Behavior Research Methods (2021) 53:1031–1045
https://doi.org/10.3758/s13428-020-01472-7

Harmonizing altered measures in integrative data analysis: A methods analogue study

Andrea M. Hussong[1] · Daniel J. Bauer[1] · Michael L. Giordano[1] · Patrick J. Curran[1]

Published online: 16 September 2020

Abstract

In the current study, we used an analogue integrative data analysis (IDA) design to test optimal scoring strategies for harmonizing alcohol- and drug-use consequence measures with varying degrees of alteration across four study conditions. We evaluated performance of mean, confirmatory factor analysis (CFA), and moderated nonlinear factor analysis (MNLFA) scores based on traditional indices of reliability (test–retest, internal, and score recovery or parallel forms) and validity. Participants in the analogue study included 854 college students (46% male; 21% African American, 5% Hispanic/Latino, 56% European American) who completed two versions of the altered measures at two sessions, separated by 2 weeks. As expected, mean, CFA, and MNLFA scores all resulted in scales with lower reliability given increasing scale alteration (with less fidelity to formerly developed scales) and shorter scale length. MNLFA and CFA scores, however, showed greater validity than mean scores, demonstrating stronger relationships with external correlates. Implications for measurement harmonization in the context of IDA are discussed.

Keywords Integrative data analysis · Data pooling · Alcohol consequences · Drug consequences · Harmonization

In the context of data pooling, measurement harmonization encompasses approaches designed to improve the comparability, in meaning and metric, of scores derived from different measures collected across studies and/or other known groups (Bauer & Hussong, 2009; Curran & Hussong, 2009; Steinberg & Thissen, 2013). Some techniques address this challenge by using meta-analytic approaches that first estimate effects of interest within study and then rely on analysis of summary statistics across studies to obtain a generalizable effect (Cooper et al., 2009). A complementary set of approaches aims to pool participants from independent studies into a single analysis with the goal of increasing statistical power, testing patterns of replication of effects across studies, and addressing questions that may be infeasible to test in the individual contributing studies alone. Data pooling techniques in this latter category include individual participant data meta-analysis (Pigott, Williams, & Polanin, 2012) and integrative data analysis (IDA; Hussong, Curran, & Bauer, 2013). These latter approaches have been used to study moderating factors of brief motivation interviewing techniques in college samples (Huh et al., 2015; Mun et al., 2015), risk trajectories underlying psychopathology shown by children of alcoholic versus non-alcoholic parents (Hussong et al., 2009; Hussong, Wirth, et al., 2007), and measurement equivalence in diagnostic symptoms of nicotine dependence (Rose, Dierker, Hedeker, & Mermelstein, 2013).

Several approaches to measurement harmonization populate this literature, most of which can be characterized as using logical harmonization (i.e., aligning 'like' items across measures from different studies based on face validity or expert judges), analytic harmonization (i.e., using psychometric analyses to test assumptions about item equivalence across instruments from different studies to obtain comparable measurement), or some combination of the two. Although logical harmonization is often a starting point in creating commensurate measures for pooling data across studies, assumptions about comparability of item performance across studies remain untested. Analytic harmonization not only tests the viability of such assumptions but also allows incorporation of item differences into scoring, creating potentially more comparable scales across studies.

In the current study, we focus on the problem of measurement harmonization in the context of IDA. Two of the most vexing sources of threats to effective harmonization in IDA

✉ Andrea M. Hussong
hussong@unc.edu

[1] University of North Carolina at Chapel Hill, UNC-CH, CB #3007, Chapel Hill, NC 27599-3007, USA

are study differences in samples and in measurement. Because these two sources are often inextricably tied in applications of IDA (i.e., pooled studies vary in both), it has been difficult to make recommendations for approaches to measurement harmonization specifically. For this reason, we compared various approaches to measurement harmonization using an analogue design in which participants from a single population completed altered measures of the same construct in four randomly assigned experimental conditions paralleling a four-study IDA.

Analytic harmonization in IDA

In the current study, we evaluated a key tool for analytic harmonization in IDA known as moderated nonlinear factor analysis (MNLFA; Bauer, 2017; Bauer & Hussong, 2009). IDA is a methodology that involves the simultaneous analysis of item-level data in pooled analyses (Curran & Hussong, 2009). In practice, IDA often uses a logical harmonization approach to align data from independent studies prior to data pooling. We have advocated, however, that analytic harmonization techniques are a necessary additional step in IDA to test whether assumptions about item alignment and thus measurement comparability across studies are consistent with the observed data. MNLFA is a highly flexible approach to analytic harmonization (though it has other applications as well, Bauer, 2017) and blends the traditions of confirmatory factor analysis (CFA; Bollen & Hoyle, 2012) and item response theory (IRT; Steinberg & Thissen, 2013) to create factor scores for measures that may differ to some degree in item content (e.g., instructions, item wording, response scales, scale length) across studies. The goal of MNLFA is not to create a set of interchangeable items across studies. Rather, the goal is to use both common items (i.e., identically administered or logically harmonized items) and unique items (i.e., items available in only one or a sub-set of studies) across studies to infer scores for an underlying "harmonized factor" representing the construct of interest. Provided that enough (but not necessarily all) items measure the construct in the same way across studies – an empirically testable assumption – harmonized factor scores can then be compared directly in subsequent analyses to test hypotheses. Importantly, scoring through MNLFA accounts for the presence of some items that, despite being logically harmonized, do not in fact measure the construct equivalently across studies.

More specifically, MNFLA assumes that the set of items administered across studies assesses a single construct defined by a shared underlying factor, η (for simplicity we assume a unidimensional model; the generalization to multifactorial structures is straightforward). For binary items, the relationship of a given item to the factor can be expressed as

$$\text{logit}\left(\mu_{ij}\right)=\nu_{ij}+\lambda_{ij}\eta_j \quad (1)$$

where μ_{ij} represents the probability that item i will be endorsed by person j. This probability is determined by a logistic relationship to the latent factor defined by an intercept ν_{ij} and a factor loading (slope) λ_{ij}, consistent with an item-level CFA or two-parameter logistic IRT model.

Study (or other covariate) differences in MNLFA are then modeled in four ways. The first two involve indices of impact or the extent to which studies differ in the latent factor mean as well as the latent factor variance. For example, study A may have a higher mean level of alcohol-related consequences (a latent construct) because it contains more high-risk individuals than study B. Similarly, study A may have a higher variance in alcohol-related consequences because the individuals in the sample reflect a wider range of risk than those in study B. Although impact is usually taken to represent valid study differences, in some cases it may also be due to measurement differences. For instance, variation in instructions or item wording across studies that results in across-the-board higher endorsement rates in study A than study B would manifest as mean impact. These two sources of impact (sample and measurement differences) are intractably intertwined but for IDA they are both considered nuisance variance ideally separated from variance in the construct of interest.

In MNLFA, differences in the latent factor mean and variance are random effects modeled as:

$$\alpha_j = \alpha_0 + \sum_{p=1}^{P}\gamma_p x_{pj} \quad (2)$$

$$\psi_j = \psi_0 \exp\left(\sum_{p=1}^{P}\beta_p x_{pj}\right) \quad (3)$$

The scale for the latent factor is typically set by fixing α_0 (the baseline factor mean) to zero and ψ_0 (the baseline factor variance) to one, standardizing the factor when all p covariates are scored zero. The estimates γ_p (factor mean impact) and β_p (factor variance impact) serve to deterministically shift the mean and variance of the factor, α and ψ, respectively, from these baseline values as a function of $\boldsymbol{x}$. This differs from a standard CFA or IRT analysis in which the mean and variance of the factor would typically each be assumed equal across persons but is similar to how "multiple groups" CFA and IRT models are scaled (i.e., standardizing the factor in a reference group). The effect of a covariate on the factor mean, denoted by γ, and variance, denoted by β, is referred to as impact.

