Published in final edited form as:
*Drug Alcohol Depend*. 2013 June 1; 130(0): 167–177. doi:10.1016/j.drugalcdep.2012.11.002.

# Item Banks for Alcohol Use from the Patient-Reported Outcomes Measurement Information System (PROMIS): Use, Consequences, and Expectancies

**Paul A. Pilkonis**[a,*], **Lan Yu**[a], **Jason Colditz**[a], **Nathan Dodds**[a], **Kelly L. Johnston**[a], **Catherine Maihoefer**[a], **Angela M. Stover**[a], **Dennis C. Daley**[a], and **Dennis McCarty**[b]

[a]Department of Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, PA 15213

[b]Department of Public Health and Preventive Medicine, Oregon Health and Science University, Portland, OR 97239

## Abstract

**Background—**We report on the development and calibration of item banks for alcohol use, negative and positive consequences of alcohol use, and negative and positive expectancies regarding drinking as part of the Patient-Reported Outcomes Measurement Information System (PROMIS).

**Methods—**Comprehensive literature searches yielded an initial bank of more than 5,000 items from over 200 instruments. After qualitative item analysis (including focus groups and cognitive interviewing), 141 items were included in field testing. Items for alcohol use and consequences were written in a first-person, past-tense format with a 30-day time frame and 5 response options reflecting frequency. Items for expectancies were written in a third-person, present-tense format with no time frame specified and 5 response options reflecting intensity. The calibration sample included 1,407 respondents, 1,000 from the general population (ascertained through an internet panel) and 407 from community treatment programs participating in the National Institute on Drug Abuse (NIDA) Clinical Trials Network (CTN).

**Results—**Final banks of 37, 31, 20, 11, and 9 items (108 total items) were calibrated for alcohol use, negative consequences, positive consequences, negative expectancies, and positive expectancies, respectively, using item response theory (IRT). Seven-item static short forms were also developed from each item bank.



*Corresponding author at postal address: Western Psychiatric Institute and Clinic, 3811 O'Hara Street, Pittsburgh, PA 15213, Telephone: 412.246.5833, Fax: 412.246.5840, pilkonispa@upmc.edu.



**Contributors**

Paul A. Pilkonis, PhD, contributed to study conception and design and took responsibility for drafting the manuscript. Lan Yu, PhD, provided data analysis and interpretation. Jason B. Colditz, BS, Nathan Dodds, BS, Kelly L. Johnston, MPH, Catherine Maihoefer, MS, LSW, Angela M. Stover, MA, Dennis C. Daley, PhD, and Dennis McCarty, PhD, contributed to study implementation (literature reviews, conceptual organization of items, item and intellectual property reviews) and data collection (focus groups, cognitive interviews, coordination of field testing for calibration). All authors reviewed and approved the final manuscript.

**Conflict of Interest**

There are no conflicts of interest for any authors.

**Conclusions—**Test information curves showed that the PROMIS item banks provided substantial information in a broad range of severity, making them suitable for treatment, observational, and epidemiological research.

### Keywords

alcohol use; alcohol consequences; alcohol expectancies; item response theory; measurement

## 1. INTRODUCTION

The Patient-Reported Outcomes Measurement Information System (PROMIS) is an NIH Roadmap initiative designed to improve self-reported outcomes using state-of-the-art psychometric methods (for detailed information, see www.nihpromis.org). PROMIS has developed and calibrated item banks assessing physical, mental, and social health, consistent with the World Health Organization's tripartite framework (Cella et al., 2007). For example, there are item banks assessing physical functioning, pain, fatigue, sleep disturbance, emotional distress (depression, anxiety, and anger), and social participation, providing a comprehensive profile of health status (Buysse et al., 2010; Cella et al., 2007, 2010; Fries et al., 2009; Pilkonis et al., 2011; Revicki et al., 2009). PROMIS is the most ambitious attempt to date to apply models from item response theory (IRT) to health-related assessment. The PROMIS methodology involves iterative steps of comprehensive literature searches; item pooling; development of conceptual frameworks; qualitative assessment of items using expert review, focus groups, and cognitive interviewing; and quantitative evaluation of items using techniques from both classical test theory (CTT) and IRT (Cella et al., 2007, 2010; Hilton, 2011; Reeve et al., 2007). We report here on the development and calibration of five item banks capturing prominent aspects of alcohol use (consumption, craving, efforts at control, internal and external triggers for drinking), negative and positive consequences of alcohol use, and negative and positive expectancies regarding drinking.

There is informative previous work using IRT models for the scaling and calibration of criteria for substance abuse and dependence, including those for alcohol (Krueger et al., 2004; Langenbucher et al., 2004; Martin et al., 2006; Saha et al., 2006). These criteria are sufficiently unidimensional for calibration with IRT models, but they are high-threshold items most appropriate for clinical samples. The use of such items results in "tests" with narrow bandwidth relevant only at the severe end of the continuum of alcohol use and consequences. From a psychometric perspective, our goal was to identify items that were more normally distributed and less positively skewed in a sample that included both members of the general population who used alcohol as well as alcohol abusers. Such items will provide more information across a broader range of the continuum of alcohol use, and for this reason, they will constitute more sensitive measures of treatment outcome and result in a single metric that could be used across treatment, observational, and epidemiological settings. Thus, we were interested in alcohol use not only at the level of clinical disorders but also at lower levels of consumption, where alcohol use may still be an important health-related behavior (or risk factor) relevant to a wide range of medical and psychiatric conditions.

## 2. METHODS

### 2.1. Development of item pool

**2.1.1. Comprehensive literature searches—**The Pittsburgh PROMIS research site developed a methodology for performing comprehensive literature searches to ensure content validity and broad coverage of the alcohol domain. We performed searches in the MEDLINE, PsycINFO, and Health and Psychosocial Instruments (HaPI) databases. Details

of the methodology are reported in Klem et al. (2009), and all search algorithms are available upon request. The searches generated 785 abstracts that could be linked to more than 200 unique measures of substance use. Cited reference searches were run on the primary reference for each measure in order to determine its acceptance and use by the scientific community. Copies of the measures were gathered from both electronic and print sources, and the measures were reviewed at the item level.

**2.1.2. Conceptual organization of items—**The initial alcohol item pool contained 5,241 items. We organized the items into conceptually meaningful categories using a hierarchical approach informed by previous empirical work (e.g., factor analyses) and clinical formulation. Previous work had divided alcohol use items into subdomains and factors relevant both to the DSM-IV categorization of alcohol use disorders, e.g., alcohol consumption, impairment associated with drinking (Green et al., 2011; Krueger et al., 2004; Muthén, 2006; Saha et al., 2007), and to broader themes surrounding the use of alcohol, e.g., precipitants to alcohol use, alcohol motives, expectancies regarding alcohol use (Jones et al., 2001; Pabst et al., 2009).

Our hierarchical structure for alcohol use included eight subdomains; consumption, craving and efforts to control drinking, triggers (internal and external), negative consequences, positive consequences, negative expectancies, positive expectancies, and general attitudes about alcohol. We also created 105 distinct facets within the subdomains. For example, within the consumption subdomain, we included facets for frequency, quantity, and patterns of alcohol use (e.g., binge versus consistent drinking).

**2.1.3. Focus groups—**To ensure comprehensive coverage of the conceptual area, we conducted focus groups and performed thematic analyses of the topics discussed (see Castel et al., 2008; Kelly et al., 2011). Members of four groups were recruited from outpatient substance use treatment programs. Two additional groups were comprised of social drinkers, community participants who reported drinking at least one alcoholic beverage in the past 30 days but who had no lifetime history of substance use disorders and no risk factors for current alcohol problems according to the National Institute on Alcohol Abuse and Alcoholism (NIAAA) screening guidelines (2003). Participants (total $n = 65$) were between the ages of 25 and 64 (mean = 45, SD = 10). They were predominantly female (68%) and members of minority groups (race = African American 58%, Caucasian 37%, Other 5%; ethnicity = Hispanic 5%). A majority had an annual household income of less than $20,000 (66%) and no formal education beyond high school (51%).

Using semi-structured scripts, facilitators prompted participants to discuss their experiences with alcohol and characteristics of problematic drinking. Research staff reviewed process notes from the groups and audio recordings, paying special attention to positive and negative appraisals (consequences of alcohol use, general expectations regarding alcohol) and contexts of drinking-related experiences. The goal was to enrich our item pool with content not represented on traditional questionnaires. For this purpose, we paid particular attention to accounts that suggested lower threshold items (e.g., did embarrassing things when drinking, rudeness, drinking routinely at the end of a busy day).

**2.1.4. Qualitative item review—**A key step in editing the item bank was qualitative review of the items done by members of the research team (see DeWalt et al., 2007, for a description of the qualitative procedures used by the PROMIS network). This process involved elimination of redundant items, items that were too narrow (often by virtue of being disease-specific), items that were confusing or vague, and items that were poorly written (e.g., double-barreled items). Our goal was to create a pool of about 250 items for field testing, with approximately 150 items for the alcohol bank and an additional 100 items

devoted to demographic characteristics, health status, medical history, history of substance use, and "legacy" measures of alcohol use and abuse (to investigate convergent validity with the new item bank). With this goal in mind, we reduced the item pool to 147 items, covering 103 of the original 105 facets.

**2.1.5. Standardization of items—**Items for alcohol use and consequences were written in a first-person, past-tense format with a 30-day time frame and 5 response options reflecting frequency (e.g., In the past 30 days, I lied about my drinking: *never, rarely, sometimes, often, almost always*). Expectancies and general attitudes about alcohol use, however, represent enduring beliefs, and as such, these items used a third-person, present-tense format, no time frame, and an intensity scale (e.g., Drinking puts people in a bad mood: *not at all*, *a little bit*, *somewhat*, *quite a bit*, *very much*). A small number of consumption items used a scale of actual number of drinks (e.g., drinks in a typical week, largest number of drinks in a single day). This standardization of items was consistent with our usual efforts to promote internal consistency across PROMIS measures (DeWalt et al., 2007; Pilkonis et al., 2011). In addition, a review of intellectual property issues was completed for all items (Berzon et al., 1994; Revicki and Schwartz, 2009). The large majority of items were generic, that is, they were similar to several extant items but not identifiable with any one in particular.

