Routledge
Taylor & Francis Group

Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Classification Consistency and Accuracy for Mixed-Format Tests

Stella Y. Kim & Won-Chan Lee

13 Mar 2019

IF: 1.043

Reporter: Yingshi Huang

Introduction

- Tests are administered for a variety of reasons:
  - to determine rank orders
  - to screen/select a certain group

CET®

fail < **425** < pass

→ a single test administration

- Classification consistency
- Classification accuracy

$P_i = p_{11} + p_{00}$

		Version B	
		pass	fail
Version A	pass	p_{11}	p_{10}
	fail	p_{01}	p_{00}

$\gamma_i = p_{11}$ / $\gamma_i = p_{00}$

		observed	
		pass	fail
true	pass	p_{11}	p_{10}
true	fail	p_{01}	p_{00}

Mixed-format tests ?

multiple-choice (MC) + free-response (FR)

- provide a rich understanding of examinee performance
- demonstrate some level of multidimensionality

The impact of construct equivalence was **negligible** (Wan, Brennan, & Lee, 2007)

➡ classical models

VS

When the testlet effect is low, the unidimensional IRT method **outperformed** bi-factor MIRT (Lafond, 2014)

➡ UIRT and MIRT

Impact of cut score location?

A cut score **near the mean or median** leads to **lower P** estimates (Huynh, 1976; Knupp, 2009; Lee, 2008; Wan et al., 2007)

As the number of classification **categories increases**, the CC and CA estimates tend to be **lower** (Berk, 1980; Feldt & Brennan, 1989; Lafond, 2014; Wan, 2006)

Present various estimation procedures

- classical test theory
- unidimensional item response theory (IRT)
- multidimensional IRT (MIRT)

Investigate the impact of multidimensionality

- real data
  - effects of dimensionality & impact of cut score location
- simulated data
  - sample size & degree of multidimensionality

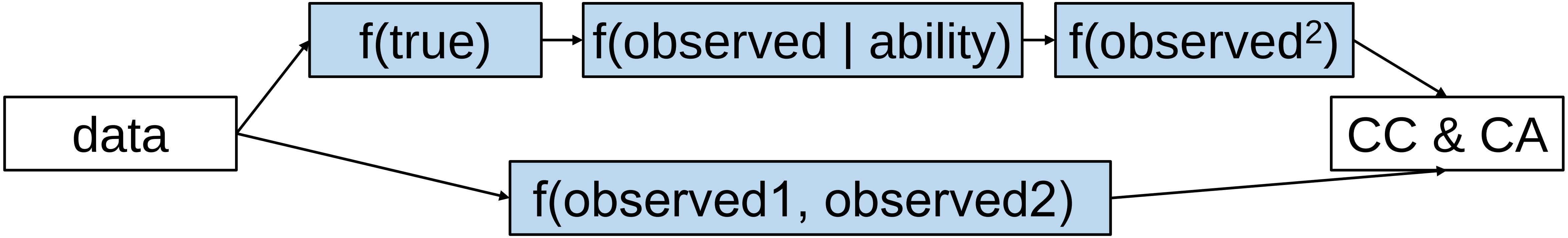
Classification Consistency and Accuracy for Mixed-Format Tests

classical approaches

- normal approximation (Peng & Subkoviak, 1980)
- Livingston-Lewis (Livingston & Lewis, 1995)
- compound multinomial (Lee, 2008)

IRT approaches

- unidimensional IRT (Lee, 2010)
- simple-structure MIRT (Knupp, 2009)
- bi-factor MIRT (LaFond, 2014)



1. Normal Approximation Procedure

- Scores from parallel forms follow a **bivariate normal distribution** with a correlation equal to test reliability, ***ρ***.

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_{y_1}\sigma_{y_2}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_{y_1}}{\sigma_{y_1}}\right)^2 - \frac{2\rho(y_1-\mu_{y_1})(y_2-\mu_{y_2})}{\sigma_{y1}\sigma_{y2}} + \left(\frac{y_2-\mu_{y_2}}{\sigma_{y_2}}\right)^2\right]\right)$$

$$f(y_1, y_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{y_1^2-2\rho y_1 y_2+y_2^2}{2(1-\rho^2)}\right)$$

$\left[c_{(j-1)}, c_j - 1\right] \rightarrow$ category U_j

$$z_{c_j} = \frac{c_j - \mu}{\sigma} \qquad z_{c_{(j-1)}} = \frac{c_{(j-1)} - \mu}{\sigma} \qquad (c_1,\, c_2,\, \ldots c_{J-1})$$

Peng & Subkoviak, 1980 *JEM*

1. Normal Approximation Procedure

- Being classified into category U_j on two parallel forms with scores Y_1 and Y_2

$$\Phi_2\left(Y_1 \epsilon U_j,\ \ Y_2 \epsilon U_j\right) = \int_{z_{c_{(j-1)}}}^{z_{c_j}} \int_{z_{c_{(j-1)}}}^{z_{c_j}} \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{2(1-\rho^2)}\right) dy_1 dy_2$$

$$P = \sum_{j=1}^{J} \Phi_2\left(\mathrm{Y}_1 \epsilon U_j, \mathrm{Y}_2 \epsilon U_j\right)$$

$P_i = p_{11} + p_{00}$

		Version B	
		pass	fail
Version A	pass	p_{11}	p_{10}
	fail	p_{01}	p_{00}

Peng & Subkoviak, 1980 *JEM*

1. Normal Approximation Procedure

- The true and observed scores follow a bivariate normal distribution with a correlation equal to **the square root of reliability, $\sqrt{\rho}$**.

$$z_{\xi_\eta} = \frac{\xi_\eta - \mu}{\sqrt{\rho}\sigma} \quad (\xi_\eta = c_j \rightarrow z_{\xi_\eta} = \frac{z_{c_j}}{\sqrt{\rho}})$$

summed score (τ) metric

$$\gamma = \sum_{\eta=j=1}^{J} \Phi_2(\tau \epsilon U_\eta, Y \epsilon U_j) = \int_{z_{c_{(j-1)}}}^{z_{c_j}} \int_{z_{\xi_{(\eta-1)}}}^{z_{\xi_\eta}} \frac{1}{2\pi\sqrt{1-\rho}} exp(-\frac{\tau^2 - 2\sqrt{\rho}\tau y + y^2}{2(1-\rho)}) d\tau dy$$

Peng & Subkoviak, 1980 *JEM*

2. Livingston-Lewis Procedure

- True scores are assumed to take the form of either a **two- or four-parameter beta distribution**.

$$f(\pi_i) = \frac{1}{B(\alpha,\beta)} * \frac{(\pi_i - a)^{\alpha-1}(b - \pi_i)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}$$

→ proportion-correct score (π) metric

the effective test length: $\tilde{n} = \boldsymbol{int}\left(\frac{(\mu - Y_{min})(Y_{max} - \mu) - \rho\sigma^2}{\sigma^2(1-\rho)}\right)$

