

An Application of Multilevel Mixture Item Response Theory Model

Sedat ŞEN *

Türker TOKER **

Abstract

Although the mixture item response theory (IRT) models are useful for heterogeneous samples, they are not capable of handling a multilevel structure that is very common in education and causes dependency between hierarchies. Ignoring the hierarchical structure may yield less accurate results because of violation of the local independence assumption. This interdependency can be modeled straightforwardly in a multi-level framework. In this study, a large-scale data set, TEOG exam, was analyzed with a multilevel mixture IRT model to account for dependency and heterogeneity in the data set. Sixteen different multilevel models (different class solutions) were estimated using the eighth-grade mathematics data set. Model fit statistics for these 16 models suggested the CB1C4 model (one school-level and four student-level latent classes) was the best fit model. Based on CB1C4 model, the students were classified into four latent student groups and one latent school group. Parameter estimates obtained with maximum likelihood estimation were presented and interpreted. Several suggestions were made based on the results.

Key Words: Item response theory, mixture models, multilevel mixture item response theory, maximum likelihood estimation, TEOG exam.

INTRODUCTION

Item response theory (IRT; Lord & Novick, 1968) models have been commonly used by practitioners for several testing applications, including test development, item analyses, test scoring, and differential item functioning. In contrast to the classical test theory that makes analyses on total score, IRT provides the opportunity to perform analyses based on individual test items. Examinee responses to each item are typically analyzed with a range of IRT models, including one-parameter, two-parameter, and three-parameter logistic models. Several extensions of these models have been proposed for the different data conditions (van der Linden & Hambleton, 1997). Successful applications of IRT models depend on meeting their assumptions. According to Embretson and Reise (2000), two major assumptions are required for estimating item parameters with IRT; local independence and appropriate dimensionality. Local independence indicates that the responses to an item are unrelated to any other item when the person's location is controlled (de Ayala, 2009). Appropriate dimensionality indicates that the IRT model has the correct number of trait level estimates for examinees (Embretson & Reise, 2000). de Ayala (2009) states another assumption that is called functional form assumption. This simply represents whether the data follow the function specified by the model. Additional assumptions may be needed for different estimation techniques.

Another characteristic of IRT involves the indeterminacy property which refers to the independence of item parameter estimates from sample characteristics and independence of person estimates from item characteristics. This property claims that item parameter estimates of a test should not differ based on the varying populations. Thus, a single homogenous population was expected in the traditional IRT model estimations. However, there may be situations that examinees can come from different

* Assoc. Prof., Harran University, Faculty of Education, Şanlıurfa-Turkey, sedatsen@harran.edu.tr, ORCID ID: 0000-0001-6962-4960

** Assist. Prof., Uşak University, Faculty of Education, Uşak-Turkey, tokerturker@hotmail.com, ORCID ID: 0000-0002-3038-7096

To cite this article:

Şen, S., & Toker, T. (2021). An application of multilevel mixture item response theory model. *Journal of Measurement and Evaluation in Education and Psychology*, 12(3), 226-238. doi: 10.21031/epod.893149

Received: 8.03.2021
Accepted: 23.07.2021

subpopulations (Lubke & Muthén, 2005). Assuming a single population can be considered a limitation of IRT models. Other alternatives should be used for such cases. A relatively new approach called mixture IRT was developed to examine and account for the possible subpopulations in the data (Mislevy & Verhelst, 1990; Rost, 1990). Mixture IRT models are analytically based on mixture models (McLachlan & Peel, 2000), and this mixture is achieved by combining an IRT model with a latent class analysis model. Unlike the quantitative information provided by IRT models, one of the strengths of mixture IRT models is to provide both quantitative and qualitative information about the items and examinees. In the presence of multiple populations, the application of traditional IRT models may yield biased results. In this case, the mixture IRT model would be the most appropriate approach.

Mixture IRT models have been used to investigate several psychometric issues, such as detection of differential item functioning (DIF; Cohen & Bolt, 2005), different response strategies (Mislevy & Verhelst, 1990), effects of testing accommodations (Cohen, Gregg, & Deng, 2005), and test speededness (Bolt, Cohen, & Wollack, 2002). Although the mixture IRT models are useful for heterogeneous samples, they are not capable of handling a multilevel structure, common in educational research. Ignoring the hierarchical structure may yield less accurate results because of violation of the local independence assumption (Lee, Cho, & Sterba, 2018). Multilevel models acknowledge that the data consisted of hierarchies by allowing for residual components at each level in the hierarchy. When the structure of data is nested, multilevel modeling provides more accurate estimates and inferences. In this regard, multilevel mixture IRT models (Asparouhov & Muthén, 2008; Cho & Cohen, 2010; Vermunt, 2008) were developed to account for possible dependency, such as can arise due to cluster or multistage sampling. Multilevel mixture IRT models extend the standard mixture IRT model to allow detection of nuisance dimensionality at different levels in the data. In the model, dependency is taken into account by incorporating continuous or categorical latent variables or both at the higher level. Multilevel mixture IRT models have been used in several studies including Bacci and Gnaldi (2015); Cho and Cohen (2010); Finch and Finch (2013); Jilke, Meuleman, and van de Walle (2015); Lee et al., (2018); Sen and Cohen (2020); Sen, Cohen, and Kim, (2018); Liu, Liu, and Li (2018); Li, Liu, and Liu, (2020); Tay, Diener, Drasgow, and Vermunt (2011); Varriale and Vermunt (2012); and Vermunt (2008, 2011). Except for Cho and Cohen (2010) and Sen et al. (2018), all of these studies used maximum likelihood estimation (MLE).

