteria endorsement (0 to 11), the number of possible combinations of the 11 criteria at each level of criteria endorsement, and the actual number of the these criteria combinations observed in the current data. In addition, it shows the proportion of possible criteria combinations that are not observed in the current data at each level of criteria endorsement, with a corresponding adjusted value that accounts for some criterion counts not having as many individuals as there are possible combinations for that count.

We begin by using this method of data presentation to depict variability in our internal, latent AUD severity indicator (factor scores derived from the above IRT analyses) across all observed combinations of criteria at each level of criterion endorsement (see Figure 1). In Figure 1 each circle reflects a different unique combination of criteria that are endorsed at a given endorsement level, with the size of the circles reflecting the number of individuals with that unique criteria combination. For example, at endorsement levels of 1 criterion and 10 criteria (see x-axis) there are 11 possible criteria combinations, and so at those levels of criteria endorsement there will be up to 11 unique circles corresponding to these 11 unique groups of individuals. The position of each circle on the y-axis represents the mean AUD IRT factor score for those with that particular criteria combination. Collapsing across

different criteria combinations within each endorsement level, overall AUD IRT factor score means for each endorsement level are shown with black squares. Since each of the external measures was generally positively skewed we opted to plot overall AUD IRT factor score medians for each endorsement level with the blue line. At each level of criteria endorsement we also provide the inter-quartile ranges of AUD IRT factor scores across all combinations in order to illustrate the relative overlap of adjacent endorsement levels. Finally, the different DSM-5 severity categories are demarcated by solid vertical lines.

The graphic for the IRT factor scores (Figure 1) shows substantial heterogeneity in severity across combinations within a given endorsement category. It also shows that this variability results in some overlap in latent severity scores between adjacent endorsement categories such that a total of 18% of individuals within any given endorsement category have higher latent IRT factor scores than some individuals in the endorsement category above them or lower scores than some individuals in the endorsement category below them.

In general, Figure 1 seems to partially support the severity grading of DSM-5, with only modest overlap between the inter-quartile ranges across adjacent severity categories. However, this may be partially due to certain structural features of IRT analysis, such that it assumes criteria additivity and ignores specific configural information. These assumptions have the propensity to limit the range of assigned factor scores and, as a result, artificially restrict the observed overlap across endorsement categories (Linden & Hambleton, 1997). Consistent with this interpretation, we note that, while the rank orderings of criterion severity were highly correlated when comparing AUD IRT thresholds and standardized regression model estimates for the external severity indicators (r = .99), the individual IRT-derived factor scores and the external composite scores at the person level were much less correlated (r = .48). Although we would expect the correlation with external correlates to be substantially lower both because they are not “gold standards” and because there is error of measurement in each of the external severity indicators that will attenuate the correlation, the difference in correlation is considerable and it is that issue we turn to next.

To contrast the theoretical approach to AUD severity using IRT with a more empirical approach, Figure 2 shows the same graphical representation but for the four external severity measures. For each of the severity measures, though most so for the drinking behavior (Figure 2A) and composite (which includes drinking behavior; Figure 2D) measures, we observe a general monotonic trend of increasing severity as the number of criteria endorsed increases. This provides partial support for operationalizing severity using just the total number of criteria endorsed. However, across the various combinations of criteria, at each level of number of criteria satisfied we find substantial differentiation across each of the different validity measures. For example, the implicit DSM-5 assumption is that any combination of three criteria should be as severe as any other three criteria. In contrast, we find that when individuals endorse exactly three criteria some combinations are associated with rather low levels of severity while other combinations are associated with much higher severity. Furthermore, when comparing adjacent levels of criteria satisfied, especially those that define DSM-5 severity thresholds (i.e., mild = 2–3, moderate = 4–5, and severe = 6+), we observe, for example, that satisfaction of exactly three criteria (i.e., mild AUD) is, in many cases, as severe as satisfaction of a combination of four other criteria. And in many

cases the satisfaction of exactly four criteria (i.e. moderate AUD severity) is *less* severe than satisfaction of exactly three, or even two, criteria (i.e. mild AUD severity). More concretely, using the inter-quartile ranges as the boundaries for overlap across endorsement categories fully 100% (lifetime diagnoses), 81% (health and well-being), 53% (drinking behavior), and 66% (composite) of individuals overlap with the severity scores of individuals in a different endorsement category. In terms of overlap between the DSM-5 severity gradings the analogous values are 100%, 73%, 33%, and 48%, respectively. Both observations, of heterogeneity within endorsement categories and comparability of different combinations of criteria between categories, are most apparent for the health and well being (Figure 2B) and lifetime diagnoses (Figure 2C) measures with their shallower slopes. However, importantly, both are still observed to a large extent in the measures that exhibit much steeper slopes. The differences in the heterogeneity between the latent IRT values and manifest external severity values across the different permutations of criteria at different categories of endorsement reflect a potential limitation of IRT approaches when characterizing AUD severity; namely, that it assumes criteria additivity and ignores specific non-additive configural information.

Exploratory Analysis of Severity in Pairs of AUD Criteria

Up to this point we have reviewed and explicitly quantified the effects that individual criteria may assert on both internal (i.e., IRT) and external indicators of AUD severity. Figures 1 and 2 show that different combinations of items are indeed associated with varying degrees of severity; however, these analyses do not resolve which particular combinations of criteria those correspond to, and if these combinations are systematically associated with more/less severity across levels of endorsement. We therefore conducted an exploratory analysis aimed at identifying particular combinations of criteria that are associated with different degrees of AUD severity focusing on the 55 possible pairs of criteria endorsement for two reasons: (1) simultaneous endorsement of two criteria is the minimum required to classify for an AUD and therefore represents the nominal transition from no AUD to mild AUD, (2) criterion pairs are more manageable in terms of the number of combinations compared to triads, quadruplets, or quintuples; and they represent the next step in configural complexity compared to singlets.

Before exploring the impact of individual criterion pairs on levels of observed severity, we first examined the evidence that there is indeed systematic configural information that accounts for variance above and beyond that which is accounted for by simple symptoms counts, as well as any additional variance that might be explained by IRT scores that incorporate criterion weights. We note that algebraically each of the 2,048 criterion configurations is formally nested within each of the possible IRT factor score combinations, which are themselves formally nested within each level of criterion counts. Therefore we can treat configuration as a random effect in separate analyses with drinking behavior, health and well-being, total diagnoses, and our composite measure as the outcomes to observe 1) the proportion of total variation that is explainable by configuration, and 2) the proportion of systematic configural variation that remains when we include symptom count and IRT factor scores into the model as predictors. We fit a pair of multilevel models in Mplus 6.0 (Muthén & Muthén, 2010) for each of the four external measures in which individuals were measured at Level 1 and criterion configuration at Level 2. We accounted

for primary sampling unit, stratum, and sampling weights as in the previous analyses. The first model was a null model that only fit a random intercept for criterion configuration such that the intraclass correlation estimated by the ratio of the configuration variance and total variance (configuration + error) indexed the proportion of total variability in a given outcome accounted for by configural information. The second model included the total number of criteria endorsed and individuals' latent factor scores from the IRT analysis as independent variables. Since both predictors are nested within specific criterion configurations, the extent to which they account for systematic variation in outcome scores will reduce the amount of variability in the random effect of configuration.