↓ two-term approximation to the compound binomial distribution

$$P_r(Y = y|\pi_i) = \binom{\tilde{n}}{y} {\pi_i}^{y}(1-\pi_i)^{\tilde{n}-y}$$

$$\Pr(Y \in U_j|\pi_i) = \sum_{y=c_{(j-1)}}^{c_j-1} \Pr(Y = y|\pi_i)$$

$[c_{(j-1)}, c_j - 1]$ → category U_j

Livingston & Lewis, 1995 *JEM*

2. Livingston-Lewis Procedure

- Due to the conditional independence assumption:

$$\boxed{\Pr(Y \in U_j | \pi_i)} = \sum_{y=c_{(j-1)}}^{c_j - 1} \Pr(Y = y | \pi_i)$$

for examinee *i*

$$P_i = \sum_{j=1}^{J} \Pr(Y_1 \in U_j, Y_2 \in U_j | \pi_i) = \sum_{j=1}^{J} \Pr(Y_1 \in U_j | \pi_i) \Pr(Y_2 \in U_j | \pi_i)$$

$$= \sum_{j=1}^{J} \left[\boxed{\Pr(Y \in U_j | \pi_i)}\right]^2.$$

for a group of examinees

$$P = \int_0^1 P_i g(\pi)\, d\pi$$

Livingston & Lewis, 1995 *JEM*

2. Livingston-Lewis Procedure

- a similar approach

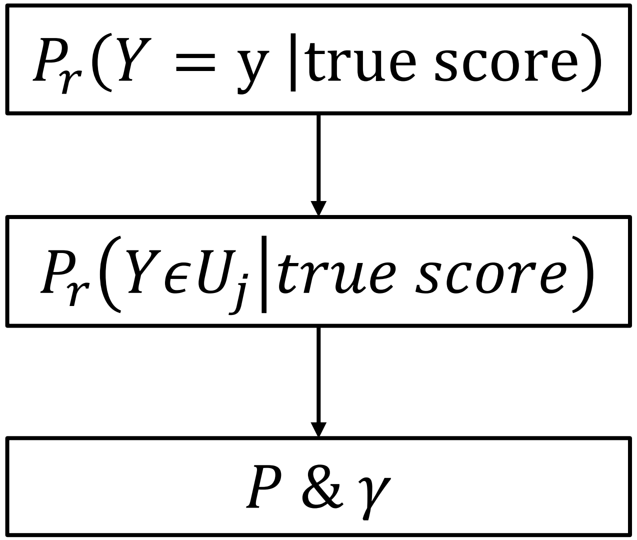
$$\Pr\left(Y \in U_j | \pi_i\right) = \sum_{y=c_{(j-1)}}^{c_j - 1} \Pr(Y = y | \pi_i)$$

for examinee *i*

$$\gamma_i = \Pr\left(Y \in U_j | \pi_i \in U_{\eta_i}\right) = \Pr\left(Y \in U_j | \pi_i\right), \text{ for } \eta_i = j$$

for a group of examinees

$$\gamma = \int_0^1 \gamma_i \, g(\pi) d\pi$$



Livingston & Lewis, 1995 *JEM*

3. Compound Multinomial Procedure

$P_r(Y = y \mid \text{true score})$

- item cluster:
  - **the same number of score categories** or **the same sub-content area**

$$\pi_{MC} = \{\pi_1, \pi_2\},\ \pi_1 + \pi_2 = 1,$$
$$\pi_{FR} = \{\pi_1, \pi_2, \dots \pi_k\},\ \pi_1 + \pi_2 + \cdots + \pi_k = 1.$$

Under the assumption of **uncorrelated** errors over the two item-format sections

$$\Pr(Y = y \mid \pi_{MCi}, \pi_{FRi}) = \sum \Pr(X_{MC} = x_{MC} \mid \pi_{MCi}) \Pr(X_{FR} = x_{FR} \mid \pi_{FRi})$$

all possible combinations of $w_{MC}X_{MC}$ and $w_{FR}X_{FR}$

Lee, 2008 *CASMA Research Report*
Lee, Brennan, & Wan, 2009 *APM*

3. Compound Multinomial Procedure

$$\Pr(Y = y|\pi_{MCi}, \pi_{FRi}) = \sum \Pr(X_{MC} = x_{MC}|\pi_{MCi})\Pr(X_{FR} = x_{FR}|\pi_{FRi})$$

$$P_r(Y \epsilon U_j|\pi_{MC_i}, \pi_{FR_i})$$

$$P_i = \sum_{j=1}^{J} \Pr(Y_1 \in U_j, Y_2 \in U_j|\pi_{MCi}, \pi_{FRi}) = \sum_{j=1}^{J} [\Pr(Y \in U_j|\pi_{MCi}, \pi_{FRi})]^2$$

$$\gamma_i = P_r(Y \epsilon U_j|\pi_{MC_i}, \pi_{FR_i})$$

equivalent to his/her actual classification based on the observed score

take the **average** of the conditional (individual) estimates

$$P = \sum_{i=1}^{N} P_i/N \qquad \gamma = \sum_{i=1}^{N} \gamma_i/N$$

Lee, 2008 *CASMA Research Report*

Lee, Brennan, & Wan, 2009 *APM*

4. Unidimensional IRT Procedure

$P_r(Y = y \mid \text{true score}) \rightarrow \theta$

$w_{MC}X_{MC}$ and $w_{FR}X_{FR}$ computed separately

$$\Pr(Y = y|\theta) = \sum \Pr(X_{MC} = x_{MC}|\theta)\Pr(X_{FR} = x_{FR}|\theta)$$

$$P_r(Y \epsilon U_j|\theta)$$

$$P_i = \sum_{j=1}^{J} P_r(Y_1 \epsilon U_j, Y_2 \epsilon U_j|\theta) = \sum_{j=1}^{J} [P_r(Y \epsilon U_j|\theta)]^2$$

$$\gamma_i = P_r(Y \epsilon U_j|\theta)$$

$$P = \int_{-\infty}^{\infty} P_i h(\theta) d(\theta)$$

$$\gamma = \int_{-\infty}^{\infty} \gamma_i h(\theta) d(\theta)$$

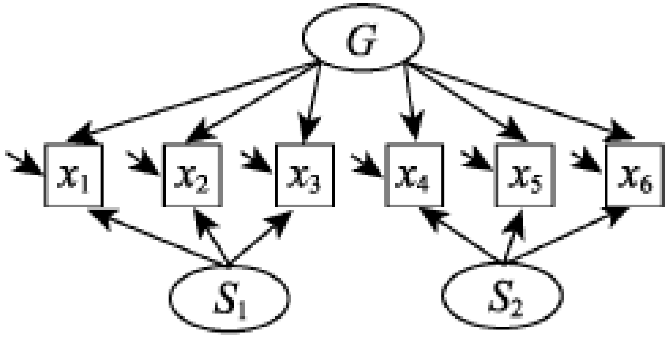
Lee, 2010 *JEM*

5. Simple-Structure MIRT Procedure

$P_r(Y = \text{y} \mid \text{true score})$ → $\boldsymbol{\theta_{MC}}$ **and** $\boldsymbol{\theta_{FR}}$**(allowed to be correlated)**

$$\Pr\left(Y = y \middle| \theta_{MC}, \theta_{FR}\right) = \sum \Pr(X_{MC} = x_{MC} | \theta_{MC}) \Pr(X_{FR} = x_{FR} | \theta_{FR})$$

6. Bi-Factor MIRT Procedure



$P_r(Y = \text{y} \mid \text{true score})$ →

$\boldsymbol{\theta_g}$ general ability

$\boldsymbol{\theta_{MC}}$ and $\boldsymbol{\theta_{FR}}$

(zero correlations)

$$\Pr\left(Y = y \middle| \theta_g, \theta_{MC}, \theta_{FR}\right) = \sum \Pr\left(X_{MC} = x_{MC} \middle| \theta_g, \theta_{MC}\right) \Pr\left(X_{FR} = x_{FR} \middle| \theta_g, \theta_{FR}\right)$$

Knupp, 2009 *Unpublished doctoral dissertation*

LaFond, 2014 *Unpublished doctoral dissertation*

Real Data Analysis

Table 1. Test information and sample sizes.

