

## Using a Two-Component Normal Mixture Distribution as a Latent Distribution in Estimating Parameters of Item Response Models

Seewoo Li\*

Yonsei University

A two-component normal mixture distribution (2NM) could be used as the latent distribution of item response models when test scores follow a bimodal distribution, an idea that was originally proposed by Mislevy (1984). The purpose of this study is to examine the performance of this aforementioned 2NM-assumption item response model (2NM-IRM) under various shapes of the true latent distributions. Computer simulation techniques were used for this purpose. A reparameterization was enacted to characterize the shape of the 2NM using only three shape parameters, through which 15 total shapes of 2NM were used as simulation conditions. The performance of 2NM-IRM was compared with the normality-assumption, empirical histogram, and Davidian-curve methods. 2NM-IRM produced the most accurate results except when the true latent distribution was practically equivalent to the normal distribution. Even when some biases were present in the shape parameter estimates, 2NM-IRM was consistent at accurately identifying the overall shape of the true latent distributions and thus provided accurate item and ability parameter estimates.

*Keywords : Latent Distribution, Bimodality, Normal Mixture Distribution, Item Response Model*

---

\* 교신저자: 이시우, 연세대학교 교육학과 석사과정, [cu@yonsei.ac.kr](mailto:cu@yonsei.ac.kr)

## I. Introduction

The normal distribution has conventionally been used as a latent distribution (distribution of a latent variable) in item response models. If the true latent distribution is not normal, the normality assumption for the latent variable could yield biased estimates of item or ability parameters in item response models under marginal maximum likelihood estimation (MMLE) procedures (Kang & Lee, 2020; Kim, 2012; Mislevy, 1984; Seong, 1990; Woods & Lin, 2009; Woods & Thissen, 2006). In this case, a non-normal distribution reflecting the true latent distribution could be applied to obtain unbiased estimates of model parameters.

A bimodal distribution of test scores can be easily found in many educational and psychological settings. For example, the Myers-Briggs Type Indicator (MBTI) scale has been known to show a degree of bimodality in each subscale (Bess & Harvey, 2002; Girelli & Stake, 1993; Rytting, Ware, & Prince, 1994). In education, learning attitudes could be the cause for a bimodal distribution of academic achievement (Meyer & Land, 2006; Yadin, 2013). In particular, bimodality of students' test score distributions was observed in approximately one out of four classrooms in Sibbald's (2014) research on math assessments in 6,943 high school classrooms. The terms "math phobia", "math anxiety", and "math gap" have been used to explain bimodally-distributed math test scores (Dodd, 1992; Geist, 2015; Rossnan, 2006). In these cases, the normality assumption for latent variables in item response models can be easily violated and may cause biases in item or ability parameter estimates.

The mixture item response model (mixture-IRT) can be an option to deal with potential bimodal test score distributions. This mixture-IRT approach usually postulates a hidden structure that separates examinees into multiple groups (von Davier & Rost, 2016). This hidden structure may not be of interest for a researcher whose main purpose is to obtain item and ability parameter estimates by treating the examinees as a single group. The latent distribution estimation method, as opposed to mixture-IRT, would conceptually be more straightforward, and could be more suitable for this purpose of estimating item and ability parameters under the single group assumption.

Several types of distributions have been proposed to represent non-normal latent distributions. Most of them were skewed unimodal distributions and were applied with Rasch models (Andersen

& Madsen, 1977; Baker & Subkoviak, 1981; Kelderman, 1984; Mellenbergh & Vijn, 1981; Thissen & Mooney, 1989). Some were nonparametric or semi-nonparametric distributions which have fewer restrictions on the shapes of latent distributions (Finch & Edwards, 2016; Mislevy, 1984; Mislevy & Bock, 1985; Woods & Lin, 2009; Woods & Thissen, 2006). The normal mixture distribution was proposed by Mislevy (1984) as a parametric approach for fitting multimodal latent distributions.

The two-component normal mixture distribution (2NM), which was one of Mislevy's (1984) proposals, could be seen as a reasonable choice when test scores are bimodally distributed, as an extension of the conventional normality assumption. The combination of two normal components can formulate a variety of distribution shapes (Titterton, Smith, & Makov, 1985). Using a real dataset with a bimodal distribution of number-correct scores, Mislevy (1984) showed that the 2NM-assumption item response model (2NM-IRM) could outperform the normality-assumption item response model (normal-IRM) when it came to recovering the number-correct score distribution.

There is a need for further simulation studies to provide additional information on appropriate conditions for applying 2NM-IRM that were not addressed by Mislevy (1984). For example, the performance of 2NM-IRM could be affected by the shape of the true latent distribution. In particular, each instance of real data analysis would likely incur a different mixing proportion of the two normal components, as well as varying differences in the mean and variance between those two normal components (Lubke & Muthén, 2005; Pastor, Barron, Miller, & Davis, 2007; Pearl, 2017; Sibbald, 2014; Yadin, 2013). Fitting 2NM-IRM to such data would thereby produce several different shapes corresponding to the underlying latent distribution. In some cases, the flexibility of 2NM-IRM may yield an accurate approximation of the true latent distribution compared to the normal-IRM, which would decrease biases in item or ability parameter estimates. In other cases, however, the additional normal component in 2NM-IRM risks overfitting the data by capturing random error (Bauer & Curran, 2004), which would conversely increase the error in model parameter estimates. Results on the performance of 2NM-IRM as to these details have not been presented in the literature.

This study evaluates the performance of 2NM-IRM in estimating item and ability parameters of unidimensional two-parameter-logistic (2PL) item response models under 15 types of true latent

distributions. The overarching purpose of this study is therefore to establish 2NM-IRM as a method of estimating item and ability parameters in the bimodal case, where a better estimation of those parameters could be achieved by means of more accurate approximations of the true latent distribution. Three shape parameters of 2NM are defined to characterize its shape. Different combinations of these shape parameters yield 15 total 2NM latent distributions to be examined in the simulation study. The extent to which 2NM-IRM captures the true shape parameter values is observed, and the performance of 2NM-IRM is compared with normal-IRM, the empirical histogram method (EHM; Mislevy, 1984; Mislevy & Bock, 1985), and Davidian-curve IRT (DC-IRT; Woods & Lin, 2009). Normal-IRM is selected as a conventional item response model, and EHM and DC-IRT are selected for their flexibility to estimate bimodal distributions.

## II. Estimating Distribution Parameters of 2NM-IRM

This section is mainly a review of Mislevy's (1984) normal mixture distribution item response model and Bock and Aitkin's (1981) MMLE procedure using the EM algorithm (MMLE-EM) with some modifications on the derivation of the estimators using a log-likelihood decomposition of the aforementioned model, an original contribution of this paper.

Assuming a latent distribution as a unidimensional 2NM, the latent distribution  $g(\theta | \tau)$  can be written as a weighted sum of two normal components (Mislevy, 1984; Titterton et al., 1985):

$$g(\theta | \tau) = g(\theta | \pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \pi \times \phi_1(\theta | \mu_1, \sigma_1^2) + (1 - \pi) \times \phi_2(\theta | \mu_2, \sigma_2^2) \quad (1)$$

where  $\theta$  is the latent variable,  $\tau$  is a vector of distribution parameters,  $\pi$  is the mixing proportion ( $0 < \pi < 1$ ),  $\phi(\theta | \mu, \sigma^2)$  is a normal component with mean  $\mu$  and variance  $\sigma^2$ , and the subscripts of 1 and 2 indicate the class of the normal components.

Since the implementation of 2NM-IRM (Mislevy, 1984) follows the MMLE-EM (Bock & Aitkin, 1981), the continuous latent space is converted into discrete values by subsetting and dividing it into  $q$  intervals. Then, quadrature points can be set as the midpoints of the intervals,

with corresponding densities calculated as,

$$A(X_k) = \frac{g(X_k|\tau)}{\sum_{k=1}^q g(X_k|\tau)} \quad (2)$$

where  $\theta = X_k$  is the midpoint of the  $k$ th interval,  $k = 1, 2, \dots, q$  (Bock & Aitkin, 1981). For this discrete density  $A(X_k)$ ,  $\sum_{k=1}^q A(X_k) = 1$  is satisfied by equation (2).

The marginal log-likelihood of the model can be expressed as,

$$\log L = \sum_{j=1}^N \log \sum_{k=1}^q L_j(X_k) A(X_k). \quad (3)$$

The quantity  $L_j(X_k) = \prod_{i=1}^n \Pr(U = u_{ij} | \theta = X_k)$  is the  $j$ th examinee's likelihood for his or her total  $n$  observed item responses with  $\theta = X_k$ , where  $\Pr(U = u_{ij} | \theta = X_k)$  is the probability of  $j$ th examinee's,  $j = 1, 2, \dots, N$ ,  $i$ th item response,  $i = 1, 2, \dots, n$ , if he or she had an ability parameter of  $\theta = X_k$ .