These analyses revealed that 12.1%, 6.3%, 10.6%, and 10.8% of the total variance, for drinking behavior, health and well-being, total diagnoses, and the composite, respectively, was accounted for by the configural information (all *p*s < .001). Adding criterion count and IRT factor scores into the model, this variation was reduced to 7.8%, 6.1%, 9.9%, and 8.0%, respectively; but fully 61.0%, 96.3%, 92.5%, and 71.6% (all *p*s < .001) of the *systematic* configural variation in drinking, health, diagnoses, and the composite remained. These analyses substantiate our interest in looking for specific configurations that are associated with systematically higher/lower severity, and confirm that the heterogeneity observed in Figure 2 across the different indicators is not simply random error.

Next focusing on just the different pairs of criteria, we conducted an omnibus test to determine if there was meaningful variability to be accounted for by the set of all 55 pairs (independent of criteria count). We did so by comparing a base model that included each of the 11 criteria as multivariate predictors of our global composite severity indicator (R^2 = 26.6%) to a model that also included each of the 55 criterion pairs as predictors (R^2 = 27.6%). The improvement in model fit was highly significant ($\chi^2(55) = 160.45$, $p < 3\times10^{-12}$). We note that while the relative improvement in R^2 was 1%, this is quite substantively meaningful in light of the low base rates of endorsement across criteria (Judd, McClelland, & Ryan, 2011).

Next, within levels of criteria endorsement, we assessed differential severity among the 55 criterion pairs by making comparisons among them based on scores from the composite severity index. Specifically, we first computed the median composite score within each level of endorsement (2 through 10), then classified each participant as either above or below the median composite score for his/her level of endorsement, and then assigned each participant to criteria pairs based on the specific criteria he/she endorsed.[2] This allowed us to calculate the proportion of times when endorsement of a given criteria pair was associated with a composite score above the median, and then compare the proportions across the different criteria pairs, viewing high proportions as indicating greater severity.[3]

[2]Note that for endorsement categories from 3 to 9, individuals' composite scores will be repeated because for individuals endorsing 3 criteria (e.g. tolerance, craving, withdrawal) they belong to 3 unique pairs (tolerance & craving, tolerance & withdrawal, craving & withdrawal).

[3]Parallel analyses for the other three severity measures are available upon request. Overall, the pattern of results for the most/least severe pairs was consistent across measures.

Table S2 in the Supplementary Materials lists those proportions within each level of endorsement, along with a weighted average of proportions across all endorsement levels, which adjusts for the number of individuals at each endorsement level for each unique pairing. Notably, the pairings with the greatest implied severity (i.e., the highest weighted proportions) tended to be include one or two criteria that were individually associated with more "moderate" as indicated by our IRT and regression analyses (e.g. tolerance, time spent using/recovering). The pairings associated with lower implied severity show a less consistent pattern, but often appear to include at least one criterion that was individually associated with relatively high severity (e.g. role interference, giving up activities). Thus, although preliminary and exploratory, these findings appear consistent with the notion that the meaning (e.g., severity) of endorsing a given AUD criterion may vary as a function of the other criteria that are endorsed along with it. We note that the highest severity pairs (not singletons) often contain criteria that are less correlated with one another compared to the bivariate associations of other criteria, whereas the lowest severity pairs tend to be relatively more correlated. This suggests that unique pairs that share less empirical overlap provide more information about AUD severity such that the endorsement of those two criteria more likely represents two separate aspects of AUD while more correlated pairs primarily give information about one singular aspect.

Discussion

The analyses presented here, while largely descriptive, serve to concretely illustrate a core limitation of the algorithm used to diagnose and grade AUDs and suggest that this limitation is likely shared by other disorders that use similar algorithms. We replicated the results of past IRT analyses (using the NESARC) and demonstrated that they closely paralleled those using our external severity indicators in order to expose the shared limitation between simple count and IRT methods of ignoring unique configural information. We show that when considering all of the possible ways that the criteria can be combined; there is a surprising amount of heterogeneity in severity *within levels of criteria endorsement*, due to variability in severity across different criteria *combinations*. This directly leads to substantial overlap in empirically defined (i.e., based on external validity measures) severity *across levels of criteria* endorsement. Our exploratory investigation showed that endorsement of specific criterion pairs (e.g. tolerance & interpersonal problems) can be systematically associated with higher severity across endorsement categories. This would not be expected under the current DSM-5 assumptions or even under the assumptions of unidimensional IRT.

Our close replication of previous NESARC IRT analyses using the NESARC (e.g., Casey et al., 2012; Dawson, Saha, & Grant, 2010; Saha, Chou, & Grant, 2006), and parallel regression analyses of the individual criteria using our external measures provides external, "real-world" validation of latent variable methods used to infer criteria severity. It simultaneously substantiates our constructed measures as valid indicators of alcohol use risk and/or severity given the more theory-based approach of IRT analyses. However, importantly, when we plot the various combinations of criteria at each level of criteria endorsement against our severity measures we observe that even the differential weighting

of individual criteria based upon IRT is likely not sufficient (e.g. up to 18% of individuals may still be diagnostic impostors/orphans).

The current system of totaling the number of criteria endorsed does seem to have some merit, as the relationship between number of criteria and severity is monotonic and it accounts for a substantial amount of the variance in severity. However, it still clearly does not account for other substantial sources of variance in severity, at least part of which we argue is systematically related to specific combinations of criteria. This observation clarifies a finding reported by Dawson, Saha, and Grant (2010) where they find, similar to us, that, *on average*, predicted severities using symptom counts, as they relate to external measures, correlate almost at unity with IRT-based severities. Whereas this result could be interpreted as providing evidence that there is relatively little to gain from more sophisticated IRT-based approaches and/or that simple symptom counts are sufficient to adequately characterize AUD severity, we would strongly discourage such interpretations because they ignore the heterogeneity within each endorsement category. This is evidenced by the improvement in prediction when adding the 55 criterion pair interactions to the base model that included the 11 criterion main effects.