Another way in which study differences can manifest in MNLFA is through differential item functioning or DIF. Like traditional CFA, intercepts and slopes vary across items (the i subscript) but, unlike traditional CFA, they can also vary across persons (the j subscript) as a function of covariates, for instance study membership. The following equations describe how the values of the item parameters can vary across studies,

generically denoted here via coding variables as $x_1, x_2, \ldots, x_P$:

$$\nu_{ij} = \nu_{0i} + \sum_{p=1}^{P} \kappa_{pi} x_{pj} \quad (4)$$

$$\lambda_{ij} = \lambda_{0i} + \sum_{p=1}^{P} \omega_{pi} x_{pj} \quad (5)$$

Item intercepts are denoted ν_{ij} and item loadings are denoted λ_{ij}, and these may vary in value over persons (j) and items (i) as a function of study membership as represented by coding variables within the set of covariates, or x's. The baseline values of the intercept and loading for an item are denoted as ν_{0i} and λ_{0i}. Parameters denoted by a κ (for item intercepts) or ω (for item loadings) reflect DIF (in the IRT tradition) or non-invariance (in the CFA tradition). Generalizing from these traditions, MNLFA provides a flexible framework for testing patterns of DIF in a set of items governed by an underlying latent factor as a function of study membership. (Indeed, MNLFA provides even greater flexibility to test whether a set of categorical and/or continuous covariates uniquely as well as interactively account for DIF; see Bauer, 2017; Bauer & Hussong, 2009). Intercept DIF indicates that item endorsement differs across studies above and beyond what one would expect based on study differences in the latent variable mean alone. Loading DIF, in contrast, indicates the extent to which items are linked to the underlying latent factor differentially across studies. This may occur, for example, due to differences in item wording across studies or to differences in how groups of individuals interpret an item (e.g., girls and boys are well known to differentially endorse 'cries a lot' as an indicator of depression, even at the same levels of underlying depression; Steinberg & Thissen, 2013). In sum, MNLFA allows us to detect study (and other covariate) differences in our measures related to factor means and variances as well as in item intercepts and slopes (loadings) and to take these differences into account when creating harmonized scores for subsequent hypothesis testing.

An analogue IDA design

The purpose of the current study was to evaluate the performance of MNLFA as a method of analytic harmonization for IDA across altered measures, holding sample characteristics constant. More specifically, we used MNLFA to harmonize addictions-related measures from a set of four "analogue" studies and evaluated the performance of resulting scores relative to traditional CFA scores (that do not take DIF into account) and to widely used mean scores (created by averaging or unit-weighting items without consideration of DIF). The analogue study design provides a novel approach to evaluating the performance of statistical techniques and augments work by our team that uses computer simulation approaches for this purpose. Based on simulation results, we know that MNLFA scores, standard CFA scores, and mean scores tend to be highly correlated (Curran, Cole, Bauer, Hussong & Gottfredson, 2016). However, we also know that MNLFA scores have modestly higher correlations with the true underlying factor than standard CFA or mean scores (Curran et al., 2016) and demonstrate considerably less bias when used in follow-up models examining predictor-criterion relationships (Curran, Cole, Bauer, Rothenberg & Hussong, 2018).

A problem inherent in simulation studies, however, is that the data must be generated according to a known model. Thus, these simulation studies show that when item responses are generated from an MNLFA model with both DIF and impact, MNLFA scores outperform CFA and mean scores that ignore DIF and impact. However, in practice, we do not know the true measurement model that underlies our observed data. More importantly, real data are messier than computer generated data and contain influences between true scores and random measurement error that produce observed response patterns. A wide variety of biasing sources are discussed in the psychometric literature (Steinberg & Thissen, 2013) and include testing conditions (e.g., lighting, noise, temperature, the presence of others), preferred response styles (e.g., for extremes, mid-scores), anchor wording on Likert scales (e.g., 'none' versus 'zero'), item context (e.g., priming due to proceeding items, fatigue due to test length), and item phrasing. Importantly, individuals in a sample will differ in the extent to which these sources of bias influence their response patterns. Thus, real data have much more noise than computer simulated data and our findings regarding the performance of statistical techniques based on computer simulation studies (a design that favors internal validity) may not easily generalize to real-world applied data analyses (a design that favors external validity).

The current study

To address this gap in the literature, we created a real-world methodological evaluation of measurement harmonization approaches targeting altered measurement. Specifically, we created four, human subjects, analogue studies to evaluate the performance of MNLFA as an analytic harmonization technique as compared to CFA and mean scores. Participants were drawn from the same population and randomly assigned to analogue study conditions in which measures were administered in increasingly altered forms, creating an analogue to a four-study IDA. Using a repeated measures design, the same participants took part in multiple studies, permitting us to evaluate the extent to which we could recover similar scores for the same people under different measurement conditions.

Using the analogue IDA design, we tested which scoring method (MNFLA, CFA and mean scores) resulted in optimal reliability and validity when conducting real-world

harmonization. Because these techniques have much promise in the field of addictions, we grounded the study in understanding college students' alcohol- and drug-related consequences. As the scope of addiction science widens, the challenge and need grow for methods that integrate findings across studies to advance a cumulative approach to building knowledge. Given tight funding environments, the growing expense and significant time investment of new data collections, and the plethora of high-quality data sets available in many areas of science, secondary data analysis is clearly a necessary and efficient platform for testing a host of scientific questions. This case has certainly been made for the study of addictions (Conway et al., 2014). Often more powerful than the secondary analysis of a single dataset, however, are techniques for data pooling that permit researchers to answer novel questions beyond those addressed by individual contributing studies. Yet optimizing techniques for data pooling remains an active area of study and a key area of needed optimization concerns measurement harmonization.

We posited that scores would have lower reliability across studies given increasing scale alteration (with less fidelity to formerly developed scales and shorter scale length in our final study condition); we also anticipated that scores based on MNLFA and CFA would show greater test–retest reliability (comparability across studies) than mean scores because simulations show them to be more highly correlated with true scores (reliability hypothesis 1). We also predicted that MNLFA scores would show fewer study differences than CFA or mean scores given that MNLFA scores account for all aspects of differential item functioning which, if ignored, might produce artificial study differences in the factor scores (harmonization hypothesis 2). Similarly, we posited that MNLFA scores would reduce study differences in associations between harmonized measures and external correlates to a greater extent than would CFA or mean scores (validity hypothesis 3). Finally, we anticipated that this pattern of findings would generalize across measures of alcohol- and drug-related consequences but that MNLFA would outperform CFA and mean score measures in terms of validity particularly when greater differential item functioning was present in a measure (generalizability hypothesis 4).

Method

Sample

In the Real Life Experiences of University Students (REAL-U) study, we created a recruitment pool from a list of 8995 undergraduates randomly sampled from university registrar records (with oversampling for males and African Americans who were underrepresented in the student body) and 57 undergraduates who contacted us directly about the study. We invited the resulting 9052 students via email to complete a screening survey. Inclusion criteria were being age 18 to 23 and reporting alcohol use in the past year; 1403 (15.4%) of those in the recruitment pool completed the screening survey prior to study closure, of whom 1141 (81.3%) were eligible. Of those eligible to participate, 854 students (75%) completed the first session and 840 completed both sessions (for a 98% retention rate). The sample of 854 participants was 46% male, multi-ethnic (21% African American, 5% Hispanic/Latino, 56% European American, 11% Asian, 6% multi-racial, and < 1% Native American/Alaskan Native or Pacific Islander or unknown), more likely to include first years and seniors (29% each) than sophomores and juniors (20% each) and had a cumulative GPA of 3.23 (SD = .52). In comparison, the larger undergraduate body of this institution was 42% male, multi-ethnic (8% African American, 5% Hispanic/Latino, 65% European American, 9% Asian, 4% multi-racial, and 3% Native American/Alaskan Native or Pacific Islander or unknown), approximately evenly split (21–27%) across matriculation status and had a cumulative GPA of 3.17. Overall, the analysis sample was highly comparable to the student body, though more ethnically diverse (by design).