Exam	Section	# of Items	Score Points	Section Weights	Score Range	*n*
German	MC	65	65	1.00	0–130	4,283
	FR	4	5, 5, 5, 5	3.25		
Chemistry	MC	50	50	1.00	0–100	17,969
	FR	7	10, 10, 10, 4, 4, 4, 4	1.0869		
French	MC	65	65	1.0344	0–130	17,067
	FR	4	5, 5, 5, 5	3.25		
U.S. History	MC	80	80	1.125	0–180	17,239
	FR	3	9, 9, 9	3.33		
Biology	MC	58	58	1.00	0–120	9,911
	FR	8	10, 10	1.5		
			4, 4, 4, 3, 3, 3	1.4285		
English	MC	55	55	1.2272	0–150	15,541
	FR	3	9, 9, 9	3.0556		
Spanish	MC	65	65	1.00	0–130	16,459
	FR	4	5, 5, 5, 5	3.25		

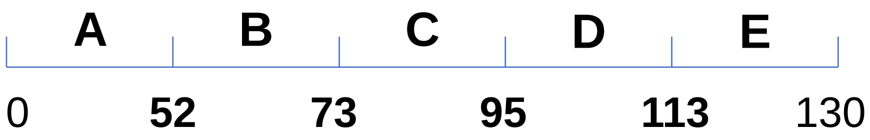
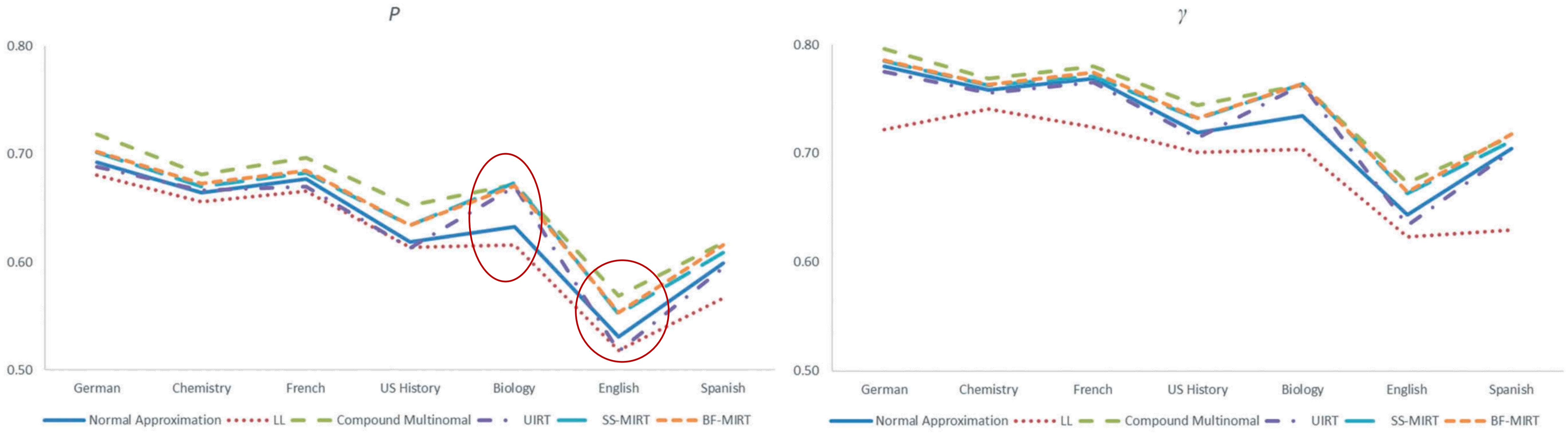


Table 2. Descriptive statistics and cut score information.

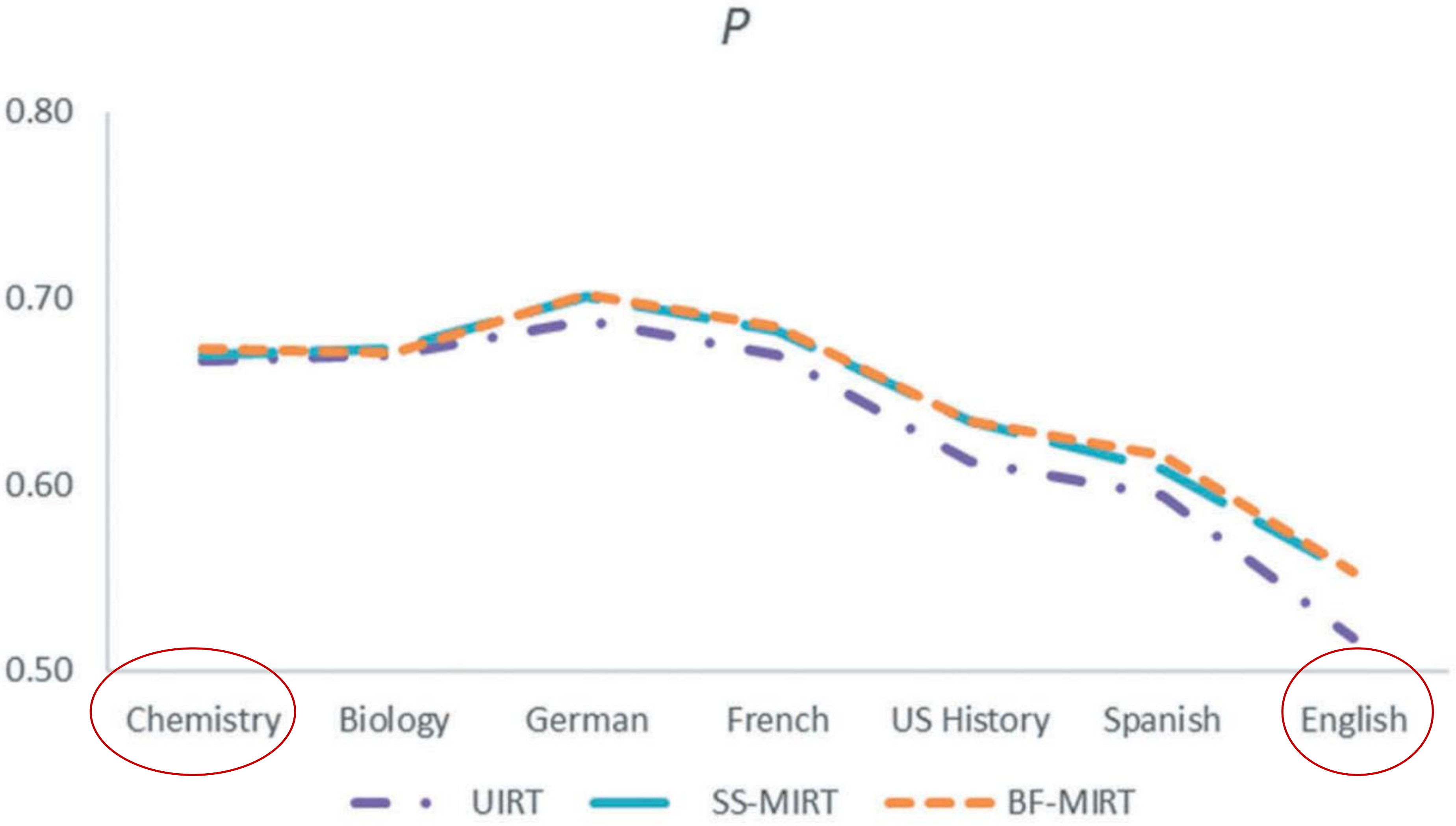
Exam	Mean	*SD*	Kurt.	Skew.	Rel.	$\hat{\rho}_{\theta_{MC}\theta_{FR}}$	Cut Score
German	90.911	25.502	2.382	−.390	.93797	.94	52, 73, 95, 113
Chemistry	44.443	19.598	2.162	.240	.92818	.97	27, 42, 58, 72
French	84.358	22.457	2.606	−.298	.91807	.92	44, 66, 88, 106
U.S. History	85.479	26.169	2.548	−.022	.91065	.89	59, 82, 97, 118
Biology	67.200	21.232	2.352	−.238	.88863	.96	33, 55, 76, 94
English	80.254	20.248	2.865	−.204	.82897	.75	54, 75, 91, 105
Spanish	93.484	18.758	3.637	−.724	.82014	.87	43, 68, 90, 107

