2021 ■ Vol. 17 ■ no. 2

Binary item CFA of Behavior Problem Index (BPI) using Mplus: A step-by-step tutorial

Minjung Kim [a] ✉, Christa Winkler [a] & Susan Talley [a]

[a]Ohio State University

Abstract ■ In social science research, we often use a measurement with items answered in two ways: yes or no, pass or fail. The goal of this paper is to provide explicit guidance for substantive researchers interested in using the binary confirmatory factor analysis (CFA) through Mplus to validate the psychometric properties of a measurement with dichotomous items. Although the implementation of binary CFA using Mplus is simple and straightforward, interpreting the results is not as simple as implementation, especially for those who have limited experiences or knowledge in the Item Response Theory (IRT). We use an illustrative example from the National Longitudinal Survey of Youth (NLSY) to demonstrate the implementation of binary CFA of the Behavior Problem Index (BPI) using Mplus and the interpretation of the analysis results.

Keywords ■ binary confirmatory factor analysis, binary CFA. **Tools** ■ Mplus.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

✉ kim.7144@osu.edu

doi 10.20982/tqmp.17.2.p141

Introduction

Confirmatory factor analysis (CFA) is a method for testing the relationships among observed items of a measure based on a theory (Brown, 2006). Traditional linear approaches to CFA assume that the responses to the items are continuous and normally distributed. However, we often encounter measures with binary responses, such as yes/no and pass/fail, in practice. When the response to a questionnaire is binary, not only is the assumption of normal distribution underlying traditional CFA approaches violated, but the interpretation of linear coefficients also needs to be adjusted; consequently, an alternative approach needs to be employed.

Implementation of binary CFA using Mplus (Muthén & Muthén, 2017) is more accessible than alternative approaches and offers more adequate option for estimating parameters with dichotomous data (Brown, 2006; Wang & Wang, 2012). As Brown (2006) notes, the framework and procedures for the binary CFA models "differ considerably" from traditional CFA approaches (Brown, 2006, p. 389). In particular, estimation of model parameters results in a threshold structure akin to that of item difficulty parameters in item response theory (IRT; Kim & Yoon, 2011). We have found that although the procedure for applying the binary CFA is simple and straightforward using Mplus, interpreting the generated output of the analysis results is not as simple as the implementation, especially for the substantive researchers who have no previous knowledge in IRT or logistic regression.

The goal of this paper is to provide explicit guidance for substantive researchers who are interested in using binary CFA through Mplus when they have no prior knowledge of IRT. We use an illustrative example from the National Longitudinal Survey of Youth (NLSY) to demonstrate the implementation of binary CFA using Mplus and interpretation of the analysis results.

Background

Binary CFA is widely employed in education, health, and social science research to validate constructs when the response options of the measured items are dichotomous (e.g., yes/no, pass/fail, diagnosed/not-diagnosed, etc.). For instance, Gonzales et al. (2017) utilized the Index of Learning Styles (ILS) measure in their study of learning behaviors of nursing students. Because responses to the ILS are binary (1 for responses in the "sensible" category and 0 for the responses in the "imaginative" category), both binary EFA and CFA were used to assess the data. More recently, a study on the political discussion networks in Britain used CFA to assess their battery of questions (Galandini & Fieldhouse, 2019). This included measures with binary responses, such as measures of mobilization value (yes/no for questions such as "As far as you know, did each of these people vote in the recent European Elections") and co-ethnicity of discussants (1 for same ethnic group as discussant, 0 for not the same ethnic group). There are other studies that use the data with continuous scale in the original form, which is converted into a binary scale given the data distribution. For example, Fernandez, Vargasm, Mahometa, Ramamurthy, and Boyle (2012) conducted a binary CFA to validate the factor structure of 24 items assessing the pain descriptor system. Although the original items were scaled to be 0 to 10, they were dichotomized to be 0 for no pain and 1 for pain since the data showed a zero-inflated distribution.

When it comes to conducting CFA with binary data, Mplus has been recommended as a convenient and ideal software (e.g. Brown, 2006; Wang & Wang, 2012). Mplus (Muthén & Muthén, 2017) is a powerful statistical software that is capable of analyzing latent variable models and other complex models. One of the reasons that Mplus is recommended for conducting the binary CFA is in part because Mplus allows users to easily switch the method of estimation from maximum likelihood (ML) to a more adequate option for non-normal items, such as, weighted least square mean- and variance-adjusted estimation method (WLSMV). Although ML is adopted as a default estimation method in general for structural equation modeling (SEM) software including Mplus, it becomes problematic when applied to binary data because the normality assumption is violated with only two discrete response options. According to Brown (2006), treating categorical or binary observed variables as continuous items can result in the attenuated estimates of the relationships (i.e., correlations) among indicators, the false identification of "pseudofactors" that are merely byproducts of item difficulty or extreme response styles, and significant bias in test statistics, standard errors, and subsequent inferences (Brown, 2006, p. 387).

In fact, Bengt Muthén of Mplus developed the approach for dichotomous and ordinal data (Muthén, 1984), and it has subsequently been referred to previously as the "golden standard" in the SEM literature (Rosseel, 2014). WLSMV is commonly identified as the optimal method for estimating models with binary and/or categorical indicators, as it provides accurate parameter estimates without requiring large sample sizes as opposed to the alternative options such as traditional weighted least squares (WLS) estimation (Beauducel & Herzberg, 2006; Flora & Curran, 2004). Mplus uses the probit link function for binary CFA, which inversely models the standard normal distribution of the probability (Wang & Wang, 2012).