**2.1.6. Cognitive interviews—**Twenty-eight participants were recruited for cognitive interviews, and items were reviewed by at least 9 individuals with a variety of characteristics: at least 3 female, 4 minority, 3 social drinkers, 5 less than high school graduate reading level, and 2 less than 9$^{th}$ grade reading level as assessed by the Wide Range Achievement Test (WRAT-4; Wilkinson, 1993; Wilkinson and Robertson, 2006). An interviewer met with participants and asked each to "think aloud" while responding to items, then prompted for feedback on the language and clarity of items and the relevance of the content. Adaptations arising from cognitive interview feedback included the removal of modifiers that increased the threshold of items (e.g., "I had a *strong* urge to continue drinking"), clarifying ambiguities (e.g., "Drinking eases *physical* pain," to differentiate this from emotional pain), and reducing the literacy demand by replacing longer words with shorter synonyms.

## 2.2. Sampling

For calibration purposes, we administered the banks to an internet (YouGov Polimetrix) sample of 1,000 participants from the general population and a clinical sample of 407 patients in treatment for substance use disorders (not limited to alcohol abuse or dependence) at three sites: Addiction Medicine Services at the University of Pittsburgh Medical Center; the CODA treatment programs in Portland, Oregon; and the Evergreen treatment clinics in Seattle, Washington. These three sites are members of the NIDA Clinical Trials Network, and they served as collaborators for the PROMIS work. YouGov Polimetrix is a national, web-based polling firm in Palo Alto, CA. Both samples were composed of participants who answered "yes" to the screening question, "*For the past 30 days, did you drink any type of alcoholic beverage?*" Given the 30-day time frame, our items were only relevant for respondents who had used some alcohol in the past month, and our IRT calibrations should be interpreted in the context of this "floor," which required some minimal exposure to alcohol. On the other hand, to ensure adequate frequencies for each response category at higher levels of severity, we enriched the internet sample with the clinical sample of identified patients. Demographic characteristics of the Polimetrix sample and the clinical sample (combined across the three sites) are summarized in Table 1.

### 2.3. Measures

The alcohol use item pool brought to field testing contained 141 items tapping domains of consumption (14 items), craving and control (14 items), triggers (24 items), negative consequences (31 items), positive consequences (21 items), negative expectancies (14 items), positive expectancies (15 items), and general attitudes (8 items). Participants also completed (a) the 6-item set of "recommended alcohol questions" developed by the NIAAA task force (2003), which assesses frequency and quantity of alcohol use and patterns of consumption over the past 12 months, and (b) a legacy instrument: the Alcohol Use Disorders Identification Test (AUDIT), a 10-item measure reflecting primarily the frequency and quantity of drinking and the negative consequences associated with alcohol use, also during the past 12 months (Saunders et al., 1993).

### 2.4. Data Analysis

**2.4.1. General strategy—**We did not expect that an item bank of 141 diverse items would reflect a single underlying dimension. Therefore, our primary goal was to identify the most robust latent constructs and to document sufficient unidimensionality for each of them so that we could proceed with IRT analyses in which the credibility of model parameters relies on the assumption of unidimensionality. There are trade-offs, however, between bandwidth (item banks that have good content validity and capture a somewhat varied pool of clinical indicators) and fidelity (item banks that are unidimensional), and we tried to strike appropriate compromises by ensuring that each measure or subset of items was suitable for unidimensional scaling without unduly narrowing the construct.

As a first step, we inspected frequency distributions of individual items for sparse cells. We then began our explorations of dimensionality by dividing the sample randomly into two subsamples, one for exploratory factor analysis (EFA, $n = 681$) and the other for subsequent confirmatory factor analysis (CFA, $n = 726$). Both EFA and CFA were conducted using Mplus 4.21 with promax rotation (Muthén and Muthén, 2006). In the CFAs, the items were treated as categorical variables, and the robust weighted least squares (WLSMV) estimator was used. Scree plots, eigenvalues, and factor loadings were examined. We focused on the ratio of eigenvalues in EFAs and the relative proportion of variance accounted for by the factors extracted. We also emphasized the magnitude of factor loadings that appeared in both EFAs and CFAs and the fit and information values reflected in IRT models.

**2.4.2. Item response theory analysis—**The most commonly used IRT model for polytomous items (i.e., items with 3 or more ordinal response categories) is the two-parameter graded response model (GRM; Samejima, 1969). The GRM has a slope parameter and $n-1$ threshold parameters for each item, where $n$ is the number of response categories. The slope parameter measures item discrimination, i.e., how well the item differentiates between higher versus lower levels of severity (or θ in IRT terms). Useful items have large slope parameters. Threshold parameters measure item difficulty, i.e., the ease versus difficulty of endorsing different response options for an item. For example, the first threshold parameter for an item tells us where along the θ scale of severity a respondent is more likely to endorse a response of "rarely" rather than "never."

Items remaining in the pool for each construct were calibrated with the two-parameter graded response model (GRM) using MULTILOG 7.03 (Thissen et al., 2003). The convergence criterion for the EM cycles was set to .0001, with the number of cycles set to 100. IRT model fit was examined for each item using the IRTFIT macro program and the option for the sum-score-based method, which uses the sum score instead of theta for computing the predicted and observed frequencies (Orlando and Thissen, 2003).

Differential item functioning (DIF) occurs when characteristics such as age, gender, or ethnicity, which may seem extraneous to the assessment of the constructs under consideration, actually do have an effect on measurement. An item is identified as functioning differentially if the item is more (or less) difficult to endorse or more (or less) discriminating in some focal group (compared to a reference group) when the different subgroups have been matched on the latent trait under investigation. With regard to demographic characteristics, we conducted DIF analyses (for both uniform and non-uniform DIF) on the basis of gender and education (high school education or less versus further educational attainment). We focused on these two variables initially because the relevant comparison groups were adequately represented; that is, our sample included 56% men versus 44% women and 38% respondents with a high school education or less versus 62% with further educational attainment. Other potential comparison groups (e.g., white versus non-white respondents, Hispanic versus non-Hispanic respondents) were less equally divided. In addition, we conducted DIF analyses on the basis of drinking behavior (higher versus lower alcohol volume in the past year, as ascertained from the NIAAA question). This analysis allowed us to examine the performance of our items based on differential exposure to alcohol and to document their consistency across the full spectrum of alcohol use. Two different DIF procedures were employed: the IRT likelihood ratio method (Thissen et al., 1993) and an ordinal logistic regression procedure (Zumbo, 1999); items were considered for removal if they showed significant DIF ($p < .01$) by both methods (Teresi et al., 2009).

**2.4.3. Parallel analysis—**Following the IRT calibrations, we used parallel analysis to document further the unidimensionality of the item banks (Horn, 1965). We generated 1,000 simulated datasets of the same size as the field-test sample for the final versions of each item bank. The medians of the eigenvalues from the simulated datasets were plotted as a comparison line against the scree plots of the actual eigenvalues. The intersection of the two lines provides a threshold for determining the number of dimensions for each item bank. If these results suggested any departures from unidimensionality, we took a further step to evaluate the significance of such departures. Using the item parameter estimates obtained from the IRT calibrations, we generated a strictly unidimensional item response dataset of the same size as the field-test sample. The simulated data were subjected to principal component analyses using Mplus (Muthén and Muthén, 2006), and the scree plots of eigenvalues from the simulated data were plotted against the eigenvalues from the observed data to examine their concordance.

**2.4.4. Concurrent calibrations with the AUDIT—**The term “calibration” has various meanings in different contexts (Angoff, 1971; Kolen and Brennan, 2004; Linn, 1993; Lord, 1980; Thissen and Wainer, 2001). Concurrent calibration refers here to estimating item parameters across multiple measures (i.e., the alcohol use item banks and the AUDIT) on a single computer run. In this case, we fixed the final item parameters for the alcohol item banks and calibrated the AUDIT with these same parameters using the GRM. This procedure places the AUDIT on the same θ scale of severity of each of the alcohol item banks. Because the items of the AUDIT focus on alcohol consumption and negative consequences of drinking, we were interested primarily in comparing the information obtained from the AUDIT with the information obtained from our new item banks for these two constructs (alcohol use and negative consequences).

# 3. RESULTS

## 3.1. Alcohol, drug, and tobacco use in the two samples

The most stable estimates of alcohol use came from the NIAAA questions for the past 12 months (2003), given that shorter term patterns of drinking were influenced by the recent decision to seek treatment among members of the clinical sample. Table 2 summarizes these data. Frequency of drinking did not distinguish the community and clinical samples, with "once a week" being the median for both groups over the past year. (In this context, please note that members of the clinical sample sought treatment for many forms of substance abuse and may not have had a primary diagnosis of alcohol abuse or dependence.) Differences emerged, however, with quantity of alcohol (e.g., number of drinks on a typical drinking day, maximum number of drinks in a 24-hour period, frequency of drinking this maximum) and patterns of drinking (e.g., more frequent binges). Participants from the clinical sample also reported getting "high" on any other substance (in the past 30 days) much more frequently than members of the community sample: 82% versus 16%. Also, 80% of the clinical sample were current smokers versus 19% for the community sample.

Given this pattern of results, we created a measure of annual alcohol volume (drinks per year), computed as the number of days drinking × numbers of drinks on a typical drinking day, which reflected a strong difference between the two groups: $M = 376$ for the community sample versus 1,005 for the clinical sample. The distribution of this variable was positively skewed, producing means that were considerably higher than the median of 165 drinks per year across the entire sample. We used a median split at 165 drinks per year for our DIF analyses of higher versus lower exposure to alcohol. In the community sample, 54% fell at or below this level, whereas 42% did so in the clinical sample.

