KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: March 30, 2016
Revision received: January 24, 2017
Accepted: March 24, 2016
OnlineFirst: April 25, 2017

www.estp.com.tr
DOI 10.12738/estp.2017.4.0250 ♦ August 2017 ♦ 17(4) ♦ 1165–1177

Research Article

Parameter Recovery for the 1-P HGLLM with Non-Normally Distributed Level-3 Residuals*

Yusuf Kara[1]
Anadolu University

Akihito Kamata[2]
Southern Methodist University

Abstract

A multilevel Rash model using a hierarchical generalized linear model is one approach to multilevel item response theory (IRT) modeling and is referred to as a one-parameter hierarchical generalized linear logistic model (1-P HGLLM). Although it has the flexibility to model nested structure of data with covariates, the model assumes the normality of the residuals (i.e., abilities) at all its levels. However, in real-world datasets, the normality assumption of the residuals may not always be sound. This study investigated the parameter recovery characteristics for the 1-P HGLLM when the normality assumption of higher-level residuals is violated. Under a three-level 1-P HGLLM, two separate simulation studies were conducted with skewed and uniformly distributed level-3 residuals. Results from both simulation studies showed that there was not a dramatic effect of the non-normal level-3 residuals on the parameter estimations. Suggestions for further research were also provided in the discussion section.

Keywords

Multilevel IRT • Hierarchical generalized linear model (HGLM) • Hierarchical measurement model • Normality violation • Parameter recovery

* The first author was supported by the Scientific and Technological Research Council of Turkey as part of 2214-A scholarship program.
An earlier version was presented at the "3rd Annual Texas Universities Educational Statistics and Psychometrics Meeting," a conference held at the University of Texas at Austin, USA, March 2015.

1 **Correspondence to:** Yusuf Kara, Department of Educational Measurement and Evaluation, Anadolu University, Eskisehir, Turkey. Email: yusufkara@anadolu.edu.tr

2 Department of Education Policy & Leadership and Department of Psychology, Southern Methodist University, Dallas, TX, USA. Email: akamata@smu.edu

Citation: Kara, Y., & Kamata, A. (2017). Parameter recovery for the 1-P HGLLM with non-normally distributed level-3 residuals. *Educational Sciences: Theory & Practice, 17*, 1165–1177. http://dx.doi.org/10.12738/estp.2017.4.0250

In educational research settings, it is common for studies to have hierarchically structured data. With such a data structure, for example, students are nested in classrooms and classrooms are nested in schools. Ignoring the nested structure of these data is equivalent to ignoring the dependency between observations within the same clusters, such as classrooms and schools. Many studies (e.g., Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) have underlined the fact that inefficient estimation of the parameters and underestimation of their standard errors occurs when this dependency is ignored. Hence, to avoid this deficiency, it is recommended to employ a modeling approach that incorporates nested data structures, such as the hierarchical linear model (HLM).

Two-step analysis is a common approach used when one is interested in estimating a dependent variable to be predicted by covariates from item response data. While abilities are estimated using an IRT model in the first step, in the second step, estimated ability parameters are used as a dependent variable in a linear model, such as HLM, if the data have a nested structure. Despite the fact that the employment of HLM in this condition seemingly gives unbiased standard error estimates compared to a single-level multiple regression, there are other potential problems related to this approach. One is that estimates may be inaccurate due to measurement error in ability estimates being ignored (Fox & Glas, 2001; Kamata, 1998). Because a one-step analysis inherently incorporates measurement error in ability estimates, the use of a one-step multilevel IRT model approach is expected to overcome this shortcoming.

