Psychological Medicine (2016), **46**, 1769–1784. © Cambridge University Press 2016
doi:10.1017/S0033291716000404

ORIGINAL ARTICLE

Meta-analysis of DSM alcohol use disorder criteria severities: structural consistency is only 'skin deep'

S. P. Lane*, D. Steinley and K. J. Sher

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

Background. Item response theory (IRT) analyses of alcohol use disorder (AUD) and other psychological disorders are a predominant method for assessing overall and individual criterion severity for psychiatric diagnosis. However, no investigation has established the consistency of the relative criteria severities across different samples.

Method. PubMed/Medline, PsycINFO, Web of Science and ProQuest databases were queried for entries relating to alcohol use and IRT. Study data were extracted using a standardized data entry sheet. Consistency of reported criteria severities across studies was analysed using generalizability theory to estimate generalized intraclass correlations (ICCs).

Results. A total of 451 citations were screened and 34 papers (30 unique samples) included in the research synthesis. The AUD criteria set exhibited low consistency in the ordering of criteria using both traditional [ICC = 0.16, 95% confidence interval (CI) 0.06–0.56] and generalized (ICC = 0.18, 95% CI 0.15–0.21) approaches. These results were partially accounted for by previously studied factors such as age and type of sample (e.g. clinical *v.* community), but the largest source of unreliability was the diagnostic instrument employed.

Conclusions. Despite the robust finding of unidimensional structure of AUDs, inconsistency in the relative severities across studies suggests low replicability, challenging the generalizability of findings from any given study. Explicit modeling of well-studied factors like age and sample type is essential and increases the generalizability of findings. Moreover, while the development of structured diagnostic interviews is considered a landmark contribution toward improving psychiatric research, variability across instruments has not been fully appreciated and is substantial.

Received 18 June 2015; Revised 3 February 2016; Accepted 9 February 2016; First published online 28 March 2016

Key words: Alcohol use disorder, generalizability theory, item response theory, meta-analyses, severity.

Introduction

It is assumed, and often supported, that diagnostic criteria are reliable and valid indicators of underlying disorders (World Health Organization, 1992; American Psychiatric Association, 2013). Indeed, with the recent changes in criteria sets and requirements for individual criterion endorsement associated with the migration from the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV; American Psychiatric Association, 2000) to DSM-5 (American Psychiatric Association, 2013), researchers have reported substantial consistencies in the dimensionality and diagnosis of a given disorder across a wide variety of disorders (Hasin *et al.* 2013; Regier *et al.* 2013; Grant *et al.* 2015), implying the structural validity (Loevinger, 1957) of the criteria sets themselves.

Latent variable methods such as item response theory (IRT; Embretson & Reise, 2000) have become the preferred approach for assessing these assumptions and for determining the relative severity of individual criteria across a wide variety of personality (Balsis *et al.* 2007; Cooper & Balsis, 2009), mood (Uebelacker *et al.* 2009) and substance use disorders (Langenbucher *et al.* 2004). Severity, from an IRT perspective, corresponds to the difficulty of endorsing a given criterion, and is directly related to its base rate of endorsement (i.e. threshold). Estimating individual criteria severities is critical because they identify the specificity of particular criteria as indicators of the underlying disorder. A number of researchers have argued both theoretically (e.g. Martin *et al.* 2008, 2011, 2014) and empirically (e.g. Cooper & Balsis, 2009; Casey *et al.* 2012; Hagman & Cohn, 2013; Lane & Sher, 2015) that disorder criteria, including alcohol use disorder (AUD), fall along a continuum of severity, with endorsement of different criteria being indicative of varying levels of disorder severity. In this way, the 'severity' from IRT, positively, though imperfectly, relates to external, real-world measures of severity (e.g. hospitalization, long-term health, persistence of disorder, comorbidity; Lane & Sher, 2015). The inclusion of criteria with varying levels of severity (i.e. difficulty) is

* Address for correspondence: S. P. Lane, Ph.D., Department of Psychological Sciences, Psychology Building, 200 South Seventh Street, Columbia, MO 65211, USA.
(Email: lanesp@missouri.edu)

considered a hallmark of optimal test development/performance (Embretson & Reise, 2000; Reise & Waller, 2009) because it ensures broad coverage of the underlying latent trait continuum and can be used to assess the trait more adequately across an entire population. Including criteria with a range of observed severities also allows for systematic investigations of particular symptoms of disorder that may be differentially diagnostic for particular groups of individuals, facilitating targeted interventions that isolate problematic cognitions and behaviors. However, others (Dawson *et al.* 2010) have suggested that individual AUD criteria severities may not confer much additional precision compared with criteria counts in determining overall disorder severity. However, we note that these suggestions are based upon analyses of a single study.

To date there has been no systematic investigation of the consistency (i.e. reliability) of the relative severities of criteria within a given disorder across studies. That is, we do not know the degree to which a criterion that is identified as severe compared with other criteria consistently surfaces as severe across repeated investigations with different study characteristics. If consistency is high then there would be confidence in the generalizability of findings for individual criteria across investigations.

However, to the extent to which consistency is low and is not accounted for by systematic differences between study characteristics, between-study variability in estimated criteria severities may be random. Such a situation would have profound implications for epidemiological and clinical studies of psychiatric disorders employing polythetic criteria. First, it indicates that even if two studies using standardized instruments found comparable prevalence rates of, say, AUD, this ostensible consistency may be illusory in that the symptom profiles of those diagnosing could be quite different despite being drawn from the same population. In such a case it would not be reasonable to conclude replicability, except in a superficial sense. Second, it also suggests that we should expect little gain in weighting criteria by their individual severities over simple criteria counts (Dawson *et al.* 2010) and that a given criteria set is not equipped to distinguish individuals across the continuum of the population distribution. Furthermore, it would suggest that previous investigations that have found differences in criteria severities as a function of age, gender, race, socio-economic status (SES), diagnostic time-frame and clinical diagnosis (see Table 1) may not be generalizable.

In contrast, if there is variability in criteria severities that is explained by particular aspects of an individual study (age, gender, etc.) then (1) the generalizability of the severities of the overall criteria set may be supported and robust to external factors, if consistency is high, and (2) the influence of those factors may be considered robust and generalizable and provide strong grounds for targeting particular criteria in different groups of individuals. With respect to this latter point, substantial prior research has argued that endorsement of individual criteria within AUD should and do differ systematically as a function of different factors. For example, in the progression from adolescence to young adulthood and extending into later adulthood, symptoms associated with physiological dependence are initially considerably more difficult to endorse (Martin *et al.* 2006, 2008) while those associated with psychosocial consequences are relatively easier to endorse (Martin *et al.* 1995). This is consistent with developmental psychopathology perspectives in which behavioral problems associated with impaired control are characteristic of the adolescent life stage and part of a more general externalizing spectrum (e.g. Martin *et al.* 2014).

In comparison, more durable neuroadaptations to a chronic drinking pattern associated with addiction (e.g. withdrawal, craving) are expected to manifest later in life (Langenbucher & Chung, 1995), although much research on tolerance shows high rates in adolescents and young adults, possibly due to both developmental factors (Silveri & Spear, 2001) and problems assessing this construct via self-report (e.g. O'Neill & Sher, 2000). Additionally, drinking-related social and health problems are likely to be more severe in adulthood owing to the various role occupancies (e.g. wage earner, spouse/partner, parent) that carry greater responsibilities, as well as the fact that the effects of alcohol exposure on certain types of organ damage (e.g. brain, liver) are cumulative and dose dependent (e.g. Mezey *et al.* 1988; Smith & Riechelmann, 2004) and that aging itself may represent a vulnerability to alcohol-related toxicity (e.g. Oscar-Berman, 2000). We note, however, that the developing brain may be especially sensitive to some kinds of alcohol-related insult (e.g. Jacobus & Tapert, 2013).

Similar examples can be observed with respect to gender, in which women are less likely to endorse criteria related to quantity and frequency of alcohol consumption such as tolerance and withdrawal (Harford *et al.* 2009; Srisurapanont *et al.* 2012), presumably due to differences in total body water and gastric alcohol metabolism (Baraona *et al.* 2001). Also, cultural differences in the consumption and availability of alcohol have been suggested to be differentially indicative of underlying disordered use (e.g. Borges *et al.* 2010; Srisurapanont *et al.* 2012). These previous findings lead to the hypothesis that such factors will account for substantial variability in individual criteria severities across studies and that ignoring them may undermine global reliability estimates.

