The Case for Cognitive Interviewing in Survey Item Validation: A Useful Approach for Improving the Measurement and Assessment of Substance Use Disorders

CASSANDRA L. BONESS, M.A.,[a,*] & KENNETH J. SHER, PH.D.[a]

[a]*Department of Psychological Sciences, University of Missouri, Columbia, Missouri*

ABSTRACT. Objective: To accurately identify substance use disorders, we must be confident of our ability to define and measure the construct itself. To date, research has demonstrated that the ways in which substance use disorder criteria are operationalized or assessed can significantly affect the information we obtain from these diagnoses. For example, differing operationalizations of the same construct, such as impaired control over substance use, can result in markedly different estimates of prevalence. This points to the need for approaches that aim to improve the validity of diagnostic assessments during the measure development phase. **Method:** We performed a scoping review of the cognitive interviewing literature, a technique that aims to provide a systematic way of identifying and reducing measurement error associated with the structure and content of assessment items. Along with this, we apply cognitive interviewing to items assessing alcohol tolerance. **Results:** We argue that cognitive interviewing is well suited for reducing measurement error in substance use disorder assessment items. **Conclusions:** Incorporating cognitive interviewing into the item generation stage of measure development for substance use disorder assessments is a worthwhile endeavor for improving validity. (*J. Stud. Alcohol Drugs, 81,* 401–404, 2020)

CONSTRUCT VALIDATION is central to the research process (Cronbach & Meehl, 1955). In addiction research, our ability to accurately assess and diagnose substance use disorders (SUDs) is largely based on the ability to define and measure the construct of interest (e.g., Flake & Fried, 2019). Decades of research have demonstrated that the definition and measurement of SUDs is neither clear nor straightforward. For example, research has demonstrated that the way in which symptoms of psychological disorders, such as SUDs, are operationalized can greatly affect the validity of diagnosis. This was clearly demonstrated by Lane et al. (2016) in a meta-analysis of alcohol use disorder criteria item response theory parameters that demonstrated the relative severity of a given criterion can vary dramatically among commonly employed diagnostic interviews. Further, differing operationalizations of the same construct (e.g., impaired control over alcohol use) can greatly affect overall prevalence rates (Boness et al., 2019) and can influence the estimation of diagnostic structure (e.g., Hoffman et al., 2018). In addition, there is evidence that items used to assess certain SUD symptoms, such as alcohol withdrawal, may be misunderstood and result in high false positive rates for that criterion (with associated false positive diagnoses; Boness et al., 2016).

Together, findings such as these suggest that it is imperative to consider the way in which a diagnostic instrument operationalizes a given criterion, as we might not be measuring what we think we are measuring. This can result in unknown or poor construct validity of our assessments. Unknown construct validity has far-reaching consequences for addressing the significant impact of SUDs in society. In research, uncertain construct validity for SUD measures means we lack the information necessary to evaluate the validity of a study's outcomes and conclusions. In clinical practice, the inability to accurately assess SUDs can result in a failure to identify those with the disorder, thus resulting in a lack of treatment or assignment to a treatment that is not effective. Therefore, greater attention to the construct validity of self-report items, particularly during the measure development stage of SUD assessment instruments and surveys, should be an extremely high priority for both research and practice.

What is cognitive interviewing?

One viable approach for improving the validity of diagnostic assessment is through the use of cognitive interviewing (sometimes called cognitive testing). Cognitive interviewing is a technique that aims to provide a systematic way to identify and reduce measurement error associated with the structure and content of interview items and is an important first step in measure development (Schwarz, 1999). It has the goal of understanding how respondents interpret a given item and select a response (i.e., the response process) and can be used to provide feedback on the candidate items and response scales, suggest alternative wording, and ensure relevant constructs are validly represented. Cognitive interviewing is a flexible approach that can be used across varied survey administration modes such as in-person, telephone, and online (see Willis, 2005).

Received: March 9, 2020. Revision: March 30, 2020.

This study was supported by National Institutes of Health Grant F31AA026177. The funders played no role in the current review.

*Correspondence may be sent to Cassandra Boness at the Department of Psychological Sciences, 146 Psychology Building, 200 South Seventh Street, Columbia, MO 65211, or via email at: clmkdb@mail.missouri.edu.