Implementing the binary CFA by switching the ML to WLSMV is extremely easy in Mplus because it just requires an additional statement of `CATEGORICAL ARE [variable names]` in the input command. Compared to its simplicity for implementing the procedure, interpreting the results requires more knowledge and experience in categorical data analysis, which can be challenging for substantive researchers or graduate students who are more accustomed to traditional CFAs with continuous data. For example, when working with a continuous indicator, a factor loading of 0.75 indicates that a 0.75 unit change in the item corresponds with every 1 unit increase in the factor score. If that same interpretation is applied to a CFA model with a binary indicator, it often results in a meaningless statement because the possible response options are 0 or 1 and it cannot go over 1 or below 0; two unit increases in latent score cannot be interpreted as 1.5 (i.e., .75*2) increase in the item.

Others have noted that conducting binary CFA models in Mplus allows for "an integrated application" of both IRT and SEM models "in a unified, general latent variable modeling framework" (Glockner-Rist & Hoijtink, 2003, p. 545). That being said, novice users who have no previous knowledge in IRT may seek guidance on how to interpret the associated model parameters, which are more closely aligned with IRT than traditional CFA.

Additionally, the continually evolving capabilities of Mplus software have resulted in the availability of new options and output for users, such as, the removal of an "experimental" weighted root mean square residual (WRMR) index, addition of the standardized root mean square residual (SRMR) fit index, and the addition of standardized output options. Many of the more comprehensive resources (e.g. Brown, 2006) predate those options, leaving users without clear guidance; conversely, our paper provides guidance that reflects all of the most current capabilities available in Mplus software (version 8) using an example dataset.

Illustrative Example for binary CFA using Mplus: NLSY79

Data

The National Longitudinal Surveys (NLS) are a set of surveys sponsored by the Bureau of Labor Statistics (BLS) of the U.S. Department of Labor. Included in that set is the National Longitudinal Survey of Youth (NLSY), which fol-

Table 1 ■ BPI subscales and items

Subscale	Factor Names in Mplus Syntax	Items	Item Names in Mplus Syntax
Anxious/ Depressed	AnxDep	Has sudden changes in mood or feeling	ad1
		Feels/complains no one loves him/her	ad2
		Is too fearful or anxious	ad3
		Feels worthless or inferior	ad4
		Is unhappy, sad, or depressed	ad5
Headstrong	Headstr	Is rather high strung, tense, and nervous	hs1
		Argues too much	hs2
		Is disobedient at home	hs3
		Is stubborn, sullen, or irritable	hs4
		Has strong temper and loses it easily	hs5
Antisocial	Antisoc	Cheats or tells lies	as1
		Bullies or is cruel/mean to others	as2
		Does not seem to feel sorry after misbehaving	as3
		Breaks things deliberately	as4
		Is disobedient at school	as5
		Has trouble getting along with teachers	as6
Hyperactive	Hyperac	Has difficulty concentrating/paying attention	hy1
		Is easily confused, seems in a fog	hy2
		Is impulsive or acts without thinking	hy3
		Has trouble getting mind off certain thoughts	hy4
		Is restless, overly active, cannot sit still	hy5
Peer Problems	PeerProb	Has trouble getting along with other children	pp1
		Is not liked by other children	pp2
		Is withdrawn, does not get involved with others	pp3
Dependent	Depend	Clings to adults	de1
		Cries too much	de2
		Demands a lot of attention	de3
		Is too dependent on others	de4

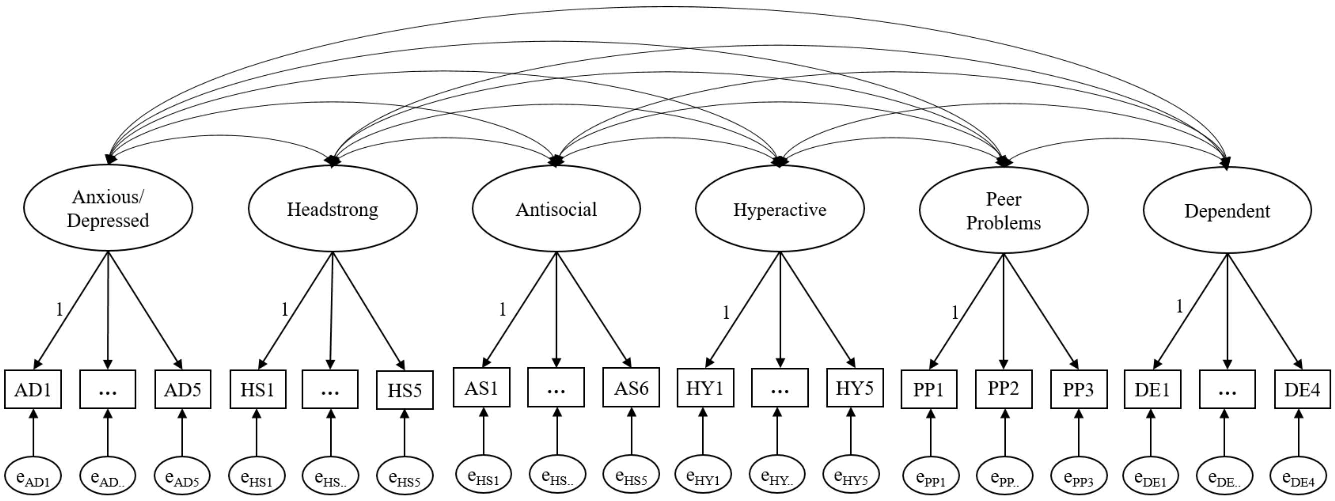
lows a sample of American youth, assessing a broad array of topics including education, training, employment, family relationships, income, health, and general attitudes. As of 2014, a total of 26 interview rounds were completed with a total of 12,686 respondents. Data from all 26 rounds of the NLSY are available publicly and accessible through the NLS Investigator (www.nlsinfo.org/investigator). Data for the present study used the 2012 administration of the NLSY79 ($N = 491$).