Some modeling approaches have been proposed for multilevel IRT models both for dichotomous and polytomous data (e.g., Fox 2001; Kamata, 1998; Maier, 2001; Williams, 2003). Among these, Kamata (1998) generalized the Rasch model as a multilevel model with a hierarchical generalized linear model (HGLM) framework. He termed this modeling approach the one-parameter hierarchical generalized linear logistic model (1-P HGLLM). With 1-P HGLLM, it is possible to incorporate person-level covariates in the model, as well as extending the model to the third level and taking cluster level covariates. This functionality of the 1-P HGLLM gives researchers the opportunity to extend it to various psychometric analyses, including conventional differential item functioning (DIF) analysis (Chu, 2002; Kamata, 1998), random-effect DIF analysis, which assumes DIF magnitude varies between clusters (Binici, 2007; Kamata & Binici, 2003), and cross-level two-way DIF analysis (Patarapichayatham, Kamata, & Kanjanawasee, 2009, 2012) where multiple sources of DIF are at different levels (e.g., person and school). A brief overview of the 1-P HGLLM is presented in the next section.

One-Parameter Hierarchical Generalized Linear Logistic Model

Multilevel generalization of the Rasch model can be based on the HGLM framework. The logit link function and a Bernoulli sampling model are employed in this modeling approach.

Two-level model. The formulation of the unconditional two-level 1-P HGLLM can be mathematized as

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_i I_i + \theta_j^{(2)}, \quad (1)$$

where γ_i is the difficulty parameter of the item i, I_i is the indicator variable for item i, and $\theta_j^{(2)}$ is the ability parameter of person j. Ability parameters, considered to be residuals in the HGLM framework, are assumed to be normally distributed with a mean of 0 and a specific variance value of τ_j.

Three-level model. In the three-level formulation of the 1-P HGLLM, there is an additional subscript k that represents the third level of the model, such as schools. The three-level unconditional model equation is

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \gamma_i I_i + \theta_{jk}^{(2)} + \theta_k^{(3)}, \quad (2)$$

where $\theta_{jk}^{(2)}$ refers to person level residuals and $\theta_k^{(3)}$ to cluster level residuals. As a result, $\theta_{jk}^{(2)} + \theta_k^{(3)}$ is equivalent to person-level abilities $\theta_j^{(2)}$ in the two-level model. With the three-level formulation, in addition to the normality assumption for person-level residuals, cluster-level residuals are also assumed to be normally distributed with a mean of 0 and a specific variance of τ_k.

Purpose of the Study

Under hierarchical linear modeling, all residuals are assumed to have normal distributions. This assumption holds for $\theta_{jk}^{(2)}$ and $\theta_k^{(3)}$ for the 1-P HGLLM. In real-world conditions, however, it is not uncommon to have non-normal residuals. For example, in a study conducted by Micceri (1989), it was seen that none of the 400 distributions of latent and observed variables taken from real educational datasets met the assumption of normality. Moreover, as was addressed in Sass, Schmitt, and Walker (2008), it is possible to have non-normal ability distributions when the individuals are not sampled randomly from the population or when the abilities are estimated from extremely easy/difficult tests. As such, investigation of the violation of the normality assumption using 1-P HGLLM will help inform researchers and practitioners of model behaviors under these conditions.

Studies have investigated the non-normal distribution condition of the residuals in the forms of multilevel models. For example, Maas and Hox (2004a; 2004b) reported the negative effects of these conditions especially on the estimation of random-effect variance parameters of the HLM. However, these effects have not been extensively investigated using multilevel IRT models, apart from in a few studies. In one, Moyer (2013) studied the effects of normality violations with two-level 1-P HGLLM. In his study, he considered the two-level model and hence suggested that the effects of the normality violations be investigated with a three-level model. Schmitt (2007) also suggested that the robustness of 1-P HGLLM to normality violations of between- and within-level residuals be evaluated. Furthermore, Dowling (2006) investigated the effects of non-normally distributed level-3 residuals with multilevel IRT in the framework proposed by Fox (2001). She considered three non-normal distributions, which were Student's *t*, gamma, and the bimodal mixture of normal. Although a gamma distribution was used to generate skewed residuals, her study included only one degree of skewness as a condition. Thus, this study investigated the effects of the violations of normality assumptions with various skewed and uniform distributions. Furthermore, instead of using the Bayesian estimation, which is already known to be more efficient for non-parametric conditions, the performance of the maximum likelihood (ML) estimation that is widely used with multilevel IRT models was evaluated.