Purpose of the Study

Large-scale data sets (e.g., TIMSS, PISA) are typically analyzed with IRT models. Recently, researchers have started to analyze such data sets using mixture IRT models to account for the heterogeneous structure underlying the examinee population (Choi, Alexeev, & Cohen, 2015; Sen et al., 2018). Although the use of mixture IRT models for large-scale data sets has increased recently, multilevel mixture item response models are seldom used compared to single-level mixture item response models (e.g., Liu et al., 2018). The data used in this study consist of a nested structure. Students are nested in schools, along with schools nested in districts. Research mentioned above provides useful information about estimates and inferences when data have subgroups. The purpose of this study is to illustrate the application of a multilevel mixture IRT model on a large-scale data set. In this study, we attempt to show how the multilevel mixture IRT model can be used to identify and describe characteristics of latent groups in the presence of a multilevel data structure.

METHOD

Multilevel mixture IRT modeling approach was used in this study to explain the heterogeneity behind the hierarchical data set under examination. Detailed explanations about the data set and analyses are presented below.

Participants and Data Set

37,276 eighth-grade students studying in one of the provinces of the South East region of Turkey constituted the participants of this study. The sample consists of students from 521 schools from 13 districts of that province. The number of students per school varied between 1 and 609. Thirteen schools with less than 10 students were excluded from the data set in order to prevent estimation errors for hierarchical data. Thus, the remaining 508 schools with 37,199 students were used as an effective sample size in this study. The responses of these students to the Mathematics section of TEOG (Transition from Basic Education to Secondary Education) exam in November 2016 were used. There were twenty multiple-choice questions in each of four different booklets (A, B, C, and D) in the TEOG exam. Each booklet was re-coded as 0 for incorrect and 1 for correct responses. In addition, empty answers were coded as incorrect answers. After re-coding the data set, it was prepared for multilevel analyses by creating the school IDs.

Data Analysis

The multilevel mixture IRT models were used to analyze the TEOG Mathematics data set in this study. The formula of multilevel mixture IRT model can be given as follows (Lee et al., 2018, p.4):

$$\text{logit}[P(y_{jki} = 1 | \theta_{jkg}, \theta_k, C_{jk})] = \alpha_{i.g.W} \theta_{jkg} + \alpha_{i.B} \theta_k - \beta_{ig} \tag{1}$$

where j and k ($k = 1, \dots, K$) represent respondents and clusters, respectively, C_{jk} is a categorical latent variable at the within level for a respondent j nested within a cluster k , $\alpha_{i.g.W}$ is a class-specific within-level item discrimination parameter, $\alpha_{i.B}$ is a between-level item discrimination parameter, β_{ig} is a class-specific item location parameter, θ_{jkg} is a class-specific within-level continuous latent variable and θ_k is a between-level continuous latent variable. Both of these two continuous latent variables are assumed to follow a normal distribution. A sample path diagram for two level mixture IRT model with five items is displayed in Figure 1. Interested readers are referred to Lee et al. (2018) for more details.

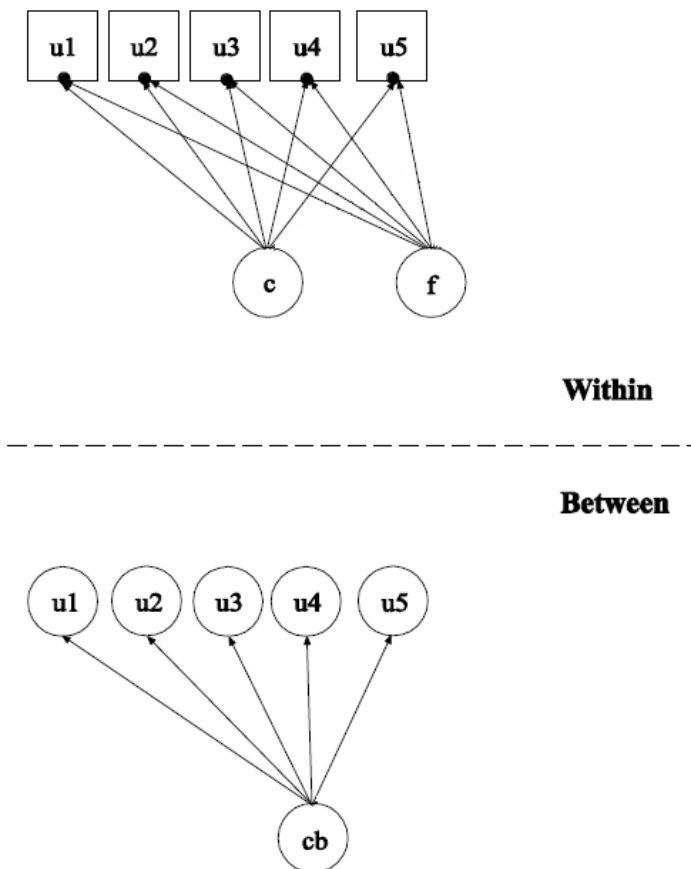


Figure 1. Diagram of the Two Level Mixture IRT Model

All analyses were conducted using Mplus 8.2 software (Muthén & Muthén, 1998-2018). Marginal maximum-likelihood estimation technique with the MLR estimator option was used for parameter estimation. For model identification, factor mean and variance were set to be 0 and 1, respectively (Muthén, 2008). The factor means in all classes were fixed to zero as the thresholds were not held equal across classes and the variances were fixed at one to set the metric of the factors. In IRT, this is usually done by fixing the factor variance to one and freeing all factor loadings. The syntax used for the final model is presented in the Appendix. TYPE = TWOLEVEL MIXTURE; ALGORITHM = INTEGRATION; options were used under ANALYSIS command in order to estimate a two level mixture IRT model. %WITHIN% and %BETWEEN% options were used to specify number of classes at each level and the relationship between items and factors under the MODEL command.

