Published in final edited form as:

Clin Psychol Sci. 2018 ; 6(4): 506–516. doi:10.1177/2167702617747657.

Criteria Definitions and Network Relations: The Importance of Criterion Thresholds

Michaela Hoffman, Douglas Steinley, Timothy J. Trull, and Kenneth J. Sher

University of Missouri, Columbia

Abstract

Across various structured diagnostic instruments, the criteria used to diagnose alcohol use disorder (AUD) are not assessed consistently. For example, different instruments often pose questions that reflect different thresholds of the underlying symptoms. We consider the criteria for *craving* and *the inability to cut down or stop drinking* to demonstrate the influence of using different thresholds for a positive symptom endorsement with respect to the estimated edges of a symptom network. Results indicate that the utilization of these differing thresholds leads to significant differences in edge weights. Generally, higher thresholds relate more strongly to lower prevalence rate criteria, and the reverse for lower thresholds. These findings have implications for reproducibility of effects in symptom networks and their generalization across studies.

Keywords

alcohol use disorder; network models; dsm 5 criteria; criteria wording; survey construction

Introduction

The diagnostic revolution in the 60s and 70s culminated in the widespread adoption of DSM-III (APA, 1980). Although this ushered in an era of more explicit criteria definitions and structured interviews designed to assess them, there has been a general lack of appreciation of variability with respect to how individual diagnostic criteria are operationalized across alternative diagnostic instruments. In the case of alcohol use disorder (AUD), and presumably other disorders as well, there is considerable variability across diagnostic instruments such that a criterion that represents a low-middle severity criterion in one instrument can be a high severity criterion in another instrument (Lane, Steinley, & Sher, 2016). This might appear to be trivial at first glance, and may often be overlooked as the researcher focuses on the final diagnoses or model characteristics. The importance of the specific way criteria are assessed however should be considered for the interpretations of the

Correspondence concerning this article should be addressed to Michaela Hoffman, Medical University of South Carolina, 67 President Street, 473N, Charleston, SC 29425. HoffmaMi@musc.edu; Phone: (843) 792-7758, Contact: HoffmanMic@missouri.edu.

This manuscript was prepared using a limited access dataset obtained from the National Institute on Alcohol Abuse and Alcoholism and does not reflect the opinions or views of NIAAA or the U.S. Government.

Michaela Hoffman is now at the Department of Psychiatry and Behavioral Sciences, Charleston Alcohol Research Center, Medical University of South Carolina.

final results, as the reliability and validity of any results is predicated on our ability to precisely identify the disorder itself.

Variation in operationalization has received relatively little attention in the empirical literature, although Regier et. al (1998) discussed the general need for improved standardization of assessment methods 20 years ago. More recently, it was demonstrated that for the diagnosis of a major depressive episode (MDE), varying the threshold of the “depressed mood all day” criterion, which is required for the diagnosis, produces very different prevalence rates (Karlsson, et al., 2010). In this case, the threshold was defined by setting the cutoff for the criterion at different response levels to the same question. Similarly, different diagnostic instruments for AUD operationalize the descriptors “recurrent” or “repeatedly” as reflecting two or more (e.g., AUDADIS-5; Hasin et al., 2015) or three or more times (SSAGA, Bucholz et al., 1994).

Beyond using the duration and frequency of specific symptoms, the severity of criterion thresholds can vary based on the symptoms used to operationalize them. For example, one can vary the threshold for the presence/absence of a diagnostic criterion by using multiple, binary items that exhibit properties of a Guttman scale. Guttman scales utilize a series of binary questions to assess a single construct. The notable feature is that items are (largely) nested within each other successively (e.g., a “yes” for question two implies the answer to question one was also “yes.”). In the context of symptoms of psychopathology, this could be viewed as degrees of severity. The implication of this definitional variability for nosological research on symptom structure is unknown, but, should it be influential, the generalizability of estimates of relations across studies employing different assessments is called into question.

Alcohol Use Disorder (AUD)

The assessment of AUDs provides an informative case to explore the relevance of criterion thresholds. In a recent meta-analysis, Lane et al. (2016) found considerable variability in the rank order of IRT thresholds across 34 published studies (30 unique datasets) with the largest source of identifiable variance attributable to the diagnostic interview employed in the study. Some of this variation is likely due to how different interviews assess the extent to which a given symptom is considered “repeated” or “recurrent” (e.g., two or more times, three or more times), some possibly due to the number of questions used to assess the symptom, and some due to the actual wording of the item.

In many survey instruments, each AUD criterion is assessed using responses from multiple items. In some instances, the items included are distinct components of the criteria. For example, one criterion asks if an individual has given up activities in order to drink. This is measured in the Alcohol Use Disorder and Associated Disabilities Interview Schedule (AUDADIS-IV; Grant, Dawson, & Hasin, 2001; AUDADIS 5; Grant et al., 2011), asking “...did you ever:”

1. Give up or cut down on activities that were important to you in order to drink - like work, school, or associating with friends or relatives?

