Published in final edited form as:

Measurement (Mahwah N.J.). 2012 ; 10(4): 212–216. doi:10.1080/15366367.2012.743401.

Old Issues in New Jacket: Power and Validation in the Context of Mixture Modeling

Gitta Lubke

Quantitative Psychology, University of Notre Dame, VU University Amsterdam

Von Davier et al. describe two analyses that aim at determining whether the constructs measured with a number of observed items are categorical or continuous in nature. The issue of types vs. traits has a long history, and is relevant in many areas of behavioral research, including personality research as emphasized by von Davier et al.

Types can be investigated using latent class models, whereas traits are modeled with latent trait or factor models. As described in von Davier et al., general modeling frameworks have been proposed that permit a comparison of these models based on fit indices. Ideally then, the decision concerning types vs. traits can be facilitated by the results of such a comparison.

The analyses presented by von Davier et al. nicely illustrate the difficulties encountered when analyzing empirical data. Two basic issues in mixture modeling, namely the power to discriminate between alternative models, and the need to validate results deserve more emphasis.

A small illustration of the impact of power on model comparisons

Mixture modeling consists of the specification of a set of alternative models that are compared using indices of model fit such as the BIC. Model specification is prone to misspecification, which in turn can easily lead to incorrect conclusions concerning the number of latent classes as well as the structural relations within class.

To illustrate the impact of power on model comparisons, consider the following small simulation. Data were generated for 10 normally distributed items using three different 2-class models, S1, S2, and S3. The items were subsequently dichotomized. S1 data had the same simple 2-factor structure in both classes, with 5 items loading on each of the factors. S2 data had the same 2-factor structure in one class, but local independence (i.e., uncorrelated items, abbreviated as LI) in the other class. LI items were generated by setting loadings on the underlying factors to zero. S3 data had 5 LI items and 5 items with non-zero loadings in class 1, and the reverse pattern in class 2 (i.e., the 5 items that are LI in class 1 load on a factor in class 2). Mean differences between classes were in the order of a Mahalanobis distance of 1.5, and were implemented as factor mean differences. As a consequence, LI items had population means equal to zero. Error variances were chosen such that the variances of the items were approximately unity before dichotomization.

All three data types are relevant in practice. S1 data correspond to two classes with substantial trait or severity differences in both classes. S2 data represent a situation where class 1 has substantial severity differences on both factors whereas class 2 does not, leading

to uncorrelated items in class 2. S3 data reflect two subtypes where class 1 shows severity differences on factor 1 but not factor 2, whereas class 2 has the reverse pattern.

Data were generated for $N=100$, or $N=1000$, and eleven models were fitted to the 3 different data types (see Table 1). All analyses were carried out using Mplus 7 (Muthén & Muthén, 1998–2012).

The results for Models M1-M8 are presented in Table 2. Note that M5 is the correct model for S1, and M7 is a correct model for S2 and S3 data since threshold differ across classes. The results illustrate the problems relating to power. First, if the set of fitted models does not include the correct model, and is limited to fitting LCA models (i.e., M1-M3), then a model with too many classes would be accepted in almost all cases. Instead of representing the true severity differences within class by factors, the item covariances are captured by additional classes. The results match previous findings (Lubke & Neale, 2006, 2008). Second, extending the set of fitted models to factor mixture models does not necessarily solve the problem. In this particular simulation $N=1000$ is necessary to select the correct model for S1 data (although $N=100$ comes close). For S2 and S3 data the incorrect MI model would be selected. The correct nMI models have more parameters, and the misfit of MI models is apparently not outweighing parsimony.

Note that correlations between BIC's of the different models over the 25 replications were all close to unity, indicating that BIC's of models fitted to any given data set followed almost always the same ordering as the averages presented in Table 2. In sum, the results imply that simple model comparisons cannot unequivocally solve the question type vs. trait since detection of the correct number of classes and of the within class structure depend on issues related to power.