## 3.2. Frequency distributions

Among the initial item pool of 141 items, there were no items with any response categories having less than 1% response (14 participants). There were 33 items (23%) having at least one response category with between 1% and 3% response. However, the sparse cells for all 33 items had at least 20 respondents, ranging from 20 (1.4%) to 39 (2.8%). Therefore, we retained all 5 response categories for all items for further analyses.

## 3.3. Factor analyses

**3.3.1. Exploratory factor analyses—**Initial EFA of the entire item pool with the first half of the sample yielded a 5-factor solution with a large primary factor reflecting several aspects of alcohol use (i.e., consumption, craving and efforts at control, internal and external triggers for drinking, and some general attitudes about alcohol). The four additional factors were organized clearly around negative consequences, positive consequences, negative expectancies, and positive expectancies regarding alcohol. The eigenvalues for these five factors were 65.2, 11.8, 6.4, 4.1, and 3.7, and they accounted for 55% of the variance among the items.

We then did a second round of EFAs, with each of the five factors examined separately. For the primary factor of alcohol use (USE), there were 60 candidate items. Fourteen items were removed because of factor loadings < .50: 2 (of 14) items from consumption, 4 (of 14) items from craving and control, 4 (of 24) items from triggers, and 4 (of 8) items from general attitudes. All of the candidate items from negative consequences (NECO, $n = 31$), positive consequences (POCO, $n = 21$), negative expectancies (NEXP, $n = 14$), and positive expectancies (PEXP, $n = 15$) had factor loadings larger than .50 in the repeat EFAs and survived into the next stage of confirmatory factor analysis.

**3.3.2. Confirmatory factor analyses—**Single-factor CFAs were performed on each of the five factors based on the items retained in the second rounds of EFAs. For the USE, NECO, and POCO factors, all items had factor loadings greater than .50. However, both the NEXP and PEXP factors had two items with loadings below this threshold, and these four items were eliminated.

### 3.4. IRT calibrations

Based on the EFA and CFA analyses, the original 141 items were trimmed to a pool of 123 items distributed across 5 factors. Thus, 5 corresponding item banks were calibrated separately using the two-parameter GRM. Following IRT calibration, 6 items were eliminated on the basis of model misfit ($p < .001$), 1 from the POCO bank, 1 from the NEXP bank, and 4 from the PEXP bank. We also examined local dependency (i.e., residual correlations) in the IRT models using the Q3 statistic (Yen, 1984), but no items were eliminated for this reason. Based on item information functions (IIFs), an additional 7 items were removed from the USE item bank because they contributed little information (i.e., the peak of the IIF was less than 0.5). Four of these items were triggers, 2 reflected general attitudes, and 1 was related to consumption. This stage of pruning left only 2 general attitude items in the USE bank, and we removed these items to provide greater consistency within the bank; they were the only items without the 30-day time frame, and they required intensity rather than frequency ratings. Finally, our analyses of DIF by gender, education, and higher versus lower exposure to alcohol in the past year identified no items that were flagged by both DIF methods, and no further items were eliminated for this reason.

Thus, the final 5 calibrated item banks included 108 items in total: 37 items for USE, 31 items for NECO, 20 items for POCO, 11 items for NEXP, and 9 items for PEXP. Using Microsoft Word, the Flesch-Kincaid readability test was performed on individual items. The mean grade level across all items was 2.9 (SD = 2.4). The item banks, together with their IRT parameters, are summarized in Tables 3–7. Test information curves (and plots of corresponding standard errors) are displayed in Figures 1–5. Information values of 10 correspond approximately to CTT reliabilities of .90. At this threshold, the effective range of measurement varied across the banks, but in all cases, they were substantial: USE, −1 to +3 SDs; NECO, −1 to +2.5; POCO, −1 to +2; NEXP, −2.5 to +2; and PEXP, −1.5 to +2. These ranges are broader than those typically found for traditional measures of alcohol abuse and dependence, which assess high levels of severity.

### 3.5. Parallel analysis

Medians of the eigenvalues from 1,000 simulated datasets were plotted against the scree plots from the actual eigenvalues for each item bank. The intersections of the simulated eigenvalues and the actual eigenvalues for NECO, NEXP, and PEXP showed only one actual eigenvalue above the simulated eigenvalue lines, which supported the presence of a single dimension for these three banks. The graphs for USE and POCO showed two actual eigenvalues above the simulated eigenvalue lines, but in both cases, the second actual eigenvalues were very close to the simulated eigenvalue lines. Nonetheless, to evaluate further the significance of any potential multidimensionality of the USE and POCO banks, we generated a strictly unidimensional dataset for each of these two banks using the final IRT item parameter estimates. We used principal components analyses to generate the first 10 eigenvalues from the simulated and actual datasets. The ratios of the first-to-second eigenvalues in the actual data for USE and POCO were 10.3 and 7.5. The ratios of the first-to-second eigenvalues in the simulated data were 26.1 for USE and 16.3 for POCO. These large ratios document the significant influence of the first factor relative to all others in both the observed and simulated data for USE and POCO, with ratios of 4 or higher often taken as persuasive evidence for unidimensionality (Reeve et al., 2007). To illustrate this point

further, we also generated scree plots of the eigenvalues from both the observed and simulated data, and the striking overlap in these plots made it clear that the assumption of unidimensionality was well supported in the observed data for USE and POCO.

### 3.6. Selection of items for short forms

For some applications where CAT is not feasible, static short forms may be a useful alternative, assuming that they provide good coverage across the relevant range of the construct being measured. To develop such short forms, we rank ordered all USE, NECO, POCO, NEXP, and PEXP items on four criteria: discrimination parameter (α), the percentage of times the item would have been selected in a simulated CAT for each item bank based on the observed data from our calibration sample, expected information under the standard normal distribution with a mean of 0 and SD of 1, and expected information under a normal distribution with a larger SD, i.e., a mean of 0 and SD of 1.5 (Choi et al., 2010). The CAT simulations were performed using the Firestar program (Choi, 2009). We selected the best performing 7 items for each bank based on the convergence of these psychometric criteria, together with decisions about the clinical importance and content balance of candidate items. Tables 3–7 display the items in rank order according to their α parameters, and the items with asterisks are those selected for the short forms.

The internal consistency of the short forms was excellent. Alpha coefficients were .95, .96, .91, .92, and .88 for USE, NECO, POCO, NEXP, and PEXP, respectively. Correlations between the theta scores derived from the short forms and their corresponding full item banks were very high: .93, .95, .97, .99, and .99 for USE, NECO, POCO, NEXP, and PEXP, respectively. The same correlations using raw scores were .91, .94, .97, .97, and .97.

### 3.7. Concurrent calibrations with the AUDIT

In order to compare the final USE and NECO item banks to the legacy measure (AUDIT), on the same metric, items from the AUDIT were calibrated concurrently (on the same computer runs) with both the USE and NECO item banks by fixing the USE and NECO item parameters to their final calibration values. Given the content of the AUDIT, which is a mix of questions regarding alcohol consumption and negative consequences of drinking, only these item banks are relevant for comparative purposes. Figures 6 and 7 display the test information curves for the full USE and NECO item banks, the USE and NECO 7-item short forms, and the 10-item AUDIT. Overall, the full USE (37 items) and NECO (31 items) banks provided the most test information, a result that is not surprising given the large number of items they contain. The performance of the USE and NECO short forms is of greater interest. Even with fewer items, they provide more information than the AUDIT across the same ranges of measurement (for USE, −0.5 to +2.5 SDs, and for NECO, −1 to +2 SDs). The comparative efficiency of the short forms and the AUDIT should be interpreted with some caution, however, because the AUDIT was “projected” onto the PROMIS item banks (whose parameters were fixed) and, in a strict sense, the measures may not be assessing exactly the same construct.

## 4. DISCUSSION

Compared with existing alcohol use instruments (which typically cover the higher end of the severity spectrum, 1–2 SDs above the mean), the PROMIS item bank for alcohol use provides more information in a range from −1 to +3 SDs among people who have had some exposure to alcohol in the past month. The item banks for negative and positive consequences and negative and positive expectancies also provide broad coverage of these constructs. The item banks for use and consequences (both positive and negative) originated from larger initial item pools and were the most robust psychometrically. The initial item

pools for expectancies (14 and 15 items for NEXP and PEXP, respectively) were smaller and suffered some attrition, based on psychometric considerations, resulting in final item banks of 11 and 9 items. In particular, PEXP was somewhat difficult to model, with a larger proportion of items fitting poorly with an IRT model.

There is a literature documenting the complexities, both conceptual and psychometric, of measures of alcohol expectancies (Aarons et al., 2003; Vik et al., 1999). Different relevant dimensions have been discussed, e.g., valence (positive versus negative) and activation (arousing versus sedating), and different contextual domains have been described, e.g., expectancies regarding personal versus social effects. Our item pools addressed the issue of valence primarily, and within that scope, included items for personal (mental and physical) and social outcomes. We were able to derive small item banks (and 7-item short forms) that provide brief and efficient measures of positive and negative expectancies, but we recognize that these measures still require careful validation.

### 4.1. IRT modeling and unidimensionality

All health-related constructs are likely to reflect some multidimensionality, either because of the complexity of the constructs or for “nuisance” reasons, e.g., subsets of items that are more closely associated than predicted by a single latent factor because of semantic similarities or because they reflect a specific subdomain. For example, we generated items that reflected physical, mental, and social consequences of drinking, and it is plausible that items within each of these content areas might be more closely related than items across these areas. It is unlikely, however, that such associations compromise their general relevance for assessing negative (or positive) consequences of alcohol use, the broader construct being measured. As Reise has argued, there is a difference between studying the dimensionality of a correlation matrix versus determining the degree to which “scores” are influenced by a single common factor. Even multidimensional data can result in scores that still reflect essentially only one common influence, and we believe that our PROMIS item banks strike a reasonable compromise in this regard (Reise et al., 2010). We were able to calibrate well-fitting IRT models with plausible parameter estimates, and our parallel analysis using the final IRT parameters also provided confirmation of the unidimensional structure of the observed data.

