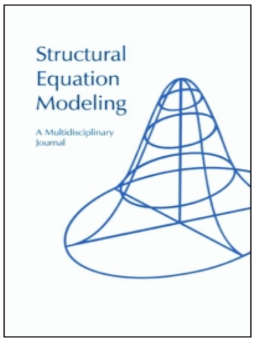

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

An Empirical Assessment of the Sensitivity of Mixture Models to Changes in Measurement

Veronica T. Cole, Daniel J. Bauer, Andrea M. Hussong & Michael L. Giordano

To cite this article: Veronica T. Cole, Daniel J. Bauer, Andrea M. Hussong & Michael L. Giordano (2017) An Empirical Assessment of the Sensitivity of Mixture Models to Changes in Measurement, Structural Equation Modeling: A Multidisciplinary Journal, 24:2, 159-179, DOI: 10.1080/10705511.2016.1257354

To link to this article: https://doi.org/10.1080/10705511.2016.1257354

Published online: 13 Jan 2017.

Submit your article to this journal

Article views: 204

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=hsem20

Download by: [University North Carolina - Chapel Hill] **Date:** 02 January 2018, At: 09:52

Structural Equation Modeling: A Multidisciplinary Journal, 24: 159–179, 2017
Copyright © Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2016.1257354

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

ARTICLES

An Empirical Assessment of the Sensitivity of Mixture Models to Changes in Measurement

Veronica T. Cole, Daniel J. Bauer, Andrea M. Hussong, and Michael L. Giordano
The University of North Carolina at Chapel Hill

This study explored the extent to which variations in self-report measures across studies can produce differences in the results obtained from mixture models. Data ($N = 854$) come from a laboratory analogue study of methods for creating commensurate scores of alcohol- and substance-use-related constructs when items differ systematically across participants for any given measure. Items were manipulated according to 4 conditions, corresponding to increasing levels of alteration to item stems, response options, or both. In Study 1, results from latent class analyses (LCAs) of alcohol consequences were compared across the 4 conditions, revealing differences in class enumeration and configuration. In Study 2, results from factor mixture models (FMMs) of alcohol expectancies were compared across 2 of the conditions, revealing differences in patterns and magnitude of the factor loadings and thresholds. The results suggest that even subtle differences in measurement can have substantively meaningful effects on mixture model results.

Keywords: factor mixture models, latent class analysis, mixture models, self-report

Increasingly popular within psychology and allied fields, finite mixture models offer the opportunity to identify latent subgroups of individuals within a population (McLachlan & Peel, 2000). For instance, one recent study used mixture models to find subtypes of individuals with schizophrenia based on comorbidity, finding three classes characterized by no comorbidity, comorbid anxiety and depression, or comorbid addiction (Tsai & Rosenheck, 2013). Another study (Crow et al., 2012) used mixture models to find six latent classes of individuals based on eating disorder symptoms, and additionally found that three of these classes were related to increased mortality risk.

Although mixture models have been applied to many behavioral phenomena, results can differ widely from one study to the next, presenting an inconsistent picture of the underlying latent structure of a given construct. One notable example is in the study of alcohol use disorder (AUD) as defined by the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed. [*DSM–5*]; American Psychiatric Association, 2013). A number of studies have attempted to uncover homogeneous classes of individuals on the basis of different patterns of the 11 *DSM*[1] diagnostic criteria as defined by either *DSM–5* or *DSM–IV* (American Psychiatric Association, 2000). The number of classes found from one application to the next ranges widely, with some studies finding two (Rinker & Neighbors, 2015), three (Beseler, Taylor, Kraemer, & Leeman, 2012; Chung & Martin, 2001; La Flair et al., 2012; La Flair et al., 2013; Mancha, Hulbert, & Latimer, 2012), four (N. Jackson et al., 2014; Wells, Horwood, & Fergusson, 2004), and five (Lynskey et al., 2005). Moreover, although most of these studies find classes on a continuum of severity that increases monotonically between classes (i.e., the classes mainly capture level of AUD liability), a few studies (Beseler et al., 2012; N. Jackson et al., 2014; Lynskey et al., 2005) find at least one class with a unique configuration of symptoms that falls outside of this continuum.

Correspondence should be addressed to Veronica T. Cole, Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC 27599. E-mail: vcole@email.unc.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hsem.

[1] Importantly, although the diagnostic classification of AUD changed from *DSM–IV* to *DSM–5*, the criteria themselves changed only by the omission of one item and addition of another.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

Of course, there are many potential reasons for these inconsistencies. For instance, results might differ depending on characteristics of the population that is sampled (e.g., college students vs. primary care patients; Lubke & Miller, 2015), as well as the well-documented effect of sample size on the power to detect classes (Lubke & Neale, 2006, 2008; Nylund, Asparouhov, & Muthén, 2007), especially when classes are unevenly sized (Tueller & Lubke, 2010).The results of finite mixture models can also be particularly sensitive to misspecification of the model assumptions and structure (Bauer & Curran, 2003, 2004; Van Horn et al., 2012). Here, however, we focus on another possibility: The results might differ across studies due to the use of different measurement instruments. A recent review by Lubke and Miller (2015) cautioned that both theory and prior simulation results point to a general sensitivity of mixture models (as well as taxometric techniques; Meehl, 1995) to the characteristics of the items included in the analysis. Often, these characteristics vary in important ways across studies. With respect to AUD symptoms, for instance, measurement can be performed via structured interviews such as the SCID (Chung & Martin, 2001), the SSAGA (Lynskey et al., 2005), or the CIDI (Wells et al., 2004), all of which assess *DSM* criteria through a set of questions designed for either particular clinical or research settings. Alternatively, some studies use paper-and-pencil or computerized surveys that elicit subjects' direct report of each of the 11 criteria (Jackson et al., 2014; La Flair et al., 2012; La Flair et al., 2013). Although the instruments used to measure AUD in these studies assess the same 11 criteria, they do so using different questions and modes of response. Thus, it is challenging to determine to what extent differences in class structure between studies are a result of differences in how AUD criteria were measured. Such differences are not isolated to research on AUD.

The question this article seeks to address is this: To what extent might differences in measurement across studies be responsible for these inconsistent results? Two prior lines of research have explored this possibility empirically, in each case by evaluating the sensitivity of longitudinal mixture models to variations in the measurement of over-time trajectories, including both the way outcomes were assessed as well as the number and timing of the assessments. First, using archival criminal offense data for 500 boys between ages 7 and 70, Eggleston, Laub, and Sampson (2004; Sampson, Laub, & Eggleston, 2004) investigated the methodological sensitivity of results obtained from a semiparametric growth model (SPGM; Nagin, 1999; Nagin & Tremblay, 2001). Eggleston and colleagues found that the number and nature of classes found by SPGM differed based on the time span over which individuals were studied, as well as the inclusion of controls for incarceration or mortality. In particular, these researchers found that the inclusion of fewer time points led to the discovery of fewer trajectory classes as well as steeper rates of increase and decrease over time. In a second line of research focused on growth mixture models with random effects (GMMs; Muthén, 2001; Muthén & Shedden, 1999), K. M. Jackson and Sher (2006) demonstrated a similar sensitivity of trajectory classes to the number and timing of repeated assessments. K. M. Jackson and Sher (2005) also found that separate GMM analyses conducted with different but related alcohol involvement constructs—AUD diagnosis, alcohol dependence symptoms, quantity-frequency, and heavy drinking—resulted in different conclusions regarding the number of latent classes. Even when holding the number of classes constant, the shapes and prevalence rates of the implied trajectories were highly dissimilar, and individual classifications were largely discordant across models. Finally, in an analysis of heavy episodic drinking (HED), the authors ran separate analyses on binary measures of HED scored according to different cut points (e.g., 5+ drinks in the past 30 days vs. 5+ drinks at least once or twice a week). Although the choice of HED cut point did not change the obtained number of classes, it did alter the relative proportion of individuals thought to belong to each class (K. M. Jackson & Sher, 2008).

In a similar spirit, the analyses reported here aim to further our knowledge on whether and to what extent mixture model results change meaningfully based on subtle differences in how constructs are operationalized and measured. Empirical investigation of this problem in real data has heretofore been challenging for one simple reason: It is rare for studies to systematically vary the measurement of a construct between or within subjects and compare results based on these measurement differences. Here, we report on the results of a unique laboratory analogue study that was designed to mimic measurement differences across studies in a number of self-report inventories in a college sample. Specifically, participants received different versions of the same measures, intended to measure precisely the same constructs but with superficial differences in instructions, wording, and response options (consistent with measurement differences commonly observed across independently conducted studies). Through two empirical examples of alcohol use consequences and alcohol expectancies, we examine the sensitivity of results from a latent class analysis (LCA) and a factor mixture model (FMM) to changes in measurement.

STUDY 1

Study 1 used an experimental design to manipulate the measurement of alcohol use consequences in a college sample. We investigated the stability of LCA results across four different experimental conditions, each corresponding to a different level of alteration of item stems and response options.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

Method

Participants

We obtained student contact information from the university registrar's office and selected a sampling frame to overrepresent African American students (the largest ethnic minority group on this campus) and men (given that 57% of the undergraduate population on this campus were women). A total of 6,000 students received an initial email inviting their participation (and for many, several follow-up emails), yielding a total of 854 study participants. To be included in the study, individuals must have been between 18 and 23 years of age and consumed alcohol in the past year. The final sample was 45% male, 58.1% European American, 21.9% African American, 10.4% Asian, 6.1% more than one race, and 3.5% some other race; across all races, 5.4% of participants were Hispanic or Latino. In addition, 28.6% of the participants were first-year students, 20.5% were sophomores, 20.0% were juniors, 28.9% were seniors, and 2.0% were nonstudents, did not specify, or were graduate students.

Measures

One of the goals of the REAL-U study was to empirically test the stability of findings in alcohol and drug use research across studies that use different versions of scales to measure the same constructs. Thus, items in the REAL-U study were manipulated in a number of different ways. Although we explain these differences in the context of the specific measure of alcohol use problems used in Study 1, items were manipulated similarly in both Studies 1 and 2.

Lifetime alcohol use consequences were measured using the Rutgers Alcohol Problems Index (RAPI; White & Labouvie, 1989), which has been shown to have very good internal consistency (α = .92), test–retest reliability (.89–.92), and criterion validity (Miller et al., 2002). In this study, an 18-item subset of the full 23-item questionnaire was used, based on the findings of Neal, Corbin, and Fromme (2006) that this was the best functioning subset of items, relatively free of both differential item functioning and local dependence between item pairs. Participants were instructed to indicate how many times they had experienced a given alcohol-related consequence (e.g., going to work or school drunk or waking up in an unfamiliar place after drinking) in their lifetime.

All items are shown in Table 1. Items were manipulated according to one of four versions, corresponding to increasing levels of perturbation in item stems and response categories. In Version 1, items were administered in their original form, using both the original item stems and response scales from the RAPI. In Version 2, half of the items appeared in their original form; half had perturbed item stems, based on items taken from another self-report measure of alcohol use consequences from the Core study (Presley, Meilman, & Lyerla, 1994). All items had the same response categories as the RAPI. Version 3 used the same item stems as Version 2, but used different response categories. By collapsing some categories, however, the response categories in Version 3 could be harmonized with Versions 1 and 2. Finally, Version 4 perturbed the remaining item stems (such that all stems now differed from Version 1). For Version 4 response categories, half the items maintained response categories from Version 3, whereas the other half used unique response categories that could not be collapsed to be equivalent to those in the other versions (taken from the Semi-Structured Assessment of Alcohol and Other Drugs; Buchholz et al., 1994).