As the latent classes are unobserved and the number of classes is unknown a priori, mixture models typically follow an exploratory approach to determine the final model. Generally, it starts with a single-class model and continues by adding a class to the model until a desirable fit is obtained. Information criteria-based relative fit indices are used to determine the best-fitting model. Three information criteria indices, Akaike's Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), and Sample-size Adjusted BIC (SABIC; Sclove, 1987), were used to determine the best model in terms of fit. In this study, the following formulas were used to calculate information criteria indices:

$$AIC = -2LL + 2p, \quad (2)$$

$$BIC = -2LL + \log(n) \times p, \quad (3)$$

$$SABIC = -2LL + \log\left(\frac{n+2}{24}\right) \times p. \quad (4)$$

where LL represents log-likelihood value, p denotes the number of estimated parameters and n is used for sample size. Multilevel mixture IRT models with different numbers of between and within level classes were compared in this study. The following 16 multilevel models were estimated: CB1C1 (one between level and one person level class), CB1C2, CB1C3, CB1C4, CB2C1, CB2C2, CB2C3, CB2C4, CB3C1, CB3C2, CB3C3, CB3C4, CB4C1, CB4C2, CB4C3, and CB4C4 where CB represents between-level class and C represents the within-level class. AIC, BIC, and SABIC indices were calculated for each of these models. The smallest value of each information criterion index was taken as indicating the best fitting model. Li, Cohen, Kim, and Cho (2009) and Preinerstorfer and Formann (2011) suggested that the BIC was more accurate than the AIC for model selection with single-level dichotomous mixture IRT models. In line with these studies, Sen et al. (2018) suggested that BIC was more accurate at the selection of multilevel mixture Rasch models. Therefore, BIC was used as the main index for model selection in this study.

RESULTS

As the multilevel mixture IRT model was used to analyze the data, the hierarchical structure of the data set was examined using the intra-class correlation (ICC; Raudenbush & Bryk, 2002) before conducting the analyses. A multilevel Rasch model was estimated based on the linear mixed-effects model approach using the *lmer* function (Bates & DebRoy, 2004). The ICC was .578, indicating school level can explain 57.8% of the total variance. As mentioned earlier, 16 different models were analyzed with the same data set. Model fit statistics for these 16 models are presented in Table 1. As shown in Table 1, CB1C4 (one school-level and four student-level latent classes) and CB3C4 had the smallest AIC values, CB1C4 and CB2C4 had the smallest BIC and SABIC values. Sen et al. (2018) suggested that BIC was more accurate at the selection of multilevel mixture Rasch models. Therefore, in view of these results, we conclude that the heterogeneity behind this real data can be explained by the CB1C4 model.

Table 1. Fit Statistics for Estimated Models

	LL	np	AIC	BIC	SABIC
CB1C1	-443868.808	40	887817.615	888158.577	888031.457
CB1C2	-437191.328	121	874624.657	875656.315	875271.777
CB1C3	-434580.273	182	869524.546	871076.297	870497.901
CB1C4	-433001.290	243	866488.580	868560.423	867788.169
CB2C1	-443038.394	121	886318.788	887350.447	886965.909
CB2C2	-436628.305	242	873740.609	875803.927	875034.851
CB2C3	-433632.697	363	867991.393	871086.369	869932.755
CB2C4	-432101.861	484	865171.721	869298.356	867760.204
CB3C1	-442778.648	182	885921.297	887473.048	886894.652
CB3C2	-436376.190	363	873478.379	876573.356	875419.741
CB3C3	-433708.326	544	868504.652	873142.853	871414.021
CB3C4	-431726.026	725	864902.053	871083.479	868779.429
CB4C1	-442648.253	243	885782.506	887854.349	887082.095
CB4C2	-436242.312	484	873452.624	877579.259	876041.107
CB4C3	-433412.992	725	868275.983	874457.410	872153.359
CB4C4	-432016.674	666	867365.347	871042.356	868925.808

Note. LL = Log-likelihood; np = number of parameters; AIC = Akaike's information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC; CB1C1 = one school level and one student level; CB4C4 = four school level and four student level; other model names on the first column follow the similar labeling rules.

Based on CB1C4 model, the students were classified into four latent student groups and one latent school group. Table 2 presents the final class counts and proportions for each latent class variable based on estimated posterior probabilities. Student level Class 4 is the dominant class (.499) based on the proportion of students within each latent school level class. It should be noted that the sum of the proportions reported in Table 2 equals 1.

Table 2. Final Class Counts and Proportions for Each Student Level Latent Class

Class	Count	Proportion
1	7597	.20379
2	3781	.10144
3	7302	.19589
4	18597	.49888

Item parameter estimates of the final model are presented in Table 3. Mplus output provided slope and intercept (threshold) parameters for within- and between-level separately. Thus, W and B subscripts were used to differentiate between the two levels. As shown in Table 3, slope (α) parameters were reported for each class at both levels. However, thresholds were obtained only for between level part. As explained by Şen, Cohen, and Kim (2020), IRT discrimination parameters are equal to slope parameters that are provided in Mplus output. However, item difficulty parameters can be obtained by dividing threshold values by slope values for each item. Item difficulty parameters for Class 4 appear to be positive and higher than those of other classes.

DISCUSSION and CONCLUSION

In this study, a multilevel mixture IRT model was presented and applied to a large-scale test dataset. The proposed model was a combination of an IRT model, a latent class model, and a multilevel model. Combining the advantages of these different techniques gives researchers a broad understanding of the concept. Analysis done at the individual level assumes one's standing is a product of the individual level. But individuals within a class might affect one another; thus, this makes them quantitatively comparable.