### 4.2. Future directions

One goal is to further validate the PROMIS alcohol use item banks in prospective designs among clinical participants starting treatment for an alcohol use disorder. The key issue to be examined is the responsiveness to change of the PROMIS measures, especially when compared to common legacy measures. A related goal is to investigate the utility of the PROMIS item banks for purposes of screening and case identification. The original mandate of PROMIS has been to develop item banks that provide a continuous metric of severity (a “ruler” for each latent construct being assessed) relevant to all chronic diseases, both medical and psychiatric. At the same time, given the brevity and efficiency of CAT and short forms, it may be useful to examine the suitability of the PROMIS item banks for screening purposes and their concordance at different thresholds with clinical diagnoses.

As described above, we did initial explorations of DIF by gender, education, and higher versus lower exposure to alcohol given the demographic breakdown of our sample. Future work could include even larger samples and examination of the influence of race, ethnicity, and age. In the case of age, the item banks themselves could benefit from expansion designed to increase their relevance across the lifespan. This work would require adding items relevant to younger respondents (e.g., adolescents from ages 13–17) and older adults (e.g., geriatric respondents). Such items could then be linked to the existing item banks

through the use of shared items to ensure the availability of a single metric across the lifespan.

## Acknowledgments

We acknowledge the work of our colleagues within the two research nodes (the Tri-State Appalachian node and the Western States node) affiliated with the NIDA Clinical Trials Network that collaborated in this work. Dennis Daley, PhD, is the principal investigator of the Tri-State Appalachian node, and clinical data from this node were collected in Pittsburgh, PA, where we are grateful for the help of Dorothy Sandstrom, MS, and Janis McDonald. Dennis McCarty, PhD, is the principal investigator of the Western States node, and clinical data from this node were collected in Portland, OR and Seattle, WA. In Portland, we thank Katharina Wiest, PhD, and in Seattle, we thank Ron Jackson, MSW, and Ester Ricardo-Bulis, MA. We also thank our collaborators from NIH: Thomas Hilton, PhD (NIDA), William Riley, PhD (NHLBI), Howard Moss, MD (NIAAA), Daniel Falk, PhD (NIAAA), and Mark Willenbring, MD (formerly of NIAAA). Ann Doucette, PhD, of George Washington University, consulted throughout the item bank development process. Angela Stover, MA, is now affiliated with the Department of Health Behavior and Health Education at the University of North Carolina, Chapel Hill.

**Role of Funding Source**

The Patient-Reported Outcomes Measurement Information System (PROMIS®) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded initially by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project have included Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Laura Lee Johnson, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Susana Serrate-Sztein, MD, and James Witter, MD, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the web site at www.nihpromis.org for additional information on the PROMIS Cooperative Group.

## References

Aarons GA, Goldman MS, Greenbaum PE, Coovert MD. Alcohol expectancies: Integrating cognitive science and psychometric approaches. Addict. Behav. 2003; 28:947–961. [PubMed: 12788267]

Angoff, WH. Scales, norms, and equivalent scores. In: Thorndike, RL., editor. Educational Measurement. 2nd Edition. Washington, D.C.: American Council on Education; 1971.

Berzon R, Patrick D, Guyatt G, Conley JM. Intellectual property considerations in the development and use of HRQL measures for clinical trial research. Qual. Life Res. 1994; 3:273–277. [PubMed: 7812280]

Buysse D, Yu L, Moul DE, Germain A, Stover A, Dodds NE, Johnston KL, Shablesky-Cade MA, Pilkonis PA. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. Sleep. 2010; 33

Castel LD, Williams KA, Bosworth HB, Eisen SV, Hahn EA, Irwin DE, Kelly MAR, Morse J, Stover A, DeWalt DA, DeVellis RF. Content validity in the PROMIS social health domain: a qualitative anaylsis of focus group data. Qual. Life Res. 2008; 17:737–749. [PubMed: 18478368]

Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short forms, and computerized adaptive assessment. Qual. Life Res. 2007; 16:133–144. [PubMed: 17401637]

Cella D, Riley W, Stone A, Rothrock N, Reeve BB, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, DeVellis R, DeWalt D, Fries JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item bank: 2005–2008. J. Clin. Epidemiol. 2010; 63:1179–1194. [PubMed: 20685078]

Choi S. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. Appl. Psychol. Measure. 2009; 33:644–645.

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual. Life Res. 2010; 19:125–136. [PubMed: 19941077]

DeWalt DA, Rothrock N, Yount S, Stone AA. on behalf of the PROMIS Cooperative Group. Evaluation of item candidates: the PROMIS qualitative item review. Med. Care. 2007; 45:S12–S21. [PubMed: 17443114]

Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J. Rheumatol. 2009; 36:2061–2066. [PubMed: 19738214]

Green BA, Ahmed AO, Marcus DK, Walters GD. The latent structure of alcohol use pathology in an epidemiological sample. J. Psychiatr. Res. 2011; 45:225–233. [PubMed: 20615513]

Hilton TF. The promise of PROMIS for addiction. Drug Alcohol Depend. 2011; 119:229–234. [PubMed: 22238781]

Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965; 30:179–185. [PubMed: 14306381]

Jones BT, Corbin W, Fromme K. A review of expectancy theory and alcohol consumption. Addiction. 2001; 96:57–72. [PubMed: 11177520]

Kelly MA, Morse JQ, Stover A, Hofkens T, Huisman E, Shulman S, Eisen SV, Becker SJ, Weinfurt K, Boland E, Pilkonis PA. Describing depression: congruence between patient experiences and clinical assessments. Br. J. Clin. Psychol. 2011; 50:46–66. [PubMed: 21332520]

Klem ML, Saghafi E, Abromitis R, Stover A, Dew MA, Pilkonis PA. Building PROMIS item banks: librarians as co-investigators. Qual. Life Res. 2009; 18:881–888. [PubMed: 19548118]

Kolen, MJ.; Brennan, RL. Test Equating, Scaling, and Linking: Methods and Practices. 2nd ed.. New York: Springer; 2004.

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

Langenbucher JW, Labouvie E, Martin CS, Sanjuan PM, Bavly L, Kirisci L, Chung T. An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. J. Abnorm. Psychol. 2004; 113:72–80. [PubMed: 14992659]

Linn RL. Linking results of distinct assessments. Appl. Measure. Educ. 1993; 6:83–102.

Lord, FM. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.

Martin CS, Chung T, Kirisci L, Langenbucher JW. Item response theory analysis of diagnostic Cciteria for alcohol and cannabis use disorders in adolescents: implications for DSM-V. J. Abnorm. Psychol. 2006; 115:807–814. [PubMed: 17100538]

Muthén B. Should substance use disorders be considered as categorical or dimensional? Addiction. 2006; 101:6–16. [PubMed: 16930156]

Muthén, LK.; Muthén, B. Mplus User's Guide. 4th ed.. Los Angeles, CA: Muthén & Muthén; 2006.

National Institute on Alcohol Abuse and Alcoholism (NIAAA). [accessed on November 4, 2011] Task Force on Recommended Alcohol Questions: National Council on Alcohol Abuse and Alcoholism Recommended Set of Alcohol Consumption Questions: October 15–16. 2003. http://www.niaaa.nih.gov/Resources/ResearchResources/TaskForce.htm

Orlando M, Thissen D. Further investigation of the performance of the S-X2: an item fit index for use with dichotomous item response theory models. Appl. Psychol. Measure. 2003; 27:289–298.

Pabst A, Baumeister SE, Kraus L. Alcohol-expectancy dimensions and alcohol consumption at different ages in the general population. J. Stud. Alcohol Drugs. 2009; 71:46–53. [PubMed: 20105413]

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system: depression, anxiety, anger. Assessment. 2011; 18:263–283. [PubMed: 21697139]

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Cella D, group obotPc. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcome

Measurement Information System (PROMIS). Med. Care. 2007; 45:S22–S31. [PubMed: 17443115]

Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. J. Pers. Asses. 2010; 92:544–559.

Revicki D, Chen W, Harnam N, Cook K, Amtmann D, Callahan LF, Jensen MP, Keefe FJ. Development and psychometric analysis of the PROMIS pain behavior item bank. Pain. 2009; 146:158–169. [PubMed: 19683873]

Revicki D, Schwartz CE. Intellectual property rights and good research practice. Qual. Life Res. 2009; 18:1279–1280. [PubMed: 19885743]

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

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

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Mono. 1969; 17

Saunders JB, Aasland OG, Babor TF, Delafuente JR, Grant M. Development of the Alcohol-Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption. Addiction. 1993; 88:791–804. [PubMed: 8329970]

Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, Lai JS, Choi SW, Hays RD, Reeve BB, Reise SP, Pilkonis PA, Cella D. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. Psychol. Sci. Q. 2009; 51:148–180. [PubMed: 20336180]

Thissen, D.; Chen, WH.; Bock, RD. Multilog (version 7). Lincolnwood, IL: Scientific Software International; 2003.

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum; 1993. p. 67-113.

Thissen, D.; Wainer, H. Test Scoring. Hillsdale, NJ: Lawrence Erlbaum Associates; 2001.

Vik PW, Carrello P, Nathan PE. Hypothesized simple factor structure for the Alcohol Expectancy Questionnaire: confirmatory factor analysis. Exp. Clin. Psychopharm. 1999; 7:294–303.

Wilkinson, GS. The Wide Range Achievement Test: Manual. Wilmington, DE: Wide Range; 1993.

