Standard Errors and Confidence Intervals of Norm Statistics for Educational and Psychological Tests

心理与教育测验中常模统计量的标准误与置信区间

Hannah E. M. Oosterhuis

Psychometrika (IF = 2.111)

报告人：黄颖诗

目录

一、问题提出

总体 → 代表性样本（标准化样本） → 常模统计量

常模统计量 → 估计总体参数 → 总体

估计总体参数 → 有偏

一、问题提出

误差会给结果带来怎样的影响？

↓

例子

一、问题提出

- 学前和幼儿园行为量表（PKBS）：社交技能水平(Merrell, 1994)
- 评分标准：

显著缺乏	中度缺乏	平均水平

5th (59) 20th (76)

- Olive & Jack & Harry

|| || ||

56 82 61

一、问题提出

a. 不考虑测量误差与抽样误差

① O & J & H的观测分数等同于真分数

② 由常模样本得出的常模统计量等同于总体参数

一、问题提出

O属于显著缺陷；J属于平均水平；H属于中度缺陷

一、问题提出

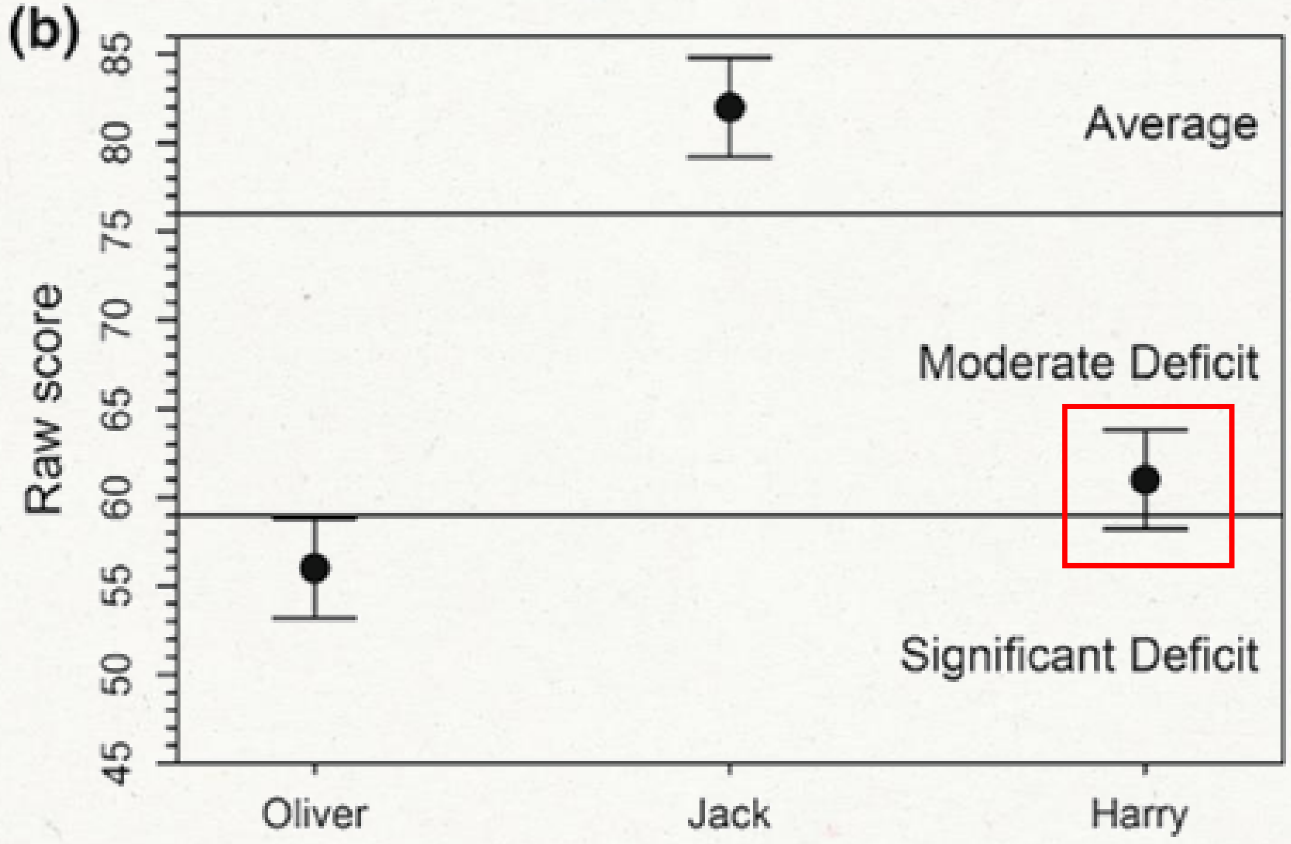
b. 考虑测量误差

① 68%CI得出O & J & H的分数带

Oliver	[53.2, 58.8]
Jack	[79.2, 84.8]
Harry	[58.2, 63.8]

