JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
DOI: 10.1111/jedm.12250
Journal of Educational Measurement

Classification Consistency and Accuracy With Atypical Score Distributions

Stella Y. Kim & Won-Chan Lee

05 September 2019

IF: 0.938

Reporter: Yingshi Huang

Classification Indices

- Agreement Index P

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

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

- Accuracy Indices

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

$$\gamma = p_{11} + p_{00}$$

$$\gamma- = p_{10}$$

$$\gamma+ = p_{01}$$

- Kappa Coefficient

		Version B		marginal proportions
		pass	fail	
Version A	pass	p_{11}	p_{10}	$p_{1\cdot}$
	fail	p_{01}	p_{00}	$p_{0\cdot}$
marginal proportions		$p_{\cdot 1}$	$p_{\cdot 0}$	1

$$p_c = p_{1\cdot} \times p_{\cdot 1} + p_{0\cdot} \times p_{\cdot 0}$$

$$\kappa = \frac{P - p_c}{1 - p_c}$$

Hambleton, & Novick, 1973 *JEM*
Cohen, 1960 *EPM*

Introduction

- a single test administration

		Version B	
		pass	fail
Version A	pass	?	?
	fail	?	?

		observed	
		pass	fail
true	pass	?	?
true	fail	?	?

- classical approaches → a potential factor: score distributions (Deng, 2011; Li, 2006)
- IRT approaches

evaluate the performance of several **non-IRT estimation procedures**
under various **“atypical” score distributions**

- Study Design

  - Study 1.

    investigate the performance of the estimation procedures under a **bimodal distribution**

  - Study 2.

    explore the effect of **structural bumpiness** in a score distribution

  - Study 3.

    examine the impact of **structural zeros** in a score distribution

Estimation Procedures

- Normal Approximation Procedure

◆ **have an identical mean and standard deviation**
◆ **smooth without bumpiness**

  - Scores from parallel forms follow a **bivariate normal distribution** with a correlation equal to test reliability, $\boldsymbol{\rho}$.

		Version B	
		pass	fail
Version A	pass	?	?
	fail	?	?

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

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

		observed	
		pass	fail
true	pass	?	?
true	fail	?	?

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

Peng & Subkoviak, 1980 *JEM*

- Program Operation (R)

```
setwd("Desktop")        #set working directory to the folder
                        where the data file is located
data <- read.table("example.dat")    #read the data file
library(pbivnorm)             #load the pbivnorm package
nm(data,.8,c(20,25))          #specify the reliability as .8
                               and two cut scores (20 and 25)
```

```
> nm(data, .8, c(20,25))
$`Binary Classifications`
                  PHI     KAPPA     GAMMA FALSE_POSITIVE FALSE_NEGATIVE
cut score 1 0.7983122 0.5888838 0.8547743     0.07661441     0.06861129
cut score 2 0.8907881 0.5330344 0.9232532     0.05150827     0.02523856

$`Simultaneous Classification`
                   PHI     KAPPA    GAMMA FALSE_POSITIVE FALSE_NEGATIVE
3 categories 0.7023259 0.4783902 0.779905      0.1267672     0.09332775
```

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

		Version B	
		pass	fail
Version A	pass	?	?
	fail	?	?

– the effective test length:

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

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

$$\Pr(Y_1 \in U_j, Y_2 \in U_j|\pi_i)$$

		observed	
		pass	fail
true	pass	?	?
true	fail	?	?

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

◆ **smooth without bumpiness**

Livingston & Lewis, 1995 *JEM*

- Program Operation (BB-CLASS)

```
LL 0.9 4    check
"LL data" f 1 2
3  140. 160.   .4  .6


121 3      141 5      161 14     181 8
122 5      142 20     162 17     182 3
123 8      143 11     163 17     183 9
124 5      144 14     164 23     184 0
125 3      145 15     165 29     185 7
126 9      146 21     166 19     186 5
127 2      147 13     167 16     187 0
128 2      148 12     168 33     188 2
129 9      149 10     169 12     189 1
130 18     150 18     170 34     190 1
131 10     151 18     171 16
132 11     152 17     172 21
133 13     153 8      173 17
134 12     154 21     174 32
135 10     155 6      175 0
136 11     156 33     176 32
137 16     157 32     177 22
138 11     158 7      178 14
139 16     159 17     179 8
140 15     160 36     180 25
```

```
***ACCURACY RELATIVE TO ACTUAL OBSERVED SCORES***

            x0        x1        x2        marg
  t0    0.15114   0.01021   0.00000   0.16135
  t1    0.06240   0.18741   0.01194   0.26174
  t2    0.00046   0.11539   0.46106   0.57690

marg    0.21400   0.31300   0.47300   1.00000

probability of correct classification = 0.79961
false positive rate = 0.02215;   false negative rate =  0.17824

***CONSISTENCY USING EXPECTED (row) VS. ACTUAL (column) OBSERVED SCORES***

            x0        x1        x2        marg
  x0    0.16806   0.04193   0.00068   0.21068
  x1    0.04527   0.20712   0.07116   0.32355
  x2    0.00066   0.06395   0.40116   0.46577

marg    0.21400   0.31300   0.47300   1.00000

pc = 0.77634;  pchance = 0.36667;  kappa = 0.64685
probability of misclassification = 0.22366
```

Brennan, 2004 *CASMA Research Report*

• Compound Multinomial Procedure

- item cluster:

◆ no process for fitting the observed score distribution

the same number of score categories or the same sub-content area

$$\Pr(Y_1 = y_1, \ldots, Y_L = y_L | \vec{\pi}_1, \ldots, \vec{\pi}_L) = \prod_{i=1}^{L} \Pr(Y_i = y_i | \vec{\pi}_i) \qquad (Z = \textstyle\sum_{i=1}^{L} w_i Y_i)$$

		Version B	
		pass	fail
Version A	pass	?	?
	fail	?	?

– $$\Pr(Z = z | \vec{\pi}_1, \ldots, \vec{\pi}_L) = \sum_{y_1, \ldots, y_L : \sum w_i y_i = z} \Pr(Y_1 = y_1, \ldots, Y_L = y_L | \vec{\pi}_1, \ldots, \vec{\pi}_L)$$

$$\sum_{h=1}^{H} \Pr(\lambda_{h-1} \leq Z < \lambda_h | \vec{\pi}_{p1}, \ldots, \vec{\pi}_{pL})^2 \qquad \lambda_1, \lambda_2, \ldots, \lambda_{H-1}$$

		observed	
		pass	fail
true	pass	?	?
true	fail	?	?