Wilkinson, GS.; Robertson, GJ. WRAT4: Wide Range Achievement Test Professional Manual. Lutz, FL: Psychological Assessment Resources; 2006.

Yen WM. Effects of local item dependence on the fit and equation performance of the three-parameter logistic model. Appl. Psychol. Measure. 1984; 2:125–145.

Zumbo, BD. A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores. Ottawa, ON: Directorate of Human Resources Research and Education, Department of National Defense; 1999.

USE

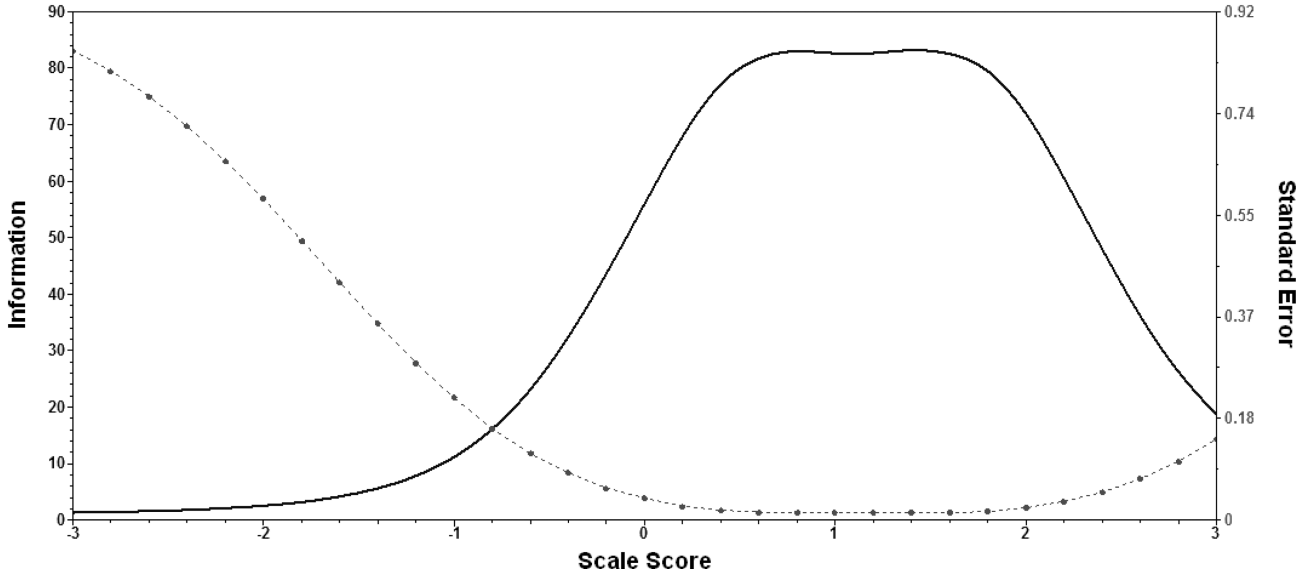


**Figure 1.**
Test information and standard error curves for the alcohol use item bank.

NEGATIVE CONSEQUENCES

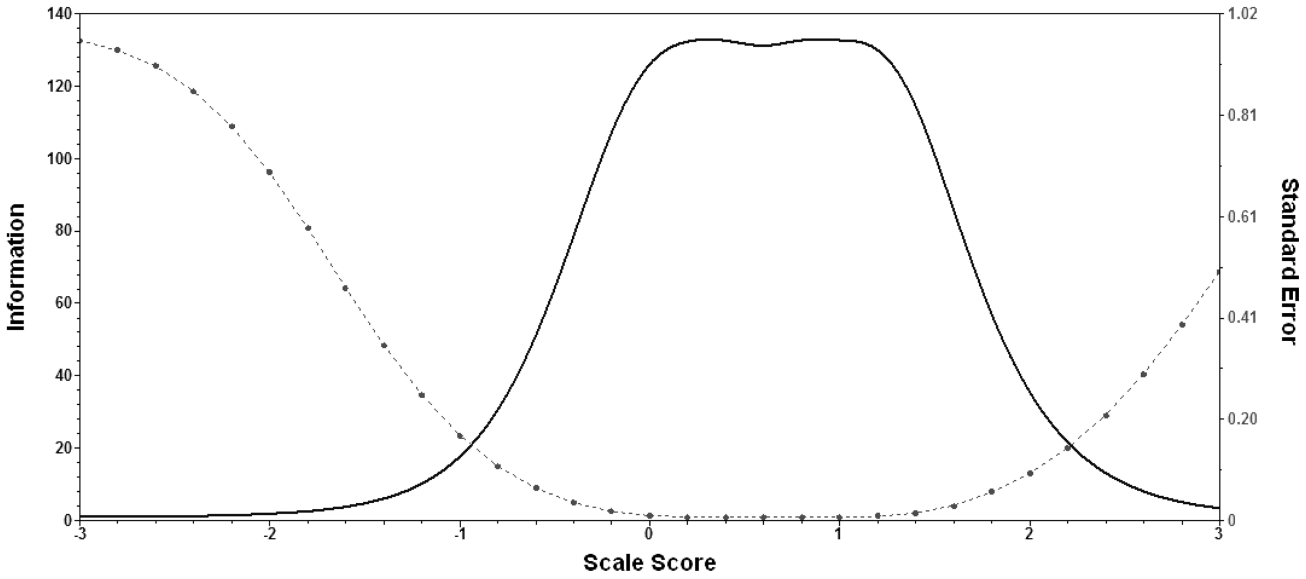


**Figure 2.**
Test information and standard error curves for the negative consequences item bank.

POSITIVE CONSEQUENCES

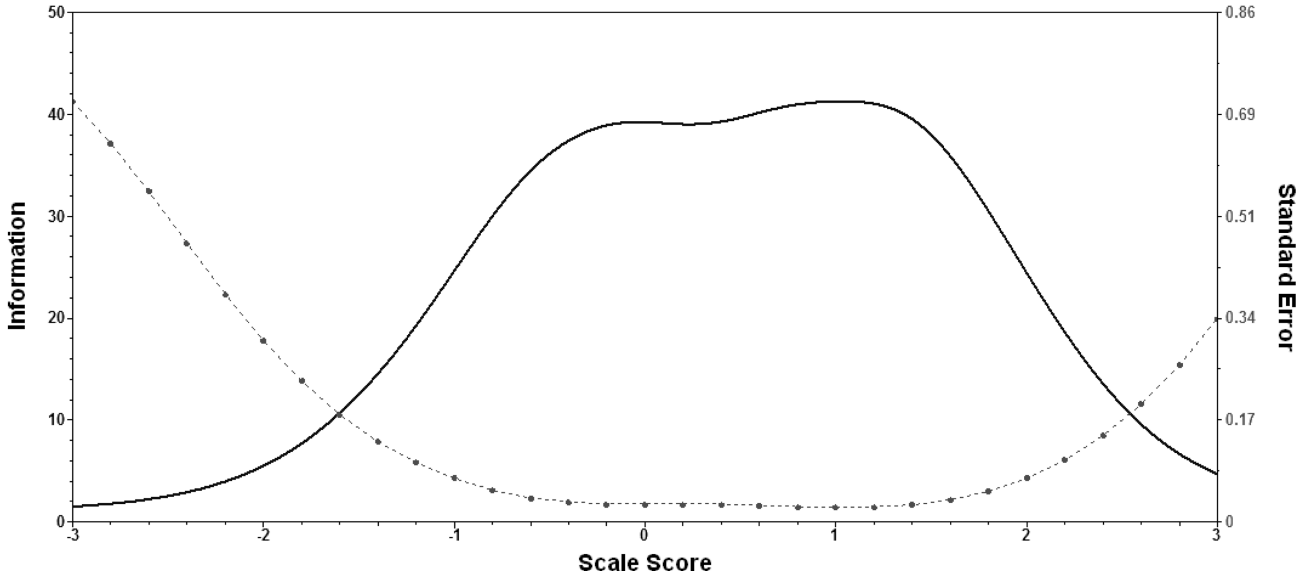


**Figure 3.**
Test information and standard error curves for the positive consequences item bank.

NEGATIVE EXPECTANCIES

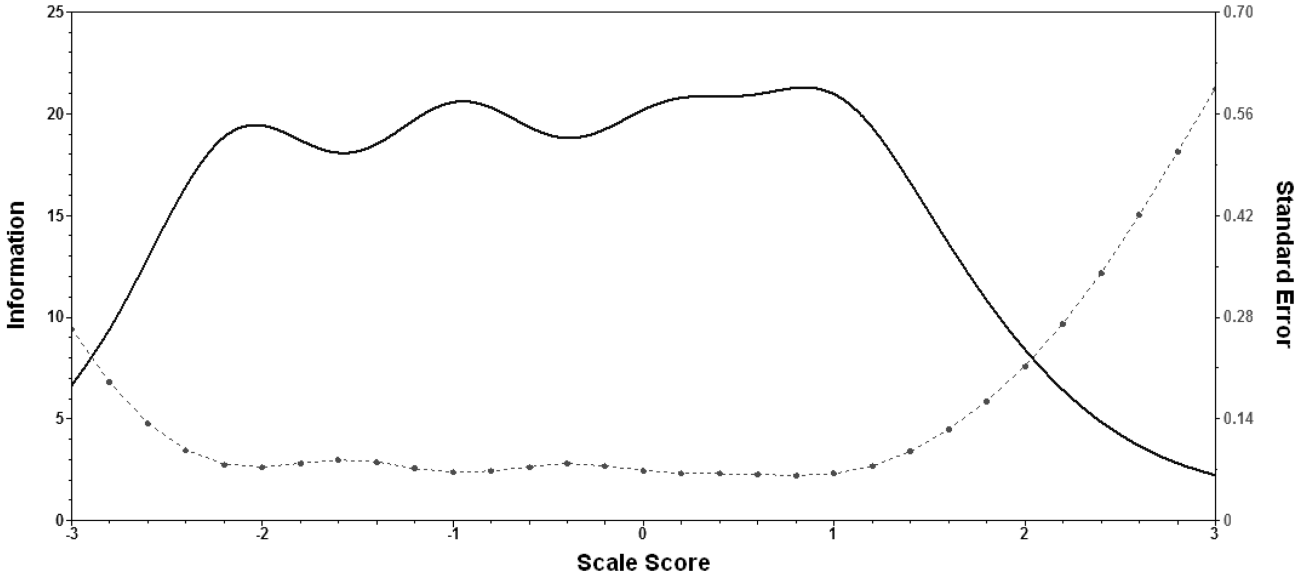


**Figure 4.**
Test information and standard error curves for the negative expectancies item bank.

# POSITIVE EXPECTANCIES

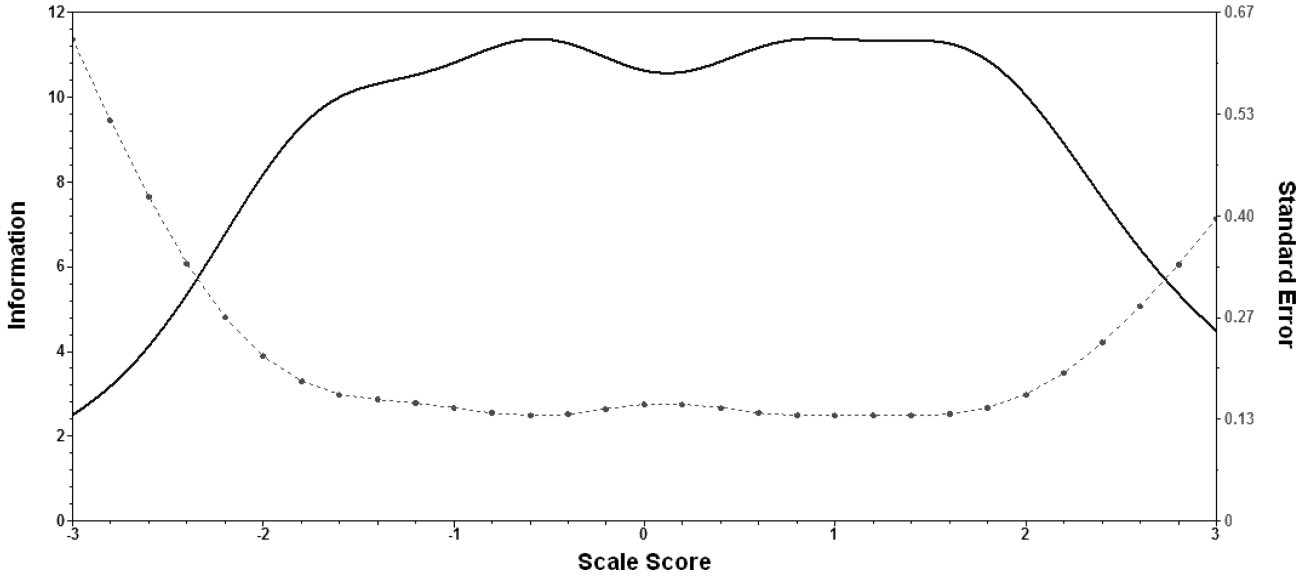


**Figure 5.**
Test information and standard error curves for the positive expectancies item bank.

USE VS. AUDIT

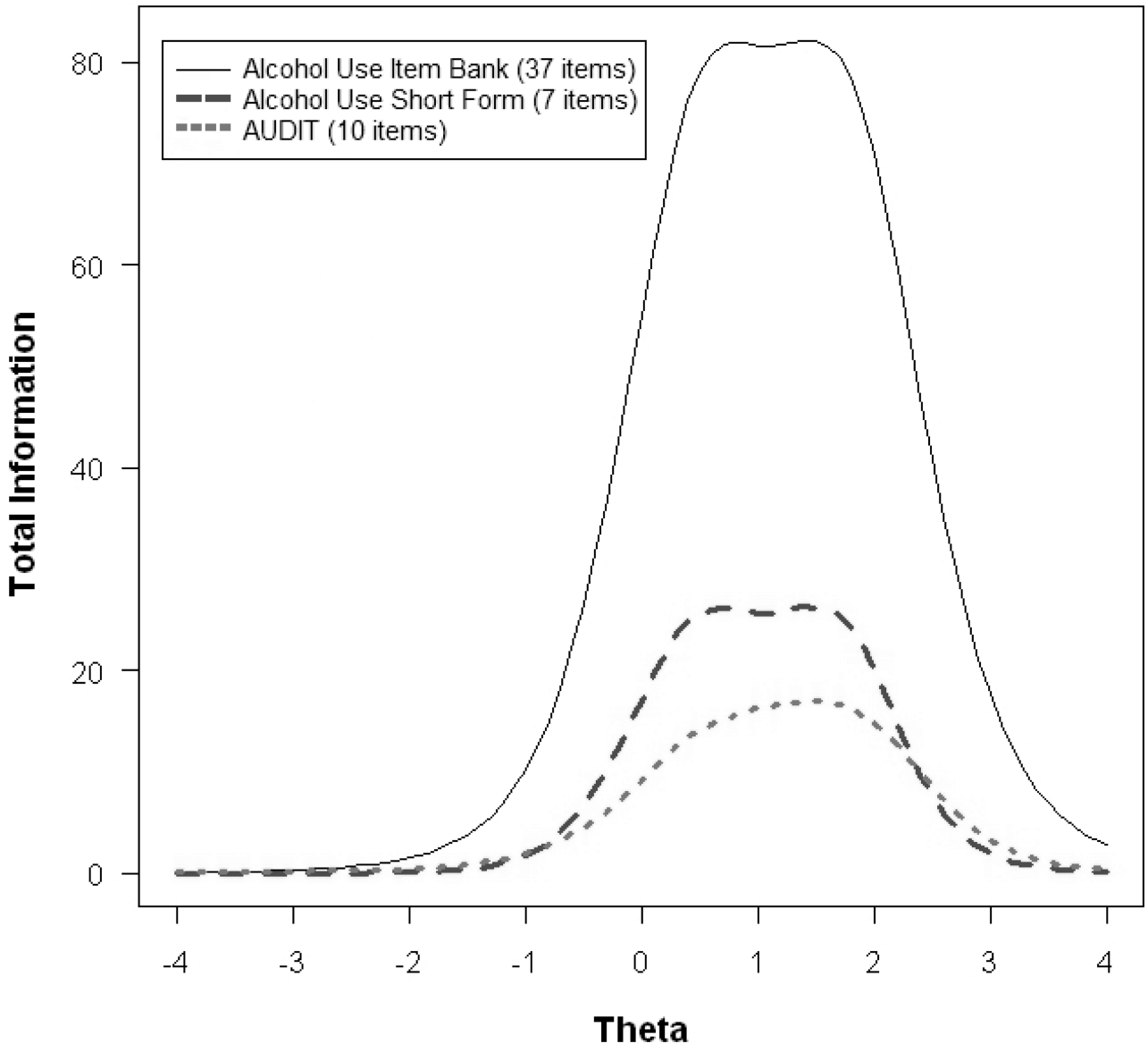


**Figure 6.**
Comparative test information curves for alcohol use: Full item bank, 7-item short form, and the AUDIT.

NEGATIVE CONSEQUENCES VS. AUDIT

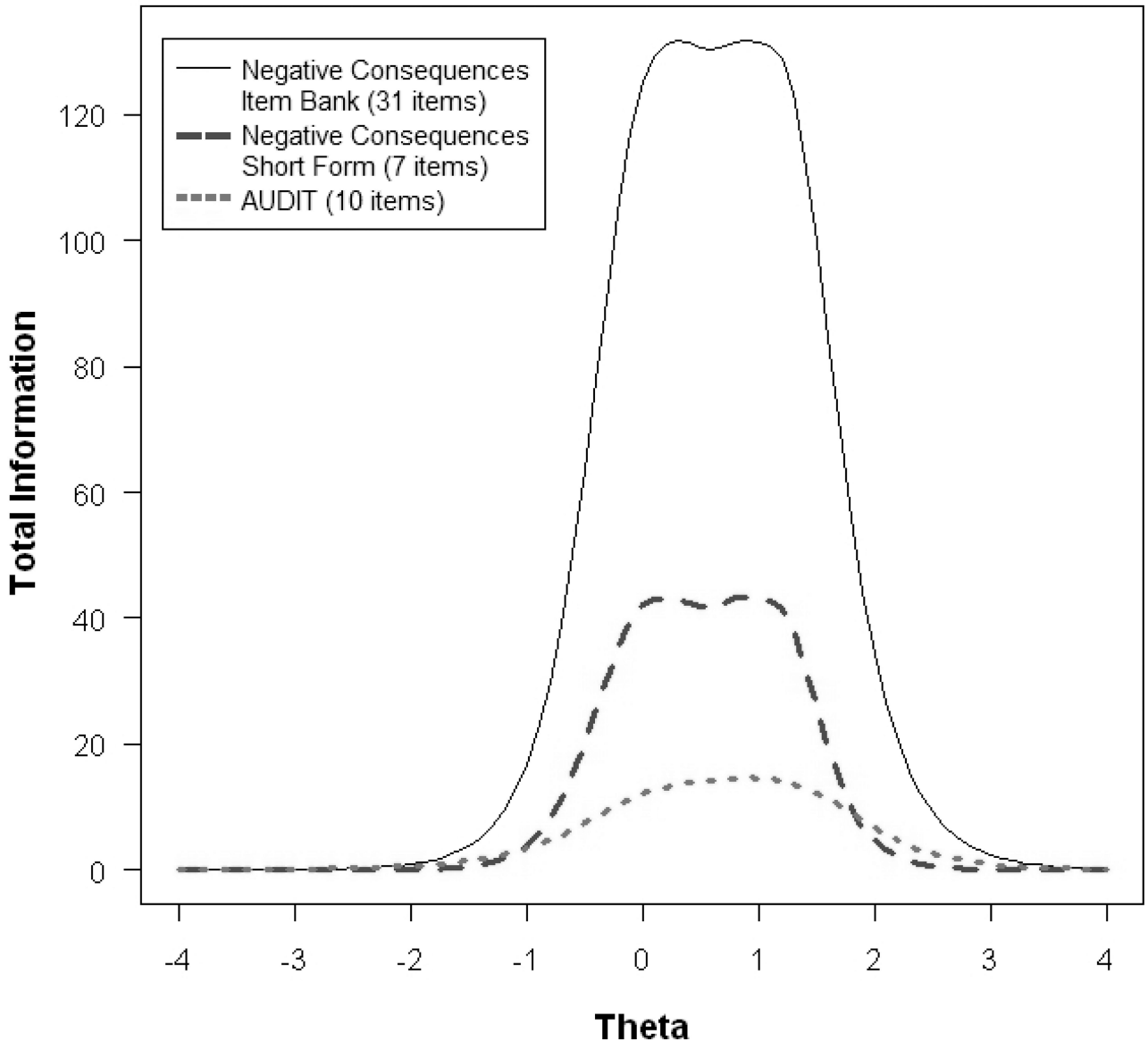


**Figure 7.**
Comparative test information curves for negative consequences: Full item bank, 7-item short form, and the AUDIT.

**Table 1**

Size and Demographic Characteristics of the Calibration Sample

| | Community Sample (*N* = 1,000) | Clinical Sample (*N* = 407) |
|---|---|---|
| **Characteristic** | **%** | **%** |
| Sex | | |
| Male | 54.3 | 60.7 |
| Ethnicity | | |
| Hispanic | 9.5 | 7.6 |
| Race | | |
| American Indian / Alaska Native | 1.2 | 3.7 |
| Asian | 1.2 | 0.2 |
| Black / African American | 8.4 | 20.6 |
| Native Hawaiian / Pacific Islander | 0.6 | 1.5 |
| White | 85.3 | 67.6 |
| Other / Multiracial | 3.3 | 6.4 |
| Education | | |
| High school diploma or less | 27.6 | 62.4 |
| Further educational attainment | 72.4 | 37.1 |
| Mean Age (*SD*) | 45.2 (15.7) | 37.2 (11.5) |

**Table 2**

Alcohol Use in the Calibration Sample (Past Year)

| | Frequency of drinking any kind of alcoholic beverage | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 or 2 times a year | 3 to 11 times a year | Once a month | 2 to 3 times a month | Once a week | Twice a week | 3 to 4 times a week | 5 to 6 times a week | Every day | |