Although, the IRT approach used by past researchers has its strengths, in that it is firmly rooted in psychometric theory and that it provides evidence for the existence of a single latent construct of AUD severity, it also has its limitations. This can be seen visually when comparing Figures 1 and 2, particularly Figure 2D. Namely, the IRT models fit up to this point have been strictly additive, such that latent severity scores are calculated as a linear combination of the 11 criterion estimates. This ignores possible multiplicative (i.e., subadditive and superadditive) effects of criteria endorsement. As a result, the variance in the factor scores for different combinations of criteria is inherently limited by the model constraints. This is why we observe the more tightly clustered configurations of criteria at each level of criteria endorsement in Figure 1. In contrast we observe substantially more variability across different combinations within a given level of criteria endorsement in Figure 2D, likely because the data are estimated from random variates (as opposed to fixed values from the IRT analysis) and there is no assumed linear model. Because the data represented in Figure 2A–D represent observed data, part of the variability is due to error in criteria endorsement, the external indicators' measurement, or both. However, by the same token, error in criteria endorsement would also be an issue in IRT analyses and, more importantly, the error variance removed from the latent factor scores in the IRT model could also be systematic across the different criteria.

While IRT approaches have been well developed in educational and other assessment contexts and there are many fruitful applications of IRT to clinical measurement Reise and Waller (2009) list a number of structural and conceptual limitations in applying IRT methods to clinical assessment compared large-scale cognitive testing contexts where there use is most established. First, the assessment of AUD, other substance use disorders, and many other psychiatric disorders (e.g., Balsis, Gleason, Woods, Oltmanns, 2007; Cooper & Balsis, 2009; Dawson, Saha, & Grant, 2010; Langenbucher, Labouvie, Martin, Sanjuan, Bavly, Kirisci, & Chung, 2004; Uebelacker, Strong, Weinstock, & Miller, 2009) is predicated on the administration of a small fixed set of criterion indicators that correlate

highly to very highly with one another. This can directly lead to criterion configurations that are systematically associated with more extreme scores on the latent construct that would otherwise be minimized with the application of many and varied criterion sets that overlapped less in their content. Second, the distribution of criterion endorsements, both singularly and as a sum score, is usually highly positively skewed (unless one is using a select group of higher severity individuals, which is itself a separate problem), which violates the implicit assumption of IRT that the latent distribution is normal. This leads directly into a third problem regarding the metric of the latent severity construct, namely that it is arguably uninterpretable. Given that it seems that the latent IRT scores may have been used to inform the demarcations for AUD severity (see Figure 1), this is grounds for concern because it is unclear what those demarcations mean. This is one reason why in the current investigation we attempted to superimpose manifest variables that had concrete and interpretable metrics onto the latent IRT space. Fourth, a critical assumption of IRT models is local independence (i.e., independence of criterion errors after adjusting for their common, latent cause). To the extent to which criteria share method variance (i.e., a common exogenous cause) or contain overlapping content that is not due to the underlying disorder, this can pose serious problems in justifying the application of IRT models to clinical diagnosis data. Indeed, we believe there is nontrivial lack of local independence in the NESARC AUD assessment.[4]

Our findings highlight the importance of considering criteria configurations beyond symptom counts and IRT-based latent severity estimates but are not particularly informative as to why this should be the case. At least a few possibilities can be considered. First, as just noted, there is some evidence for violations of the assumption of local independence which, by itself, suggests that a simple linear model may not adequately reproduce the latent structure of the criteria set and, therefore, tend to create psychometric pressures towards super or subadditivity among those criteria that tend to violate this assumption. It is often difficult to determine how to “lump” or “split” diagnostic criteria but it is worth noting that the designers of other diagnostic systems (e.g., ICD-10; WHO, 1992) for SUDs while overlapping considerably with DSM-IV and DSM-5, chose to combine “give up activities” and “time spent” into a single criterion. Without judging which is the better option, it is worth noting that the issue is not trivial and leads one to question whether one approach implicitly “double dips” or the other systematically undercounts. Such lumping or splitting of criteria is a general issue in diagnostic research that, from our perspective, has received far too little empirical attention.

Another possibility is that a given criterion, by itself, is highly fallible and prone to being a false positive. However, a second criterion, besides being informative in its own right tends to validate the first symptom. That is, the validity of the first criterion is, probabilistically, either confirmed or undermined, by the endorsement (or lack of it) of the second criterion. In such a case, the combination of endorsing both symptoms would imply a degree of

[4]We tested the assumption of local independence in our IRT analysis of the current data. Of the 55 residual correlations 9 were found to result in substantial increments in model fit ($\chi^2(1) > 10.00$, $p < .002$). Most notable of these was the positive residual correlation of .36 between role interference and interpersonal problems. These two criteria overlap substantially in their content, particularly with regard to social obligations, and present a clear violation of local independence. The other 8 correlations were smaller, but still substantial, ranging from $.13 < |r| < .35$.

nonadditivitiy. Here we are aware of previous research suggesting just that (O'Neill & Sher, 2000). In that longitudinal study of college students, a baseline measure of withdrawal in the absence of tolerance predicted a much more benign course of drinking over seven years than tolerance alone or tolerance plus withdrawal. We speculated that reports of withdrawal in the absence of tolerance might have reflected false positive reports (perhaps misdiagnosed hangover) and suspect that the phenomenon of one criterion validating another might be not limited to this particular situation.

It is our view that each of the potential pitfalls we have identified when using either a simple count or an IRT approach to AUD diagnosis, either categorical or severity graded, can very likely be readily extended to other diagnostic algorithms across personality, mood, and substance use disorders in DSM-5 and ICD-10. If there is heterogeneity in the base rates of endorsement of criteria and in their associations with one another the same propensity for meaningful configural variance seems high. How this might be affected by the size of given disorders' criterion sets, their thresholds for diagnosis, and their degree of hierarchical diagnosis structures (i.e. necessary and/or sufficient criteria) remains unclear. However, to the degree that such configural complexity does exist there are broad implications for both clinical assessment and basic research.

The primary implication is that the categorical diagnosis and severity grading (e.g. in the case of DSM-5 substance use disorders) should be taken as no more than a very crude guide, since it is likely that some individuals deemed to have "mild" levels of disorder might have observed levels of severity that are likely higher than some other individuals deemed to have "moderate" or even "severe" levels of disorder (see Figures 2A–2D). Even in our results where severity was indexed with IRT-derived AUD factor scores (Figure 1), there is nontrivial overlap between adjacent criterion counts with substantial overlap between mild and moderate severities criterion counts and moderate and severe criterion counts. This phenomena has previously been described by Cooper and Balsis (2009) and would appear to be a general property of any severity grading system that assumes (1) equal weighting and (2) a linear model, despite clear differences in base rates of endorsement across criteria. We speculate that if one wants to employ severity grading of diagnoses, an approach that uses an IRT estimate of latent severity will represent an incremental improvement, even if slight (Dawson et al., 2010), over the DSM-5 approach assuming that valid data exist for estimating criteria thresholds. However, such an approach is unlikely to "solve" the problem since some degree of nonadditivity seems likely and because there is still considerable variability in how specific criteria are assessed across diagnostic instruments, thus rendering latent trait estimates (and manifest criterion counts) highly fallible (along with the larger diagnostic enterprise).

