Drug and Alcohol Dependence 194 (2019) 59–68

Contents lists available at ScienceDirect

Drug and Alcohol Dependence

journal homepage: www.elsevier.com/locate/drugalcdep

Full length article

Approaches for creating comparable measures of alcohol use symptoms: Harmonization with eight studies of criminal justice populations

Andrea M. Hussong[a,*,1], Nisha C. Gottfredson[a,2], Dan J. Bauer[a,1], Patrick J. Curran[a,1], Maleeha Haroon[a,1], Redonna Chandler[b,3], Shoshana Y. Kahana[b,4], Joseph A.C. Delaney[c,5], Frederick L. Altice[d,6], Curt G. Beckwith[e,7], Daniel J. Feaster[f,8], Patrick M. Flynn[g,9], Michael S. Gordon[l,10], Kevin Knight[g,11], Irene Kuo[h,12], Lawrence J. Ouellet[i,13], Vu M. Quan[j,14], David W. Seal[k,15], Sandra A. Springer[d,16]

[a] University of North Carolina at Chapel Hill, United States
[b] National Institute on Drug Abuse/National Institutes of Health, United States
[c] University of Washington, Seattle, United States
[d] Yale School of Medicine, United States
[e] Alpert Medical School of Brown University, United States
[f] University of Miami, United States
[g] Texas Christian University, United States
[h] The George Washington University, United States
[i] University of Illinois at Chicago, United States
[j] Johns Hopkins University, United States
[k] Tulane University School of Public Health and Tropical Medicine, United States
[l] Friends Research Institute, United States

ARTICLE INFO

Keywords:
Data pooling
Drinking severity
Integrative data analysis

ABSTRACT

Background: With increasing data archives comprised of studies with similar measurement, optimal methods for data harmonization and measurement scoring are a pressing need. We compare three methods for harmonizing and scoring the AUDIT as administered with minimal variation across 11 samples from eight study sites within the STTR (Seek-Test-Treat-Retain) Research Harmonization Initiative. Descriptive statistics and predictive

* Corresponding author at: Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC 27599-3270, USA.
E-mail addresses: hussong@unc.edu (A.M. Hussong), gottfredson@unc.edu (N.C. Gottfredson), dbauer@email.unc.edu (D.J. Bauer), curran@unc.edu (P.J. Curran), maleeha@unc.edu (M. Haroon), shoshana.kahana@nih.gov (S.Y. Kahana), jacd@uw.edu (J.A.C. Delaney), frederick.altice@yale.edu (F.L. Altice), CBeckwith@Lifespan.org (C.G. Beckwith), dfeaster@biostat.med.miami.edu (D.J. Feaster), ibr@tcu.edu (P.M. Flynn), mgordon@friendsresearch.org (M.S. Gordon), k.knight@tcu.edu (K. Knight), ikuo@gwu.edu (I. Kuo), ljo@uic.edu (L.J. Ouellet), vquan1@jhu.edu (V.M. Quan), dseal@tulane.edu (D.W. Seal), sandra.springer@yale.edu (S.A. Springer).
[1] Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC 27599-3270, USA.
[2] 319C Rosenau Hall, CB #7440, Chapel Hill, NC 27599, USA.
[3] National Institute on Drug Abuse, 6001 Executive Boulevard, Room 5265, Rockville, MD 20892-9581, USA.
[4] Division of Biomedical Research Workforce, Office of Extramural Research, Office of the Director, National Institutes of Health, 6705 Rockledge Drive, Room 3534, Bethesda, MD 20892-7963, USA.
[5] Department of Epidemiology, University of Washington, USA.
[6] Yale University AIDS Program, 135 College Street, Suite 323, New Haven, CT 06510-2283, USA.
[7] The Miriam Hospital, 164 Summit Ave, Providence, RI 02906, USA.
[8] 1120 N.W. 14th Street, Clinical Research Building – Room 1064, Miami, Florida 33136, USA.
[9] Institute of Behavioral Research, Texas Christian University, TCU Box 298740, Fort Worth, TX 76129, USA.
[10] Friends Research Institute, Inc., 1040 Park Avenue, Suite 103, Baltimore, MD 21201, USA.
[11] Institute of Behavioral Research, Texas Christian University, Box 298740, Fort Worth, TX 76129, USA.
[12] George Washington University, Milken Institute School of Public Health, 950 New Hampshire Avenue NW, Suite 500, Washington, DC 20052, USA.
[13] Epi/Bio/COIP, School of Public Health (MC 923), University of Illinois at Chicago, 1603 W. Taylor Street, Chicago, IL 60612, USA.
[14] Johns Hopkins University, Center for Global Health, 415 N Washington St., Rooms 333-341, Baltimore, MD 21205, USA.
[15] 1440 Canal Street, Suite 2200, Mailstop #8319, Office 2224, New Orleans, LA, 70112, USA.
[16] Yale School of Medicine, Section of Infectious Diseases, Yale AIDS Program, 135 College Street, Suite 323, New Haven, CT 06510, USA.

https://doi.org/10.1016/j.drugalcdep.2018.10.003
Received 4 September 2018; Received in revised form 16 October 2018; Accepted 17 October 2018
Available online 23 October 2018

Data harmonization

validity results for cut-scores, sum scores, and Moderated Nonlinear Factor Analysis scores (MNLFA; a psychometric harmonization method) are presented.

Methods: Across the eight study sites, sample sizes ranged from 50 to 2405 and target populations varied based on sampling frame, location, and inclusion/exclusion criteria. The pooled sample included 4667 participants (82% male, 52% Black, 24% White, 13% Hispanic, and 8% Asian/ Pacific Islander; mean age of 38.9 years). Participants completed the AUDIT at baseline in all studies.

Results: After logical harmonization of items, we scored the AUDIT using three methods: published cut-scores, sum scores, and MNLFA. We found greater variation, fewer floor effects, and the ability to directly address missing data in MNLFA scores as compared to cut-scores and sum scores. MNLFA scores showed stronger associations with binge drinking and clearer study differences than did other scores.

Conclusions: MNLFA scores are a promising tool for data harmonization and scoring in pooled data analysis. Model complexity with large multi-study applications, however, may require new statistical advances to fully realize the benefits of this approach.

1. Introduction

The National Institute on Drug Abuse (NIDA) and other National Institutes of Health partners initiated the STTR (Seek-Test-Treat-Retain, referencing the HIV treatment cascade) Research Harmonization Initiative to create the largest prospective data collection and harmonization effort (Chandler et al., 2015). The initiative serves as an example for creating partnerships between government agencies and study investigators at the point of project inception with the goal of eventual data harmonization (i.e., creating comparable measures in different data sets for pooled analyses across studies). The collaboration involves 22 NIDA-funded studies that test interventions involving the STTR treatment cascade for criminal justice-involved (n = 12) and vulnerable populations with drug abuse (n = 10). Workgroups comprised of NIDA staff and study investigators identified core constructs and shared assessments that were administered across studies (with some exceptions). The result of this collaboration is a rich data set managed by a data coordinating center at the University of Washington.