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

Table 1. *Descriptive characteristics and median Spearman correlations between IRT criteria thresholds from published investigations*

Article	Sample	Instrument	Sample size	Time-frame	No. of criteria	Median ρ (range)[a]
Casey *et al.* (2012)	NESARC wave 2	AUDADIS-IV	22 177[b]	Current	11	0.30 (−0.60 to 0.93)
Dawson *et al.* (2010)[c]	NESARC wave 1	AUDADIS-IV	26 946[b]	Current	11	0.38 (−0.39 to 1.00)
Saha *et al.* (2006)[c]	NESARC wave 1	AUDADIS-IV	22 526[d]	Current	10	0.33 (−0.53 to 1.00)
Saha *et al.* (2007)[c]	NESARC wave 1	AUDADIS-IV	20 846[e]	Current	11	0.42 (−0.30 to 0.99)
Shmulewitz *et al.* (2010)[c]	Israeli households	AUDADIS-IV	1066	Current	11	0.43 (−0.52 to 0.87)
			1160	Lifetime	11	0.35 (−0.11 to 0.85)
Keyes *et al.* (2011)	NLAES	AUDADIS-IV	18 352[e]	Current	12	0.36 (−0.56 to 0.85)
Preuss *et al.* (2014)	WHO/ISBRA	AUDADIS-based	711	Lifetime	11	0.26 (−0.65 to 0.86)
	Australia		104			0.18 (−0.43 to 0.89)
	Brazil		212			0.51 (−0.60 to 0.96)
	Canada		227			0.15 (−0.68 to 0.88)
	Finland		86			0.21 (−0.53 to 0.89)
	Japan		82			0.38 (−0.57 to 0.86)
Mewton *et al.* (2011*a*); Proudfoot *et al.* (2006)	NSMHWB	CIDI V2.0	7746	Current	11	0.31 (−0.52 to 0.93)
Mewton *et al.* (2011*b*)	NSMHWB	CIDI V2.0	853[f]	Current	11	0.44 (−0.37 to 0.89)
McCutcheon *et al.* (2011)	COGA	SSAGA	8605	Lifetime	9	0.40 (−0.52 to 1.00)
	Non-DUI men		3056			0.41 (−0.52 to 1.00)
	Non-DUI women		3894			0.47 (−0.52 to 1.00)
	DUI men		1330			0.38 (−0.47 to 1.00)
	DUI women		325			0.38 (−0.47 to 1.00)
Beseler *et al.* (2010)	College students	Survey-specific	353	Current	10[g]	0.37 (−0.21 to 0.74)
Hagman & Cohn (2011)	College students	CIDI-SAM	396	Current	11	0.45 (−0.38 to 0.79)
Ehlke *et al.* (2012)	NSDUH 2009	SAMHSA	4605[h]	Current	11	0.06 (−0.54 to 0.97)
Kuerbis *et al.* (2013*b*)	NSDUH 2009	SAMHSA	3412[i]	Current	11	0.16 (−0.48 to 0.90)
Hagman & Cohn (2013)	NSDUH 2009	SAMHSA	3806[j]	Current	11[k]	−0.31 (−0.68 to 0.98)
Rose *et al.* (2012)	NSDUH 2002–2008	SAMHSA	9356[l]	Current	11	−0.14 (−0.68 to 0.48)
Harford *et al.* (2009)	NSDUH 2002–2005	SAMHSA	133 231	Current	11	−0.05 (−0.65 to 0.94)
	Men, age 12–17 years		11 651			−0.18 (−0.56 to 0.98)
	Men, age 18–25 years		27 377			−0.09 (−0.64 to 0.96)
	Men, age 26+ years		25 872			0.04 (−0.64 to 0.99)
	Women, age 12–17 years		12 304			−0.08 (−0.51 to 0.97)
	Women, age 18–25 years		29 331			0.04 (−0.62 to 0.99)
	Women, age 26+ years		26 696			0.30 (−0.52 to 0.90)
Srisurapanont *et al.* (2012)	Thai-NMH survey	MINI-Thai	3718	Current	7	0.54 (−0.54 to 0.93)
	Men		3174			0.25 (−0.45 to 0.52)
	Women		544			0.57 (−0.54 to 0.96)
	Adolescents		272			0.29 (−0.49 to 0.71)
	Adults		3446			0.54 (−0.43 to 0.96)
Duncan *et al.* (2011)	MOAFTS	SSAGA	2835	Lifetime	11	0.19 (−0.72 to 0.71)
	Women, age 18–20 years		1158			0.23 (−0.75 to 0.99)
	Women, age 21–25 years		1677			0.19 (−0.72 to 0.99)
Derringer *et al.* (2013)	MTFS and SAGE	SSAGA	6597	Lifetime	7	0.54 (−0.57 to 0.93)
Gilder *et al.* (2011)[m]	American Indians	SSAGA	530	Lifetime	10	0.28 (−0.29 to 0.86)
Gelhorn *et al.* (2008)	Mixed adolescents[n]	CIDI-SAM	5587	Lifetime	11	0.33 (−0.18 to 0.85)
Bond *et al.* (2012); Borges *et al.* (2010, 2011); Cherpitel *et al.* (2010)	ED patients	CIDI V1.0	3191	Current	12	0.09 (−0.58 to 0.70)
	Argentina		662			0.16 (−0.45 to 0.80)
	Mexico		547			0.17 (−0.49 to 0.80)
	Poland		1098			−0.01 (−0.75 to 0.73)
	USA		884			0.18 (−0.42 to 0.73)
Hasin *et al.* (2012)[c]	Clinical	PRISM	543	Current	11	0.26 (−0.66 to 0.89)
Langenbucher *et al.* (2004)	Clinical	CIDI-SAM	372	Lifetime	9	0.31 (−0.41 to 0.76)

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

Table 1 (*cont.*)

Article	Sample	Instrument	Sample size	Time-frame	No. of criteria	Median ρ (range)[a]
Wu *et al.* (2009)	Clinical	DSM-IV checklist	462	Current	7	0.32 (−0.61 to 0.82)
Wu *et al.* (2012)	Clinical	DSM-IV checklist	671	Current	7	0.50 (−0.68 to 1.00)
Martin *et al.* (2006)	Clinical adolescents	SCID	464	Lifetime	11	0.26 (−0.41 to 0.85)
Edwards *et al.* (2013)	VATSPSUD	SCID	7454	Lifetime	11	0.49 (−0.14 to 0.68)
Kuerbis *et al.* (2013*a*)	SARD	SCID	461	Lifetime	11	0.40 (−0.48 to 0.93)

IRT, Item response theory; NESARC, National Epidemiological Study on Alcohol and Related Conditions; AUDADIS-IV, Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV; NLAES, National Longitudinal Alcohol Epidemiologic Study; WHO/ISBRA, World Health Organization/International Society on Biomedical Research Collaborative Study; NSMHWB, National Survey of Mental Health and Well-Being (Australia); CIDI, Composite International Diagnostic Interview; COGA, Collaborative Study on the Genetics of Alcoholism; SSAGA, Semi-Structured Assessment for the Genetics of Alcoholism; DUI, driving under the influence; SAM, Substance Abuse Module; NSDUH, National Survey of Drug Use and Health; SAMHSA, Substance Abuse and Mental Health Services Administration; Thai-NMH Survey, Thai National Mental Health Survey; MINI-Thai, Mini International Neuropsychiatric Inventory, Thai module; MOAFTS, Missouri Adolescent Female Twin Study; MTFS, Minnesota Twin Family Study; SAGE, Study of Addiction: Genes and Environment; ED, emergency department, PRISM, Psychiatric Research Interview for Substance and Mental Disorders; DSM, Diagnostic and Statistical Manual of Mental Disorders; SCID, Structured Clinical Interview for DSM-IV; VATSPSUD, Virginia Adult Twin Study of Psychiatric and Substance Use Disorders; SARD, Substance Abuse Research Demonstration.

[a] Spearman rank-order correlation. Values represent median correlations between reported threshold estimates. Values in parentheses represent the range of correlations across samples. Note that estimates are probably positively biased due to imposed constraints on severity parameters in articles where multiple subsamples were analysed and differential item functioning assessed.

[b] Past-year drinkers.

[c] We could not confirm the reported metric for the IRT parameters, but based on the description and software used an IRT parameterization seemed likely.

[d] $\geqslant$12 Drinks in the past year and ever drank 5+ drinks on $\geqslant$1 occasion.

[e] $\geqslant$12 Drinks in the past year.

[f] Young adult (18–24 years) subsample only.

[g] Authors created a combined measure of interpersonal and legal problems criteria.

[h] College students.

[i] Age 50+ years.

[j] Non-college, age 18–25 years.

[k] Tolerance severity not reported.

[l] Adolescent and young adult drinkers (12–21 years) only.

[m] We selected the authors' 'once per month' binge drinking criteria for comparison of the IRT thresholds. Using the other criteria resulted in trivially different associations.

[n] Combination of community, adjudicated and clinical individuals.

Although receiving little attention, another factor that may account for variability in criteria severities across studies is the diagnostic instrument employed to assess AUD criteria. Historically, prior to the advent of structured (and semi-structured) diagnostic interviews, psychiatric diagnosis was plagued with unreliability across clinicians, cultures and age groups, amongst other factors (Cooper *et al.* 1972; Sartorius *et al.* 1974; Aboraya *et al.* 2006). Since then, the adoption of structured interviews, also known as the 'operational revolution', has been credited with dramatic improvements in the internal and re-test reliabilities of diagnostic interviews (i.e. within-test reliability). However, to the extent that there is limited generalizability across different interviews and samples (i.e. between-test reliability), it is difficult to compare results from different studies. This is especially critical given recent initiatives concerning the reproducibility of research findings across science as a whole (Nosek & Lakens, 2013; Collins & Tabak, 2014; Makel & Plucker, 2014). While individual studies are capable of identifying factors such as gender, age and ethnicity that could have an impact on criterion severities, because studies rarely employ more than one diagnostic instrument at a time, meta-analysis across studies is needed to evaluate the contribution of diagnostic instrument to variability in criterion severity.