Table 3. Item Parameter Estimates of the Final Model

Item	Class 1			Class 2			Class 3			Class 4		
	$\alpha_{1,W}$	$\alpha_{1,B}$	β_1	$\alpha_{2,W}$	$\alpha_{2,B}$	β_2	$\alpha_{3,W}$	$\alpha_{3,B}$	β_3	$\alpha_{4,W}$	$\alpha_{4,B}$	β_4
1	2.406	1.185	-1.240	0.390	0.263	-0.034	2.191	0.880	-0.711	-0.851	0.025	0.723
2	2.008	1.199	-2.217	1.276	0.902	0.533	-0.166	0.456	-0.676	0.124	0.064	1.177
3	1.558	1.052	-3.589	1.655	0.684	-0.849	2.124	1.215	-2.233	0.105	0.170	0.071
4	0.508	0.819	-2.021	-0.004	0.264	0.330	0.897	0.534	0.714	0.314	0.147	1.058
5	2.345	0.987	-0.578	1.243	0.612	0.634	2.037	1.003	-0.565	-0.513	0.004	0.935
6	1.742	0.739	0.275	4.516	1.995	-1.989	0.518	0.448	0.746	0.466	0.047	0.573
7	1.132	0.566	0.102	4.954	2.244	-1.349	0.386	0.253	1.297	0.732	0.041	0.705
8	0.943	0.761	-1.687	2.031	1.035	-0.244	0.345	0.493	0.461	0.454	0.104	0.984
9	1.963	0.901	-0.985	3.017	1.301	-0.565	1.579	0.883	0.039	-0.013	0.015	1.095
10	0.922	0.729	-1.330	1.262	0.819	1.296	0.718	0.603	1.278	-0.260	-0.067	1.110
11	0.754	0.665	-0.966	0.623	0.442	1.479	1.313	0.576	0.797	0.013	0.095	1.161
12	1.223	0.752	-0.204	1.177	0.620	0.407	0.356	0.061	0.316	0.146	0.086	1.163
13	1.173	0.739	-0.910	-1.176	-0.140	0.775	0.689	0.688	1.198	0.167	0.015	1.554
14	1.237	0.740	1.211	0.176	0.061	2.073	0.454	0.227	1.681	-0.279	-0.100	1.056
15	1.259	0.500	0.435	1.473	1.012	0.656	-0.179	0.088	0.308	0.137	0.061	1.480
16	1.083	0.630	0.554	-0.061	0.162	-0.435	0.324	0.303	1.905	-0.518	-0.018	1.676
17	0.963	0.543	0.707	1.132	0.850	1.413	-0.690	-0.037	1.482	-0.164	-0.140	1.513
18	1.338	0.798	-0.351	-0.355	0.113	0.794	0.358	0.222	0.243	-0.213	0.044	1.378
19	1.266	0.688	0.910	1.851	0.766	-0.056	0.090	0.040	1.592	0.306	0.032	1.052
20	0.590	0.586	0.173	-0.011	0.161	0.878	-0.148	-0.063	1.046	-0.188	-0.006	1.469

First, an ICC value was calculated to see the ratio of the between-cluster variance to the total variance. This was done to see the proportion of the total variance in Y that is accounted for by the hierarchy. Later, different models were analyzed for model fit purposes. Using both BIC and SABIC indices one model was selected from 16 competing models.

Similar to the Vermunt (2008) study, it was found that there were differences in average latent abilities across schools. However, when a student’s ability was controlled, there were no differences in the individual item performances between schools. At this point, a detailed analysis including covariates might answer the question of why there were differences in average latent abilities across schools. Additionally, the Mplus software used in this study can estimate even more complex models; this model can be extended by adding continuous and categorical latent variables both at student and school levels while noting possible practical problems.

The proposed model can be useful for educational researchers when data are multilevel. Moreover, if there are concerns about heterogeneity in datasets when validity is the main issue for cross-cultural studies using large-scale assessment data. Also, the model can be a handful when researchers’ main interest is investigating the possible latent structures that share the same measurement model within the population. The main advantage of the proposed model is it can infer person-level measurement class along with the hierarchical class at the same time.

The multilevel mixture IRT models are becoming more popular among researchers. It is suggested that studies using the model should consider some requirements of multilevel mixture IRT models. The sample size requirement is one of the main concerns for researchers. This is mainly built on two blocks: the numbers of items and sample sizes required at each level. Simulation studies showed that $n = 5$ to 30 person-level units and 30 to 500 hierarchical-level units are required (Lukočienė, Varriale, & Vermunt, 2010).

