Examining Construct Congruence for Psychometric Tests: A Note on an Extension to Binary Items and Nesting Effects

Educational and Psychological

Measurement

2018, Vol. 78(1) 167–174

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416655379

journals.sagepub.com/home/epm

Tenko Raykov¹, George A. Marcoulides²,
Dimitar M. Dimitrov^{3,4}, and Tatyana Li¹

Abstract

This article extends the procedure outlined in the article by Raykov, Marcoulides, and Tong for testing congruence of latent constructs to the setting of binary items and clustering effects. In this widely used setting in contemporary educational and psychological research, the method can be used to examine if two or more homogeneous multicomponent instruments with distinct components measure the same construct. The approach is useful in scale construction and development research as well as in construct validation investigations. The discussed method is illustrated with data from a scholastic aptitude assessment study.

Keywords

binary item, clustering effect, congruence, construct, latent variable modeling, multidimensionality, nested models, unidimensionality

Binary items and observations stemming from individuals that are nested in higher-order units represent arguably the rule rather than the exception in the majority of contemporary educational and behavioral studies. In these investigations, researchers

¹Michigan State University, East Lansing, MI, USA

²University of California, Santa Barbara, CA, USA

³George Mason University, Fairfax, VA, USA

⁴National Center for Assessment, Riyadh, Saudi Arabia

Corresponding Author:

Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443A Erickson Hall, East Lansing, MI 48824, USA.

Email: raykov@msu.edu

are oftentimes confronted with the query whether two or more multicomponent instruments or psychometric tests (for simplicity frequently referred to as “instruments” or “tests” below) can be viewed or treated as evaluating the same underlying construct of substantive interest. This query becomes of special relevance in settings characterized by a single measurement occasion and instruments that are each unidimensional and consist of distinct components. The question is also important in construct validation studies where it may be of particular interest to find out whether a set of homogeneous tests consisting of different elements are evaluating the same latent dimension.

Recently, Raykov, Marcoulides, and Tong (2015) presented a latent variable modeling procedure that responds to the above-mentioned queries. Their method was developed for the case of (a) continuous instrument components, whereby (b) the studied subjects were in addition assumed to be independent of each other. Conditions (a) and (b) place, however, significant restrictions on the general applicability of that testing approach. The reason is that constraints (a) and (b) cannot be expected to be fulfilled routinely or perhaps even frequently in present-day educational and psychological studies.

This note extends the procedure described in Raykov et al. (2015) to the case of binary items and clustering of studied persons, which are two circumstances that are typical for instance in ability testing when examined students are randomly sampled from classes, schools, or districts say (generically referred to as “Level 2” units below; e.g., Raudenbush & Bryk, 2002). The method discussed in the remainder of this note has therefore substantially wider applicability than that earlier procedure when one is interested in examining if two or more homogeneous tests with distinct items are evaluating the same ability or perhaps two or more underlying constructs, while accounting for the highly discrete nature of these items and what may often be seen as natural nesting of examined subjects within higher-order units.

Background, Notation, and Assumptions

In order to accomplish the aims of this note, we assume throughout that a set of q ($q > 1$) multicomponent instruments, or tests, consisting of distinct binary items is available, whereby each instrument is unidimensional (i.e., comprising congeneric components; e.g., Jöreskog, 1971). We denote these items as $Y_{11}, Y_{12}, \dots, Y_{1p_1}; Y_{21}, Y_{22}, \dots, Y_{2p_2}; \dots$; and $Y_{q1}, Y_{q2}, \dots, Y_{qp_q}$, respectively ($p_j > 1; j = 1, \dots, q$). On the assumption of a latent normal variable (LNV) underlying each item (e.g., B. O. Muthén, 1984), denoted correspondingly $Y_{11}^*, Y_{12}^*, \dots, Y_{1p_1}^*; Y_{21}^*, Y_{22}^*, \dots, Y_{2p_2}^*; \dots$; and $Y_{q1}^*, Y_{q2}^*, \dots, Y_{qp_q}^*$, the following single-factor model is thus applicable for the k th test:

$$y_k^* = \tau_k + \Lambda_k \eta_k + \delta_k, \quad (1)$$

where y_k^* is the $p_k \times 1$ vector of its items and η_k is the construct evaluated by the instrument, that is, the common factor behind the LNVs $Y_{k1}^*, Y_{k2}^*, \dots, Y_{kp_q}^*$

pertaining to the items in the k th test; in addition, τ_k is the $p_k \times 1$ vector of associated thresholds, Λ_k is the $p_k \times 1$ vector of factor loadings of its indicators, and δ_k is the $p_k \times 1$ vector of pertinent unique factors ($k = 1, \dots, q$; e.g., Raykov & Marcoulides, 2011). We similarly assume that the $q \times q$ latent covariance matrix $\Phi = \text{Cov}(\eta)$ is positive definite, where $\eta = (\eta_1, \dots, \eta_q)$ is the $q \times 1$ vector of studied constructs (with $\text{Cov}(\cdot)$ denoting covariance matrix of the vector in parentheses, and priming symbolizing transposition). In this article, the individuals that provide data on all q tests' components need not be independent of each other (i.e., the individuals are allowed to be nested or clustered within Level 2 units, such as say examined students being nested within schools).

Testing Construct Congruence with Binary Items and Clustering Effects

When the set of q tests described in the previous section are administered to random samples of studied persons from higher-order units (Level 2 units), the relevant model postulated is based on the q sets of Equations (1) and is as follows (e.g., Rabe-Hesketh & Skrondal, 2012; for convenience, the same generic notation is used next as in (1)):

$$y_{kij}^* = \tau_k + \Lambda_k \eta_{kij} + \delta_{kij}, \quad (2)$$

where $i = 1, \dots, n_j$ is the subject index, n_j is the sample size of the j th Level 2 unit, $j = 1, \dots, J$, and J is the number of Level 2 units (typically randomly sampled from what may be seen as a population of Level 2 units available to study; $J > 1$, $k = 1, \dots, q$).

The concerns of this note would be addressed if one were in a position to examine whether all latent constructs in the vector η correlate perfectly among themselves (i.e., all their bivariate correlations are 1), in which case one could claim that in practical terms the tests could be treated as measuring the same underlying construct (see also Note 2). Using Proposition 1 in Raykov et al. (2015), which is generally applicable regardless of the scale of observed variables (items), it is readily found that the model defined in Equations (2) with these correlations fixed at 1, is nested in the same model without the latter $q(q-1)/2$ constraints.

Fitting the model defined by the q sets of Equations (2) as well as any model nested in it while accounting for the binary nature of the observed variables can be carried out using the well-known weighted least squares method (WLS method; e.g., Bollen, 1989; B. O. Muthén, 1984), implemented in the popular latent variable modeling (LVM; B. O. Muthén, 2002) software Mplus (L. K. Muthén & Muthén, 2015). Accounting for the clustering effect of persons within Level 2 units can be achieved then by employing the procedure described in Asparouhov (2005). Specifically, the WLS estimator is used, which yields consistent parameter estimates, and the variance of the WLS estimates are adjusted to produce corrected standard errors and chi-square test of model fit, to account for the dependence among observations within clusters. Testing the perfect correlation restrictions in model (2) subsequently, which render as

nested in it the single-factor model for all $p_1 + p_2 + \dots + p_q$ items under consideration (e.g., Raykov et al., 2015) is straightforwardly carried out using the corrected chi-square difference test (L. K. Muthén & Muthén, 2015).¹ The latter test is similarly implemented and readily available in Mplus, and the pertinent command files needed for its execution are provided in the appendix.

We illustrate next the discussed procedure for examining latent construct congruence with binary items and nesting effects using empirical data from a scholastic aptitude study.

Illustration on Data

To demonstrate the utility and applicability of the discussed procedure, we use data from the General Aptitude Test–Mathematics Part (GAT-M) that is developed and administered by the National Center for Assessment (NCA) in Saudi Arabia as a part of standardized large-scale assessments of high school graduates applying to Saudi colleges.