Procedures

Participants completed two testing sessions separated by 2 weeks (T1 and T2 in Table 1). In each session, participants completed one of two surveys (A or B) that each contained some scales that were altered from their original form across surveys to test hypotheses about harmonization (including alcohol- and drug-consequences measures, see Table 1, point 1) and others that were held constant over surveys in their original form to serve as validity measures (see Table 1, point 2). Across the two surveys, alcohol- and drug-consequence measures were administered in four forms within increasing alterations (described below). Each altered measure represented a different analogue "study" (studies 1–4 in Table 1). To avoid participant fatigue and excessive redundancy, survey A contained measures for study 1 and study 3 and survey B contained measures for study 2 and study 4. Participants were randomly assigned to one of four conditions (AA, BB, AB, BA) that fully crossed survey administration (A or B) by session (T1 or T2; see Table 1). In these sessions, participants completed consent procedures (first session only), their randomly assigned computerized survey (A or B), and a lab task (second session only). Participants unable to attend the second session in person completed batteries online (n = 17). Sessions lasted 75–90 min and participants received a $20 and a $25 incentive for completing each session, respectively.

Table 1 Key elements of analogue IDA design

		EACH PERSON WAS RANDOMIZED TO ONE OF FOUR EXPERIMENTAL CONDITIONS							
		AA		BB		AB		BA	
SURVEY	ANALOGUE STUDY	SESSION T1	SESSION T2	SESSION T1	SESSION T2	SESSION T1	SESSION T2	SESSION T1	SESSION T2
A	1	X	X			X			X
	3	X	X			X			X
B	2			X	X		X	X	
	4			X	X		X	X	

Point 1. The analogue study design included altered measures for alcohol- and drug-consequences to test generalizability of findings

Point 2. Validity measures included descriptive and injunctive peer norms, parent attitudes, negative urgency, positive urgency, and sensation seeking

Point 3. Alcohol- and drug-consequence measures were scored using three techniques: mean, CFA factor, and MNLFA scores

Point 4. Analytic samples varied to match design features: test–retest analyses used participants in conditions AA and BB; parallel forms analysis used participants in conditions AB and BA as did validity analyses

Altered measures for alcohol- and drug-related consequences

Original measures were drawn from the PHEN-X battery (Conway et al., 2014); all had previously demonstrated reliability and validity. We used the Rutgers Alcohol Problems Inventory (the RAPI; White & Labouvie, 1989) to assess alcohol-related consequences (18 items) and a parallel version to assess drug-related consequences (23 items). To simulate variation in measures administered across four independent studies that might comprise a data harmonization project, we created four versions of each measure. These four versions contained one of two sets of item stems for each measure (those from the original measures versus altered versions we created) and one of two response scales (those from the original measure and an altered response scale created to be logically harmonizable with the original response scale). We made alterations to item wording, directions, and response scales to form measures in studies 2–4 based on other established consequences measures (Fromme, Stroot, & Kaplan, 1993; Ham, Stewart, Norton, & Hope, 2005; Leigh & Stacey, 1993). Across the four studies, alterations to the original battery were increasingly severe and cumulative; they included administration of the original unaltered measure (study 1), altered stems for half of the items (study 2), the collapsible response scale for all items (study 3), and dropping half of the items (those with formerly altered stems) and adding altered stems for the remaining items along with the collapsible response scale (resulting in a short form for study 4). Study 1 and 3 versions of the measure appeared in separate parts of survey A and study 2 and 4 versions of the measures appeared in separate parts of survey B. (See Table 1 for design overview and Table 2 for item crosswalk for alcohol-related consequences).

Standard validity measures

To evaluate validity of scores derived from psychometric harmonization of the four studies, we assessed common correlates of substance use in college students including descriptive and injunctive peer norms, parent attitudes, negative urgency, positive urgency, and sensation seeking (see Table 1, point 2). We again selected measures from the PHEN-X battery with demonstrated reliability and validity. Traditional scoring procedures were used for each standard measure.

To evaluate descriptive norms, injunctive norms, and parent attitudes, participants completed an expanded measure of norms based on that developed in the Monitoring the Future Study (Ennett et al., 2006 as based on Johnston, O'Malley, Bachman & Schulenberg, 2013). Participants answered nine items about friends' use (descriptive norms), friends' perceptions (injunctive norms), and parents' attitudes regarding the participant using alcohol, tobacco, marijuana/hashish, Ritalin, OxyContin/pain killers, and other drugs using a 5-point response scale. We created a mean score to index peer descriptive norms (Session 1: M = 1.35, SD = 0.58, $\alpha = 0.84$; Session 2: M = 1.32, SD = 0.58, $\alpha = 0.84$), peer injunctive norms (Session 1: M = 2.30, SD = 0.66, $\alpha = 0.87$; Session 2: M = 2.32, SD = 0.65, $\alpha = 0.86$) and parent attitudes (Session 1: M = 1.46, SD = 0.37, $\alpha = 0.76$; Session 2: M = 1.49, SD = 0.44, $\alpha = 0.84$) in subsequent analyses.

Participants also completed the Urgency Premeditation Planning Sensation Seeking Impulsivity Scale-Revised (UPPS-R, Lynam, Smith, Cyders, Fischer, & Whiteside, 2007). For the current manuscript, we included scores for three subscales assessing negative urgency (12 items), sensation seeking behavior (12 items), and positive urgency (14 items). Participants completed the measure using a 4-point Likert scale ranging from "agree strongly" to "disagree

Table 2 Alcohol-related consequence measure used in IDA analogue design

	Study 1	Study 2	Study 3	Study 4	MNLFA Item DIF by Study ALCOHOL	MNLFA Item DIF by Study DRUG
	Same Response Scale for Studies 1 and 2: None (0), 1–2 times (1), 3–5 times (2), More than 5 Times (3), Refuse to answer (.)		Same Response Scale for Studies 3 and 4: Never (0), Once (1), Twice (2), 3–5 times (3), 6–9 times (4), 10 or more times (5), Refuse to answer (.)		C1: Contrast Code comparing Studies 1–3 vs. 4 C2: Studies 1–2 vs. 3 C3: Study 1 vs. 2	
1	Got into fights with other people (friends, relatives, strangers)	Got into fights with other people (friends, relatives, strangers)	All item stems the same in studies 2 and 3.	Gotten into physical fights when drinking	Threshold DIF: C1	
2	Went to work or school high or drunk	Gone to class or a job when drunk		(Empty Gray Box indicates Dropped Item)		
3	Caused shame or embarrassment to someone	Made others ashamed by your drinking behavior or something you did when drinking			Threshold DIF: C2	
4	Neglected your responsibilities	Neglected your responsibilities		Neglected your obligations, your family, or your work for two or more days in a row because you were drinking	Threshold DIF: C1,C2	Threshold DIF: C1 Loading DIF: C1
5	Relatives avoided you	Family members rejected you because of your drinking				
6	Felt that you needed more alcohol than you used to in order to get the same effect	Felt that you needed more alcohol than you used to in order to get the same effect		Needed to drink more and more to get the effect you want		
7	Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)		Tried to cut down or quit drinking or using alcohol		
8	Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking		Felt sick, shaky, or depressed when you stopped drinking		
9	Noticed a change in your personality	Acted in a very different way or did things you normally would not do because of your drinking			Loading DIF: C2 Threshold DIF: C2,C3	
10	Felt that you had a problem with alcohol	Felt that you had a problem with alcohol		Thought you might have a drinking problem		
11	Wanted to stop drinking but couldn't	Tried unsuccessfully to stop drinking				
12	Suddenly found yourself in a place that you could not remember getting to	Awakened the morning after some drinking the night before and could not remember a part of the evening.			Threshold DIF: C3	Threshold DIF: C2, C3
13	Passed out or fainted suddenly	Passed out after drinking			Threshold DIF: C3	Threshold DIF: C3
14	Had a fight, argument, or bad feeling with a friend	Had a fight, argument, or bad feeling with a friend		Drinking created problems between you and a near relative or close friend	Threshold DIF: C1	
15	Kept drinking when you promised yourself not to	Kept drinking when you promised yourself not to		Could not stop drinking without difficulty after one or two drinks	Threshold DIF: C1	Threshold DIF: C3
16	Felt you were going crazy	Your drinking made you feel out of control even when you were sober				
17	Felt physically or psychologically dependent on alcohol	Felt physically or psychologically dependent on alcohol		Thought you were dependent on alcohol		
18						