Method

Simulation Conditions and Data Generation

Two simulation studies were conducted by generating item response data based on the 3-level 1-P HGLLM (Equation 2). For the first study, skewed level-3 residuals were generated from five beta distributions with differently shaped parameters. All of these distributions were negatively skewed with degrees of skewness ranging from −.40 to −2.35. To meet the assumption of zero mean for the residuals, generated beta variates were linearly transformed in such a way that the mean of the distribution would be zero, with the transformation maintaining the degrees of skewness. In the second study, level-3 residuals were generated from various uniform distributions. The aim of the second study was to mimic a condition in which cluster abilities are distributed as a flatter distribution than normal distribution. Thus, we could expect to observe more extremely low- and high-cluster-ability parameters than normal distribution. Uniformly distributed level-3 residuals were also linearly transformed to have a zero mean. For both studies, the level-2 residuals were generated from a normal distribution with a mean of 0 and a variance of .80 for all conditions. This specific value was chosen to control the intra-class correlation (ICC) and the total variance of the ability parameters (see next paragraph).

In both studies, a condition with normally distributed level-3 residuals was included to compare the results from non-normal residual conditions. The number of clusters and the cluster size were set to 100 and 50, respectively, based on Maas and Hox (2004a), who report that these numbers are generally sufficient to efficiently estimate both fixed and random parameters. Furthermore, the ICC value was set to .20, which can be considered a medium clustering effect (Dowling, 2006) and is commonly found in multilevel data in educational research settings (Hox & Maas, 2001). To satisfy the ICC value, randomly generated and linearly transformed level-3 residuals were further linearly transformed to have a variance of .20. A larger variance was assumed for person-level τ_j, while a total variance of person-level and cluster-level residual parameters was fixed at 1 $\tau_j + \tau_k = 1$. This is analogous to fixing the variance of ability parameters to 1, which is a common practice in many IRT modeling applications. The number of items was set to 21 in both studies. The difficulty parameters of the items were set between −2.5 to 2.5 at .25 increments.

Analyses

In all, 50 replications of data generation and analysis were performed for each condition in each study. Means and standard deviations of the correlation coefficients between the true and estimated item difficulty parameters, as well as residual values, were calculated for all 50 replications. Additionally, bias, root mean squared error (RMSE) and standard error (SE) were calculated for the variance parameters of both levels.

To obtain a detailed evaluation of parameter recovery, RMSE, SE, and bias for logit values of all item responses were calculated and averaged across items. In this way, the RMSE, SE, and bias values based on logits were computed for each student. For a more concise report, these values were then averaged for each of the predefined five ability intervals, which were (−∞, −1.95), (−1.95, −0.65), (−0.65, 0.65), (0.65, 1.95], and (1.95, ∞). It was made certain to cover at least 100 students in the extreme intervals (especially for the leftmost interval due to negative skewness) when specifying the limits. After some trials, −1.95 was specified as the proper negative extreme and other intervals were shaped by using 1.30 increments from this value. Consequently, the means and standard deviations of RMSE, SE, and bias were computed for each of the five ability intervals. All analyses were conducted using Mplus 7.11 (Muthén & Muthén, 1998/2012), employing ML with the robust standard errors (MLR) estimation technique.

Findings

Skewed Distributions

Results for the first simulation study are summarized in Table 1 and Table 2. As can be seen in Table 1, the means of the correlations between true and estimated difficulty parameters were close to 1.0 for all conditions. Standard deviation values for these correlations were very low, indicating that large correlation values between the true and estimated difficulty parameters was consistently obtained across 50 replications.