Using the lower bound maximization technique of the EM algorithm (Harpaz & Haralick, 2006), the marginal log-likelihood in equation (3) can be decomposed as,

$$\begin{aligned} \log L &= \sum_{k=1}^q \sum_{j=1}^N \gamma_{jk} \log L_j(X_k) + \sum_{k=1}^q \sum_{j=1}^N \gamma_{jk} \log A(X_k) - \sum_{k=1}^q \sum_{j=1}^N \gamma_{jk} \log \gamma_{jk} \\ &= \log L_{item} + \log L_{distribution} - \text{constant} \end{aligned} \quad (4)$$

where  $\gamma_{jk}$  is the  $j$ th examinee's expected probability of having a latent ability parameter in the  $k$ th interval. The  $\gamma_{jk}$  is calculated in the expectation step (E-step) of the MMLE-EM procedures through Bayes' theorem (Bock & Aitkin, 1981).

The log-likelihood of the distribution parameter, which is the second term in the last line of equation (4), can be decomposed by the lower bound maximization technique of the EM algorithm (Harpaz & Haralick, 2006):

$$\log L_{distribution} \propto \sum_{k=1}^q \hat{f}_k \eta_k \log [\pi \phi_1(X_k | \mu_1, \sigma_1^2)] + \sum_{k=1}^q \hat{f}_k (1 - \eta_k) \log [(1 - \pi) \phi_2(X_k | \mu_2, \sigma_2^2)] + \text{constant} \quad (5)$$

where  $\eta_k$  is the expected probability of  $\theta = X_k$  belonging to the first normal component, and  $\hat{f}_k = N^{-1} \sum_{j=1}^N \gamma_{jk}$  is the expected number of examinees in the  $k$ th interval. To calculate these values of  $\eta_k$ , we initiate another EM algorithm within the M-step of the main EM procedure. In the E-step of this secondary EM algorithm, the  $\eta_k$  is calculated through Bayes' theorem treating distribution parameters in the vector  $\tau$  as known (Mislevy, 1984). Using these decomposed log-likelihoods shown in equations (4) and (5), solutions identical to those obtained directly from equation (3), as in Mislevy (1984), can be obtained with fewer steps.

The estimates,  $\hat{\pi}$ ,  $\hat{\mu}_1$ ,  $\hat{\mu}_2$ ,  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$ , can be obtained by differentiating equation (5) with respect to each parameter and setting the first derivative equal to zero:

$$\hat{\pi} = \frac{\sum_{k=1}^q \hat{f}_k \eta_k}{\sum_{k=1}^q \hat{f}_k} = \frac{\sum_{k=1}^q \hat{f}_k \eta_k}{N} \quad (6)$$

$$\hat{\mu}_1 = \frac{\sum_{k=1}^q \hat{f}_k \eta_k X_k}{\sum_{k=1}^q \hat{f}_k \eta_k} \quad (7)$$

$$\hat{\sigma}_1^2 = \frac{\sum_{k=1}^q \hat{f}_k \eta_k (X_k - \hat{\mu}_1)^2}{\sum_{k=1}^q \hat{f}_k \eta_k} \quad (8)$$

$$\hat{\mu}_2 = \frac{\sum_{k=1}^q \hat{f}_k (1 - \eta_k) X_k}{\sum_{k=1}^q \hat{f}_k (1 - \eta_k)} \quad (9)$$

$$\hat{\sigma}_2^2 = \frac{\sum_{k=1}^q \hat{f}_k (1 - \eta_k) (X_k - \hat{\mu}_2)^2}{\sum_{k=1}^q \hat{f}_k (1 - \eta_k)} \quad (10)$$

Solutions for  $\hat{\pi}$ ,  $\hat{\mu}_1$ , and  $\hat{\mu}_2$  are identical to those presented by Mislevy (1984). There, Mislevy (1984) pooled the variances of the two normal components for simplicity, whereas  $\sigma^{2_1}$  and  $\sigma^{2_2}$  are estimated separately in this paper. If heterogeneity of variance exists between the two normal components (Bauer & Curran, 2004; Sibbald, 2014; Yadin, 2013), unequal variances of normal components in 2NM-IRM could be more appropriate to reflect the true latent distribution.

### III. The Three Shape Parameters of 2NM

The shape of a 2NM distribution can be characterized by the mixing proportion and the differences in mean and variance between the two normal components. With a slight modification of a reparameterization method suggested by Behboodian (1970), three shape parameters of 2NM are defined and specified in this paper to reflect these characteristics. This reparameterization method separates shape parameters from the overall mean and the overall variance, which implies that changes in shape parameter values do not affect values of the overall mean and the overall variance, and vice versa. This property enables researchers to assign a scale to a latent variable by fixing the overall mean and the overall variance to some constants, which would also help stabilize the convergence of the MMLE-EM procedures.

The first is the standardized mean difference (SMD) parameter, which is defined as the standardized distance between the two means of each normal component:

$$\delta \equiv \frac{\mu_2 - \mu_1}{\bar{\sigma}} \quad (11)$$

where  $\bar{\sigma}$  is the overall standard deviation of the 2NM. Without loss of generality,  $\mu_1 < \mu_2$ , or equivalently  $\delta > 0$ , is assumed to resolve labeling indeterminacy of two normal components.

The second parameter is the mixing proportion (MP) parameter  $\pi$ , which represents the proportion contribution of the first normal component to the overall mixture distribution. The value  $(1 - \pi)$ , therefore, corresponds to the proportion contribution of the second normal component. The definition and application of the MP parameter is identical to that of Mislevy

(1984).

The final parameter is the ratio-of-standard-deviation (RSD) parameter, which is defined as ratio between the standard deviation of the second normal component and that of the first normal component:

$$\zeta \equiv \frac{\sigma_2}{\sigma_1} \quad (12)$$

The RSD parameter can represent variance homogeneity ( $\zeta=1$ ) or the degree of variance heterogeneity ( $\zeta \neq 1$ ) of the two normal components.

The MP and RSD parameters are identical to the reparameterizations suggested by Behboodian (1970), whereas the SMD parameter was modified from the quantity used in Behboodian (1970) by introducing  $\bar{\sigma}$ . In this way, the SMD parameter is not affected by the magnitudes of the overall means and overall variances.

The overall mean and variance of 2NM can also be obtained from the original 2NM distribution parameters  $\pi$ ,  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1^2$ , and  $\sigma_2^2$ :  $\bar{\mu} = \pi\mu_1 + (1-\pi)\mu_2$  and  $\bar{\sigma}^2 = \pi\sigma_1^2 + (1-\pi)\sigma_2^2 + \pi(1-\pi)(\mu_2 - \mu_1)^2$ . When item and ability parameters are estimated together, the scale of a latent variable of 2NM-IRM can be assigned by fixing  $\bar{\mu}$  and  $\bar{\sigma}$  to some constants, which is a typical way of assigning a scale to latent variables in MMLE-EM procedures (Woods, 2014). In Mislevy's (1984) implementation of 2NM-IRM, the scale of the latent variable was determined by pre-specified item parameter values, and the metric issue was not further addressed.

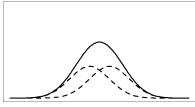
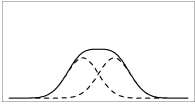
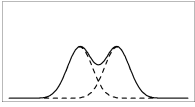
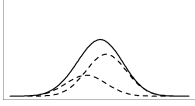
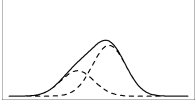
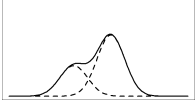
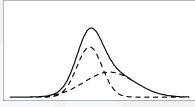
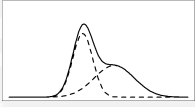
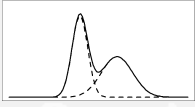
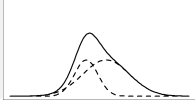
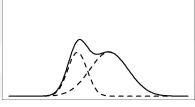
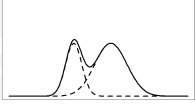
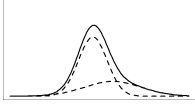
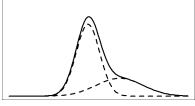
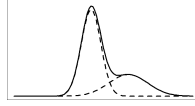
The estimates  $\hat{\delta}$  and  $\hat{\zeta}$  can be obtained simply by substituting  $\mu_1$ ,  $\mu_2$ ,  $\bar{\sigma}$ ,  $\sigma_1$ , and  $\sigma_2$  in equations (11) and (12) with their estimates. In addition, the derivation of  $\hat{\pi}$  was shown in equation (6) and can be obtained in that manner.

## IV. Methods

### 1. Simulation conditions



<Table 1> 15 types of 2NM used for true latent distributions

		Distribution type		
		$\delta$		
$\zeta$	$\pi$	0.894	1.414	1.664
		$(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 1$	$(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 2$	$(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 3$
1	$\frac{1}{2}$	 Skewness=0.0, Kurtosis=2.9	 Skewness=0.0, Kurtosis=2.5	 Skewness=0.0, Kurtosis=2.0
	$\frac{1}{3}$	 Skewness=-0.1, Kurtosis=2.9	 Skewness=-0.2, Kurtosis=2.7	 Skewness=-0.3, Kurtosis=2.4
	$\frac{1}{2}$	 Skewness=0.6, Kurtosis=3.6	 Skewness=0.6, Kurtosis=2.8	 Skewness=0.5, Kurtosis=2.1
	$\frac{1}{3}$	 Skewness=0.4, Kurtosis=3.1	 Skewness=0.3, Kurtosis=2.4	 Skewness=0.1, Kurtosis=2.1
	$\frac{2}{3}$	 Skewness=0.8, Kurtosis=4.4	 Skewness=1.0, Kurtosis=3.9	 Skewness=1.0, Kurtosis=3.4

Note. Equations regarding  $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2}$  are only satisfied when  $\pi=1/2$  and  $\zeta=1$  (Eisenberger, 1964; Titterton et al., 1985). Each dashed line is the density of a single normal component and the solid lines are the densities of 2NM. Values for skewness and kurtosis were rounded to the nearest tenth. For each graph, the range of the horizontal axis is  $(-4,4)$  and that of the vertical axis is  $(0,0.65)$ .