Although large cross-national collaborations involving data harmonization are not new to the sciences (Dubrow and Tomescu-Dubrow, 2016), the STTR collaboration is among an increasing number of studies applying data harmonization methods to address public health issues ranging from aging (Gatz et al., 2015) and cancer (Adams et al., 2015) to treatments for college problem drinking (Huh et al., 2015) and adolescent depression (Brown et al., 2016). Data harmonization methods use resources efficiently to create large samples with greater statistical power and increased heterogeneity for testing population differences (Curran et al., 2017; Curran and Hussong, 2009; Hussong et al., 2013). The simultaneous analysis of data from different studies also permits direct evaluation of the extent to which effects replicate across independent studies. Harmonization studies, however, are not without their challenges (Schaap et al., 2011), particularly those involving data harmonization when pooled studies use non-identical measures.

Approaches to data harmonization can be categorized into logical versus analytic methods (Granda et al., 2010). Logical harmonization methods describe rule-based processes for making decisions about which items or measures in one study align with those in another. For example, some studies assume that measures administered identically in different studies are by default directly comparable across studies (even when larger batteries, assessment techniques, samples differ). One approach to harmonization is to rely on agreement among expert raters to determine whether item wording, instructions or response scales are (or can be made) comparable by collapsing across categories (i.e., a 5-point Likert response scale ranging from none to very often is collapsed to match a 2-point response scale ranging from none to any). Logical harmonization efforts yield pooled data sets with intuitively comparable items. Researchers then often turn to traditional scoring methods (i.e., using pre-determined scale cut-scores or sum scores) to create variables from the pooled data that are used in subsequent analyses.

Although widely used and easy to implement, cut-off and sum scores have well-documented psychometric limitations (Wainer and Thissen, 2001) that may be amplified in the context of integrative data analysis (Curran et al., 2018b). These scoring methods fail to capture meaningful individual variation due to which items on an instrument are endorsed (e.g., feeling guilty about drinking and alcohol-related injuries would be equally weighted indicators of problem drinking), how the measure is administered across studies (i.e., minor variation in item wording, item order, instructions or response scales as well as differences in administration due to language or preceding assessment tasks), and how individuals may differ from one another in how they interpret and respond to individual items (i.e., males may be more likely to endorse binge drinking as females even at the same level of drinking severity). Analytic harmonization methods use psychometric approaches to test assumptions about comparable item performance across studies and to create commensurate measures through equating procedures (Bauer and Hussong, 2009; Granda et al., 2010). One approach to analytic harmonization is moderated non-linear factor analysis (MNLFA, Bauer and Hussong, 2009; Curran et al., 2014). Applied within an Integrative Data Analysis framework, MNLFA draws upon psychometric theory to address the problem of creating comparable measures across studies. More specifically, MNLFA combines elements of nonlinear factor analysis and item response theory (IRT) to evaluate whether items perform comparably across pooled studies and to account for differences in item performance when creating variables for subsequent analyses. MNLFA models, however, can be complex to estimate and can require additional time and expertise (though see recent advances in partial automation of this model, Gottfredson et al., 2018).

To evaluate these relative benefits of the MNLFA approach, we compare scores drawn from three approaches to scoring the AUDIT (Babor et al., 2001) as administered similarly in eight STTR Criminal Justices (STTR-CJ) study sites. These include cut-scores, sum scores, and MNLFA scores. The AUDIT, created as a screening tool for problematic alcohol use, is strongly correlated with alcohol diagnoses (with median sensitivity and specificity rates of 0.86 and .89, Reinert and Allen, 2002). Research using the AUDIT relies on both cut-scores (Babor, 2001; Saunders et al., 1993, suggest a cut-score of 8) and sum scores (Chen et al., 2006; Knibbe et al., 2006) to provide dichotomous and continuous indicators of drinking severity, respectively.

Within- and between-study differences in the AUDIT can be accounted for when creating MNLFA scores by including these as covariates within the fitted model. More importantly, by including specific study membership as a covariate in these models we can account for clustering and differences that might exist across studies in AUDIT score performance due to differences in placement of AUDIT within a battery, modifications to AUDIT items, and/or differences in study populations that might interpret items differently. In contrast, cut-scores or sum scores are not informed by covariates. Core parameters in the MNLFA include the mean and variance of the latent factor (in this case drinking

severity) and the factor loadings and intercepts for the set of (AUDIT) items. Whereas an item intercept represents the predicted value of an AUDIT item when the latent variable is held at zero (i.e., items with higher intercepts are more commonly endorsed), the factor loading represents the predicted increment in a subject's response to the item associated with a one-unit increase in the latent variable (i.e., items with higher slopes are more strongly related to the latent construct or more discriminating between individuals with different levels of underlying drinking problems). In MNLFA, covariates (such as study membership) may be added to influence the parameters that define the latent factor. Given that the STTR CJ study samples differ in factors that are associated with drinking severity (e.g., male/female, race/ethnicity, HIV status; Galvan et al., 2002; Grant et al., 2004; Nolen-Hoeksema, 2004), we can include these covariates to determine whether study differences persist after accounting for sampling variation on these characteristics.

Using MNLFA, we can evaluate whether covariate differences are evident in overall drinking severity (both the mean and variance), in the extent to which individual items are indicative of drinking severity (factor loadings), and in individual item endorsement for those at the same level of drinking severity (item intercepts). Importantly, MNLFA permits us to differentially weight items in creating composite scores as a function of individual factors (e.g., male/female, race/ethnicity, HIV status) and study design features (e.g., study membership). Recent computer simulation studies have demonstrated that factor scores obtained in this way are psychometrically superior (i.e., in terms of true score recovery and predictor-criterion recovery) to those derived from more traditional methods that do not incorporate covariates (Curran et al., 2018a, 2016). Here we extend those simulation findings with an application to real data from a complex data harmonization project.

In sum, we contrast three scoring approaches for creating comparable measures for the AUDIT across 11 samples from eight of the STTR CJ cohort sites (note some sites contributed multiple subsamples). These approaches include traditional AUDIT cut-scores, sum scores based on logically harmonized AUDIT items, and MNLFA factor scores that account for posited differences in the structure of measurement. We evaluate differences between scores generated from these approaches by (a) evaluating differences in item performance as a function of study membership and individual difference factors (e.g., male/female, race/ethnicity, and HIV status of participants), (b) comparing descriptive differences across scores in means and sample variation, and (c) testing differences in how scores are associated with a measure of 30-day binge drinking.

2. Methods

2.1. Participants

The product of the STTR Data Collection and Harmonization Initiative developed by the National Institute of Allergy and Infectious Diseases (NIAID), the National Institute on Drug Abuse (NIDA), the National Institute of Mental Health (NIMH), and the Office of AIDS Research (OAR), the STTR CJ cohort emerged through a collaboration with researchers who harmonized and pooled data from 12 independent studies funded to evaluate the STTR treatment cascade for CJ involved individuals (Chandler et al., 2017). Participants for the current analysis come from eight CJ study sites that administered the AUDIT at baseline. Because some study sites collected more than one sample but used the same measures, we combine samples from the same site in our harmonization analysis. Table 1 reports demographic characteristics of each sample. Sample sizes ranged from 50 to 2405 and target populations varied based on sampling frame, location, and inclusion/exclusion criteria (see Table 2). The pooled sample included 4667 participants and was 82% male, 52% Black, 24% White, 13% Hispanic, and 8% Asian or Pacific Islander, with a mean age of 38.86 (range 18–74).