Measure

The Behavior Problems Index (BPI; Peterson & Zill, 1986) is comprised of 28 items intended to measure type, frequency, and range of specific behavioral problems in children ages four and older (National Longitudinal Surveys, n.d.). The 28 items are clustered on 6 subscales, each representing problem behavior commonly displaying by children (Table 1). Those subdomains include antisocial behavior, anxiousness/depression, headstrongness, hyperactivity, immature dependency, and peer conflict/social withdrawal (National Longitudinal Surveys, n.d.). Figure 1 presents the SEM diagram of CFA model for the BPI measure with 6 subscales loaded on 28 items. Each ellipse represents the latent construct (e.g., Anxious/Depressed, Headstrong, Antisocial, etc.) that are measured by 28 items, represented in rectangles. The directional arrow ($\rightarrow$) represents the factor loadings and the bidirectional arrow ($\leftrightarrow$) represents the correlation between the latent factors. As shown in the diagram, all latent factors are correlated while the errors of the observed items are uncorrelated across all items, which is the default setting of Mplus. For the measurement model identification, Mplus constrains the first loading for each factor to equal 1 by default (see Figure 1).

For the purposes of the NLSY79, the BPI was administered to mothers with children between the ages four and

Figure 1 ■ SEM diagram of CFA model for the BPI measure.



fourteen. Mothers were asked to recall whether it was (1) "often true," (2) "sometimes true," or (3) "not true" that their child exhibited the target behaviors in the previous three months. Dichotomized recoding of the original items (0 = behavior not reported; 1 = behavior reported sometimes or often) was then used to compute and report the subscores.

Preparing Input for Binary CFA

Preparing Mplus input to conduct CFA is quite similar with the one for the regular CFA with continuous scaled items. As shown in Listing 1, there are a few of simple modifications to the Mplus input file to conduct the binary CFA. Those modifications include: (1) use of raw data input file, (2) identification of categorical variables in VARIABLE command, (3) application of the WLSMV estimation method in ANALYSIS command, and (4) inclusion of alternative difference test for model fit (Table 2).

First, it is required that the input data file referenced in the DATA command be a raw dataset (see Listing 1). Traditional CFA models can be conducted using either raw data or a variance-covariance matrix (S) as the input file. However, binary CFA models do not use a standard variance-covariance matrix computed from the raw observed variables. Instead, estimated correlations of the underlying latent trait ($y*$) are used to estimate binary CFA models. Thus, complete raw data must be available and used as the input file when conducting binary CFA in Mplus.

Second, the most crucial change to the Mplus input file pertains to the `VARIABLE` command. All the variables in the dataset are listed after a `names are` statement and variables included in the model are listed after `usevariables are` option. A similar process is used to identify categorical (or in this case, binary) variables in the model. Those variables are identified using the `categorical are` option (see Listing 1, `VARIABLE` command). This statement must be included in the input code for Mplus to identify the binary variables and it automatically changes the ML estimation method to WLSMV. Another option to change the estimation method is by identifying the estimator using `estimator is` command under `ANALYSIS` (see Listing 1 `ANALYSIS` command).

Finally, for model fit comparison purposes, an optional `SAVEDATA` command can be added to the input. Within that SAVEDATA command, a statement is included to export data necessary to conduct an appropriate difference test (in lieu of the standard chi-square difference test, which should not be used with binary CFA models). This can be used only for comparison of nested models, meaning that one model is a more constrained model of the other. Additional details regarding use of the DIFFTEST option for model comparison in binary CFA is addressed below.

Interpretation of Output for Binary CFA

In this paper, we focus on the five primary portions of Mplus output to interpret the results of binary CFA: univariate descriptives, model fit indices, thresholds, factor loadings, and r-square values, following the order of the sections printed in the Mplus output.