2. Give up or cut down on activities that you were interested in or that gave you pleasure in order to drink?

These two questions assess different aspects of the same criterion, and one may exist without the other. However, there are other criteria whose sub-items appear to assess different severities of the same symptom. Though not technically Guttman scales in structure, a number of items assessing AUD might be termed “Guttman-like” in that, for the most part, a positive response to one predicates a positive response to the other. The added severity information from these Guttman-like questions is, however, lost on these items when the scoring requires that simply “one or the other.” Final estimates of AUD severity are assessed using the total number of criteria met, ignoring responses to the implicit severity of the item in the subset of items assessing the criterion. Scoring algorithms aside, it is this quality of the AUDADIS that permits an examination of the extent to which the threshold for determining the presence or absence of an individual symptom/criterion can affect analyses. The subsets of two AUD items in particular provide excellent opportunities for exploring the significance and impact of variation in criterion threshold: alcohol craving and the inability to cut down on alcohol use. The National Epidemiological Surveys on Alcohol and Related Conditions, NESARC Wave 2 (Grant et al., 2007) and NESARC III (Grant et al., 2015), provide large samples with which to test the differences between these AUDADIS items. Prior to analyzing these data, we describe the two criteria in detail.

Craving

The *craving* criterion is defined by the DSM-5 as “craving, or a strong desire or urge to use alcohol” (American Psychiatric Association, 2013; pp. 490-491). In diagnostic interviews, however, the wording of the actual questions for this criterion varies. There exist larger scales used to assess craving, such as the 14 item Desires for Alcohol Questionnaire (DAQ; Love et al., 1998, Kramer et al., 2010) including a range of phrases such as “I wanted a drink so much I could almost taste it,” and “I thought drinking was pleasant.” More often in larger, comprehensive surveys fewer items are used. For example, in the Composite International Diagnostic Interview 3.0 (CIDI 3.0; Kessler et al., 2013), *craving* is assessed by the single item “Was there ever a time in your life when you often had such a strong desire to drink that you couldn't stop yourself from taking a drink or found it difficult to think of anything else?” Similarly, in the Semi-Structured Assessment of the Genetics of Alcoholism (SSAGA; Bucholz et al., 1994) *craving* refers to “In situations where you couldn't drink, did you ever have such a strong desire for it that you couldn't think of anything else?” In contrast, for someone to endorse craving on the AUDADIS, one needs only to endorse one of two items asking “...did you ever:”

1. Feel a very strong desire or urge to drink?
2. Want a drink so badly that you couldn't think of anything else?

Here, the first statement is arguably a lower threshold of *craving*, thus the two craving items reflect a Guttman-like quality, where statement (1) could be seen as a lower threshold and prerequisite to statement (2), which is indeed revealed in endorsement rates. The low threshold cut down item has a prevalence of 12.31% in NESARC Wave 2 and 13.43% in NESARC - III, and the high threshold has corresponding prevalence rates of 2.69% and

3.46%. In NESARC Wave 2 and NESARC III, 80.87% and 89.25% of those who endorse the high threshold item also endorse the low threshold item. Correspondingly, only 16.65% and 17.86% of those who endorse the low threshold item also endorse the high threshold item.

Cut Down

The inability to cut down or stop drinking (*cut down*) provides a second example of a diagnostic criterion assessed with statements that differ in severity in the NESARC studies. *Cut down* was included in the DSM-IV as one of the alcohol dependence criteria and was retained in the DSM-5. It is defined by the DSM-5 as “a persistent desire or unsuccessful efforts to cut down or control alcohol use,” (American Psychiatric Association, 2013, p. 490). The “persistent” component of this implies multiple occurrences of the symptom without a strictly defined count, and this is reflected in its varying assessments. For example, the CIDI 3.0 asks “Were there times when you tried to stop or cut down on your drinking and found that you were not able to do so?” (Kessler et al., 2013). The SSAGA-I instrument asks “Have you 3 or more times wanted to stop or cut down on drinking?” followed by “Were you always able to stop or cut down when you wanted to?” Here, the criterion is met if the answer to the first is “yes” or the answer to the second is “no” (Bucholz et al., 1994).

In the National Longitudinal Alcohol Epidemiology Survey (NLAES; Grant et al., 1994), using an earlier version of the AUDADIS, the persistence of the *cut down* items is assessed first asking if the experience had ever happened and following up by asking how many times. Individuals qualified for this criterion by either endorsing both statements or reporting multiple occasions of either in the follow-up question. Both the AUDADIS-IV and 5 collapse the statement and follow up by asking “...did you ever:”

1. More than once want to stop or cut down on your drinking?
2. More than once TRY to stop or cut down on your drinking but found you couldn't do it?

As with the other criteria, *cut down* is scored positively if either item is endorsed (and all respondents are asked both questions). Though, in the structure of the interview, these items are not strictly Guttman scales, (1) does appear to measure a lesser degree of the criterion, and logically it could be argued that in most cases a person would *want* to cut down before they *tried*. This is reflected in the endorsement rates in the NESARC data. The low threshold craving item has a prevalence rate of 4.01% and 10.69%, and the high .83% and 2.14% in NESARC Wave 2 and III, respectively. Of those who report the high threshold cut down item, 86.55% for NESARC Wave 2, and 88.21% for NESARC III report the low threshold item as well, demonstrating that it rarely appears alone. On the other hand, only 18.89% and 22.73% of those who positively endorse the low threshold item also endorse the high threshold item.