Table 2 (continued)

	Study 1	Study 2	Study 3	Study 4	MNLFA Item DIF by Study ALCOHOL	MNLFA Item DIF by Study DRUG
	Was told by a friend, neighbor or relative to stop or cut down drinking	Near relative or close friend worried or complained about your drinking				
ITEMS ON DRUG USE CONSEQUENCES MEASURE ONLY						
19	Had a strong desire or urge to use drugs	Wanted to use drugs so badly you couldn't think about anything else				Threshold DIF: C2, C3
20	Your drug use lead to health, social, legal, or financial problems			Have you had medical problems, fights with family or friends, or engaged in other illegal activities due to your drug use		
21	Failed to do what was normally expected of you because of your drug use			Neglected your responsibilities		Threshold DIF: C1
22	A friend or relative or anyone else has expressed concern about your drug use	Been told by a friend that you have a drug problem				Threshold DIF: C3 Loading DIF: C3
23	Tried and failed to control, cut down, or stop using drugs			Wanted to stop using drugs but couldn't		

Note: Items 1–18 appeared on both the alcohol- and drug-consequence scales (replacing references to alcohol with drugs) and items 19–23 appeared only on the drug-consequence scale. Columns 2–5 report logical harmonization of similar items across the four conditions in the IDA analogue study. Columns 6 and 7 report results of MNLFA DIF analyses for alcohol- and drug-consequence items, respectively. Shaded boxes highlight item stems that were changed for forms 2–3 and dropped in form 4 versus unhighlighted boxes for item stems changed only for form 4

strongly". We created mean scores to index each subscale in subsequent analyses (Positive Urgency – Session 1: M = 1.84, SD = 0.46, α = 0.86; Session 2: M = 1.84, SD = 0.51 , α = 0.90; Negative Urgency – Session 1: M = 2.04; SD = 0.58, α = 0.88; Session 2: M = 2.03, SD = 0.63, α = 0.90; Sensation Seeking – Session 1: M = 2.76, SD = 0.52, α = 0.80; Session 2: M = 2.76, SD = 0.56, α = 0.84).

Analytic plan

Our analytic plan included three steps: (1) scoring, (2) testing our reliability hypothesis, and (3) testing our harmonization, validity, and generalizability hypotheses simultaneously.

Step 1: Scoring In Step 1, we computed mean, CFA, and MNLFA scores for the alcohol and drug consequence measures in the altered battery; resulting in a total of 12 scores per person for each measure (4 analogue study scores x 3 scoring algorithms; see point 3; Table 1). To do so, we first used logical harmonization strategies to create a pooled data set that included each of the altered measures by aligning conceptually similar items and collapsing response options in the altered scale to resemble the original scale (see Table 2). Mean scores resulted from averaging all available item responses for a given scale within participant in the pooled data set.

Procedures for analytical harmonization and scoring followed those outlined elsewhere (e.g., Curran et al., 2016) and included creating a calibration sample by randomly sampling one study (or altered measure) for each participant using data from the first testing session (to create a sample with independent observations; resulting in n = 214, 226, 200 and 214 for studies 1–4, respectively). We then reviewed descriptive analyses for all items as a function of study membership to identify potential differential item functioning and conducted exploratory factor analyses within and across study to identify unidimensional scales (to facilitate MNLFA).

We then derived traditional CFA factor scores from a unidimensional factor analysis in which all items loaded on a single underlying factor, factor means were set to zero and variance to one, and all loadings freely estimated in the pooled data set; however, no study effects (in impact or DIF) were modeled. Since all items were measured on an ordinal scale, estimation proceeded via marginal maximum likelihood with a logit link function. Once the model had been fit to the calibration sample, the obtained estimates were treated as fixed and known and used to compute factor scores for the rest of

the sample. We estimated these models using Mplus (Muthén & Muthén, 1998–2017; more details below).

To conduct and derive scores from MNLFA, we followed an iterative approach to model specification that has been shown to perform reasonably well for related models (Navas-Ara & Gómez Benito, 2002; Oort, 1998) and in prior MNLFA applications (e.g., Curran et al., 2016) using Mplus. Given that study samples were equated on all other covariates by design, we only included study membership as a covariate in these models. As a result, MNLFA models were similar to a multiple groups factor analysis or IRT model in which study membership is modeled as a grouping indicator (Bauer, 2017). As with the CFAs, models were fit to the calibration sample and then the estimates were used to generate scores for all remaining observations.

More specifically, for each factor, we fit a baseline unidimensional MNLFA model allowing the factor mean and variance to vary by study and tested this model against a model in which the intercept and loading for one item at a time was free to vary across studies (iteratively) while all other item parameters were held invariant. For these analyses, we included contrast codes for study membership to test for increasing alteration of measures across studies (rather than simply study differences) comparing (a) studies 1 to 3 versus 4, (b) studies 1 and 2 versus 3, and (c) study 1 versus 2. Using likelihood ratio tests, we first identified the item for which DIF would result in the largest improvement in fit. We retained DIF for this item and then determined whether allowing for DIF in a second item would significantly improve model fit. Allowing for DIF in the second item that would most improve model fit, we then considered a third item, and so on, until no further significant improvement in model fit could be obtained. Finally, we removed nonsignificant DIF terms (based on Wald tests), other than lower-order terms involved in higher-order effects.

Step 2: Reliability analyses Step 2 in our analytic plan tested the reliability hypothesis. We estimated reliability of resulting mean, CFA and MNLFA scores (to test hypothesis 1) using the reliability subsample who received the same survey at both sessions (e.g., survey A or B at both session 1 and 2; $n = 432$; see Point 4, Table 1). We first examined internal reliability using Cronbach's alpha estimates for mean scores and marginal reliability estimates for CFA and MNLFA scores. We then examined test–retest reliability by estimating correlations between session 1 and 2 scores. Finally, we estimated original score recovery (akin to parallel forms reliability) by estimating correlations between scores from studies 2–4 and the original scale as administered in study 1. For this final reliability analysis, we used the recovery subsample of participants who completed both survey A or B on separate sessions (to allow for a wash-out period) to test score recovery over 2 weeks for harmonized scores in altered measures for studies 2 and 4 (i.e., survey A then B or vice-versa over sessions 1 and 2; $n = 402$). However, because the altered measure for study 3 was also administered on survey A (with the original measure), when evaluating score recovery for this study we only included participants who completed survey A at both sessions, randomly sampling whether the altered or original scores came from session 1 or 2 ($n = 208$).