This paper presents a general outline of multilevel mixture IRT model. The approach presented in this study has multiple theoretical and methodological advantages. Multilevel mixture IRT deals with issues of latent class models and measurement under one single model. In conclusion, the model can be used where researchers suspect latent structures within the data, when data are hierarchical, also when there is a need for cross-cultural comparisons. The results showed that these student and school-level classes are interpretable and uniquely explain how different latent ability structures spread across individuals and schools.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723. doi: 10.1109/TAC.1974.1100705
- Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing.
- Bacci, S., & Gnaldi, M. (2015). A classification of university courses based on students' satisfaction: An application of a two-level mixture item response model. *Quality & Quantity*, 49(3), 927-940. doi: 10.1007/s11135-014-0101-0
- Bates, D. M., & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, 91(1), 1-17. doi: 10.1016/j.jmva.2004.04.013
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39(4), 331-348. doi: 10.1111/j.1745-3984.2002.tb01146.x
- Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, 35(3), 336-37. doi: 10.3102/1076998609353111
- Choi, Y. J., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, 15(3), 239-253. doi: 10.1080/15305058.2015.1007241
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2), 133-148. Retrieved from <https://www.jstor.org/stable/20461782>
- Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities: Research and Practice*, 20(4), 225-233. doi: 10.1111/j.1540-5826.2005.00138.x
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Publications.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, 73(6), 973-993. doi: 10.1177/0013164413494776
- Jilke, S., Meuleman, B., & van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, 75(1), 36-48. doi: 10.1111/puar.12318
- Lee, W. Y., Cho, S. J., & Sterba, S. K. (2018). Ignoring a multilevel structure in mixture item response models: Impact on parameter recovery and model selection. *Applied Psychological Measurement*, 42(2), 136-154. doi: 10.1177/0146621617711999
- Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33(5), 499-518. doi: 10.1177/0146621608326422
- Li, M., Liu, Y., & Liu, H. (2020). Analysis of the problem-solving strategies in computer-based dynamic assessment: The extension and application of multilevel mixture IRT model. *Acta Psychologica Sinica*, 52(4), 528-540. doi: 10.3724/SP.J.1041.2020.00528
- Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology*, 9, Article 1372. doi: 10.3389/fpsyg.2018.01372
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10(1), 21-39. doi: 10.1037/1082-989X.10.1.21
- Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, 40(1), 247-283. Retrieved from <https://www.jstor.org/stable/41336886>
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215. doi: 10.1007/BF02295283
- Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In Hancock, G. R., & Samuelsen, K. M. (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age Publishing, Inc.
- Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus users guide* (7th ed.). Los Angeles, CA: Author.
- Preinerstorfer, D., & Formann, A. K. (2011). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, 65(2), 251-262. doi: 10.1111/j.2044-8317.2011.02020.x

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271-282. doi: 10.1177/014662169001400305
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464. Retrieved from <https://www.jstor.org/stable/2958889>
- Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343. doi: 10.1007/BF02294360
- Şen, S., & Cohen A. S. (2020). The impact of test and sample characteristics on model selection and classification accuracy in the multilevel mixture IRT model. *Frontiers in Psychology*, 11, Article 197. doi: 10.3389/fpsyg.2020.00197
- Şen, S., Cohen, A. S., & Kim, S. H. (2018). Model selection for multilevel mixture Rasch models. *Applied Psychological Measurement*, 43(4), 1-18. doi: 10.1177/0146621618779990
- Şen, S., Cohen, A., & Kim, S.-H. (2020). A short note on obtaining item parameter estimates of IRT models with Bayesian estimation in Mplus. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 11(3), 266-282. doi: 10.21031/epod.693719
- Tay, L., Diener, E., Drasgow, F., & Vermunt, J. K. (2011). Multilevel mixed-measurement IRT analysis: An explication and application to self-reported emotions across the world. *Organizational Research Methods*, 14(1), 177-207. doi: 10.1177/1094428110372674
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, 47(2), 247-275. doi: 10.1080/00273171.2012.658337
- Vermunt, J. K. (2008). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics*, 37(3&4), 285-299.
- Vermunt, J. K. (2011). Mixture models for multilevel data sets. In J. Hox & K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 59-81). New York: Routledge.

Çok Düzeyli Karma Madde Tepki Kuramı Modelinin Bir Uygulaması

Giriş

Madde tepki kuramı (MTK; Lord & Novick, 1968) modelleri, uygulayıcılar tarafından test geliştirme, madde analizi, test puanlama ve farklılaşan madde fonksiyonu dahil olmak üzere çeşitli test uygulamalarında yaygın olarak kullanılmaktadır. Toplam puan üzerinden yapılan analizlere dayanan klasik test teorisinin aksine, MTK, bireysel test maddelerine dayalı analizler yapma fırsatı sunar. Sınava girenlerin doğru-yanlış şeklinde kodlanan her bir maddeye verdiği yanıtlar tipik olarak bir parametrelili, iki parametrelili ve üç parametrelili lojistik modelleri içeren bir dizi MTK modeliyle analiz edilir. Farklı veri koşulları için bu modellerin çeşitli uzantıları önerilmiştir (van der Linden & Hambleton, 1997). MTK modellerinin uygulamalarının başarısı varsayımlarının karşılanmasına bağlıdır. Embretson ve Reise'e (2000) göre, MTK ile madde parametrelerini tahmin etmek için iki ana varsayım gereklidir; yerel bağımsızlık ve uygun boyutluluk. Yerel bağımsızlık, kişinin konumu kontrol edildiğinde bir maddeye verilen yanıtların başka herhangi bir madde ile ilgisi olmadığını gösterir (de Ayala, 2009). Uygun boyutluluk, MTK modelinin sınava giren kişiler için doğru sayıda özellik düzeyi tahminine sahip olduğunu gösterir (Embretson & Reise, 2000). de Ayala (2009), işlevsel form varsayımı olarak adlandırılan başka bir varsayım belirtir. Bu, verilerin model tarafından belirtilen işlevi takip edip etmediğini gösterir. Farklı tahmin teknikleri için ek varsayımlar gerekebilir.

MTK'nın diğer bir özelliği, madde parametre tahminlerinin örneklem özelliklerinden ve kişi tahminlerinin madde özelliklerinden bağımsızlığına atıfta bulunan değişmezliktir. Bu özellik, bir testin madde parametresi tahminlerinin değişen popülasyonlara göre farklılık göstermemesi gerektiğini iddia etmektedir. Bu nedenle, geleneksel MTK modeli tahminlerinde tek bir homojen popülasyon varsayılır. Ancak, sınava girenlerin farklı alt popülasyonlardan gelebileceği durumlar olabilir (Lubke ve Muthén,

2005). Bu durumda tek bir popülasyonun MTK modellerinin bir sınırlaması olarak kabul edilebileceği varsayılır. Bu tür durumlar için başka alternatif modeller kullanılmalıdır. Verilerdeki olası alt popülasyonları incelemek ve hesaba katmak için karma MTK adı verilen nispeten yeni bir yaklaşım geliştirilmiştir (Mislevy & Verhelst, 1990; Rost, 1990). Karma MTK modelleri analitik olarak karma modellere (McLachlan & Peel, 2000) dayalıdır ve karma MTK modeli bir MTK modeli ile bir örtük sınıf analizi modeli birleştirilerek elde edilir. MTK modelleri tarafından sağlanan nicel bilginin aksine, karma MTK modellerinin güçlü yönlerinden biri, maddeler ve sınava giren kişiler hakkında hem nicel hem de nitel bilgi sağlamasıdır. Birden fazla popülasyonun varlığında, geleneksel MTK modellerinin uygulanması yanlış sonuçlar verebilir. Bu durumda, karma MTK modeli daha uygun bir yaklaşım olacaktır.