For the current analyses, all items were recoded in all versions as binary. In Versions 1 and 2, responses of *none* were coded as 0, and all other responses were coded as 1. In Versions 3 and 4 under the 5-point scale, responses of *never* were coded as 0 and all other responses were coded as 1. Note that, for items measured under the 4-point scale in Version 4, this harmonization was imperfect, as the lowest category was *0–2 times*. However, the collapsing of items was intended not only to ensure comparability of solutions across measurement version and model complexity, but also to avoid problems with sparseness given the presence of a number of low-frequency response patterns in the original response scale. Thus, for all versions, a response of 0 generally indicated not having experienced a given alcohol use problem, and a response of 1 indicated having experienced this problem at least once, except in half the Version 4 items, for which it indicated three or more times.

Procedure

On each study visit, participants completed two versions of all measures on a computer. Versions of the RAPI were paired systematically such that participants either completed Versions 1 and 3 (denoted Battery A) or Versions 2 and 4 (denoted Battery B) within a given visit. Each battery was designed to be completed in roughly 75 minutes and participants completed a set of additional measures at Visit 2. Participants were compensated $20 for completion of Visit 1 and $25 for completion of Visit 2.

Participants were randomized to one of four conditions determining the combination and order of batteries they completed. As shown in Table 2, they completed either Battery A at Visit 1 and Battery B at Visit 2 (AB; *n* = 196), Battery B at Visit 1 and Battery A at Visit 2 (BA; *n* = 212), Battery A at Visit 1 and Visit 2 (AA; *n* = 213), or Battery B at Visit 1 and Visit 2 (BB; *n* = 219). Also shown in Table 2, for each version of the measure, data were split to form two analysis samples, denoted sample x and sample y. For instance, sample 1x included data on Version 1 from the first visit for individuals assigned to condition AA as well as the second visit for individuals assigned to BA. Sample 1y, in turn, included data from Version 1 from the second visit for individuals assigned to AA, as well as the first visit for individuals assigned to BA.

Partitioning the sample in this way was helpful for a number of reasons. First, the splitting of each measurement version into two equally sized analysis samples afforded the

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

TABLE 1
Study 1: Summary of All Alcohol Problems Items Used

	Version 1 (Battery A)	*Version 2 (Battery B)*	*Version 3 (Battery A)*	*Version 4 (Battery B)*
Response Scale Item	*None (0), 1–2 Times (1), 3–5 Times (2) More Than 5 Times (3)*	*None (0), 1–2 Times (1), 3–5 Times (2), More Than 5 Times (3)*	*Never (0), Once (1), Twice (2), 3–5 Times (3), 6–9 Times (4), 10 or More Times (5)*	*Never (0), Once (1), Twice (2), 3–5 Times (3), 6–9 Times (4), 10 or More Times (5) (for Nonitalicized items); 0–2 Times (0), 3–4 Times (1), 5–9 Times (2), 10 or More Times (3) (for Italicized Items)*
1	Got into fights with other people (friends, relatives, strangers)	Got into fights with other people (friends, relatives, strangers)	Got into fights with other people (friends, relatives, strangers)	Gotten into physical fights when drinking
2	Went to work or school high or drunk	Gone to class or a job when drunk	Gone to class or a job when drunk	*Gone to class or a job when drunk*
3	Caused shame or embarrassment to someone	Made others ashamed by your drinking behavior or something you did when drinking	Made others ashamed by your drinking behavior or something you did when drinking	*Made others ashamed by your drinking behavior or something you did when drinking*
4	Neglected your responsibilities	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
5	Relatives avoided you	Family members rejected you because of your drinking	Family members rejected you because of your drinking	*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	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 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: Have you tried to cut down or quit drinking or using alcohol or other drugs?
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	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	Acted in a very different way or did things you normally would not do because of your drinking	*Acted in a very different way or did things you normally would not do because of your drinking*
10	Felt that you had a problem with alcohol	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	Tried unsuccessfully to stop drinking	*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	Awakened the morning after some drinking the night before and could not remember a part of the evening	*Awakened the morning after some drinking the night before and could not remember a part of the evening*
13	Passed out or fainted suddenly	Passed out after drinking	Passed out after drinking	*Passed out after drinking*
14	Had a fight, argument, or bad feeling with a friend	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
15	Kept drinking when you promised yourself not to	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
16	Felt you were going crazy	Your drinking made you feel out of control even when you were sober	Your drinking made you feel out of control even when you were sober	*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	Felt physically or psychologically dependent on alcohol	Thought you were dependent on alcohol
18	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	Near relative or close friend worried or complained about your drinking	Near relative or close friend worried or complained about your drinking

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

TABLE 2
Study 1: Summary of the Composition of Each Analysis Sample

	Condition/Visit								
	AA (N = 213)		*BB (N = 219)*		*AB (N = 196)*		*BA (N = 212)*		
Analysis Sample	*1*	*2*	*1*	*2*	*1*	*2*	*1*	*2*	*N*
1x	1							1	425
1y		1			1				409
2x			2			2			415
2y				2			2		431
3x	3							3	425
3y		3			3				409
4x			4			4			415
4y				4			4		431

Note. Numbers in the body of the table refer to the measurement version sampled within a given battery.

opportunity to replicate LCA results for each version, establishing a baseline level of stability for model results when there is no measurement perturbation. Second, each analysis sample included data on a given measurement version from both Visits 1 and 2, to balance order effects; for instance, analysis sample 2x included data from measurement version 2 taken from Group BB at Visit 1 and AB at Visit 2. Finally, the overlap between analysis samples allowed for the examination of class assignment stability within and between measurement versions. For instance, half of the members of analysis sample 2x came from Group BB; thus, they were also in analysis samples 2y, 4x, and 4y. The other half of analysis sample 2x came from Group AB; thus, were also in analysis samples 1y and 3y. This permitted within-group comparisons (i.e., by comparing class assignments between 2x and 2y) and between-group comparisons (i.e., by comparing class assignments between 2x and 1y, 3y, 4x, and 4y).

Analyses

LCA (Clogg & Goodman, 1984; Lazarsfeld & Henry, 1968) models were fit to binary alcohol use consequence items separately for each of the eight analysis samples using M*plus* version 7.2 (Muthén & Muthén, 2015). A latent class model consists of classes defined by categorical observed variables, which are assumed conditionally independent given class membership. Let i index subjects (where $i = 1, \ldots, N$), q index binary items (where $q = 1, \ldots, Q$), k index latent classes (where $k = 1, \ldots, K$), and p index covariates (where $p = 1, \ldots, P$). Define the vector of item responses for subject i as $\mathbf{y}_i$, with individual elements y_{iq} that represent subject i's response to the qth binary item. Then the latent class analysis model is given by:

$$P(\mathbf{y}_i = 1) = \sum_{k=1}^{K} \pi_k P(\mathbf{y}_i = 1 | c_{ik} = 1) \qquad (1)$$

where c_{ik} is an indicator variable that takes on a value of 1 if subject i is a member of class k and 0 otherwise, and π_k is the prevalence of class k, subject to the constraints that π_k ranges from 0 to 1, and $\sum_{k=1}^{K} \pi_k = 1$.[2] The class-specific probability mass function for subject i under class k is

$$P(\mathbf{y}_i = 1 | c_{ik} = 1) = \prod_{q=1}^{Q} P\left(y_{iq} = 1 | c_{ik} = 1\right). \qquad (2)$$

Critically, this formulation implies conditional independence of indicators, given class membership. This assumption can be relaxed to allow continuous factors to account for local dependence between pairs of items (Reboussin, Ip, & Wolfson, 2008), or for substantively meaningful factors to be defined on the basis of multiple indicators, as in the FMM presented in Study 2. Although preliminary analyses determined that local dependence might exist between some item pairs for some of the models under consideration, the offending item pairs were not consistent across models and incorporating local dependence proved computationally intractable. Thus, we proceeded with the typical LCA formulation here.

Item endorsement probabilities, as well as potential covariate effects, were compared across the eight analyses. Additionally, to gauge the agreement between the modal classifications given by the optimal model for each version of the measure, the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) was computed for each pairwise combination of LCA solutions both within-version/between-subsample (e.g., comparing the solution for analysis samples 1x and 1y), and between-version/within-subsample (e.g., comparing the solution for analysis samples 1x and 2x). The ARI measures the concordance between two partitions of the same data, adjusting for chance, and ranges from −1 to 1, with values closer to 1 indicating greater agreement between the two classifications (Steinley, 2004).

Results

Class Enumeration

Model fit statistics informing class enumeration are presented in Table 3. Initially, class enumeration was informed by consideration of Akaike's information criterion (AIC; Akaike, 1998), Bayesian information criterion (BIC;

[2] Importantly, class membership probabilities can be affected by covariates (Huang & Bandeen-Roche, 2004). Preliminary analyses included gender, African American and Asian race, and Hispanic or Latino ethnicity as covariates affecting class membership. However, neither class enumeration nor the LCA solutions themselves (i.e., item endorsement patterns and class prevalence rates) changed with the inclusion of these covariates. Thus, in the interest of parsimony, we exclusively consider an unconditional model here, so class membership probabilities π_{ik} do not vary over individuals and become overall prevalence rates π_k.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

TABLE 3
Study 1: Fit Indexes for Latent Class Analysis Models With Different Numbers of Classes Under Each Measurement Version