To illustrate this problem, consider the newly added craving criterion for AUD in DSM-5. In the NESARC, this criterion is assessed by positively endorsing one of two symptoms: (1) wanting a drink so badly you could not think about anything else (12 month prevalence = .8%), or (2) feeling a very strong desire to drink (12 month prevalence = 4.1%). Without considering whether the lower or higher threshold item is the more valid, an instrument that assesses only the more severe definition of craving will provide a higher implied severity for that criterion and for various combinations of criteria that contain this item. Comparing

epidemiological data from different studies, we see that some studies employ only a high threshold assessment (e.g. those using the Composite International Diagnostic Interview [CIDI; WHO, 1997]; see Cherpitel et al., 2010; Mewton, Slade, McBride, Grove, Teesson, 2011a; or an earlier version of the AUDADIS as used in the National Longitudinal Alcohol Epidemiologic Study; see Keyes et al., 2011) and the IRT analyses of the criteria set from these studies would invariably provide a different rank ordering of criteria severity. In a nutshell, that is the "dirty little secret" of most research on diagnostic criteria. The foundational units of most research is at the criterion level but, depending upon how the criterion is assessed via specific signs and symptoms, and depending upon where the threshold for meeting a criterion is set, the prevalences of various criteria are likely to vary both in their absolute and relative (to other criteria) prevalence and implied severity. This applies not only to craving as a criterion, but arguably any of the 11 criteria where endorsement is determined using multiple items (or similar items varying in form). From our perspective, this is one of the major problems that occurred with the wholesale adoption of an IRT framework originally developed in the context of educational testing into the area of psychiatric diagnosis. This problem is likely to be greatest when using either latent severities (from "internal" IRT estimates) or estimated severities with respect to external criteria (e.g. our drinking, health, and diagnosis indicators) that are derived on the basis of different samples and instruments. That is, while an IRT-based approach is probably preferable to a straight criterion count, it is still based on the assumption of additivity and results are likely conditional on sample and diagnostic instrument. Empirically derived estimates, although not as closely tied to the model constraints of an IRT-approach are also likely to be highly dependent on sample and instrument.

To illustrate the severity of this problem, we found that our rank orderings of criteria severity were highly correlated (*r*s = .99 or 1.00) with those of four previous IRT investigations (Casey et al., 2012; Dawson, Saha, & Grant, 2010; Saha, Chou, & Grant, 2006; Saha, Stinson, & Grant, 2007); but those and our study were all conducted using the same data. If we compare our rank order findings with those from IRT analyses of different samples we find much less agreement. We observe the highest rank-order correlations between our IRT thresholds and those of other investigations when the same or a similar diagnostic instrument is used *and* when the samples are broadly representative of the general population. The above example concerning craving is corroborated, given that correlations are markedly attenuated (though still sometimes large; *r* range = .21 – .86) between our IRT threshold rankings and those of analyses using the Composite International Diagnostic Interview (CIDI; WHO, 1997), where only the high threshold craving item was used. Furthermore, studies with specific participant subpopulations or idiosyncratic instruments tend to produce much lower and even negative rank-order correlations compared to our severity findings.

A likely candidate in explaining some of these discrepancies in relative severity other than what might be caused by differential criterion assessment is that of ascertainment bias. For example, it seems plausible that adolescents and younger adults (e.g., college students) may be less likely to endorse role-related criteria (e.g., role interference) due to having fewer responsibilities than older adults. This would result in role interference becoming a more

"severe" criterion for younger subpopulations. This same argument could also be applied to males versus females, as differences in endorsement may not necessarily reflect sex differences in the severity of certain AUD symptoms, but rather sex differences in the most likely symptom-level manifestations of AUD. Similarly, there are cross-national differences in the enforcement of laws regarding driving under the influence of alcohol (e.g. Australia), which may lead to lower endorsement rates of criteria like hazardous use (e.g. Mewton et al., 2011a; 2011b).

The other major clinical implication is that within the DSM-5, there is considerable implied overlap in severity between individuals who are subthreshold for AUD (i.e., meeting only one criterion) and those meeting the diagnostic threshold (e.g., meeting two criteria; see Figures 2A–2D). We suspect that this would present a similar or perhaps an even bigger issue within disorders that rely completely on a categorical diagnosis and/or have higher criteria thresholds such that there is more sub-threshold theoretical variability. We have previously argued that the 2/11 algorithm for AUD diagnosis is "too lenient" (Martin, Steinley, Verges, & Sher, 2011) but our findings highlight the problem with proposing any simple cut-off based upon criterion counts alone. Indeed, the grading of severity might be particularly important with the move to DSM-5 given that there has been substantial professional debate about the implications of what some believe is an overly liberal threshold for diagnosis. Hasin and colleagues (2013, p. 841) note that "concerns about the threshold should be addressed by indicators of severity, which clearly indicate that cases vary in severity." There is no easy solution to this problem although future research should consider a range of alternative approaches for improving where to place the diagnostic boundary between disorder and lack of disorder. One recently proposed approach is to require both harm (as manifested by consequences) and dysfunction (as indicated by withdrawal or compulsive use; Wakefield & Schmitz, 2014). This may be a promising approach as Wakefield and Schmitz (2014) show that this type of modification results in estimates that are more plausible than those published in the epidemiological literature. We note, however, that the consequences of adopting such an approach are potentially problematic for multiple reasons. These include the fact that harm and dysfunction are highly context and resource dependent, that effects may be delayed (e.g., alcohol-related liver disease and cancers), and because attributing causality of a consequence to substance use (as opposed to generalized externalizing psychopathology) can be devilishly hard and often impossible (Martin, Langenbucher, Chung, & Sher, in press). Consequently, although we believe the call to require symptoms indicative of dysfunction (i.e., the presence of an acquired dependence process) as holding promise, it is not clear that an equally compelling case can be made for harm. However, the case for moving away from treating all criteria as equal seems compelling.