Univariate descriptives. This section contains the distribution statistics for the binary variables. Listing 2 shows

Listing 1 ■ Mplus input file for binary CFA model

```
TITLE: CFA Binary NLSY BPI 2012;
DATA: file is NLSY79 BPI 2012.dat;
      format is f7.0 f5.0 32f2.0;
VARIABLE:
  names are child parent ad1 ad2 hs1 as1 ad3 hs2 hy1 hy2 as2 hs3 as3
            pp1 hy3 ad4 pp2 hy4 hy5 hs4 hs5 ad5 pp3 as4 de1 de2 de3
            de4 na1 na2 na3 na4 as5 as6;
  usevariables are ad1 ad2 hs1 as1 ad3 hs2 hy1 hy2 as2 hs3 as3 pp1
                   hy3 ad4 pp2 hy4 hy5 hs4 hs5 ad5 pp3 as4 de1 de2
                   de3 de4 as5 as6;
  categorical are ad1 ad2 hs1 as1 ad3 hs2 hy1 hy2 as2 hs3 as3 pp1
                  hy3 ad4 pp2 hy4 hy5 hs4 hs5 ad5 pp3 as4 de1 de2
                  de3 de4 as5 as6;
  missing are blank;
ANALYSIS: estimator is WLSMV;
MODEL: AnxDep   by ad1 ad2 ad3 ad4 ad5;
       Headstr  by hs1 hs2 hs3 hs4 hs5;
       Antisoc  by as1 as2 as3 as4 as5 as6;
       Hyperac  by hy1 hy2 hy3 hy4 hy5;
       PeerProb by pp1 pp2 pp3;
       Depend   by de1 de2 de3 de4;
OUTPUT: standardized;
SAVEDATA: difftest is DiffTestData.dat;
```

the partial output for the first four variables in the dataset[1] following the order of the variables in `names are` statement. The categories marked as "Category 1" and "Category 2" are in the same order as they are coded. For instance, the example dataset coded 0 for "no" and 1 for "yes," so "Category 1" would be the values marked "no" or 0. In Listing 2, for AD1 (Anxious/Depressed Question 1), there were 229 responses for "no" ("Category 1") and 262 "yes" responses, which equate to proportions of .466 and .534 of the valid responses. However, the Univariate Proportions and Counts section does not take into account any missing responses. In the example, AS1 (Antisocial Question 1) contains 359 responses to "no" and 129 responses for "yes." This leaves three missing responses which are not listed in this section. Likewise, the proportions are reflective of this, listing 0.736 and 0.264 as the valid proportions as opposed to .731 and .263, the proportions if the missing values were included (.006).

Model fit assessment. In traditional CFA models, numerous indices are available to evaluate model fit. Commonly used indices include the chi-square (χ^2) test of model fit, root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and standardized root mean square residual (SRMR). Scholars have suggested standard benchmarks for these indices to serve as indicators of good model fit, including a nonsignificant χ^2 test, RMSEA less than 0.050, CFI/TLI greater than 0.950, and SRMR less than 0.080 (Hu & Bentler, 1999). As is customary, RMSEA, SRMR, CFI, and TLI indices do not have tests of statistical significance attached; thus, it cannot be determined based on those indices whether a better fitting model provides a statistically significant better fit. These same standard model fit indices, RMSEA, CFI/TLI, and SRMR are currently reported when conducting binary CFA in Mplus. In earlier versions of Mplus (up to version 7), the weighted root mean square residual (WRMR) was reported instead of the SRMR when WLS family of estimation is employed. Given that the WRMR has shown to perform poorly for situations with large sample sizes (DiStefano, Liu, Jiang, & Shi, 2018, 3), it is now replaced by SRMR from version 8.

Listing 3 presents the partial output of Mplus for the model fit information. Relying on established benchmarks (e.g. Hu & Bentler, 1999), the model indices suggest overall good fit of the model based on RMSEA (0.039 < 0.050), CFI, (0.964 > 0.950), and TLI (0.959 > 0.950). While the SRMR fell

[1] To simplify the presentation, we only show the partial (first four variables) output in Listing 2, while all 28 items are printed in the original output. The full Mplus output is available on the online repository.

Listing 2 ■ Selected Mplus output: Univariate proportions and counts for categorical variables

```
UNIVARIATE PROPORTIONS AND COUNTS FOR CATEGORICAL VARIABLES
    AD1
      Category 1    0.466         229.000
      Category 2    0.534         262.000
    AD2
      Category 1    0.821         403.000
      Category 2    0.179          88.000
    HS1
      Category 1    0.790         388.000
      Category 2    0.210         103.000
    AS1
      Category 1    0.736         359.000
      Category 2    0.264         129.000
                        ...
```

Listing 3 ■ Selected Mplus output: Model fit indices

```
MODEL FIT INFORMATION

Number of Free Parameters                      71

Chi-Square Test of Model Fit

          Value                           580.114*
          Degrees of Freedom                  335
          P-Value                          0.0000
*   The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used
    for chi-square difference testing in the regular way.  MLM, MLR and WLSM
    chi-square difference testing is described on the Mplus website.  MLMV, WLSMV,
    and ULSMV difference testing is done using the DIFFTEST option.

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                          0.039
          90 Percent C.I.                   0.033  0.044
          Probability RMSEA <= .05          1.000

CFI/TLI
          CFI                               0.964
          TLI                               0.959

Chi-Square Test of Model Fit for the Baseline Model
          Value                          7129.371
          Degrees of Freedom                  378
          P-Value                          0.0000

SRMR (Standardized Root Mean Square Residual)
          Value                             0.082

Optimum Function Value for Weighted Least-Squares Estimator
          Value                    0.48812935D+00
```

Table 2 ■ Summary of modifications to Mplus syntax for binary CFA

	Modification	Description	Required?
Data input	Add `File is...` option to `DATA:` command	Utilizes raw data input file (as opposed to variance/covariance matrix)	Required
Variable specification	Add `Categorical are...` option to `VARIABLE:` command	Identifies binary variables in the dataset	Required
Estimator specification	Add `Estimator = WLSMV;` to `ANALYSIS:` command	Applies the WLSMV estimator	Optional, as this is the Mplus default with specification of categorical variables
Model comparison	Apply `DIFFTEST is...` option to the `ANALYSIS:` command (and conduct corresponding process with nested model for comparison)	Conducts a robust chi-square difference test (that is appropriate for WLSMV estimation)	Optional, if interested in statistical test of model fit for nested models

slightly outside the benchmark of good fit (0.082 > 0.080), it was still within an acceptable range. A summary of all reported model fit indices and their corresponding interpretations are presented in Table 3.