Results

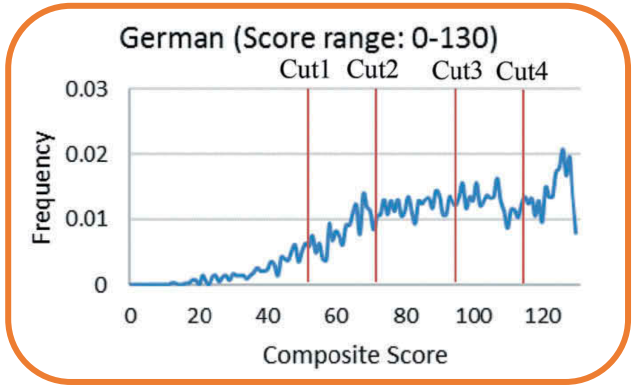
- Comparison of Estimation Procedures (multilevel classification)

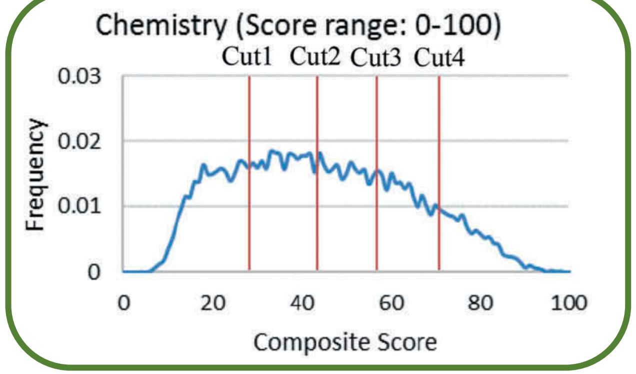


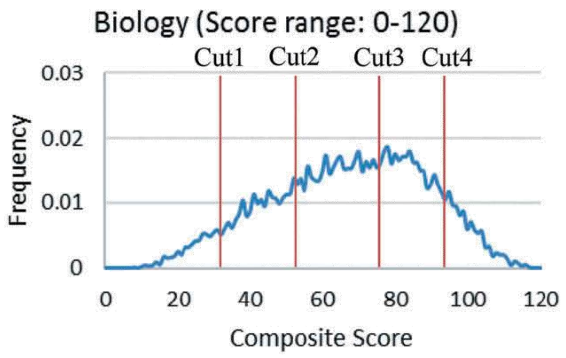
- Effects of Dimensionality (item-format effects)