One intriguing possibility is to move away from an approach that defines criteria nominally (and thus, failing to resolve degree of severity at the symptom level), to an approach that grades severity of each symptom. Such approaches are common in clinical medicine where grading the health of a newborn (Apgar, 1953) or level of consciousness (e.g. Teasdale & Jennett, 1974), among many other constructs, are based on the sum of ordered, polytomous ratings across criteria. Such an approach, although not without its own problems, directly addresses the problem of where to place the threshold of endorsement for a criterion by

acknowledging that most clinical criteria of interest are themselves dimensional. We are not arguing that such dimensional ratings should necessarily be predicated on the frequency of a given symptom being experienced, since such frequency-weighted symptom counts show modest improvements in incremental validity (e.g. Dawson & Grant, 2010; Hasin et al., 2012; Hurlbut & Sher, 1992). However, a strong case can be made that each DSM-5 AUD criterion is dimensional and resolving this dimensionality could only serve to enhance the assessment of severity better at the syndrome level. Such an approach is already consistent with some DSM-5 diagnostic algorithms (e.g. major depressive disorder, bipolar disorder, 2013). While the introduction of structured and semi-structured diagnostic interviews has been hailed as a major advance in psychiatric research, differences among instruments have been largely ignored and the differences in IRT findings across studies noted above suggest that there is substantial instrument-related variation in item performance, highlighting problems with severity grading based either on criterion counts or latent-trait measures.

With respect to more basic research on psychological disorders our findings suggest that selecting or differentiating clinical (and comparison) groups on the basis of (either DSM-5 or ICD-10) criterion counts or IRT latent trait scores, or using such counts/scores as a proxy for risk could seriously undermine the investigation of processes of interest. For example, the use of such strategies implicitly assumes a common underlying cause or set of additive causes to a disorder and ignores that different causes may interact depending on their severity. This is exactly the configural information that we find can be important for diagnosis. As such ignoring it may be ignoring information useful for better tracking the etiology of various diagnoses. Similarly, to the extent that specific criteria consistently manifest as important indicators in various configurations, but not necessarily on their own, they may be uncovering unique processes that cut across diagnostic domains and can be used to inform transdiagnostic criteria. Also, as mentioned above, sub-/super-additive clusters could be useful for identifying common mechanisms across a criterion that counts or IRT analysis would not be able to identify, leading to novel and more focused investigations.

Primary limitations of the current work are those associated with epidemiological studies in the general population using lay interviewers and highly structured interviews and the lack of external validating criteria that are not based upon self-report (e.g., behavioral tests, biomarkers). Additionally, although there were two waves of data collection in NESARC, craving was assessed only at the second wave, precluding informative analysis of syndrome persistence as another validity criterion. Although such limitations are common to this type of research, such ubiquity does not undermine their importance. Moreover, the large size and population-based sampling employed in NESARC enable the types of analyses reported here. However, even with the twenty-thousand-plus past-year drinkers in NESARC at Wave 2, a large number of possible symptom configurations were not observed (see Table 2) and we expect that these missing cells are most likely random and not structural zero cell counts. However, we have confidence that the “missing” configurations represent what is likely a vanishingly small percent of the general population.

Broadly, this research indicates that grading severity of diagnosis by criterion counts is more imprecise than is typically recognized, and in some cases may not reliably distinguish

between individuals with varying levels of a given disorder. Moreover, such severity grading is likely to be moderated by the nature of the assessment instrument. While the focus of this paper has been on severity among those above diagnostic threshold, the same implications hold for moving from subthreshold to threshold cases. In terms of basic research on the etiology, course, and correlates of various disorders, and applied research on intervention, poorly resolved severity grading can reduce the ability to identify factors associated with underlying severity and to evaluate treatment and prevention programs that view psychopathology as quantitative phenotypes. From a clinical perspective, the degree of likely heterogeneity of severity among individuals with the same criterion counts and similar levels of severity across individuals with very different criterion counts urge caution for using such grading in clinical decision-making about level of care or healthcare policies based on severity.

A number of alternative approaches can be entertained including use of generalized, including non-linear and non-parametric, applications of latent trait scores (Reiss & Waller, 2009), developing empirical severity scores that employ configural information, and grading the severity of each criterion separately before aggregating them into an overall severity index. To the extent that it is important to resolve syndrome severity, these and other alternative approaches should be seriously considered despite the strong tradition of using simple counts.

The lack of strict additivity of symptoms also provides empirical grounds for understanding the nature of how symptoms could interact and, along with other psychometric techniques, provide a tool for identifying those criteria which alone or in combination with other criteria degrade the diagnostic systems. Thus, we view the types of analyses presented here as more than a technique to reveal problems with the current approach to diagnosis but as a potential tool for understanding diagnostic performance and optimizing criteria sets and algorithms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Matthew R. Lee, Douglas Steinley, and Thomas A. Widiger for insightful and helpful comments on earlier drafts of the manuscript.

This research was supported by NIH/NIAAA grants K05AA017242 and T32AA013526 to Kenneth J. Sher.

References

American Psychiatric Association. DSM 5. American Psychiatric Association; 2013.

Apgar V. A proposal for a new method of evaluation of the newborn. Current Research in Anaesthesia and Analgesia. 1953; 32:260–267.

Balsis S, Gleason ME, Woods CM, Oltmanns TF. An item response theory analysis of DSM-IV personality disorder criteria across younger and older age groups. Psychology and Aging. 2007; 22:171–185. [PubMed: 17385993]

Bollen K, Lennox R. Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin. 1991; 110:305–314.

Burns L, Teesson M. Alcohol use disorders comorbid with anxiety, depression and drug use disorders: Findings from the Australian National Survey of Mental Health and Well Being. Drug and Alcohol Dependence. 2002; 68:299–307. [PubMed: 12393224]

Casey M, Adamson G, Shevlin M, McKinney A. The role of craving in AUDs: Dimensionality and differential functioning in the DSM-5. Drug and Alcohol Dependence. 2012; 125:75–80. [PubMed: 22516145]

Cherpitel CJ, Borges G, Ye Y, Bond J, Cremonte M, Moskalewicz J, Swiatkiewicz G. Performance of a craving criterion in DSM alcohol use disorders. Journal of Studies on Alcohol and Drugs. 2010; 71:674–684. [PubMed: 20731972]

Chung T, Martin CS. Classification and short-term course of *DSM-IV* cannabis, hallucinogen, cocaine, and opioid disorders in clinical adolescents. Journal of Consulting and Clinical Psychology. 2005; 73:995–1004. [PubMed: 16392973]

Compton WM, Dawson DA, Goldstein RB, Grant BF. Crosswalk between DSM-IV dependence and DSM-5 substance use disorders for opioids, cannabis, cocaine and alcohol. Drug and Alcohol Dependence. 2013; 132:387–390. [PubMed: 23642316]

Cooper LD, Balsis S. When less is more: How fewer diagnostic criteria can indicate greater severity. Psychological Assessment. 2009; 21:285–293. [PubMed: 19719341]

Dawson DA, Grant BF. Should symptom frequency be factored into scalar measures of alcohol use disorder severity? Addiction. 2010; 105:1568–1579. [PubMed: 20569231]

Dawson DA, Goldstein RB, Grant BF. Differences in the profiles of DSM-IV and DSM-5 alcohol use disorders: Implications for clinicians. Alocoholism: Clinical and Experimental Research. 2013; 37:E305–E313.