Table 1
Results for the Recovery of the Difficulties, Person and Cluster Level Residuals and Variance Parameters with Level-3 Skewed Distributions

		Distributions / Skewness Values					
Parameters	Results	Normal 0.0118	Beta(14, 5) −0.4803	Beta(14, 3) −0.6939	Beta(14, 2) −0.9493	Beta(14, 1) −1.3009	Beta(14, .4) −2.2220
Item Difficulties	Cor. mean	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997
	Cor. sd	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Person Level Residuals	Cor. mean	0.8403	0.8416	0.8406	0.8412	0.8407	0.8413
	Cor. sd	0.0041	0.0045	0.0034	0.0045	0.0043	0.0042
Cluster Level Residuals	Cor. mean	0.9439	0.9491	0.9492	0.9493	0.9484	0.9498
	Cor. sd	0.0099	0.0087	0.0084	0.0089	0.0103	0.0079
Within Cluster Variance	Bias	0.0597	0.0631	0.0669	0.0694	0.0666	0.0688
	RMSE	0.0631	0.0656	0.0695	0.0723	0.0700	0.0722
	SE	0.0204	0.0178	0.0186	0.0204	0.0218	0.0216
Between Cluster Variance	Bias	0.0060	0.0075	0.0042	0.0062	0.0059	0.0019
	RMSE	0.0164	0.0164	0.0150	0.0168	0.0145	0.0115
	SE	0.0153	0.0146	0.0144	0.0156	0.0133	0.0113

The means of the correlation coefficients between person-level residuals (abilities) were similar across all conditions, at approximately .84. They differed only at the third decimal place. For cluster-level residuals (abilities), mean correlation values were greater than those from the person level, exceeding .94 in all conditions. In addition, the correlations for skewed distributions were slightly greater than the normal distribution, which was unexpected.

For the person-level variance parameter, bias and RMSE values were higher with skewed distributions. However, there was no systematic increasing trend depending on the degree of skewness for all the bias, RMSE, and SE (see Figure 1). Additionally, SE values for the first two skewed distributions were found to be slightly greater than those for the normal distribution.

All bias, RMSE, and SE values for cluster-level variance were quite low (bias $< .01$, RMSE $< .02$, and SE $< .02$). On the other hand, it is worth emphasizing that there was no systematic increasing trend depending on the degree of skewness for cluster-level variance parameters. Moreover, bias, RMSE, and SE had the lowest values with the most skewed distribution condition, which went against our expectation.

When the results for logits were inspected, it was found that the absolute values of the bias means were generally greater for the two ability intervals in tails than the three middle intervals (see Table 2). In addition, logit values were overestimated for those intervals with less absolute ability, while they were underestimated for those intervals with greater absolute ability values. For the middle ability interval, bias values were less under normal distribution conditions than under all other skewed distribution conditions. This being said, however, no clear increasing trend was observed for these values relative to the degree of skewness.