Step 3: Harmonization and validity analyses In Step 3, we evaluated hypotheses 2 and 3, which posited that MNFLA scores would differ less between studies than scores from other methods and that MNLFA scores would show greater associations with external validity measures. These hypotheses were evaluated by estimating regression models within a structural equation modeling framework accounting for unreliability in scores from the altered battery (Bollen, 1989). These models were fit to the recovery subsample of individuals assigned to the AB or BA conditions for whom scores were available for all four studies, using cluster-robust standard errors to account for dependence. The harmonized scores for alcohol and drug use consequences served as outcomes, with separate models fit for each type of scores. Our six validity measures from the standard battery served as predictors in separate models, along with contrast-coded study membership, and the interaction of study membership and the validity measure. We evaluated the extent to which scores differed as a function of severity of measurement alteration by testing for study main effects (harmonization hypothesis). We also tested whether differences in validity emerged over scoring method (differences in the main effects of validity measures on scores). Given that the same population is assessed in each study, significant interactions between study membership and validity measures when predicting scores would indicate a failure of harmonization; therefore, we also tested whether such interactions were significant as a further test of our harmonization hypothesis. Finally, we evaluated our generalizability hypothesis by comparing results of all analyses across the alcohol- and drug-consequence measures.

Results

Step 1: Scoring

Exploratory factor analysis Consistent with expectations, results of exploratory factor analysis confirmed that both alcohol- and drug-consequence measures conformed to a unidimensional factor structure and inspective of descriptive statistics indicated that items had sufficient variance for estimation.

Table 3 Summary of MNLFA DIF findings and four-study analogue design

CONSTRUCT	Does MNLFA mis-detect DIF for items with no changes in stem or response scale? Results from contrast code comparing study 1 versus study 2	Does MNLFA detect changes in the item stem? Results from contrast code comparing study 1 versus study 2	DOES MNLFA detect changes in response scale changes (only, not stems)? Results from contrast code comparing study 3 versus studies 1–2	DOES MNLFA detect changes in items with both stem and response scale changes? Results from contrast code comparing study 3 versus studies 1–2 and from contract code comparing study 4 versus studies 1–3	
	Items where DIF not expected because item was identical over condition	Items where DIF expected because stem changed only	Items with DIF expected because of response scale change only,	C2: Items with changes in stem and response scale,	C1: Subset of Items and ALL had change in stem and response scale,
Alcohol Consequences	0 items of 9 show DIF	3 items out of 9 show DIF	1 item out of 9 shows DIF	2 items out of 9 show DIF	4 items out of 9 show DIF
Drug Consequences	1 item out of 12 show DIF	4 items out of 11 show DIF	0 items out of 12 show DIF	2 items out of 11 show DIF	2 items out of 12 show DIF

MNLFA scoring For alcohol-consequences, MNLFA did not identify study differences in the factor mean or variance (all $p > .05$). In addition, MNLFA identified study DIF in eight of 18 items (as indicated in Table 3), primarily reflecting intercept DIF with only one item (item 9) showing loading DIF. For drug-consequences, MNLFA identified marginally significant study differences in the factor mean (i.e., $\widehat{\gamma} = 0.15$, z = 1.65, $p = .099$ for study 1 and 2 versus 3; $\widehat{\gamma} = .26$, z = 1.67, $p = .096$ for studies 1 to 3 versus 4) but not variance as well as study DIF in seven of 23 items, with only two items showing loading DIF.

Results of the DIF analyses are largely consistent with the intended four study analogue design (see summary in Table 3). We did not detect study DIF for items that were identical over study conditions for the alcohol consequence measures (see Table 3, column 2) and only did so for one of 12 items for drug consequences. Rates of DIF detection were higher for items altered over study conditions than would be expected to show DIF by design. This pattern of DIF was more sensitive to changes in item stems than to those in response scale. For example, for alcohol-consequences, only one of nine items with an altered response scales displayed DIF (Table 3, column 4) whereas three of nine items with altered stems (column 3) and six of 18 items with both altered item stems and response scales (columns 5 and 6) displayed DIF. For drug-consequences, no items with an altered response scale displayed DIF whereas four of 11 items with altered stems and four of 23 items with both altered stems and response scales displayed DIF.

Step 2: Reliability analyses

We used results of CFA and MNLFA models to create scores for participants on all four study measures they completed (two at Session 1 and two at Session 2) that we then subjected to reliability analyses. We calculated three forms of reliability to evaluate each of the four study measures when using mean, CFA, and MNLFA scores, respectively (see Table 4). The three forms of reliability include internal, (2-week) test–retest, and parallel forms.

Internal reliability For internal consistency, higher indices were evident for mean scores (which were based on Cronbach's alpha) versus CFA and MNLFA scores (which were based on marginal reliability estimates and generally comparable). The necessary difference in method for calculating internal consistency across scoring methods may account for this result, particularly given evidence that Cronbach's alpha may yield biased reliability estimates due to unrealistic assumptions (McNeish, 2018). However, we include estimates of Cronbach's alpha here for comparison with commonly used applications in the

Table 4 Score internal, test–retest, and parallel forms reliability by study and scoring method

Construct	Study 1	Study 2	Study 3	Study 4
Alcohol consequences				
Internal: MNLFA Scores	.80	.82	.80	.58
Internal: CFA Scores	.81	.82	.78	.62
Internal: Mean Scores	.84	.88	.89	.84
Test–retest: MNLFA Scores	.81	.87	.80	.80
Test–retest: CFA Scores	.81	.87	.80	.81
Test–retest: Mean Scores	.82	.88	.85	.81
Parallel Forms: MNLFA Scores	---	.84	.77	.72
Parallel Forms: CFA Scores	---	.83	.78	.71
Parallel Forms: Mean Scores	---	.83	.82	.72
Drug consequences				
Internal: MNLFA Scores	.74	.72	.67	.52
Internal: CFA Scores	.72	.70	.68	.55
Internal: Mean Scores	.92	.94	.94	.93
Test–retest: MNLFA Scores	.83	.83	.84	.69
Test–retest: CFA Scores	.83	.83	.84	.70
Test–retest: Mean Scores	.84	.81	.83	.77
Parallel Forms: MNLFA Scores	---	.79	.78	.66
Parallel Forms: CFA Scores	---	.79	.78	.65
Parallel Forms: Mean Scores	---	.72	.78	.64

Note: Study 1 includes original measures; study 2 includes alterations to half of item stems; study 3 includes study 2 item stem and an altered response scale for all item; study 4 includes altered stems for all items and study 3 response scale

field. In addition, as one would expect, internal consistency estimates were lower for study 4 that had fewer items than studies 1–3 (Wainer & Thissen, 2001). For alcohol consequences, reliability indices were also somewhat higher for study 2 than for other studies. For drug consequences, reliability indices varied little across studies for mean scores but were progressively lower across study perturbations (1 to 4) for MNLFA and CFA scores.

Test–retest reliability Test–retest reliability indices were similar across scoring method, with some isolated variation (i.e., lower for MNLFA scores versus others in study 3 for alcohol consequences; higher for mean versus other scores in study 4 for drug consequences). Test–retest reliability indices were also higher for study 2 than for other studies across all scoring methods for alcohol-consequences and lower for study 4 for drug-consequences (similar to internal consistency results).