| Community | 4.4% | 13.1% | 6.6% | 21.4% | 11.0% | 15.7% | 14.9% | 7.6% | 5.3% | |
| Clinical | 6.6% | 14.1% | 10.1% | 12.9% | 7.8% | 11.6% | 14.1% | 10.6% | 12.1% | |
| | Number of alcoholic drinks on a typical day when drinking | | | | | | | | | |
| | 1 drink | 2 drinks | 3 to 4 drinks | 5 to 6 drinks | 7 to 8 drinks | 9 to 11 drinks | 12 to 15 drinks | 16 to 18 drinks | 19 to 24 drinks | 25+ drinks |
| Community | 29.1% | 34.6% | 21.1% | 8.2% | 3.1% | 1.2% | 1.0% | 0.2% | 0.5% | 0.9% |
| Clinical | 8.2% | 21.5% | 24.0% | 16.2% | 7.2% | 11.2% | 5.0% | 2.0% | 2.0% | 2.5% |
| | Largest number of drinks containing alcohol within a 24-hour period | | | | | | | | | |
| | 1 drink | 2 drinks | 3 drinks | 4 drinks | 5 to 7 drinks | 8 to 11 drinks | 12 to 17 drinks | 18 to 23 drinks | 24 to 35 drinks | 36+ drinks |
| Community | 7.9% | 18.2% | 13.0% | 16.5% | 18.9% | 13.9% | 7.2% | 1.8% | 1.4% | 1.2% |
| Clinical | 4.2% | 5.5% | 8.0% | 12.5% | 17.5% | 17.5% | 14.0% | 8.8% | 7.5% | 4.5% |
| | Frequency of drinking the largest number of drinks | | | | | | | | | |
| | 1 or 2 times in a year | 3 to 11 times a year | Once a month | 2 to 3 times a month | Once a week | Twice a week | 3 to 4 times a week | 5 to 6 times a week | Every day | |
| Community | 49.9% | 21.5% | 10.8% | 7.0% | 4.5% | 3.1% | 1.4% | 0.5% | 1.3% | |
| Clinical | 30.9% | 16.0% | 13.0% | 12.5% | 5.7% | 7.2% | 5.2% | 5.0% | 4.5% | |
| | Frequency of having at least 4 (for females) or 5 (for males) drinks containing alcohol within a two-hour period | | | | | | | | | |
| | Never | 1 or 2 days a year | 3 to 11 days a year | One day a month | 2 to 3 days a month | One day a week | Two days a week | 3 to 4 days a week | 5 to 6 days a week | Every day |
| Community | 60.7% | 15.6% | 8.5% | 4.3% | 3.1% | 2.5% | 1.9% | 1.8% | 1.0% | 0.6% |
| Clinical | 29.5% | 13.8% | 9.8% | 5.5% | 7.5% | 4.2% | 7.2% | 11.2% | 6.2% | 5.0% |

**Table 3**

Calibrated Alcohol Use Items

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Never vs. Rarely | Rarely vs. Sometimes | Sometimes vs. Often | Often vs. Almost always |
| I had trouble controlling my drinking * | 4.70 | 0.43 | 0.84 | 1.33 | 1.75 |
| I had trouble stopping drinking when I wanted to | 4.17 | 0.51 | 0.91 | 1.37 | 1.86 |
| It was difficult to get the thought of drinking out of my mind * | 3.85 | 0.46 | 0.92 | 1.45 | 1.90 |
| It was difficult for me to stop drinking after one or two drinks * | 3.76 | 0.23 | 0.65 | 1.17 | 1.58 |
| I felt I needed help for my drinking | 3.62 | 0.71 | 0.94 | 1.34 | 1.79 |
| I spent too much time drinking * | 3.22 | 0.24 | 0.85 | 1.49 | 2.00 |
| I drank more than planned * | 3.18 | −0.09 | 0.52 | 1.30 | 1.88 |