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

In the current investigation, we focus on the consistency of IRT-estimated criteria severities across studies of DSM AUD. We chose AUD because there is a well-developed literature applying IRT models to AUD diagnostic criteria across a variety of samples using different measurement instruments, whereas similar studies are considerably less common for other substance use disorders (Hasin *et al.* 2013) and personality and mood disorders. Recent research suggests that the consistency of AUD criteria severities across studies may be questionable due to factors as banal as survey content (Lane & Sher, 2015). The purpose of this meta-analysis is to synthesize the findings of the relative severities of AUD criteria, to gauge the extent to which the influential literature on IRT studies of AUD is generalizable, and to identify factors that lead to inconsistencies in which criteria are estimated as more/less severe.

The specific research question we address leads to a different approach to meta-analysis than is typically employed because the current interest is not individual effect estimates and their variability across studies, but rather the consistency in the relative ordering of criterion severity estimates across studies. It is appropriate to characterize consistency across studies using an intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979). However, since we are interested in factors that contribute to (in)consistency beyond the individual criteria themselves, we estimate generalized ICCs using generalizability theory (GT; Cronbach *et al.* 1972; Brennan, 2001).

Method

Data sources and study selection

The PubMed/MEDLINE, Web of Science, ProQuest (including dissertation abstracts) and PsycINFO electronic databases were searched from 1 January 1977 up to 1 May 2015 using the search criteria 'item response theory' or 'differential item function', and 'alcohol use disorder', 'alcohol abuse' or 'alcohol dependence' (including variations of each phrase). The year 1977 was the year that the ninth revision of the International Classification of Diseases (ICD; World Health Organization, 1977, 1978) was released and subsumed the period in which the DSM and ICD began to assess AUDs using specific criteria sets. A total of 451 citations were identified by the search criteria, 314 of which were unique after duplicate records were removed. The abstracts for each of the 314 articles were screened for a focus on AUD and the use of IRT methodology. The IRT approach is often considered superior to a simple sum/average of indicator variables because it allows for the differential weighting of the indicators in the overall trait score. The individual weights estimated by an IRT analysis are known as the discriminations (i.e. slopes) and the estimated base rates of endorsement are known as the severities (i.e. thresholds). In the current investigation we focus specifically on severity parameters (see online Supplementary material for a brief discussion on criteria discriminations).

To be included in the meta-analysis, a paper needed to report discrimination and severity parameter estimates from an IRT analysis that assessed DSM-III, DSM-IIIR, DSM-IV, DSM-5, ICD-9 or ICD-10 AUD criteria (World Health Organization, 1977, 1978, 1992; American Psychiatric Association, 1980, 1987, 2000, 2013). We limited the search to two-parameter models (2PL; see online Supplementary material). See Fig. 1 for a flow diagram of the search and inclusion process. We identified a total of 34 published papers (30 unique samples) that performed IRTs on 49 different subsamples (see Table 1 and online Supplementary material). For clarity we refer to the 49 different IRT analyses conducted on a single sample within a given article or on multiple subsamples within the same article as individual 'studies' as they are the primary unit of measurement. Though our date range and consideration of various diagnostic systems were very inclusive, we note that the earliest study included in the meta-analysis was published in 2004 (Langenbucher *et al.* 2004), and all included studies assessed either DSM-IV or DSM-5 criteria sets, even if using instruments designed for ICD-9/10 criteria (e.g. Cherpitel *et al.* 2010).

Data extraction

The main outcome measures were the estimated severities for each criterion. The following additional information was extracted from all of the articles: authors, year of publication, sample characteristics (i.e. age and gender composition, clinical *v.* general population; we did not include SES, race or education in the analyses because there were very few studies that either reported sufficient information to be coded or split results by these groups), sample size, diagnostic instrument, diagnosis time-frame, number of criteria assessed, and reporting metric (unstandardized, standardized, IRT parameterized; see online Supplementary material). Two independent raters systematically parsed each article and coded the aforementioned variables. Agreement for the coding of criteria severities and discriminations was nearly perfect (ICCs ranged from 0.98 to 1.00). Agreement for the sample descriptive information was very good to excellent (the range for κ was 0.86 to 1.00; see online Supplementary material for additional details). Age was categorized into five

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

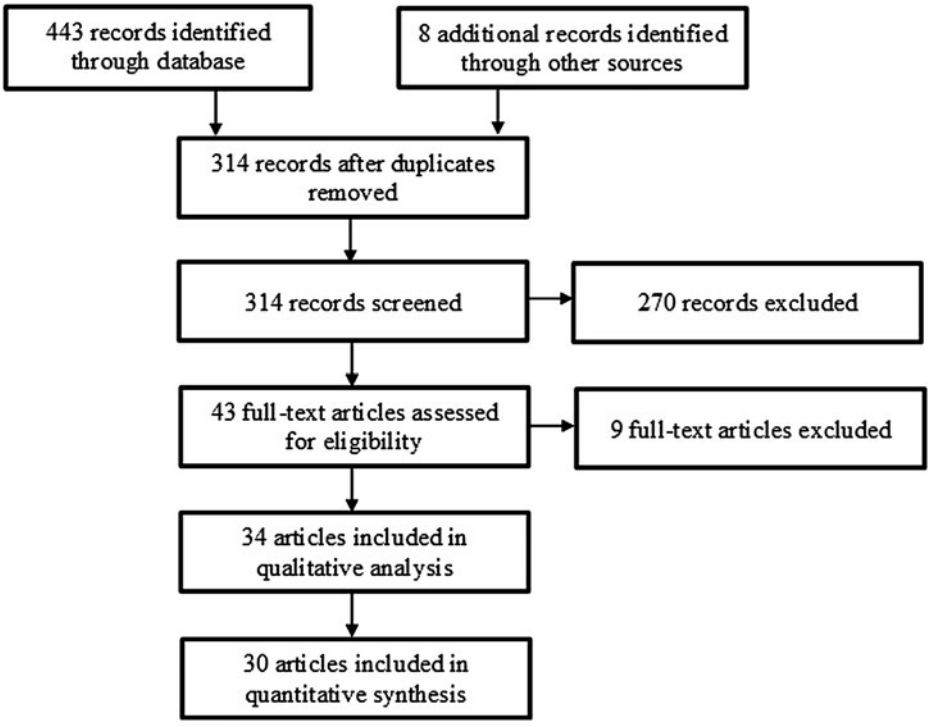

Fig. 1. Flow diagram of identification and selection of studies.

groups (<18 years, primarily between 18–30 years, primarily between 30–50 years, >50 years, representative of the population 18 years or older), gender into five groups (exclusively men, primarily men, approximately equal men and women, primarily women, exclusively women), population into three groups (clinical, general population, a combination of clinical and general population), instrument into seven groups [Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS), Composite International Diagnostic Interview (CIDI), Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA), Substance Abuse and Mental Health Services Administration (SAMHSA), Psychiatric Research Interview for Substance and Mental Disorders (PRISM), Structured Clinical Interview for DSM-III-R (SCID), Other; Spitzer & Williams, 1985; Bucholz *et al.* 1994; World Health Organization, 1997; Grant *et al.* 2003; Hasin *et al.* 2006; Substance Abuse and Mental Health Services Administration, 2006], and time-frame into two groups (current, lifetime). When there was disagreement the two coders and first author (S.P.L.) jointly reviewed the article to resolve inconsistencies. Table 1 lists the relevant information for each of the included publications.