Parallel forms reliability Finally, parallel forms reliability indices were also fairly consistent across scoring method. However, parallel forms reliability indices decreased with increasing study perturbations (studies 2–3), suggesting that we were successful in designing studies with diminishing comparability with the original measure across the three altered measure study conditions. Given that reliability decreased similarly for all three scoring methods, MNLFA does not appear to enhance fidelity of scores relative to other methods.

Summary In sum, we found few differences in reliability as a function of scoring method, but we did find lower reliability as expected for shorter (and more altered) measures.

Step 3: Harmonization, validity, and generalizability

We built regression models to test these three hypotheses simultaneously by predicting scores from study main and interactive effects (harmonization) and expected correlates of substance use (validity), and we evaluated the comparability of findings across alcohol- and drug-use consequence analyses (generalizability). Specifically, we fit regression models in which scores (mean, CFA, or MNLFA) were separately predicted by each of the six validity variables. Each model also included contrast-coded study indicators as well as the interaction between the validity measure and the study indicators. The six validity measures included peer descriptive norms, peer injunctive norms, parent attitudes, negative urgency, positive urgency, and sensation seeking. Study contrast codes mirrored those used for scoring in MNLFA models. We estimated separate regression models for each harmonized construct (2), validity measure (6), and scoring method (3), resulting in 36 models total. Given multiple testing, we applied a Benjamini–Hochberg correction (Benjamini & Hochberg, 2000) to control the false discovery rate.

Results of these initial models found numerous study interactions showing that study moderated relations between harmonized scores and validity measures across scoring conditions. This outcome was unexpected given that the sample in each study was drawn from the same population and so effects should not vary between studies. Such an outcome may reflect a failure of harmonization. Prior literature, however, suggests that differential reliability may also account for differences in the strength of predictive correlations. Given differential reliability of harmonized measures across study condition and scoring method, we re-estimated validity analyses correcting for score unreliability. All regression models were thus estimated in Mplus as single indicator SEMs with a latent variable representing the 'true' factor score, measured by the estimated score, regressed on the set of predictors (see Kline, 2015 pp. 214–215 for an example). The measurement model was estimated such that intercepts, loadings, and unique variances were fixed values determined by score reliability (alpha for mean scores and marginal reliability for CFA/MNLFA scores), as well as means and variances of factor scores. Results for predictive validity analyses (reported as squared semi-partial correlations) for both alcohol- and drug-

Table 5 Validity results for alcohol- and drug-use consequences

	Alcohol-use consequences			Drug-use consequences		
	Mean	CFA	MNLFA	Mean	CFA	MNLFA
Peer use						
Peer Use	.177***	.229***	.229***	.133***	.269***	.266***
S1v2	.000	.000	.000	.001	.000	.001
S12v3	.000	.000	.000	.000	.000	.001**
S123v4	.002***	.000	.000	.000	.001	.004**
Peer Use x S1v2	.002***	.000	.000	.001	.000	.000
Peer Use x S12v3	.000	.000	.000	.002	.000	.000
Peer Use x S123v4	.001	.000	.000	.000	.001	.000
Model R^2	0.235	0.229	0.233	0.142	0.27	0.298
Peer attitudes						
Peer Attitudes	.138***	.163***	.164***	.120***	.240***	.240***
S1v2	.001	.000	.001	.000	.000	.001
S12v3	.000	.000	.000	.001	.001	.000
S123v4	.001	.000	.000	.001	.001	.001
Peer Attitudes x S1v2	.002**	.000	.001	.000	.000	.000
Peer Attitudes x S12v3	.000	.000	.000	.002	.001	.000
Peer Attitudes x S123v4	.001	.000	.000	.002	.001	.000
Model R^2	0.194	0.164	0.168	0.13	0.242	0.272
Parent attitudes						
Parent Attitudes	.036***	.045***	.046***	.035***	.086***	.085***
S1v2	.000	.000	.000	.001	.002	.001
S12v3	.000	.000	.000	.000	.000	.002**
S123v4	.005**	.002	.001	.000	.000	.002
Parent Attitudes x S1v2	.000	.000	.000	.001	.002	.001
Parent Attitudes x S12v3	.000	.000	.000	.000	.001	.001
Parent Attitudes x S123v4	.000	.002	.002	.000	.000	.000
Model R^2	0.093	0.051	0.055	0.041	0.087	0.119
Negative urgency						
Negative Urgency	.104***	.122***	.121***	.049***	.064***	.064***
S1v2	.000	.000	.000	.000	.000	.000
S12v3	.002**	.000	.002**	.000	.000	.002
S123v4	.000	.000	.000	.000	.000	.005**
Negative Urgency x S1v2	.000	.000	.000	.000	.000	.000
Negative Urgency x S12v3	.000	.000	.001	.000	.000	.000
Negative Urgency x S123v4	.001	.000	.000	.001	.000	.001
Model R^2	0.157	0.123	0.126	0.055	0.064	0.097
Positive urgency						
Positive Urgency	.102***	.110***	.109***	.047***	.049***	.049***
S1v2	.000	.000	.000	.000	.000	.000
S12v3	.001	.000	.001	.001	.000	.002
S123v4	.001	.000	.000	.000	.000	.004**
Positive Urgency x S1v2	.000	.000	.000	.000	.000	.000
Positive Urgency x S12v3	.000	.000	.000	.000	.000	.000
Positive Urgency x S123v4	.000	.000	.000	.000	.000	.001
Model R^2	0.155	0.111	0.113	0.053	0.049	0.081
Sensation seeking						
Sensation Seeking	.031***	.037***	.037***	.021**	.042***	.041***

Table 5 (continued)

	Alcohol-use consequences			Drug-use consequences		
	Mean	CFA	MNLFA	Mean	CFA	MNLFA
S1v2	.003	.002	.002	.002	.002	.001
S12v3	.002**	.000	.001	.001	.001	.002**
S123v4	.000	.000	.000	.001	.001	.000
Sensation Seeking x S1v2	.001	.002	.002	.002	.002	.001
Sensation Seeking x S12v3	.001	.000	.001	.000	.001	.001
Sensation Seeking x S123v4	.001	.000	.000	.001	.001	.000
Model R^2	0.086	0.039	0.043	0.03	0.045	0.075

Note: S1v2 represents the contrast code for membership in study 1 versus study 2 and S12v3 and S123v4 represents those for studies 1 and 2 versus 3 and studies 1–3 versus 4, respectively. Entries are squared semi-partial correlations for each predictor (unique variance explained by that effect), except for rows labeled Model R^2, which report the multiple squared correlation for each model (variance explained by the set of predictors). Note that * indicates significant at $p < .05$; ** < .01; and *** < .001

consequence models using reliability correction are reported in Table 5.

Harmonization findings For alcohol-consequence models, there were no study main effects with either CFA or, with one exception, MNLFA scores, though there were study differences in the mean scores in three of six predictive validity models (peer use, peer attitudes, and sensation seeking). Given the lack of impact (i.e., mean differences in the latent factor comprising alcohol-consequences) found in the MNLFA measurement model, we would not expect study differences to emerge in our predictive models and thus the main effects of study suggest poorer harmonization of the alcohol-consequence measure in mean scores than in CFA and MNLFA scores. The exception to this pattern is the significant effect of study (1 and 2 versus 3) when predicting mean and MNLFA scores with negative urgency. This aberration could reflect type I error even though we used corrections for alpha inflation or could indicate limits or boundary conditions on the benefits of MNLFA over other scoring methods.