Dawson DA, Grant BF, Stinson FS, Zhou Y. Effectiveness of the derived Alcohol Use Disorders Identification Test (AUDIT-C) in screening for alcohol use disorders and risk drinking in the U.S. general population. Alcoholism: Clinical and Experimental Research. 2005; 29:844–854.

Dawson DA, Saha TD, Grant BF. A multidimensional assessment of the validity and utility of alcohol use disorder severity as determined by item response theory models. Drug and Alcohol Dependence. 2010; 107:31–38. [PubMed: 19782481]

Galatzer-Levy IR, Bryant RA. 636,120 ways to have posttraumatic stress disorder. Perspectives on Psychological Science. 2013; 8:651–662. [PubMed: 26173229]

Gilder DA, Gizer IR, Ehlers CL. Item response theory analysis of binge drinking and its relationship to lifetime alcohol use disorder symptom severity in an American Indian community sample. Alcoholism: Clinical & Experimental Research. 2011; 35:984–995.

Grant BF, Dawson DA, Stinson FS, Chou PS, Kay W, Pickering R. The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): Reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. Drug and Alcohol Dependence. 2003; 71:7–16. [PubMed: 12821201]

Grant, BF.; Kaplan, KD. Source and accuracy statement: Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Rockville, MD: National Institute on Alcohol Abuse and Alcoholism; 2005.

Grant, BF.; Moore, TC.; Kaplan, KD. Source and accuracy statement: Wave 1 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2003.

Grant BF, Stinson FS, Dawson DA, Chou SP, Dufour MC, Compton W, Kaplan K. Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders. Alcohol Research & Health. 2006; 29:107–120.

Grant BF, Stinson FS, Dawson DA, Chou PS, Ruan WJ. Co-occurrence of *DSM–IV* personality disorders in the United States: Results from the National Epidemiologic Study on Alcohol and Related Conditions. Comprehensive Psychiatry. 2005; 46:1–5. [PubMed: 15714187]

Hasin DS, Fenton MC, Beseler C, Park JY, Wall MM. Analyses related to the development of DSM-5 criteria for substance use related disorders: 2. Proposed DSM-5 criteria for alcohol, cannabis, cocaine and heroin disorders in 663 substance abuse patients. Drug and Alcohol Dependence. 2012; 122:28–37. [PubMed: 21963333]

Hasin DS, O'Brien CP, Auriacombe M, Borges G, Buckolz K, Budney A, et al. DSM-5 criteria for substance use disorders: Recommendations and rationale. American Journal of Psychiatry. 2013; 170:834–851. [PubMed: 23903334]

Hurlbut SC, Sher KJ. Assessing alcohol problems in college students. Journal of American College Health. 1992; 41:49–58. [PubMed: 1460173]

Judd, CM.; McClelland, GH.; Ryan, CS. Data analysis: A model comparison approach. Routledge: 2011.

Kahler CW, Strong DR. A Rasch model analysis of *DSM-IV* alcohol abuse and dependence items in the National Epidemiological Survey on Alcohol and Related Conditions. Alcoholism: Clinical and Experimental Research. 2006; 30:1165–1175.

Keyes KM, Krueger RF, Grant BF, Hasin DS. Alcohol craving and the dimensionality of alcohol disorders. Psychological Medicine. 2011; 41:629–640. [PubMed: 20459881]

Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comoridity Survey Replication. Archives of General Psychiatry. 2005; 62:593–602. [PubMed: 15939837]

Krueger RF, Nichol PE, Hicks BM, Markon KE, Patrick CJ, McGue M. Using latent trait modeling to conceptualize an alcohol problems continuum. Psychological Assessment. 2004; 16:107–119. [PubMed: 15222807]

Langenbucher JW, Martin CS, Hasin DS, Helzer JE. Alcohol abuse: Adding content to category. Alcoholism: Clinical and Experimental Research. 1996; 20:270A–275A.

Langenbbucher J, Martin CS, Labouvie E, Sanjuan PM, Bavly L, Pollock NK. Toward the *DSM-5*: The withdrawal-gate model versus the *DSM-IV* in the diagnosis of alcohol abuse and dependence. Journal of counseling and Clinical Psychology. 2000; 68:799–809.

Langenbucher JW, Labouvie E, Martin CS, Sanjuan PM, Bavly L, Kirisci L, Chung T. An application of Item Response Theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. Journal of Abnormal Psychology. 2004; 113:72–80. [PubMed: 14992659]

Linden, WJ.; Hambleton, RK. Handbook of modern item response theory. New York: 1997.

Martin CS, Chung T, Kirisci L, Langenbucher JW. Item response theory analysis of diagnostic criteria for alcohol and cannabis use disorders in adolescents: Implications for *DSM-IV*. Journal of Abnormal Psychology. 2006; 115:807–814. [PubMed: 17100538]

Martin CS, Langenbucher JW, Chung T, Sher KJ. Truth or consequences in the diagnosis of substance use disorders. Addiction. (in press).

Martin CS, Steinley DL, Verges A, Sher KJ. Letter to the Editor: The proposed 2/11 symptom algorithm for DSM-5 substance-use disorders is too lenient. Psychological Medicine. 2011; 41:2008–2010. [PubMed: 21557890]

McCutcheon VV, Agrawal A, Heath AC, Edenberg HJ, Hesselbrock VM, Schuckit MA, Kramer JR, Bucholz KK. Functioning of alcohol use disorder criteria among men and women with arrests for driving under the influence of alcohol. Alcoholism: Clinical & Experimental Research. 2011; 35:1985–1993.

Mewton L, Slade T, McBride O, Grove R, Teesson M. An evaluation of the proposed DSM-5 alcohol use disorder criteria using Australian national data. Addiction. 2011a; 106:941–950. [PubMed: 21205055]

Mewton L, Teesson M, Slade T, Cottler L. Psychometric performance of DSM-IV alcohol use disorders in young adulthood: Evidence from an Australian general population sample. Journal of Studies on Alcohol and Drugs. 2011b; 72:811–822. [PubMed: 21906508]

Muthén, LK.; Muthén, BO. Mplus: Statistical analysis with latent variables: User's guide. Muthén & Muthén; 2010.