Karma MTK modelleri, farklılaşan madde fonksiyonunun tespiti (DIF; Cohen & Bolt, 2005), farklı yanıt stratejileri (Bolt, Cohen, & Wollack, 2002; Mislevy & Verhelst, 1990) test düzenlemelerinin etkileri (Cohen, Gregg, & Deng, 2005) ve test hızının etkileri (Bolt ve diğerleri, 2002) gibi çeşitli psikometrik sorunları araştırmak için kullanılmıştır. Karma MTK modelleri heterojen örneklem için kullanışlı olsa da eğitim araştırmalarında yaygın olan çok düzeyli bir yapıyı hesaba katmamaktadır. Hiyerarşik yapıyı göz ardı etmek, düzey içi gözlemler arası bağımsızlık varsayımının ihlali nedeniyle daha yanlış sonuçlar verebilir (Lee, Cho, & Sterba, 2018). Çok düzeyli modeller, hiyerarşideki her düzeyde artık bileşenlere izin vererek verilerin hiyerarşilerden oluştuğunu kabul etmektedir. Veri yapısı iç içe olduğunda çok düzeyli modeller daha doğru tahminler ve çıkarımlar yapılmasını sağlamaktadır. Bu bağlamda, hiyerarşik veya çok düzeyli örneklemeden kaynaklanabilecek olası bağımlılığı hesaba katmak için çok düzeyli karma MTK modelleri (Asparouhov & Muthén, 2008; Cho & Cohen, 2010; Vermunt, 2008) geliştirilmiştir. Çok düzeyli karma MTK modelleri, verilerdeki farklı düzeylerde rahatsız edici boyutluluğun saptanmasına izin vermek için standart karma MTK modelini genişletir. Modelde, bağımlılık, sürekli veya kategorik örtük değişkenleri veya her ikisini üst düzeyde dahil ederek hesaba katılır. Çok düzeyli karma MTK modelleri son yıllarda birçok araştırmada kullanılmaya başlamıştır (Bacci & Gnaldi, 2015; Cho & Cohen, 2010; Finch & Finch, 2013; Jilke, Meuleman, & van de Walle, 2015; Lee ve diğerleri, 2018; Liu, Liu, & Li, 2018; Sen & Cohen 2020; Sen, Cohen, & Kim, 2018; Tay, Diener, Drasgow, & Vermunt, 2011; Varriale & Vermunt 2012; Vermunt, 2008). Cho ve Cohen (2010) ve Sen ve diğerleri (2018) dışında tüm bu çalışmalar maksimum olabilirlik tahminini (MLE) yöntemini kullanmışlardır.

Büyük ölçekli veri setleri (örneğin, TIMSS, PISA) tipik olarak MTK modelleriyle analiz edilir. Son zamanlarda araştırmacılar, incelenen popülasyonun altında yatan heterojen yapıyı hesaba katmak için bu tür veri setlerini karma MTK modelleri kullanarak analiz etmeye başlamışlardır (Choi, Alexeev, & Cohen, 2015; Sen ve diğerleri, 2018). Büyük ölçekli veri kümeleri için karma MTK modellerinin kullanımı son zamanlarda artmış olsa da çok düzeyli karma MTK modelleri, tek düzeyli karma MTK modellerine kıyasla nadiren kullanılmaktadır. Bu çalışmada kullanılan veriler iç içe bir yapıdan oluşmaktadır. Öğrenciler okullarda, okullar ise ilçeler içerisinde gruplanmaktadır. Yukarıda bahsedilen araştırmalarda, veri setleri alt gruplardan oluştuğunda çok düzeyli modellerin daha doğru tahminler ve çıkarımlar sağladığı vurgulanmaktadır. Bu çalışmanın amacı, çok düzeyli bir karma MTK modelinin hiyerarşik yapıya sahip büyük ölçekli bir veri setine uygulanmasını göstermektir. Bu çalışmada, çok düzeyli bir veri yapısının varlığında örtük sınıfların özelliklerini tanımlamak ve açıklamak için çok düzeyli karma MTK modelinin nasıl kullanılabileceğini göstermeye çalışıyoruz.

Yöntem

Bu çalışmada incelenen hiyerarşik veri setinin ardındaki heterojenliği açıklamak için çok düzeyli karma MTK modelleme yaklaşımı kullanılmıştır. Türkiye'nin Güneydoğu bölgesi illerinden birinde öğrenim gören 37,276 sekizinci sınıf öğrencisi bu çalışmanın katılımcılarını oluşturmaktadır. Örneklem, o ilin 13 ilçesinde yer alan 521 okuldaki öğrencilerden oluşmaktadır. Okul başına öğrenci sayısı 1 ile 609 arasında değişmiştir. 10'dan az öğrencisi olan 13 okul hiyerarşik veriler için tahmin hatalarını önlemek amacıyla veri setinden çıkarılmıştır. Böylece, 37,199 öğrenci ile kalan 508 okul bu çalışmada etkin örneklem büyüklüğü olarak kullanılmıştır. Örnek analizlerde bu öğrencilerin Kasım 2016'da TEOG sınavının Matematik bölümüne verdikleri yanıtlar kullanılmıştır. TEOG sınavında dört

farklı kitapçığın (A, B, C ve D) her birinde yirmi çoktan seçmeli soru vardır. Her kitapçık yanlış yanıt için 0 ve doğru yanıt için 1 olarak yeniden kodlanmıştır. Ayrıca boş cevaplar yanlış cevap olarak kodlanmıştır. Veri seti yeniden kodlandıktan sonra okul kimlikleri oluşturularak çok düzeyli analizlere hazırlanmıştır.