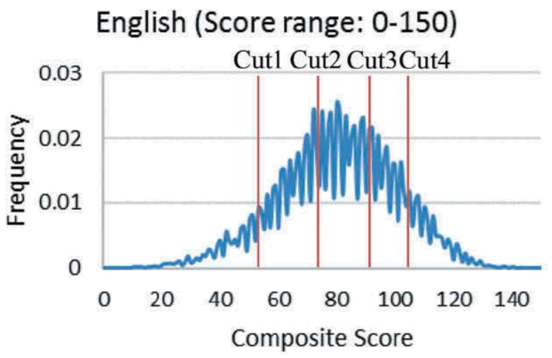


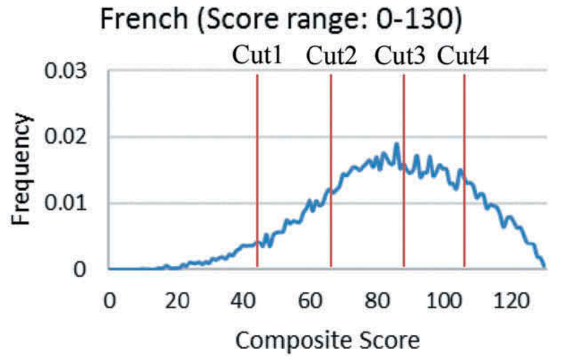
• Impact of Cut Score Location

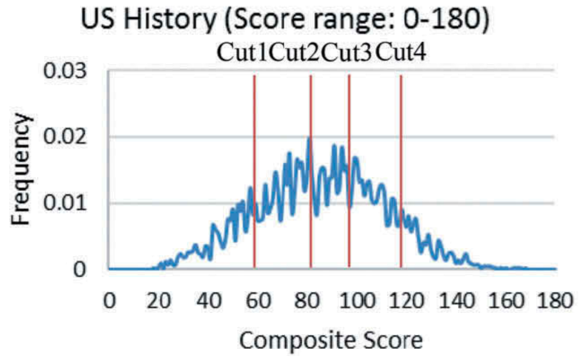










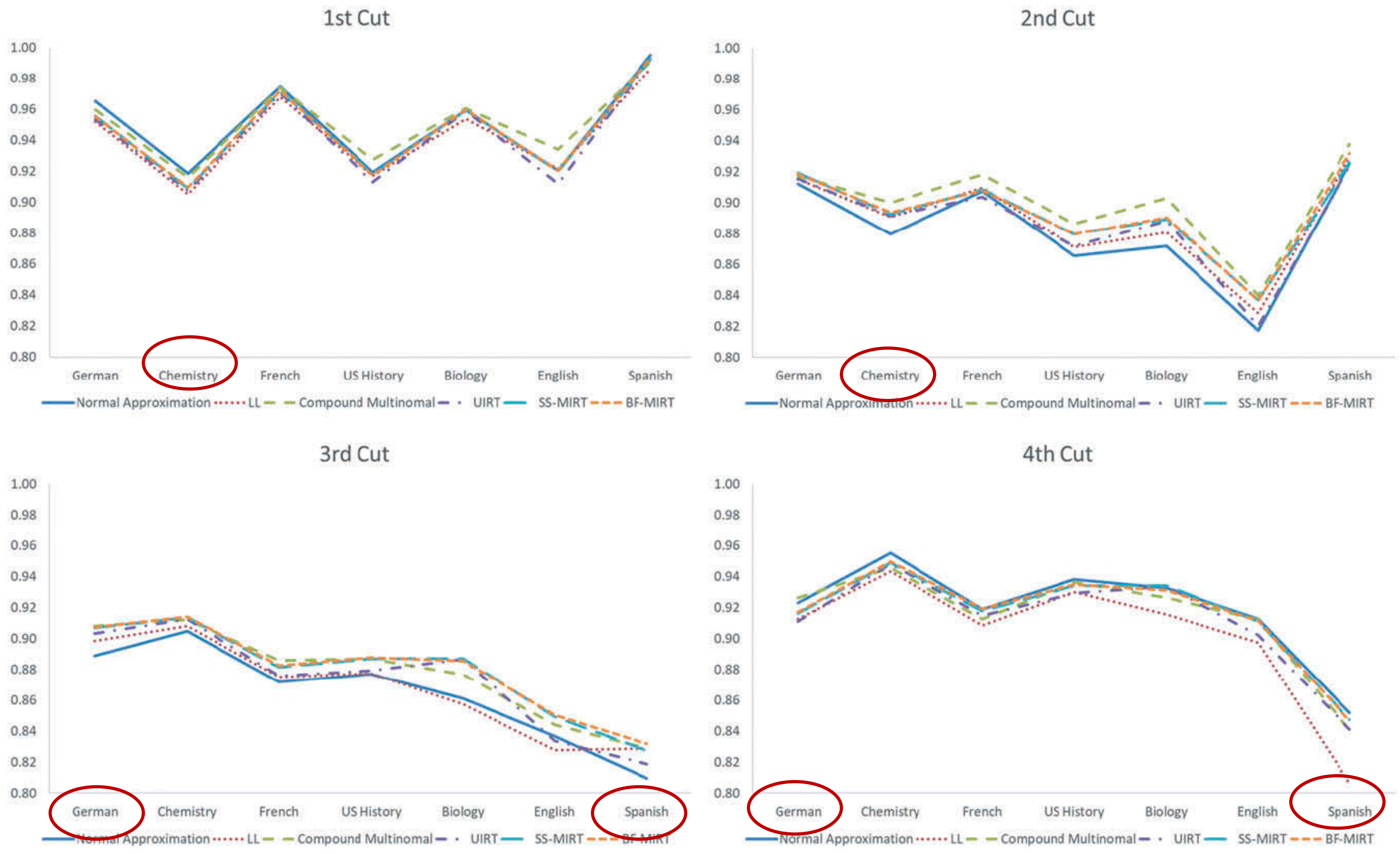


Spanish and German: negatively skewed distribution

Chemistry: positively skewed distribution

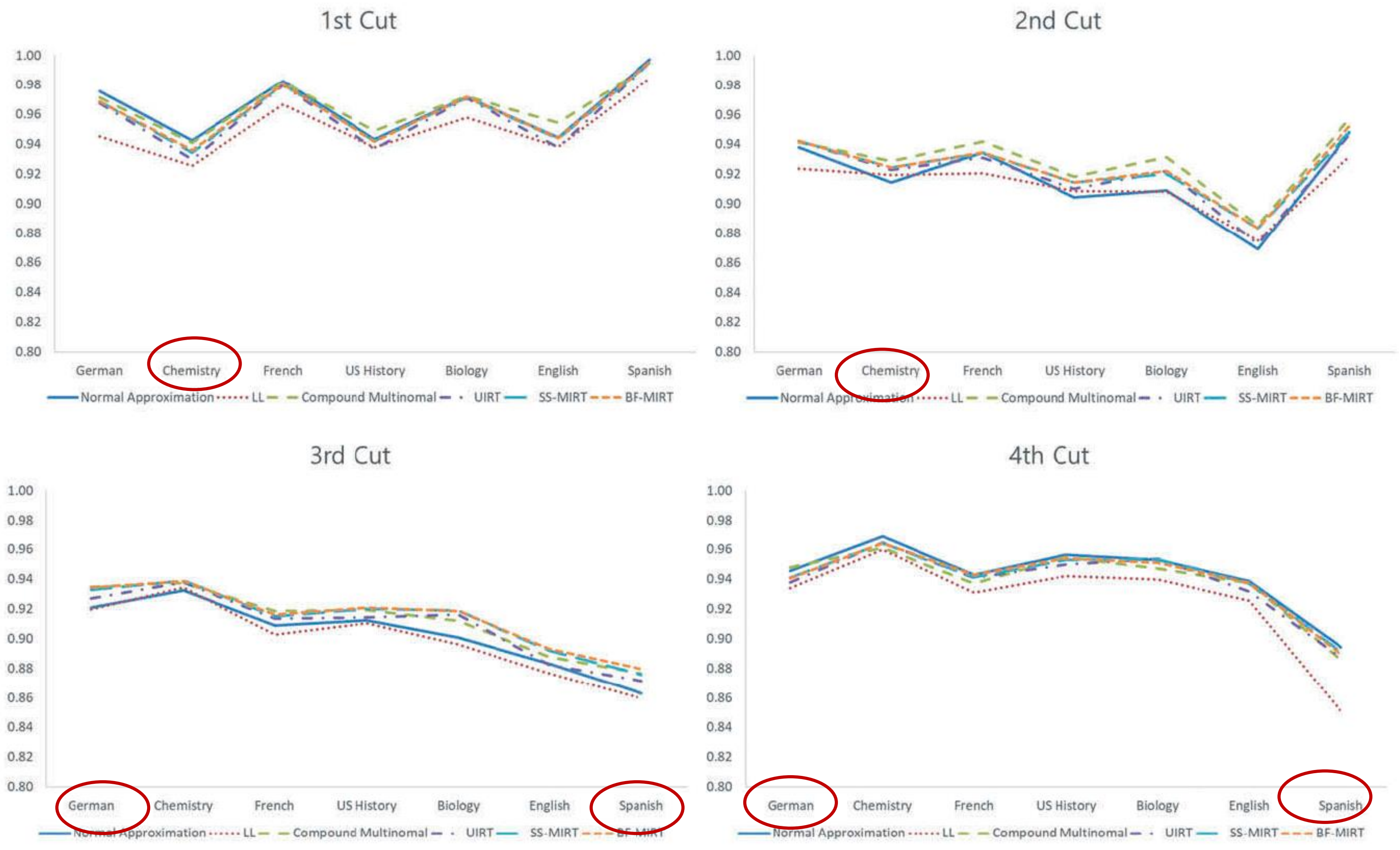
Exam	Mean	*SD*	Kurt.	Skew.	Rel.	$\hat{\rho}_{\theta_{MC}\theta_{FR}}$	Cut Score
German	90.911	25.502	2.382	−.390	.93797	.94	52, 73, 95, 113
Chemistry	44.443	19.598	2.162	.240	.92818	.97	27, 42, 58, 72
French	84.358	22.457	2.606	−.298	.91807	.92	44, 66, 88, 106
U.S. History	85.479	26.169	2.548	−.022	.91065	.89	59, 82, 97, 118
Biology	67.200	21.232	2.352	−.238	.88863	.96	33, 55, 76, 94
English	80.254	20.248	2.865	−.204	.82897	.75	54, 75, 91, 105
Spanish	93.484	18.758	3.637	−.724	.82014	.87	43, 68, 90, 107

- Impact of Cut Score Location



P estimates for binary classifications.

- # Impact of Cut Score Location



γ estimates for binary classifications.

Simulated Data Analysis

- Using the **simple-structure MIRT** model

- In the **item pool**, there were 657 MC items (3PLM) and 14 FR items (GRM) scored 0–10
  - 90 MC : scored 0–1
  - 3 FR : scored 0–10
  - Section weights of 1:3, score range of 0–180

- Four **cut scores**: 59, 82, 97, 118

- Manipulated variables
  - degree of multidimensionality: $\hat{\rho}_{\theta_{MC}\theta_{FR}}$ = 0.80 or 0.95
  - sample size: *N* = 500 or 3000

Criterion classification indexes (β)

the item pool

(657 MC & 14 FR)

90 MC & 3 FR

a large group of examinees

(N = 1,000,000)

θ_{MC} & θ_{FR} × 500 / 3000

Generate item responses

(for the two forms)