For drug-consequence models a different pattern emerged. No study main effects or interactions were found in validity models predicting mean or CFA scores. For models using MNLFA scores, main effects of study were found in five of six models, suggesting higher scores in studies 1–3 versus 4 (in models evaluating relations with peer use, negative urgency and positive urgency) and/or studies 1–2 versus 3–4 (for peer use, parent attitudes, and sensation seeking). Given evidence of marginally significant impact in the drug-consequence models, we could speculate that study differences found in the MNLFA were appropriately detected but missed in the mean and CFA scores models. Alternatively, MNLFA scores may be more sensitive to systematic measurement bias than CFA and mean scores, a possibility that deserves future consideration.

Validity findings Because the study interactions, even when present, accounted for relatively little variance, we can compare the magnitude of the main effects of the validity measures to evaluate differences in the construct validity of the scores. Importantly, for all six validity variables the squared semi-partial correlations for both alcohol and drug consequences were higher for CFA and MNLFA scores than mean scores. In some cases, these differences were modest (e.g., positive urgency), but for other validity variables the differences were substantial. For example, peer use uniquely explained 27% of the variance in CFA and MNLFA scores for drug consequences, a twofold increase over the 13% explained in mean scores. For almost all drug-consequence models, the multiple R^2 values obtained from the full models (also including study main effects and study by validity variable interactions) also favor CFA and MNLFA scores over mean scores. (The sole exception to this pattern being positive urgency, where mean scores resulted in higher R^2 than CFA scores but not MNLFA scores.) In contrast, for alcohol-consequence scores R^2 estimates tended to be higher for mean scores than for CFA and MNLFA scores. This difference may be due to the greater number of study main effects detected for mean scores with alcohol consequences (as reflected in the higher squared semi-partial correlations for study for mean scores than CFA or MNLFA scores for all six validity variables). As noted, given equivalent populations in all four studies, these study main effects can be largely discounted as spurious consequences of poor scoring methodology.

Summary In sum, study interactions in our harmonization analyses suggest potential for differential replication of effects across studies that, due to study design, could be attributed to

measurement differences across study (and again, poorer harmonization). Although we found no study interactions in any drug use consequence models (regardless of scoring method), we did find study interactions for alcohol consequences with two of six validity variables when using mean scores but not when using CFA or MNLFA scores. This occurred for analyses with peer descriptive norms and injunctive norms in which these measures were somewhat more strongly correlated with mean scores for alcohol use consequences in study 1 (using the standard measure) than in the study 2 (using the slightly altered measure). In terms of validity analyses, we found some differences as a function of scoring method which we explore further below.

Discussion

In the current study, we used an analogue IDA study to test optimal scoring strategies for harmonizing measures with varying degrees of alteration in measurement across four study conditions, holding study differences in sample characteristics constant. We evaluated score performance based on traditional indices of reliability (test–retest, internal, and score recovery or parallel forms) and construct validity (strength of relations to external correlates). As expected, mean, CFA and MNLFA scores all resulted in scales with lower reliability given increasing scale alteration (with less fidelity to formerly developed scales) and shorter scale length. However, we did not find expected stronger test–retest and internal reliability indices for MNLFA and CFA as compared to mean scores. We also did not find that MNLFA and CFA scores from studies with altered measures were superior in recovering (showed higher correlations with) scores based on standard measures as compared to mean scores. Overall, we did not see the expected advantages in score recovery with MNLFA versus CFA and mean scores on any reliability index within the current analogue design. However, the full story is more complex.

Perhaps more important is score validity, as highly reliable scores with little validity provide little utility in advancing science. Importantly, differences in score reliability over method may masquerade as differences in validity if not taken into account. We tested validity in relation to harmonization, relationships to external correlates, and generalizability hypotheses. We evaluated the harmonization hypothesis by testing whether there were differences across studies in harmonized scores; we predicted that such study differences would be less evident with more successful harmonization and that MNLFA scores, which account for study differences in DIF, would show fewer study differences than CFA or mean scores. For alcohol use consequence models, study differences were indeed evident in some validity analyses using mean scores but not for those using CFA or, with one exception, MNLFA scores. For drug use consequence models, study differences were found in some validity models using MNLFA scores but not CFA or mean scores.

In general, main effect study differences in scores could reflect true impact (i.e., study differences in the true mean of the construct) that should not disappear with appropriate harmonization. Although the IDA analogue design equates the four study conditions on sampling (one source of impact), other sources of impact could remain. For example, study differences may have occurred as a function of ordering of altered measures that represent study conditions (i.e., studies 3 and 4 were the second times the scale was given in one battery but also versions of the scale with a change in response scale that may change endorsement probabilities). MNLFA results suggest that such differences may have occurred to a limited extent for our drug consequence scale. Although no study contrasts were significantly associated with the latent factor mean for alcohol use consequences (t = – .19 to .54, all $p > .10$), two of three study contrasts were marginally significant predictors of the latent factor mean for drug use consequences (study 1 vs. 2, t = .49, $p = .626$; studies 1 and 2 vs. 3, t = 1.65, $p = .099$; studies 1, 2, and 3 vs. 4, t = 1.67, $p = .096$). For this reason, we might speculate that the mean scores erroneously picked up study differences in alcohol consequences while MNLFA (and CFA) scores did not whereas mean (and CFA) scores failed to pick up modest study differences in drug consequences while MNLFA scores did not.

This would not explain, however, the one study effect found for predicting MNLFA alcohol-consequence scores from negative urgency. This deviation to the pattern may reflect type I error, although we did employ corrections for alpha inflation in interpreting our findings. This may also indicate that there are exceptions or boundary conditions under which MNLFA scores outperform mean scores in terms of validity analyses, meaning that MNLFA scores may only outperform mean scores under certain modeling contexts. This possibility deserves further study.

Also, interesting, the study differences we do see in the predictive models are making the same study comparisons that include changes in response scales or both response scales/items stems by design (studies 1 and 2 vs. 3 or studies 1–3 vs. 4) and not between studies where scale differences were just in item stems (studies 1 and 2). This pattern mirrors that for DIF in our MNLFA findings. Such results may indicate that greater caution is needed in harmonizing scales with different response scales across studies and that logical harmonization may be insufficient to account for what would then result in study impact differences in later predictive models.

Addressing our validity hypothesis, we posited that MNLFA scores would be more strongly associated with external correlates and would show fewer differences in these associations over studies than would CFA or mean scores. For

both our alcohol and drug use consequence models, MNLFA and CFA consequence scores showed stronger associations with validity measures, though there was little difference between MNLFA and CFA scores in this regard. Our simulation results, however, show that this need not always be the case and that under other conditions MNLFA scores can be expected to outperform CFA scores in capturing relationships with other variables (Curran et al., 2018).

Study differences in associations (indicated by study interactions) were only evident in results of alcohol use consequences. We saw failure of pure replication (as evidenced by significant study interactions) for mean scores but not for MNLFA and CFA scores, suggesting that less optimal measurement harmonization (using mean scores) can result in misleading predictive models. Moreover, this pattern only became evident in analyses with reliability corrected harmonized scores, suggesting that reliability corrections for harmonized scores may be important for avoiding detection of spuriously different effects across studies (or the false conclusion of failure to replicate). This finding has implications for IDA and replication analyses in general when measures have differential reliability across studies, a potentially important topic for further research.