– $$\Pr(\lambda_{h-1} \leq Z < \lambda_h | \boxed{\vec{\pi}_{p1}, \ldots, \vec{\pi}_{pL}})$$

① observed proportion correct score

$$\hat{\pi}_o = \frac{x}{k} = \overline{x}$$

② regressed-score

$$\hat{\pi}_r = (1 - \rho^2)\,\mu + \rho^2\,\overline{x}$$

✓ optimal estimate

$$\hat{\pi}_w = w\,[(1 - \rho^2)\,\mu + \rho^2\,\overline{x}] + (1 - \boxed{w})\,\overline{x}$$

$$= (1 - \rho)\,\mu + \rho\,\overline{x}$$

$$\frac{1}{\sqrt{\rho^2 + 1}}$$

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

- Program Operation (MULT-CLASS)

```
$Number of item sets
2
$Set1: weight, #items, #score points, score points, data file
1
8
4
2 4 6 8
mixpoly.dat
$Set2: weight, #items, #score points, score points, data file
2
20
2
0 1
mixdich.dat
$#categories, observed cut scores, true cut scores
5
30 49 65 90
30 49 65 90
$Conditional results (Yes=1, No=0)
1
$Output file
mix.out
$Bias correction (Yes=1, No=0)
1
```

```
********************************************************************
                Overall Classification Consistency
--------------------------------------------------------------------
consistency (phi)     = 0.62527
1-phi                 = 0.37473
chance probability    = 0.29565
kappa                 = 0.46799
********************************************************************

********************************************************************
                 Overall Classification Accuracy
--------------------------------------------------------------------
accuracy (gamma)      = 0.72649
false positive error = 0.15871
false negative error = 0.11480
********************************************************************

********************************************************************
                       Conditional Results
--------------------------------------------------------------------
    Obs      Total        phi      gamma     false+     false-
      1      28.00    0.50102    0.47489    0.52511    0.00000
      2      46.00    0.55133    0.66244    0.33653    0.00102
      3      52.00    0.54295    0.66815    0.02188    0.30997
      4      82.00    0.75725    0.86183    0.11891    0.01927
      5      46.00    0.51883    0.62658    0.36633    0.00709
      6      56.00    0.59007    0.74572    0.09854    0.15574
      7      40.00    0.84784    0.91803    0.07025    0.01172
      8      52.00    0.54472    0.67004    0.02113    0.30883
      9      96.00    0.73934    0.84594    0.00000    0.15406
     10      54.00    0.65940    0.79125    0.02843    0.18032
********************************************************************
```

Lee, 2008 (Program MULT-CLASS)

Method

- Simulation Conditions
  - **IRT** was used to simulate data
  - **Test length** was fixed to 40 for all 3 studies
  - **Cut scores:** 50% (cut1), 65% (cut2), and 80% (cut3) of the maximum possible score
  - **sample size:** 100, 1,000, and 5,000

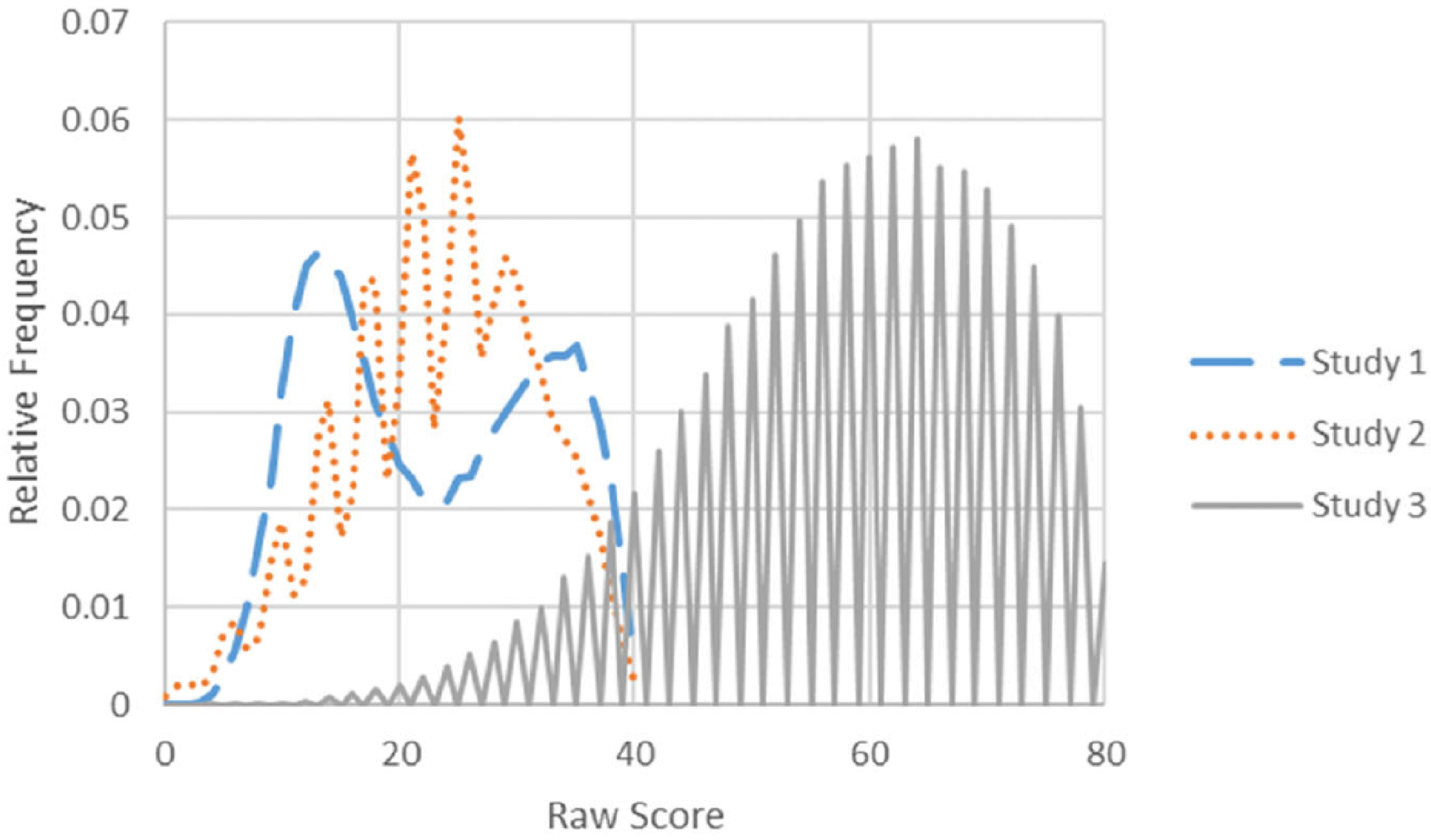
- Item parameters

Table 1

Item Pool Information

	Score Range	# of Items	Item Parameters	Range of Item Parameters	Mean of Item Parameters	*SD* of Item Parameters
(3PLM) MC Item Pool	0–1	657	*a*	.1080 ~ 2.2350	.8155	.3162
			b	–4.7450 ~ 3.0217	–.3116	1.0484
			c	.0226 ~ .5452	.1833	.0999
(GRM) FR Item Pool	0–10	8	*a*	.9577 ~ 1.2195	1.0805	.1042
			*b*1	–2.3404 ~ –1.5463	–1.8350	.3143
			*b*2	–1.5311 ~ –.7330	–1.1473	.3093
			*b*3	–.8770 ~ –.1012	–.5477	.3161
			*b*4	–.4364 ~ .4628	–.0192	.3287
			*b*5	–.0242 ~ .9923	.4537	.3647
			*b*6	.4140 ~ 1.5239	.9242	.4098
			*b*7	.8410 ~ 2.0878	1.4036	.4808
			*b*8	1.3209 ~ 2.7192	1.9261	.5733
			*b*9	1.8451 ~ 3.5369	2.5538	.7155
			*b*10	2.5148 ~ 4.2761	3.2736	.7826
	0–5	21	*a*	.6361 ~ 1.7053	1.016	.3232
			*b*1	–5.5202 ~ –1.9127	–2.9879	.8873
			*b*2	–3.5113 ~ –1.0895	–2.0399	.7638
			*b*3	–2.8014 ~ .0576	–.9828	.7588
			*b*4	–1.2324 ~ 1.5885	.2848	.7437
			*b*5	.2988 ~ 3.0753	1.4529	.7953