Data analysis

We calculated a traditional consistency-based ICC for any single randomly chosen study [i.e. ICC(3,1)] using a two-way mixed design in which criteria (treated as a fixed factor) was crossed with each individual study (treated as a random factor; Shrout & Fleiss, 1979). However, this approach is limited in that it cannot accommodate unbalanced designs (i.e. not all studies assessed the same criteria) and resulting missing data. We therefore opted for a GT approach. Given that different investigations used different samples of individuals where underlying severity is expected to be different (e.g. representative *v.* clinical populations) we performed all analyses on the raw severity estimates as well as on severity estimates that were standardized within study in order to eliminate systematic variance due to mean differences in criteria severities across studies. Doing so yielded the same pattern of results (see online Supplementary material). The basic GT model is the same analysis of variance model used to generate traditional ICCs (equation 1),

$$P_{cs} = \mu + C_c + S_s + e_{cs} \quad (1)$$

Here, P_{cs} is the severity parameter estimate for criterion, c, from study, s; μ is the grand mean for all severity parameter estimates. C_c is the tendency for a criterion to generally be more or less severe across samples, and S_s is the tendency for a study s to produce higher or lower severities across criteria. Variance components are estimated for this model using a multilevel model in which random effects are estimated for criterion (C_c) and study (S_s) using restricted maximum

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

likelihood estimation. The analog to the traditional ICC (3,1) where the interest is in the reliability of the estimated severity for a fixed criterion set for any randomly selected investigation is then (Cranford *et al.* 2006):

$$R_{1F} = \frac{\sigma^2_{CRITERION}}{\sigma^2_{CRITERION} + \sigma^2_{ERROR}} \tag{2}$$

We then constructed an expanded GT analysis that estimated additional variance components for diagnostic instrument, diagnosis time-frame, sample type, gender and age, as well as their interactions with criterion (equation 3),

$$\begin{aligned} P_{csitnmag} = {} & \mu + C_c + S_s + I_i + T_t + N_n + M_m + A_a + G_g \\ & + (CI)_{ci} + (CT)_{ct} + (CN)_{cn} + (CM)_{cm} + (CA)_{ca} \\ & + e_{csitnmag} \end{aligned} \tag{3}$$

In this model $P_{csitnmag}$ is the severity parameter estimate for criterion, *c*, from study, *s*, where study was indexed by using a specific instrument (*i*), measuring AUD diagnosis on a current or lifetime time-frame (*t*), assessing clinical or non-clinical individuals (*n*), containing primarily male/female (*m*) and younger/older (*a*) participants, and being part of a group of studies that used the same or a partially overlapping sample (*g*). μ, C_c and S_s are as above but now we include effects for instrument (I_i), diagnosis time-frame (T_t), sample population (N_n), gender composition (M_m), age group (A_a) and being part of a group of studies that used overlapping samples (G_g). Importantly, we include two-way interaction terms between criterion and instrument, time-frame, sample population, gender and age as these may be systematic sources of variability across investigations. These interactions are analogous to moderators in traditional meta-analysis. The corresponding ICC(3,1), given that we are interested in the consistency of criterion severities for a single randomly selected study that is not due to instrument, time-frame, population, gender or age, is:

$$R_{1R} = \frac{\sigma^2_{CRITERION}}{\left(\begin{array}{l} \sigma^2_{CRITERION} + \sigma^2_{CRITERION*INSTRUMENT} \\ + \sigma^2_{CRITERION*TIMEFRAME} + \sigma^2_{CRITERION*CLINICAL} \\ + \sigma^2_{CRITERION*GENDER} + \sigma^2_{CRITERION*AGE} + \sigma^2_{ERROR} \end{array}\right)} \tag{4}$$

The five interaction effects with criterion are included as sources of variance in the denominator because, while in the classic case these effects are assumed to be zero and equation 4 degenerates to equation 2, there may be genuine variation in those components that should be considered random if the interest is purely in absolute ordering of criteria across investigations.

Results

Fig. 2*a* depicts the raw severity estimates plotted for each IRT analysis with criteria ordered by their median ranking across investigations. Fig. 2*b* shows the same data but with criteria severities standardized within study to more clearly illustrate the reliability across studies and the different sources of systematic unreliability. The plots reveal considerable variability in severities across the different samples, even when standardized; but there is a systematic linear increase (especially in the standardized case) indicative of some degree of reliability. However, as highlighted by the different line patterns, indicating instrument type, line shading, indicating measurement time-frame, and markers, indicating sample composition, visually there appears to be systematic differences due to instrument, time-frame and sample composition.

Table 1 presents the median Spearman rank-order correlation and the range between each study and all of the others (see online Supplementary material for full bivariate table). Table 2 shows the estimated variance components for the basic (model 1) and expanded (model 2) models. First, we note that the ICC estimate for the reliability of criteria severities for any randomly selected study using a two-way mixed model is 0.16, with an associated 95% confidence interval (CI) of 0.06–0.56 (Shrout & Fleiss, 1979). The parallel estimate and corresponding 95% bootstrapped CI from the multilevel model (model 1), which can accommodate all available data, was 0.27 (95% CI 0.24–0.29). However, when we fit the expanded model, the estimated ICC is reduced to 0.18 (95% CI 0.15–0.21). This is due in part to systematic variance in criteria severities that is associated with particular instruments ($\sigma^2 = 0.09$, S.E. = 0.02, $p < 0.001$), the age range of the participants ($\sigma^2 = 0.07$, S.E. = 0.02, $p < 0.001$) and measurement of AUD in clinical, population-based or mixed samples [$\sigma^2 = 0.02$, S.E. = 0.01, $p = 0.068$; see Higgins *et al.* (2003) for interpreting significance of random effects in meta-analysis (suggested cut-off $p < 0.10$)]. Differences in criteria severities due to diagnosis time-frame and gender composition were not observed (p's > 0.215). The online Supplementary material contains additional analyses in which variance components are estimated on standardized data and where criteria severities are weighted by the relative size of a sample across included studies. Overall, standardizing (*z*) and weighting (*w*) increase the estimates of consistency, but estimates are still quite low ($ICC_w = 0.28$; $ICC_z = 0.20$; $ICC_{wz} = 0.26$), and systematic instrument, diagnostic sample and age effects are still observed.

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

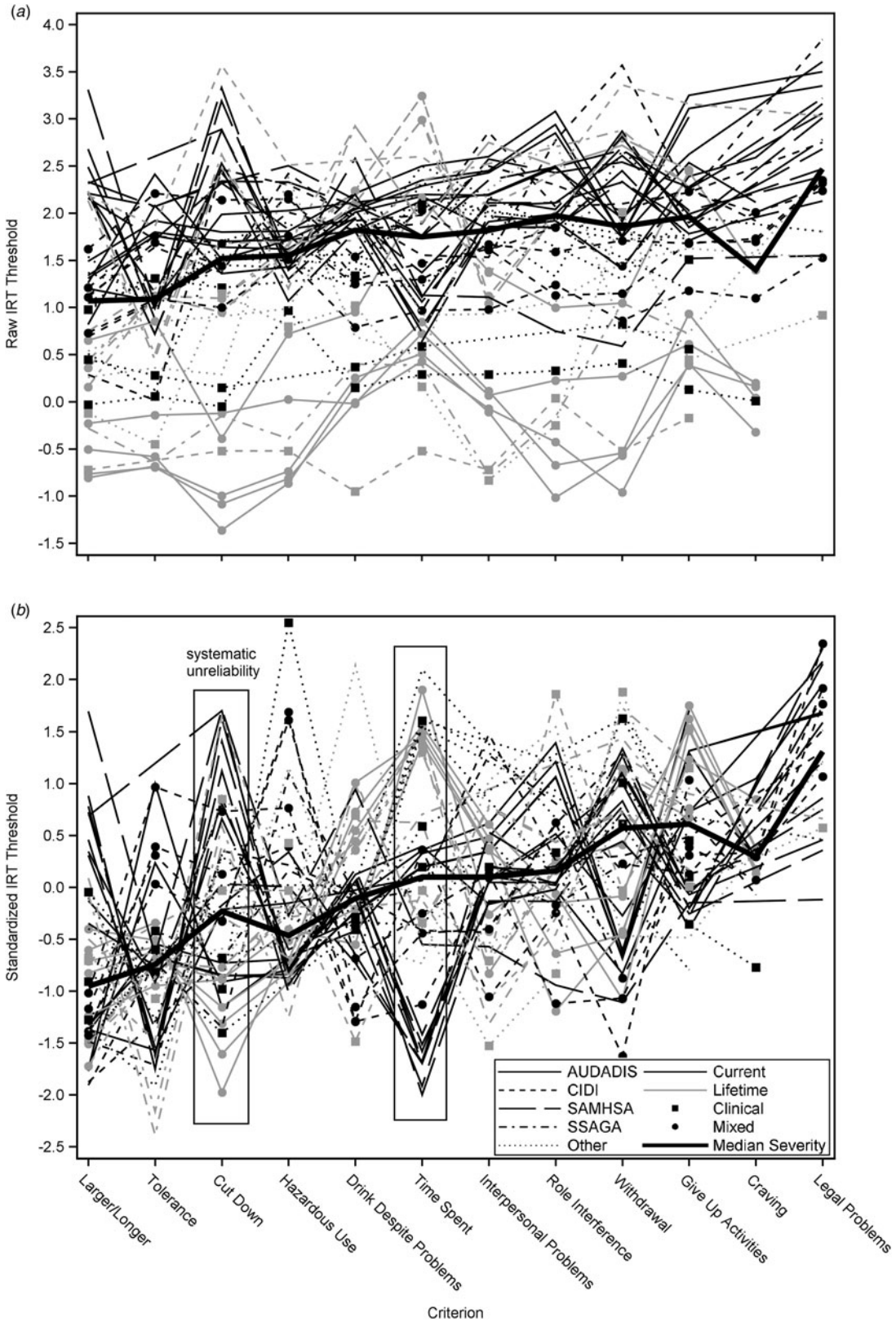

Fig. 2. Raw (*a*) and standardized (*b*) thresholds for Diagnostic and Statistical Manual of Mental Disorders (DSM) alcohol use disorder criteria for the 49 studies. IRT, Item response theory; AUDADIS, Alcohol Use Disorder and Associated Disabilities Interview Schedule; CIDI, Composite International Diagnostic Interview; SAMHSA, Substance Abuse and Mental Health Services Administration; SSAGA, Semi-Structured Assessment for the Genetics of Alcoholism.