In sum, this work demonstrates that MNLFA may be successfully employed in IDA to create harmonized measures. There are clear cases in which MNLFA scores are superior to mean scores. Although the two scoring systems in the current study were comparable in reliability, validity appears to be more accurate (for recovering impact) and stronger (based on semi-partial correlations) for MNLFA than for mean scores. This may be particularly true when harmonizing measures in which different response scales have been logically harmonized. However, it remains unclear if there are cases where MNLFA scores are superior to CFA scores. We would suspect this to be the case when greater DIF is present, either greater in magnitude than in the current study or occurring in more complex patterns (additional covariates and study interactions). However, in our simulation work the biggest differentiator of score performance has been study impact. As such, we may expect MNLFA scores to outperform CFA scores when we have more sampling variation or assessment variation that could impact factor means. Little mean impact was expected in the current design given equivalent sampling for each study, limiting our ability to see this potential advantage of MNLFA relative to CFA. Nevertheless, some evidence of this possibility is seen in the marginally significant study effects in the predictive validity analyses we found for drug-consequences that mimic the MNLFA model but were not recovered in CFA (or mean) scores. We may also see differences in MNLFA and CFA score performance if we were to consider sources of study impact and DIF in addition to study differences (e.g., sex differences); the exclusion of such effects in CFA models may create greater differences in scale performance relative to scores derived from MNLFA models that include such effects.

Strengths of the study include the use of an experimental design to evaluate the performance of statistical models with real-world data, the pairing of the current study with our prior computer simulation results, and the consideration of these methods within the context of broadly used substance use measures. Limitations include the need to administer two versions of altered measures in the same survey to reduce participant burden, introducing potential order effects and a smaller sample size for examining reliability estimates for study condition 3 than for other conditions. Moreover, Cronbach's alpha is a problematic estimate for internal validity due to its reliance on unrealistic assumptions resulting in shrunken estimates (McNeish, 2018) and comparisons across different forms of internal reliability estimates are challenging to interpret. Nonetheless, these results lead us to ask not whether MNLFA scores are better than CFA and mean scores, but when. The answer is likely to be complicated but suggests important future directions for research, including the role of reliability correction in scoring to redress study differences in measurement in replication and IDA studies, potential differences in score performance as a function of sampling differences across pooled studies, and the integration of such analytic harmonization approaches into more complex predictive models. But without this knowledge, we believe the most prudent approach remains MNLFA, given that there is little cost of MNLFA relative to CFA beyond model complexity and that, based on our simulation results, there are likely contexts other than those explored here where MNLFA would still be expected to outperform CFA.

Acknowledgements The ideas and findings presented in this study have not been presented or published elsewhere. This research was made possible by a grant from the National Institute on Drug Abuse awarded to Daniel J. Bauer (1R01DA034636-01A1). The opinions in this paper are those of the authors and do not reflect those of the National Institutes of Health.

Open science statement None of the data for the experiment reported here are publicly available, though the materials include original measures posted in the Phen-X measurement archive at https://www.phenx.org/ and data may be shared upon request, given compliance with IRB requirements. None of the experiments was preregistered.

References

Bauer, D.J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22,* 507–526. doi: https://doi.org/10.1037/met0000077.

Bauer, D.J. & Hussong, A.M (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14*, 101–125. doi:https://doi.org/10.1037/a0015583

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*(1), 60–83. https://doi.org/10.2307/1165312

Bollen, K.A. (1989) Structural Equations with Latent Variables. John Wiley and Sons, Inc., New York. https://doi.org/10.1002/9781118619179

Bollen, K.A. & Hoyle, R.H. (2012). Latent variables in structural equation modeling. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling,* pp. 56–67. New York: The Guilford Press.

Conway, K. P., Vullo, G. C., Kennedy, A. P., Finger, M. S., Agrawal, A., Bjork, J. M., … Sher, K. J. (2014). Data compatibility in the addiction sciences: An examination of measure commonality. *Drug and Alcohol Dependence*, *141*. 153–158. https://doi.org/10.1016/j.drugalcdep.2014.04.029

Cooper, H., Hedges, L., & Valentine, J. (2009). *The Handbook of Research Synthesis and Meta-analysis*. New York: Russell Sage Foundation.

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. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 827–844. DOI: https://doi.org/10.1080/10705511.2016.1220839

Curran, P.J., Cole, V.T., Bauer, D.J., Rothenberg, A.W., & Hussong, A.M. (2018). Recovering predictor-criterion relations using covariate-informed factor score estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 25*, 860–875.

Curran, P.J., & Hussong, A.M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14,* 81–100. https://doi.org/10.1037/a0015914

Ennett, S. T., Bauman, K. E., Hussong, A., Faris, R., Foshee, V. A., Cai, L., & DuRant, R. H. (2006). The peer context of adolescent substance use: Findings from social network analysis. *Journal of Research on Adolescence*,*16*(2), 159–186.

Fromme, K., Stroot, E., & Kaplan, D. (1993). Comprehensive effects of alcohol: Development and psychometric assessment of a new expectancy questionnaire. *Psychological Assessment, 5*(1), 19–26.

Ham, L.S., Stewart, S.H., Norton, P.J., & Hope, D.A. (2005). Psychometric assessment of the Comprehensive Effects of Alcohol Questionnaire: Comparing a brief version to the original full scale. *Journal of Psychopathology and Behavioral Assessment 27*(3), 141–158.

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. *Alcoholism Clinical and Experimental Research, 39,* 919–931. https://doi.org/10.1111/acer.12714

Hussong, A. M., Flora, D. B., Curran, P. J., Chassin, L. A., Zucker, R. A. (2008). Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents. *Development and Psychopathology, 20*(1), 165–193. https://doi.org/10.1017/S0954579408000084

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology*, *9*, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Hussong, A. M., Wirth, R. J., Edwards, M. C., Curran, P. J., Chassin, L. A, & Zucker, R. A. (2007). Externalizing symptoms among children of alcoholic parents: Entry points for an antisocial pathway to alcoholism. *Journal of Abnormal Psychology*, *116*(3), 529–42. https://doi.org/10.1037/0021-843X.116.3.529

Johnston, L. D., O'Malley, P. M., Bachman, J. G., and Schulenberg, J. E., (2013). *Monitoring the Future national survey results on drug use, 1975–2012: Volume 2, College students and adults ages 19–50.* Ann Arbor: Institute for Social Research, the University of Michigan.)

Kline, R.B. (2015). *Principles and Practice of Structural Equation Modeling* (4th). New York: Guilford Press.

Leigh, B.C Stacey, A.W. (1993). Alcohol outcome expectancies: Scale construction and predictive utility in higher order confirmatory models. *Psychological Assessment, 5,* 216–229.

Lynam, D.R., Smith, G.T., Cyders, M.A., Fischer, S., & Whiteside, S.A. (2007). The UPPS-P: a multidimensional measure of risk for impulsive behavior. Unpublished technical report.

Mun, E.-Y., de la Torre, J., Atkins, D. C., White, H. R., Ray, A. E., Kim, S.-Y., … Huh, D. (2015). Project INTEGRATE: An integrative study of brief alcohol interventions for college students. *Psychology of Addictive Behaviors*, *29*(1), 34–48. https://doi.org/10.1037/adb0000047

Muthén, L.K. and Muthén, B.O. (1998–2017). *Mplus User's Guide.* Eighth Edition. Los Angeles, CA: Muthén & Muthén

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5,* 107–124. https://doi.org/10.1080/10705519809540095

Pigott, T., Williams, R., & Polanin, J. (2012). Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis and Methods 3*, 257–268. https://doi.org/10.1002/jrsm.1051

Rose, J.R., Dierker, L.C., Hedeker, D., & Mermelstein, R. (2013). An integrated data analysis approach to investigating measurement equivalence of DSM nicotine dependence symptoms. *Drug and Alcohol Dependence, 129,* 25–32.

Steinberg, L. & Thissen, D. (2013). Item response theory. In J. S. Comer & P. C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology.* (pp. 336–373). New York, NY: Oxford University Press.

Wainer, H. & Thissen, D. (2001). True score theory: The traditional method. In H. Wainer and D. Thissen, (Eds.), *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum

White, H.R., & Labouvie, E.W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol, 50*, 30–37.