Users who want to compare different models traditionally turn to the chi-square difference test as a test of absolute model fit. However, as noted in the Mplus output (Listing 3), the chi-square values in the output cannot be directly used for the difference test. Instead, when using WLSMV estimation, users must utilize the DIFFTEST option, which conducts an alternative robust chi-square difference test (Asparouhov & Muthén, 2006). The DIFFTEST is conducted via a two-step process and can be applied to nested models only (Asparouhov & Muthén, 2018). This process will be demonstrated with two sample models: (1) the model of interest with the BPI items loading onto 6 distinct factors, and (2) an alternative model with the BPI items loading onto only 1 factor.

In the first step, the least restrictive model (i.e., with the most parameters) is fit first. In this case, that is Model 1 where the BPI items are loaded onto 6 latent factors. The SAVEDATA command is added to the input file to read,

```
SAVEDATA: DIFFTEST is filename.dat;
```

as shown in Listing 1. In the second step, the most restrictive model (i.e., with fewer parameters) is fit. In this case, that is Model 2 where the BPI items are loaded onto only 1 latent factor. Listing For this model, the DIFFTEST option is added to the ANALYSIS command to read,

```
ANALYSIS: DIFFTEST is filename.dat;.
```

This statement should reference the same data file produced from the first step. Once both models are fit, the DIFFTEST results are output to the model fit section in the second step for interpretation. If the chi-square test for difference testing output in the second step produces a significant result, then it can be determined that the restrictions imposed by Model 2 significantly worsen the fit, as compared to Model 1. Such is the case with this dataset ($\chi^2(15) = 138.79, p < .001$), leading us to determine that Model 1 (with the BPI items loaded onto 6 latent factors) is a significantly better fitting model.

Model results. Following the model fit indices, estimated model parameters are provided in the Mplus output. Listings 4 and 5 present the partial output for the unstandardized and standardized parameters, respectively.

Unstandardized parameters. Unstandardized solutions using the default setting of Mplus quantify the relationship between indicators and latent constructs via the raw metric of the marker variable, which is the first indicator of each latent factor. When WLSMV is employed, parameter estimates produced from binary models are unlike those most users are accustomed to in standard CFA. In Mplus, binary CFA models rely on probit links, which are similar to logit links, to estimate the probability of a discrete outcome. Thus, factor loadings no longer reflect standard CFA interpretations; instead, the unstandardized factor loadings are essentially probit regression coefficients, where each loading reflects the slope coefficient obtained when the observed binary indicator is regressed on the underlying continuous latent variable (Wang & Wang, 2012). The probit coefficients of the factor loadings represent the change in the z-score underlying the latent value of the items (y*s; Figure 2) for every unit change in the latent factor. The probit model uses the standard normal cumulative distribution to identify what a particular z-value

Table 3 ■ Summary of model fit indices and corresponding interpretations

Model Fit Index	Estimate in Mplus Output	Benchmark for Good Model Fit (Hu & Bentler, 1999)	Interpretation
Chi-square (χ^2) test of model fit	$\chi^2(335) = 580.114$, $p < .001$	$p > .05$	N/A for binary CFA models estimated with WLSMV. Instead, robust χ^2 test produced from DIFFTEST should be interpreted.
Root mean square error of approximation (RMSEA) [95% confidence interval]	0.039 [0.033, 0.044]	0.00 to 0.05	RMSEA < 0.050 indicates good model fit.
Comparative fit index (CFI)	0.964	0.95 to 1.00	CFI > 0.950 indicates good model fit.
Tucker-Lewis index (TLI)	0.959	0.95 to 1.00	TLI > 0.950 indicates good model fit.
Standardized root mean square residual (SRMR)	0.082	0.00 to 0.08	SRMR > 0.080 indicates fit not ideal, though only deviates slightly from benchmark of good fit.

(in probit units) translates to as a probability. Since the marker variable approach (i.e., fixing the factor loading to be 1 for the first indicator) is used by default, the first factor loading for AnxDep factor (AD1) is 1.000 with the standard error of 0.000, indicating that the parameter is fixed, but not estimated. On the other hand, the estimated probit coefficient of 0.975 for the second item (AD2) variable on the AnxDep factor in Listing 4 can be interpreted as for a one unit increase in the true (latent) anxiety/depression measure, the z-score for the latent score of AD2 increases by 0.975.

Standardized parameters. In applied research, the standardized solutions are most often reported for traditional CFA models given their consistent interpretation across different datasets (Brown, 2006). Standardized solutions quantify the relationships using the unstandardized observed indicators and the standardized latent variable of interest (Brown, 2006). With adding the command, `OUTPUT: standardized;`, Mplus provides multiple options for standardization, including STDYX, STDY, and STD depending on the standardization of the exogenous (X) and endogenous (Y) variables (Muthén & Muthén, 2017). For binary CFA models, however, all three sets of standardized options generate the identical results because all indicators are considered to be endogenous variables loaded on the latent factor. Listing 5 presents the partial output of Mplus for the standardized factor loadings. The resulting standardized factor loading can be squared to yield the proportion of variance in y* that is explained by the latent factor.