2.2. Procedures

Studies took place across various CJ settings, including jail, prison, and community supervision, and included both persons living with HIV (PLWH) and HIV-uninfected individuals. For the current analyses, we examined only data collected at baseline in each sample. (More information available at Chandler et al., 2017, 2015).

2.3. Measures

Particpicipants completed surveys in which they reported their demographic characteristics including age, gender identity (categorized here as male and female, given that not all studies measured transgender status and individuals were categorized as the gender to which they had transitioned), and race/ethnicity (collapsed to reflect white,

Table 1
Demographic characteristics for STTR CJ cohort study participants.

Study	N	Self-Identified Gender, %[a]		Race/ethnicity, %				Age, M (SD)
	Female	Male	White	Black	Asian/PI		Multiethnic	
BRIGHT								
BRIGHT 1	2405	18	82	28	52	0	4	38.99 (29.96)
BRIGHT 2	100	22	78	3	90	0	3	45.22 (7.76)
CARE								
CARE + Parole	138	20	80	41	19	0	6	34.41 (10.30)
CARE + RCT	112	42	58	4	88	0	8	32.90 (10.16)
CARE + RIDOC	250	6	94	47	15	0	8	42.23 (9.56)
NEW HOPE	122	19	81	12	23	0	1	45.64 (8.09)
START TOGETHER	195	0	100	22	46	1	1	37.90 (10.86)
STRIDE								
STRIDE 1	50	31	69	0	94	0	4	53.02 (6.36)
STRIDE 2	109	54	46	1	37	0	0	50.83 (9.04)
STT								
STT JAIL	276	20	80	8	81	0	1	40.91 (12.02)
STT PRISON	90	12	88	3	91	0	0	43.72 (10.34)
SUCCESS	56	11	89	9	70	0	7	37.16 (8.61)
VISTA[b]	378	0	100	0	0	100[b]	0	35.59 (5.97)

Note: CJ = Criminal Justice; STTR = Seek, Test, Treat, and Retain.

[a] Transgender participants were coded as the gender (male/female) they identified as transitioning to (due to small number of transgender participants across studies).

[b] Study took place in Vietnam and all participants were Asian.

Table 2
Inclusion/Exclusion Criteria and AUDIT Questions for STTR CJ Cohort Studies.

Sample	Inclusion/Exclusion Criteria[a]	Differences from Standard AUDIT Questionnaire
BRIGHT 1	Unknown HIV status; on probation or parole in Baltimore, MD or Providence/Pawtucket, RI and residence in the Baltimore or Providence/Pawtucket area throughout the study period	Items 9-10 have a response scale of (0) Never; (1) Less than monthly; (2) Monthly; (3) Weekly; (4) Daily or almost daily
BRIGHT 2	HIV-infected; on probation or parole in Baltimore, MD and planned residence in greater Baltimore area throughout the study period	Same alteration as BRIGHT 1.
CARE + Parole	Aged 18+; self-reported as HCV negative or unknown; on probation or parole; English speaking	Retrospection period of "in the previous 3 months"; binge criteria of 5 drinks for men or MTF transgender participants; criteria of 4 drinks for women or FTM participants
CARE + RCT	Aged 18+; HIV-infected; released from the correctional facility or half-way house ≤6 months ago and living in Washington, DC metropolitan community (not a restricted setting, e.g. half-way house) or currently detained in jail with anticipated release to community (not a restricted setting); reading at 8th grade level and English-speaking	Retrospection period of "in the 12 months before you were incarcerated"; binge criteria same as CARE + parole
CARE + RIDOC	Aged 18+; self-reported as HCV negative and documented HCV infection during Department of Corrections time with anticipated release between 3 and 12 weeks from enrollment; English-speaking	Retrospection period of "in the 3 months prior to incarceration"; binge criteria same as CARE + Parole
NEW HOPE	Aged 18+, HIV-infected, meeting DSM-IV criteria for opioid dependence, within CT corrections system and not pending trial for a felony, within 30 days of being released to greater New Haven, Hartford, Waterbury or Springfield areas or 30 days after release; English- or Spanish-speaking, no liver failure or grade IV hepatitis, no active opioid withdrawal, no receipt of methadone or buprenorphine/naloxone for treatment of opioid dependency, no participation in pharmacotherapy trial in the previous 30 days	Retrospection period of "in the year prior to this incarceration"
START TOGETHER	Men aged 18+ with either unknown or believed negative HIV status within 90 days of release from NY detention center entering a residential substance abuse treatment program	n/a
STRIDE 1	Aged 18+; HIV-infected; meeting DSM-IV criteria for opioid dependence; resident of Washington, DC with eligibility for medical entitlements; English- or Spanish-speaking; no current opiate medications for chronic pain conditions or need to be placed on such medications; no current methadone doses over 30 mg/day, no AST and ALT > 5× the ULN; no pregnancy or breast-feeding; no liver dysfunction; no suicidal ideation; no participation in pharmacotherapy trial in the previous 30 days	n/a
STRIDE 2	Aged 18+; HIV-infected; meeting DSM-IV criteria for opioid dependence; resident of Washington, DC with eligibility for medical entitlements; English-speaking	n/a
STT JAIL	Aged 18+; HIV-infected; detained in IL corrections system (jail); expecting to reside in Chicago after release	Retrospection period of "past 6 months"; binge criteria of 5 or more drinks
STT PRISON	Aged 18+; HIV-infected; recently released from IL corrections system (prison); enrollment was within 60 days of release	Retrospection period of "in the 6 months before you were in prison this last time"; binge criteria of 5 or more drinks
SUCCESS	Aged 18+; HIV-infected; detained or sentenced in jail or detention center and likely to leave within 6 weeks; no recent participation in randomized trial to improve retention in HIV care; English-speaking	Retrospection period of "in the 12 months before you came to jail"
VISTA	Aged 18+; recently released from drug treatment centers; HIV-infected; male; meet criteria for ART by Vietnam's national guidelines; not currently on ART; not enrolled in other HIV interventional research study	Retrospection phrased as "in the last 12 months" instead of "in the last year"

[a] Information in this column taken from Table 2 in Chandler et al. (2017). N/A denotes AUDIT was administered without any changes from the original instrument.

black, Asian, Hispanic, and other).

Participants also completed the AUDIT to assess problem drinking. Across studies, we followed standard AUDIT response scoring (see appendix) and coded response scales to range from 0 to 4 for items 1–8 assessing frequency of alcohol use, quantity of alcohol consumed during drinking episodes, binge drinking, inability to stop drinking once started, failing to meet obligations due to drinking, needing a first drink to get started in the morning, feelings of guilt or remorse following drinking, and memory loss due to drinking; and response scales of 0 (no), 2 (yes), or 4 (not in the past year) for items 9–10 assessing whether the participant or another indivdual had been injured as a result of drinking and whether another individual has been concerned about the participant's drinking habits. The standard AUDIT asks individuals to recall and answer these questions based on drinking habits over the past year, with items 9–10 also adding a score for drinking severity that happened outside of this one year retrospection period (Babor et al., 2001). Seven of the eight STTR CJ studies had response scales that were harmonizable with standard AUDIT scoring. Some samples, however, received an altered version of the original AUDIT measure (see Table 2) involving a change in the length of the reporting window (up to 12 months), change in the wording of the binge item question (to capture sex differences in recommended cut-points), or change in the response scale. Therefore, though standard AUDIT scores could be generated easily for nine of the eleven samples, they may reflect different criteria across samples. Such variations highlight the need to consider more principled approaches to developing comparable measures.