- Score distribution



- **Study 1**

  all 40 items were MC items combining Normal (−1.8,√.8) and Normal (.8,√.8)

- **Study 2**

  4 FR items
  scored 0, 1, 4, 5, 8, and 10

- **Study 3**

  20 MC items and 4 FR items scored 0–5
  weights of 2 for each section

- Criterion classification indexes (α)

the item pool

(657 MC & 29 FR)

Study1: 40 MC
Study2: 4 FR
Study3: 20 MC & 4 FR

100 / 1000 / 5000 examinees (fixed)

100 times

Generate item responses

(for the two forms)

assign the examinees into categories

Compute classification indexes

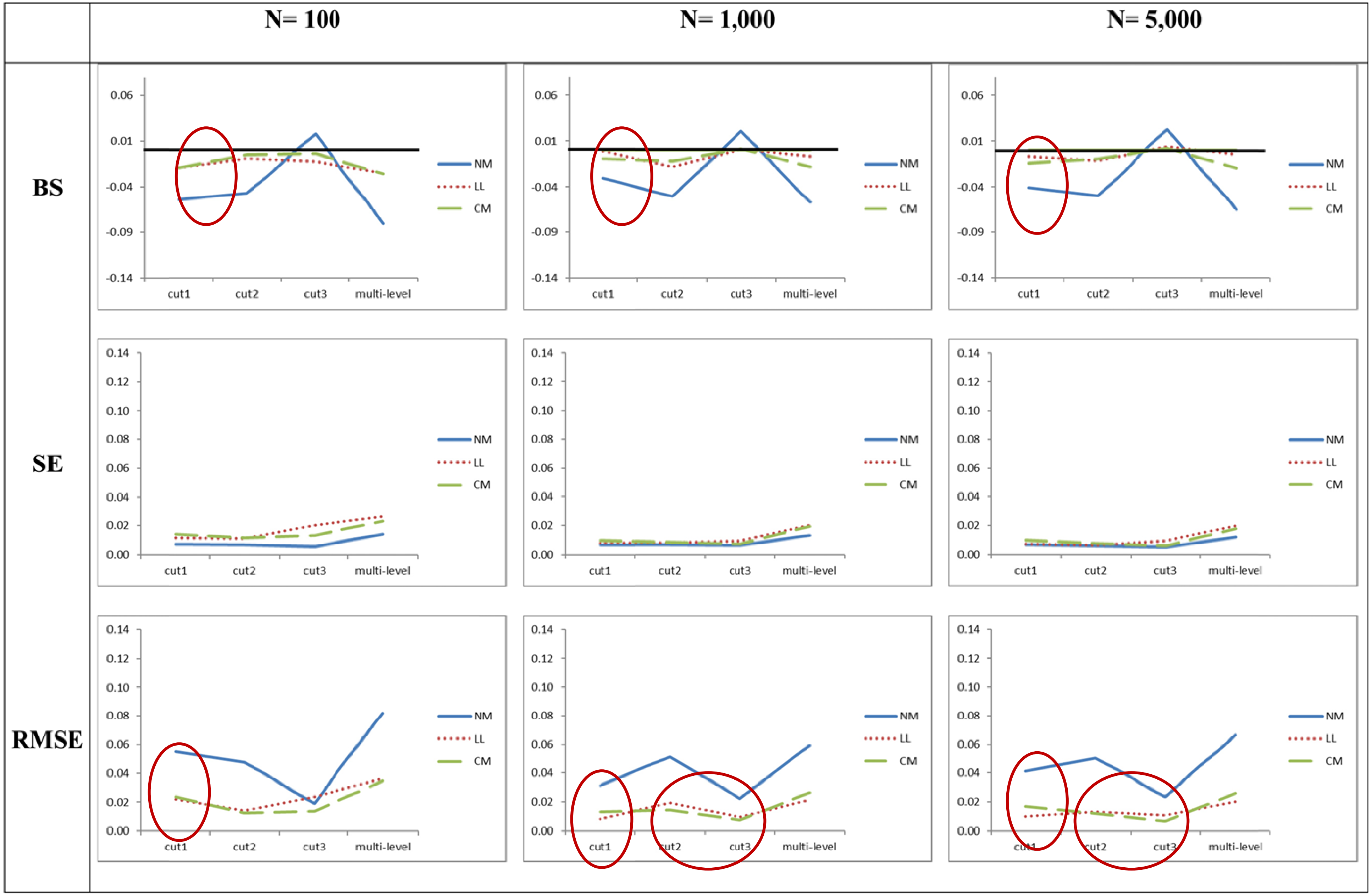
- the criterion **classification consistency**:
  - ✓ the average of classification consistency values
- the criterion **classification accuracy**:
  - ✓ based on their true score and observed score for only one form
  - ✓ the average of classification accuracy values
- random error: $SE = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\hat{\alpha}_i - \bar{\hat{\alpha}})^2}$
- systematic error: $BS = \bar{\hat{\alpha}} - \alpha$
- overall error: $RMSE = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\hat{\alpha}_i - \alpha)^2}$