assign the examinees into categories

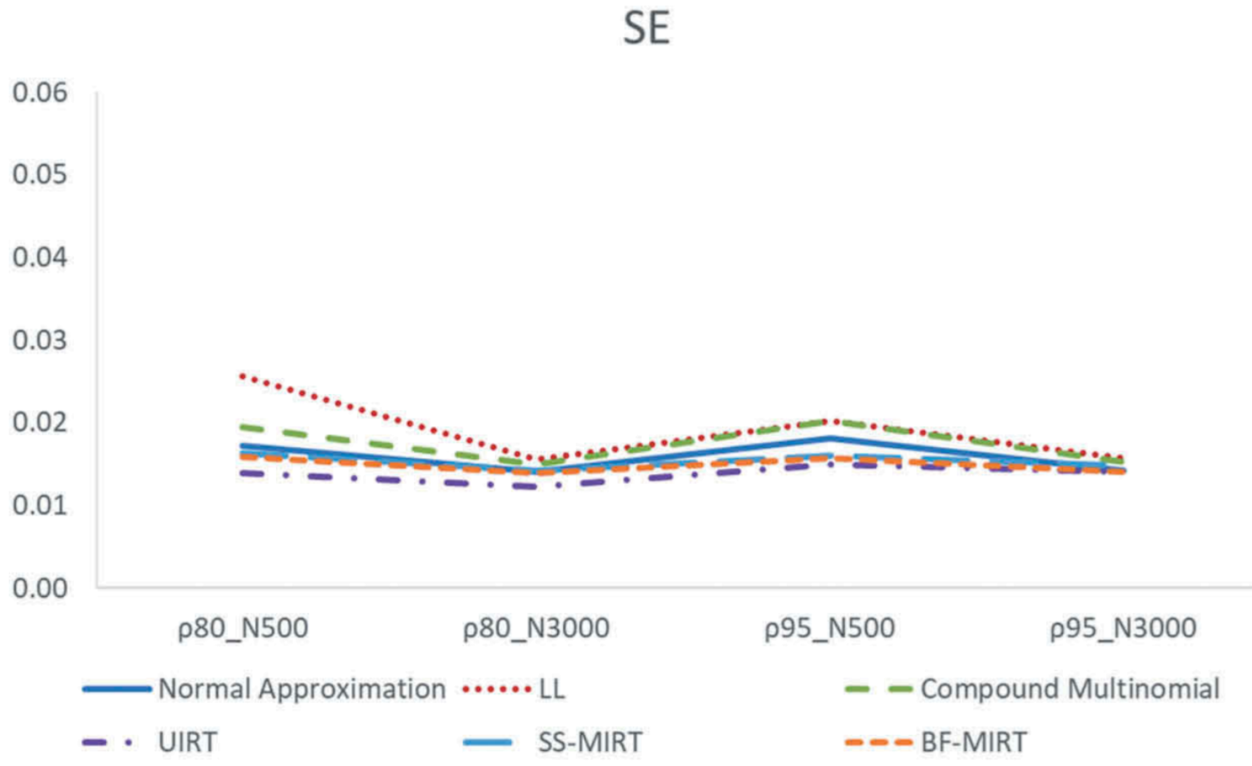
Compute classification indexes

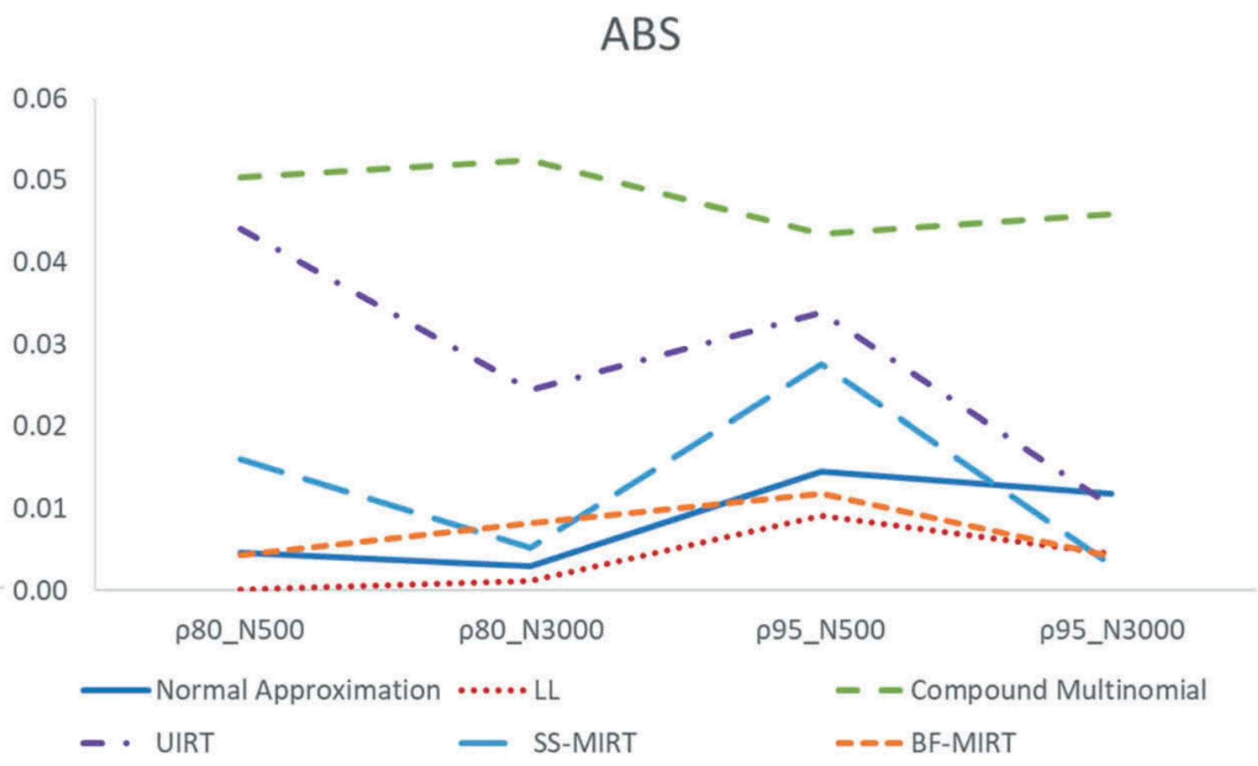
- repeated **100 times**
- the criterion **classification consistency**:
  - ✓ the average of classification consistency values
- the criterion **classification accuracy**:
  - ✓ based on their true score and observed score for only one form
  - ✓ the average of classification accuracy values
- random error: $SE(\beta) = \sqrt{\frac{1}{R}\sum_{r}^{R}\left(\hat{\beta}_r - \bar{\hat{\beta}}\right)^2}$
- systematic error: $ABS(\beta) = \left|\bar{\hat{\beta}} - \beta\right|$
- overall error:

$$RMSE(\beta) = \sqrt{\frac{1}{R}\sum_{r}^{R}\left(\hat{\beta}_r - \beta\right)^2} = \sqrt{SE(\beta)^2 + ABS(\beta)^2}$$