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

Table 2. *Variance component estimates for basic and expanded models*

	Model 1 (equation 1)		Model 2 (equation 2)	
Parameter	Estimate (S.E.)	*p*	Estimate (S.E.)	*p*
$\sigma^2_{\text{CRITERION}}$	0.109 (0.050)	0.015	0.073 (0.054)	0.087
σ^2_{SAMPLE}	0.610 (0.131)	<0.001	0.041 (0.025)	0.047
$\sigma^2_{\text{INSTRUMENT}}$			0.074 (0.107)	.245
$\sigma^2_{\text{TIME-FRAME}}$			0.215 (0.380)	0.286
$\sigma^2_{\text{CLINICAL}}$			0.386 (0.482)	0.212
σ^2_{GENDER}			0.006 (0.018)	0.360
σ^2_{AGE}			0.000[a]	–
σ^2_{GROUP}			0.405 (0.134)	0.001
$\sigma^2_{\text{CRITERION} \times \text{INSTRUMENT}}$			0.095 (0.024)	<0.001
$\sigma^2_{\text{CRITERION} \times \text{TIME-FRAME}}$			0.009 (0.011)	0.216
$\sigma^2_{\text{CRITERION} \times \text{CLINICAL}}$			0.020 (0.013)	0.068
$\sigma^2_{\text{CRITERION} \times \text{GENDER}}$			0.000 (0.005)	0.486
$\sigma^2_{\text{CRITERION} \times \text{AGE}}$			0.070 (0.021)	<0.001
σ^2_{ERROR}	0.302 (0.020)	<0.001	0.129 (0.010)	<0.001
ICC				
Estimate (95% CI)[b]	0.265 (0.244–0.288)		0.185 (0.154–0.207)	

S.E., Standard error; ICC, intraclass correlation; CI, confidence interval.

[a] Variance component could not be estimated.

[b] CIs were calculated using 1000 bootstrapped resamples.

While the major findings of this study are the generalizability coefficients presented above, which demonstrate the overall poor consistency of the relative severities of AUD criteria across the published literature and the large, systematic effects associated with both the diagnostic instrument employed and age, it is useful to isolate those criteria that tend to produce highly replicable relative severities and those that show more variation. Examining the random effects of which criteria are significantly more/less severe than the average severity across studies ($\sigma^2 = 0.07$, S.E. = 0.05, $p = 0.087$), only tolerance ($b = -0.42$, $p = 0.021$) and legal problems ($b = -0.39$, $p = 0.041$) were consistently significantly less and more severe criteria, respectively. The associated random-effects estimates from the criterion × instrument interaction indicated that many of the observed differences were localized to three instruments. Of the 84 random-effect estimates for the criterion × instrument interaction (12 criteria × 7 instruments), 13 reached significance ($p < 0.10$). Of those, five were associated with the AUDADIS, with withdrawal ($b = -0.44$, $p = 0.015$), hazardous use ($b = -0.40$, $p = 0.032$) and quitting/cutting down ($b = -0.34$, $p = 0.057$) estimated as less severe criteria than in the average study, and legal problems ($b = 0.44$, $p = 0.041$) and giving up important activities ($b = 0.40$, $p = 0.029$) estimated as more severe. In the SSAGA, tolerance ($b = -0.37$, $p = 0.054$) and larger/longer ($b = -0.34$, $p = 0.073$) were less severe while withdrawal ($b = 0.54$, $p = 0.005$) and time spent ($b = 0.43$, $p = 0.026$) were more severe. In the SAMHSA time spent ($b = -0.49$, $p = 0.007$) was less severe than in the average study, while larger/longer ($b = 0.69$, $p < 0.001$) and quitting/cutting down ($b = 0.42$, $p = 0.023$) were more severe. Lastly, time spent ($b = 0.41$, $p = 0.029$) was a more severe criterion for studies using other, survey specific, instruments.

While the random-effect variance of sample composition (i.e. clinical, mixed, population-based) was significant, there were no individual criteria that were significantly associated with greater/less severity for different types of samples. In contrast, for the 60 individual random-effect estimates for differences in criteria severities depending on age group (12 criteria × 5 age groups), seven were statistically significant. Studies that assessed adolescents had a tendency to find social problems as a less severe criterion ($b = -0.33$, $p = 0.066$) and quitting/cutting down as a more severe criterion ($b = 0.52$, $p = 0.002$). Studies that predominantly assessed young adults were more likely to find that tolerance ($b = -0.61$, $p < 0.001$) and time spent ($b = -0.33$, $p = 0.046$) were less severe criteria, while withdrawal ($b = -0.37$, $p = 0.054$) was a more severe criterion than average. Lastly, in samples that assessed representative populations quitting/cutting down was a less severe criterion ($b = -0.32$, $p = 0.053$) and tolerance was a more severe criterion ($b = 0.28$, $p = 0.084$).

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

Discussion

The introduction of structured and semi-structured diagnostic interviews tied to modern diagnostic criteria such as the DSM-III (American Psychiatric Association, 1980) was spurred by seminal studies showing the unreliability of psychiatric diagnosis such as the International Pilot Study of Schizophrenia (Sartorius *et al.* 1974). While more explicit diagnostic criteria and interviews tailored to assess them represented a major advance over the less structured assessments and vaguer diagnostic systems that characterized the pre-modern era, there has been very little attention paid to the implications of generalizability of the diagnostic criteria themselves across interviews and operationalizations more generally. The validation work that has been done (both within- and between interviews) has focused almost exclusively on diagnosis or symptom count (e.g. Grant *et al.* 1995, 2003, 2015; Chatterji *et al.* 1997; Hasin *et al.* 1997; Vrasti *et al.* 1998; Canino *et al.* 1999; Ruan *et al.* 2008), which ignores which criteria specifically are judged as present or absent. Therefore, it is possible for two interviews to be highly reliable in that they identify the same individuals as having a disorder, but the actual criteria being met are different.

Indeed, our primary finding is that the AUD criteria set evidences very low levels of consistency of criterion severity from one study to the next, and a considerable amount of this inconsistency appears attributable to the specific assessment employed and the age groups assessed. Thus, much of the inconsistency is not random and is a function of systematic study characteristics. Some symptoms such as withdrawal are highly severe symptoms by some assessments (e.g. SSAGA) and 'middling' severity symptoms by other assessments (e.g. AUDADIS). As such, we cannot make strong, generalizable statements about which criteria are intrinsically and universally likely to be more or less severe based upon the extensive, extant literature. However, the results do suggest a degree of local generalizability when such factors are taken into account. We do not view this lack of consistency as an inherent problem of the diagnostic systems themselves (DSM-IV and DSM-5) as much as the underappreciated issue of the operationalization of these criteria in research interviews (and by extension, in clinical practice where even greater variability of assessment may be likely) and the more appreciated issues associated with assessing different populations using the same criteria.

The highly significant criterion × instrument interaction is particularly notable because it suggests that although all of the AUD instruments are based on the same set of criteria definitions, subtle, and what otherwise might be considered trivial differences in wording or administration lead to marked differences in the level and ordering of criteria. Previous research has demonstrated that even ostensibly trivial changes in a diagnostic interview can lead to wildly different lifetime prevalence estimates of AUD (Vergés *et al.* 2011). The findings here suggest an even more serious concern, that the findings of a strong positive manifold among diagnostic criteria robustly found across instruments and samples and that support a unidimensional structure (Hasin *et al.* 2013) represents only 'skin deep' replication of the latent structure of AUD. That is, there is consistency in that only one latent factor is required to explain covariances among criteria, but a lack of consistency in the structure of that factor. This indicates low generalizability of the form of diagnosis when study features are ignored, especially among those who diagnose at low and moderate levels of DSM-5 diagnostic severity. This issue is potentially highly important if one is interested in specific criteria endorsements to guide treatment selection, such as those central to theories of physiological dependence (e.g. withdrawal, craving; Robinson & Berridge, 1993; Langenbucher *et al.* 2000). This is similarly important from a research perspective in apportioning variance to individual criteria due to exogenous variables or in considering alternative models of diagnosis (e.g. network models; Cramer *et al.* 2010; Borsboom & Cramer, 2013).

More expected is the highly significant criterion × age group interaction. A number of researchers have been interested in the structure of AUD in younger individuals and how if differs from adults (see Table 1). Some have considered the systematic differences in endorsement of specific criteria between adolescents and young adults to be due to measurement error, namely tolerance and withdrawal (Caetano & Babor, 2006). Consistent with those interpretations, our analysis suggests that tolerance is a lower threshold criterion for young people to endorse. However, contrary to those findings, ours suggest that withdrawal is on average a higher threshold criterion for young people. One way to reconcile these opposing findings, which was similarly advanced by Caetano & Babor (2006), is by considering the diagnostic instrument used to assess the criteria. They suggest, as we find, that withdrawal is a relatively easy criterion to endorse in the AUDADIS compared with other instruments. Also, a majority of the adolescents included in the current meta-analysis completed instruments (e.g. SSAGA, SAMHSA) where withdrawal is a higher severity criterion that is more difficult to endorse across all age groups.