Results

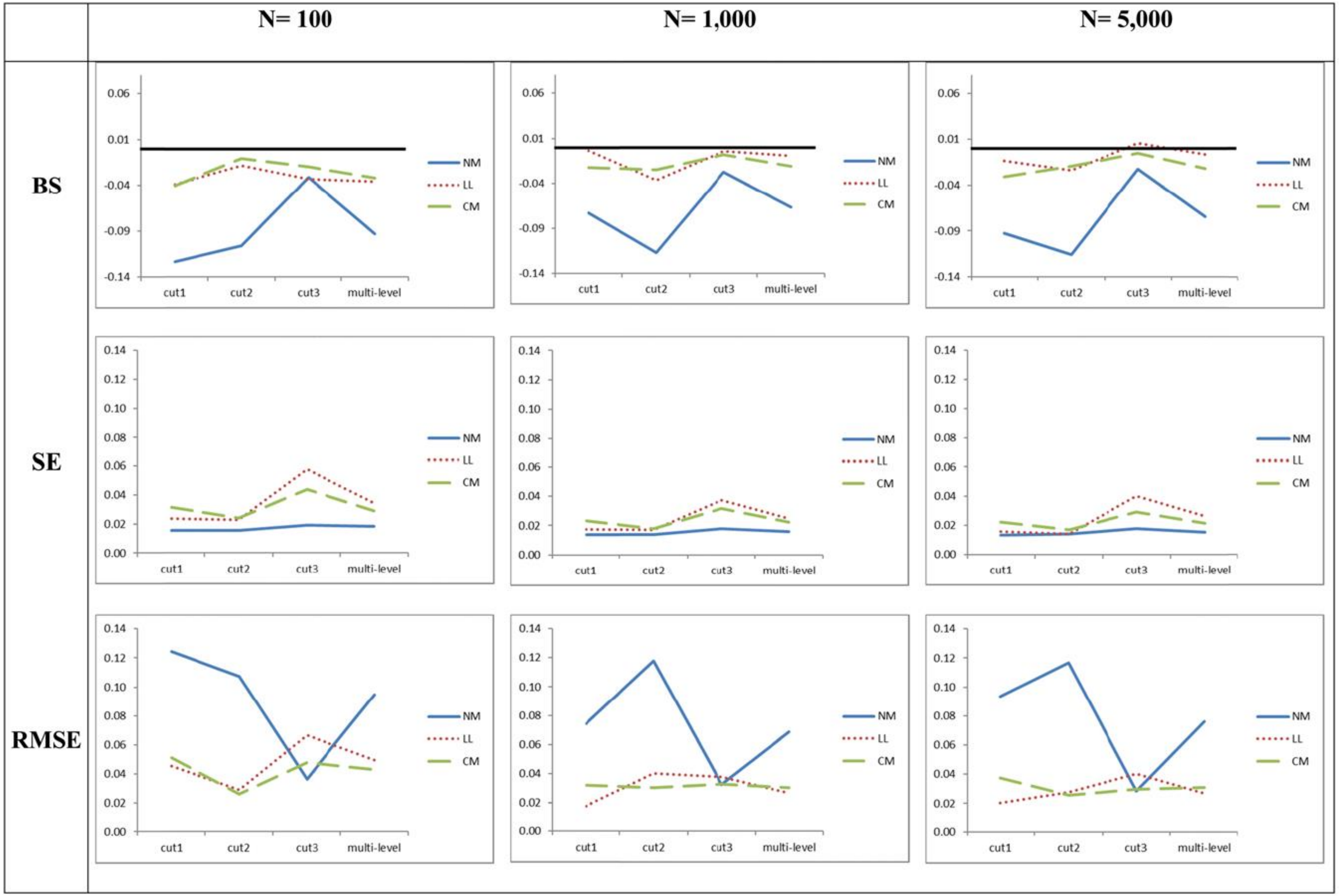
Table 3
Criterion Classification Indices

Study	N	Cons. Index	Cut1	Cut2	Cut3	Multilevel	Acc. Index	Cut1	Cut2	Cut3	Multilevel
Study 1	100	P	.9383	.9291	.8986	.7776	γ	.9573	.9521	.9303	.8413
	1,000		.9051	.9264	.8959	.7419		.9340	.9502	.9260	.8124
	5,000		.9175	.9288	.8960	.7557		.9435	.9508	.9259	.8225
Study 2	100		.7564	.6975	.7978	.4143		.8247	.7815	.8540	.5089
	1,000		.7447	.6976	.8180	.4215		.8161	.7778	.8734	.5176
	5,000		.7423	.7038	.8182	.4238		.8147	.7831	.8728	.5214
Study 3	100		.8497	.7825	.8235	.5148		.8969	.8491	.8735	.6309
	1,000		.8346	.7790	.8469	.5212		.8816	.8416	.8939	.6295
	5,000		.8311	.7874	.8433	.5219		.8808	.8452	.8859	.6250
Study 1	100	Kappa	.8739	.8576	.7603	.6763	$\gamma+$	.0195	.0283	.0361	.0832
	1,000		.8065	.8501	.7333	.6299		.0275	.0233	.0380	.0880
	5,000		.8329	.8553	.7354	.6452		.0209	.0281	.0370	.0849
Study 2	100		.3659	.3920	.3013	.2120		.1229	.1122	.0253	.2361
	1,000		.3686	.3834	.3250	.2181		.1243	.0927	.0240	.2211
	5,000		.3781	.3930	.3105	.2200		.1176	.0932	.0234	.2144
Study 3	100		.5032	.5645	.5163	.3476		.0737	.0497	.0483	.1666
	1,000		.5249	.5594	.5265	.3577		.0723	.0723	.0335	.1726
	5,000		.5263	.5756	.5086	.3584		.0665	.0647	.0327	.1586
Study 1	100						$\gamma-$	.0232	.0196	.0336	.0755
	1,000							.0385	.0265	.0360	.0996
	5,000							.0356	.0211	.0371	.0926
Study 2	100							.0524	.1064	.1208	.2551
	1,000							.0596	.1295	.1025	.2612
	5,000							.0677	.1238	.1037	.2642
Study 3	100							.0294	.1012	.0782	.2025
	1,000							.0461	.0861	.0726	.1980
	5,000							.0528	.0900	.0814	.2163

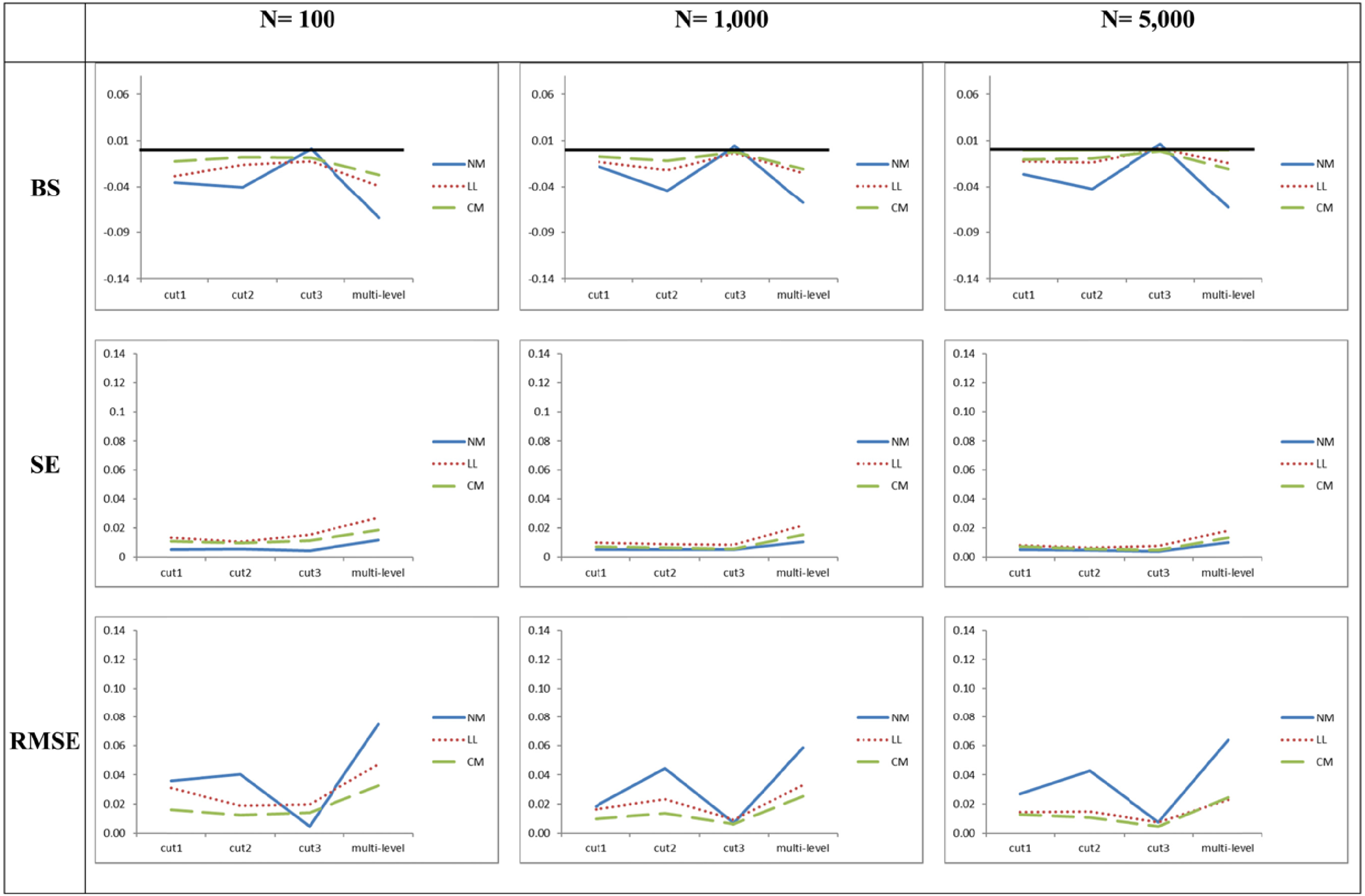
- Study 1: Bimodal Distribution (Agreement index P)