Performance of 2NM-IRM was evaluated under 15 types of true latent distributions created by combinations of the aforementioned shape parameters, as shown in <Table 1>. Values for the SMD parameter  $\delta$  were selected using the quantity  $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2}$ . Note that the denominator

of the quantity  $\sqrt{\sigma_1\sigma_2}$  and the overall standard deviation  $\bar{\sigma} = \sqrt{\pi\sigma_1^2 + (1-\pi)\sigma_2^2 + \pi(1-\pi)(\mu_2 - \mu_1)^2}$  are not identical. This criterion was adopted because bimodality of 2NM appears as  $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} > 2$ , given that  $\zeta = 1$  and  $\pi = 0.5$  (Eisenberger, 1964; Titterton et al., 1985). Corresponding SMD parameters for  $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 1, 2, 3$  are  $\delta = 0.894, 1.414, 1.664$ . The reason for selecting these values of 1, 2, and 3 for the simulation conditions is to consider a case in which the distribution is clearly unimodal ( $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 1$ ), a case in which the distribution is borderline bimodal ( $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 2$ ), and a case in which the distribution is clearly bimodal ( $(\mu_2 - \mu_1)/\sqrt{\sigma_1\sigma_2} = 3$ ). Values for the MP parameter  $\pi$  were selected to represent equal ( $\pi = 1/2$ ) and unequal ( $\pi = 1/3, 2/3$ ) proportions of the two normal components. Values for the RSD parameter  $\zeta$  were similarly selected to represent homogeneous variances ( $\zeta = 1$ ) and heterogeneous variances ( $\zeta = 2$ ) of the two normal components. The cases for which  $\zeta = 1$  and  $\pi = 2/3$  were excluded as they are functionally identical to those of  $\zeta = 1$  and  $\pi = 1/3$  when  $\delta$  is fixed. The overall mean and variance for each of these 15 true latent distributions were set at 0 and 1, respectively.

## 2. Data generation

Item discrimination and difficulty parameters were generated each from  $a \sim \text{Uniform}(0.8, 2.5)$  and  $b \sim N(0, 1^2)$ , with difficulty parameters truncated at  $\pm 2$ . The item parameters generated from these conditions can be considered good when the mean and variance of the overall latent distribution are 0 and 1, respectively (De Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991). The truncation of the difficulty parameters could reflect typical test settings in which items that are exceedingly difficult or easy are excluded from the test (Hambleton et al., 1991).

Following from the simulation designs of Woods and Lin (2009) and Woods and Thissen (2006), the number of examinees and items were fixed to 1000 and 25, respectively. For each of the 15 simulation conditions, 100 dichotomous item response datasets were generated along with 100 corresponding sets of 1000 ability parameters generated from the true latent distribution. A new set of item parameters was generated for each of the 100 data generation iterations, and the same item parameters were used for each of the 15 simulation conditions per iteration. All of the values were generated by built-in functions in R (R Core Team, 2021).

### 3. Analysis

For each of the 1500 datasets (i.e. 100 datasets for each of the 15 simulation conditions), unidimensional 2PL models were fitted using normal-IRM, EHM, DC-IRT, and 2NM-IRM. Normal-IRM, EHM, and 2NM-IRM were each fitted once per dataset, whereas DC-IRT was fitted ten times per dataset. The best DC-IRT was selected according to the Hannan-Quinn (HQ) criterion (Hannan & Quinn, 1979):  $HQ = -2\log L + 2p(\log(\log n))$ , where  $\log L$  is the log-likelihood of the model,  $p$  is the number of parameters estimated, and  $n$  is the number of examinees (Woods & Lin, 2009). The performance of the best DC-IRT was compared with the other methods.

The scale of the latent variable was assigned by setting the mean and variance of the latent distribution to 0 and 1, respectively. The latent space was then subset and divided to 121 equally spaced intervals and the corresponding quadrature points were located from  $-6$  to  $6$  in increments of  $0.1$ . This quadrature scheme was designed to encompass most of the possible latent ability values (Houts & Cai, 2020). In EHM, these quadrature points could change through standardization; hence, initial quadrature points were recovered at each MMLE-EM cycle by linear interpolation and extrapolation. This procedure is incorporated in ‘flexMIRT 3.62’ (Cai, 2020).

The expected *a posteriori* (EAP) scores were computed using the same quadrature scheme, and EAP scores were used as ability parameter estimates.

The convergence threshold for the MMLE-EM procedure was set to  $10^{-4}$  for the maximum item parameter change. In Fisher-scoring iterations for item parameter estimation, that convergence threshold was reduced to  $10^{-7}$ . In the secondary EM algorithm used for 2NM shape parameter estimation, the threshold was set to  $10^{-4}$  on the change of SMD parameter  $\delta$ . The maximum number of MMLE-EM iterations was set to 200 for normal-IRM, DC-IRT, and 2NM-IRM, and 1000 for EHM, as EHM tends to require more iterations (Woods & Lin, 2009).

Initial values for the item discrimination and difficulty parameters were set to 1.65 and 0, respectively, where each value is the mean of its prior distribution. The initial latent density of the first MMLE-EM cycle is the standard normal distribution. In the secondary EM algorithm for estimating shape parameters of 2NM, initial values were set to  $\delta = \sqrt{2}$ ,  $\pi = 0.5$ , and  $\zeta = 1$  so as to remove preference on skewness (positive or negative) and modality (unimodal or bimodal).

The software program ‘R 4.1.1’ (R Core Team, 2021) was used to fit item response models. R functions were created directly for the purpose of this study by referring to estimation equations from Baker and Kim (2004), and Hastie, Tibshirani, and Friedman (2009), as previously published programs for fitting 2NM-IRM could not be found. <Appendix 4> contains specific details on these functions. Normal-IRM, EHM, and 2NM-IRM were implemented using these functions. For normal-IRM and EHM, outputs from these functions were practically equivalent to those obtained from ‘flexMIRT 3.62’ (Cai, 2020); under normal and bimodal distributions, maximum differences in item and ability parameter estimates were less than  $10^{-3}$  and  $10^{-2}$ , respectively. DC-IRT was implemented using the *mirt* package (Chalmers, 2012) in R.

#### 4. Evaluation criteria

For 2NM-IRM, recovery of the 2NM shape parameters was evaluated by observing the bias of each parameter:

$$bias(\hat{\tau}) = \frac{\sum_{r=1}^{100} (\hat{\tau}_r - \tau)}{100} \quad (13)$$

where  $\tau$  is a shape parameter of 2NM ( $\delta$ ,  $\pi$ , or  $\zeta$ ), and  $\hat{\tau}_r$  is an estimate of  $\tau$  from the  $r$ th dataset. The empirical standard errors were calculated for each bias value to test their statistical significance.

The performances of these methods were evaluated with respect to recovery of (a) latent density, (b) item characteristic curve (ICC), and (c) ability parameter. The integrated squared error (ISE) was selected as an evaluation index for latent density recovery (Woods & Lin, 2009). ISE measures the discrepancy between the true latent distribution and its estimated distribution. ISE can therefore be calculated as,

$$ISE(\hat{g}) = \int \{\hat{g}(\theta) - g(\theta)\}^2 d\theta \approx \sum_{k=1}^{121} \frac{1}{10} \{\hat{g}(X_k) - g(X_k)\}^2 \quad (14)$$

where  $g(\theta)$  is the true latent density with its estimate  $\hat{g}(\theta)$ , and  $1/10$  is the width of a quadrature interval. The integration was approximated by summation using the quadrature scheme

of this study. For normal-IRM,  $\hat{g}$  was replaced with the standard normal distribution because normal-IRM does not estimate latent densities.

The root mean squared error (RMSE) measured the accuracy of the ICC and ability parameter estimates. The RMSE for the estimated ICC was calculated as,

$$RMSE(\widehat{ICC}) = \sqrt{\frac{1}{25} \sum_{i=1}^{25} \sum_{k=1}^{121} \{(\hat{P}_i(X_k) - P_i(X_k))^2 A(X_k)\}} \quad (15)$$

where  $A(X_k)$  is the discrete posterior density at  $\theta = X_k$  as in equation (2),  $P_i(X_k)$  is the probability of a correct response to the  $i$ th item for an examinee having latent ability  $\theta = X_k$ , and  $\hat{P}_i(X_k)$  is the estimate of  $P_i(X_k)$ . The RMSE for EAP scores was calculated as,

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{1000} \sum_{j=1}^{1000} (\hat{\theta}_j - \theta_j)^2} \quad (16)$$

where  $\theta_j$  is the  $j$ th examinee's ability parameter, and  $\hat{\theta}_j$  is the EAP score for the  $j$ th examinee.

To aggregate the values of the evaluation criteria from 100 datasets for each simulation condition, the median was reported for ISE, because its empirical distribution is typically skewed (Woods & Lin, 2009), and the mean was reported for RMSE. Smaller ISE and RMSE indicate better performance of the method.