The study that was more challenging to logically harmonize with the standard AUDIT was the BRIGHT study (which consisted of data from both BRIGHT 1 and BRIGHT 2 samples). We equated two items (9 and 10) whose response scales differed from the original format and coded responses of '1' (less than monthly) as 2 ('yes, but not in the past year') and answers of > 1 (more than monthly) as '4' (yes, in the past year). We recognize, however, that a response of "less than monthly" could possibly indicate that it happened in the past year, making the scoring not entirely accurate for this population. We test this harmonization assumption in one of our scoring strategies.

We analyzed the AUDIT data using three different scoring approaches, including traditional cut-scores. To create cut-scores, sum scores using response scales for all ten items were examined and scores of 8 or more (for men) or 4 or more (for women and those over 60 years of age) were coded as positive screens (following Babor et al., 2001),

yielding a dichotomous indicator (0/1) of hazardous/harmful alcohol use. Additional scoring approaches are described below.

Participants also completed measures assessing their recent alcohol use based on a single item from the Alcohol Severity Index (McLellan et al., 1980). The following studies measured pre-incarceration alcohol use: CARE + RCT, CARE + RIDOC, NEW HOPE, STT JAIL, STT PRISON, and SUCCESS; whereas others asked about alcohol use prior to baseline (regardless of timing with respect to incarceration). In addition, timeframes for recent alcohol use varied, including 30 days (for NEW HOPE and START), 90 days (for BRIGHT, STT), and one year (for CARE + RCT and SUCCESS). Timeframe was not specified for STRIDE and START. Following, Matsuzaki et al. (2018), these items were collapsed across varying time frames. Dichotomized responses indicated engaging in recent alcohol use (1) or not (0).

2.4. Analytic plan

The analyses proceeded in three steps. First, we generated scores for the AUDIT based on three procedures: (1) cut-scores following logical harmonization of AUDIT items, (2) sum score based on summing the 10 AUDIT items, and (3) factor scores based on MNLFA. The first two techniques assume logical harmonization is sufficient to address differences in samples and measurement that may otherwise mascquerade as true score variance on the AUDIT. The last technique tests assumptions regarding logical harmonization and incorporates differences in item performance across studies into scoring. Second, we conducted descriptive statistics to examine similarities and differences in the scores derived from each of the three techniques. And, third, we examined predictive validity of all three scores relative to an indicator of 30-day binge drinking.

3. Results

3.1. Creating sum and MNLFA scores

We examined item distributions within and across studies and conducted a series of exploratory factor analyses (EFA) to assess dimensionality and local dependence (item redundancy). Based on these analyses, we selected items and transformed variables as follows in preparation for scoring. To create MNLFA and sum scores, we transformed all items to a binary response scale (yes/no) to avoid sparseness. We also eliminated quantity and frequency items in MNLFA and sum scores because they were redundant with the binge item and because they split off to form a separate factor (with the binge item) in an EFA. Without these items, the remaining items formed a unidimensional factor and were used in sum score as well as MNLFA scoring. (Percent endorsement of AUDIT items are listed for each study in Table 3.) An extension of factor analysis and item response theory, MNLFA generates scores that reflect not simply how many items were endorsed but which pattern of items were endorsed. Additionally, MNLFA allows covariates to directly affect both the mean and variance of the latent variable, referred to as *impact*, as well as measurement parameters linking items to the latent variable, referred to as *differential item functioning* (DIF). Details of our scoring analyses are presented in previous work (Curran et al., 2016).

Along with study membership, we attempted to control for demographic factors (e.g., race/ethnicity, male/female, whether participants were incarcerated, and whether participants reported on current or past use) as these may inform item performance and enhance score precision. However, these models did not converge (even when we limited the covariates to just race/ethnicity and male/female), probably due to substantial imbalances across studies giving rise to excessive multicollinearity with study membership (i.e., resulting in confounding of the study membership variables and covariates). Therefore, the MNLFA models only control for sample effects.

Sample membership was effect-coded to capture study differences. The largest study site (BRIGHT) was used as the reference group. Sample membership was allowed to affect both the mean and variance of the latent factor (assessing differences in drinking severity, impact) as well as the intercepts and loadings of selected items (assessing differences in item performance across studies, DIF). Following established model building strategies[16], factor mean and variance impact models were run separately and DIF was tested separately for each item. At this stage, any impact with $p < .10$ and either intercept or loading DIF with $p < .05$ was retained to provide a more inclusive impact model and avoid false detection of DIF due to omission of small but real impact effects. If only intercept DIF was detected, then only this parameter was considered further. If loading DIF was detected, then both the intercept and loading DIF parameters were carried forward. The identified effects were next evaluated simultaneously in a cumulative model. We then sequentially trimmed effects to arrive at a final scoring model, using a $p < .10$ threshold for impact and $p < .05$ threshold for DIF (using a familywise false detection rate of $p < .05$ for DIF). To avoid Type II error, nonsignificant effects from this final, trimmed model were not re-trimmed, so some effects in the final scoring model were non-significant.

There were some differences in factor mean (i.e., impact) and variance by sample. CARE and START had positive mean impact indicating that these two studies had higher MNLFA factor scores on the AUDIT than the average score across studies. NEW HOPE had negative mean impact. A small, negative mean impact effect of STRIDE was included in the final scoring model (as per trimming guidelines) but was not significant in the final model. No study had significant variance impact on the latent factor, although two were included in the final scoring model: NEW HOPE and STT.

There were also some sample differences in item factor means

Table 3
Item endorsements by STTR Sample.

	CARE	NewHope	STT	BRIGHT	Start	Stride	Success	Vista
How often have you had [4/5/6 – study-specified binge criteria] or more drinks on one occasion?	61%	28%	49%	29%	54%	31%	48%	53%
How often during [retrospective period] were you not able to stop drinking once you had started?	37%	20%	24%	10%	32%	18%	27%	100 %[a]
How often during [retrospective period] have you failed to do what was normally expected from you because of drinking?	33%	20%	25%	11%	27%	20%	30%	24%
How often during [retrospective period] have you needed a first drink in the morning to get yourself going after a heavy drinking session?	27%	19%	15%	7%	24%	17%	21%	9%
How often during [retrospective period] have you had a feeling of guilt or remorse after drinking?	36%	26%	25%	15%	32%	19%	30%	22%
How often during [retrospective period] have you been unable to remember what happened the night before because you had been drinking?	37%	18%	20%	15%	24%	19%	30%	10%
Have you or someone else been injured as a result of your drinking?	29%	6%	18%	12%	20%	8%	11%	21%
Has a relative, friend, doctor, or other health worker been concerned about your drinking or suggested you cut down?	37%	21%	26%	19%	33%	8%	25%	60%

[a] This item was scored as missing in the VISTA study because responses in the pooled data set indicated 100% endorsement in this study. The item was present in all other studies and included in analyses.