Network Analysis

To examine these effects, we utilize an emerging modeling tool for epidemiological studies of psychopathological symptoms: network analysis. This is of particular use here due to its focus on the symptoms or criteria themselves. Network models of psychological symptoms

are increasingly utilized as they allow the researcher an alternative to the classical latent variable models of disorders (Hofmann et al., 2016; Fried et al., 2017). In these networks, symptoms interact with and influence each other rather than reflecting manifestations of an underlying (latent) disorder (Borsboom & Cramer, 2013). Generally, a network is comprised of a set of *nodes*, here, representing the symptoms or criteria of interest, and *edges* that describe how strongly they relate to each other. Unlike traditional networks where the edges are observed (for a review of traditional social networks, see Wasserman & Faust, 1994), in symptom or criteria networks they must be inferred from data. This methodology has already been applied to many studies of specific psychiatric disorders such as substance use disorders (e.g., Anker et al., 2017, Rhemtulla et al., 2016) or depression (e.g., Boschloo et al., 2016).

With the focus on the symptoms themselves, these models provide a framework with which to consider diagnostic criteria individually, as opposed to focusing on the larger diagnosis. This can be done by considering each of an individual criterion's connections within the network. The presence of an edge between two criteria indicates they are significantly associated with each other. Within the realm of network analysis there also exist specific measures used to evaluate the role of a node within an entire network called centrality statistics. Centrality is a measure of importance, and in a network this can take multiple forms. The three commonly used indices to study symptom networks are betweenness, closeness, and strength. *Betweenness centrality* is calculated by counting how often each node lies on shortest paths between all pairs of other nodes in the network, giving higher scores to those that are intermediaries more often. *Closeness centrality* describes how close, on average, the given node is to any other node in the graph. It is calculated for each node as the inverse of the sum of all the edge weights to the other nodes. The *strength* of a node, also called its degree, is simply the sum of the weights of its adjacent edges.

The Present Study

In this paper, we are interested in examining the effect of the different thresholds of a criterion on the estimated edges and centrality estimates of an AUD criteria network. Utilizing network estimation techniques allow us to characterize the way the remaining criteria differentially relate to different thresholds of the criterion in question. Finding that estimates of various network parameters are affected by the severity of the criteria would have important implications for determining the likely reproducibility of research across instruments, as the wording utilized varies.

In the AUDADIS interviews used for NESARC Wave 2 and NESARC III, we focus on the *craving* and *cut down* items as they reflect the Guttman-like quality discussed above. Although counted equally in the scoring of the data (endorsement of either leads to meeting the criterion), the items could be expected to identify people with very different AUD severities. Because of this, we expect them individually to relate differently to other symptoms in the network. This allows us to examine how the threshold used to determine presence/absence of a symptom might affect the structure of the criteria used to make a given diagnosis, and to investigate the effect of varying the threshold for determining the presence/absence of a symptom on both generalizability and reproducibility.

Method

Samples

Two National Epidemiological Survey on Alcohol and Related Conditions (NESARC) data sets were included in this analysis. These are nationally representative samples of civilian, non-institutionalized adults in the U.S. population conducted by the National Institute on Alcohol Abuse and Alcoholism (NIAAA). NESARC Wave 2 was a follow up to the original survey, taken in 2004-2005 ($n=34,653$; Grant et al., 2007). NESARC-III was conducted in 2012-2013 with a new sample of $n=36,309$ (Grant et al., 2015). Participants under the age of 21 were removed from the sample, 24 from Wave 2 and 1,597 from III. Of this sample only those who reported they had consumed at least one drink of alcohol in the past year were included, leaving $n=22,160$ and $n=24,773$ observations, respectively. In this final sample, the average age of participants was $45.88(\sigma=15.89)$ for Wave 2 and $44.40(\sigma=16.03)$ for III. The sample from NESARC Wave 2 consisted of 46.86% males and was 63.35% white, 15.44% black and 17.36% Hispanic. The sample from NESARC-III was 46.57% male and 56.03% white, 19.74% black and 18.57% Hispanic.

Measures

The DSM-5 defines 11 criteria for alcohol use disorder¹: larger amounts/longer time spent drinking than intended (*larger/longer*), inability to cut down (*cut down*), excessive time spent drinking or recovering from drinking (*time spent*), *craving*, failure to fulfill major roles and obligations (*role interference*), use despite causing social/interpersonal problems (*social problems*), giving up activities in order to drink (*give up activities*), getting involved in risky situations while or after drinking (*hazardous use*), use despite causing physical/psychological problems (*phys/psych problems*), *tolerance* and *withdrawal*. The AUDADIS-IV (Grant et al., 2001) provides the wording for the questions used to assess these criteria used in the NESARC Wave 2 survey, and the updated AUDADIS 5 (Hasin et al., 2015) was used for NESARC III². The AUDADIS is fully structured diagnostic interview that utilizes the DSM-IV criteria for mental disorders, with the focus on alcohol use disorder (AUD). In this study, only the past year occurrences of the criteria are considered. The unweighted³ sample prevalence rates (for past year drinkers over 21) can be found in Table 1.