Bu çalışmada TEOG Matematik veri setinin analizinde çok düzeyli karma MTK modelleri kullanılmıştır. Tüm analizler Mplus 8.2 (Muthén & Muthén, 1998-2018) yazılımı kullanılarak gerçekleştirilmiştir. Parametre tahmini için marjinal maksimum olabilirlik kestirim tekniğinin sağlam versiyonu (MLR) kullanılmıştır. Model tanımlaması için faktör ortalaması ve varyansı sırasıyla 0 ve 1 olarak ayarlanmıştır (Muthén, 2008). Eşikler sınıflar arasında eşit tutulmadığından ve faktörlerin metriğini ayarlamak için varyanslar bire sabitlendiğinden, tüm sınıflardaki faktör ortalamaları sıfıra sabitlendi. MTK'da bu genellikle faktör varyansını bire sabitleyerek ve tüm faktör yüklerini serbest bırakarak yapılır.

Örtük sınıflar gözlemlenmediğinden ve sınıf sayısı önceden bilinmediğinden, karma model uygulamalarında nihai modeli belirlemek için keşfedici bir yaklaşım izlenir. Genellikle tek sınıflı bir modelle başlanır ve istenen bir uyum elde edilinceye kadar modele bir sınıf eklenerek devam edilir. En uygun modeli belirlemek için bilgi kriterlerine dayalı göreceli uyum (bilgi kriteri) indeksleri kullanılır. En iyi modeli belirlemek için Akaike'nin bilgi kriteri (AIC; Akaike, 1974), Bayesci bilgi kriteri (BIC; Schwarz, 1978) ve örneklem düzeltmeli BIC (SABIC; Sclove, 1987) olmak üzere üç bilgi kriteri indeksi kullanılmıştır.

Bu çalışmada, farklı sayıda düzey arası ve sınıf içi sınıflara sahip çok düzeyli karma MTK modelleri karşılaştırılmıştır. Hem öğrenci hem de okul düzeyindeki farklı sınıf kombinasyonlarına dayalı olarak 16 çok düzeyli model tahmin edilmiştir: CB1C1 (biri öğrenci düzeyi sınıf ve bir okul düzeyi sınıf), CB1C2, CB1C3, CB1C4, CB2C1, CB2C2, CB2C3, CB2C4, CB3C1, CB3C2, CB3C3, CB3C4, CB4C1, CB4C2, CB4C3 ve burada CB, düzeyler arası sınıfı temsil eder ve C, düzey içi sınıfı temsil eder. Bu modellerin her biri için AIC, BIC ve SABIC indeksleri hesaplanmıştır. Her bilgi kriteri indeksinin en küçük değeri, en uygun modeli gösterecek şekilde alınmıştır. Li, Cohen, Kim ve Cho (2009) ve Preinerstorfer ve Formann (2011), BIC'nin, tek seviyeli iki kategorili karma MTK modellerinin seçiminde AIC'den daha doğru olduğunu öne sürmüşlerdir. Bu çalışmalar doğrultusunda Sen ve diğerleri (2018), BIC'nin çok düzeyli karma Rasch modellerinin seçiminde diğer indekslerden daha iyi performans gösterdiğini belirtmişler. Bu nedenle, BIC bu çalışmada model seçiminde ana indeks olarak kullanılmıştır.

Sonuç ve Tartışma

Verilerin analizinde çok düzeyli karma MTK modeli kullanıldığından, analizler yapılmadan önce veri setinin hiyerarşik yapısı sınıf içi korelasyon (ICC; Raudenbush & Bryk, 2002) değeri hesaplanarak incelenmiştir. Çok düzeyli bir Rasch modeli, *lmer* fonksiyonu kullanılarak doğrusal karma etkiler modeli yaklaşımına dayalı olarak tahmin edilmiştir (Bates & Debroy, 2004). ICC değeri .578 olarak kestirilmiştir, bu da okul düzeyinin toplam varyansın %57.8'ini açıklayabileceğini gösteriyor. Daha önce de belirtildiği gibi, aynı veri seti ile 16 farklı model analiz edilmiştir. Bu 16 model için model uyum istatistikleri Tablo 1'de sunulmuştur. Tablo 1'de gösterildiği gibi, CB1C4 (1 okul düzeyinde ve 4 öğrenci düzeyinde örtük sınıf) ve CB3C4 en küçük AIC değerlerine sahipken, CB1C4 ve CB2C4 en küçük BIC ve SABIC değerlerine sahiptir. Sen ve diğerleri (2018), BIC'nin çok düzeyli karma Rasch modellerinin seçiminde daha doğru olduğunu öne sürmüştür. Bu nedenle, bu sonuçlar ışığında, bu gerçek verilerin arkasındaki heterojenliğin CB1C4 modeli ile açıklanabileceği sonucuna varılmıştır.