(intercept) and loadings by sample. There was no significant loading DIF in the final scoring model. This indicates that across this set of studies, all items were related to the underlying factor. STT was negatively associated with the intercept of Injury, indicating that, at the same level of drinking severity, participants in the STT study were more likely to endorse this item than were participants in other studies (i.e., the item was easier to endorse in STT). CARE and STT were positively associated with the intercept for the Binge item, indicating that this item was more severe in these studies or that, at the same level of drinking severity, participants in these studies were less likely to endorse this item than were participants in other studies (i.e., the item was harder to endorse). NEW HOPE and STRIDE (both specifically recruited persons with opioid disorders and did not exclude for comorbid AUDs/SUDs) were negatively associated with Binge (easier to endorse). Being in the VISTA study was not significantly related to the item assessing that alcohol caused you to fail to engage in expected behaviors but the effect was included in the final scoring model.

3.2. Comparison of MNLFA, sum, and AUDIT cut-off scores

Sum scores were more zero-inflated (i.e., piling up at zero) than MNLFA score, which had greater variation (Fig. 1 shows MNLFA (top) vs. sum scores (bottom) by study membership.) Whereas 52% of participants had a sum score of zero, only 7% of participants had MNLFA scores below -1.25 (i.e., in the lowest bar of the histogram shown in Fig. 2). Study effects are also variable across MNLFA and sum scores, though floor effects are clearly visible in sum scores and, to a lesser extent, in MNLFA scores. Sum scores have a mean of 1.69 and sd = 2.41; N = 310 people have missing scores due to missing responses to some items. MNLFA scores (distributed on a standard normal and thus with a different metric than sum scores), have a mean of -0.18 and sd = .93; no scores are missing due to the use of full information maximum likelihood to estimate scores from available items and study membership indicators.

The correlation between sum scores and MNLFA scores is $r = 0.90$. Fig. 2 shows the associations amongst the three scoring methods with univariate distributions on the diagonal. As seen here, MNLFA scores capture greater individual variation than either sum scores or cut-scores and permit covariates to better differentiate (and predict) individuals' drinking severity. As shown in Fig. 3, people who do not meet the cutoff criteria tend to have very low AUDIT sum scores (mean = median = mode = 0) and those who do meet criteria have a very wide range of sum scores.

3.3. Predictive validity of the three scoring methods

Table 4 compares results of a predictive validation analysis for the three AUDIT scoring methods using 30-day binge drinking as an outcome in separation regresion analyses. A direct comparison of regression coefficients was not possible because cutoff scores, sum scores, and MNLFA scores were differently scaled. To get around this problem, we rescaled the coefficients in the following way. We first calculated the mean sum score and the mean MNLFA score for individuals with cutoff = 0 and the mean sum score and mean MNLFA score for individuals with cutoff = 1. We took the difference of the mean sum score above and below the cutoff threshold and did the same thing for

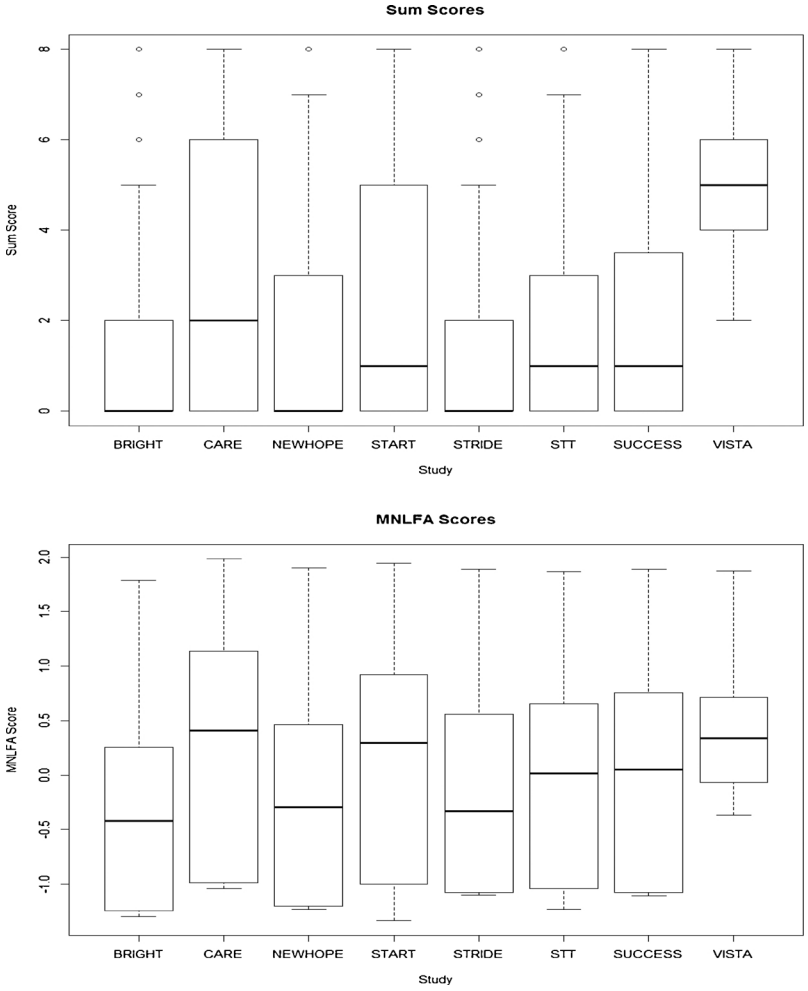

Fig. 1. MNLFA scores (top) and sum scores (bottom) plotted by study.