Cognitive interviews are typically conducted by trained interviewers who are familiar with the constructs of interest and measurement objectives. Interviewers can use several interviewing techniques including think-aloud and verbal probing procedures. The think-aloud procedure instructs participants to vocalize their thoughts as they select their response to an item. The interviewer pays careful attention to the response to identify misunderstandings or item flaws. With this procedure, open-ended questions with minimal bias in phrasing are used to clarify answers (e.g., "tell me more about that"). In the verbal probing procedure, the interviewer asks the target question and the participant responds. The interviewer then follows up by probing for specific information, such as asking participants to (a) paraphrase the question, (b) discuss the thought processes used to choose an answer (e.g., what kinds of situations were you thinking about in coming up with your answer?), and/or (c) define key terms (e.g., what is a hangover?) (Fowler, 1995). Probing can be concurrent or retrospective and probes can be standardized (i.e., predetermined for every participant) or nonstandardized (i.e., specific to each participant). Thus, there is an immense amount of flexibility with verbal probing (see Willis, 2005, for a thorough discussion).

Cognitive interviewing has been largely overlooked in psychology and the study of psychopathology. A quick search of the literature revealed only a few publications that used cognitive interviewing for addressing issues related to the assessment of SUDs (e.g., Chung & Martin, 2005; Mewton et al., 2014; Slade et al., 2013 [the latter two used the same sample]). These publications, focused on adolescents and young adults, highlighted several common areas of misunderstanding related to the assessment of SUDs. One example is impaired control, which is typically conceptualized as the result of addiction-related compulsion. This is often assessed via criteria such as drinking larger amounts or longer than intended. Young adults, although endorsing drinking larger amounts or longer than intended, frequently reported that this behavior was due to social or other noncompulsion-based reasons (e.g., Chung & Martin, 2005; Slade et al., 2013), suggesting false positives for impaired control among this age group. Tolerance is another criterion that is likely to be misunderstood, particularly among adolescents. Chung and Martin (2005) demonstrated that the extent of tolerance reported by adolescents was more reflective of expected changes in sensitivity to the effects of alcohol rather than a clinically significant increase in consumption that is typically suggestive of physiological dependence. Although important research with significant implications, none of these publications were used in the context of item-refinement for the development of an assessment instrument, but rather to elucidate the problems with existing instruments (e.g., World Mental Health Survey Initiative version of the Composite International Diagnostic Interview; Kessler & Ustün, 2004) and, in some cases, make suggestions for revisions (e.g., to DSM-5 diagnostic criteria; Slade et al., 2013). Thus, cognitive interviewing has been largely neglected as a part of SUD measure development.

Despite this oversight, we believe that cognitive interviewing should be considered as a part of the standard measure development process.[1] Guidelines for the construction of objective tests (e.g., Clark & Watson, 2019) recognize the importance of conceptual and psychometric analysis during the item generation stage of scale construction. However, clinical psychology has leaned heavily toward psychometric analyses using approaches such as factor analysis and item response theory. Although this is crucially important, psychometric analysis alone overlooks the possibility of conceptual problems with items. Thus, we argue that cognitive interviewing is an ideal complement to psychometric approaches in the item generation stage of measure construction and is necessary if we wish to avoid building our diagnostic structures on weak foundations.

Using cognitive interviewing to improve the construct validity of SUD items

Given some of the issues unique to the diagnosis of SUDs (e.g., the difficulty in assessing alcohol withdrawal; Boness et al., 2019), cognitive interviewing is especially well suited for dealing with these challenges. This approach allows the researcher to gain a deeper understanding of how respondents are interpreting items, including any areas in need of refinement or additional clarification. It can also aid in the evaluation of measure instructions and response scales, providing feedback that is often treated as less important to scale construction but that, in practice, can have a significant impact on the usefulness of a scale (e.g., Simms et al., 2019). Further, we argue that SUD researchers are particularly well suited to carry out these interviews given their substantive expertise in the construct of interest.

An example of cognitive interviewing with alcohol tolerance

To offer an applied example, we conducted cognitive interviews with a small group of participants ($N = 10$) to evaluate how tolerance items from the National Epidemiologic

[1]Although other qualitative approaches (e.g., focus groups) can certainly be used for the same purpose of revising items, we believe that cognitive interviewing offers several unique benefits such as its flexible application to different types of surveys and its emphasis on gaining unbiased information from participants. In focus groups, for example, it is well known that biases can arise from the presence of other participants in the group (e.g., Vogt et al., 2004). This is an important consideration when asking about SUDs, which can be difficult for participants to talk about openly (e.g., due to stigma). At a minimum, cognitive interviewing could be more readily combined with other qualitative approaches in the measure development process.