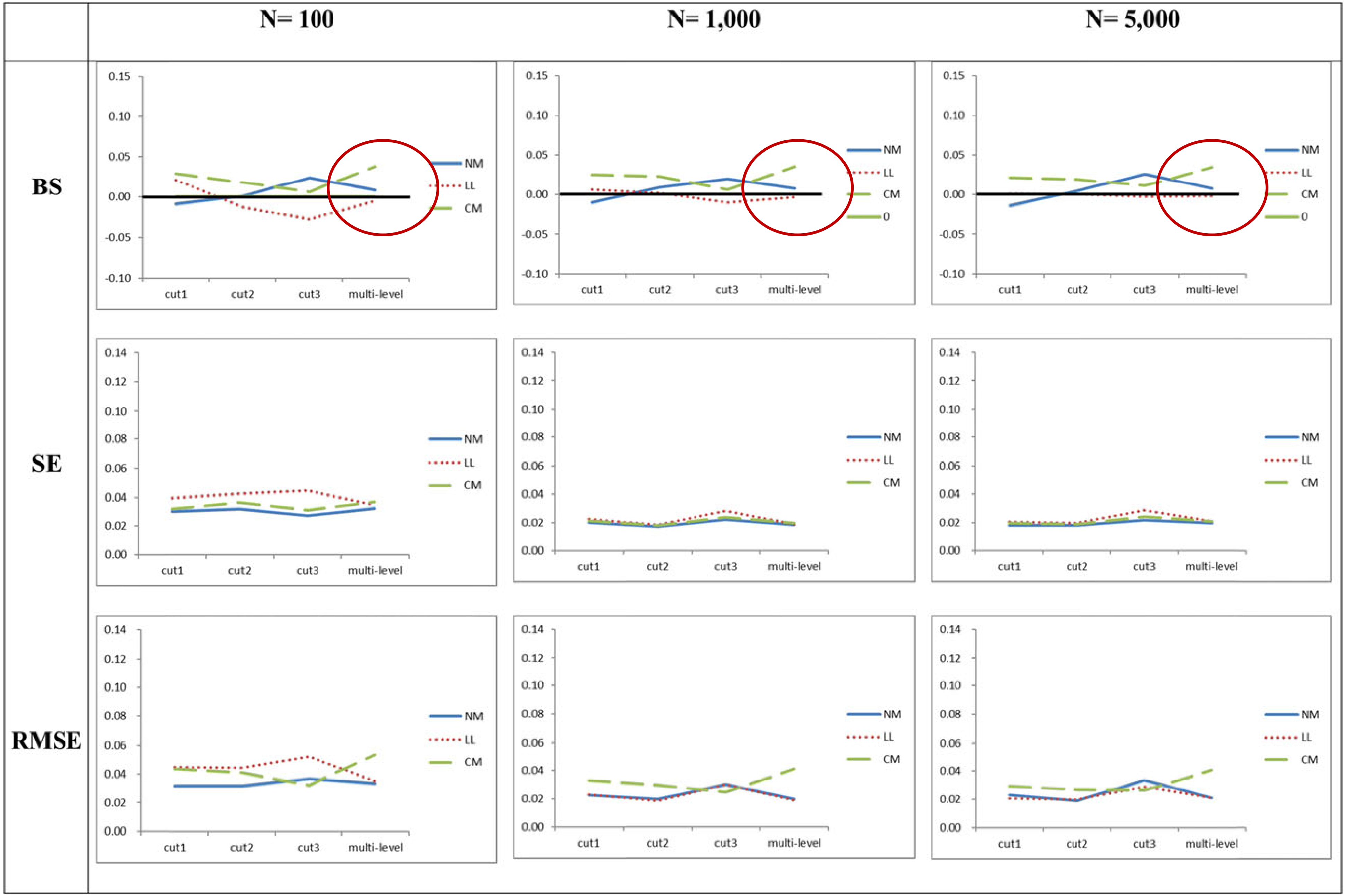
- Study 1: Bimodal Distribution (Kappa coefficient)