Version 1		*Sample 1x*				*Sample 1y*			
K	*Parameters*	*LL*	*BIC*	*LMR*	*LMR p Value*	*LL*	*BIC*	*LMR*	*LMR p Value*
1	18	−3243.53	6595.96	N/A	N/A	−3212.3	6532.67	N/A	N/A
2	37	−2803.87	5831.57	879.334	.0002	−2693.38	5608.89	1037.850	**< .0001**
3	56	−2678.06	**5694.91**	251.611	.001	−2592.6	**5521.42**	201.545	.2465
4	75	−2621.27	5696.27	113.581	**.0053**	−2548.97	5548.23	87.267	.2487
5	94	−2585.74	5740.16	71.055	.241	−2507.43	5579.23	83.077	.4613
6	113	−2561.03	5805.68	49.423	.4393	−2479.25	5636.95	56.352	.254
7	132	−2538.57	5875.71	44.921	.4845	−2456.58	5705.67	45.350	.7603
Version 2		*Sample 2x*				*Sample 2y*			
K	*Parameters*	*LL*	*BIC*	*LMR*	*LMR p Value*	*LL*	*BIC*	*LMR*	*LMR p Value*
1	18	−3644.5	7397.41	N/A	N/A	−3600.69	7310.48	N/A	N/A
2	37	−3022.69	6268.26	1243.600	**< .0001**	−3014.34	6252.96	1172.690	< .0001
3	56	−2871	6079.31	303.393	.0578	−2887.87	6115.17	252.954	.011
4	75	−2808.58	**6068.91**	124.840	.0418	−2815.92	**6086.46**	143.883	< .0001
5	94	−2757.86	6081.92	101.439	.0359	−2772.45	6114.67	86.949	**.0006**
6	113	−2723.27	6127.19	69.179	.1838	−2741.71	6168.36	61.480	.8012
7	132	−2693.2	6181.49	60.146	.3856	−2711.21	6222.53	60.799	.1744
Version 3		*Sample 3x*				*Sample 3y*			
K	*Parameters*	*LL*	*BIC*	*LMR*	*LMR p Value*	*LL*	*BIC*	*LMR*	*LMR p Value*
1	18	−3021.04	6150.84	N/A	N/A	−3079.64	6267.18	N/A	N/A
2	37	−2491.42	5206.41	1059.240	< .0001	−2488.04	5197.86	1183.200	< .0001
3	56	−2369.44	**5077.27**	243.946	.0076	−2360.95	**5057.57**	254.181	.0021
4	75	−2312.36	5077.92	114.160	**.0042**	−2311.36	5072.27	99.177	**.0006**
5	94	−2280.97	5129.96	62.779	.7179	−2280.57	5124.57	61.586	.0812
6	113	−2246.75	5176.32	68.446	.1493	−2250.85	5179.01	57.273	.9061
7	132	−2224.82	5247.27	43.864	.2427	−2223.80	5238.81	55.908	1
Version 4		*Sample 4x*				*Sample 4y*			
K	*Parameters*	*LL*	*BIC*	*LMR*	*LMR p Value*	*LL*	*BIC*	*LMR*	*LMR p Value*
1	18	−2334.85	4778.08	N/A	N/A	−2299.78	4708.5	N/A	N/A
2	37	−1851.77	3926.31	966.164	< .0001	−1882.04	3988.00	835.487	< .0001
3	56	−1767.44	**3871.86**	168.653	**.0047**	−1779.99	**3898.89**	204.097	**< .0001**
4	75	−1724.56	3900.69	85.564	.0681	−1748.11	3950.12	63.767	.2147
5	94	−1698.87	3963.71	51.378	.2086	−1726.81	4022.51	42.592	.3062
6	113	−1674.3	4028.97	49.139	.2814	−1705.88	4095.65	43.381	1
7	132	−1654.68	4104.14	39.224	.3878	−1688.91	4176.70	35.234	.6435

Note. LL = log-likelihood; BIC = Bayesian information criterion; LMR = Lo–Mendell–Rubin test statistic testing the null hypothesis that a model with *K* – 1 classes fits as well as a model with *K* classes; LMR *p* value = the *p* value for the LMR statistic. Entries corresponding to the value of *K* favored by a given fit index are shown in bold.

Schwarz, 1978), Vuong Lo–Mendell–Rubin likelihood ratio test (LMR; Lo, Mendell, & Rubin, 2001; Vuong, 1989), and the bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000). However, ultimately only the BIC and LMR *p* value were considered as criteria because, with very few exceptions, neither the AIC nor the BLRT favored a value of *K* within the range of models considered (i.e., they continued to support more classes even at seven classes). When there was disagreement between the BIC and LMR, the BIC was generally favored, given that previous simulation work supported its accuracy in detecting the correct number of classes (Nylund et al., 2007; Tofighi & Enders, 2008).

For all models (Table 3), fit indexes showed varying levels of agreement between and within measurement versions. The BIC was minimized for K = 3 classes in all measurement versions except for Version 2, in which the BIC favored a four-class solution in both analysis samples 2x and 2y. In analysis samples 1x and 1y, the three-class model was only narrowly favored over a four-class model

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

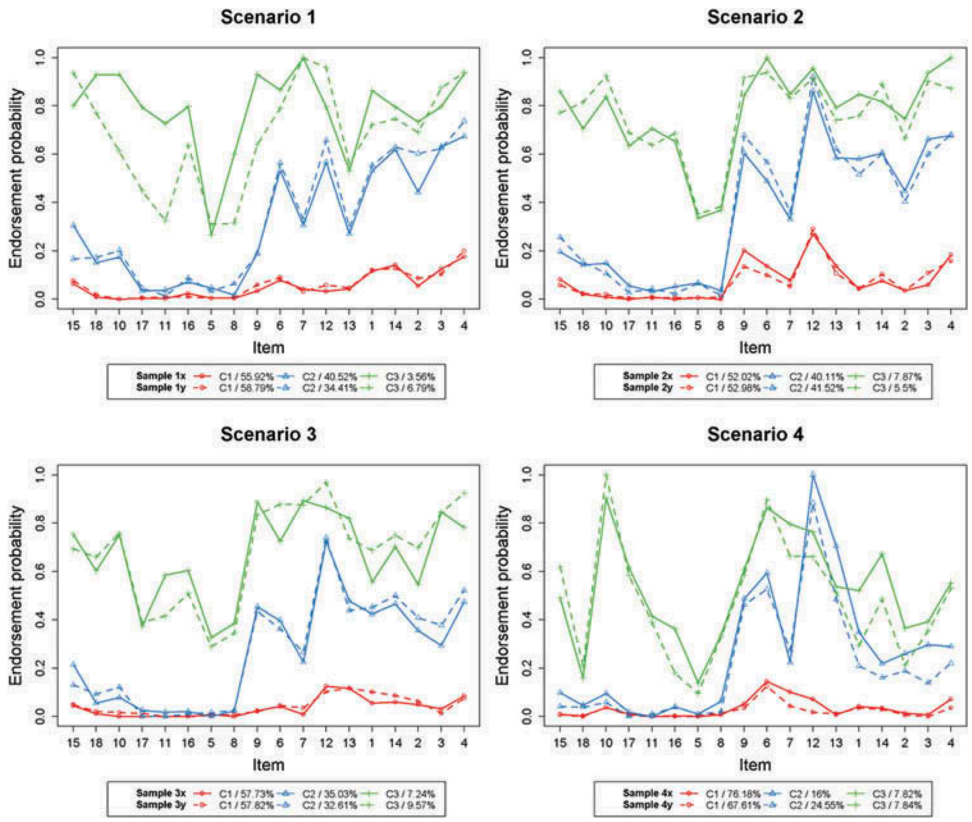

FIGURE 1 Study 1: The three-class model under all measurement versions.

by the BIC. The LMR test was highly inconsistent across analysis samples for Versions 1 and 2, favoring a two-class solution in samples 1y and 2x, a four-class solution in sample 1x, and a five-class solution in sample 2y. However, in Versions 3 and 4, the LMR was in agreement across samples, favoring a four-class solution in Version 3 and a three-class solution in Version 4. Solutions with more than three classes were generally unstable in sample 4x; each solution had at least one extremely small class in which parameters could not be freely estimated.

In summary, the BIC generally favored a three-class solution in Versions 1, 3, and 4, and a four-class solution in Version 2. The LMR favored anywhere between two and five classes, with little consistency between and within measurement versions.[3] Given this mixed support, we considered both the three- and four-class solutions across all versions.

Endorsement Probabilities and Class Prevalences

Figures 1 and 2 show model-implied endorsement probabilities for all three-class and four-class solutions in each measurement version. Note that, due to the instability of the four-class solution in Version 4x, it is not considered further and only Version 4y is presented. To present items in a way that facilitates their interpretation, the optimal order of items on the x-axis was determined using a hierarchical clustering algorithm (Fraley & Raftery, 2002). This algorithm groups together items that were highly correlated with one another in a full-sample analysis.

Three-class solutions. Item endorsement patterns for all three-class solutions are shown in Figure 1. Across all measurement versions, the three-class solution identified one class (Class 1) comprising the majority of the sample, which was characterized by generally low probabilities of endorsing all items. In Versions 1, 2, and 3, the classes largely captured differences in overall level of endorsement, with Class 3 endorsing most items with high probability and Class 2 endorsing roughly half the items (those on the left side of the x-axis) with low probability and roughly half (those on the right side) with intermediate to high probability. The items that were endorsed most frequently by this class generally pertained to either loss of control (e.g., Items 6, 12, and 13) or social consequences (e.g., Items 1, 14, and 4). Two items, Item 9 (V1: "Noticed a change in your personality"; V2, V3, V4: "Acted in a very different way or did things you normally would not do because of your drinking") and Item 12 (V1: "Suddenly found yourself in a place that you could not

[3] This lack of consistency within version raises questions about the use of LMR LRT for class enumeration in general.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

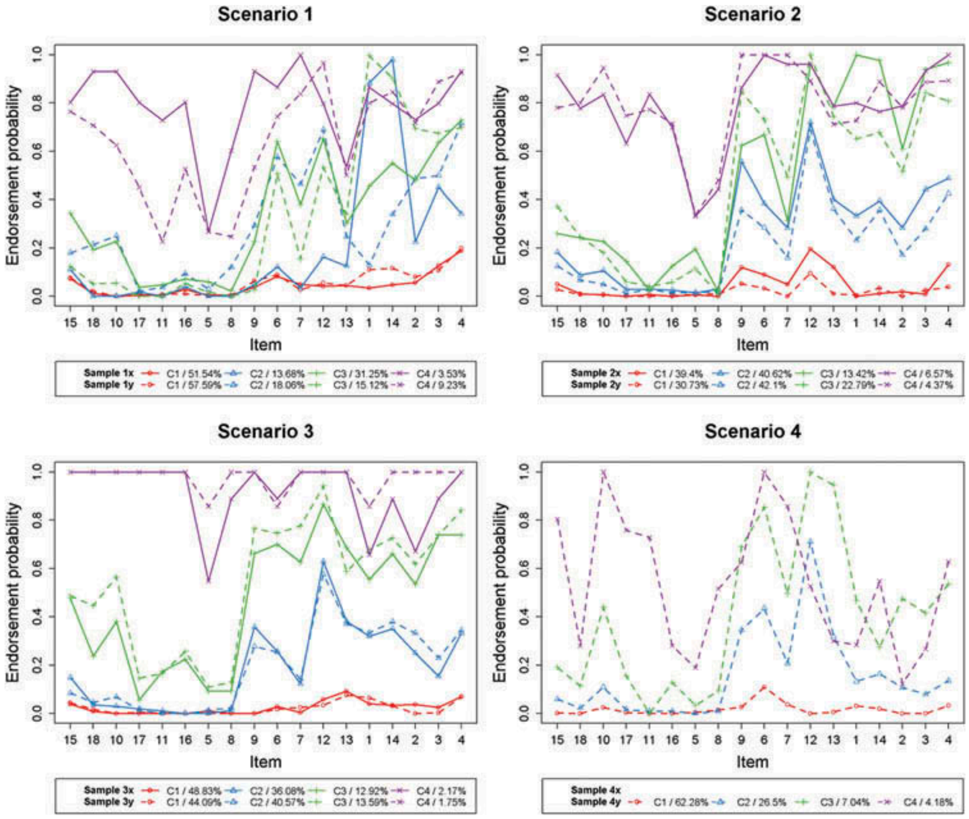

FIGURE 2 Study 1: The four-class model under all measurement versions.

remember getting to"; V2, V3, V4: "Awakened the morning after some drinking the night before and could not remember a part of the evening"), appeared to be less frequently endorsed by this intermediate group in Version 1 than in all other versions. By contrast, items infrequently endorsed by this group relative to Class 3 generally pertained to symptoms of dependence (Items 10, 11, 15, 16, and 17) or family or close relations disapproving of one's drinking (Items 5 and 18).