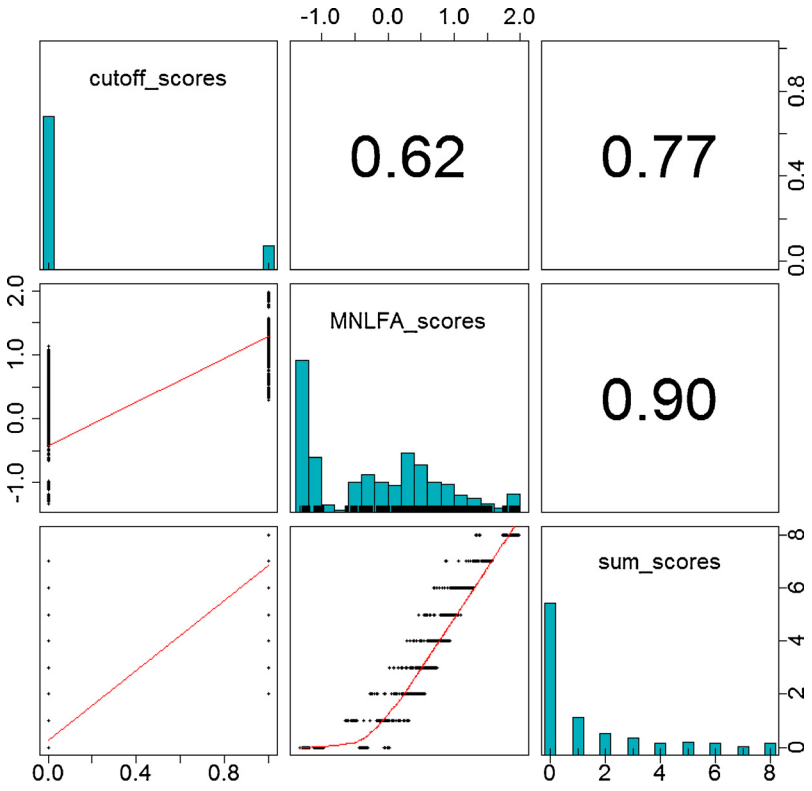

Fig. 2. Comparison of MNLFA, sum scores, and AUDIT cut-off scores.
NOTE: The diagonal is the univariate distribution for each type of AUDIT score (cutoff_scores = traditional threshold scoring approach; MNLFA_scores = moderated nonlinear factor analysis method for integrative data analysis; sum_scores = a count of symptoms endorsed). The lower diagonal depicts bivariate associations among scores, along with a loess curve to indicate the bivariate density of observations. The upper diagonal provides correlations among the scores.

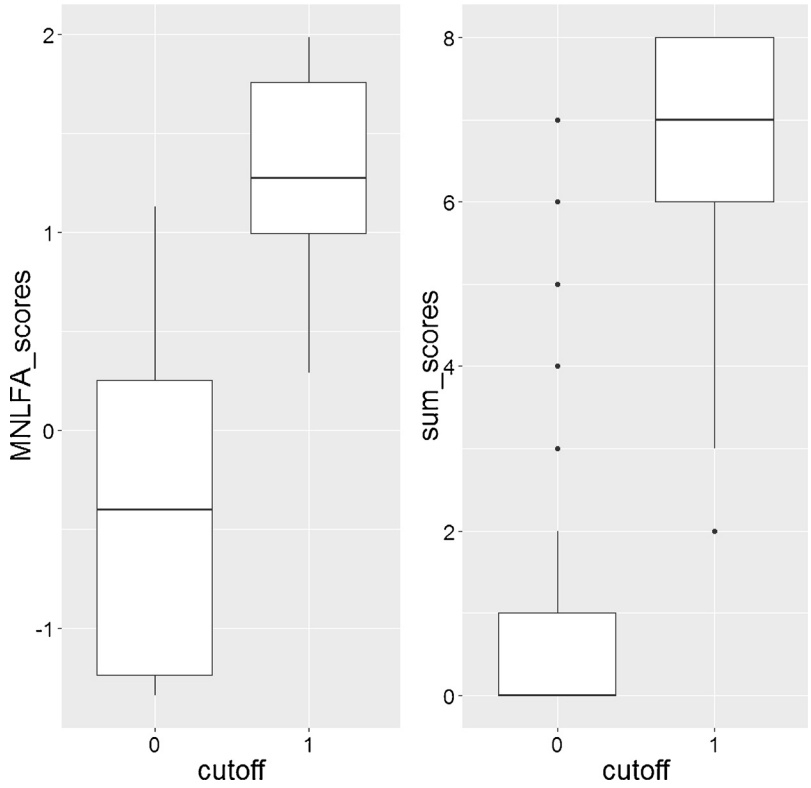

Fig. 3. Distribution of sum scores (left) and MNLFA scores (right) for those who do and do not meet AUDIT cut-score threshold.
NOTE: cutoff_scores = traditional threshold scoring approach; MNLFA_scores = moderated nonlinear factor analysis method for integrative data analysis; sum_scores = a count of symptoms endorsed.

MNLFA scores. Finally, we multiplied the regression coefficients by these difference values to obtain effect size estimates that are on a comparable scale. For sum scores, this was the difference between a sum score of 4.92 (for those above the cutoff) and a sum score of 1.48 (for those below the cutoff). For MNLFA scores, this was the difference between an MNLFA score of 0.97 and an MNLFA score of -.25. We found an average equivalent effect of 3.33 for sum scores and of 6.10 for MNLFA scores (as compared to 3.86 for cutoff scores). Thus, the

Table 4
Multiple Regression Results Testing Predictive Validity of Three AUDIT Scoring Methods: Any 30 day binge drinking.

	Cutoff Logistic Regression Results B (SE)	Sum Score Ordinary Least Squares Regression Results B (SE)	MNLFA Score Ordinary Least Squares Regression Results B (SE)
Intercept	**−1.41 (.15)**	**−1.95 (.18)**	**−.79 (.22)**
Female	**−.90 (.07)**	**−.27 (.06)**	**−.17 (.08)**
Black	**−.45 (.14)**	**−.50 (.18)**	**−.44 (.22)**
Hispanic	−.22 (.16)	−.25 (.20)	−.20 (.25)
White	−.25 (.15)	−.33 (.18)	−.03 (.23)
CARE	**1.02 (.12)**	**1.32 (.16)**	**.73 (.16)**
New Hope	**−.49 (.25)**	−.23 (.42)	−.57 (.32)
STT	**.91 (.09)**	.05 (.18)	**.74 (.19)**
Start Together	**.99 (.15)**	−.20 (.36)	−.01 (.41)
Stride	.07 (.23)	−.26 (.38)	−.27 (.42)
Success	**1.07 (.29)**	−.15 (.69)	**.39 (.64)**
Vista	**.63 (.18)**	−1.18 (2.42)	**−1.55 (.43)**
AUDIT Score	**3.86 (.29)**	**.97 (.04)**	**4.98 (.26)**
AUDIT Score x CARE	−.41 (.44)	**−.58 (.06)**	**−3.40 (.29)**
AUDIT Score x New Hope	−.71 (1.01)	**−.47 (.11)**	**−3.40 (.42)**
AUDIT Score x STT	.43 (.59)	**.77 (.14)**	**1.02 (.58)**
AUDIT Score x Start Together	–	**.96 (.33)**	1.72 (1.30)
AUDIT Score x Stride	1.43 (1.09)	.28 (.21)	−.10 (.90)
AUDIT Score x Success	–	.59 (.45)	.75 (1.87)
AUDIT Score x Vista	–	.75 (.83)	**2.83 (1.01)**

Note. The BRIGHT study site was used as the comparison group in analyses because it had the largest sample size. Estimates for cells with dashes could not be obtained due to empirical underidentification. Bold parameters are significant at $p < .05$ but given that scores are on different metrics, values from the different regressions are not directly comparable. All effects are reported controlly for other reported variables in the model.

effect for MNLFA scores was nearly double that of cut-off and average scores.

4. Discussion

In the current study, we demonstrate that MNLFA is a feasible approach to analytic harmonization for large data pooling projects such as that comprised of the eight sites from the 11 STTR CJ samples. Unlike cut-off and sum scores, MNLFA scores allow differences in measure performance across studies to be accounted for when creating scores. Although some applications of MNLFA also allow scoring to consider differences in other individual and study characteristics, model complexity and large study differences did not permit us to take advantage of this feature. Advances are needed in using MNLFA in more complex data pooling contexts to address this challenge in the future.