Table 2
Results for the Recovery of the Logits with Level-3 Skewed Distributions

		Distributions / Skewness Values					
Results	Ability intervals	Normal 0.0118	Beta(14, 5) −0.4803	Beta(14, 3) −0.6939	Beta(14, 2) −0.9493	Beta(14, 1) −1.3009	Beta(14, .4) −2.2220
Bias Means	(−∞,−1.95]	0.5537	0.5156	0.5488	0.5177	0.5301	0.4248
	(−1.95, −0.65]	0.2224	0.2253	0.2138	0.2160	0.2196	0.2447
	(−0.65, 0.65]	0.0014	0.0108	0.0069	0.0124	0.0114	0.0162
	(0.65, 1.95]	−0.2281	−0.2375	−0.2258	−0.2335	−0.2357	−0.2412
	(1.95, ∞)	−0.5318	−0.5754	−0.5718	−0.5648	−0.5799	−0.5991
Bias Sd’s	(−∞,−1.95]	0.1782	0.2106	0.1880	0.1955	0.2107	0.3223
	(−1.95, −0.65]	0.1378	0.1376	0.1449	0.1489	0.1502	0.1471
	(−0.65, 0.65]	0.1318	0.1273	0.1317	0.1272	0.1263	0.1191
	(0.65, 1.95]	0.1379	0.1310	0.1259	0.1205	0.1280	0.1228
	(1.95, ∞)	0.1785	0.1745	0.1752	0.1484	0.1600	0.1471
RMSE Means	(−∞,−1.95]	0.6856	0.6657	0.6866	0.6644	0.6736	0.6478
	(−1.95, −0.65]	0.4818	0.4844	0.4824	0.4858	0.4860	0.4958
	(−0.65, 0.65]	0.4301	0.4292	0.4309	0.4316	0.4304	0.4275
	(0.65, 1.95]	0.4832	0.4867	0.4814	0.4828	0.4850	0.4865
	(1.95, ∞)	0.6718	0.7025	0.7018	0.6914	0.7080	0.7159
RMSE Sd’s	(−∞,−1.95]	0.1441	0.1517	0.1481	0.1481	0.1580	0.1809
	(−1.95, −0.65]	0.0739	0.0722	0.0744	0.0728	0.0754	0.0711
	(−0.65, 0.65]	0.0472	0.0465	0.0464	0.0468	0.0478	0.0456
	(0.65, 1.95]	0.0725	0.0754	0.0704	0.0696	0.0732	0.0724
	(1.95, ∞)	0.1361	0.1413	0.1446	0.1232	0.1340	0.1237
SE Means	(−∞,−1.95]	0.3883	0.3925	0.3938	0.3945	0.3899	0.4071
	(−1.95, −0.65]	0.4091	0.4103	0.4119	0.4131	0.4113	0.4094
	(−0.65, 0.65]	0.4100	0.4103	0.4108	0.4127	0.4118	0.4107
	(0.65, 1.95]	0.4075	0.4090	0.4101	0.4089	0.4084	0.4085
	(1.95, ∞)	0.3914	0.3879	0.3925	0.3882	0.3943	0.3816
SE Sd’s	(−∞,−1.95]	0.0411	0.0451	0.0431	0.0395	0.0383	0.0523
	(−1.95, −0.65]	0.0425	0.0425	0.0423	0.0418	0.0436	0.0413
	(−0.65, 0.65]	0.0410	0.0417	0.0411	0.0426	0.0428	0.0413
	(0.65, 1.95]	0.0399	0.0414	0.0412	0.0410	0.0420	0.0418
	(1.95, ∞)	0.0444	0.0402	0.0419	0.0396	0.0439	0.0421

RMSE mean values were found to be considerably greater than .05 in all distribution conditions. The pattern of these values is similar to the pattern of the bias means: greater at the extremes and less in the middle intervals. SE mean values were less at the extremes and greater at the middle, in contrast to the bias and RMSE means. This was perhaps due to the shrinkage of the expected a posteriori (EAP) estimators for the residuals. However, it is worth noting that these differences were not large, and SE values did not fluctuate as much between intervals as bias and RMSE means did. Last, RMSE mean values were generally higher for skewed distributions than for normal distribution, especially at the extreme ability intervals. However, there was no clear trend in these values relative to the degree of skewness.

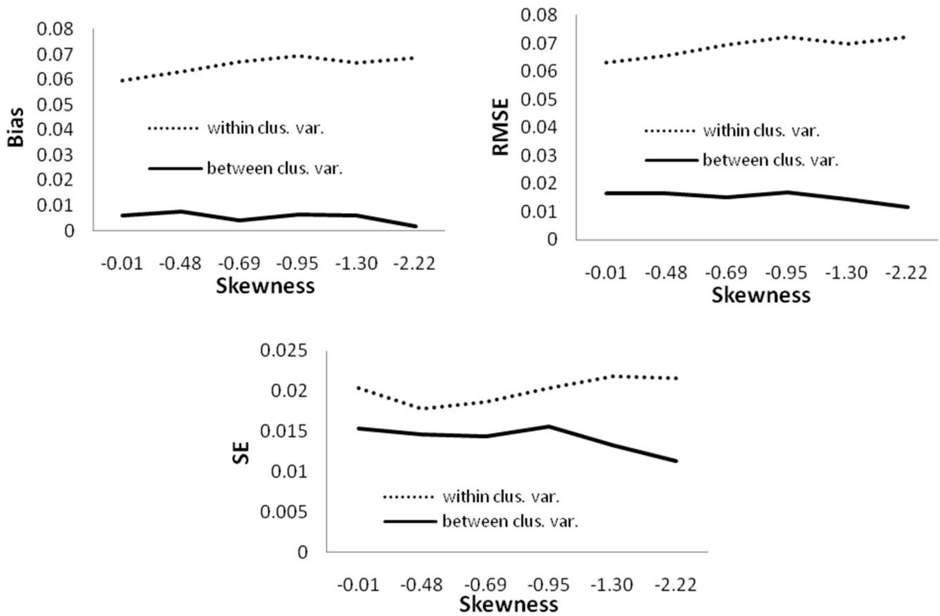

Figure 1. Bias, RMSE and SE trends for between and within-cluster variances.