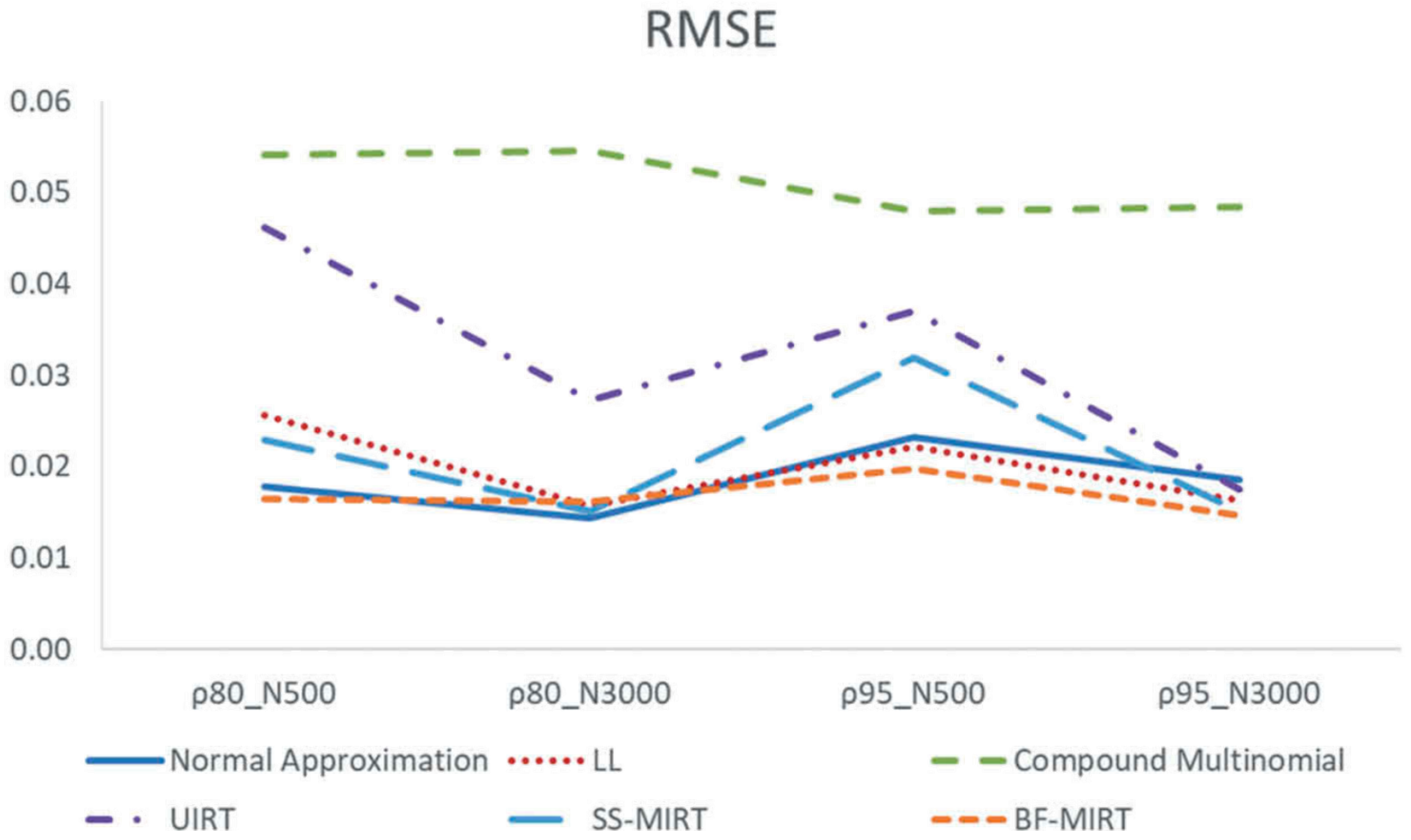
Results for *P*

- Comparison of Estimation Procedures (multilevel classification)

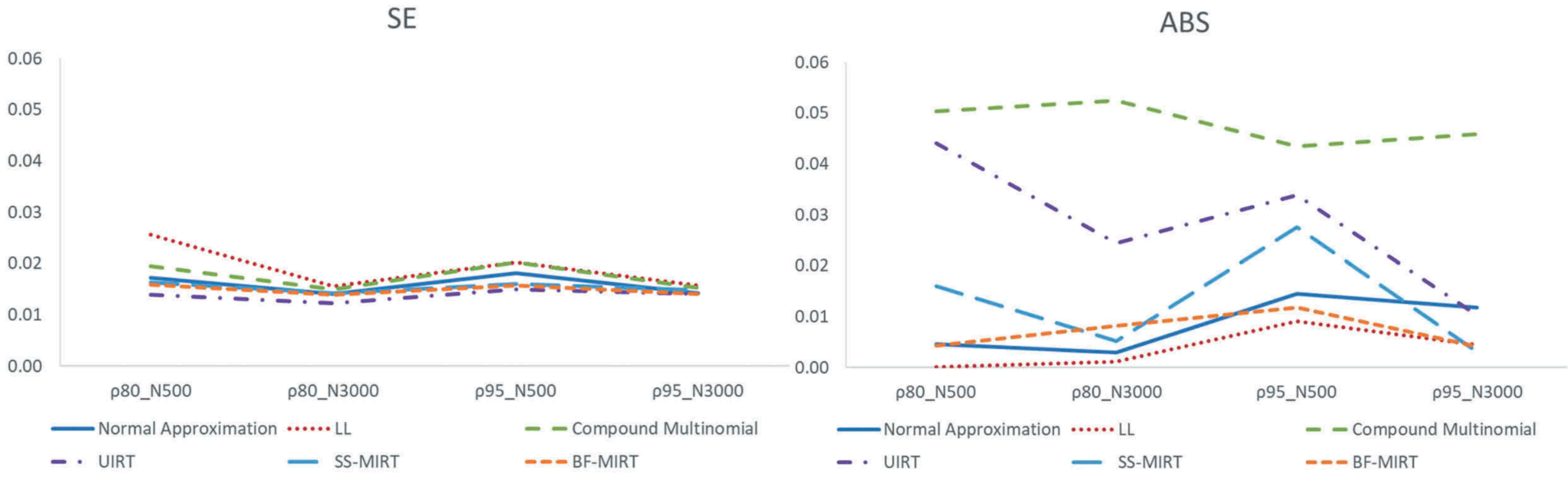




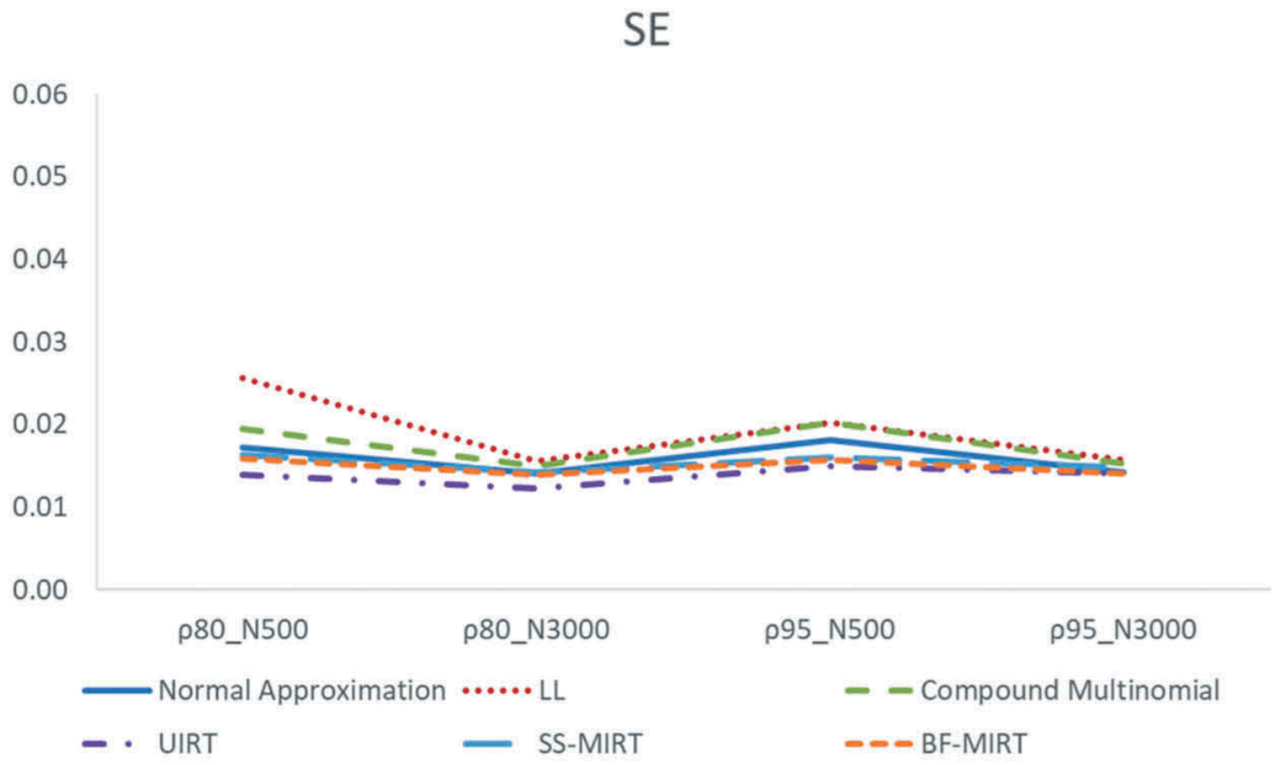
- Comparison of Estimation Procedures (multilevel classification)