| I had an urge to continue drinking once I started | 3.07 | −0.16 | 0.34 | 1.02 | 1.58 |
| I drank too much * | 3.03 | 0.00 | 0.74 | 1.50 | 2.02 |
| I drank because I was lonely | 2.94 | 0.44 | 0.82 | 1.52 | 2.04 |
| I drank because I was angry with myself | 2.92 | 0.63 | 1.11 | 1.73 | 2.28 |
| I drank because I was bored | 2.87 | 0.18 | 0.63 | 1.41 | 2.09 |
| I drank heavily at a single sitting * | 2.85 | −0.02 | 0.70 | 1.46 | 2.08 |
| I had cravings for alcohol | 2.80 | 0.09 | 0.65 | 1.43 | 2.07 |
| I drank because I was sad | 2.78 | 0.33 | 0.78 | 1.51 | 2.13 |
| I drank because I was depressed | 2.71 | 0.27 | 0.72 | 1.42 | 2.03 |
| I had urges to drink | 2.67 | −0.19 | 0.53 | 1.33 | 2.12 |
| I drank because I had nothing to do | 2.63 | 0.27 | 0.73 | 1.49 | 2.19 |
| I drank because someone made me angry | 2.61 | 0.46 | 0.95 | 1.80 | 2.29 |
| I drank because I was irritable | 2.57 | 0.33 | 0.83 | 1.63 | 2.27 |
| I drank because I was nervous | 2.55 | 0.39 | 0.90 | 1.71 | 2.29 |
| I drank because I was annoyed | 2.51 | 0.33 | 0.90 | 1.72 | 2.42 |
| I felt that I should cut down on my drinking | 2.48 | 0.18 | 0.65 | 1.35 | 1.86 |
| I finished several drinks fast to get a quick effect | 2.47 | 0.17 | 0.76 | 1.47 | 2.04 |
| I drank throughout the day | 2.47 | 0.54 | 1.13 | 1.78 | 2.28 |

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Never vs. Rarely | Rarely vs. Sometimes | Sometimes vs. Often | Often vs. Almost always |
| I became drunk or intoxicated | 2.33 | −0.03 | 0.96 | 1.52 | 1.87 |
| I spent a whole weekend drinking | 2.26 | 0.77 | 1.27 | 1.62 | 1.93 |
| On a typical day when I drank alcohol, I had…[1] | 2.11 | 0.12 | 0.91 | 1.54 | 2.14 |
| I drank because I felt tense | 1.98 | 0.00 | 0.60 | 1.67 | 2.48 |
| The largest number of drinks that I had in a single day was…[1] | 1.92 | −0.61 | 0.17 | 0.78 | 1.34 |
| I used alcohol and other drugs together, to get high | 1.81 | 0.40 | 0.86 | 1.59 | 2.28 |
| In a typical week, I drank…[2] | 1.74 | 0.63 | 1.27 | 1.79 | 2.31 |
| I drank because I had physical pain | 1.69 | 0.76 | 1.26 | 2.12 | 2.99 |
| I drank when I was alone | 1.52 | −0.12 | 0.53 | 1.42 | 2.17 |
| I drank when I arrived at home | 1.48 | −0.16 | 0.52 | 1.59 | 2.36 |
| I drank because I deserved it | 1.30 | 0.07 | 0.77 | 2.13 | 2.95 |
| I drank at the end of a busy day | 1.24 | −0.99 | −0.12 | 1.44 | 2.63 |

[1]Response options = 1–2 drinks; 3–4 drinks; 5–6 drinks; 7–10 drinks; More than 10 drinks

[2]Response options = 1–7 drinks; 8–14 drinks; 15–21 drinks; 22–28 drinks; More than 28 drinks

*Note*. Items are rank-ordered on the basis of their slope (discrimination) parameters.