## V. Results

### 1. Biases in shape parameter estimates of 2NM-IRM

We observed the bias of each shape parameter of 2NM to determine how well 2NM-IRM captured the true shape parameters under the 15 true latent 2NM distributions. Biases of shape parameter estimates obtained from 2NM-IRM are listed in <Table 2>. For the symmetric distributions ( $\zeta = 1$  and  $\pi = 0.5$ ), biases were almost equal to zero, and none of them were statistically significant.

〈Table 2〉 Biases in shape parameter estimates of 2NM-IRM

Parameter			$bias(\hat{\zeta})$	$bias(\hat{\pi})$	$bias(\hat{\delta})$
$\zeta$	$\pi$	$\delta$			
1	$\frac{1}{2}$	0.894	0.02 (0.20)	0.00 (0.02)	0.00 (0.37)
		1.414	0.00 (0.11)	0.00 (0.04)	-0.01 (0.10)
		1.664	0.00 (0.08)	0.00 (0.02)	0.00 (0.03)
	$\frac{1}{3}$	0.894	-0.04 (0.22)	<b>0.16 (0.02)</b>	-0.07 (0.39)
		1.414	-0.14 (0.18)	<b>0.14 (0.06)</b>	-0.18 (0.24)
		1.664	-0.06 (0.32)	0.06 (0.12)	-0.09 (0.22)
	$\frac{1}{2}$	0.894	0.08 (0.36)	0.01 (0.09)	0.01 (0.20)
		1.414	-0.10 (0.58)	0.04 (0.08)	0.07 (0.17)
		1.664	-0.05 (0.63)	0.01 (0.04)	0.01 (0.08)
2	$\frac{1}{3}$	0.894	-0.09 (0.65)	0.11 (0.11)	0.11 (0.24)
		1.414	-0.52 (0.53)	0.14 (0.08)	0.10 (0.08)
		1.664	<b>-0.72 (0.31)</b>	<b>0.10 (0.04)</b>	0.01 (0.04)
	$\frac{2}{3}$	0.894	0.09 (0.32)	0.00 (0.11)	0.03 (0.34)
		1.414	0.13 (0.47)	-0.02 (0.08)	-0.05 (0.29)
		1.664	0.20 (0.57)	-0.02 (0.06)	-0.07 (0.21)

Note. For each bias value, empirical standard error was presented inside the parenthesis. Statistically significant biases were emphasized with bold font, using significance level of  $\alpha=.05$ .

For the asymmetric distributions, the shape parameter estimates tended to be more biased. In cases of a small MP parameter ( $\pi=1/3$ ) and moderate or large RSD parameter ( $\zeta=1, 2$ ), four total scenarios emerged with statistically significant biases. The bias of  $\hat{\pi}$  was statistically significant in the three cases where  $\zeta=1$ ;  $\pi=1/3$ ;  $\delta=0.894, 1.414$  and  $\zeta=2$ ;  $\pi=1/3$ ;  $\delta=1.664$ . Also, the bias of  $\hat{\zeta}$  was statistically significant when  $\zeta=2$ ;  $\pi=1/3$ ;  $\delta=1.664$ . In addition, considerable but non-statistically significant bias ( $bias(\hat{\zeta})=-0.52$ ) was observed when  $\zeta=2$ ,  $\pi=1/3$ , and  $\delta=1.414$ . Across all conditions, none of the biases in SMD parameter  $\delta$  were statistically significant.

Biases in RSD and MP parameters tended to be negatively correlated. Similarly, except for two

conditions ( $\zeta = 1$ ;  $\pi = 1/2, 1/3$ ;  $\delta = 0.894$ ), individual estimates of RSD and MP parameters were negatively correlated within each simulation condition. For example, a negative correlation of  $-0.75$  was observed between  $\hat{\zeta}$  and  $\hat{\pi}$  for the condition where both  $bias(\hat{\zeta})$  and  $bias(\hat{\pi})$  were statistically significant ( $\zeta = 2$ ,  $\pi = 1/3$ , and  $\delta = 1.664$ ). These results may be attributed to the fact that the thickness of distribution tails, as well as the skewness, mainly depend on both the RSD and MP parameters, meaning that the small sample size in the tails could have been a source of the bias in the parameters.

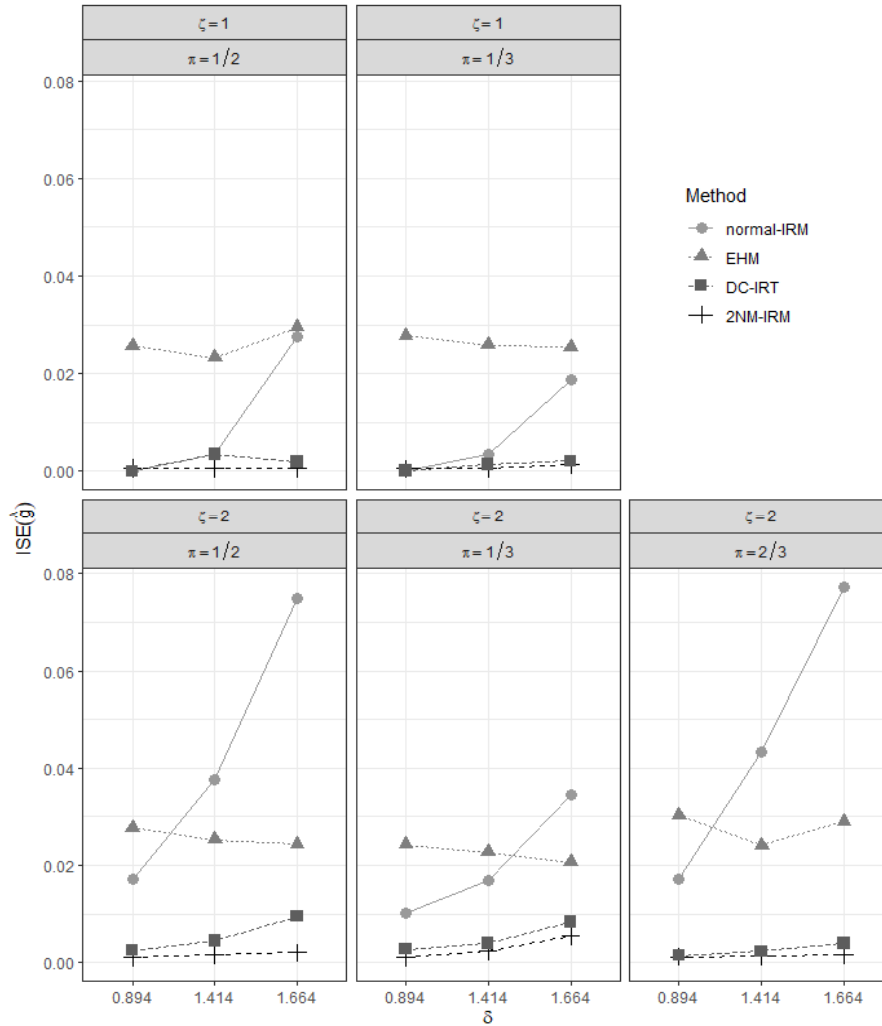
The empirical standard error of  $bias(\hat{\zeta})$  tended to be large compared to those of  $bias(\hat{\pi})$  and  $bias(\hat{\delta})$ . The empirical standard error of  $bias(\hat{\delta})$  decreased as the true parameter value of  $\delta$  increased.

## 2. Performance of 2NM-IRM compared to normal-IRM, EHM, and DC-IRT

The performance of 2NM-IRM was examined in comparison to normal-IRM, EHM, and DC-IRT. Performance of each method was evaluated with respect to latent density recovery, ICC recovery, and ability parameter recovery, with corresponding evaluation criteria  $ISE(\hat{g})$ ,  $RMSE(\widehat{ICC})$ , and  $RMSE(\hat{\theta})$ .

[Figure 1] shows ISE values. ISE values for 2NM-IRM were the smallest except for the following two conditions:  $\zeta = 1$ ;  $\pi = 1/2$ ;  $\delta = 0.894$  and  $\zeta = 1$ ;  $\pi = 1/3$ ;  $\delta = 0.894$ . In these two simulation conditions, ISE values for normal-IRM and DC-IRT were smaller than those for 2NM-IRM, and the ISE values between the true latent distributions and standard normal distribution were less than  $10^{-4}$ . These small ISE values for the normal-IRM and DC-IRT may be attributed to the overall minute differences between the true latent distributions and the standard normal distribution, as 99.5% of DC-IRT models were simplified to normal distribution models according to the HQ criterion. In contrast, 2NM-IRM may have captured random error and yielded misleading results for latent density estimation compared to models that simply assume normality.

The ISE values for EHM, DC-IRT, and 2NM-IRM were relatively consistent across simulation conditions, whereas those for normal-IRM varied according to the combination of 2NM shape parameters. Also, except for the two aforementioned conditions ( $\zeta = 1$ ;  $\pi = 1/2, 1/3$ ;  $\delta = 0.894$ )



Note. The ISE values for normal-IRM are constants, since normal-IRM does not estimate latent distributions.