Survey on Alcohol and Related Conditions-III (NESARC-III; Grant et al., 2014) perform and to recommend revisions. NESARC-III was chosen because it is used to derive population-based estimates of DSM-5 SUD diagnoses. Participants were recruited from the community through fliers posted around a Midwest town and online. They had to be 18 or older and have met heavy drinking criteria[2] within the past 30 days to be eligible. Participants were largely female (90%) and had an average age of 20.7 (*SD* = 1.6), and all had graduated high school. Items comprised the following: (a) did you find that your usual number of drinks had much less effect on you than it once did; (b) did you find that you had to drink much more than you once did to get the effect you wanted; (c) did you drink as much as a fifth of liquor in one day, that would be about 20 drinks, or 3 bottles of wine, or as much as 3 six-packs of beer in a single day; and (d) did you increase your drinking because the amount you used to drink didn't give you the same effect anymore? All questions referred to the past 12 months, and the response scale was "yes" or "no" for each item. Interviews were conducted using a combination of the think-aloud and verbal probe procedures.

The most notable issue that arose across the items was related to the time frame of comparison, which varied markedly across participants. For example, when asked, "In the past 12 months, did you find that your usual number of drinks had much less effect on you than it once did?" some participants reported that they were comparing their usual number of drinks now to when they first started drinking. However, others reported that they were comparing their usual number of drinks now to their freshman year of college (i.e., several years earlier), to the beginning of the current year (i.e., a few months before the interview), to one year ago, and to the start of the academic year (i.e., 4–5 months earlier). This is problematic given that respondents were answering on the basis of different time frames. Misinterpretation of items across respondents is thought to affect the estimation of population-based prevalence rates (e.g., Boness et al., 2016) and could contribute, in part, to the so-called treatment gap (Drummond et al., 2011), the difference between estimated prevalence in the population and treatment utilization.

In line with the use of cognitive interviewing for item refinement, we feel most of the tolerance items could be revised to more clearly define the time frame of interest in order to increase consistency in reporting and, thus, improve validity. As such, for the item described above, we recommended the following revision for consideration: "Compared to when you first started drinking regularly, do you find that a given number of drinks has much less effect on you now?"[3] Although cognitive interviewing is useful for identifying potential pitfalls and revising items accordingly, it is only intended to provide information about items rather than to ensure validity (Willis, 2005).

Limitations of cognitive interviewing and alternative approaches

Limitations of cognitive interviewing include being time- and labor-intensive and requiring specific training in interviewing. Some have argued against its use on these grounds alone, which is problematic given that various solutions have been offered to address issues related to the time- and labor-intensity required. For example, the Response Process Evaluation method uses open-ended metasurveys to more rapidly gather information about how participants interpret items, make revisions, and retest the revision across new samples of respondents (Wolf et al., 2019). This is similar to cognitive interviewing but is more efficient for some applications. Other techniques, such as the Questionnaire Appraisal System Checklist (Willis, 2005), have been developed as simpler and more efficient alternatives to full cognitive interviews. These offer many of the benefits, such as the ability to identify items with potential pitfalls (e.g., making inappropriate assumptions about respondents), while decreasing some of the mastery required to conduct cognitive interviews.

Others (e.g., Willis & Artino, 2013) have argued that cognitive interviewing, particularly the think-aloud procedure, is an unnatural and a difficult process for respondents to engage in given that it is not a typical way of processing information. It can, therefore, require significant participant training before the interview. Although this criticism may be true for some participants and for some questions, there are few other options for gaining insight into the response process. Even with the Response Process Evaluation method, respondents are asked to report on their response process. Further, some have maintained that cognitive interviewing could cause survey developers to overthink their items. This potential problem is easily addressed by focusing on the identification of clear trends across respondents before making changes to any given item (Willis, 2005). In addition, this criticism requires reflection on the balance between under- and over-consideration of item qualities. We would argue that overthinking items is worth the risk when compared with the alternative of not considering how items are performing at all for fear of "overthinking."

A final issue to consider is that items may have different meanings and interpretations for different populations. Thus, it is imperative to consider the identification and recruit-

[2]Heavy drinking was defined as (a) having at least five separate heavy drinking occasions (i.e., 5+ drinks for males; 4+ drinks for females) OR (b) exceeding NIAAA's weekly limit guidelines (i.e., >14 drinks for males; >7 drinks for females; U.S. Department of Health and Human Services & U.S. Department of Agriculture, 2015).

[3]Importantly, though, these findings were only from one round of cognitive interviewing. In practice, several rounds of interviewing (e.g., three rounds of 10–15 participants) would be conducted to test item revisions iteratively.

ment of participants from appropriate subpopulations for testing items. In most cases, respondents should have the characteristics of interest for the survey. However, in some cases, such as with SUD diagnosis, the questionnaire may be intended for use in the general population. Thus, recruiting participants with and without the construct of interest (e.g., heavy drinking) is useful for ensuring that the majority of individuals who will be administered the items do, in fact, understand them (Willis, 2005).