National Institute on Alcohol Abuse and Alcoholism. The Physicians' Guide to Helping Patients with Alcohol Problems. Rockville, MD: Department of Health and Human Services; 1995. NIH Publication No. 95-3769

O'Neill S, Sher KJ. Physiological alcohol dependence symptoms in early adulthood: A longitudinal perspective. Experimental and Clinical Psychopharmacology. 2000; 8:493–508. [PubMed: 11127421]

Pollock NK, Martin CS. Diagnostic orphans: Adolescents with alcohol symptoms but no *DSM–IV* diagnosis. American Journal of Psychiatry. 1999; 156:897–901. [PubMed: 10360129]

Regier DA, Farmer ME, Rae DS, Locke BZ, Keith SJ, Judd LL, Goodwin FK. Comoridity of mental disorders with alcohol and other drug use: Results from the Epidemiologic Catchment Area (ECA) study. Journal of the American Medical Association. 1990; 264:2511–2518. [PubMed: 2232018]

Reise SP, Waller NG. Item response theory and clinical measurement. Annual Review of Clinical Psychology. 2009; 5:27–48.

Saha TD, Chou SP, Grant BF. Toward an alcohol use disorder continuum using item response theory: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. Psychological Medicine. 2006; 36:931–941. [PubMed: 16563205]

Saha TD, Stinson FS, Grant BF. The role of alcohol consumption in future classifications of alcohol use disorders. Drug and Alcohol Dependence. 2007; 89:82–92. [PubMed: 17240085]

Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. The Lancet. 1974; 304:81–84.

Uebelacker LA, Strong D, Weinstock LM, Miller IW. Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. Psychological Medicine. 2009; 39:591–601. [PubMed: 18588740]

Wakefield JC, Schmitz MF. How many people have alcohol use disorders?: Using the harmful dysfunction analysis to reconcile prevalence estimates in two community surveys. Frontiers in Psychiatry. 2014; 5:10. [PubMed: 24550847]

Ware, JE.; Kosinski, M.; Turner-Bowker, DM.; Gandek, B. How to score version 2 of the SF-12 health survey (with supplement documenting version 1). QualMetric Incorporated; 2002.

Widiger TA, Trull TJ. Plate tectonics in the classification of personality disorder: shifting to a dimensional model. American Psychologist. 2007; 62:71–83. [PubMed: 17324033]

World Health Organization. Composite International Diagnostic Interview—version 2.0. Geneva: WHO; 1997.

World Health Organization. The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines. Vol. 1. World Health Organization; 1992.

World Health Organization. Global status report on alcohol and health. 2011.

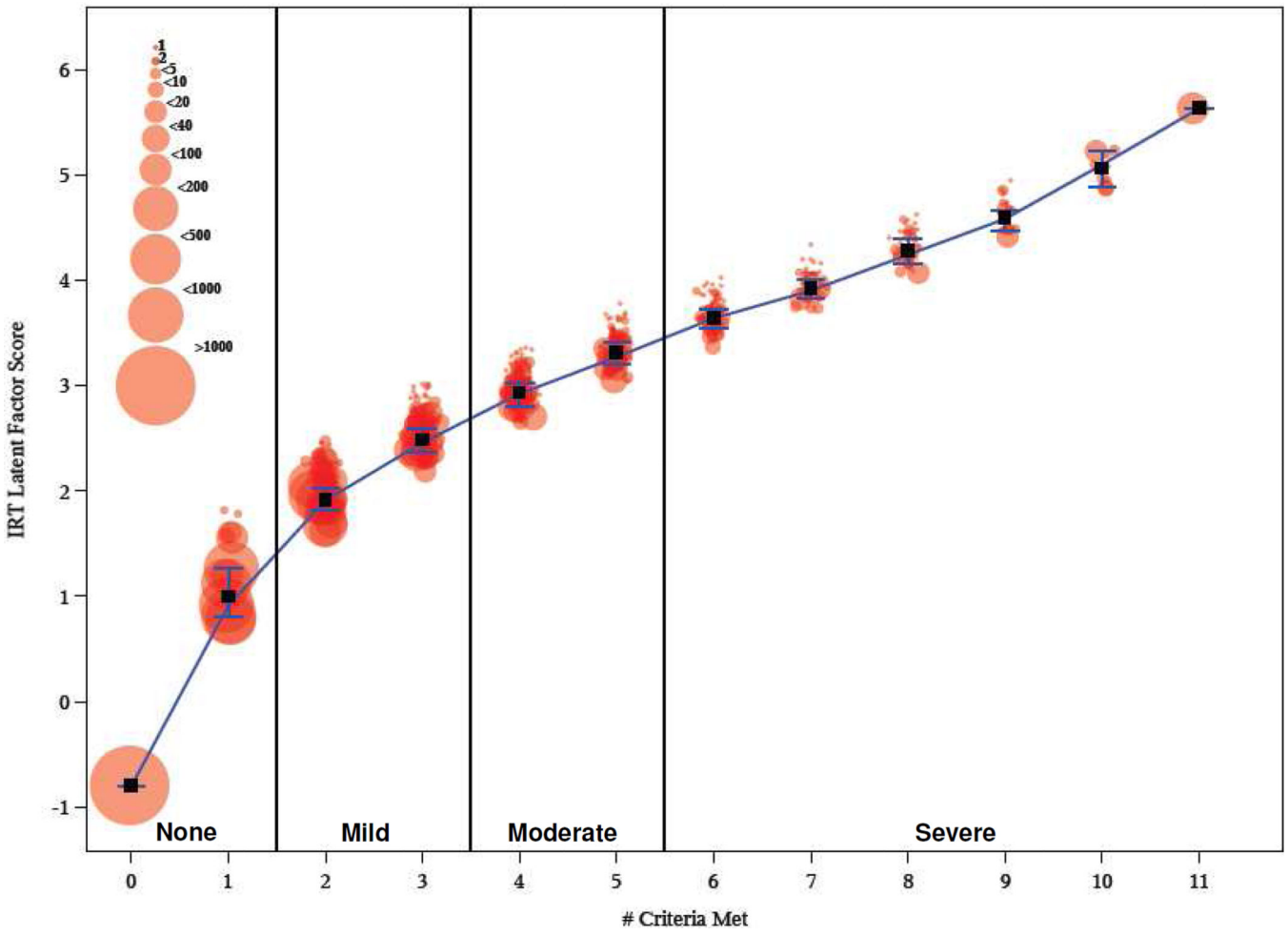

Figure 1.

Plot of the different criteria permutation (from 0 to 11) and corresponding average IRT latent factor severity scores across groups. *Note*: Connected line represents the weighted median value across all combinations of criteria at each level of number of criteria endorsed for the four different severity measures. Vertical lines at each level represent 25th and 75th percentiles. Squares denote the weighted mean at each level across all combinations. Circles represent weighted mean values of each of the severity measures for particular unique combinations of criteria at each level. Circles are jittered for ease of presentation.