Thresholds. One of the most noticeable differences in the binary CFA output is the inclusion of thresholds in the model results. Thresholds connect the binary indicators with their underlying continuous latent variable; specifically, they separate adjacent response categories by designating the point on the underlying latent construct at which respondents are likely to transition from responding with a 0 to responding with a 1 on a particular item (Wang & Wang, 2012; Brown, 2006). Thresholds are essentially the z-scores related to the probability estimates of responses to a particular item (Finney & DiStefano, 2013) and thus can be either positive or negative. They can be translated to probabilities of endorsing a particular item by consulting a z-table for the standard normal distribution. For users familiar with IRT frameworks, thresholds are akin to item difficulty parameters (Curran, Edwards, Wirth, Hussong, & Chassin, 2007).

There are 28 total thresholds in the complete Mplus output produced in this example, with one threshold for each binary item. The first threshold shown in the output (Listings 4 and 5), AD1$1 has a value of -0.084. This threshold of -0.084 is a z-value on the underlying latent trait (y*), and thus indicates that a respondent is likely to transition from an answer of “No” (0) to an answer of “Yes” (1) on item AD1 once their underlying latent anxiety/depression score (y*) exceeds that value. Using the standard normal distribution, this z-score can be translated to a probability. Using a z-table, one can determine that a threshold of

Listing 4 ■ Partial Mplus output for the unstandardized model results

```
MODEL RESULTS
                                                          Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 ANXDEP   BY
    AD1                1.000      0.000    999.000    999.000
    AD2                0.975      0.087     11.272      0.000
    AD3                0.992      0.079     12.568      0.000
    AD4                1.014      0.087     11.617      0.000
    AD5                1.196      0.084     14.286      0.000
                                    ...
Thresholds
    AD1$1             -0.084      0.057     -1.489      0.136
    AD2$1              0.918      0.066     13.884      0.000
                                    ...
```

Listing 5 ■ Partial Mplus output for the standardized model results

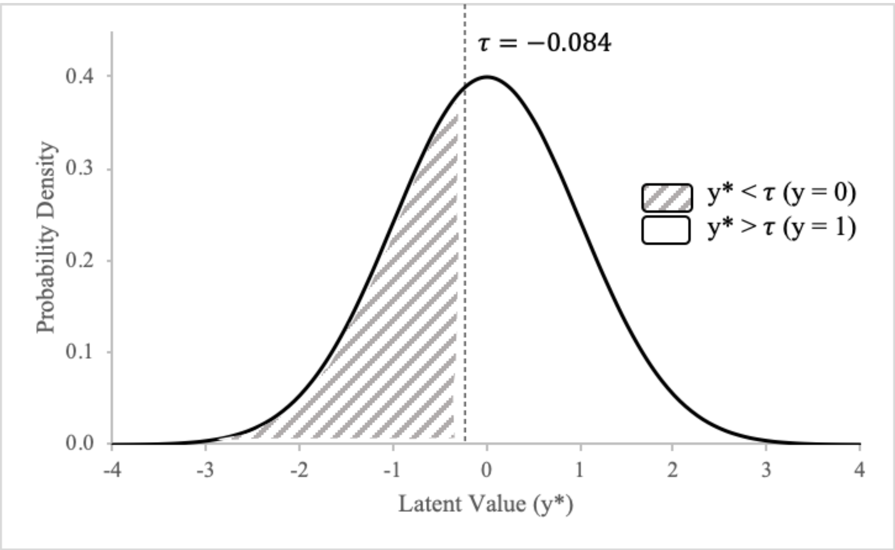
```
STANDARDIZED MODEL RESULTS

STDYX Standardization
                                                          Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value
 ANXDEP   BY
    AD1                0.738      0.040     18.395      0.000
    AD2                0.719      0.050     14.470      0.000
    AD3                0.732      0.045     16.101      0.000
    AD4                0.748      0.050     14.988      0.000
    AD5                0.883      0.037     23.780      0.000
                                    ...
Thresholds
    AD1$1             -0.084      0.057     -1.489      0.136
    AD2$1              0.918      0.066     13.884      0.000
                                    ...
```

-0.084 translates to a probability of 0.467; in other words, there is a 46.7% probability that AD1 = 0 and thus a 53.3% probability that AD1 = 1. The threshold structure for item AD1 in relation to the underlying latent variable is illustrated in Figure 2. Similarly, the next threshold, AD2$2, indicates that a respondent is likely to transition from an answer of “No” (0) to an answer of “Yes” (1) on item AD2 once their latent anxiety/depression score (y*) exceeds the threshold value of 0.918. That z-value also translates to a 0.821, or 82.1%, probability that AD2 = 0 and thus a 17.9% probability that AD2 = 1. Since thresholds are the z-values on the latent trait, they are identical regardless of the standardization (see Listings 4 and 5).