Items included in the short form are marked with an asterisk.

**Table 4**

Calibrated Negative Consequences Items

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Never vs. Rarely | Rarely vs. Sometimes | Sometimes vs. Often | Often vs. Almost always |
| Drinking created problems between me and others * | 5.54 | −0.06 | 0.36 | 0.80 | 1.19 |
| I disappointed others when I drank * | 5.01 | −0.08 | 0.31 | 0.77 | 1.21 |
| Others had trouble counting on me when I drank | 4.91 | 0.07 | 0.45 | 0.84 | 1.27 |
| I was unreliable after I drank * | 4.80 | −0.10 | 0.33 | 0.83 | 1.29 |
| Others complained about my drinking * | 4.78 | 0.08 | 0.44 | 0.90 | 1.27 |
| I had trouble keeping appointments after I drank | 4.75 | 0.13 | 0.47 | 0.82 | 1.23 |
| I used poor judgment when I drank * | 4.70 | −0.38 | 0.11 | 0.67 | 1.15 |
| I was criticized about my drinking | 4.68 | 0.02 | 0.37 | 0.80 | 1.29 |
| I was inconsiderate when I drank | 4.30 | −0.17 | 0.38 | 0.92 | 1.40 |
| I said or did embarrassing things when I drank * | 4.03 | −0.35 | 0.15 | 0.88 | 1.35 |
| I took risks when I drank | 4.00 | −0.23 | 0.18 | 0.72 | 1.41 |
| I got in an argument when I drank | 3.96 | −0.10 | 0.44 | 1.00 | 1.43 |
| I felt angry when I drank | 3.89 | −0.07 | 0.42 | 1.03 | 1.43 |
| I lied about my drinking | 3.87 | 0.19 | 0.46 | 0.91 | 1.32 |
| I had trouble trusting other people when I drank | 3.64 | −0.08 | 0.43 | 0.90 | 1.37 |
| I got confused when I drank | 3.63 | −0.09 | 0.44 | 1.04 | 1.51 |
| I felt guilty when I drank | 3.51 | −0.11 | 0.32 | 0.87 | 1.25 |
| My problems seemed worse when I drank | 3.48 | −0.08 | 0.41 | 0.91 | 1.34 |
| I looked sloppy when I drank | 3.44 | −0.11 | 0.41 | 1.01 | 1.54 |
| I felt sad when I drank | 3.38 | −0.18 | 0.33 | 0.97 | 1.48 |
| I had trouble getting things done after I drank * | 3.22 | −0.32 | 0.19 | 0.82 | 1.36 |
| I was clumsy when I drank | 3.22 | −0.47 | 0.14 | 0.88 | 1.48 |
| I was critical of myself when I drank | 3.20 | −0.24 | 0.30 | 0.89 | 1.43 |
| I felt anxious when I drank | 3.15 | 0.01 | 0.56 | 1.18 | 1.64 |
| I was loud when I drank | 2.90 | −0.40 | 0.14 | 0.87 | 1.31 |

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Never vs. Rarely | Rarely vs. Sometimes | Sometimes vs. Often | Often vs. Almost always |
| I worried when I drank | 2.90 | −0.10 | 0.48 | 1.10 | 1.69 |
| I felt nervous when I drank | 2.70 | 0.17 | 0.71 | 1.41 | 1.84 |
| I got sick when I drank | 2.57 | 0.07 | 0.67 | 1.22 | 1.86 |
| I had a hangover after I drank | 2.46 | −0.42 | 0.27 | 0.99 | 1.52 |
| I felt dizzy when I drank | 2.24 | −0.27 | 0.36 | 1.19 | 1.86 |
| I had a headache after I drank. | 2.15 | −0.50 | 0.20 | 1.05 | 1.68 |

*Note*. Items are rank-ordered on the basis of their slope (discrimination) parameters.

Items included in the short form are marked with an asterisk.

**Table 5**

Calibrated Positive Consequences Items

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Never vs. Rarely | Rarely vs. Sometimes | Sometimes vs. Often | Often vs. Almost always |
| I felt outgoing when I drank * | 3.59 | −0.60 | −0.13 | 0.72 | 1.34 |
| I felt confident when I drank * | 3.32 | −0.46 | 0.04 | 0.83 | 1.45 |
| I had more fun when I drank * | 3.19 | −0.68 | −0.16 | 0.79 | 1.34 |
| I felt creative when I drank * | 2.79 | −0.12 | 0.46 | 1.24 | 1.87 |
| I fit in better when I drank | 2.75 | −0.14 | 0.32 | 1.24 | 1.90 |
| I was able to express myself better when I drank * | 2.72 | −0.23 | 0.28 | 1.11 | 1.69 |
| It was easier to talk to people when I drank | 2.72 | −0.57 | −0.15 | 0.80 | 1.44 |
| My future seemed better when I drank | 2.71 | 0.20 | 0.80 | 1.52 | 2.02 |
| I felt at ease when I drank * | 2.66 | −0.97 | −0.42 | 0.60 | 1.32 |
| I felt good about myself when I drank * | 2.65 | −0.47 | 0.02 | 0.97 | 1.46 |
| I felt happy when I drank | 2.61 | −1.00 | −0.48 | 0.62 | 1.31 |
| I felt like I could do anything when I drank | 2.61 | 0.19 | 0.67 | 1.32 | 1.83 |
| I felt comfortable around others when I drank | 2.49 | −0.88 | −0.39 | 0.64 | 1.23 |
| I enjoyed life when I drank | 2.45 | −0.89 | −0.36 | 0.72 | 1.39 |
| I could relax when I drank | 2.35 | −1.26 | −0.69 | 0.50 | 1.29 |
| I felt relaxed when I drank | 2.31 | −1.31 | −0.71 | 0.44 | 1.26 |
| I calmed down when I drank | 2.21 | −0.64 | −0.10 | 0.98 | 1.78 |
| I felt a sense of control when I drank | 2.06 | −0.17 | 0.36 | 1.26 | 1.72 |
| I slept better after I drank | 1.74 | −0.72 | −0.05 | 0.91 | 1.66 |
| I had more desire for sex when I drank | 1.70 | −0.44 | 0.23 | 1.29 | 2.00 |

*Note.* Items are rank-ordered on the basis of their slope (discrimination) parameters.

Items included in the short form are marked with an asterisk.

*Drug Alcohol Depend*. Author manuscript; available in PMC 2014 June 01.

**Table 6**

Calibrated Negative Expectancies Items

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Not at all vs. A little bit | A little bit vs. Somewhat | Somewhat vs. Quite a bit | Quite a bit vs. Very much |
| People are careless when they drink * | 3.32 | −2.12 | −0.97 | 0.17 | 1.07 |
| People make bad decisions when they drink * | 3.19 | −2.23 | −1.13 | 0.00 | 0.88 |
| People are irresponsible when they drink * | 3.15 | −2.06 | −0.85 | 0.20 | 1.04 |
| People do things they regret while drinking * | 3.00 | −2.32 | −1.06 | 0.05 | 0.93 |
| People are rude when they drink * | 2.64 | −1.98 | −0.68 | 0.67 | 1.58 |
| People have trouble thinking when they drink * | 2.47 | −2.14 | −0.98 | 0.29 | 1.23 |
| People are pushy when they drink * | 2.42 | −1.89 | −0.66 | 0.69 | 1.65 |
| People feel sick the day after drinking | 2.04 | −1.97 | −0.66 | 0.69 | 1.67 |
| Drinking is harmful to mental health | 2.02 | −1.71 | −0.63 | 0.46 | 1.12 |
| Drinking can be harmful to physical health | 1.90 | −2.61 | −1.33 | −0.17 | 0.68 |
| People are selfish when they drink. | 1.89 | −1.40 | −0.35 | 0.92 | 1.88 |

*Note.* Items are rank-ordered on the basis of their slope (discrimination) parameters.

Items included in the short form are marked with an asterisk.

**Table 7**

Calibrated Positive Expectancies Items

| Item Stem | Slope (Discrimination) | Location Thresholds | | | |
|---|---|---|---|---|---|
| | | Not at all vs. A little bit | A little bit vs. Somewhat | Somewhat vs. Quite a bit | Quite a bit vs. Very much |
| People have more fun at social occasions when they drink * | 2.63 | −1.70 | −0.54 | 0.70 | 1.59 |
| People feel happy when they drink * | 2.51 | −1.80 | −0.53 | 0.86 | 1.95 |
| People are outgoing when they drink * | 2.38 | −1.82 | −0.64 | 0.60 | 1.62 |
| Alcohol makes it easier to talk to people * | 2.20 | −1.33 | −0.35 | 0.76 | 1.69 |
| People forget their problems when they drink * | 1.87 | −1.13 | −0.18 | 1.02 | 1.96 |
| Drinking improves a person's mood * | 1.86 | −1.02 | 0.15 | 1.74 | 2.68 |
| People have more desire for sex when they drink * | 1.65 | −1.37 | −0.29 | 0.97 | 1.94 |
| Drinking eases physical pain | 1.41 | −1.04 | 0.15 | 1.55 | 2.77 |
| People sleep better when they drink | 1.26 | −1.13 | 0.13 | 1.56 | 2.76 |

*Note.* Items are rank-ordered on the basis of their slope (discrimination) parameters.

Items included in the short form are marked with an asterisk.