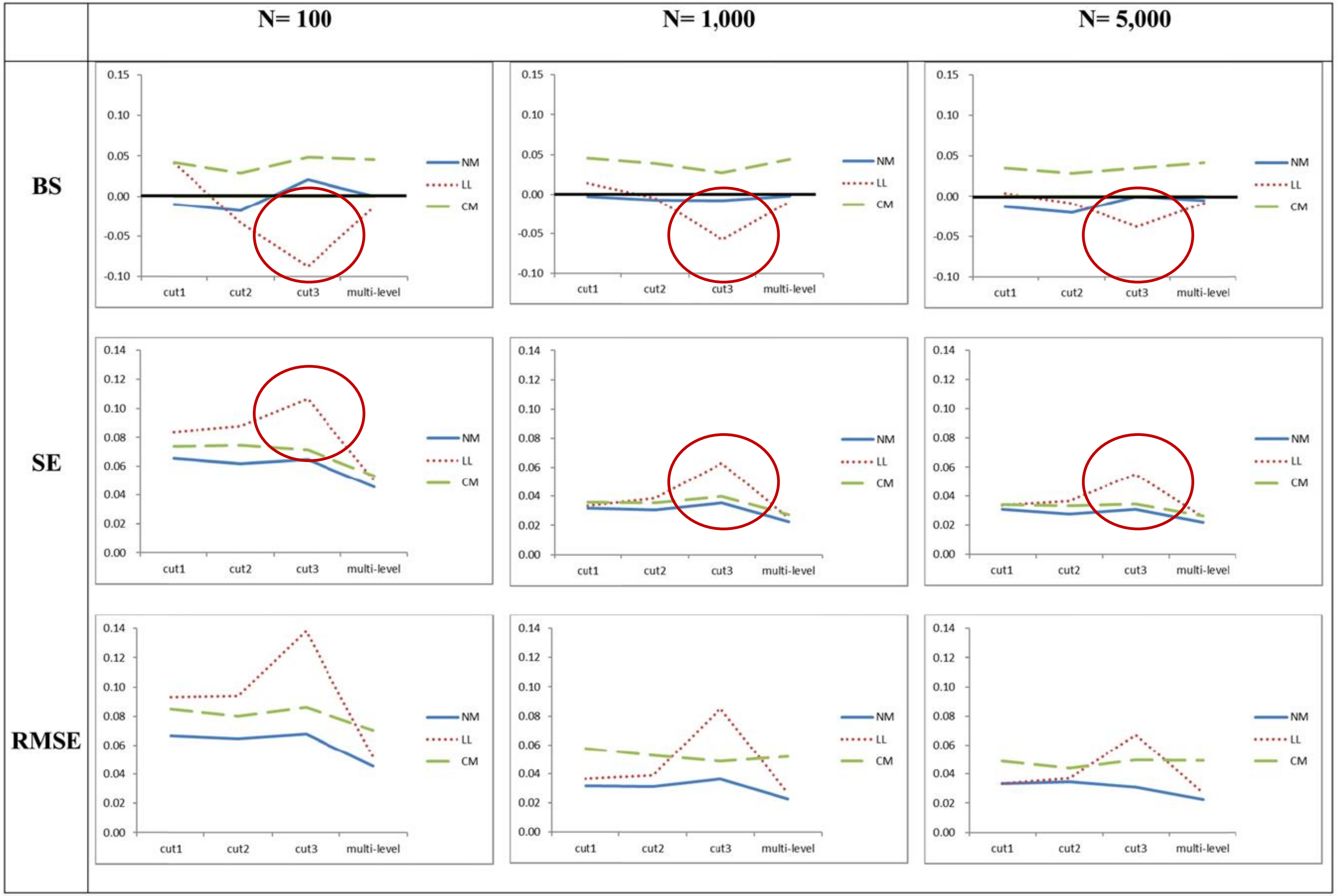
- Study 1: Bimodal Distribution (Gamma index)



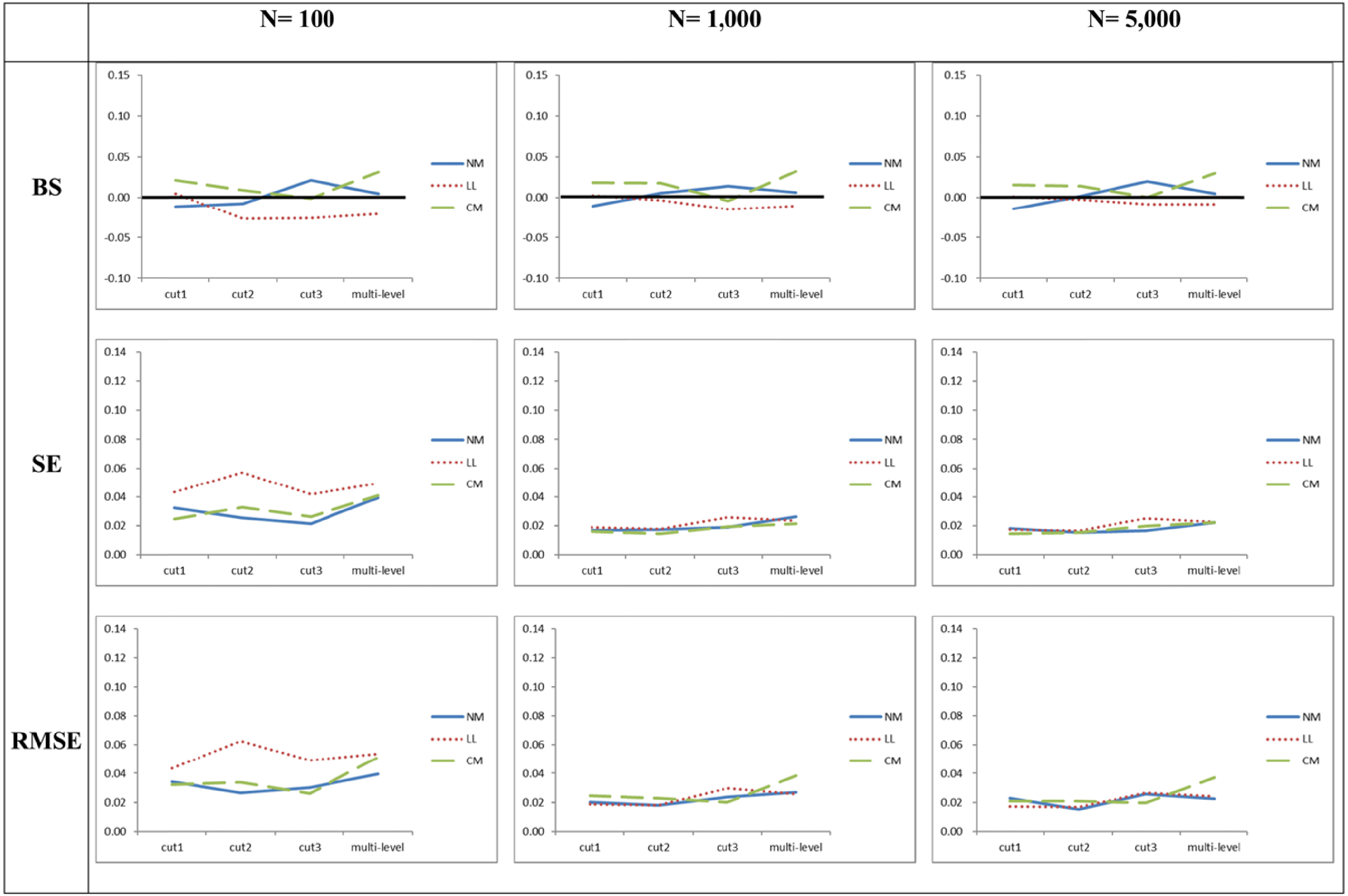
- Study 2: Distribution with Structural Bumpiness (Agreement index P)