To quantify the overall extent of the similarity between versions in endorsement patterns, the Euclidean distances between within-class endorsement probabilities (averaged across samples) were calculated for each pair of versions. These values are shown in the top half of Table 4. Differences between versions in Class 1 were generally small, corresponding to the generally low levels of endorsement in all versions. Differences in Classes 2 and 3 were greatest between Version 4 and the other three versions. In Version 4, Class 2 was characterized by a lower probability of endorsing Items 3, 4, and 14, but a higher probability of endorsing Item 12, than in Versions 1 and 2. Additionally, in Version 4, Class 3 was characterized by lower endorsement probabilities on Items 2, 3, 16, and 18 than in the other versions. Of these items, Versions 2 and 4 used the same stems for all but Item 4 (V2: "Neglected your responsibilities"; V4: "Neglected your obligations, your family, or your work for two or more days in a row because you were drinking") and Item 14 (V2: "Had a fight, argument, or bad feeling with a friend"; V4: "Drinking created problems between you and a near relative or close friend"). This commonality suggested that changes to the stems for these items did not solely account for the differences observed for this version.

It was of interest to compare Version 2 to Versions 1 and 3, because Version 2 had the same response options as Version 1 but 50% different item stems, and the same item stems as Version 3 but different response options. Although the squared Euclidean distances did not indicate differentially close relationships between Version 2 and either Version 1 or 3, visual inspection of Figure 1 suggested that the general pattern of item endorsements might be somewhat closer between Versions 2 and 3 than between Versions 1 and 2. This impression was further supported by the fact that there was greater concordance between Versions 2 and 3 in the rankings of item endorsement probabilities relative to one another (Spearman's $\rho = .92$ for Class 2, $\rho = .88$ for Class 3) than between Versions 1 and 2 ($\rho = .85$ for Class 2, $\rho = .74$ for Class 3). Versions 2 and 3 appeared to differ mainly in the severity of the items, with Items 9 and 12 being endorsed more frequently in Version 2 than in Version 3. Additionally, the prevalence of the low class relative to the intermediate class was different between these

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

TABLE 4
Study 1: Euclidean Distance Between Profile Solutions of Each Version Under Each Latent Class

	Class 1					*Class 2*					*Class 3*					*Class 4*			
	V1	*V2*	*V3*	*V4*		*V1*	*V2*	*V3*	*V4*		*V1*	*V2*	*V3*	*V4*		*V1*	*V2*	*V3*	*V4*
3-Class Solution																			
V1	0				V1	0				V1	0								
V2	0.0888	0			V2	0.4015	0			V2	0.1773	0							
V3	0.0392	0.0712	0		V3	0.3471	0.2793	0		V3	0.2826	0.2629	0						
V4	0.0621	0.1137	0.037	0	V4	1.066	0.717	0.316	0	V4	1.706	1.762	1.146	0					
4-Class Solution																			
V1	0				V1	0				V1	0				V1	0			
V2	0.0472	0			V2	0.3710	0			V2	0.9062	0			V2	0.2862	0		
V3	0.0418	0.0222	0		V3	0.3663	0.0976	0		V3	1.0723	0.3693	0		V3	1.4901	0.7783	0	
V4	0.062	0.037	0.022	0	V4	0.917	0.305	0.2276	0	V4	1.502	0.975	0.728	0	V4	0.928	0.718	1.4	0

TABLE 5
Study 1: Adjusted Rand Indexes for All Models

	3-Class Models					*4-Class Models*			
Version	*1y*	*2y*	*3y*	*4y*	*Version*	*1y*	*2y*	*3y*	*4y*
1x	**0.4264**	0.3634	0.3011	0.432	1x	**0.3365**	0.2105	0.2729	0.3912
2x	0.4604	**0.5999**	0.3436	0.2897	2x	0.3113	**0.3888**	0.3714	0.226
3x	0.3787	0.3106	**0.3677**	0.2765	3x	0.3449	0.1781	**0.3536**	0.2859
4x	0.2126	0.4305	0.2528	**0.6177**	4x	—	—	—	—

two versions, with Version 3 placing more subjects in the low class and fewer in the intermediate class than Version 3.

Four-class solutions. Item endorsement patterns for all four-class solutions, with the exception of sample 4x, are shown in Figure 2. Here solutions were considerably less stable within versions than in the three-class case, particularly for the high symptomatology classes with low prevalence rates, posing some challenge to interpretation. However, two things were particularly noteworthy with respect to the prevalence rates of each class. First, as in the three-class solutions, the four-class solutions identified one class (Class 1) that was characterized by low endorsement probabilities for all items; however, the prevalence of this class varied widely across versions, with Version 2 placing the smallest portion of the sample into this class and Version 4 placing the largest portion of the sample into this class. Second, also as in the three-class solutions, all solutions here found a class (Class 4) with generally high levels of endorsement; however, the prevalence of this class varied widely within and between versions and was generally quite small (with a maximum prevalence of 9.23% in sample 1y). In Version 3, this extremely high level of endorsement was uniform across most items, whereas in Versions 1 and 2 there was considerably more variation among item endorsements.

A number of interesting differences between versions emerged with respect to the two intermediate classes. As in the three-class solutions, there was some support for the conclusion that the general shapes of the classes in Version 2 were more similar to those of Version 3 than to Version 1, with greater concordance in rank order between Versions 2 and 3 in item endorsement rates for Class 2 (ρ = .93 between Versions 2 and 3, ρ = .81 between Versions 2 and 1) but not for Class 3 (ρ = .89 between Versions 2 and 3, ρ = .90 between Versions 2 and 1). As in the three-class solution, Versions 2 and 3 were differentiated largely by item severity, with generally higher endorsement rates for a number of items in the intermediate classes in Version 2 than Version 3. Unlike in the three-class solution, here the Euclidean distance between Versions 2 and 3 was smaller than that between Versions 1 and 2 for all classes.[4] Additionally, Version 4's difference from the other versions in endorsement patterns was not as pronounced as in the three-class solution. Although there were some differences between Version 4 and Versions 1 and 2, particularly in Class 3, these differences were somewhat challenging to interpret because of within-version differences in endorsement patterns.

Finally, some of the most consistent differences between versions were seen in class prevalence. In particular, the prevalence of the low-endorsement class was lowest in Version 2 and highest in Version 4. Both Versions 2 and 3 placed a majority of the sample in intermediate classes (i.e., Classes 2 and 3). By contrast, both Versions 1 and 4 placed a majority of the sample in Class 1, with intermediate classes being somewhat smaller and less stable across analysis samples.

Class Assignments

Table 5 shows the ARIs for modal class assignments in both three-class and four-class solutions. Diagonal elements indicate the stability of class assignments within different samples in the same measurement version, with the exception of Version 4, in which only sample 4y is considered. In the three-class model, within-version class membership was most stable within Version 4. Version 2 showed similar levels of within-version stability in the three-class solutions, but was less stable in the four-class solution; thus, in addition to being supported by the BIC as balancing fit and parsimony, the three-class solution appeared to be particularly reliable in Version 2. As discussed earlier, the general shape of Version 2's endorsement profiles corresponded more to that of Version 3, with which it shared common item stems, than to that of Version 1, with which it shared response options. However, the ARI did not present the impression that class membership was especially stable from Versions 2 to 3 in either the three- or four-class versions.

Summary

Study 1 examined LCAs of alcohol consequences under four different measurement versions. There were some

[4] As in the three-class solutions, Euclidean distance was computed on probabilities averaged across samples for a given version (e.g., samples 1x and 1y), with the exception of Version 4, in which only sample 4y was used.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

differences across versions in class enumeration, with the BIC indicating that a three-class solution fit best in all versions other than Version 2, in which the four-class solution was favored. Both the three- and four-class solutions showed some degree of difference in item endorsement patterns and class prevalence rates across measurement versions. In both the three- and four-class solutions, differences from Version 1 in item endorsement patterns within each class generally increased with greater levels of measurement perturbation, with Version 4, corresponding to the highest level of item alteration, showing the greatest difference in the shapes of the class endorsement profiles. There were differences in class prevalence rates across versions, although these differences did not correspond directly to the degree of item perturbation, particularly in the four-class case. In particular, whereas Versions 2 and 3 placed a large portion of the sample into classes characterized by intermediate levels of item endorsement, the low-endorsement class was considerably larger in Version 4, in both the three-class and four-class solutions. This finding is particularly interesting given that, in all versions aside from Version 4, a response of 0 always corresponds to a subject never having experienced a given consequence; in Version 4, a response of 0 might correspond to *never* (for items originally measured using the 5-point scale) or *0–2 times* (for items originally measured using the 4-point scale). Thus, the high prevalence of the low-endorsement class in Version 4 might reflect the higher threshold required to endorse items originally measured using the 4-point response scale. In sum, the results obtained from LCA models differed in a number of important ways across variations in measurement. We now examine the extent to which this sensitivity is also observed for FMMs.

STUDY 2

Study 2 used the REAL-U data described in Study 1 to investigate the stability of FMM results across two highly disparate measurement versions. This study focused exclusively on differences across Versions 1 and 4, the two most dissimilar experimental conditions in the study, in the nature of FMM results. Class enumeration has been shown to be highly sensitive to measurement in GMMs, a special case of FMM (K. M. Jackson & Sher, 2005), and we investigated the possibility of different numbers of classes being chosen in Versions 1 and 4. Unlike in GMM, however, in FMM measurement parameters (factor loadings and thresholds) are freely estimated; thus, it was of primary interest to determine whether and to what extent the measurement properties of items within and between classes differed on the basis of alterations to items.

Method

Participants

Participants ($N = 854$) were the same as those in Study 1.

Measures

Alcohol expectancies were measured using 14 items in two subscales, relating to tension reduction and sociability; these items are shown in Table 6. These items were drawn from a larger pool of 17 items administered in the REAL-U study, but three items were removed due to problematic characteristics, including cross-loadings or local dependence, in one or both versions in preliminary analyses. Tension reduction items were taken from the corresponding subscale on the 9-item Alcohol Outcome Expectancies scale, which has good internal consistency (α = .89; Kushner, Sher, Wood, & Wood, 1994). Sociability items were taken from the corresponding subscale in the Brief Comprehensive Effects of Alcohol (B-CEOA; Fromme, Stroot, & Kaplan, 1993); these items show fair internal consistency (α = .81). Items were manipulated according to the same measurement versions as in Study 1, with the exception that here the instructions were also altered between measurement versions.

Data were collapsed to a 3-point ordinal scale in both versions to enhance comparability and to eliminate sparse categories that could cause estimation difficulties. Original response options are shown in Table 6. In Version 1, responses were originally measured using a 5-point scale, ranging from *not at all* to *a lot*. These options were recoded so that a response of *not at all* or *a little bit* was coded as 1, *somewhat* was coded as 2, and *quite a bit* or *a lot* was coded as 3. Half of the items in Version 4 were measured using a 4-point scale ranging from *disagree* to *agree* and the other half with a 5-point scale ranging from *no chance* to *certain to happen*. Items on the 4-point scale were recoded so that a response of *disagree* or *slightly disagree* was coded as 1, *slightly agree* was coded as 2, and *agree* was coded as 3. Items on the 5-point scale were recoded so that a response of *no chance* or *very unlikely* was coded as 1, *unlikely* was coded as 2, and *very likely* or *certain to happen* were coded as 3. Although these response options are clearly not harmonizable to categories with identical meanings across scales, such situations are not unusual when comparing results across studies in the absence of a gold standard measure and this experimental condition was meant to mimic such conditions.