Other researchers suggest that the differing criteria endorsements as a function of age can be explained by developmental factors relating to various life role transitions (Christo, 1998; Martin *et al.* 2008, 2011, 2014).

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

Our results are similarly consistent with these interpretations, and suggest that if researchers and clinicians explicitly (or implicitly) account for the observation that different symptoms carry different meaning for underlying severity depending on a participant's/patient's age (i.e. withdrawal is more severe for younger individuals while social consequences are more severe for older adults), then the generalizability of overall diagnoses should be increased. Indeed, if such age effects were to be explicitly estimated in research studies, and we assume that such accommodations can be made in the clinic, the estimated reliability coefficient in our analysis would increase from 0.18 to 0.36. However, we note that overall generalizability would still be low. If we in addition were able to account for differences associated with criteria endorsement for individuals belonging to various clinical subpopulations compared with those who are generally healthy, the generalizability estimate would further increase to 0.41, but is still low. However, substantively these effects can still be meaningful. Adjusting for and otherwise understanding that endorsement of certain criteria confer different information about where individuals are in the progression of alcohol use problems, as indexed by where they are likely to be in their drinking career and if they are probably experiencing problems due to other related causes, is likely to still be of substantial interest in guiding further research and practice with respect to treatment.

We did not observe differences in the severities of individual criteria as a function of gender or diagnosis time-frame, both of which could be hypothesized in light of previous research (Harford *et al.* 2009; Shmulewitz *et al.* 2010). However, we do not necessarily view this as a failure to replicate. A majority of the studies included in our analysis contained approximately equal numbers of men and women ($n = 25$, 51%) and assessed past-year AUD ($n = 31$, 63%), resulting in relatively little variability compared with the other moderators. Furthermore, these factors were highly correlated with the particular study they were associated with [e.g. National Epidemiological Study on Alcohol and Related Conditions (NESARC), National Survey of Drug Use and Health (NSDUH)], leading to little unique variability. Thus, the current analysis is not ideally suited for identifying such effects based on the relatively small sample of available IRT studies of AUD and the largely within-study nature of these factors – the latter of which is common to all meta-analyses barring access to raw data from individual studies (Cooper *et al.* 2009).

Substance use disorders and AUDs in particular have seen the most use of IRT in assessing criteria severity (Langenbucher *et al.* 2004; Hasin *et al.* 2013). Personality, mood, anxiety and psychotic disorders have been explored less extensively, but to the extent to which there is variability in sample characteristics or assessment instrument, we might expect the same low levels of agreement in criteria severity that we find with respect to AUDs.

One strategy for mitigating criteria severity inconsistency is through the use of integrative data analysis techniques (IDA; Curran *et al.* 2008; Curran & Hussong, 2009; Hofer & Piccinin, 2009). Such methods allow for underlying overlap between instruments from different investigations to be estimated and compared in light of their structural differences. These approaches could be used to optimize model estimation of severity parameters such that they are maximally consistent across samples. For such an approach to be successful, there needs to be sufficient overlap with respect to the specific wording of items, probes of items, and thresholds for determining whether or not a given symptom fulfills the criterion. Witkiewitz *et al*. (2016) recently conducted an IDA of data from four alcohol treatment studies, integrating data across four diagnostic instruments, and even after harmonizing across instruments identified differences in symptom endorsement as a function of age, treatment status and gender. Their findings suggest that studies utilizing different diagnostic interviews/instruments can be harmonized to maximize generalizability of diagnosis symptom structure, and underscore the importance of taking demographic factors into account (e.g. age, gender, culture) in making general statements about individual symptom severity.

We investigated why the SSAGA may have exhibited such different criterion severity ordering compared with the other instruments. We observed that for a number of criteria (e.g. withdrawal), symptoms must have occurred at least three times in the past year, whereas other instruments (e.g. AUDADIS) require that it only occurred more than once to qualify for endorsement of the criteria. This could have resulted in withdrawal and other criteria obtaining relatively higher thresholds/severities in the IRT analyses compared with criteria that did not, because they were incrementally harder to endorse. Additionally, another reason why withdrawal may be more severe in the SSAGA relative to, say, the AUDADIS is because it stipulates that individuals must have experienced the items 'for most of the day for 2 days or longer'. Other instruments such as the AUDADIS and CIDI do not require that the items be experienced for such a long duration. We also note that steps are taken within the SSAGA to exclude experiences of hangover from qualifying towards withdrawal with fewer safeguards in the AUDADIS (Grant *et al.* 2003). We withhold judgment as to which operationalization is preferred, but rather note that

https://doi.org/10.1017/S0033291716000404 Published online by Cambridge University Press

resolution of these issues should improve the generalizability of findings across all studies. The analyses presented here are useful as exploratory tools to identify where differences between instruments, age composition and other factors might lead to differences in which criteria are deemed more/less severe.

One consideration that bears mention is that some criteria, in a classical test theory sense (Crocker & Algina, 1986), may inherently be more reliable in the way they are measured by virtue of how many items are used to assess them. This can also be the case for overall criteria sets within instruments. Furthermore, while the end result is a binary criterion endorsement (not a graded criterion score), such additional assessments (and qualification questions used by some instruments) can be used to provide resolution to criteria measurement. We were unable to assess the possible influence of these features given that individual criterion assessment is highly confounded within instrument, though both may explain part of the instrument variability we observed.

In general, we find that which AUD criteria are more or less likely to be endorsed is not consistent across studies. This is critical given the preferential focus on certain criteria in theoretical models and intervention research (e.g. Langenbucher *et al.* 2000; De Bruijn *et al.* 2005). While factors such as age and type of sample explain part of this inconsistency (Caetano & Babor, 2006; Martin *et al.* 2008, 2011, 2014), the typically overlooked factor of diagnostic instrument/interview, which is typically assumed to be essentially interchangeable, accounts for a majority of explainable unreliability.

While traditional psychiatric diagnosis is being actively challenged by alternative models such as the Research Domain Criteria (RDoC) initiative [National Institute of Mental Health, 2008; see also Litten *et al.* (2015) for an extension of the RDoC framework to addictive disorders], future research should be aware of how instrument properties may account for observed results (Sher, 2015). The issues we identify regarding instrument are not necessarily limited to traditional diagnosis but can extend to dimensional assessments as well.

With the results from our analysis in mind, if researchers were to *a priori* adjust for the systematic differences in criteria severities due to all of the factors we modeled (most notably instrument), the estimate of the reliability of criteria severities increases from 0.18 to 0.67, a substantial increase that suggests that the generalizability of their findings may be reasonable. However, the problem with diagnostic instruments remains because criterion operationalization had the largest impact on criteria severities, and which operationalization is preferred is an open and understudied topic.

Conclusion

There are strong reasons to question the broad generalizability of criteria severities from any individual IRT study of AUD, and, likely, other psychiatric disorders, without taking into account systematic factors. Some factors have been increasingly studied (e.g. age, gender) while others may be less recognized but even more important (e.g. instrument). The fact that there is considerable variability associated with particular diagnostic instruments highlights the need for further standardization of how diagnostic items are operationalized and administered. To the extent that these measurement concerns are rooted in assessment and sampling variability, even ostensibly alternative approaches to diagnosis (e.g. RDoC; National Institute of Mental Health, 2008) need to be attentive to underlying structural validity concerns.

Supplementary material

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/S0033291716000404

Acknowledgements

S.P.L. has had full access to the data and conducted the analyses under the supervision of K.J.S. and D.S. The project described was supported by grants R01AA024133, K05AA01724 and T32AA013526 from the National Institute on Alcohol Abuse and Alcoholism to K.J.S. and grant R01AA023248 from the National Institute on Alcohol Abuse and Alcoholism to D.S. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institute on Alcohol Abuse and Alcoholism.

Declaration of Interest

None.

References

Aboraya A, Rankin E, France C, El-Missiry A, John C (2006). The reliability of psychiatric diagnosis revisited: the clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry* (*Edgmont*) **3**, 41–50.

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Association: Washington, DC.

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

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn., text revision. American Psychiatric Association: Washington, DC.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. American Psychiatric Association: Washington, DC.

Balsis S, Gleason ME, Woods CM, Oltmanns TF (2007). An item response theory analysis of DSM-IV personality disorder criteria across younger and older age groups. *Psychology and Aging* **22**, 171–185.

Baraona E, Abittan CS, Dohmen K, Moretti M, Pozzato G, Chayes ZW, Schaefer C, Lieber CS (2001). Gender differences in pharmacokinetics of alcohol. *Alcoholism: Clinical and Experimental Research* **25**, 502–507.

Beseler CL, Taylor LA, Leeman RF (2010). An item-response theory analysis of DSM-IV alcohol-use disorder criteria and "binge" drinking in undergraduates. *Journal of Studies on Alcohol and Drugs* **71**, 418–423.

Bond J, Ye Y, Cherpitel CJ, Borges G, Cremonte M, Moskalewicz J, Swiatkiewicz G (2012). Scaling properties of the combined ICD-10 dependence and harms criteria and comparisons with DSM-5 alcohol use disorder criteria among patients in the emergency department. *Journal of Studies on Alcohol and Drugs* **73**, 328–336.