Uniform Distributions

Table 3
Results for the Recovery of the Difficulties, Person and Cluster Level Residuals and Variance Parameters with Level-3 Uniform Distributions

		Distributions					
Parameters	Results	Normal	Unif (−1, 1)	Unif (−1.5, 1.5)	Unif (−2, 2)	Unif (−2.5, 2.5)	Unif (−3.5, 3.5)
Item Difficulties	Cor. mean	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997
	Cor. sd	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Person Level Residuals	Cor. mean	0.8401	0.8414	0.8399	0.8409	0.8413	0.8417
	Cor. sd	0.0041	0.0044	0.0044	0.0039	0.0041	0.0045
Cluster Level Residuals	Cor. mean	0.9471	0.9512	0.9476	0.9492	0.9492	0.9483
	Cor. sd	0.0075	0.0075	0.0070	0.0083	0.0080	0.0060
Within Cluster Variance	Bias	0.0600	0.0658	0.0662	0.0640	0.0672	0.0688
	RMSE	0.0634	0.0687	0.0697	0.0665	0.0698	0.0715
	SE	0.0203	0.0200	0.0218	0.0183	0.0190	0.0192
Between Cluster Variance	Bias	0.0065	0.0090	0.0129	0.0079	0.0099	0.0106
	RMSE	0.0162	0.0201	0.0201	0.0192	0.0179	0.0198
	SE	0.0148	0.0180	0.0154	0.0176	0.0149	0.0167

The results for the second simulation study with uniform distributions are presented in Table 3 and Table 4. As can be seen in Table 3, difficulty parameters were recovered well in all widths of the uniform distribution. The mean correlation coefficients for person-level residuals were lower than those for cluster-level residuals for all conditions. For both person- and cluster-level residuals, mean correlation coefficients were higher under uniform distribution conditions than those obtained under the normal distribution condition. Bias and RMSE values for person-level variance parameters were between .06 and .07. These values were greater for uniform distribution conditions than for the normal distribution condition. Cluster-level variance seemed to be better recovered than person-level variance, according to the bias, RMSE, and SE values, which were lower than .05 for all conditions.

In Table 4, the results were found to be similar to the first study in terms of the recoveries of logit values. Larger bias and RMSE values were obtained for extreme ability intervals, while the values for the middle intervals were less under all conditions. Conversely, SE values showed opposite behavior to bias and RMSE values, namely greater values occurred for the middle ability intervals.

Table 4
Results for the Recovery of the Logits with Level-3 Uniform Distributions