Procedure

The experimental procedure and study design were the same as those in Study 1. However, a different subsampling strategy was used to generate analysis samples in Versions 1 and 4. In this analysis, comparing results between Versions 1 and 4 was of primary interest; for this reason, and to maximize

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

TABLE 6
Study 2: Summary of All Alcohol Expectancies Items Used

		Version 1 (Battery A)	*Version 4 (Battery B)*
Instructions		The following items describe some effects of alcohol. Because alcohol affects people in different ways, we would like to know which of these effects you experience when you drink alcohol. Based on your own drinking experience, how much do you expect each of these effects when drinking alcohol? (If you have never consumed alcohol, indicate how you might expect alcohol to affect you if you had several drinks.)	Choose from DISAGREE TO AGREE depending on whether you expect the effect to happen to you IF YOU WERE UNDER THE INFLUENCE OF ALCOHOL. These effects will vary, depending on the amount of alcohol you typically consume. Check one answer after each statement. There are no right or wrong answers. (If you have never consumed alcohol, indicate how you might expect alcohol to affect you if you had several drinks).
Response scale		*Not at all* (0), *A little bit* (1), *Somewhat* (2), *Quite a bit* (3), *A lot* (4)	*Disagree* (1), *Slightly Disagree* (2), *Slightly Agree* (3), *Agree* (4) (for nonitalicized items). *No chance* (0), *Very unlikely* (1), *Unlikely* (2), *Very likely* (3), Certain to happen (4) (for italicized items)
Item	*Factor*		
1	Tension reduction	Drinking helps me to relax.	I would feel calm.
2	*Tension reduction*	Drinking helps me forget problems at work or school.	*I would be able to take my mind off my problems.*
3	*Tension reduction*	Drinking helps me feel better about myself.	*I would be more satisfied with myself.*
4	Tension reduction	Drinking helps me forget my worries.	I would feel less worried.
5	*Tension reduction*	Drinking helps me feel better when I'm feeling down.	*I would feel less depressed.*
6	Tension reduction	Drinking helps me relax when I'm tense.	I would be less tense.
7	*Tension reduction*	Drinking helps me to calm down when I'm angry.	*I would feel less hostile.*
8	*Tension reduction*	Drinking helps me deal with boredom.	*I would be less likely to have negative moods or feelings.*
9	Tension reduction	Drinking helps me express my opinions and ideas better.	I would be able to discuss or argue a point more forcefully.
10	*Sociable*	Drinking helps me act sociable.	*I would be more sociable.*
11	Sociable	Drinking helps me talk to people.	I would talk to people more easily.
12	Sociable	Drinking helps me to be friendly.	I would be friendlier.
13	*Sociable*	Drinking helps me to be talkative.	*I would be more "chatty."*
14	Sociable	Drinking helps me to be outgoing.	I would be more likely to be courageous.
15	*Sociable*	Drinking helps me to be humorous.	*I would be more likely to have my humorous side come out.*
16	*Sociable*	Drinking helps me express my feelings.	*I would more easily open up and express my feelings.*
17	Sociable	Drinking helps me feel energetic.	I would feel better physically.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

sample size, only one large subsample was investigated for Versions 1 and 4. Data came from both groups who received a given measurement version at Visit 1, as well as whichever nonredundant group received that measurement version at Visit 2. Thus, data for Version 1 came from Groups AB and AA at Visit 1 and Group BA at Visit 2, yielding a total $N = 635$; data for Version 4 came from Groups BA and BB at Visit 1 and Group AB at Visit 2, yielding a total $N = 641$. Thus, because Groups AB and BA were common to both Versions 1 and 4, the samples overlap greatly, with 65.1% of individuals in Version 4 also measured under Version 1, reducing the extent to which differences obtained across the two versions might reflect simple sampling variability (because the majority of the two samples consisted of the same individuals).

Analyses

FMMs were fit to ordinal alcohol expectancies items, assuming that a two-factor structure held in all classes. A brief description of this model follows, but see Lubke and Muthén (2005, 2007), Lubke and Neale (2008), or Muthén (2006) for a more complete description of the FMM.

As in the LCA presented in Study 1, we define y_{iq} as subject i's response to qth item and c_{ik} as an indicator variable that takes on a value of 1 if subject i is a member of class k and 0 otherwise. Within each class, items are assumed to be affected by a set of R continuous, normally distributed factors η_i according to a common factor model. As the data in this study were three-level ordinal variables, we implemented a cumulative logit model specification for the regression of the indicators on the latent factors.

Define $P(y_{iq} \leq j)$ as the probability of endorsing any response option up to and including j, where $j = 1$ or 2 (because the cumulative probability for $j = 3$ is by definition 1.0). This cumulative probability is calculated by marginalizing across continuous and categorical latent variables as follows:

$$P(y_{iq} \leq j) = \sum_{k=1}^{K} \pi_k \int P(y_{iq} \leq j | c_{ik} = 1, \boldsymbol{\eta}_i) \partial \boldsymbol{\eta}_i \tag{3}$$

where $P(y_{iq} \leq j | c_{ik} = 1, \boldsymbol{\eta}_i)$ is the probability of endorsing any response option up to and including j on item q given subject i's values of the continuous and categorical latent variables, and π_k is the probability that subject i is a member of class k, subject to the constraints that π_k ranges from 0 to 1, and $\sum_{k=1}^{K} \pi_k = 1$.

Within a given class, the distribution of $\boldsymbol{\eta}_i$ is assumed multivariate normal with $R \times 1$ mean vector $\boldsymbol{\mu}_k$ and $R \times R$ covariance matrix $\boldsymbol{\psi}_k$. The class-specific cumulative probability $P(y_{iq} \leq j)$ is related to the latent factors as follows:

$$\text{logit}\left(P(y_{iq} \leq j | c_{ik} = 1, \boldsymbol{\eta}_i)\right) = \tau_{kjq} - \boldsymbol{\lambda}_{kq} \boldsymbol{\eta}_i. \tag{4}$$

Class-specific measurement parameters are defined as in a common factor model: τ_{kjq} is a class-specific threshold parameter for response j on item q, and the $R \times 1$ vector $\boldsymbol{\lambda}_{kq}$ contains class-specific factor loadings that transmit the effect of latent variables $\boldsymbol{\eta}_i$ onto the cumulative logit for item q.

One of the strengths of FMM is that it allows for the assessment of measurement invariance (Mellenbergh, 1989; Meredith, 1993; Vandenberg & Lance, 2000) across latent classes in the population (Lubke & Muthén, 2005). In particular, one might be interested in whether certain segments of the population display fundamental differences in the organization or measurement of the underlying factors relative to other segments. To evaluate this question, we estimated FMMs assuming three distinct levels of measurement invariance, corresponding to configural invariance, weak metric invariance, and strong metric invariance across classes. The least restrictive of these models, the configural invariance model, assumes only that the pattern of factor loadings is the same across classes. The weak metric invariance model assumes equality of factor loadings across classes, and the strong metric invariance model additionally assumes equality of item thresholds across classes.

Although a review of measurement invariance testing in FMM is outside the scope of this work (see Clark et al., 2013; Lubke & Neale, 2008; Muthén, 2006), there are a few issues that distinguish measurement invariance testing in FMM from the evaluation of factor models fit to multiple observed groups. First, it is critical to note that in FMM the composition of classes could change on the basis of the level of measurement invariance assumed. Thus, although it is possible to compare the fit between, for example, models assuming weak versus strong measurement invariance across classes, the individuals within each class could shift between models, complicating their comparison in a more substantive sense. Second, even the number of classes deemed optimal for the data might differ depending on the invariance restrictions imposed on the model. That is, on one hand, the number of classes might be underestimated by assuming too low a level of invariance, owing to the inclusion of unnecessary model parameters. On the other, the number of classes might be overestimated by assuming too high a level of invariance, due to the potential for additional latent classes to compensate for model misspecification. Further, given the complexity of FMMs, information criteria such as the BIC might erroneously favor a more constrained model over the correct, noninvariant model (Lubke & Neale, 2008). As such, although we present comparisons of model fit here, we also note that hypothesis tests in FMMs must always be interpreted cautiously.

Results

Values of BIC used in determining K are shown in Table 7. The configural invariance model could not support a solution with more than two classes. In both measurement versions, BIC favored the two-class solution over either a

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

TABLE 7
Study 2: All Values of BIC Used in Model Selection

	BIC	
K	*Version 1*	*Version 4*
Strong metric invariance		
1	12650.544	12229.431
2	12648.558	12205.890
3	12667.305	12222.515
Weak metric invariance		
1	12650.544	12229.431
2	**12604.522**	**12138.856**
3	12677.704	12211.276
Configural invariance		
1	12650.544	12229.431
2	12637.701	12147.705

Note. BIC = Bayesian information criterion. Note that the one-class strong metric invariance, weak metric invariance, and configural invariance models are the same.

one- or three-class solution for the weak and strong invariance models and there were indications of estimation problems with three classes.

Given that the weak, strong, and configural invariance models all supported a two-class solution in both measurement versions, likelihood ratio tests (LRTs) were consulted in determining the optimal level of invariance. Despite having potentially limited substantive interpretability, as discussed earlier, the LRT nevertheless provides a useful comparison between models in terms of their overall balance of fit and parsimony. In both measurement versions, the two-class strong metric invariance model fit significantly worse than the two-class weak metric invariance model: Version 1, $\chi^2(26) = 142.18$, $p < .001$; Version 4, $\chi^2(23) = 171.52$, $p < .001$. The weak metric invariance model, in turn, fit significantly worse than the configural invariance model in both versions: Version 1, $\chi^2(12) = 42.36$, $p < .001$; Version 4, $\chi^2(12) = 34.83$, $p < .001$. We considered allowing for partial weak invariance across classes, but partial weak invariance was also rejected relative to configural invariance.[5] The disagreement between the LRT results, which favored the configural invariance model, and the BICs, which favored a weak invariance model, underscores the challenges in making meaningful comparisons between FMMs with different levels of invariance. Thus, despite the fact that the two-class weak metric invariance model was favored by the BIC relative to the two-class configural invariance model, we proceeded in interpreting the two-class configural invariance model in both versions.

The Two-Class Configural Invariance Solution

In both versions, the two-class solution divided the sample into relatively large classes in which the sociability and tension reduction factors were positively correlated. In Version 1, 46.54% of the sample fell into Class 1, in which the factors were correlated at $r = .825$; 53.46% of the sample fell into Class 2, in which the tension reduction and sociability factors were correlated at $r = .412$. In Version 4, 59.20% of the sample fell into Class 1, in which the tension reduction and sociability factors were correlated at $r = .721$; 40.80% of the sample fell into Class 2, in which the factors were correlated at $r = .628$. For the subset of individuals measured using both Versions 1 and 4, the ARI comparing modal class membership estimates under the two measurement versions was .0014, indicating no concordance between the two versions.

Factor loadings. Standardized loadings are shown in Figure 3; note that the background of the plot is shaded for items originally measured using the 5-point scale. We first considered the loadings for tension reduction (top panels), then sociability (bottom panels). In Version 1, loadings for the tension reduction factor were generally weaker in Class 2 than Class 1. Whereas Class 1 was characterized by consistently high loadings for all items on the tension reduction factor, in Class 2 Items 6, 7, and 8 (V1: "Drinking helps me relax when I'm tense," "Drinking helps me to calm down when I'm angry," "Drinking helps me deal with boredom," respectively) appeared particularly weak. By contrast, in Version 4, loadings for the tension reduction factor were relatively close in both Classes 1 and 2. Also different from Version 1 is the fact that in Version 4 the same general pattern of loadings—with Items 5 and 8 showing a slightly stronger relationship to the latent factor than the other items—held across both classes.