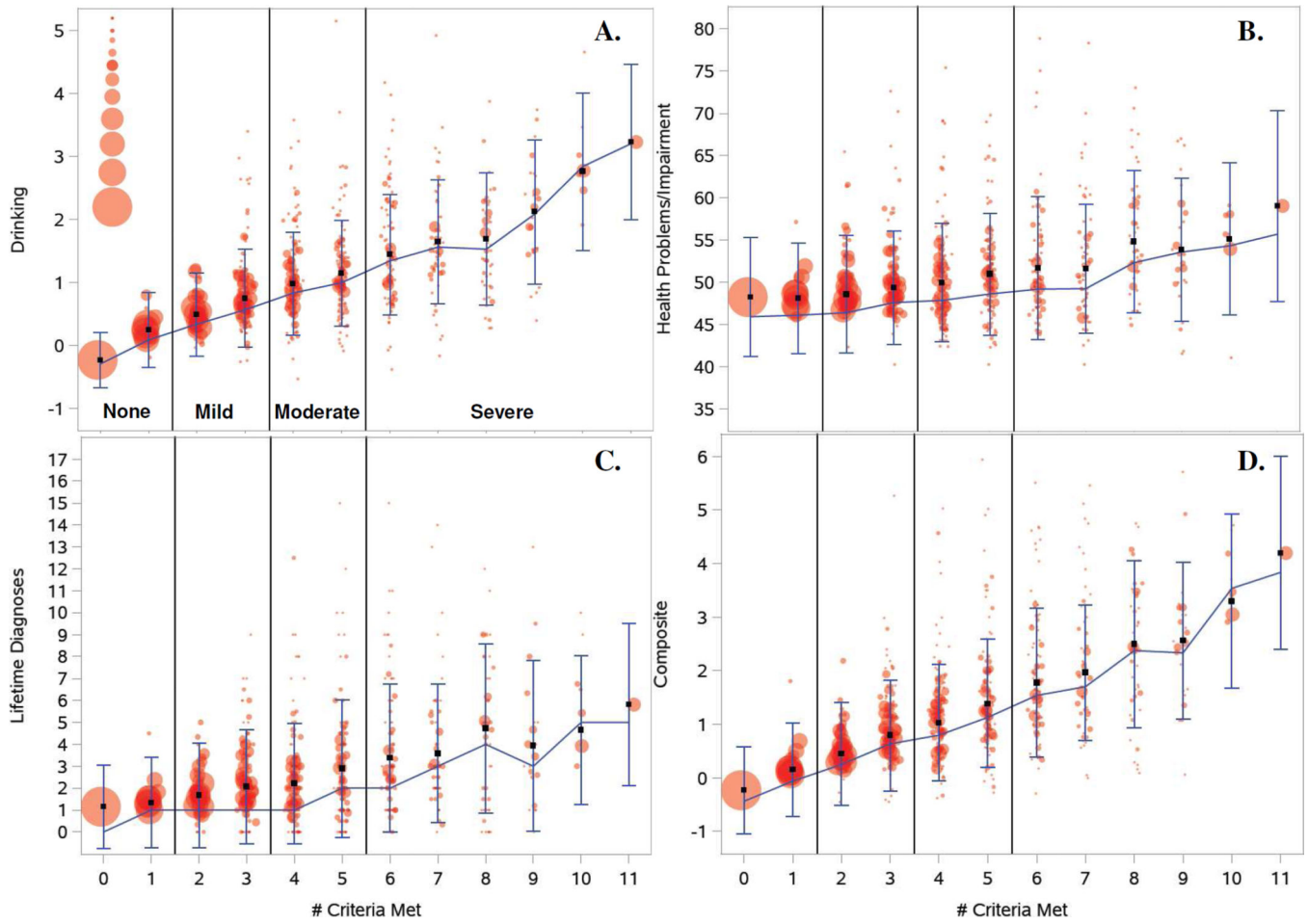

Figure 2.

Plots of the different criteria permutations (from 0 to 11) against (A.) behavioral drinking problems, (B.) overall health problems, (C.) number of total Axis I and Axis II disorders, and (D.) constructed global severity measure. *Note*: Connected line represents the weighted median value across all combinations of criteria at each level of number of criteria endorsed for the four different severity measures. Vertical lines at each level represent 25th and 75th percentiles. Squares denote the weighted mean at each level across all combinations. Circles represent weighted mean values of each of the severity measures for particular unique combinations of criteria at each level. From smallest to largest, circle sizes correspond to 1, 2, ≤5, ≤10, ≤20, ≤40, ≤100, ≤200, ≤500, ≤1000, and >1000 individuals within each unique combination of criteria. Circles are jittered for ease of presentation.

Table 1

Prevalence rates and Item Response Theory estimates for the 11 AUD criteria.

		Threshold		Discrimination	
Criteria	**Prevalence**	**Estimate**	**SE**	**Estimate**	**SE**
Larger/longer	14.5%	1.050	.015	.845	.008
Cut down	12.4%	1.144	.015	.763	.010
Hazardous use	11.1%	1.206	.019	.732	.011
Tolerance	8.2%	1.376	.017	.721	.011
Withdrawal	7.9%	1.402	.017	.815	.009
Drink despite problems	5.3%	1.606	.020	.895	.008
Craving	4.3%	1.712	.021	.832	.010
Time spent	3.0%	1.859	.024	.898	.008
Interpersonal problems	2.5%	1.945	.024	.902	.008
Role interference	1.1%	2.265	.030	.923	.008
Give up activities	1.0%	2.299	.035	.927	.009

Note. Parameter estimates are standardized. All estimates are significant at $p < .001$. SE = standard error.

Table 2

Number of possible and actual combinations of criteria at each level of endorsement.

# Criteria Endorsed	# Individuals	Possible Combinations	Actual Combinations	Proportion Missing[a]	Adjusted Proportion Missing[b]
0	15,789	-	-	-	-
1	2,972	11	11	.00	.00
2	1,465	55	41	.25	.25
3	751	165	82	.50	.50
4	441	330	104	.68	.69
5	286	462	100	.78	.65
6	165	462	79	.83	.52
7	97	330	54	.84	.44
8	84	165	42	.75	.50
9	60	55	25	.55	.53
10	46	11	8	.27	.18
11	21	1	1	.00	.00
Total	22,177	2048	548	.73	.57

[a] Proportion represents absolute proportion of combinations missing ignoring that there may be more combinations than individuals in a given category

[b] Proportion represents the adjusted proportion for when the number of possible combinations in a category exceeded the number of individuals in that category. In this case the number of individuals was used as the denominator under the assumption that if all criteria are equal and additive in severity then all combinations should be equally likely to be observed.