		Distributions					
Results	Ability intervals	Normal	Unif (−1, 1)	Unif (−1.5, 1.5)	Unif (−2, 2)	Unif (−2.5, 2.5)	Unif (−3.5, 3.5)
Bias Means	(−∞,−1.95]	0.5748	0.5760	0.6083	0.5716	0.5963	0.5830
	(−1.95, −0.65]	0.2275	0.2319	0.2260	0.2267	0.2281	0.2309
	(−0.65, 0.65]	−0.0064	−0.0002	−0.0030	−0.0037	−0.0076	−0.0001
	(0.65, 1.95]	−0.2258	−0.2228	−0.2313	−0.2203	−0.2283	−0.2214
	(1.95, ∞)	−0.5239	−0.5318	−0.5589	−0.5706	−0.5611	−0.5559
Bias Sd's	(−∞,−1.95]	0.1970	0.1950	0.1901	0.1934	0.1837	0.1902
	(−1.95, −0.65]	0.1367	0.1368	0.1337	0.1375	0.1330	0.1378
	(−0.65, 0.65]	0.1332	0.1313	0.1336	0.1331	0.1323	0.1309
	(0.65, 1.95]	0.1445	0.1382	0.1308	0.1317	0.1381	0.1293
	(1.95, ∞)	0.1846	0.1894	0.1973	0.1637	0.1920	0.1691
RMSE Means	(−∞,−1.95]	0.7008	0.7057	0.7285	0.7011	0.7222	0.7154
	(−1.95, −0.65]	0.4843	0.4847	0.4823	0.4845	0.4821	0.4845
	(−0.65, 0.65]	0.4302	0.4279	0.4300	0.4317	0.4286	0.4304
	(0.65, 1.95]	0.4866	0.4825	0.4851	0.4805	0.4859	0.4797
	(1.95, ∞)	0.6620	0.6681	0.6905	0.6970	0.6861	0.6888
RMSE Sd's	(−∞,−1.95]	0.1596	0.1595	0.1595	0.1586	0.1540	0.1540
	(−1.95, −0.65]	0.0732	0.0753	0.0727	0.0733	0.0722	0.0744
	(−0.65, 0.65]	0.0479	0.0470	0.0472	0.0473	0.0462	0.0455
	(0.65, 1.95]	0.0762	0.0736	0.0710	0.0721	0.0752	0.0702
	(1.95, ∞)	0.1489	0.1591	0.1640	0.1367	0.1633	0.1429
SE Means	(−∞,−1.95]	0.3814	0.3898	0.3854	0.3881	0.3930	0.3970
	(−1.95, −0.65]	0.4095	0.4080	0.4089	0.4099	0.4076	0.4078
	(−0.65, 0.65]	0.4096	0.4078	0.4092	0.4113	0.4082	0.4103
	(0.65, 1.95]	0.4110	0.4095	0.4098	0.4104	0.4109	0.4092
	(1.95, ∞)	0.3877	0.3893	0.3875	0.3877	0.3796	0.3944
SE Sd's	(−∞,−1.95]	0.0435	0.0422	0.0367	0.0448	0.0395	0.0441
	(−1.95, −0.65]	0.0414	0.0405	0.0403	0.0416	0.0419	0.0404
	(−0.65, 0.65]	0.0417	0.0416	0.0418	0.0412	0.0405	0.0413
	(0.65, 1.95]	0.0417	0.0414	0.0419	0.0420	0.0412	0.0428
	(1.95, ∞)	0.0402	0.0384	0.0479	0.0430	0.0392	0.0408

Discussion

This study explored the effects of a violation of the normal distribution of level-3 residuals with three-level 1-P HGLLM. Item response data with various skewed and uniformly distributed level-3 residuals were generated, and model parameters were estimated with the MLR estimator that is available with Mplus.

According to the results, item difficulty parameters, which are the fixed parameters of the model, were not affected from the violation of the level-3 residual normality. This result seems reasonable, because it has been reported by other researchers that the violation of the normality of the cluster level residuals does not have an effect on the estimation of fixed parameters in HLM when the ML-based estimators are employed (Maas & Hox, 2004b; Raudenbush & Bryk, 2002; Shieh, 1999). The recovery of the person-level residuals was worse than the cluster-level residuals for all distribution conditions, even with normal distribution. Moreover, since slightly higher correlations between true and estimated values were observed, estimations were found to be slightly better for non-normal distributions than for normal distributions for both levels' residuals.

Cluster-level variance parameters seemed not to be affected by the distributional violations according to the evaluations the bias, SE, and RMSE values. However, the quality of the person-level variance estimates was slightly worse compared to the cluster level variance estimates. Nevertheless, there was not a dramatic effect of the non-normal level-3 residuals on the estimation of both variance parameters. This result is parallel to a finding of Maas and Hox (2004b), who reported that non-normality seems to effect SEs of variance parameter estimates rather than point estimates of parameters in HLM. Furthermore, Dowling (2006) employed Bayesian estimation for the two-parameter multilevel IRT model and reported efficient level-2 and -three variance estimates under non-normal distributed cluster-level residuals with a moderate ICC value.