Differences across classes in factor loadings for sociability also varied between Versions 1 and 4. In Version 1, Items 10, 11, 12, and 14 had similar standardized factor loadings across classes (V1: "Drinking helps me to act sociable," "Drinking helps me talk to people," "Drinking helps me to be friendly," and "Drinking helps me to be outgoing," respectively). The other three items, Items 15, 16, and 17 (V1: "Drinking helps me to be humorous," "Drinking helps me express my feelings," and "Drinking helps me feel energetic," respectively), showed somewhat weaker loadings in Class 2 than Class 1. A different pattern of class differences in loadings emerged in Version 4. Similar to Version 1, loadings for Items 11 and 12 were close to invariant across classes, whereas the loading for Item 17 was considerably weaker in Class 2. However, Items 10, 15, and 16 actually showed higher loadings in Class 2 than

[5] To determine the optimal partial weak invariance model, item-by-item tests of loading noninvariance across classes were conducted, following the IRT-LR-DIF strategy (Thissen, 2001). In both versions, the resulting partial weak invariance model was still rejected relative to the configural invariance model by the LRT.

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

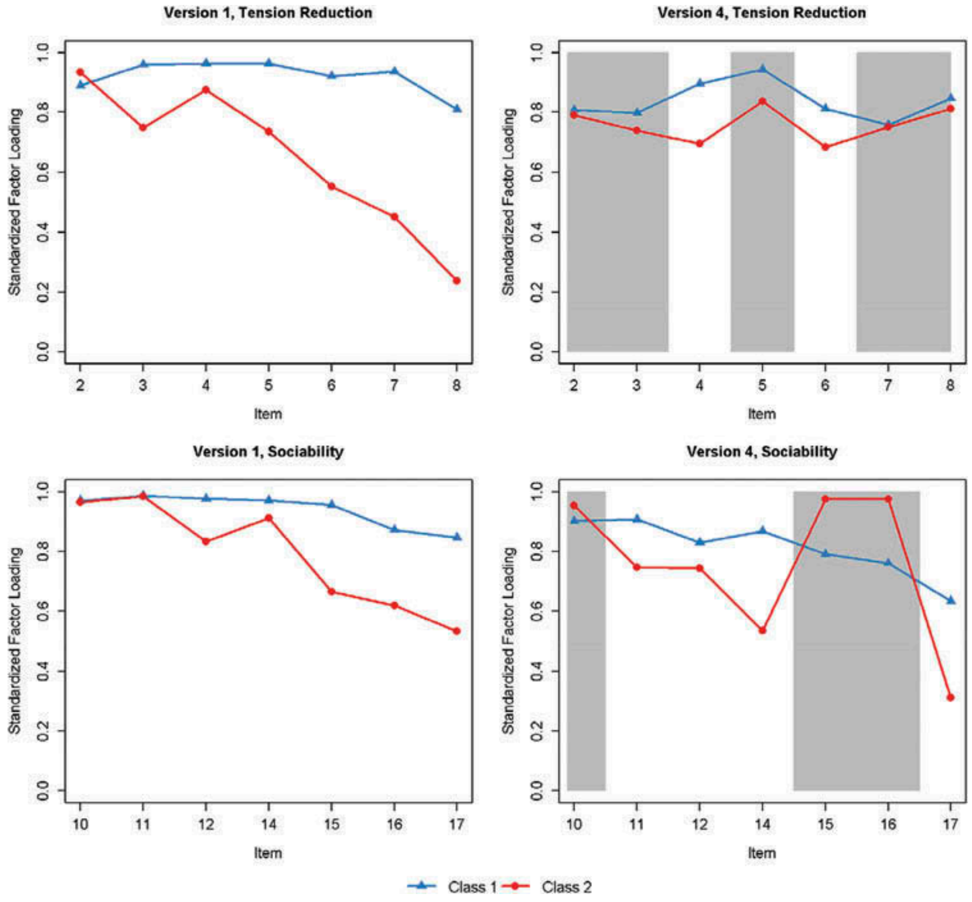

FIGURE 3 Study 2: Standardized factor loadings for all items in factor mixture model in Versions 1 and 4. *Note*. Items with a gray background in Version 4 were measured using the 5-point scale.

Class 1. As shown in Table 6 and via shading in Figure 3, these were the three (of four) items for the sociability factor that were originally measured using a different response scale (0 = *no chance* to 4 = *certain to happen*) than the other items (1 = *strongly disagree* to 4 = *strongly agree*). Thus, it might be the case that, in Version 4, differences between the classes in the measurement of sociability might have reflected a method factor corresponding to differences in response scales across subsets of items.

Thresholds. Figure 4 shows the thresholds for $y_{iq} = 1$ and $y_{iq} = 2$ across classes and versions; again, the background of the plot is shaded for items originally measured using the 5-point scale. In Version 4, a few thresholds were fixed at either positive or negative 15 in Class 2, with the threshold for $y_{iq} = 2$ for Items 5 and 7 fixed at 15, and those for $y_{iq} = 1$ for Items 11, 12, and 14 fixed at −15, for members of Class 2. This reflects a boundary condition in which the probability of endorsing $y_{iq} = 3$ on Items 5 and 7 was functionally zero, and endorsing either $y_{iq} = 2$ or $y_{iq} = 3$ for Items 11, 12, and 14 was functionally one.

In Version 1, both thresholds were consistently lower for members of Class 2 than Class 1, indicating members of Class 2 endorsed these items at higher levels. By contrast, in Version 4, differences between classes in thresholds occurred almost exclusively (with the exception of items with thresholds that were fixed at boundary values) in items measured using the 5-point response scale. In particular, thresholds for $y_{iq} = 1$ are lower in Class 2, and thresholds for $y_{iq} = 2$ are higher in Class 2, on all items that show a difference in Version 4. As such, in Version 4, collapsing the top two response categories in the 5-point response scale (*very likely* and *certain to happen*) might have decreased the portion of the sample endorsing $y_{iq} = 3$. Furthermore, class differences in thresholds for $y_{iq} = 1$ appeared somewhat more pronounced for items measuring the sociability factor, whereas class differences for $y_{iq} = 2$ were larger for items measuring the tension reduction factor. These differences, especially for the tension reduction items, were most pronounced on items originally measured using the 5-point response scale.

Expected Score Curves

To jointly consider loadings and thresholds, expected score curves are shown in Figures 5 and 6 for Versions 1 and 4,

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

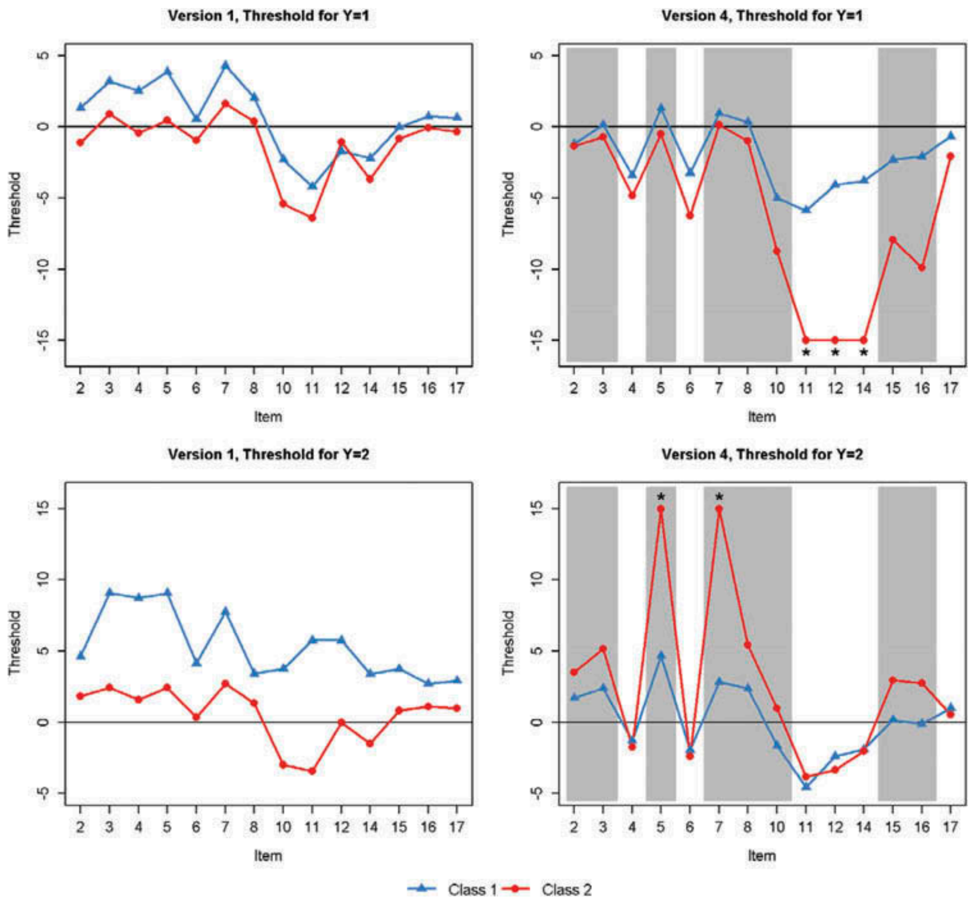

FIGURE 4 Study 2: Thresholds for all items for factor mixture models in Versions 1 and 4. *Note*. Items with a gray background in Version 4 were measured using the 5-point scale; asterisks indicate parameter fixed at boundary value.

respectively. For ordinal items, expected score curves weight each response option (here $j = 1, 2$, and 3) by their endorsement probabilities to obtain the expected value of y_{iq} at a range of values for the underlying latent trait (Hill et al., 2007). In Version 1, most items differed uniformly between classes such that members of Class 2 had lower expected scores than members of Class 1 across all values of the latent variable. However, for a few items (particularly Items 7, 8, 15, 16, and 17), the relationship between subjects' values of the latent factor and their expected score was considerably weaker for members of Class 2 than Class 1, corresponding to the lower loadings for these items in Class 2. In Version 4, items measured on the 4-point scale appeared to show a weaker relationship to the latent variable, as well as much higher expected scores across all levels of the latent variable, in Class 2; examining expected score curves shows that, for these items, endorsing a higher response category was extremely likely even at low levels of the latent variable. By contrast, items measured on the 5-point scale differed in a number of ways between Classes 1 and 2. For items measuring the tension reduction factor, members of Class 2 showed either a weaker relationship between the latent factor and the expected score (Items 2, 3, and 8), or truncation of the range of expected scores, with virtually no probability of responding $y_{kq} = 3$ (Items 5 and 7). For items measuring the sociability factor (Items 10, 15, and 16), although loadings and thresholds were generally different between Classes 1 and 2 (with generally higher loadings, lower first thresholds, and higher second thresholds in Class 2), these parameters nevertheless combined to produce relatively similar expected score curves; this is not an uncommon finding when comparing expected score curves between groups, particularly with ordinal items (Oshima, Kushubar, Scott, & Raju, 2009; Raju, Van der Linden, & Fleer, 1995).