Network Estimation

The *eLasso* algorithm creates networks from binary data and can be performed using the *IsingFit* package in R (van Borkulo et al., 2016). A full description is given by van Borkulo et. al. (2014), but here we briefly summarize the process. It begins with the set binary variables representing the criteria, indicating the presence or absence of each in the participants. Selecting one criterion at a time, a regularized logistic regression model is fit

¹A full description of the 11 criteria can be found in Diagnostic and statistical manual of mental disorders (DSM-5®) (American Psychiatric Association, 2013, pp. 490-491).

²For the two criteria under consideration, craving and cut down, the AUDADIS IV and 5 do not differ in the items used. For the remaining criteria, only role interference, hazardous use, and social problems differ: each has an additional item used for assessment added to the AUDADIS 5.

³Although the weights reflect the correction of the collected sample to match the known population, the network estimation methodology cannot yet incorporate this information. For that reason, in this paper, we describe the unweighted prevalence rates for the sample because they best relate to the conducted network analyses.

using the remaining criteria as the predictor variables. The resulting regression coefficients are used as the edge weights in the final network. To create a sparse network, where not all edges are present between pairs of criteria, a Lasso estimator is used to reduce weights to zero where conditional independence is implied between criteria. Finally, to create a symmetric matrix, the estimates for each pair are averaged (e.g. the edge between *cut down* and *larger/longer* is the average of the coefficient where *cut down* is the dependent variable and where *larger/longer* is the dependent variable)⁴.

Analysis

Four networks were estimated for each data set⁵. Networks were created considering the alteration of one criterion at a time, so that when we are studying the differences in the *craving* threshold, the *cut down* criterion is held constant using the original scoring (where a subject qualifies if they have an affirmative response to either statement), and the *craving* criterion is held constant to study the differences in the *cut down* threshold. When each is held constant, the original scoring rule is used which counts the criterion as present if the respondent indicates either or both of the statement are true.

In terms of edge weights, it seems unlikely that the changes in one criterion will extend much beyond their adjacent edges. Due to the pairwise nature of the estimation algorithm and the large sample, these remaining edges, between criteria not modified, are expected to be fairly stable. Thus, there are ten potential edges of interest for each of the criteria being modified: each describing the strength of the relationship (or lack thereof) between the *cut down* or *craving* criterion and the remaining criteria. While the differences in the estimated relations can be observed by considering the correlation of the edge weights, a larger number of observations is required to provide meaningful confidence intervals for the differences. A bootstrapping technique is employed for this purpose. The procedure begins by taking a random sample of the data with replacement (with the sample size of $n=10,000$). From this sample, both the high (e.g., using the high-threshold item) and the low (e.g., using the low threshold item) networks are estimated for a criterion. The data are resampled and the networks estimated 100 times, providing a large enough sample of estimates for inference. It gives us 100 estimates of each of the edges to conduct statistical analyses and provide meaningful confidence intervals on the correlations, as well as t-tests for the differences in specific criterion edges. For the correlations, the Intraclass Correlation coefficient (ICC, with 95% confidence intervals) for random raters is used (Shrout & Fleiss, 1979). All node centrality measures (betweenness, closeness, and strength) are calculated within the R package *qgraph* (Epskamp et al., 2012).

Results

As the relationships in these networks are pairwise, the rest of the network is not appreciably changed by the modification of one criterion. This is reflected in the correlations using all

⁴In addition to the methods described here, the bootnet R package was utilized to examine the accuracy of edges estimated for each network, creating 95% confidence intervals for each edge (Epskamp, Borsboom, & Fried, 2016). The results for each of the eight estimated networks indicated a high consistency, with very narrow confidence intervals, thus we proceeded with our planned analysis.

⁵The AND rule is used by default & the default value for the hyperparameter gamma in the extended BIC is .25. See van Borkulo et al. (2014) for details.

the estimated relations in the network (as the network is symmetric, there are 55 total edges estimated). For NESARC Wave 2, the corresponding edges of the networks using the low versus high threshold of the *craving* criterion correlated $ICC=.93$ (confidence interval=.89, .96) and the *cut down* networks correlated $ICC=.96$ (confidence interval=.93, .97). For NESARC-III the *craving* networks correlate $ICC=.89$ (confidence interval=.82,.94) and the *cut down* networks $ICC=.98$ (confidence interval=.96,.99).

There is some variation in the percentages of the sample who meet criteria for AUD by reporting two or more criteria. When only the low threshold criteria are included the rates are roughly the same, but a slight decrease can be seen where the high threshold items are used. In NESARC Wave 2, 15.5% of the sample diagnosed with AUD when the original “either or” criteria were used. For both *craving* and *cut down* the low threshold was the same, but using only the high *craving* threshold, the percentage drops to 15.1%, and using the high *cut down* threshold it drops to 13.1%. In NESARC III, 19.2% of the sample diagnosed with AUD, and moving to the low threshold only for both items this is reduced to 19.1%. Using only the high threshold *craving* and *cut down* items the percentage drops to 17.8% and 17.2% respectively.