Specifically, the data come from the responses of 5,913 high school graduates on 12 dichotomously scored multiple-choice items on the GAT-M scales of *Algebra* (four items) and *Comprehension* (eight items). The participants are nested within 222 schools, treated here as random clusters, thus rendering a two-level hierarchical structure of the data. Using the procedure outlined earlier in this note, we are interested in this section in testing whether the constructs of Algebra and Comprehension are congruent in the context of this GAT-M data, that is, whether or not they may be seen for practical purposes as collapsing into a single construct.²

We thus commence by fitting the two-factor model to all $p = 12$ measures, with the algebra and comprehension constructs that are postulated in it only loading on the first 4 and on the last 8 of these items, respectively (using the weighted least squares method; e.g., Bollen, 1989; see first command file in the appendix). This yields the following goodness-of-fit indexes: chi-square = 204.928, degrees of freedom (df) = 53, root mean square error of approximation (RMSEA) = .022, with a 90% confidence interval (CI) of (.019, .025). These results can be considered indicative of a tenable model (e.g., Browne & Cudeck, 1993). We can therefore conclude that these findings suggest the two-factor model as plausible for the analyzed data set.

Fitting next the alternative single-factor model of relevance here, for all 12 items of the two domains under consideration, similarly leads to the following tenable goodness-of-fit indexes: chi-square = 206.438, $df = 54$, RMSEA = .022 (.019, .025) (see second command file in the appendix). These results suggest the plausibility of the hypothesis that a single construct underlies the analyzed 12 binary items.

To complete the application of the LVM procedure for examining construct congruence discussed in this note, we conduct the associated likelihood ratio test of the null hypothesis $H_0: \rho = 1$, for the correlation ρ of the two factors in the first fitted model (this test statistic is contained in the output of the nested model, i.e., the single-factor model, obtained with the second Mplus command file in the appendix).

Accordingly, the test statistic of the null hypothesis (H_0) is 1.479, $df = 1$, $p = .224$. We stress thereby that under the null hypothesis being tested the fixed correlation parameter is at the boundary of the parameter space (e.g., Raykov et al., 2015). Hence, the relevant p value for testing H_0 is $p^* = .112$ and thus not significant (e.g., Rabe-Hesketh & Skrondal, 2012). We therefore cannot reject but rather consider as retainable the null hypothesis H_0 , which as indicated before is equivalent to that of construct congruence, that is, to the statement of $q = 1$ factor underlying all $p = 12$ measures under consideration (see also Note 2).

Conclusion

This note extended the latent variable modeling procedure for construct congruence testing outlined in Raykov et al. (2015) to the case of binary items and nesting effects. With the present extension, a researcher can test whether two or more latent constructs evaluated by corresponding unidimensional multicomponent instruments or tests with distinct items can be considered effectively collapsing in empirical terms. These concerns can arise in scale construction and development studies, as well as in instrument, test, or construct validation investigations. The method is readily utilized with popular statistical software, and is directly applicable with incomplete data sets under the missing completely at random assumption (e.g., Enders, 2010). Furthermore, the procedure is straightforwardly applicable with highly discrete items that have say up to five to seven possible response options (values in the analyzed sample data set; for the case with more options, it may be recommendable to use the procedure in Raykov et al., 2015, employing the robust maximum likelihood estimation method).

The approach discussed in this note has several limitations. One, it is based on the assumption of large samples, since its method of parameter estimation and model testing is weighted least squares that itself is grounded in an asymptotic statistical theory (e.g., Bollen, 1989). Second, as assumed throughout, each of the analyzed q tests needs to be unidimensional (and consisting of components that are distinct across these q instruments). We do not see any main obstacle to extend the method of this article to the case of (individual) general structure tests, when interest lies in examining congruence of a subset of their underlying constructs across instruments, but this extension is beyond the confines of the present note. Third, while this is not really a limitation per se, it is important to keep in mind that under the tested null hypothesis of unit correlation(s) (in absolute value), the latter parameter(s) is on the boundary of the pertinent parameter space and hence the p value obtained with the used test statistic needs to be halved (when $q = 2$, or correspondingly modified in the general case; e.g., Rabe-Hesketh & Skrondal, 2012; see also Raykov et al., 2015).