Borges G, Cherpitel CJ, Ye Y, Bond J, Cremonte M, Moskalewicz J, Swiatkiewicz G (2011). Threshold and optimal cut-points for alcohol use disorders among patients in the emergency department. *Alcoholism: Clinical and Experimental Research* **35**, 1270–1276.

Borges G, Ye Y, Bond J, Cherpitel CJ, Cremonte M, Moskalewicz J, Swiatkiewicz G, Rubio-Stipec M (2010). The dimensionality of alcohol use disorders and alcohol consumption in a cross-national perspective. *Addiction* **105**, 240–254.

Borsboom D, Cramer AO (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology* **9**, 91–121.

Brennan RL (2001). *Generalizability Theory*. Springer: New York.

Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI Jr., Reich T, Schmidt I, Schuckit MA (1994). A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *Journal of Studies on Alcohol* **55**, 149–158.

Caetano R, Babor TF (2006). Diagnosis of alcohol dependence in epidemiological surveys: an epidemic of youthful alcohol dependence or a case of measurement error? *Addiction* **101**, 111–114.

Canino G, Bravo M, Ramírez R, Febo VE, Rubio-Stipec M, Fernández RL, Hasin D (1999). The Spanish Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS): reliability and concordance with clinical diagnoses in a Hispanic population. *Journal of Studies on Alcohol* **60**, 790–799.

Casey M, Adamson G, Shevlin M, McKinney A (2012). The role of craving in AUDs: dimensionality and differential functioning in the DSM-5. *Drug and Alcohol Dependence* **125**, 75–80.

Chatterji S, Saunders JB, Vrasti R, Grant BF, Hasin D, Mager D (1997). Reliability of the alcohol and drug modules of the Alcohol Use Disorder and Associated Disabilities Interview Schedule – Alcohol/Drug-Revised (AUDADIS-ADR): an international comparison. *Drug and Alcohol Dependence* **47**, 171–185.

Cherpitel CJ, Borges G, Ye Y, Bond J, Cremonte M, Moskalewicz J, Swiatkiewicz G (2010). Performance of a craving criterion in DSM alcohol use disorders. *Journal of Studies on Alcohol and Drugs* **71**, 674–684.

Christo G (1998). A review of reasons for using or not using drugs: commonalities between sociological and clinical perspectives. *Drugs: Education, Prevention, and Policy* **5**, 59–72.

Collins FS, Tabak LA (2014). NIH plans to enhance reproducibility. *Nature* **505**, 612–613.

Cooper H, Hedges LV, Valentine JC (2009). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd edn. Russell Sage Foundation: New York.

Cooper JE, Kendell RE, Gurland BJ, Sharpe L, Copeland J (1972). *Psychiatric Diagnosis in New York and London*. Oxford University Press: London, UK.

Cooper LD, Balsis S (2009). When less is more: how fewer diagnostic criteria can indicate greater severity. *Psychological Assessment* **21**, 285–293.

Cramer AO, Waldorp LJ, van der Maas HL, Borsboom D (2010). Comorbidity: a network perspective. *Behavioral and Brain Sciences* **33**, 137–150.

Cranford JA, Shrout PE, Iida M, Rafaeli E, Yip T, Bolger N (2006). A procedure for evaluating sensitivity to within-person change: can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin* **32**, 917–929.

Crocker L, Algina J (1986). *Introduction to Classical and Modern Test Theory*. Wadsworth: Bellemont, CA.

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley: New York.

Curran PJ, Hussong AM (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods* **14**, 81–100.

Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA (2008). Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Developmental Psychology* **44**, 365–380.

Dawson DA, Saha TD, Grant BF (2010). A multidimensional assessment of the validity and utility of alcohol use disorder severity as determined by item response theory models. *Drug and Alcohol Dependence* **107**, 31–38.

De Bruijn C, Van Den Brink W, De Graaf R, Vollebergh WA (2005). The craving withdrawal model for alcoholism: towards the DSM-V. Improving the discriminant validity of alcohol use disorder diagnosis. *Alcohol and Alcoholism* **40**, 314–322.

Derringer J, Krueger RF, Dick DM, Agrawal A, Bucholz KK, Foroud T, Grucza RA, Hesselbrock MN, Hesselbrock V, Kramer J, Nurnberger JI Jr., Schuckit M, Bierut LJ, Iacono WG, McGue M (2013). Measurement invariance of DSM-IV alcohol, marijuana and cocaine dependence between community-sampled and clinically over-selected studies. *Addiction* **108**, 1767–1776.

Duncan AE, Agrawal A, Bucholz KK, Sartor CE, Madden PA, Heath AC (2011). Deconstructing the architecture of alcohol abuse and dependence symptoms in a community sample of late adolescent and emerging adult women: aAn item response approach. *Drug and Alcohol Dependence* **116**, 222–227.

Edwards AC, Gillespie NA, Aggen SH, Kendler KS (2013). Assessment of a modified DSM-5 diagnosis of alcohol use disorder in a genetically informative population. *Alcoholism: Clinical and Experimental Research* **37**, 443–451.

Ehlke SJ, Hagman BT, Cohn AM (2012). Modeling the dimensionality of DSM-IV alcohol use disorder criteria in a nationally representative sample of college students. *Substance Use and Misuse* **47**, 1073–1085.

Embretson SE, Reise SP (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates: Mahwah, NJ.

Gelhorn H, Hartman C, Sakai J, Stallings M, Young S, Rhee S, Corley R, Hewitt J, Hopfer C, Crowley T (2008). Toward DSM-V: an item response theory analysis of the diagnostic process for DSM-IV alcohol abuse and dependence in adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry* **47**, 1329–1339.

Gilder DA, Gizer IR, Ehlers CL (2011). 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 and Experimental Research* **35**, 984–995.

Grant BF, Dawson DA, Stinson FS, Chou PS, Kay W, Pickering R (2003). 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* **71**, 7–16.

Grant BF, Goldstein RB, Smith SM, Jung J, Zhang H, Chou SP, Pickering RP, Ruan WJ, Huang B, Saha TD, Aivadyan C, Greenstein E, Hasin DS (2015). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-5 (AUDADIS-5): reliability of substance use and psychiatric disorder modules in a general population sample. *Drug and Alcohol Dependence* **148**, 27–33.

Grant BF, Harford TC, Dawson DA, Chou PS, Pickering RP (1995). The Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS): reliability of alcohol and drug modules in a general population sample. *Drug and Alcohol Dependence* **39**, 37–44.

Hagman BT, Cohn AM (2011). Toward DSM-V: mapping the alcohol use disorder continuum in college students. *Drug and Alcohol Dependence* **118**, 202–208.

Hagman BT, Cohn AM (2013). Using latent variable techniques to understand DSM-IV alcohol use disorder criteria functioning. *American Journal of Health Behavior* **37**, 565–574.

Harford TC, Yi HY, Faden VB, Chen CM (2009). The dimensionality of DSM-IV alcohol use disorders among adolescent and adult drinkers and symptom patterns by age, gender, and race/ethnicity. *Alcoholism: Clinical and Experimental Research* **33**, 868–878.

Hasin D, Carpenter KM, McCloud S, Smith M, Grant BF (1997). The Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS): reliability of alcohol and drug modules in a clinical sample. *Drug and Alcohol Dependence* **44**, 133–141.

Hasin D, Samet S, Nunes E, Meydan J, Matseoane K, Waxman R (2006). Diagnosis of comorbid psychiatric disorders in substance users assessed with the Psychiatric Research Interview for Substance and Mental Disorders for DSM-IV. *American Journal of Psychiatry* **163**, 689–696.

Hasin DS, Fenton MC, Beseler C, Park JY, Wall MM (2012). 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* **122**, 28–37.

Hasin DS, O'Brien CP, Auriacombe M, Borges G, Bucholz K, Budney A, Compton WM, Crowley T, Ling W, Petry NM, Schuckit M, Grant BF (2013). DSM-5 criteria for substance use disorders: recommendations and rationale. *American Journal of Psychiatry* **170**, 834–851.

Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003). Measuring inconsistency in meta-analyses. *BMJ* (*Clinical Research ed.*) **327**, 557–560.

Hofer SM, Piccinin AM (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods* **14**, 150–164.

Jacobus J, Tapert SF (2013). Neurotoxic effects of alcohol in adolescence. *Annual Review of Clinical Psychology* **9**, 703–721.

Keyes KM, Krueger RF, Grant BF, Hasin DS (2011). Alcohol craving and the dimensionality of alcohol disorders. *Psychological Medicine* **41**, 629–640.

Kuerbis AN, Hagman BT, Morgenstern J (2013*a*). Alcohol use disorders among substance dependent women on temporary assistance with needy families: more information for diagnostic modifications for DSM-5. *American Journal on Addictions* **22**, 402–410.

Kuerbis AN, Hagman BT, Sacco P (2013*b*). Functioning of alcohol use disorders criteria among middle-aged and older adults: implications for DSM-5. *Substance Use and Misuse* **48**, 309–322.