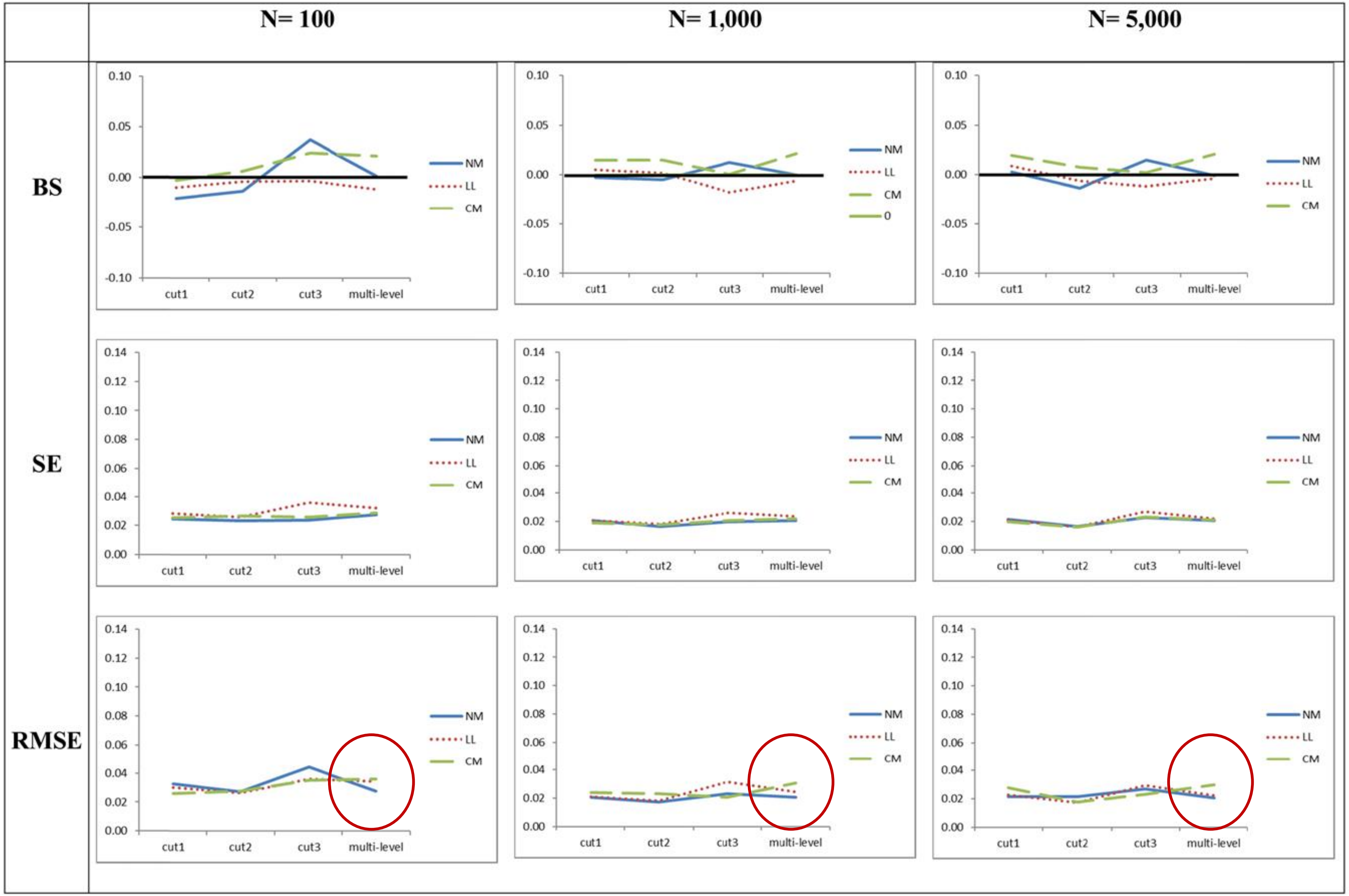
- Study 2: Distribution with Structural Bumpiness (Kappa coefficient)



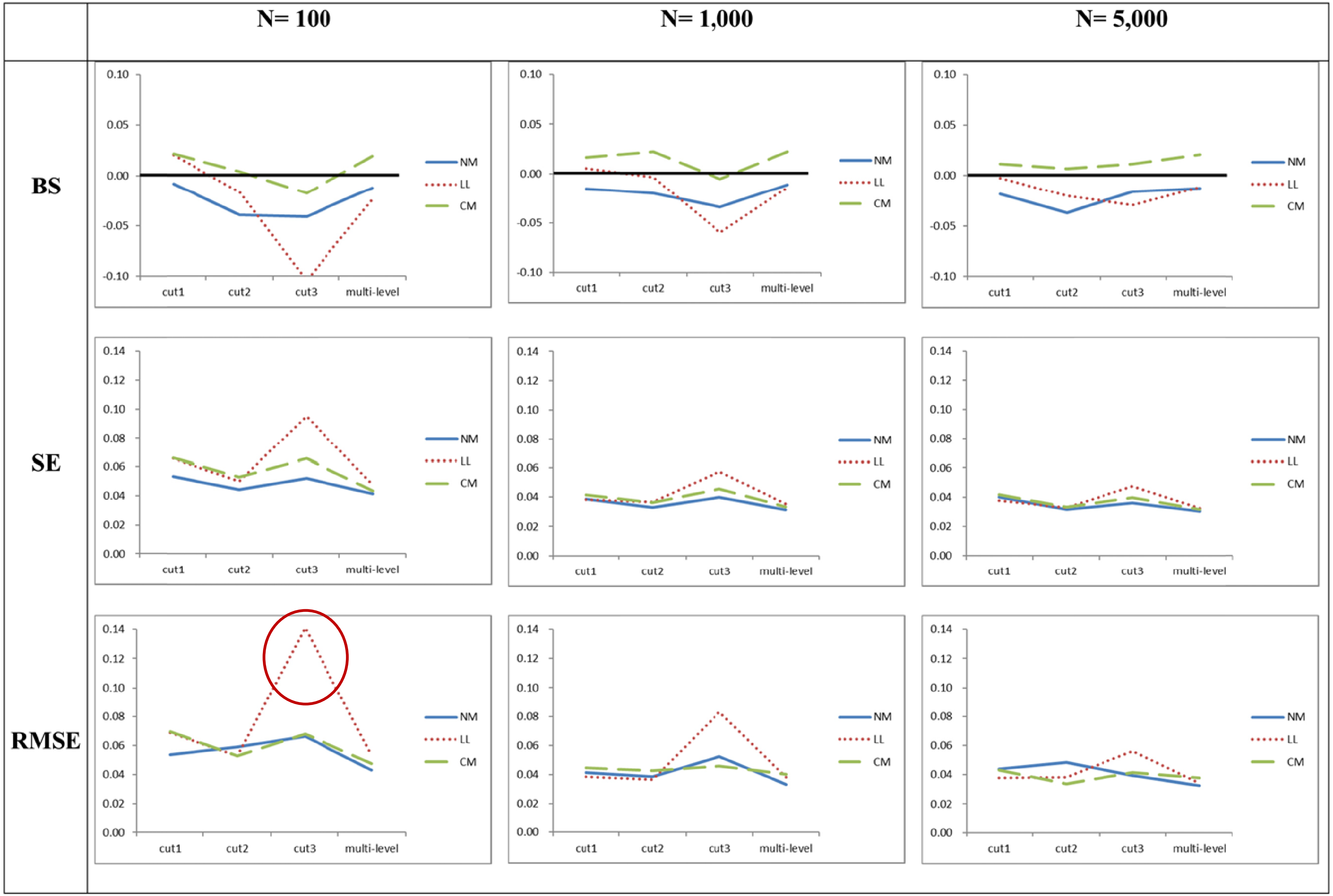
- Study 2: Distribution with Structural Bumpiness (Gamma index)