Model results. R-square. Finally, R-square is also given in the Mplus output (Listing **??**), which represents the proportion of variance in each of the observed measures that is explained by the latent factor. The R-square is computed by squaring the standardized factor loadings. For example, the standardized factor loading of item AD1 from Listing 5 is squared to be .545 (AD1 = $.738^2 = .544$), indicating that 54.4% of the latent value of AD1 item is explained by the shared variance among items through latent factor, AnxDep (Brown, 2006). Accordingly, the R-square estimate is always a value between 0 and 1, with 0 indicating that 0% of the variance in the observed measure is explained by the latent factor and 1 indicating that 100% of the variance in the observed measure is explained by the latent factor. However, this interpretation does not hold for cases of binary CFA. Rather, in binary CFA, the R-square value is computed from the underlying latent variable ($y*$) as op-

Figure 2 ■ Visual depiction of threshold (τ) for item AD1 on the latent response distribution. Based on the standard normal distribution of the underlying continuous latent trait ($y*$), $y* < \tau$ ($y = 0$): Where a respondent's value on the underlying continuous latent trait ($y*$) is less than the threshold (τ), their response on the binary observed variable (y) is expected to be 0. $y* > \tau$ ($y = 1$): Where a respondent's value on the underlying continuous latent trait ($y*$) is greater than the threshold (τ), their response on the binary observed variable (y) is expected to be 1.



posed to the observed variable, resulting in the proportion of variance in $y*$ (not y) explained by the latent factor. Thus, the R-square estimate of 0.544 for AD1 in this example (Figure 2) indicates that 54.4% of the variance in the underlying continuous latent variable (y*) is explained by the latent factor. Due to this change in interpretation, some authors suggest relying more heavily on interpreting model coefficients than R-square values for binary CFA models (UT-Austin, 2012).

Interpreting the results using probability. According to the Mplus tutorial by UT-Austin (2012), the R-square value combined with the corresponding threshold and the standardized factor loading coefficient can generate more meaningful information, which is the expected probability of case having a value of 0 or 1. The conditional probability of a $y = 0$ response given the factor η can be written as:

$$\begin{aligned} &P(y_{ij} = 0 \mid \text{factor\_value}) = \\ &\quad F\Big((\text{threshold} - \text{factor\_loading} * \text{factor\_value})\times \\ &\qquad 1/\sqrt{1 - \text{R\_square}}\Big), \end{aligned}$$

where F is the cumulative normal distribution function. For example, if you want to obtain the estimated probability for responding 'no' to item AD1 (Has sudden changes in mood or feeling) when the latent score is 0, you can compute the value using the formula above:

$$\begin{aligned} P(y_{ij} = 0 \mid 0) &= F\left((-.084 - .738 \times 0) \times \frac{1}{\sqrt{1 - .544}}\right) \\ &= F(-.084 \times 1.481) \\ &= F(-.124) \end{aligned}$$

which can be converted into the probability of .451 following the z-distribution.

The conversion from the z-score to the probability can be computed by hand or with the help of user-friendly functions in Excel, R, or other statistical software. Sample functions for conversions in both Excel and R are provided in Table 4. For example, if you put the value -.124 to the excel spreadsheet with the function `=norm.s.dist(-.124, true)`, it gives you the converted value of .451 as the corresponding probability value. It can be interpreted as the expected probability of saying 'no' to item AD1 is about .451 when the person has an average level of anxiety/depression latent score. If the latent score is increased to 1, the probability substantially decreases to .112 [i.e., $P(y_{ij} = 0 \mid 0) = F\left((-.084 - .738 \times 1) \times 1/\sqrt{1 - .544}\right) = F(-.822 \times 1.481) = F(-1.217) =$ probability value of .112]. Table 5 presents an example result for the probability of endorsing the item to zero when the latent score is 0 and 1 for five items of the latent factor AnxDep.

Table 4 ■ Conversions from z-score to probabilities and vice versa

	Excel	R
Conversion from original unit to probability		
z-value to probability	=norm.s.dist(Z-VALUE, true)	pnorm(Z-VALUE)
Conversion from probability to original unit		
probability to z-value	=norm.s.inv(P)	qnorm(P)

Note. Capitalized words and abbreviations indicate numeric estimates to be inserted into the function.

Table 5 ■ Probability of endorsing the item to 'no' for the anxiety/depression measure

Item	Threshold	Factor loading	R-square	Residual	z-value for factor score 0	z-value for factor score 1	p-value for factor score 0	p-value for factor score 0
AD1	-0.084	0.738	0.544	0.456	-0.124	-1.217	0.451	0.112
AD2	0.918	0.719	0.518	0.482	1.322	0.287	0.907	0.613
AD3	0.737	0.732	0.536	0.464	1.082	0.007	0.860	0.503
AD4	1.113	0.748	0.56	0.440	1.678	0.550	0.953	0.709
AD5	1.041	0.883	0.779	0.221	2.214	0.336	0.987	0.632

Note. The z-value is calculated using the function: $P(y_{ij} = 0 \mid factor\_score) = F[(threshold - \text{factor\_loading} * \text{factor\_score}) \times 1/\sqrt{1 - \text{r\_square}}]$.

Discussion

Binary data with yes/no responses are common in many applied research studies. Binary CFA using the WLSMV estimation embedded in Mplus is popularly used to validate the factor structure of the measurement with binary indicators. While conducting binary CFA using Mplus is simple and straightforward, interpreting the generated results seems quite challenging to substantive researchers who are not familiar with the measurement of discrete items. This paper is designed to provide easy-to-follow guidance for novice users on how to conduct binary CFA using Mplus and how to interpret the analysis results. As demonstrated here using the NLSY79 dataset, binary CFA models require changes to model specification, model fit assessment, and model results interpretation when compared to their continuous CFA counterparts.