Looking specifically at the edges related to the modified criteria, we begin to see the impact of the varying criteria thresholds. Figure 1 depicts the networks estimated from the full sample, and the relevant edges are colored in blue. Here, the changes in the thickness and darkness of the edges between the networks using the high and low threshold are notable. Table 1 contains the estimates of the network relations from the full data, as well as the mean and standard deviation of the bootstrapped estimates, and the difference between the high and low thresholds. The edges in the table are ordered by their prevalence weights averaged over both data sets, in descending order. A broad trend can be seen here in the differences between the high and low threshold edges for both modified criteria: the more prevalent criteria tend to have stronger connections to the low threshold items, and the rarer criteria are more associated with those with higher thresholds.

Considering the modification of the *craving* criterion, the low and high threshold networks edges are correlated $ICC=.25$ (confidence interval= -.47,.75) and $ICC= -.09$ (confidence interval= -.69,.56) for NESARC Wave 2 and III. These correlations indicate little similarity in the estimated edges, but the confidence intervals reflect a considerable amount of uncertainty for these estimates (based on only ten values). For this reason, we consider t-tests of the bootstrapped edges (with a Bonferroni correction of the significance for ten estimates). Except for *social problems* in NESARC III, the t-tests for the differences are all significant. The largest increases from the high to the low threshold in both data sets are the edges to the three most prevalent criteria: *larger/longer*, *cut down*, and *hazardous use*. Conversely, the largest decreases (indicating stronger relation to the high threshold item) in both data sets are found in the least prevalent items: *give up activities* and *role interference*.

The correlations for the edges estimated with different thresholds for *cut down* are $ICC=.74$ (confidence interval=.27,.93) for NESARC Wave 2 and $ICC=.89$ (confidence interval=.62,.97) for NESARC III. In the NESARC Wave 2 data, the results continue the general trend mentioned before, with higher relations between the low threshold and more prevalent items,

with all differences significant except the edges to *craving*, *phys/psych* and *role interference*. In NESARC-III there is not as clear of a trend in terms of the sign of the changes. The edges to *withdrawal*, *craving*, and *social problems* are all not significantly different. We do see the same trend for the prevalent *hazardous use* criterion, where it relates more strongly to the lower threshold item, and the rarer *give up activities* criterion is again more related to the higher threshold *cut down* statement.

The centrality measures showed little variation between the networks overall. This was anticipated, as the centrality measures would be expected to be fairly stable when using the same data with only one node varying at a time. Though small, the difference does appear to always favor low threshold items as being more central, as can be seen in the estimated centrality statistics in Table 2. A better evaluation of the effects of symptom thresholds on centrality measures would involve permuting the thresholds of all criteria systematically and estimating the distribution of each centrality parameter across these permutations. We are prevented from doing this, however, in that in the AUDADIS, there are not multiple thresholds for each criterion that would display the Guttman-like qualities needed.

Discussion

Since the dawn of the modern diagnostic era and the introduction of DSM-III, research on psychiatric diagnosis at the formal diagnostic level, the latent variable level, and the network level of analysis appear to tacitly assume that, for all intents and purposes, that the diagnostic instrument employed is not a critical factor in research. Although the sequence of steps involved in translating a set of diagnostic criteria into a structured interview have been thoughtfully characterized (Robins & Cottler, 2004), research on differences across diagnostic instruments has tended to focus at the syndromal level (assessed categorically or by criteria counts; e.g., Grant et al., 2015) and not the symptom/criterion level (Samet et al., 2007). Lane et al.'s (2016) demonstration of large variation in the relative thresholds of different AUD criteria across diagnostic instruments highlights the potential importance of considering individual criterion or symptom thresholds when examining symptom network structure. Consequently, examining the robustness of symptom networks across alternative operationalizations of symptoms would seem like a high priority for assessing reproducibility of findings and having confidence in the generalization of findings.

For both *cut down* and *craving*, when we vary the threshold for determining its presence or absence, we observe different patterns of interaction with the other symptoms of AUD. The lower threshold *craving* statement has a much higher prevalence rate, and when it is used in network analyses the relationships tend to be stronger with other criteria that are more common. When only the higher threshold *craving* statement is used, the criterion becomes much rarer and the relationships with other low prevalence rate criteria are strengthened, while ties to the more common items are weakened. This trend might be expected given the nature of the high and low thresholds. The high threshold items indicate a more severe degree of the criterion, which likely reflects a more severe disorder. *Giving up activities*, a rarer criterion, characterizes someone who is choosing alcohol over other important or pleasurable activities. This would reflect a greater degree of impact on their lives than the more common criterion, *hazardous use*, which identifies those whose drinking has led them

into risky situations. As demonstrated in our analyses, *giving up activities* is more closely associated with both higher thresholds of *cut down* and *craving*, and *hazardous use* more associated with lower.