Recommendations

Thus, we recommend that cognitive interviewing become more common place in development of SUD assessment instruments. Concretely, we suggest that cognitive interviewing be incorporated in the item generation stage of any new scale development or measure construction undertaking. Further, we believe that cognitive interviewing, or similar methods, should be used during the construction of any new assessment instrument because it offers particular benefits given the complexity of assessing SUD. Improved validity of SUD diagnosis can advance the precision of prevalence estimates derived via epidemiologic surveys as well as aid in identifying those at risk and in need of treatment.

References

Boness, C. L., Lane, S. P., & Sher, K. J. (2016). Assessment of withdrawal and hangover is confounded in the Alcohol Use Disorder and Associated Disabilities Interview Schedule: Withdrawal prevalence is likely inflated. *Alcoholism: Clinical and Experimental Research, 40,* 1691–1699. doi:10.1111/acer.13121

Boness, C. L., Lane, S. P., & Sher, K. J. (2019). Not all alcohol use disorder criteria are equally severe: Toward severity grading of individual criteria in college drinkers. *Psychology of Addictive Behaviors, 33,* 35–49. doi:10.1037/adb0000443

Chung, T., & Martin, C. S. (2005). What were they thinking? Adolescents' interpretations of DSM-IV alcohol dependence symptom queries and implications for diagnostic validity. *Drug and Alcohol Dependence, 80,* 191–200. doi:10.1016/j.drugalcdep.2005.03.023

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31,* 1412–1427. doi:10.1037/pas0000626

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. doi:10.1037/h0040957

Drummond, C., Gual, A., Goos, C., Godfrey, C., Deluca, P., Von Der Goltz, C., . . . Kaner, E. (2011). Identifying the gap between need and intervention for alcohol use disorders in Europe. *Addiction, 106, Supplement 1,* 31–36. doi:10.1111/j.1360-0443.2010.03335.x

Flake, J. K., & Fried, E. I. (2019, January 17). *Measurement schmeasurement: Questionable measurement practices and how to avoid them.* Retrieved from https://psyarxiv.com/hs7wm/

Fowler, F. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Grant, B. F., Amsbary, M., Chu, A., Sigman, R., Kali, J., Sugawana, Y., & Chou, P. S. (2014). *Source and accuracy statement: National Epidemiologic Survey on Alcohol and Related Conditions-III (NESARC-III).* Rockville, MD: National Institute on Alcohol Abuse and Alcoholism.

Hoffman, M., Steinley, D., Trull, T. J., & Sher, K. J. (2018). Criteria definitions and network relations: The importance of criterion thresholds. *Clinical Psychological Science, 6,* 506–516. doi:10.1177/2167702617747657

Kessler, R. C., & Ustün, T. B. (2004). The world mental health (WMH) survey initiative version of the World Health Organization (WHO) composite international diagnostic interview (CIDI). *International Journal of Methods in Psychiatric Research, 13,* 93–121. doi:10.1002/mpr.168

Lane, S. P., Steinley, D., & Sher, K. J. (2016). Meta-analysis of DSM alcohol use disorder criteria severities: Structural consistency is only 'skin deep.' *Psychological Medicine, 46,* 1769–1784. doi:10.1017/S0033291716000404

Mewton, L., Slade, T., Teesson, M., Memedovic, S., & Krueger, R. F. (2014). Improving the diagnostic criteria for alcohol use disorders through survey methodology and cognitive interviewing. *International Journal of Methods in Psychiatric Research, 23,* 359–371. doi:10.1002/mpr.1448

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *The American Psychologist, 54,* 93–105. doi:10.1037/0003-066X.54.2.93

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31,* 557–566. doi:10.1037/pas0000648

Slade, T., Teesson, M., Mewton, L., Memedovic, S., & Krueger, R. F. (2013). Do young adults interpret the DSM diagnostic criteria for alcohol use disorders as intended? A cognitive interviewing study. *Alcoholism: Clinical and Experimental Research, 37,* 1001–1007. doi:10.1111/acer.12063

U.S. Department of Health and Human Services & U.S. Department of Agriculture. (2015). *Dietary guidelines for Americans.* Retrieved from http://health.gov/dietaryguidelines/2015/guidelines/

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16,* 231–243. doi:10.1037/1040-3590.16.3.231

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Willis, G. B., & Artino, A. R., Jr. (2013). What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. *Journal of Graduate Medical Education, 5,* 353–356. doi:10.4300/JGME-D-13-00154.1

Wolf, M. G., Ihm, E. D., Maul, A., & Taves, A. (2019, July 23). *Survey item validation.* Retrieved from https://doi.org/10.31234/osf.io/k27w3