The recovery of the logits was neither accurate nor efficient in all distribution conditions. This was especially obvious for extreme ability intervals. Williams (2003) reported a similar conclusion in a study in which she evaluated polytomous multilevel data. She conducted her analysis using HLM 5 (Raudenbush, Bryk, Cheong, & Congdon, 2000), employing two-step estimation, which incorporates PQL and Bayes modal estimation, to obtain residual values. The MLR technique in Mplus, which was adopted in this study, estimates residual values (factor scores under the SEM framework) using the EAP technique. This inconsistency may be attributable to shrinkage towards the mean, which is known for Bayesian estimation of residuals.

Last, it is suggested that the effects of the non-normally distributed higher-level residuals with shapes of distributions other than those considered in this study be examined. Additionally, this study fixed the ICC values, cluster sizes, number of clusters, and variance magnitudes to modest values. It is suggested that these values be varied in future studies to monitor the combined effects of non-normal distributions using other factors.

References

Binici, S. (2007). *Random effect differential item functioning via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3282570)

Chu, K. (2002). *Equivalent group test equating with the presence of differential item functioning* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3065477)

Dowling, N. M. (2006). *An investigation of the robustness of multilevel item response theory models to violations of distributional assumptions* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3245602)

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 269–286. http://dx.doi.org/10.1007/BF02294839

Fox, J. P. (2001). *Multilevel IRT: A Bayesian perspective on estimating parameters and testing statistical hypothesis* (Doctoral dissertation, University of Twente, Enschede, The Netherlands). Retrieved from http://doc.utwente.nl/58675/1/thesis_Fox.pdf

Hox, J. J. (2010). *Multilevel anal*ysis. New York, NY: Routletge.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*, 157–174. http://dx.doi.org/10.1207/S15328007SEM0802_1

Kamata, A. (1998). *Some generalizations of the Rasch model: An Application of the hierarchical generalized linear model* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9922331)

Kamata, A., & Binici, S. (2003, July). *Random effect DIF analysis via hierarchical generalized linear modeling*. Paper presented at the biannual meeting of Psychometric Society, Sardinia.

Maas, C. J. M., & Hox, J. J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58,* 127–137. http://dx.doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427–440. Retrieved from https://dspace.library.uu.nl/bitstream/handle/1874/23144/Maas(2004)_influence%20of%20violations.pdf?sequence=1

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307–330. http://dx.doi.org/10.3102/10769986026003307

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. http://dx.doi.org/10.1037/0033-2909.105.1.156

Moyer, E. (2013). *Examining the invariance of item and person parameters estimated from multilevel measurement models when distribution of person abilities are non-normal* (Doctoral dissertation, University of Texas at Austin). Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/21290/MOYER-DISSERTATION-2013.pdf?sequence=1&isAllowed=y

Muthén, L. K., & Muthén, B. O. (1998/2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2009). Cross-level two-way differential item functioning analysis model by multilevel Rasch modeling. *Journal of Research Methodology & Cognitive Science*, *7,* 15–21. Retrieved from https://www.tci-thaijo.org/index.php/RMCS/article/view/46760

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2012). Evaluation of model selection strategies for cross-level two-way differential item functioning. *Educational and Psychological Measurement*, *72*(1), 44–51. http://dx.doi.org/10.1177/0013164411409743

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2000). *HLM 5* [Computer Program]. Chicago, IL: Scientific Software International.

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*(1), 67–89. http://dx.doi.org/10.1080/08957340701796415

Schmitt, T. A. (2007). *An evaluation of parameter recovery using the 1 parameter hierarchical generalized linear logistic model for dichotomous responses* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9922331)

Shieh, Y. Y. (1999). *An evaluation of mixed effects multilevel modeling under conditions of error term non normality* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 304528487)

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Williams, N. J. (2003). Item *and person parameter estimation using hierarchical generalized linear models and polytomous item response theory models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 305299058)