Summary

Study 2 examined FMMs of alcohol expectancies under the two most disparate measurement versions in the study (Version 1 and Version 4). Models with two factors corresponding to tension reduction (Factor 1) and sociability (Factor 2) were considered, with varying numbers of classes and levels of measurement invariance across classes. In both measurement versions, fit statistics favored a two-class model imposing only configural invariance between classes. Class prevalence rates were

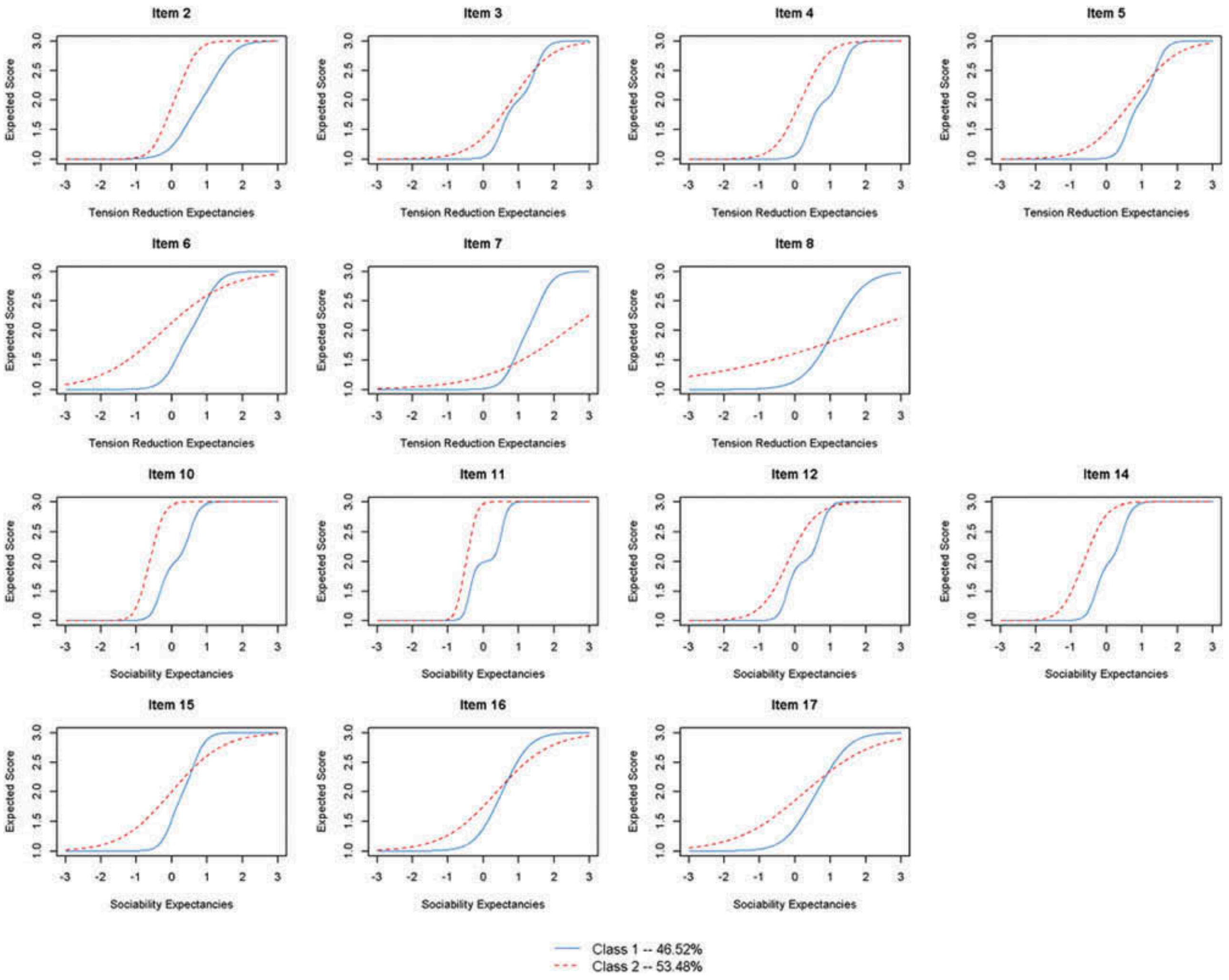

FIGURE 5 Study 2: Expected score curves in Version 1.

relatively similar across measurement versions. The two measurement versions, however, diverged greatly in the item parameter differences seen between the two classes. In Version 1, the two classes differed most in the tension reduction factor, with Class 2 showing considerably weaker loadings on a number of items than Class 1. By contrast, in Version 4 the two classes were most different in the sociability factor. Differences across classes in item parameters were strongest for those that had originally been measured using the 5-point response scale in Version 4; these items showed weaker loadings and higher thresholds in Class 2. This difference is of particular interest because it is not uncommon for item sets to include items measured using multiple response scales (e.g., when combining items from multiple scales within the same study or pooling data across multiple studies that use different response scales; Hussong, Curran, & Bauer, 2013).

DISCUSSION

This report examined the effects of differences in item wording and response scales on the nature of results obtained from mixture models. The nature of these effects was explored through two studies, which took advantage of an experimental design in which measurement was empirically manipulated. In Study 1, separate latent class analyses of binary alcohol use problem items were conducted across four measurement versions, which differed in terms of item stems, response categories, or both. In Study 2, FMMs of a two-factor alcohol expectancies scale were conducted on subjects from two different measurement versions.

Neither study found particularly strong differences between measurement versions in class enumeration. Although model fit indexes offered only equivocal support for either a three- or four-class solution in Study 1, the results from these model fit indexes did not differ reliably

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

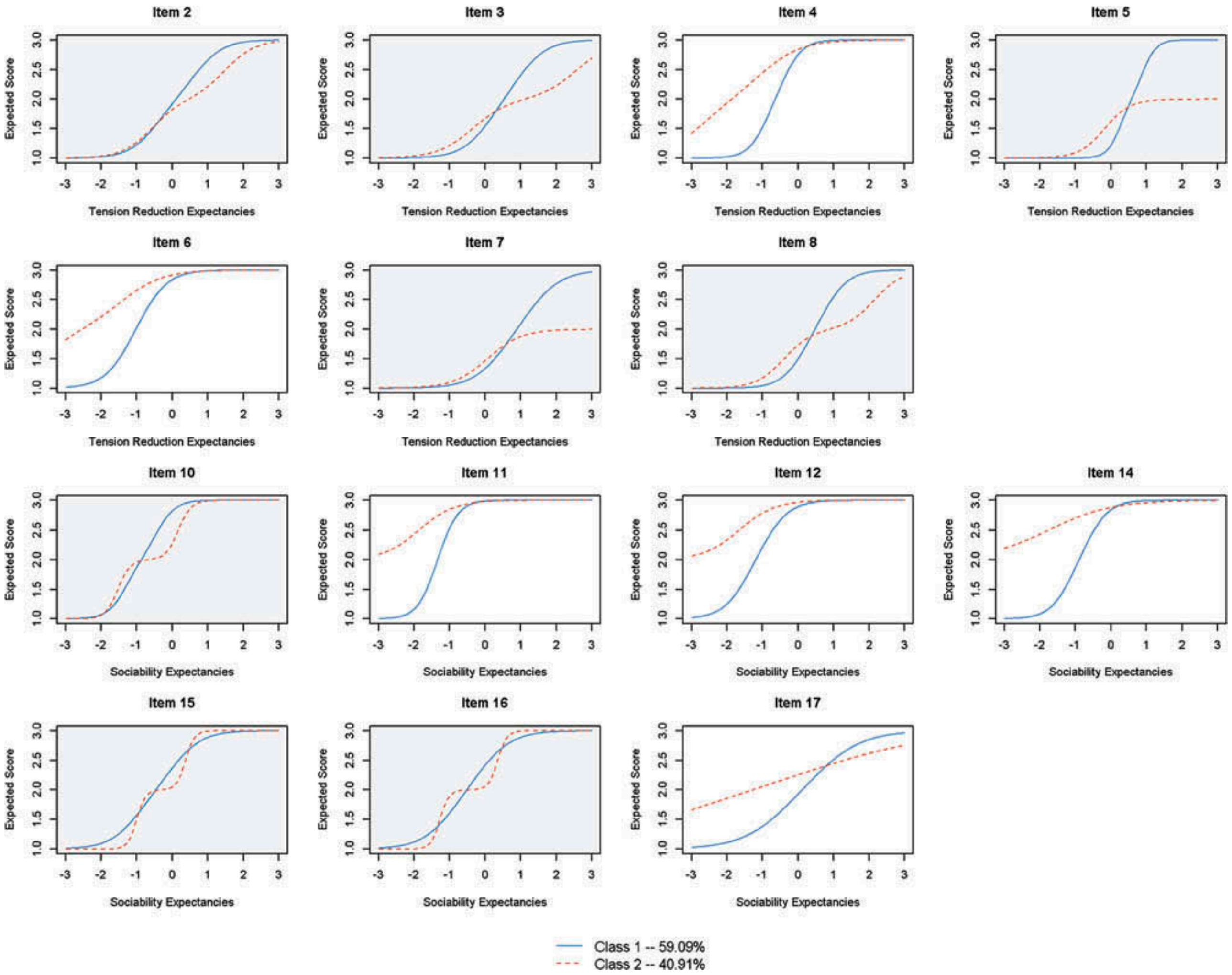

FIGURE 6 Study 2: Expected score curves in Version 4. *Note*. Items with a gray background in Version 4 were measured using the 5-point scale.

between measurement versions. In Study 2, the same number of classes, as well as the same factor structure of items (i.e., configural invariance across classes), was unequivocally favored by model fit indexes in both Versions 1 and 4. This is consistent with the results of K. M. Jackson and Sher (2005), who found that only more extreme differences in the operationalization of alcohol involvement resulted in a different number of classes being selected in a GMM.

However, where measurement differences did become relevant was in the overall configuration of the classes observed. In Study 1, although classes with low levels of alcohol problem endorsement and classes with high levels of alcohol problem endorsement were found in all measurement versions, differences in item characteristics primarily changed the shape of intermediate classes. These results suggest that differences across studies in the measurement of alcohol problems might manifest as substantively distinct findings, particularly with respect to item endorsement for intermediate classes, as well as the prevalence for each class.

Likewise, in Study 2, the two measurement versions showed differences in factor loadings and thresholds, which were somewhat stronger on items with response categories that differed across versions. Whereas loadings were most disparate across classes for the tension reduction factor in Version 1, differences in loadings were larger for the sociability factor in Version 4. The difference we observed across classes in likelihood of endorsing items at higher scale points based on differences in items' original response scale is consistent with the findings of K. M. Jackson and Sher (2008), that choosing different cut points for categorical measures changes the configuration of classes in latent class growth analysis. However, it is worth noting that, unlike these authors, we did not see differences in the prevalence of classes on the basis of different response categories.

Taken together, these findings demonstrate that the results of mixture models might change on the basis of decisions made in the measurement of the construct of interest. In this way, these results contextualize the frequent disagreement between studies

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

in the obtained number and nature of latent classes, suggesting that this disagreement might reflect differences across studies in measurement, as opposed to true differences across studies in the nature of the latent classes themselves. These discrepancies create a serious barrier to a cumulative understanding of a number of constructs in the behavioral sciences, including but not limited to the case of examining latent classes based on AUD criteria that we discussed at the outset. These barriers could potentially be surmounted by more careful consideration of the ways in which constructs are measured when conducting and interpreting mixture model results. Most concretely, researchers might be well advised to undertake sensitivity analyses to show that a given latent class structure is replicable across minor perturbations of measurement. Such alterations to measurement could be made either during data collection (e.g., administering different response scales or item stems to different subsets of participants) or after (e.g., fitting a mixture model on the same data multiple times, each time collapsing response options differently).