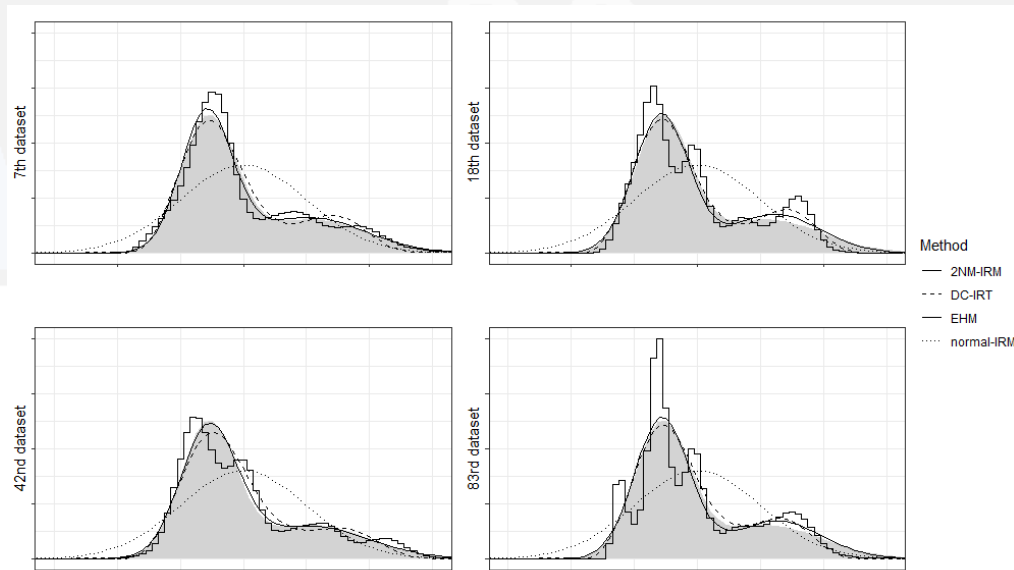
(Figure 1)  $ISE(\hat{y})$  for normal-IRM, EHM, DC-IRT, and 2NM-IRM

where the ISE values of DC-IRT were smaller than those of 2NM-IRM, the order of ISE values from EHM, DC-IRT, and 2NM-IRM was consistent; ISE values for 2NM-IRM were the smallest and those for EHM were the largest. The ISE of normal-IRM increased with the increase of SMD and RSD parameters, while the effect of the MP parameter on the ISE of normal-IRM differed with the SMD and RSD parameters. The largest ISE of 2NM-IRM was observed from the condition where both  $bias(\hat{\zeta})$  and  $bias(\hat{\pi})$  were statistically significant ( $\zeta = 2$ ,  $\pi = 1/3$ , and



$\delta = 1.664$ ).

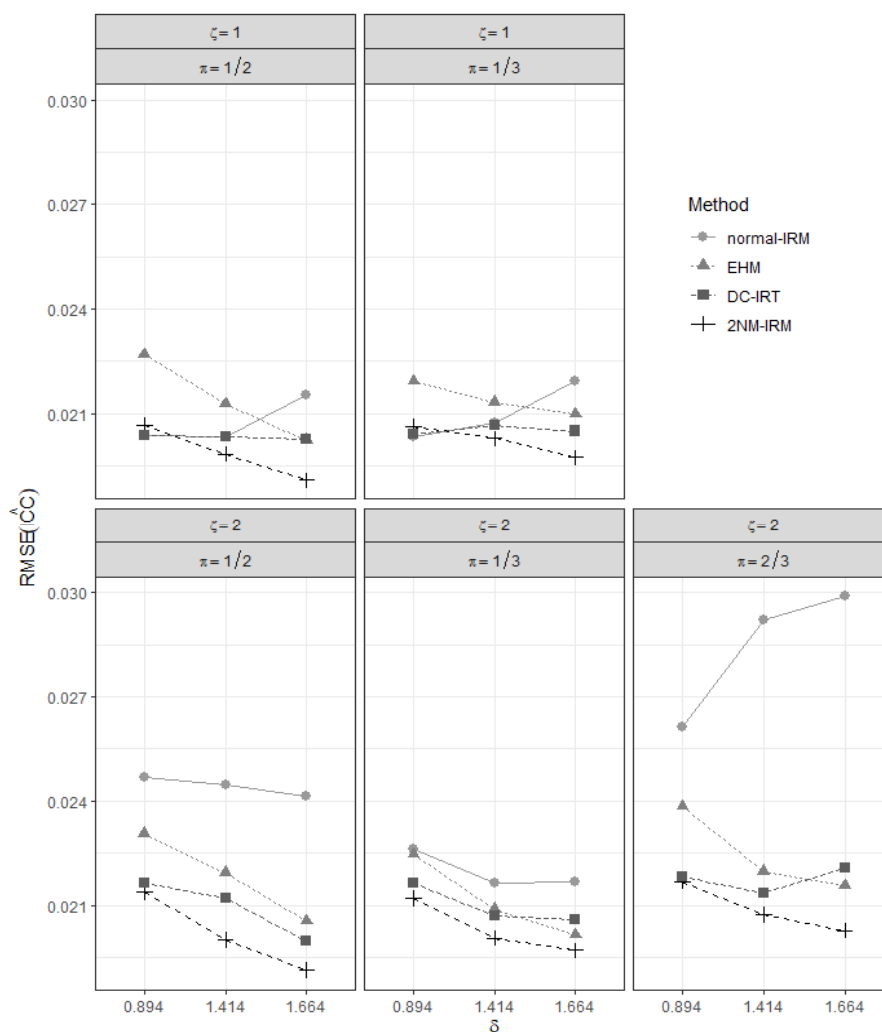
{Figure 2} visually illustrates randomly selected estimated densities from EHM, DC-IRT, and 2NM-IRM, as well as the standard normal distribution and the true latent distribution, for the simulation condition of  $\zeta = 2$ ,  $\pi = 2/3$ , and  $\delta = 1.664$ . This simulation condition was selected for visual illustration because it showed rather notable differences in performance of models with regard to  $ISE(\hat{g})$ . It can be seen that the estimated densities from 2NM-IRM (smooth solid lines) are the closest to the true distribution (shaded gray regions), with those from DC-IRT (dashed lines) being the next closest to the true distribution. The jagged densities from EHM (histogram-like solid lines) tended to overfit the data and deviated from the true distribution. In comparing {Figures 1 and 2}, it can be seen that  $ISE(\hat{g})$  was an adequate measure for assessing the similarity between the estimated densities and the true latent distribution.



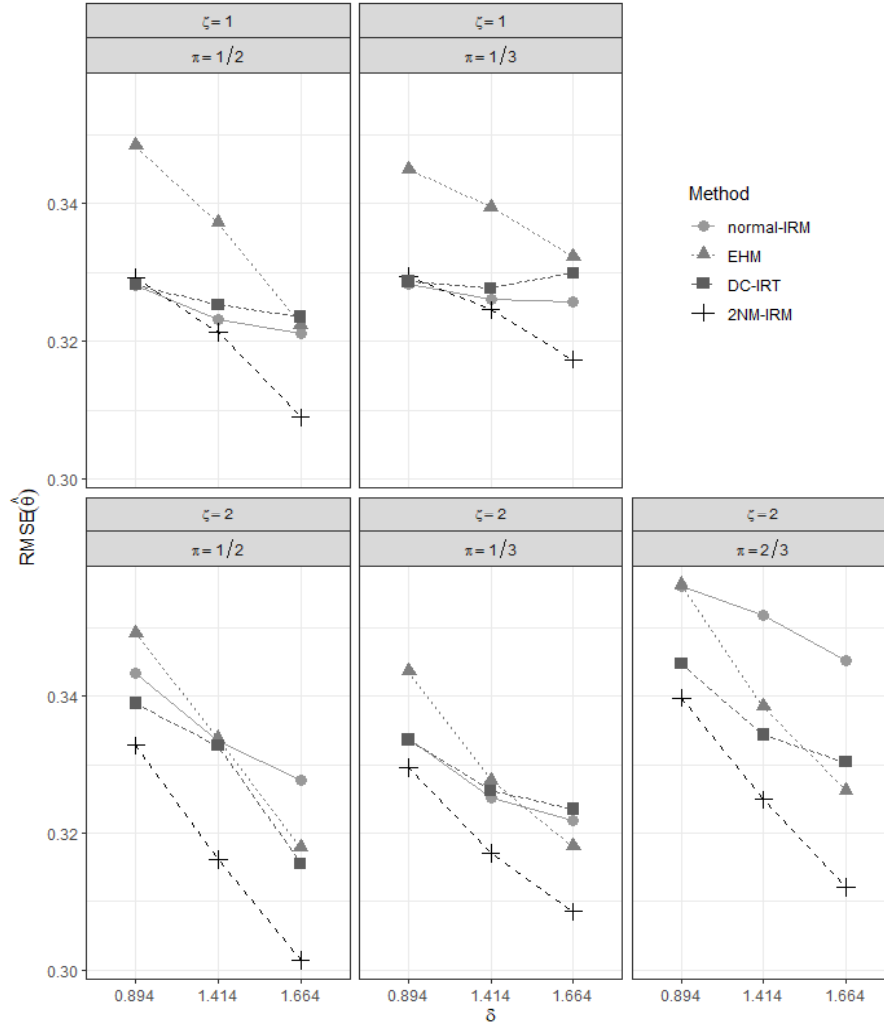
*Note.* The shaded gray regions represent the true distribution. The histogram-like solid lines represent the estimated densities from EHM, and the smooth solid lines represent the estimated densities from 2NM-IRM. The dotted lines from normal-IRM are the standard normal distribution. Among 100 datasets for the simulation condition of  $\zeta = 2$ ,  $\pi = 2/3$ , and  $\delta = 1.664$ , 7<sup>th</sup>, 18<sup>th</sup>, 42<sup>nd</sup>, and 83<sup>rd</sup> datasets were randomly selected. Densities are plotted on the  $\theta$  range of  $(-3, 3)$ .