While our primary motivation for undertaking the analyses reported here was methodological, it is important to consider that there may be substantive interpretations to these findings. That is, the true relational structure of symptoms may depend upon the severity of the symptom as operationalized. Low threshold symptom X_L might only be related to symptom Y because of its association with a high threshold manifestation of symptom X_H ; that is, there is no “direct association” between X_L and Y if the path from X_L to X_H is modeled. For network models like those in this study, such potential issues could be addressed by modeling the networks with alternative methods that do not limit the analyses to binary data (e.g. the polychoric correlation graphs in *qgraph*, or mixed graphical models, *MGM*, Hasselback & Waldorp, 2015). However, most research utilizes the binary outcome variables to assess criteria and disorders, implicitly assuming that alternative operationalizations of a criterion are functionally interchangeable. Additionally, symptom-level bias in the assessment of criteria associated with a range of covariates (e.g., differential item functioning) could distort estimates of edge weights and centrality measures. These considerations highlight the importance of not taking symptom and criterion assessments at face value and of considering both the severity of symptoms overall and population mixtures that could differentially bias such severity.

Until now we have only considered the differences between relative thresholds of the examined criteria. The identification of a Guttman-like variable was accomplished using the crosstabulations of the sub-items, finding cases where there are a large number of people who report the “low threshold” only and very few who report only the “high threshold.” However, these interview items were not developed as true Guttman scales, and in the cases examined here we only have two levels for comparison. What happens between these two “thresholds” is unknown, and thus the findings here could reflect either categorical or continuous differences in the “severity” of the symptoms. For example, the actual attempts to cut down on drinking, and particularly their subsequent failures, might have social consequences that the desire alone would not have. Indeed, in our analysis the *social/interpersonal* problems criterion was more strongly associated with the higher threshold *cut down* item. This might indicate that the two subitems are reflecting categorically different symptoms. If the underlying construct is continuous, however, this might reflect instead a critical threshold of sorts where the increasing severity of one symptom gets to a point where it begins to cause other problems. While our analyses are limited in what we can say about these possibilities, they indicate the potential for further investigation into the individual criterion definitions.

Extending beyond the comparison here, it is important to highlight there are potentially numerous differences across instruments, each of which could affect the type of finding reported here or findings as basic as prevalence. It has been previously demonstrated that the simple change in the structure of the survey likely resulted in an increase of lifetime AUD prevalence from 18% to 30% largely due to how it affected responding to the *hazardous use* item (Verges et al., 2011). Other survey instruments assess the same criteria using entirely

different wording. The results presented here reinforce the need for a more critical look at variation across diagnostic instruments and the potential for generalizations across studies beyond their effect on obtained network structures.

While we demonstrated the effect of different AUD criteria wording using network analysis, it is likely that the same issues will plague other multivariate methods that are used to analyze multidimensional binary data sets. For instance, it is highly likely that the thresholding would have an influence on both the number of classes and class structure of a latent class analysis. Additionally, varying the thresholds would influence the representation achieved via correspondence analysis. Likewise, we would expect sensitivity in the results of factor models that used different thresholds to operationalize the criteria. As such, this should serve as a siren's call to take care in how criteria are assessed and operationalized and further supports the notion that sophisticated models are not able to “fix” shortcomings at the assessment level.

Although the focus of the current analyses is on AUD, as noted above, a similar situation appears to be present for major depression. There is reason to suspect that variation in operationalization of diagnostic criteria is relevant for many if not most psychiatric conditions. While the effects of using more or less “severe” operationalizations on prevalence is obvious, the importance of variation in the operationalization of criteria is less obvious but profoundly important to understanding the associations observed *among* symptoms/criteria. Consequently, not only must we not assume generalizability across studies utilizing different diagnostic instruments, but we should more carefully consider the wide variation in diagnostic procedures that characterize modern research. As noted by Lane et al. (2016), variation in diagnostic instruments that are foundational to much psychiatric research is large and unappreciated, and even the recovery of similar psychometric structures across studies can reflect mere “skin deep” reproducibility.

Acknowledgments

Support for this research was provided by NIH grants R01 AA024133, K05 AA017242, and T32 AA 013526, to Kenneth J. Sher, and R01 AA023248 to Douglas Steinley. In addition, we acknowledge the contribution of NIAAA's Intramural Research Program who designed, fielded, curated the NESARC-III data base and provided the limited-access data used in the present analyses.

References

- American Psychiatric Association. Diagnostic and statistical manual (DSM-III). Washington, DC: American Psychiatric Association; 1980.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, D.C: American Psychiatric Association; 2013.
- Anker JJ, Forbes MK, Almquist ZW, Menk JS, Thuras P, Unruh AS, Kushner MG. A Network Approach to Modeling Comorbid Internalizing and Alcohol Use Disorders. *Journal of abnormal psychology*. 2017
- Borsboom D, Cramer AO. Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*. 2013; 9:91–121.
- Boschloo L, van Borkulo CD, Borsboom D, Schoevers RA. A Prospective Study on How Symptoms in a Network Predict the Onset of Depression. *Psychotherapy and Psychosomatics*. 2016; 85(3):183–184. [PubMed: 27043457]

- Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI Jr, Schuckit MA. 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*. 1994; 55(2):149–158. [PubMed: 8189735]
- Epskamp S, Borsboom D, Fried EI. Estimating psychological networks and their accuracy: a tutorial paper. *arXiv preprint arXiv*. 2016:1604.08462.
- Epskamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*. 2012; 48(4):1–18.
- Fried EI, van Borkulo CD, Cramer AO, Boschloo L, Schoevers RA, Borsboom D. Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*. 2017; 52(1):1–10. [PubMed: 27921134]
- Fruchterman TM, Reingold EM. Graph drawing by force-directed placement. *Software: Practice and experience*. 1991; 21(11):1129–1164.
- Grant BF, Dawson DA, Hasin DS. The Alcohol Use Disorders and Associated Disabilities Interview Schedule—Version for DSM-IV (AUDADIS-IV). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2001.
- Grant BF, Goldstein RB, Chou SP, et al. The Alcohol Use Disorder and Associated Disabilities Interview Schedule—Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition Version (AUDADIS-5). Rockville, MD: National Institute on Alcohol Abuse and Alcoholism; 2011.
- Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Hasin DS. Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions III. *JAMA psychiatry*. 2015; 72(8):757–766. [PubMed: 26039070]
- Grant BF, Goldstein RB, Smith SM, Jung J, Zhang H, Chou SP, et al. Aivadyan C. 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*. 2015; 148:27–33. [PubMed: 25595052]
- Grant BF, Kaplan K, Moore T, Kimball J. 2004–2005 Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions: Source and Accuracy Statement. Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2007.
- Grant BF, Peterson A, Dawson DA, Chou SP. Source and accuracy statement for the National Longitudinal Alcohol Epidemiologic Survey (NLAES). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 1994.
- Haslbeck J, Waldorp LJ. mgm: Structure Estimation for time-varying mixed graphical models in high-dimensional data. *arXiv preprint arXiv*. 2015:1510.06871.
- Hofmann SG, Curtiss J, McNally RJ. A complex network perspective on clinical science. *Perspectives on Psychological Science*. 2016; 11(5):597–605. [PubMed: 27694457]
- Karlsson L, Marttunen M, Karlsson H, Kaprio J, Hillevi A. Minor change in the diagnostic threshold leads into major alteration in the prevalence estimate of depression. *Journal of affective disorders*. 2010; 122(1):96–101. [PubMed: 19615754]
- Kessler RC, Calabrese JR, Farley PA, Gruber MJ, Jewell MA, Katon W, et al. Shillington AC. Composite International Diagnostic Interview screening scales for DSM-IV anxiety and mood disorders. *Psychological medicine*. 2013; 43(08):1625–1637. [PubMed: 23075829]
- Kramer JR, Chan G, Hesselbrock VM, Kuperman S, Bucholz KK, Edenberg HJ, et al. Bierut LJ. A principal components analysis of the abbreviated Desires for Alcohol Questionnaire (DAQ). *Journal of studies on alcohol and drugs*. 2010; 71(1):150–155. [PubMed: 20105425]
- Lane SP, Sher KJ. Limits of Current Approaches to Diagnosis Severity Based on Criterion Counts An Example With DSM-5 Alcohol Use Disorder. *Clinical Psychological Science*. 2015; 3:819–835. [PubMed: 26783505]
- Lane SP, Steinley D, Sher KJ. Meta-analysis of DSM alcohol use disorder criteria severities: structural consistency is only ‘skin deep’. *Psychological medicine*. 2016; 46(8):1769. [PubMed: 27019218]
- Robins Lee N, Cottler Linda B. Making a Structured Psychiatric Diagnostic Interview Faithful to the Nomenclature. *Am J Epidemiol*. 2004; 160(8):808–813. DOI: 10.1093/aje/kwh283 [PubMed: 15466503]
- Love A, James D, Willner P. A comparison of two alcohol craving questionnaires. *Addiction*. 1998; 93:1091–1102. [PubMed: 9744139]

- Rhemtulla M, Fried EI, Aggen SH, Tuerlinckx F, Kendler KS, Borsboom D. Network analysis of substance abuse and dependence symptoms. *Drug and alcohol dependence*. 2016; 161:230–237. [PubMed: 26898186]
- Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, Norquist GS. Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy. *Archives of General Psychiatry*. 1998; 55(2):109–115. [PubMed: 9477922]
- Samet S, Waxman R, Hatzenbuehler M, Hasin DS. Assessing addiction: Concepts and instruments. *Addiction science & clinical practice*. 2007; 4(1):19. [PubMed: 18292706]
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86(2):420. [PubMed: 18839484]
- van Borkulo CD, Borsboom D, Epskamp S, Blanken TF, Boschloo L, Schoevers RA, Waldorp LJ. A new method for constructing networks from binary data. *Scientific Reports*. 2014; 4
- van Borkulo C, Epskamp S, van Borkulo MC. Package ‘IsingFit’. 2016
- Vergés A, Littlefield AK, Sher KJ. Did lifetime rates of alcohol use disorders increase by 67% in ten years? A comparison of NLAES and NESARC. *Journal of Abnormal Psychology*. 2011; 120:468–477.
- Wasserman S, Faust K. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press; 1994.

Figure 1.

Graphs of the networks created using the qgraph package (Epskamp et al., 2012). This utilizes the Fruchterman & Reingold (1991) layout method to place the nodes. This layout was made to be optimal for NESARC Wave 2 and NESARC – III data, and held constant for all of the graphs. In the networks, thicker and darker lines are used to indicate larger edges (stronger relationships). The relevant edges for each graph are colored in blue shades, and the remaining edges in grey shades.