In this simulation a limited number of standard mixture models were fitted to the data, similar to (but obviously more limited than) the strategy followed in von Davier et al. All models included specific constraints on the structural relations between items within class, varying from very constrained (i.e., LI) to somewhat less constrained (i.e., factor structures with class specific loadings and thresholds). In practice, appropriate constraints are usually not known a priori. Ideally, more lenient exploratory models should be used that do not require constraints on relations between items or subscales within class. However, standard maximum likelihood methods that are commonly used for the estimation of mixture models are limited with respect to the (empirical) identifiability of parameters. For instance, it is usually not possible to estimate unconstrained covariance matrices within class once the number of items exceeds 10 and the number of classes exceeds 2.

Recent developments might alleviate this issue. Bayes estimation of Structural Equation Models permits fitting more lenient models (Muthén & Asparouhov, 2012). The main idea is instead of fixing certain parameters to zero, small variance priors are provided that permit these parameters to deviate from zero. When using Bayes estimation for mixture models priors can be provided for all covariances between items within class. Essentially this allows the estimation of the full covariance matrix of observed items within each class, albeit limited by the specified variance of the priors. Choosing the variance very small approximates results from standard LCA with LI whereas very large variance priors will result in the same estimation problems as in ML. Models that are estimated with a Bayes estimator can be compared using posterior predictive p-values.

Model M9 in the simulation is a 2-class LCA model with priors for the parameters of the within class covariance matrices. Models M10 and M11 are the 2- and 3-class LCA models implementing LI, and are the same as M1 and M2, but estimated using Bayesian estimation

in order to obtain posterior predictive p-values for comparison. Results are shown in Table 3.

The 2-class LCA model with deviation from LI fits the data well in all six settings. Whether the 3-class LCA LI model would be considered a viable alternative depends on sample size and the true within class structure. The N=100 S1 and S2 data are fitted well by the 3-class model, whereas the N=1000 data are not, again highlighting the impact of sample size on model comparisons. Results using posterior predictive p-values for the 2- and 3-class LCA models are largely consistent with using ML estimation and BIC (but note a divergence for S3 N=100). Although M9 fitted the data well in all settings, recovery of the within class structure was not always successful. The reverse pattern of the S3 data (class 1 high on correlated items 1–5, low on uncorrelated items 6–10, class 2 opposite pattern) was not detected; instead, the two classes represented groups scoring either low or high on all items. Clearly, much more work is needed to evaluate the utility of Bayes LCA for the recovery of within class structures. In this small simulation only one prior was used. The sensitivity of results to the chosen prior needs to be evaluated by comparing priors with different variances.

Conclusion

The main difference between models with latent categorical variables representing types and latent continuous variables representing traits is the way covariation between observed item scores is exploited. Latent class models attribute covariation of items entirely to mean differences between classes as within each class items are constrained to be independent. Factor or latent trait models for a single homogeneous population attribute all covariation to one or more continuous latent variables. Hybrid models seek to allocate covariation to mean differences between classes as well as factors, and serve to accommodate more gradual individual differences between individuals within a given class.

Much has been written about determining the number of classes in mixture settings (starting with Pearson, see for instance McLachlan and Peel, 2000). The problem of model comparisons can be reframed as finding the correct number of classes while modeling the data within class appropriately. In the likelihood function of mixture models the within class model specification is confounded with the number of classes, with the effect that models with more constraints on the covariation between items within class usually require more classes to fit the data. LI is an extreme constraint (zero covariation), whereas for instance class invariant factor (co)variances is more lenient, but can still be overly restrictive.

Although the presented simulation is very limited, it illustrates the impact of sample size and within class structure on mixture model comparisons. One of the conclusions is therefore that, just as in more basic statistical analyses, it is desirable to provide a small power study to accompany an empirical analysis. Results of power analyses can help substantiate many of the decisions made during the model fitting process. For instance, von Davier et al. make the ‘practical’ decision to collapse response categories. This is a very common issue in applied analyses, but without a power analysis the effects of such a decision on finding an appropriate model are unknown. On the one hand, multiple response categories may contain important information concerning more continuous variability within class, but on the other hand require the estimation of many more within class parameters. As shown in a previous study, the latter can lead to incorrectly accepting models with too few classes (Lubke & Neale, 2008).