{Figure 2} Visual illustrations of randomly selected estimated densities when  $\zeta = 2$ ,  $\pi = 2/3$ , and  $\delta = 1.664$

[Figure 3] and [Figure 4] show RMSEs for estimated ICCs and EAP scores. Similar to the results of ISE, RMSEs for estimated ICCs and EAP scores for 2NM-IRM were the smallest except for the two aforementioned conditions ( $\zeta=1$ ;  $\pi=1/2, 1/3$ ;  $\delta=0.894$ ), where the difference between the true latent distributions and the standard normal distribution was less than  $10^{-4}$ . Normal-IRM and DC-IRT yielded smaller RMSE values than 2NM-IRM for these two exceptional conditions.



(Figure 3)  $RMSE(\widehat{ICC})$  for normal-IRM, EHM, DC-IRT, and 2NM-IRM



(Figure 4)  $RMSE(\hat{\theta})$  for normal-IRM, EHM, DC-IRT, and 2NM-IRM

$RMSE(\widehat{ICC})$  and  $RMSE(\hat{\theta})$  from 2NM-IRM decreased with the increase of the SMD parameter, which can be attributed to less extreme ability values generated from the small kurtosis of true latent distributions. This pattern of decrease in RMSE with the increase of SMD parameter was more apparent in  $RMSE(\hat{\theta})$ .

EHM produced smaller  $RMSE(\widehat{ICC})$  compared to normal-IRM except for the four conditions ( $\zeta = 1$ ;  $\pi = 1/2, 1/3$ ;  $\delta = 0.894, 1.414$ ) where ISE values of normal-IRM were less than 0.01. This

tendency of better ICC recovery of EHM compared to normal-IRM is parallel to the results presented in Kim (2012). DC-IRT produced smaller  $RMSE(\widehat{ICC})$  values compared to EHM, except for the conditions of unequal proportion, large SMD and RSD parameters ( $\zeta=2$ ;  $\pi=1/3, 2/3$ ;  $\delta=1.664$ ), which reflects the results presented in Woods and Lin (2009).

## VI. Discussions and Conclusions

This study examined the performance of 2NM-IRM in estimating item and ability parameters through the accurate approximation of the true latent distribution. The 2NM-IRM approach, which was proposed by Mislevy (1984), can be a reasonable choice when the distribution of test scores is presumed bimodal. Various shapes of latent distributions can be observed by combinations of two heterogeneous latent groups (Lubke & Muthén, 2005; Pastor et al., 2007; Pearl, 2017; Sibbald, 2014; Yadin, 2013), which could affect the performance of 2NM-IRM. This study provided some information on distributional conditions for implementing 2NM-IRM and interpreting its results under 15 types of true latent distributions formulated by 2NM.

Results showed that, under the simulation design of this study, the performance of 2NM-IRM was better than the other methods in 13 conditions, and it remained acceptable for the two other conditions where the true latent distributions were close to the standard normal distribution. This would imply that the performances of 2NM-IRM ought to be satisfactory, and item and ability parameter estimates from 2NM-IRM would be fairly accurate under various types of true latent 2NM distributions. However, when the shape of the true latent 2NM closely resembles the normal distribution, normal-IRM performed better than 2NM-IRM under the simulation design of this study. In such cases, rather than truly identifying the subtle difference between the normal distribution and the true latent distribution, 2NM-IRM may have simply captured random error when estimating the latent distributions. If researchers presume that two latent groups are not sufficiently heterogeneous with respect to the latent variable  $\theta$ , cautions are needed for the implementation of 2NM-IRM.

In each simulation condition, the item response model that produced the smallest ISE value also provided the smallest RMSE values. This result supports conclusions from existing studies on

latent distributions of IRT that the proper identification of the true latent distribution plays a critical role in obtaining accurate item and ability parameter estimates (Kang & Lee, 2020; Kim, 2012; Mislevy, 1984; Seong, 1990; Woods & Lin, 2009; Woods & Thissen, 2006). The ISE values from 2NM-IRM were consistently less than 0.006, whereas the largest ISE value was 0.077 from normal-IRM. Well-identified latent distributions from 2NM-IRM could have possibly yielded decent performance of 2NM-IRM.

For most of the simulation conditions, biases in shape parameter estimates of 2NM-IRM were not statistically significant. However, results of this study suggest that if the true latent 2NM had a small MP parameter and a moderate or large RSD parameter (or, conversely, a large MP parameter and a moderate or small RSD parameter), even with small ISE values, individual estimates of the RSD and MP parameters could be biased. Biases in the estimates of RSD and MP parameters may have arisen from sparse information in the tails of the distribution, and from the interdependency of the RSD and MP parameters in determining the thickness of tails and skewness of the 2NM. In such cases, caution is required when interpreting estimated latent density using the RSD and MP parameters.

Biases in item and ability parameter estimates can deteriorate the validity of an analysis (Woods & Thissen, 2006). Selecting an appropriate item response model for a given set of data could reduce biases in item and ability parameter estimates. If a researcher has information that the distribution of test scores is bimodal, 2NM-IRM may be a better option to reflect the shape of the underlying latent distribution, as opposed to normal-IRM. In this case, more accurate item and ability parameter estimates could be obtained from 2NM-IRM.

The 2NM-IRM approach could have some advantages over DC-IRT even if their performances are similar: 2NM-IRM suffices by building only one model, while DC-IRT generally builds ten models and chooses the best among them. Fitting item response models using 2NM-IRM, as compared to DC-IRT, would thereby decrease computation time substantially. Interpreting 2NM through shape parameters would also be easier than interpreting the semi-nonparametric curve given by DC-IRT.

Cautions are needed in comparing the performance of 2NM-IRM to those of EHM and DC-IRT, as 2NM-IRM is a parametric method, and EHM and DC-IRT are nonparametric and semi-nonparametric methods. While the effectiveness of 2NM-IRM has been well established in

this study, further research is necessary to assess whether 2NM-IRM maintains these advantages when the number of items and examinees differ greatly from the setup of this study.

Further studies can be carried out to examine the robustness of 2NM-IRM in item and ability parameter estimation. Theoretically, if the true latent distribution deviates from 2NM, item or ability parameter estimates from 2NM-IRM are likely to be biased. Despite this addition of bias, the parametric nature of 2NM-IRM is likely to have lower variance than its semi-nonparametric or fully nonparametric counterparts. Likewise, the additional normal component in 2NM-IRM enhances its flexibility, thereby allowing for it to reduce bias as compared to its more traditional counterparts such as normal-IRM. Since systematic bias and variance of a model constitute the overall error (Hastie et al., 2009), 2NM-IRM may be able to perform fairly well in some situations even if the true latent distribution deviates from 2NM.

Mixtures of additional normal components, such as a three-component normal mixture distribution, could be applied in estimating latent distributions of item response models. However, whether a trimodal latent distribution is a scenario that is truly observed in practice is questionable. In addition, several issues can arise if three or more normal components are incorporated for estimating latent distributions. Interpretation of the distribution may not be straightforward, for there would be at least eight distribution parameters. Convergence of distribution parameter estimation may not be guaranteed. For instance, a variance from one of the normal components could diverge to positive infinity. Similar shapes of the latent distribution could be obtained by different sets of distribution parameters, which would complicate interpretation of the latent distribution. In such cases, cautions would be needed for interpreting a distribution by its parameters, which threatens one of the advantages of parametric assumption.

Further researches can be carried out to extend the application of 2NM-IRM. For example, performance of 2NM-IRM under the three-parameter-logistic (3PL) item response model, Samejima's (1997) graded model or Bock's (1972) nominal model could be examined. In the extension of 2NM-IRM to multidimensional item response models, several assumptions on covariance matrices of normal components could also be explored.