Table 1

Prevalence rates and results of the analyses. The columns labeled “Estimate” contain the edges estimated using the entire sample. The “Bootstrap” columns contain the mean and standard deviation of the bootstrap estimates of each edge. The “Difference” columns are calculated by subtracting the edge estimated for the high threshold from the edge estimated from the low threshold.

Craving										Cut Down			
Sample	Low Threshold		High Threshold		Difference	Low Threshold		High Threshold		Difference			
	Estimate	Bootstrap	Estimate	Bootstrap		Estimate	Bootstrap	Estimate	Bootstrap				
NESARC Wave 2													
Larger/longer	13.8%	0.92	.89(.16)	0.83	.69(.53)	.20 *	1.22	1.18(.09)	1.10	1.03(.23)	.15 *		
Cut down	12.8%	0.83	.81(.15)	0	.19(.32)	.62 *	-	-	-	-	-		
Hazardous use	10.6%	0.63	.61(.14)	0	0(0)	.61 *	0.46	.44(.09)	0	.00(.05)	.44 *		
Withdrawal	7.8%	0.66	.64(.15)	0.94	.81(.40)	.18 *	0.67	.66(.12)	0.60	.56(.22)	.09 *		
Craving	4.2%	-	-	-	-	-	0.80	.78(.16)	0.81	.81(.25)	-.02		
Tolerance	6.8%	0.30	.28(.18)	0.47	.41(.33)	.14 *	1.28	1.25(.10)	0.67	.71(.18)	.53 *		
Phys/psych	5.1%	0.96	.98(.17)	0.73	.76(.48)	.22 *	1.21	1.20(.14)	1.16	1.17(.20)	.03		
Social problems	2.4%	0.84	.81(.22)	1.12	1.03(.44)	-.22 *	0.46	.39(.26)	0.93	.90(.22)	-.51 *		
Time spent	2.8%	0.96	.95(.22)	0.64	.67(.40)	.28 *	0.26	.22(.23)	0	.11(.16)	.11 *		
Role interference	1.1%	0	.05(.14)	0.47	.43(.35)	-.38 *	0	.12(.26)	0.22	.18(.24)	-.06		
Give up activities	1.0%	0.62	.62(.37)	0.79	.85(.40)	-.23 *	0.69	.57(.43)	0.76	.75(.38)	-.18 *		
NESARC III													
Larger/longer	14.8%	1.10	1.10(.11)	0.46	.25(.33)	.85 *	1.28	1.26(.09)	1.42	1.36(.19)	-.10 *		
Cut down	13.8%	0.56	.55(.10)	0.22	.20(.24)	.35 *	-	-	-	-	-		
Hazardous use	11.2%	0.74	.73(.11)	0	-.01(.09)	.74 *	0.31	.32(.10)	0	-.04(.14)	.36 *		
Withdrawal	10.9%	0.86	.88(.10)	0.81	.80(.28)	.08 *	0.42	.40(.11)	0.48	.46(.20)	-.06		
Craving	10.9%	-	-	-	-	-	0.57	.55(.11)	0.55	.57(.21)	-.02		
Tolerance	8.0%	0.46	.44(.12)	0.28	.23(.20)	.21 *	0.97	.96(.10)	0.66	.67(.15)	.29 *		
Phys/psych	7.3%	0.70	.68(.15)	0.88	.90(.27)	-.22 *	0.72	.72(.12)	0.78	.78(.17)	-.06 *		
Social problems	6.4%	0.78	.80(.15)	0.77	.74(.23)	.05	0.63	.61(.12)	0.70	.65(.19)	-.05		

Cut Down											
Craving											
	Sample	<u>Low Threshold</u>			<u>High Threshold</u>			<u>Difference</u>			<u>Difference</u>
		Estimate	Bootstrap	Estimate	Bootstrap	Estimate	Bootstrap	Estimate	Bootstrap		
Time spent	5.3%	0.84	.85(.15)	0.44	.43(.26)	.41*	0	.09(.13)	0	.03(.08)	.05*
Role interference	2.1%	0.63	.57(.28)	1.12	1.08(.24)	-.52*	0.29	.19(.22)	0.23	.29(.21)	-.09*
Give up activities	2.0%	0.46	.39(.29)	1.01	1.03(.25)	-.64*	0.53	.41(.27)	0.87	.83(.21)	-.43*

Note:

* p<.005 for the t-test of the difference in bootstrapped edges.

Table 2

The centrality statistics of the nodes for the symptoms craving (Crave) and inability to cut down (Cut Down). The unstandardized centrality statistics are always lower (or equal) for the criteria calculated using the higher threshold.

	Low		High	
	Standardized	Unstandardized	Standardized	Unstandardized
<u>Crave</u>				
Betweenness	-0.74	0	-0.69	0
Closeness	-0.75	0.07	-1.06	0.07
Strength	-0.41	6.71	-0.80	6.00
<u>Cut Down</u>				
Betweenness	0.59	4	-0.47	0
Closeness	0.03	0.08	-0.37	0.08
Strength	-0.21	7.04	-0.67	6.25