Although the three scores were notably correlated, they showed important differences in their distributional properties. As compared to cut-off and sum scores, MNLFA scores captured more individual variation in drinking severity and addressed problems with floor effects in AUDIT sum scores. Although MNLFA and sum score were highly correlated, there were differences in score distributions such that MNLFA scores had greater variability and less missing data than sum scores. MNLFA scores also captured greater sub-threshold differences in drinking severity that fell below the AUDIT cut-score as compared to sum scores.

Comparisons of measurement performance across these STTR CJ studies suggest that the AUDIT as a measure of drinking severity was highly robust. Still, MNLFA results did find evidence of modest study differences, with some studies having higher rates of drinking severity (i.e., impact) and higher rates of endorsing individual items (despite equated overall drinking severity rates, i.e., intercepts). These are likely in part due to differences in timeframes assessed across the studies. There were no differences in how items related to drinking severity (i.e., loadings or differential item functioning) across studies. Nonetheless, prior work shows that failure to include study differences in impact and intercepts can lead to biased scores (Curran et al., 2016).

Results of our predictive analysis are consistent with this possibility. The association between MNLFA scores and an indicator of past 30 days binge drinking was notably stronger than those between cut-off and sum scores with this indicator. In addition, truncated variation in cut-off scores, but not sum or MNLFA scores, led to problems evaluating study differences in the prediction of binge drinking. Study differences in AUDIT scores associated with binge drinking were detected in analyses using sum and MNLFA scores but not in those using cut-off scores. More main effect study differences, however, were evident in cut-off score analyses than sum and MNLFA score analyses. These results suggest that greater meaningful individual variation captured in MNLFA scores (versus cut-off and sum scores) resulted in stronger associations with binge drinking and (when compared to cut-off scores) detection of study differences in AUDIT-binge drinking associations.

In sum, results of the current study indicate that analytic harmonization methods like MNLFA outperform these other traditional scoring methods and caution against pooling data without attention to measurement differences. Limitations of the current study, however, include the inability to include covariates in the MNLFA scoring analyses due to model and data complexity and, unlike computer simulation studies, the lack of a "true score" model (to indicate which of these findings are best capturing known associations in the data). Nonetheless, computer simulation studies suggest that MNLFA scores contain less bias than sum scores (Curran et al., 2016). Cut-off scores also included drinking frequency and quantity (following established guidelines) though MNLFA and traditional CFA scores did not. Thus, as the STTR collaboration recognizes the need to consider multiple approaches to data harmonization within the rapidly developing field (see Chandler et al., 2017 for a discussion of item response theory co-calibration), we suggest that MNLFA may be a useful tool in conducting analytic harmonization analysis in future data pooling efforts.

Grant support

The opinions in this paper are those of the authors and do not reflect those of the National Institute on Drug Abuse, the National Institutes of Health, or the Department of Health and Human Services. Research reported in this publication is the result of secondary data analysis and was supported by and administrative supplement to 1R01DA034636-01A1 and K01 DA035153 from the National Institute on Drug Abuse. The STTR collaborative was funded by 5U01DA037702 and primary data collection for STTR studies was supported by grants R01DA030771, R01DA030762, 5R01DA030793, R01 DA030768, 5U01DA037702, R01 DA030747, and R01 DA030776. Dr. Chandler was substantially involved in U01 DA037702, consistent with her role as Scientific Officer. She had no substantial involvement in the other cited grants. The authors thank the other investigators, the staff, and particularly the participants of the individual STTR studies for their valuable contributions. A full list of participating STTR investigators and institutions can be found at http://www.sttr-hiv.org. All authors have reviewed this publication and consented for publication.

Contributors have all substantively added to the manuscript through a coordinated publications process associated with the STTR collaborative either through analyses, manuscript review, data coordination, hypotheses generation, and interpretation of findings. All authors have approved the final article.

Conflict of interests

None

Acknowledgements

None.

Appendix A. The Alcohol Use Disorders and Identification Test

Introduction: "Now I am going to ask you some questions about your use of alcoholic beverages during this past year."

1. How often do you have a drink containing alcohol?
   (0) Never (1) Monthly or less (2) Two to four times a month (3) Two to three times a week (4) Four or more times a week
2. How many drinks containing alcohol do you have on a typical day when you are drinking?
   (0) 1 or 2 (1) 3 or 4 (2) 5 or 6 (3) 7 or 9 (4) 10 or more
3. How often do you have six or more drinks on one occasion?
   (0) Never (1) Less than monthly (2) Monthly (3) Weekly (4) Daily or almost daily
4. How often during the last year have you found that you were not able to stop drinking once you had started?
   (0) Never (1) Less than monthly (2) Monthly (3) Weekly (4) Daily or almost daily
5. How often during the last year have you failed to do what was normally expected from you because of drinking?
   (0) Never (1) Less than monthly (2) Monthly (3) Weekly (4) Daily or almost daily
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?
   (0) Never (1) Less than monthly (2) Monthly (3) Weekly (4) Daily or almost daily
7. How often during the last year have you had a feeling of guilt or remorse after drinking?
   (0) Never (1) Less than monthly (2) Monthly (3) Weekly (4) Daily or almost daily
8. How often during the last year have you been unable to remember what happened the night before because you had been drinking?
   (0) Never (1) Less than monthly (2) Monthly (3) Weekly (4) Daily or almost daily
9. Have you or someone else been injured as a result of your drinking?
   (0) No (2) Yes, but not in the last year (4) Yes, during the last year
10. Has a relative, friend, doctor, or other health worker been concerned about your drinking or suggested that you should cut down?
    (0) No (2) Yes, but not in the last year (4) Yes, during the last year

References

Adams, R.N., Mosher, C.E., Blair, C.K., Snyder, D.C., Sloane, R., Demark-Wahnefried, W., 2015. Cancer survivors' uptake and adherence in diet and exercise intervention trials: an integrative data analysis. Cancer 121, 77–83. https://doi.org/10.1002/cncr.28978.

Babor, T.F., 2001. The Alcohol Use Disorders Identification Test. pp. 1–37.

Babor, T.F., Higgins-biddle, J.C., Saunders, J.B., Monteiro, M.G., 2001. The Alcohol Use Disorders Identification Test: Guidelines for Use in Primary Care. World Health Organization, Geneva, Switzerland.

Bauer, D.J., Hussong, A.M., 2009. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. Psychol. Methods Multi-Study Methods Build. Cumul. Psychol. Sci. 14, 101–125. https://doi.org/10.1037/a0015583.

Brown, C.H., Brincks, A., Huang, S., Perrino, T., Cruden, G., Pantin, H., Howe, G., Young, J.F., Beardslee, W., Montag, S., Sandler, I., 2016. Two-year impact of prevention programs on adolescent depression: an integrative data analysis approach. Prev. Sci. 74–94. https://doi.org/10.1007/s11121-016-0737-1.