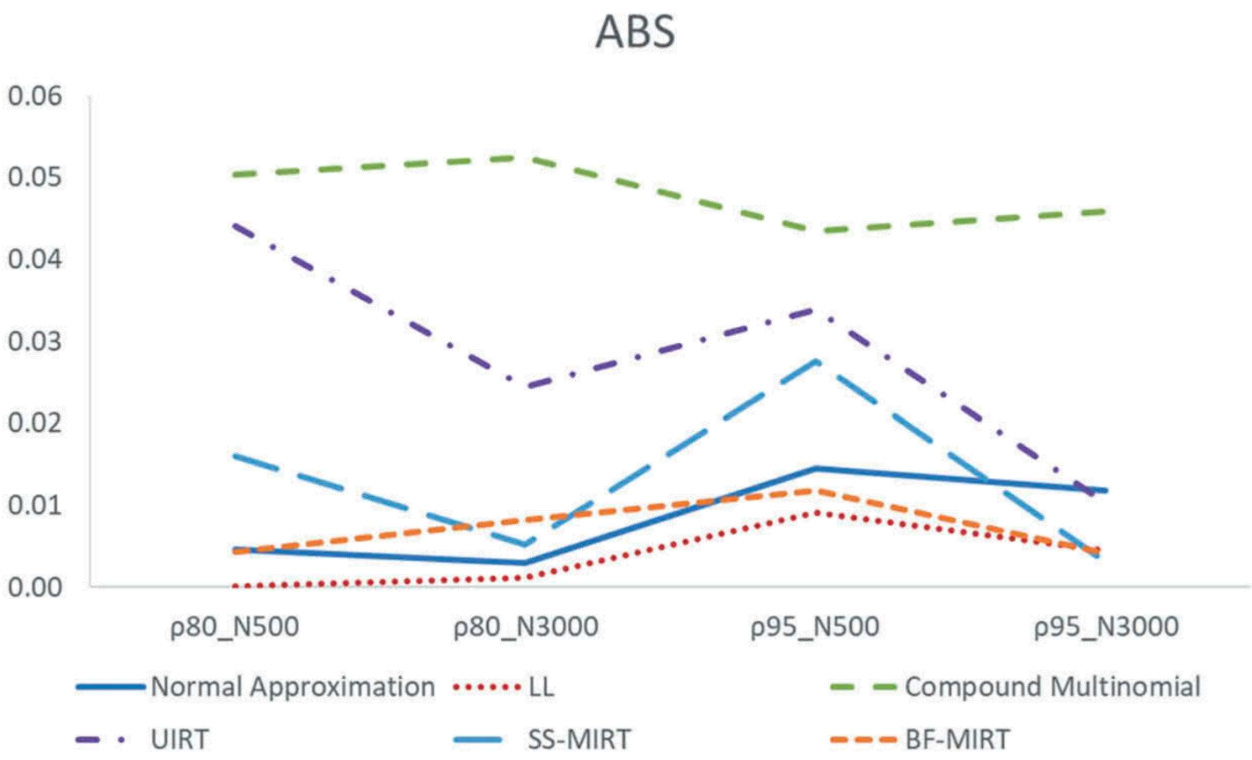


- Correlation Between MC and FR Scores (multidimensionality)

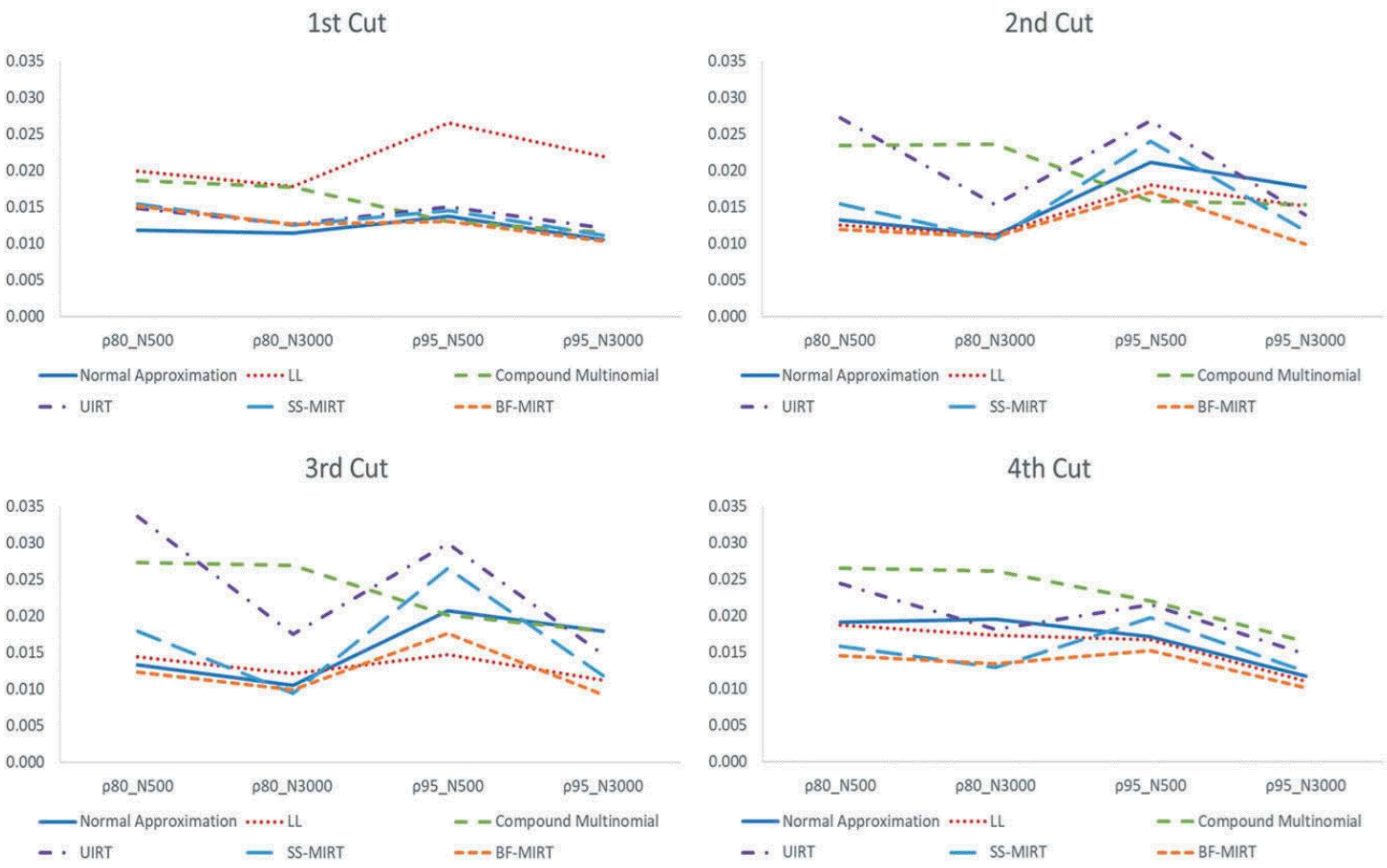


- Sample Size


SE
0.06
0.05
0.04
0.03
0.02
0.01
0.00
p80_N500
p80_N3000
p95_N500
p95_N3000
Normal Approximation
LL
Compound Multinomial
UIRT
SS-MIRT
BF-MIRT


ABS
0.06
0.05
0.04
0.03
0.02
0.01
0.00
p80_N500
p80_N3000
p95_N500
p95_N3000
Normal Approximation
LL
Compound Multinomial
UIRT
SS-MIRT
BF-MIRT

- Cut Score Location (binary classifications)



Discussion

- real data
  - All of the classical and IRT procedures show **similar patterns** across different exams.
  - The shape of the observed-score distribution influences classification indices while **interacting with** the position of the cut score.
  - As data become more multidimensional, unidimensional IRT yielding **lower P and γ** estimates than MIRT.
- simulated data
  - The largest SE was associated with **LL**, followed by the compound multinomial method.
  - The **compound multinomial procedure** and unidimensional IRT resulted in the largest bias.
  - Unidimensional IRT revealed **larger error** than bi-factor MIRT and simple-structure MIRT.

Limitations

- **Generalization** of the results is somehow limited.
- The criterion established for the simulation study might **favor the generating model**.
- It would be worth exploring some other models such as **full MIRT models**.

Thanks for listening!

Yingshi Huang 2020/04/15