## References

- Andersen, E., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, 42(3), 357-374.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press.
- Baker, F. B., & Subkoviak, M. J. (1981). Analysis of test results via log-linear models. *Applied Psychological Measurement*, 5(4), 503-515.
- Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological Methods*, 9(1), 3-29.
- Behboodian, J. (1970). On the modes of a mixture of two normal distributions. *Technometrics*, 12(1), 131-139.
- Bess, T. L., & Harvey, R. J. (2002). Bimodal score distributions and the Myers-Briggs Type Indicator: fact or artifact?. *Journal of Personality Assessment*, 78(1), 176-186.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, 37(1), 29-51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443-459.
- Cai, L. (2020). flexMIRT® version 3.62: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(1), 1-29.
- De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications.
- Dodd, A. W. (1992). Insights from a math phobic. *The Mathematics Teacher*, 85(4), 296-298.
- Eisenberger, I. (1964). Genesis of bimodal distributions. *Technometrics*, 6(4), 357-363.
- Finch, H., & Edwards, J. M. (2016). Rasch model parameter estimation in the presence of a nonnormal latent trait using a nonparametric Bayesian approach. *Educational and Psychological Measurement*, 76(4), 662-684.
- Geist, E. (2015). Math anxiety and the “math gap”: How attitudes toward mathematics disadvantages students as early as preschool. *Education*, 135(3), 328-336.
- Girelli, S. A., & Stake, J. E. (1993). Bipolarity in Jungian type theory and the Myers-Briggs type indicator. *Journal of Personality Assessment*, 60(2), 290-301.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
- Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), 190-195.
- Harpaz, R., & Haralick, R. (2006). The EM algorithm as a lower bound optimization technique. *CUNY Ph. D. Program in Computer Science Technical Reports*, 1-14.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
- Houts, C. R., & Cai, L. (2020). flexMIRT® user's manual version 3.6: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.
- Kang, H., Lee, G. (2020). The influences of non-normality of a latent variable, the number of test items, and the number of examinees on IRT parameter estimation using a Davidian curve. *Journal of Educational Evaluation*, 33(2), 533-559.
- Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, 49(2), 223-245.
- Kim, S. (2012). Effects on the item parameter estimation of empirical estimation of the underlying ability distribution in IRT model parameter estimation using the BILOG-MG program. *Journal of Educational Evaluation*, 25(2), 317-336.
- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10(1), 21-39.
- Mellenbergh, G. J., & Vijn, P. (1981). The Rasch model as a loglinear model. *Applied Psychological Measurement*, 5(3), 369-376.
- Meyer, J., & Land, R. (2006). *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge*. Routledge.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49(3), 359-381.
- Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In Weiss, D. J. (Ed.). *Proceedings of the 1982 item response theory and computerized adaptive testing conference* (pp. 189-202). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Conference.
- Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, 32(1), 8-47.
- Pearl, J. (2017). Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3), 370-389.
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for



- Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Rossnan, S. (2006). Overcoming math anxiety. *Mathitudes*, 1(1), 1-4.
- Rytting, M., Ware, R., & Prince, R. A. (1994). Bimodal distributions in a sample of CEOs: Validating evidence for the MBTI. *Journal of Psychological Type*, 31, 16-23.
- Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, 14(3), 299-311.
- Sibbald, T. (2014). Occurrence of bimodal classroom achievement in Ontario. *Alberta Journal of Educational Research*, 60(1), 221-225.
- Thissen, D., & Mooney, J. A. (1989). Loglinear item response models, with applications to data from social surveys. *Sociological Methodology*, 19, 299-330.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.
- von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In *Handbook of item response theory* (Vol. 1, pp. 421-434). Chapman and Hall/CRC.
- Woods, C. M. (2014). Estimating the latent density in unidimensional IRT to permit non-normality. In *Handbook of item response theory modeling* (pp. 78-102). Routledge.
- Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, 33(2), 102-117.
- Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2), 281-301.
- Yadin, A. (2013). Using unique assignments for reducing the bimodal grade distribution. *ACM Inroads*, 4(1), 38-42.

© 논문접수: 2021. 10. 28 / 수정본 접수: 2021. 12. 08 / 게재승인: 2021. 12. 19

— 저 자 소 개 —

· 이시우 : 연세대학교에서 교육학, 응용통계학 이중 전공으로 학사학위 취득 후, 동 대학원 교육학과에서 석사과정에 재학 중임. 연구 관심 분야는 문항반응이론, 일반화선형혼합모형, 통계적학습기법, 구조방정식, 일반화가능도이론 등임. cu@yonsei.ac.kr

〈요 약〉

## 문항반응모형 모수 추정에서 2-요소 정규혼합분포의 잠재분포로서의 활용

이 시 우

연세대학교

검사 점수가 이분분포를 따르는 경우 Mislevy(1984)는 2-요소 정규혼합분포(2NM)를 문항반응모형의 잠재분포로 활용하는 방법을 제안했다. 본 연구의 목적은 실제 잠재분포가 분포의 모수 값에 따라 형성된 다양한 모양의 2NM일 때 2NM-가정-문항반응모형(2NM-IRM)의 기능을 모의실험을 통해 검토하는 것이다. 본 연구에서는 2NM의 모양의 특징을 나타내고 모양을 결정하는 3개의 모양 모수(shape parameter)를 정의하였으며, 이 3개의 모양 모수 값들에 의해 결정된 총 15가지 모양의 2NM을 모의실험 조건으로 활용하였다. 2NM-IRM의 기능은 정규성-가정-문항반응모형(normal-IRM), 경험적 히스토그램 방법(EHM), Davidian-curve IRT(DC-IRT)와의 비교를 통해 검토되었다. 연구 결과, 2NM-IRM은 실제 잠재분포가 표준정규분포와 매우 유사한 경우를 제외하고는 다른 방법들보다 더 정확한 추정 결과를 산출하는 것으로 나타났다. 모양 모수 추정치에 편의가 존재하는 경우에도 2NM-IRM은 실제 잠재분포의 전반적인 모양을 일관적으로 정확하게 추정하고, 이에 따라 정확한 문항모수 및 능력모수 추정치를 제공하는 것으로 나타났다.

주제어 : 잠재분포, 이분분포, 정규혼합분포, 문항반응모형

⟨Appendix 1⟩  $ISE(\hat{g})$  for normal-IRM, EHM, DC-IRT, and 2NM-IRM

Parameter			$ISE(\hat{g})$			
$\zeta$	$\pi$	$\delta$	normal-IRM	EHM	DC-IRT	2NM-IRM
1		0.894	0.0000	0.0257	0.0000	0.0006
		$\frac{1}{2}$	0.0035	0.0233	0.0035	0.0007
		1.664	0.0277	0.0295	0.0020	0.0006
		0.894	0.0001	0.0278	0.0001	0.0005
		$\frac{1}{3}$	0.0034	0.0258	0.0015	0.0006
		1.664	0.0187	0.0255	0.0021	0.0014
		0.894	0.0171	0.0277	0.0025	0.0011
		$\frac{1}{2}$	0.0375	0.0253	0.0044	0.0016
		1.664	0.0748	0.0244	0.0095	0.0021
2		0.894	0.0102	0.0243	0.0027	0.0011
		$\frac{1}{3}$	0.0170	0.0227	0.0040	0.0023
		1.664	0.0344	0.0206	0.0083	0.0055
		0.894	0.0171	0.0303	0.0014	0.0012
		$\frac{2}{3}$	0.0433	0.0241	0.0023	0.0013
		1.664	0.0773	0.0290	0.0039	0.0016

Note. The ISE values for normal-IRM are constants, since normal-IRM does not estimate latent distributions.

⟨Appendix 2⟩  $RMSE(\widehat{ICC})$  for normal-IRM, EHM, DC-IRT, and 2NM-IRM

Parameter			$SE(\widehat{ICC})$			
$\zeta$	$\pi$	$\delta$	normal-IRM	EHM	DC-IRT	2NM-IRM
1		0.894	0.0204	0.0227	0.0204	0.0207
		$\frac{1}{2}$	0.0203	0.0213	0.0203	0.0198
		1.664	0.0216	0.0203	0.0203	0.0191
		0.894	0.0204	0.0219	0.0204	0.0206
		$\frac{1}{3}$	0.0207	0.0213	0.0207	0.0203
		1.664	0.0220	0.0210	0.0205	0.0197
		0.894	0.0247	0.0231	0.0216	0.0214
		$\frac{1}{2}$	0.0245	0.0219	0.0212	0.0200
		1.664	0.0242	0.0206	0.0200	0.0191
2		0.894	0.0226	0.0225	0.0216	0.0212
		$\frac{1}{3}$	0.0216	0.0209	0.0207	0.0201
		1.664	0.0217	0.0202	0.0206	0.0197
		0.894	0.0261	0.0238	0.0218	0.0217
		$\frac{2}{3}$	0.0292	0.0220	0.0214	0.0207
		1.664	0.0299	0.0216	0.0221	0.0203

〈Appendix 3〉  $RMSE(\hat{\theta})$  for normal-IRM, EHM, DC-IRT, and 2NM-IRM

Parameter			SE( $\hat{\theta}$ )			
$\zeta$	$\pi$	$\delta$	normal-IRM	EHM	DC-IRT	2NM-IRM
1		0.894	0.328	0.348	0.328	0.329
		$\frac{1}{2}$	0.323	0.337	0.325	0.321
		1.664	0.321	0.322	0.324	0.309
	$\frac{1}{3}$	0.894	0.328	0.345	0.329	0.329
		1.414	0.326	0.340	0.328	0.325
		1.664	0.326	0.332	0.330	0.317
	$\frac{1}{2}$	0.894	0.343	0.349	0.339	0.333
		1.414	0.333	0.334	0.333	0.316
		1.664	0.328	0.318	0.316	0.301
2	$\frac{1}{3}$	0.894	0.334	0.344	0.334	0.330
		1.414	0.325	0.328	0.326	0.317
		1.664	0.322	0.318	0.323	0.309
	$\frac{2}{3}$	0.894	0.356	0.356	0.345	0.340
		1.414	0.352	0.338	0.334	0.325
		1.664	0.345	0.326	0.330	0.312