Chandler, R., Gordon, M.S., Kruszka, B., Strand, L.N., Altice, F.L., Beckwith, C.G., Biggs, M.L., Cunningham, W., Chris Delaney, J.A., Flynn, P.M., Golin, C.E., Knight, K., Kral, A.H., Kuo, I., Lorvick, J., Nance, R.M., Ouellet, L.J., Rich, J.D., Sacks, S., Seal, D., Spaulding, A., Springer, S.A., Taxman, F., Wohl, D., Young, J.D., Young, R., Crane, H.M., 2017. Cohort profile: seek, test, treat and retain United States criminal justice cohort. Subst. Abuse Treat. Prev. Policy 12. https://doi.org/10.1186/s13011-017-0107-4.

Chandler, R.K., Kahana, S.Y., Fletcher, B., Jones, D., Finger, M.S., Aklin, W.M., Hamill, K., Webb, C., 2015. Data collection and harmonization in HIV research: the seek, test, treat, and retain initiative at the National Institute on Drug Abuse. Am. J. Public Health 105, 2416–2422. https://doi.org/10.2105/AJPH.2015.302788.

Chen, M., Miller, B.A., Grube, J.W., Waiters, E.D., 2006. Music, substance use and aggression. J. Stud. Alcohol 67, 373–381. https://doi.org/10.1097/OGX.0000000000000256.Prenatal.

Curran, P.J., Cole, V., Bauer, D.J., Hussong, A.M., Gottfredson, N., 2016. Improving factor score estimation through the use of observed background characteristics. Struct. Eq. Model. 23, 827–844. https://doi.org/10.1080/10705511.2016.1220839.

Curran, P.J., Cole, V., Giordano, M., Georgeson, A.R., Hussong, A.M., Bauer, D.J., 2017. Advancing the study of adolescent substance use through the use of integrative data analysis. Eval. Health Prof. https://doi.org/10.1177/0163278717747947.016327871774794.

Curran, P.J., Cole, V.T., Bauer, D.J., Rothenberg, W.A., Hussong, A.M., 2018a. Recovering Predictor-Criterion Relations Using Covariate-Informed Factor Score Estimates.

Curran, P.J., Cole, V., Bauer, D.J., Rothenberg, W.A., Hussong, A.M., 2018b. Recovering predictor-criterion relations using covariate-informed factor score estimates. Struct. Eq. Model. A Multidiscip. J. 25, 860–875.

Curran, P.J., Hussong, A.M., 2009. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol. Methods Multi-Study Methods Build. Cumul. Psychol. Sci. 14, 81–100. https://doi.org/10.1037/a0015914.

Curran, P.J., McGinley, J.S., Bauer, D.J., Hussong, A.M., Burns, A., Chassin, L., Sher, K., Zucker, R., 2014. A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. Multivariate Behav. Res. 49. https://doi.org/10.1080/00273171.2014.889594.

Dubrow, J.K., Tomescu-Dubrow, I., 2016. The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field. Qual. Quant. Int. J. Methodol. 50, 1449–1467. https://doi.org/10.1007/s11135-015-0215-z.

Galvan, F.H., Bing, E.G., Fleishman, J.A., London, A.S., Caetano, R., Burnam, M.A., Longshore, D., Morton, S.C., Orlando, M., Shapiro, M., 2002. The prevalence of alcohol consumption and heavy drinking among people with HIV in the United States: Results from the HIV cost and services utilization study. J. Stud. Alcohol 63, 179–186. https://doi.org/10.15288/jsa.2002.63.179.

Gatz, M., Reynolds, C.A., Finkel, D., Hahn, C.J., Zhou, Y., Zavala, C., 2015. Data harmonization in aging research: not so fast. Exp. Aging Res. 41, 475–495. https://doi.org/10.1080/0361073X.2015.1085748.

Gottfredson, N.C., Cole, V.T., Giordano, M.L., Bauer, D.J., Hussong, A.M., Ennett, S.T., 2018. Simplifying the Implementation of Modern Scale Scoring Methods With an Automated R Package: Automated Moderated Nonlinear Factor Analysis (aMNLFA). Manuscript submitted for publication. .

Granda, P., Wolf, C., Hadorn, R., 2010. Harmonizing survey data. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Pennell, B.-E., Smith, T.W. (Eds.), Survey Methods in Multinational, Multiregional, and Multicultural Contexts., Wiley Series in Survey Methodology. John Wiley & Sons Inc, Hoboken, NJ, US, pp. 315–332. https://doi.org/10.1002/9780470609927.ch17.

Grant, B.F., Dawson, D.A., Stinson, F.S., Chou, S.P., Dufour, M.C., Pickering, R.P., 2004. The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. Drug Alcohol Depend. 74, 223–234.

Huh, D., Mun, E., Larimer, M.E., White, H.R., Ray, A.E., Rhew, I.C., Kim, S., Jiao, Y., Atkins, D.C., 2015. Brief motivational interventions for college student drinking may not be as powerful as we think: an individual participant-level data meta-analysis. Alcohol. Clin. Exp. Res. 39, 919–931. https://doi.org/10.1111/acer.12714.

Hussong, A.M., Curran, P.J., Bauer, D.J., 2013. Integrative data analysis in clinical psychology research. Annu. Rev. Clin. Psychol. 9, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522.

Knibbe, R.A., Derickx, M., Kuntsche, S., Grittner, U., Bloomfield, K., 2006. A comparison of the alcohol use disorder identification test (Audit) in general population surveys in nine european countries. Alcohol Alcohol. 41. https://doi.org/10.1093/alcalc/agl072.

Matsuzaki, M., Vu, Q.M., Gwadz, M., Delaney, J.A.C., Kuo, I., Perez Trejo, M.E., Cunningham, W.E., Cunningham, C.O., Christopoulos, K., 2018. Perceived access and barriers to care among illicit drug users and hazardous drinkers: findings from the seek, test, treat and retain data harmominization initiative (STTR). BMC Public Health 18, 366. https://doi.org/10.1186/s12889-018-5291-2.

McLellan, A.T., Luborsky, L., O'Brien, C.P., Woody, G.E., 1980. An improved diagnostic instrument for substance abuse patients: the addiction severity index. J. Nerv. Ment. Dis. 168, 26–33.

Nolen-Hoeksema, S., 2004. Gender differences in risk factors and consequences for alcohol use and problems. Clin. Psychol. Rev. 24, 981–1010.

Reinert, D.F., Allen, J.P., 2002. The alcohol use disorders identification test (AUDIT): a review of recent research. Alcohol. Clin. Exp. Res. 26, 272–279.

Saunders, J.B., Aasland, O.G., Babor, T.F., De La Fuente, J.R., Grant, M., 1993. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II.

Addiction 88, 791–804. https://doi.org/10.1111/j.1360-0443.1993.tb02093.x.
Schaap, L.A., Peeters, G.M., Dennison, E.M., Zambon, S., Nikolaus, T., Sanchez-Martinez, M., Musacchio, E., Van Schoor, N.M., Deeg, D.J.H., 2011. European project on osteoarthritis (EPOSA): methodological challenges in harmonization of existing data from five European population-based cohorts on aging. BMC Musculoskelet. Disord. 12. https://doi.org/10.1186/1471-2474-12-272.
Wainer, H., Thissen, D., 2001. True score theory: the traditional method. In: Thissen, David, Wainer, H. (Eds.), Test Scoring. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, pp. 23–72.