Bu çalışmada, büyük ölçekli bir test veri setine uygulanmış çok düzeyli bir karma MTK modeli sunulmuştur. Önerilen model, bir MTK modeli, örtük bir sınıf modeli ve çok düzeyli bir modelin bir kombinasyonudur. Bu farklı tekniklerin avantajlarını birleştirmek, araştırmacılara kavramı geniş bir şekilde anlamalarını sağlar. Bireysel düzeyde yapılan analiz, kişinin duruşunun bireysel seviyenin bir ürünü olduğunu varsayar. Ancak bir sınıftaki bireyler birbirlerini etkileyebilir, bu da onları nicel olarak karşılaştırılabilir kılar.

Bu çalışmada ilk olarak, kümeler arası varyansın toplam varyansa oranını görmek için ICC (Raudenbush & Bryk, 2002) değeri hesaplanmıştır. Bu, hiyerarşi tarafından hesaplanan toplam varyans oranını görmek için yapıldı. Daha sonra model uyumu açısından farklı modeller analiz edilmiştir. Bilgi kriteri indekslerine dayanarak, 16 alternatif model arasından en düşük uyum indeksine dayalı olan model seçilmiştir. Vermunt (2008) çalışmasına benzer şekilde, okullar arasında ortalama örtük yeteneklerde farklılıklar olduğu bulundu. Bununla birlikte, bir öğrencinin yeteneği kontrol edildiğinde, okullar arasında bireysel madde performanslarında hiçbir fark yoktu. Bu noktada, ortak değişkenleri içeren ayrıntılı bir analiz, okullar arasında ortalama örtük yeteneklerde neden farklılıklar olduğu sorusuna cevap verebilir. Ek olarak, bu çalışmada kullanılan yazılım daha karmaşık modelleri tahmin edebilir, bu model olası pratik problemlere dikkat çekerken hem öğrenci hem de okul düzeyinde sürekli ve kategorik örtük değişkenler ekleyerek genişletilebilir.

Önerilen model, veriler çok düzeyli olduğunda eğitim araştırmacıları için yararlı olabilir. Dahası, veri kümeleriyle ilgili heterojenlikle ilgili endişeler varsa, geçerlilik büyük ölçekli değerlendirme verileri kullanan kültürler arası çalışmalar için ana konu olduğunda bu modeller kullanışlı olabilir. Ayrıca, araştırmacıların asıl ilgi alanı olası örtük yapıların popülasyon içinde aynı ölçme modelini paylaştığını araştırmak olduğunda model yetersiz olabilir. Önerilen modelin temel avantajı, hiyerarşik sınıfla birlikte kişi düzeyinde ölçme sınıfını aynı anda çıkarabilmesidir.

Çok düzeyli karma MTK modelleri, araştırmacılar arasında daha popüler hale gelmektedir. Modeli kullanan çalışmaların, çok düzeyli karma MTK modellerinin bazı gereksinimlerini dikkate alması önerilir. Örneklem büyüklüğü gereksinimi, araştırmacılar için ana endişelerden biridir. Bu, temel olarak iki blok üzerine inşa edilmiştir: her seviyede gerekli olan madde sayısı ve örneklem boyutları. Simülasyon çalışmaları, $n = 5$ ila 30 kişi düzeyinde birim ve 30 ila 500 hiyerarşik düzeyde birim gerektiğini göstermiştir (Lukočienė, Varriale, & Vermunt, 2010).

Bu makale, çok düzeyli karma MTK modelinin genel bir taslağını sunar. Bu çalışmada sunulan yaklaşımın birçok teorik ve metodolojik avantajı vardır. Çok düzeyli karma MTK, örtük sınıf modelleri ve tek bir model altında ölçüm konularını ele alır. Sonuç olarak, model, araştırmacıların verilerdeki örtük yapılardan şüphelendikleri durumlarda, veriler hiyerarşik olduğunda ve kültürler arası karşılaştırmalara ihtiyaç olduğunda da kullanılabilir. Sonuçlar, bu öğrenci ve okul düzeyindeki sınıfların yorumlanabilir olduğunu ve farklı örtük yetenek yapılarının bireyler ve okullar arasında nasıl yayıldığını benzersiz bir şekilde açıkladığını göstermiştir.

Appendix. Mplus Syntax for Final Model (CB1C4)

TITLE: This is an example of a two-level mixture IRT model with one between-level class and four within-level classes

```
VARIABLE: NAMES ARE u1-u20 clus;
          USEVARIABLES = u1-u20;
          CATEGORICAL = u1-u20;
          CLASSES = cb(1) c(4);
          BETWEEN = cb;
          CLUSTER = clus;
DATA: FILE = ALLCOMBINEDMPLUS.txt;
ANALYSIS: TYPE = TWOLEVEL MIXTURE;
          ALGORITHM = INTEGRATION;
          PROCESSORS = 2;
MODEL:
    % WITHIN%
    % OVERALL%
    f BY y1-y20;

    %cb#1.c#1%
    f BY y1-y20*;
    [f@0];f@1;

    %cb#1.c#2%
    f BY y1-y20*;
    [f@0];f@1;

    %cb#1.c#3%
    f BY y1-y20*;
    [f@0];f@1;

    %cb#1.c#4%
    f BY y1-y20*;
    [f@0];f@1;

    %BETWEEN%
    % OVERALL%
    fb BY y1-y20;
    fb@1;

    %cb#1.c#1%
    fb BY y1-y20*;
    [y1$1-y20$1];
    [fb@0];

    %cb#1.c#2%
```

fb BY y1-y20*;
[y1\$1-y20\$1];
[fb@0];

%cb#1.c#3%
fb BY y1-y20*;
[y1\$1-y20\$1];
[fb@0];

%cb#1.c#4%
fb BY y1-y20*;
[y1\$1-y20\$1];
[fb@0];

SAVEDATA: file is cb1c4.txt; SAVE IS FSCORES;
OUTPUT: TECH1 TECH8;