One critical limitation of this work is that, although measurement was manipulated experimentally, it is unknown whether and to what extent any of the latent class solutions obtained represents the truth, as data were not simulated to have any particular latent class structure. Despite this limitation, these findings suggest that variations in measurement must be considered in the interpretation of mixture model results, particularly when two sets of results differ. Although one, both, or neither of these sets of results might represent the true latent class structure of the construct under study, such a determination is not possible to make without accounting for the potentially biasing effects of differences in measurement.

FUNDING

This work was supported by National Institutes of Health grants F31 DA040334 (Fellow: Veronica T. Cole) and R01 DA034636 (Principal Investigator: Daniel J. Bauer). The content is solely the responsibility of the authors and does not represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

REFERENCES

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Parzen, E., Tanabe, K., Kitagawa, G.(Eds.) *Selected papers of Hirotugu Akaike* (pp. 199–213). New York, NY: Springer.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338–363. doi:10.1037/1082-989X.8.3.338

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3–29. doi:10.1037/1082-989X.9.1.3

Beseler, C. L., Taylor, L. A., Kraemer, D. T., & Leeman, R. F. (2012). A latent class analysis of *DSM–IV* alcohol use disorder criteria and binge drinking in undergraduates. *Alcoholism: Clinical and Experimental Research*, *36*, 153–161. doi:10.1111/acer.2011.36.issue-1

Bucholz, K. K., Cadoret, R., Cloninger, C., Dinwiddie, S. H., Hesselbrock, V. M., Nurnberger, J. I., … Schuckit, M. A. (1994). A new, semi-structured psychiatric interview for use in genetic linkage studies: A report on the reliability of the SSAGA. *Journal of Studies on Alcohol*, *55*, 149–158. doi:10.15288/jsa.1994.55.149

Chung, T., & Martin, C. S. (2001). Classification and course of alcohol problems among adolescents in addictions treatment programs. *Alcoholism: Clinical and Experimental Research*, *25*, 1734–1742. doi:10.1111/acer.2001.25.issue-12

Clark, S. L., Muthén, B., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Structural Equation Modeling*, *20*, 681–703. doi:10.1080/10705511.2013.824786

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*, 762–771. doi:10.1080/01621459.1984.10477093

Crow, S. J., Swanson, S. A., Peterson, C. B., Crosby, R. D., Wonderlich, S. A., & Mitchell, J. E. (2012). Latent class analysis of eating disorders: Relationship to mortality. *Journal of Abnormal Psychology*, *121*, 225–231. doi:10.1037/a0024455

Eggleston, E. P., Laub, J. H., & Sampson, R. J. (2004). Methodological sensitivities to latent class analysis of long-term criminal trajectories. *Journal of Quantitative Criminology*, *20*(1), 1–26. doi:10.1023/B:JOQC.0000016696.02763.ce

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631. doi:10.1198/016214502760047131

Fromme, K., Stroot, E. A., & Kaplan, D. (1993). Comprehensive effects of alcohol: Development and psychometric assessment of a new expectancy questionnaire. *Psychological Assessment*, *5*(1), 19–26. doi:10.1037/1040-3590.5.1.19

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2007). Practical issues in the application of item response theory: A demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Medical Care*, *45*, S39–S47. doi:10.1097/01.mlr.0000259879.05499.eb

Huang, G. H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, *69*, 5–32. doi:10.1007/BF02295837

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. doi:10.1007/BF01908075

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. doi:10.1146/annurev-clinpsy-050212-185522

Jackson, K. M., Bucholz, K. K., Wood, P. K., Steinley, D., Grant, J. D., & Sher, K. J. (2014). Towards the characterization and validation of alcohol use disorder subtypes: Integrating consumption and symptom data. *Psychological Medicine*, *44*(1), 143–159.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors*, *19*, 339–351. doi:10.1037/0893-164X.19.4.339

Jackson, K. M., & Sher, K. J. (2006). Comparison of longitudinal phenotypes based on number and timing of assessments: A systematic comparison of trajectory approaches II. *Psychology of Addictive Behaviors*, *20*, 373–384. doi:10.1037/0893-164X.20.4.373

Jackson, K. M., & Sher, K. J. (2008). Comparison of longitudinal phenotypes based on alternate heavy drinking cut scores: A systematic

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018

comparison of trajectory approaches III. *Psychology of Addictive Behaviors*, *22*, 198–209. doi:10.1037/0893-164X.22.2.198

Jackson, N., Denny, S., Sheridan, J., Fleming, T., Clark, T., Teevale, T., & Ameratunga, S. (2014). Predictors of drinking patterns in adolescence: A latent class analysis. *Drug and Alcohol Dependence*, *135*, 133–139. doi:10.1016/j.drugalcdep.2013.11.021

Kushner, M. G., Sher, K. J., Wood, M. D., & Wood, P. K. (1994). Anxiety and drinking behavior: Moderating effects of tension-reduction alcohol outcome expectancies. *Alcoholism: Clinical and Experimental Research*, *18*, 852–860. doi:10.1111/acer.1994.18.issue-4

La Flair, L. N., Bradshaw, C. P., Storr, C. L., Green, K. M., Alvanzo, A. A., & Crum, R. M. (2012). Intimate partner violence and patterns of alcohol abuse and dependence criteria among women: A latent class analysis. *Journal of Studies on Alcohol and Drugs*, *73*, 351–360. doi:10.15288/jsad.2012.73.351

La Flair, L. N., Reboussin, B. A., Storr, C. L., Letourneau, E., Green, K. M., Mojtabai, R., … Crum, R. M. (2013). Childhood abuse and neglect and transitions in stages of alcohol involvement among women: A latent transition analysis approach. *Drug and Alcohol Dependence*, *132*, 491–498. doi:10.1016/j.drugalcdep.2013.03.012

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767–778. doi:10.1093/biomet/88.3.767

Lubke, G. H., & Miller, P. J. (2015). Does nature have joints worth carving? A discussion of taxometrics, model-based clustering and latent variable mixture modeling. *Psychological Medicine*, *45*, 705–715. doi:10.1017/S003329171400169X

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21–39. doi:10.1037/1082-989X.10.1.21

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, *14*, 26–47. doi:10.1080/10705510709336735

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood?. *Multivariate Behavioral Research*, *41*(4), 499–532.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, *43*, 592–620. doi:10.1080/00273170802490673

Lynskey, M. T., Nelson, E. C., Neuman, R. J., Bucholz, K. K., Madden, P. A., Knopik, V. S., … Heath, A. C. (2005). Limitations of DSM–IV operationalizations of alcohol abuse and dependence in a sample of Australian twins. *Twin Research and Human Genetics*, *8*, 574–584. doi:10.1375/twin.8.6.574

Mancha, B. E., Hulbert, A., & Latimer, W. W. (2012). A latent class analysis of alcohol abuse and dependence symptoms among Puerto Rican youth. *Substance Use & Misuse*, *47*, 429–441. doi:10.3109/10826084.2011.643525

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, *50*, 266–275. doi:10.1037/0003-066X.50.4.266

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143. doi:10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. doi:10.1007/BF02294825

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., & Marlatt, G. A. (2002). Test–retest reliability of alcohol measures: Is there a difference between Internet-based assessment and traditional methods? *Psychology of Addictive Behaviors*, *16*(1), 56–63. doi:10.1037/0893-164X.16.1.56

Muthén, B. (2001). Latent variable mixture modeling. In Marcoulides, G. A. & Schumacker, R. E. (Eds.), *New Developments and Techniques in Structural Equation Modeling*. (pp. 1–33). Mahwah, NJ: Larwence Erlbaum.

Muthén, B. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction*, *101*(Suppl. 1), 6–16. doi:10.1111/j.1360-0443.2006.01583.x

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463–469. doi:10.1111/j.0006-341X.1999.00463.x

Muthén, L. K., & Muthén, B. O. (2015). *M*plus *user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, *4*, 139–157. doi:10.1037/1082-989X.4.2.139

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, *6*, 18–34. doi:10.1037/1082-989X.6.1.18

Neal, D. J., Corbin, W. R., & Fromme, K. (2006). Measurement of alcohol-related consequences among high school and college students: Application of item response models to the Rutgers Alcohol Problem Index. *Psychological Assessment*, *18*, 402–414. doi:10.1037/1040-3590.18.4.402

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569. doi:10.1080/10705510701575396

Oshima, T., Kushubar, S., Scott, J., & Raju, N. (2009). *DFIT8 for Windows user's manual: Differential functioning of items and tests*. St. Paul, MN: Assessment Systems Corporation.

Presley, C. A., Meilman, P. W., & Lyerla, R. (1994). Development of the core alcohol and drug survey: Initial findings and future directions. *Journal of American College Health*, *42*, 248–255. doi:10.1080/07448481.1994.9936356

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353–368. doi:10.1177/014662169501900405

Reboussin, B. A., Ip, E. H., & Wolfson, M. (2008). Locally dependent latent class models with covariates: An application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*, 877–897. doi:10.1111/rssa.2008.171.issue-4

Rinker, D. V., & Neighbors, C. (2015). Latent class analysis of DSM–5 alcohol use disorder criteria among heavy-drinking college students. *Journal of Substance Abuse Treatment*, *57*, 81–88. doi:10.1016/j.jsat.2015.05.006

Sampson, R. J., Laub, J. H., & Eggleston, E. P. (2004). On the robustness and validity of groups. *Journal of Quantitative Criminology*, *20*, 37–42. doi:10.1023/B:JOQC.0000016698.36239.91

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Steinley, D. (2004). Properties of the Hubert-Arable adjusted Rand index. *Psychological Methods*, *9*, 386. doi:10.1037/1082-989X.9.3.386

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. [Documentation for computer program]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In Hancock, G. R. (Ed.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.

Tsai, J., & Rosenheck, R. A. (2013). Conduct disorder behaviors, childhood family instability, and childhood abuse as predictors of severity of adult homelessness among American veterans. *Social Psychiatry and Psychiatric Epidemiology*, *48*, 477–486. doi:10.1007/s00127-012-0551-4

Tueller, S., & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, *17*(2), 165–192.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. doi:10.1177/109442810031002

Van Horn, M. L., Smith, J., Fagan, A. A., Jaki, T., Feaster, D. J., Masyn, K., … Howe, G. (2012). Not quite normal: Consequences of violating the assumption of normality in regression mixture models. *Structural Equation Modeling*, *19*, 227–249. doi:10.1080/10705511.2012.659622

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*, 307–333. doi:10.2307/1912557

Wells, J. E., Horwood, L. J., & Fergusson, D. M. (2004). Drinking patterns in mid-adolescence and psychosocial outcomes in late adolescence and early adulthood. *Addiction*, *99*, 1529–1541. doi:10.1111/add.2004.99.issue-12

White, H. R., & Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*, *50*(1), 30–37. doi:10.15288/jsa.1989.50.30

Downloaded by [University North Carolina - Chapel Hill] at 09:52 02 January 2018