With regard to model specification, users only need to identify data as categorical and apply the WLSMV estimator. Mplus makes these changes incredibly easy, with only simple modifications to the input syntax required (e.g., the addition of the `Categorical are...` option to the `DATA:` command). With regard to model fit, most indicators (specifically RMSEA, SRMR, CFI, and TLI) can be interpreted the same for continuous and binary CFA models. However, an alternative chi-square difference test must be used to compare the nested binary CFA models. While it is available, previous studies have found evidence that SRMR does not perform well when conducting CFA with WLSMV estimation for categorical variables (Yu, 2002; Garrido, Abad, & Ponsoda, 2016). The newly added SRMR from Version 8 of Mplus contains the updated function on its computation, which requires researcher's attention to examine the performance of it for evaluating the model fit.

Finally, the most notable differences between continuous and binary CFA models are in the interpretation of model results. This is due to the reliance of binary CFA models on an underlying continuous distribution of the latent variable (y*), as opposed to solely the observed indicator (y). This distinction results in the need for alternative interpretations of parameter and R-square estimates, as well as the addition of a threshold structure for binary CFA models. While these differences can present new challenges for substantive researchers more accustomed to continuous models, they are necessary for ensuring accurate and meaningful CFA models when working with binary data.

References

Asparouhov, T., & Muthén, B. (2006). Robust chi square difference testing with mean and variance adjusted test statistics.

Asparouhov, T., & Muthén, B. (2018). *Nesting and equivalence testing for structural equation models*. Structural Equation Modeling: A Multidisciplinary Journal. Retrieved from https://doi.org/10.1080/10705511.2018.1513795

UT-Austin. (2012). *Mplus tutorial*. Austin: MPlus.

Listing 6 ■ Selected Mplus output: R-square

```
R-SQUARE
    Observed                                       Two-Tailed   Residual
    Variable        Estimate       S.E.  Est./S.E.   $p$-Value     Variance
    AD1                0.544      0.059      9.198      0.000      0.456
    AD2                0.518      0.072      7.235      0.000      0.482
...
```

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in cfa. *Structural Equation Modeling*, *13*(2), 186–203. Retrieved from https://doi.org/10.1207/s15328007sem1302_2

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Washington: The Guildford Press. Retrieved from https://doi.org/10.5860/choice.44-2769

Curran, P. J., Edwards, M. C., Wirth, R. J., Hussong, A. M., & Chassin, L. (2007). The incorporation of categorical measurement models in the analysis of individual growth. In N. C. T.D. Little J.A. Bovaird (Ed.), *Modeling contextual effects in longitudinal studies* (pp. 94–125). New York: Psychology Press.

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 453–466. doi:10.1080/10705511.2017.1390394

Fernandez, E., Vargasm, R., Mahometa, M., Ramamurthy, S., & Boyle, G. J. (2012). Descriptors of pain sensation: A dual hierarchical model of latent structure. *The Journal of Pain*, *13*(6), 3. Retrieved from https://doi.org/10.1016/j.jpain.2012.02.006

Finney, S., & DiStefano, C. (2013). *Dealing with nonnormality and categorical data in structural equation modeling*. A second course in structural equation modeling. Greenwich, CT: Information Age.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466–499. Retrieved from https://doi.org/10.1037/1082-989X.9.4.466

Galandini, S., & Fieldhouse, E. (2019). Discussants that mobilise: Ethnicity, political discussion networks and voter turnout in britain. *Electoral Studies*, *57*, 163–173. Retrieved from https://doi.org/10.1016/j.electstud.2018.12.003

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? some cautionary findings via monte carlo simulation. *Psychological methods*, *21*(1), 93–99. doi:http://dx.doi.org/10.1037/met0000064

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*(4), 544–565. Retrieved from https://doi.org/10.1207/S15328007SEM1004_4

Gonzales, L. K., Glaser, D., Howland, L., Clark, M. J., Hutchins, S., Macauley, K., . . . Ward, J. (2017). Assessing learning styles of graduate entry nursing students as a classroom research activity: A quantitative research study. *Nurse Education Today*, *48*, 55–61. Retrieved from https://doi.org/10.1016/j.nedt.2016.09.016

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1–55. Retrieved from https://doi.org/10.1080/10705519909540118

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212–228. Retrieved from https://doi.org/10.1080/10705511.2011.557337

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. Retrieved from https://doi.org/10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (Eighth). Los Angeles, CA: Muthén & Muthén.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and the Family*, *99*, 295–307. Retrieved from https://doi.org/10.2307/352397

Rosseel, Y. (2014). *Structural equation modeling with categorical variables: Using R for personality research*. Racoon City: Academic Press.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using mplus*. Mawhaw: Wiley. Retrieved from http://doi.org/10.1002/9781118356258

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (vol. 30)*. Los Angeles, CA: University of California, Los Angeles.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on the journal's web site.

Citation

Kim, M., Winkler, C., & Talley, S. (2021). Binary item CFA of Behavior Problem Index (BPI) using Mplus: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *17*(2), 141–153. doi:10.20982/tqmp.17.2.p141

Received: 30/01/2021 ∼ Accepted: 06/06/2021