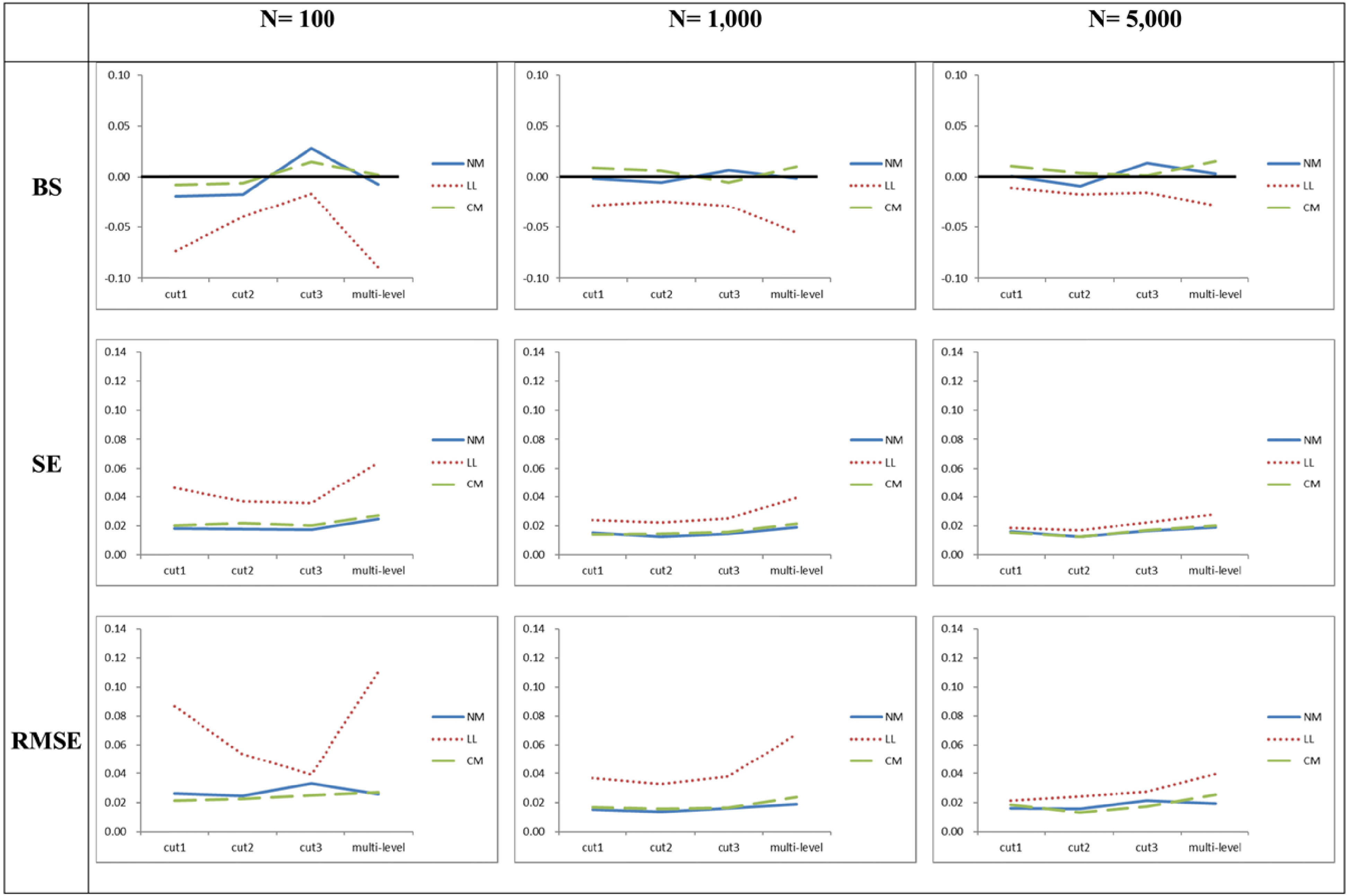
- Study 3: Distribution with Structural Zeros (Agreement index P)



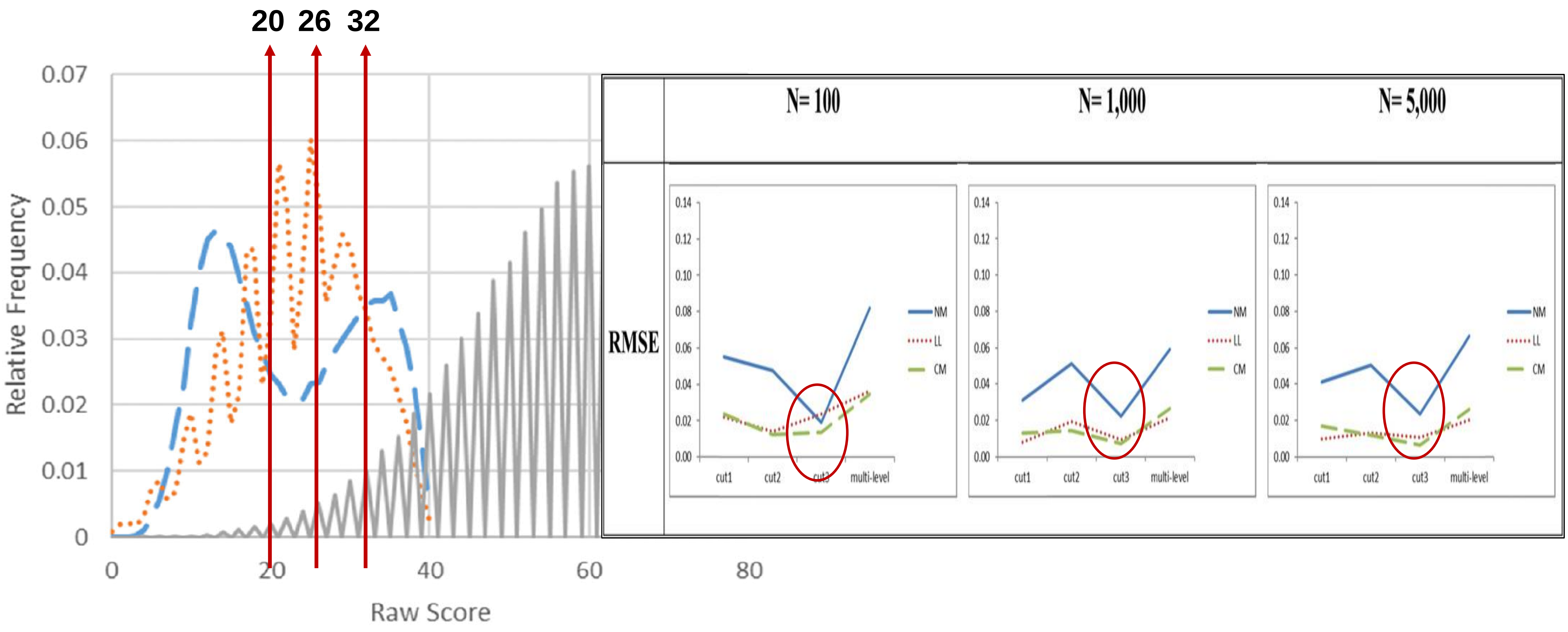
• Study 3: Distribution with Structural Zeros (Kappa coefficient)



- Study 3: Distribution with Structural Zeros (Gamma index)



- Impact of Cut Score Location



Discussion

- it seems prudent to advise that the user **consider factors such as**:
  - the type of score distribution
  - sample size
  - cut score location
  - the unique assumptions about test form parallelism
  - data structure

Limitations

- might not exactly reflect data structure observed in **real data**
- **only non-IRT** procedures were investigated
- **only a single test length** was considered
- the criterion classification indices were **obtained for each sample-size condition** separately

THE END
YINGSHI HUANG

Thanks for listening!