In conclusion, this note outlined a readily and widely applicable latent variable modeling procedure for examining construct congruence (construct collapsibility) in settings with highly discrete items and nesting of examined persons, which may be considered the rule rather than exception in much of contemporary educational and

behavioral research. In conjunction with the method outlined in Raykov et al. (2015), the present one provides empirical scientists in these and cognate disciplines with a readily applicable approach to examining practical collapsibility of latent constructs evaluated by multiple homogeneous instruments comprising distinct components/items, whether with categorical or (approximately) continuous distributions and regardless of whether or not there is any clustering effect of the studied persons.

Appendix

Mplus Command Files

```

TITLE:      TESTING COLLAPSABILITY OF LATENT ABILITIES
            EVALUATED WITH BINARY ITEMS, IN SETTINGS WITH
            CLUSTERING/NESTING EFFECT. Step 1 - full model (two-
            factor model in the example used in the illustration section) .

DATA:      FILE = <name of raw data file> ;

VARIABLE:  NAMES = ID Y1-Y12 SCHOOL;
            USEV = Y1-Y12;
            CATEGORICAL = Y1-Y12;
            CLUSTER = SCHOOL;

ANALYSIS:  TYPE = COMPLEX;

MODEL:     ALGEBRA BY Y1-Y4;
            COMPREHN BY Y5-Y12;

SAVEDATA:  DIFFTEST = DERIVATIVES.DAT;

TITLE:      TESTING COLLAPSABILITY OF LATENT ABILITIES
            EVALUATED WITH BINARY ITEMS, IN SETTINGS WITH
            CLUSTERING/NESTING EFFECT. Step 2 - nested model (sin-
            gle-factor model in the example used in the illustration
            section) .

DATA:      FILE = <name of raw data file> ;

VARIABLE:  NAMES = ID Y1-Y12 SCHOOL;
            USEV = Y1-Y12;
            CATEGORICAL = Y1-Y12;
            CLUSTER = SCHOOL;

ANALYSIS:  TYPE = COMPLEX;
            DIFFTEST = DERIVATIVES.DAT;

MODEL:     F BY Y1-Y12;

```

Note. After a title, raw data file name and variable name assignment (selection), all analyzed binary items are declared as categorical, and the clustering variable (Level 2 identifier) is stated. The ANALYSIS command requests application of the robust procedure in Asparouhov (2005) accounting for clustering effect, the model is described in the MODEL command, and in the first command file above the needed auxiliary statistics for the corrected difference test are requested in the SAVEDATA command. In the second command file above, conducting this test is requested in the

ANALYSIS command. (See, e.g., Raykov & Marcoulides, 2006, for further discussion of Mplus commands.)

Acknowledgments

Thanks are due to B. Muthén for a valuable discussion on model fitting and parameter estimation, and to B. Tong for important comments on construct congruence. We are grateful to the leadership of the National Center for Assessment in Riyadh, Saudi Arabia, for providing the illustrative data in this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The following developments in this article are not affected if $\rho = -1$ were to hold for a pair of instruments (tests). As is well known, most of the time in empirical educational and behavioral research observed measures are positively correlated (or can be correspondingly recoded to be so). Hence, without loss of generality we can confine the discussion in this article to the case of positive factor correlation(s).
2. We stress that not rejecting the null hypothesis of perfect latent correlations (see also Note 1) does not guarantee that the underlying two or more constructs are strictly speaking identical. Rather, such a finding is only sufficient to declare them linear transformations of each other (within the limits of statistical inference). The particular manner in which such a finding could be further interpreted in an empirical study is deferred to substantive experts in the pertinent subject-matter domain.

References

- Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*, 411-434.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Thousand Oaks, CA: Sage.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variables indicators. *Psychometrika, 49*, 115-132.

- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29, 81-117.
- Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.
- Raudenbush, S. W., & Bryk, T. (2002). *Hierarchical linear and nonlinear modeling*. Thousand Oaks, CA: Sage.
- Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- Raykov, T., Marcoulides, G. A., & Tong, B. (2015). Do two or more multi-component instruments measure the same construct? Testing construct congruence using latent variable modeling. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164415604705