In addition a power analysis, results of an individual analysis need to be validated. It is surprising how few studies that utilize model comparisons make an attempt to validate their

results. Validation can take different forms. One possibility is to use class assignment in subsequent post hoc tests to investigate whether the classes are related to external variables in a meaningful way. A correct way of using class assignment has been described by Vermunt (2010), and is conveniently implemented in Mplus 7 (see Asparouhov & Muthén, 2012). Obviously, results can also be validated in an independent data set. Mixture models are mainly exploratory tools, and results need to be accompanied by additional information such as power analyses and validation attempts in order to be useful as evidence in favor of types or traits.

References

- Asparouhov T, Muthén B. Auxiliary Variables in Mixture Modeling: A 3-Step Approach Using Mplus. Mplus Web Notes: No. 15. 2012 <http://www.statmodel.com/examples/webnotes/webnote15.pdf>.
- Lubke GH, Neale MC. Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*. 2008; 43:592–620. PMID: 20165736. [PubMed: 20165736]
- Lubke GH, Neale MC. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*. 2006; 41:499–532.
- McLachlan, GJ.; Peel, D. *Finite mixture models*. New York: Wiley; 2000.
- Muthén B, Asparouhov T. Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*. 2012; 17:313–335. [PubMed: 22962886]
- Muthén, LK.; Muthén, BO. *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén; 1998–2012.
- Vermunt JK. Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Political Analysis*. 2010; 18:450–469.

Table 1

Fitted Models

M1	21	2-class latent class model with local independence (LI)
M2	32	3-class latent class model with LI
M3	43	4-class latent class model with LI
M4	21	2-factor single class model
M5	27	2-factor 2-class model, measurement invariant (i.e., same loadings and thresholds in the two classes)
M6	33	2-factor 3-class model, measurement invariant (i.e., same loadings and thresholds in the two classes)
M7	40	2-factor 2-class model, measurement non-invariant (i.e., loading and thresholds class specific)
M8	59	2-factor 3-class model, measurement non-invariant (i.e., loading and thresholds class specific)
M9	111	2-class latent class model with small deviations from LI, Bayes estimation
M10	21	2-class latent class model with LI, Bayes estimation
M11	32	3-class latent class model with LI, Bayes estimation

Note: The number of estimated parameters are shown in the second column.

Table 2

Average BIC for M1-M6 over 25 simulation replications

	M1	M2	M3	M4	M5	M6	M7	M8
S1 N=100	888.97	796.12	756.99	665.17	668.23	691.20	714.84	785.08
S1 N=1000	8342.36	7099.53	6206.95	5887.95	5855.45	5901.51	5948.83	6056.34
S2 N=100	971.82	932.49	924.12	876.88	850.62	871.01	894.46	958.20
S2 N=1000	9212.28	8524.28	7972.99	8135.28	7784.78	7799.79	7861.94	7945.99
S3 N=100	1048.27	907.27	936.00	903.89	874.95	896.29	911.79	979.92
S3 N=1000	9940.04	8021.63	8022.41	8316.56	7886.75	7914.19	7964.44	8047.53

Note: S1, S2, and S3 correspond to the three different true data types. Acronyms M1-M6 are explained in Table 1. Results for M6 fitted to N=1000 data are not presented due to many errors in individual runs that correctly indicated a too high level of model complexity.

Table 3

Average predictive p-values and their standard deviation over 25 replications

	M9	M10	M11
S1 N=100	0.1941 (0.034)	0.0015 (0.003)	0.1305 (0.081)
S1 N=1000	0.1833 (0.020)	0 (0)	0 (0)
S2 N=100	0.2854 (0.064)	0.0687 (0.077)	0.1720 (0.140)
S2 N=1000	0.2210 (0.027)	0 (0)	0 (0)
S3 N=100	0.2684 (0.027)	0.0023 (0.005)	0.4577 (0.043)
S3 N=1000	0.3479 (0.020)	0 (0)	0.3305 (0.048)