〈Appendix 4〉 Software R functions for estimating parameters of normal-IRM, EHM, and 2NM-IRM

```
#####
# Preliminary functions
#####
# ICC
P <- function(theta,a=1,b,c=0){
  c+(1-c)*(1/(1+exp(-a*(theta-b))))
}
# Likelihood
logLikeli <- function(item, data, theta){
  if(length(theta)!=1){
    L <- matrix(nrow = nrow(data), ncol = ncol(data))
    for(i in 1:ncol(data)){
      for(j in 1:nrow(data)){
        Ll[,i] <- data[j,i]*log(P(theta = theta[j],a=item[i,1],b=item[i,2]))+(1-data[j,i])*log(1-P(theta = theta[j],a=item[i,1],b=item[i,2]))
      }
    }
  }else{
    L <- matrix(nrow = nrow(data), ncol = ncol(data))
    for(i in 1:ncol(data)){
      Ll[,i] <- data[,i]*log(P(theta = theta,a=item[i,1],b=item[i,2]))+(1-data[,i])*log(1-P(theta = theta,a=item[i,1],b=item[i,2]))
    }
  }
  return(rowSums(L))
}
# PDFs
```

---

```

dnormal <- function(x, mean=0, sd=1){
  (2*pi)^(-0.5)/sd*exp(-(x-c(mean))^2/(2*sd^2))
}

dist2 <- function(x, prob = 0.5, d = 0, sd_ratio = 1, overallmean=0, overallsd=1){
  m1 <- -(1-prob)*d+overallmean
  m2 <- prob*d+overallmean
  s1 <- sqrt((overallsd^2-prob*(1-prob)*d^2)/(prob+(1-prob)*sd_ratio^2))
  s2 <- s1*sd_ratio
  density <- prob*dnormal(x, m1, s1)+(1-prob)*dnormal(x, m2, s2)
  return(density)
}

# E step
Estep <- function(item, data, range = c(-6,6), q = 121, prob = 0.5, d = 0, sd_ratio = 1, Xk=NULL, Ak=NULL){
  if(is.null(Xk)) {
    Xk <- seq(range[1],range[2],length=q)
  }
  if(is.null(Ak)) {
    Ak <- dist2(Xk, prob, d, sd_ratio)/sum(dist2(Xk, prob, d, sd_ratio))
  }
  Pk <- matrix(nrow = nrow(data), ncol = q)
  for(i in 1:q){
    Pk[i,] <- exp(logLikeli(item = item, data = data, theta = Xk[i]))*Ak[i]
  }
  Pk <- Pk/rowSums(Pk)
  rik <- t(data)%*%Pk
  fk <- colSums(Pk)
  return(list(Xk=Xk, Ak=Ak, fk=fk, rik=rik,Pk=Pk))
}

# M1 step
M1step <- function(E, item, max_iter=500, threshold=0.0000001){
  item_estimated <- matrix(nrow = nrow(item), ncol = ncol(item))
  se <- matrix(nrow = nrow(item), ncol = ncol(item))
  X <- E$Xk
  r <- E$rik
  f <- E$fk
  #####item parameter estimation#####
  for(i in 1:nrow(item)){
    iter <- 0
    div <- 3
    par <- item[i,]
    #####Newton-Raphson#####
    repeat{
      iter <- iter+1
      p <- P(theta = X, a=par[1], b=par[2])
      W <- p*(1-p)
      par[1] <- max(0.1,par[1])
      X_ <- X-par[2]
      L1 <- c(sum(X_*(r[i,]-f*p)), -par[1]*sum(r[i,]-f*p)) #the 1st derivative of the marginal likelihood
      d <- sum(-par[1]^2*f*W)
      b <- sum(par[1]*X_*f*W)
      a <- sum(-X_^2*f*W)
      inv_L2 <- matrix(c(d,b,-b,a), ncol = 2)/(a*d-b^2) #inverse of the 2nd derivative of the marginal likelihood
      diff <- inv_L2%*%L1
      if( sum(abs(diff)) > div ){
        par <- par-div/sum(abs(diff))*diff
      } else {
        par <- par-diff
        div <- max(abs(diff))
      }
    }
    if( div <= threshold | iter > max_iter) break
  }
}

```

---

---

```

    }
    item_estimated[i,] <- par
    se[i,] <- sqrt(-c(inv_L2[1,1], inv_L2[2,2]))
  }
  return(list(item_estimated, se))
}

# M2 step
M2step <- function(E, max_iter=200){
  X <- E$Xk
  f <- E$fk
  prob <- 0.5
  m1 <- -0.7071
  m2 <- 0.7071
  s1 <- 0.7071
  s2 <- 0.7071
  iter <- 0
  # EM algorithm - Gaussian Mixture
  repeat{
    iter <- iter+1
    resp1 <- f*(prob*dnormal(X,m1,s1))/(prob*dnormal(X,m1,s1)+(1-prob)*dnormal(X,m2,s2))
    resp2 <- f- resp1
    new_m <- c(resp1*X/sum(resp1), resp2*X/sum(resp2))
    diff <- m2-m1-new_m[2]+new_m[1]
    m1 <- new_m[1]
    m2 <- new_m[2]
    s1 <- sqrt(as.vector(resp1*X-as.vector(m1))^2/sum(resp1))
    s2 <- sqrt(as.vector(resp2*X-as.vector(m2))^2/sum(resp2))
    prob <- sum(resp1)/sum(f)
    if( abs(diff) < 0.0001 | iter > max_iter) break
  }
  d_raw <- m2-m1
  sd_ratio <- s2/s1
  s2total <- prob*s1^2+(1-prob)*s2^2+prob*(1-prob)*d_raw^2
  d <- d_raw/sqrt(s2total)
  return(c(prob,d_raw,d,sd_ratio,m1,m2))
}

#####
# The main function
#####
Bimodal_MMLE <- function(initialitem, data, range=c(-2,2), q=19, latent_dist="Normal", max_iter=200, threshold=0.0001, stepsizeH=1){
  Options=list(initialitem=initialitem,data=data,range=range,q=q,latent_dist=latent_dist,max_iter=max_iter,threshold=threshold,stepsizeH=stepsizeH)
  I <- initialitem
  Xk <- seq(range[1],range[2],length=q)
  Ak <- NULL
  iter <- 0
  diff <- 10
  prob <- 0.5
  d <- 1
  sd_ratio <- 1
  if(latent_dist=="Normal"){
    while(iter < max_iter & diff > threshold){
      iter <- iter +1
      E <- Estep(item=initialitem, data=data, q=q, prob=0.5, d=0, sd_ratio=1, range=range)
      M1 <- M1step(E, item=initialitem)
      initialitem <- M1[[1]]
      diff <- max(abs(I-initialitem))
      I <- initialitem
      cat("\n","Method = ",latent_dist,", EM cycle = ",iter,", Max-Change = ",diff,sep="");flush.console()
    }
  }
  Ak <- E$Ak

```

---

---

```

}
if(latent_dist=="EHM"){
  while(iter < max_iter & diff > threshold){
    iter <- iter + 1
    E <- Estep(item=initialitem, data=data, q=q, prob=0.5, d=0, sd_ratio=1, range=range, Xk=Xk, Ak=Ak)
    M1 <- M1step(E, item=initialitem)
    initialitem <- M1[[1]]
    if(is.null(Ak)){
      Ak <- E$fk/sum(E$fk)
    }
    Ak <- Ak+stepsizeH*(E$fk/sum(E$fk)-Ak)
    Xk <- E$Xk
    m <- Xk%%Ak
    s <- (Xk-c(m))^2%%Ak
    Xk <- (Xk-as.vector(m))/sqrt(as.vector(s))
    ap <- approx(Xk, y = Ak, xout = seq(range[1],range[2],length=q), rule=2)
    Xk <- ap$x
    Ak <- ap$y/sum(ap$y)
    diff <- max(abs(I-initialitem))
    I <- initialitem
    cat("\n","Method = ",latent_dist," EM cycle = ",iter," Max-Change = ",diff,sep="");flush.console()
  }
}
if(latent_dist=="Mixture"){
  while(iter < max_iter & diff > threshold){
    iter <- iter + 1
    E <- Estep(item=initialitem, data=data, q=q, prob=prob, d=d, sd_ratio=sd_ratio, range = range)
    M1 <- M1step(E, item=initialitem)
    initialitem <- M1[[1]]
    M2 <- M2step(E)
    prob = M2[1];d = M2[3];sd_ratio = M2[4]
    diff <- max(abs(I-initialitem))
    I <- initialitem
    cat("\n","Method = ",latent_dist," EM cycle = ",iter," Max-Change = ",diff,sep="");flush.console()
  }
  Ak <- E$Ak
}
EAP <- as.numeric(E$Pk%%E$Xk)
logL <- 0
for(i in 1:q){
  logL <- logL+sum(logLikeli(initialitem, data, theta = Xk[i])*E$Pk[,i])
}
E$Pk[E$Pk==0]<- .Machine$double.xmin
Ak[Ak==0] <- .Machine$double.xmin
logL <- logL + as.numeric(E$fk%%log(Ak)) - sum(E$Pk*log(E$Pk))
return(list(par_est=initialitem,
            se=M1[[2]],
            fk=E$fk,
            iter=iter,
            prob=prob,
            d=d,
            sd_ratio=sd_ratio,
            quad=Xk,
            diff=diff,
            Ak=Ak,
            Pk=E$Pk,
            theta = EAP,
            logL=-2*logL,
            Options = Options))
}

```

---