Lane SP, Sher KJ (2015). Limits of current approaches to diagnosis severity based on criterion counts: an example with DSM-5 alcohol use disorder. *Clinical Psychological Science* **3**, 819–835.

Langenbucher J, Chung T (1995). Onset and staging of DSM-IV alcohol dependence using mean age and survival-hazard methods. *Journal of Abnormal Psychology* **104**, 346–354.

Langenbucher J, Martin CS, Labouvie E, Sanjuan PM, Bavly L, Pollock NK (2000). Toward the DSM-V: the withdrawal-gate model *versus* the DSM-IV in the diagnosis of alcohol abuse and dependence. *Journal of Consulting and Clinical Psychology* **68**, 799–809.

Langenbucher JW, Labouvie E, Martin CS, Sanjuan PM, Bavly L, Kirisci L, Chung T (2004). An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. *Journal of Abnormal Psychology* **113**, 72–80.

Litten RZ, Ryan ML, Falk DE, Reilly M, Fertig JB, Koob GF (2015). Heterogeneity of alcohol use disorder: understanding

mechanisms to advance personalized treatment. *Alcoholism: Clinical and Experimental Research* **39**, 579–584.

Loevinger J (1957). Objective tests as instruments of psychological theory: monograph supplement 9. *Psychological Reports* **3**, 635–694.

Makel MC, Plucker JA (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher* **43**, 304–316.

Martin CS, Chung T, Kirisci L, Langenbucher JW (2006). Item response theory analysis of diagnostic criteria for alcohol and cannabis use disorders in adolescents: implications for DSM-V. *Journal of Abnormal Psychology* **115**, 807–814.

Martin CS, Chung T, Langenbucher JW (2008). How should we revise diagnostic criteria for substance use disorders in the DSM-V? *Journal of Abnormal Psychology* **117**, 561–575.

Martin CS, Kaczynski NA, Maisto SA, Bukstein OM, Moss HB (1995). Patterns of DSM-IV alcohol abuse and dependence symptoms in adolescent drinkers. *Journal of Studies in Alcohol* **56**, 672–680.

Martin CS, Langenbucher JW, Chung T, Sher KJ (2014). Truth or consequences in the diagnosis of substance use disorders. *Addiction* **109**, 1773–1778.

Martin CS, Sher KJ, Chung T (2011). Hazardous use should not be a diagnostic criterion for substance use disorders in DSM-5. *Journal of Studies on Alcohol and Drugs* **72**, 685–686.

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

Mewton L, Slade T, McBride O, Grove R, Teesson M (2011*a*). An evaluation of the proposed DSM-5 alcohol use disorder criteria using Australian national data. *Addiction* **106**, 941–950.

Mewton L, Teesson M, Slade T, Cottler L (2011*b*). 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* **72**, 811–822.

Mezey E, Kolman CJ, Diehl AM, Mitchell MC, Herlong HF (1988). Alcohol and dietary intake in the development of chronic pancreatitis and liver disease in alcoholism. *American Journal of Clinical Nutrition* **48**, 148–151.

National Institute of Mental Health (2008). *National Institute of Mental Health Strategic Plan* (NIH publication no. 08-6368). US Government Printing Office: Washington, DC.

Nosek BA, Lakens D (2013). Call for proposals: special issue of *Social Psychology* on 'Replications of important results in social psychology'. *Social Psychology* **44**, 59–60.

O'Neill SE, Sher KJ (2000). Physiological alcohol dependence symptoms in early adulthood: a longitudinal perspective. *Experimental and Clinical Psychopharmacology* **8**, 493–508.

Oscar-Berman M (2000). Neuropsychological vulnerabilities in chronic alcoholism. In *Review of NIAAA's Neuroscience and Behavioral Research Portfolio* (ed. A. Noronha, M. J. Eckardt and K. Warren), pp. 437–471. National Institute on Alcohol Abuse and Alcoholism (NIAAA) Research Monograph No. 34. NIAAA: Bethesda, MD.

Preuss UW, Watzke S, Wurst FM (2014). Dimensionality and stages of severity of DSM-5 criteria in an international sample of alcohol-consuming individuals. *Psychological Medicine* **44**, 3303–3314.

Proudfoot H, Baillie AJ, Teeson M (2006). The structure of alcohol dependence in the community. *Drug and Alcohol Dependence* **81**, 21–26.

Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto J, Kuhl EA, Kupfer DJ (2013). DSM-5 field trials in the United States and Canada, part II: test–retest reliability of selected categorical diagnoses. *American Journal of Psychiatry* **170**, 59–70.

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

Robinson TE, Berridge KC (1993). The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Research Reviews* **18**, 247–291.

Rose JS, Lee CT, Selya AS, Dierker LC (2012). DSM-IV alcohol abuse and dependence criteria characteristics for recent onset adolescent drinkers. *Drug and Alcohol Dependence* **124**, 88–94.

Ruan WJ, Goldstein RB, Chou SP, Smith SM, Saha TD, Pickering RP, Dawson DA, Huang B, Stinson FS, Grant BF (2008). The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of new psychiatric diagnostic modules and risk factors in a general population sample. *Drug and Alcohol Dependence* **92**, 27–36.

Saha TD, Chou SP, Grant BF (2006). Toward an alcohol use disorder continuum using item response theory: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychological Medicine* **36**, 931–941.

Saha TD, Stinson FS, Grant BF (2007). The role of alcohol consumption in future classifications of alcohol use disorders. *Drug and Alcohol Dependence* **89**, 82–92.

Sartorius N, Shapiro R, Jablensky A (1974). The international pilot study of schizophrenia. *Schizophrenia Bulletin* **1**, 21–34.

Sher KJ (2015). Moving the alcohol addiction RDoC forward. *Alcoholism: Clinical and Experimental Research* **39**, 591–591.

Shmulewitz D, Keyes K, Beseler C, Aharonovich E, Aivadyan C, Spivak B, Hasin D (2010). The dimensionality of alcohol use disorders: results from Israel. *Drug and Alcohol Dependence* **111**, 146–154.

Shrout PE, Fleiss JL (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86**, 420–428.

Silveri MM, Spear LP (2001). Acute, rapid, and chronic tolerance during ontogeny: observations when equating ethanol perturbation across age. *Alcoholism: Clinical and Experimental Research* **25**, 1301–1308.

Smith ES, Riechelmann H (2004). Cumulative lifelong alcohol consumption alters auditory brainstem potentials. *Alcoholism: Clinical and Experimental Research* **28**, 508–515.

Spitzer RL, Williams JBW (1985). *Structured Clinical Interview for DSM-III-R (SCID)*. Biometrics Research Department, New York State Psychiatric Institute: New York.

Srisurapanont M, Kittiratanapaiboon P, Likhitsathian S, Kongsuk T, Suttajit S, Junsirimongkol B (2012). Patterns

of alcohol dependence in Thai drinkers: a differential item functioning analysis of gender and age bias. *Addictive Behaviors* **37**, 173–178.

Substance Abuse and Mental Health Services Administration (2006). U.S. Dept. of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies. National Survey On Drug Use and Health, 2005 [Computer file]. ICPSR04596-v1. Research Triangle Institute [producer], Research Triangle Park, NC, Inter-University Consortium for Political and Social Research [distributor], Ann Arbor, MI, 2006-11-16.

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

Vergés A, Littlefield AK, Sher KJ (2011). Did lifetime rates of alcohol use disorders increase by 67% in 10 years? A comparison of NLAES and NESARC. *Journal of Abnormal Psychology* **120**, 868–877.

Vrasti R, Grant BF, Chatterji S, Üstün BT, Mager D, Olteanu I, Badoi M (1998). Reliability of the Romanian version of the alcohol module of the WHO Alcohol Use Disorder and Associated Disabilities: Interview Schedule–Alcohol/Drug-Revised. *European Addiction Research* **4**, 144–149.

Witkiewitz K, Hallgren KA, O'Sickey AJ, Roos CR, Maisto SA (2016). Reproducibility and differential item functioning of the alcohol dependence syndrome construct across four alcohol treatment studies: an integrative data analysis. *Drug and Alcohol Dependence* **158**, 86–93.

World Health Organization (1977). *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death, Ninth Revision*, vol. 1. WHO: Geneva.

World Health Organization (1978). *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death, Ninth Revision*, vol. 2. WHO: Geneva.

World Health Organization (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*, vol. 1. WHO: Geneva.

World Health Organization (1997). *Composite International Diagnostic Interview – version 2.0*. WHO: Geneva.

Wu LT, Blazer DG, Woody GE, Burchett B, Yang C, Pan JJ, Ling W (2012). Alcohol and drug dependence symptom items as brief screeners for substance use disorders: results from the Clinical Trials Network. *Journal of Psychiatric Research* **46**, 360–369.

Wu LT, Pan JJ, Blazer DG, Tau B, Stitzer ML, Brooner RK, Woody GE, Patkar AA, Blaine JD (2009). An item response theory modeling of alcohol and marijuana dependences: a National Drug Abuse Treatment Clinical Trials Network study. *Journal of Studies on Alcohol and Drugs* **70**, 414–425.