② 由常模样本得出的常模统计量等同于总体参数

一、问题提出



O属于显著缺陷；J属于平均水平；不确定H属于中度缺陷还是显著缺陷

一、问题提出

c. 考虑测量误差与抽样误差

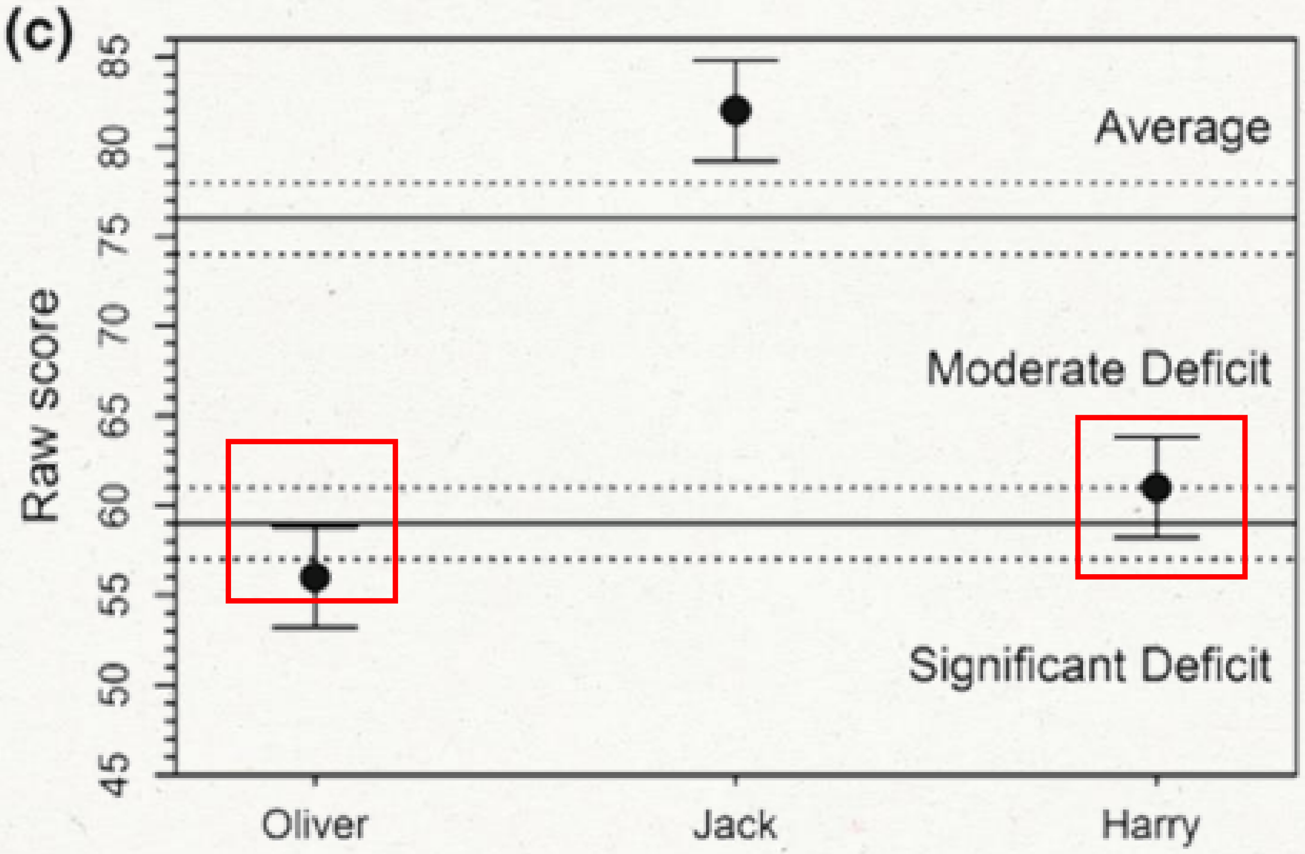
① 68%CI得出O & J & H的分数带

Oliver	[53.2, 58.8]
Jack	[79.2, 84.8]
Harry	[58.2, 63.8]

② 95%CI得出百分等级的边界

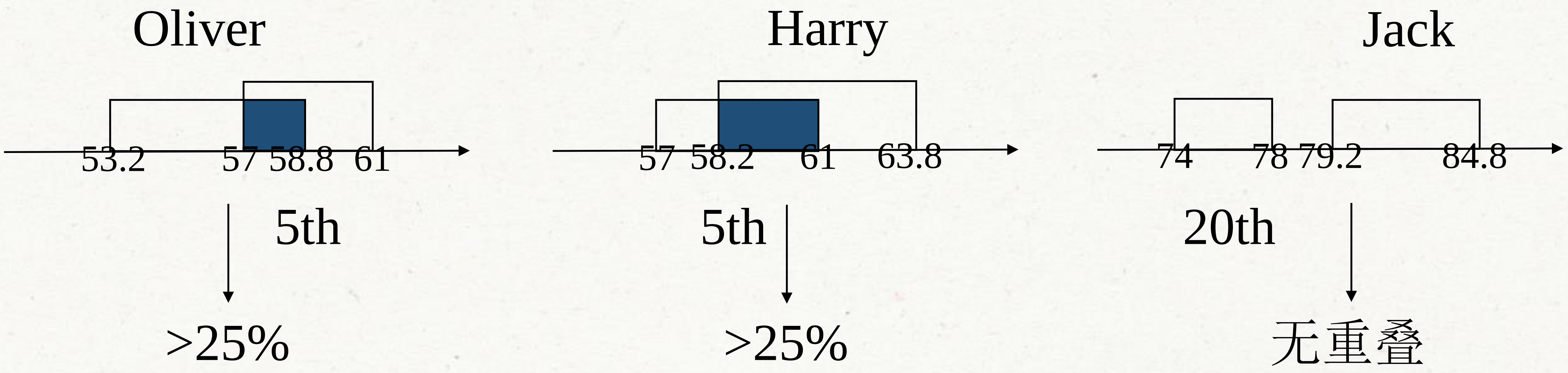
5th	[57, 61]
20th	[74, 78]

一、问题提出

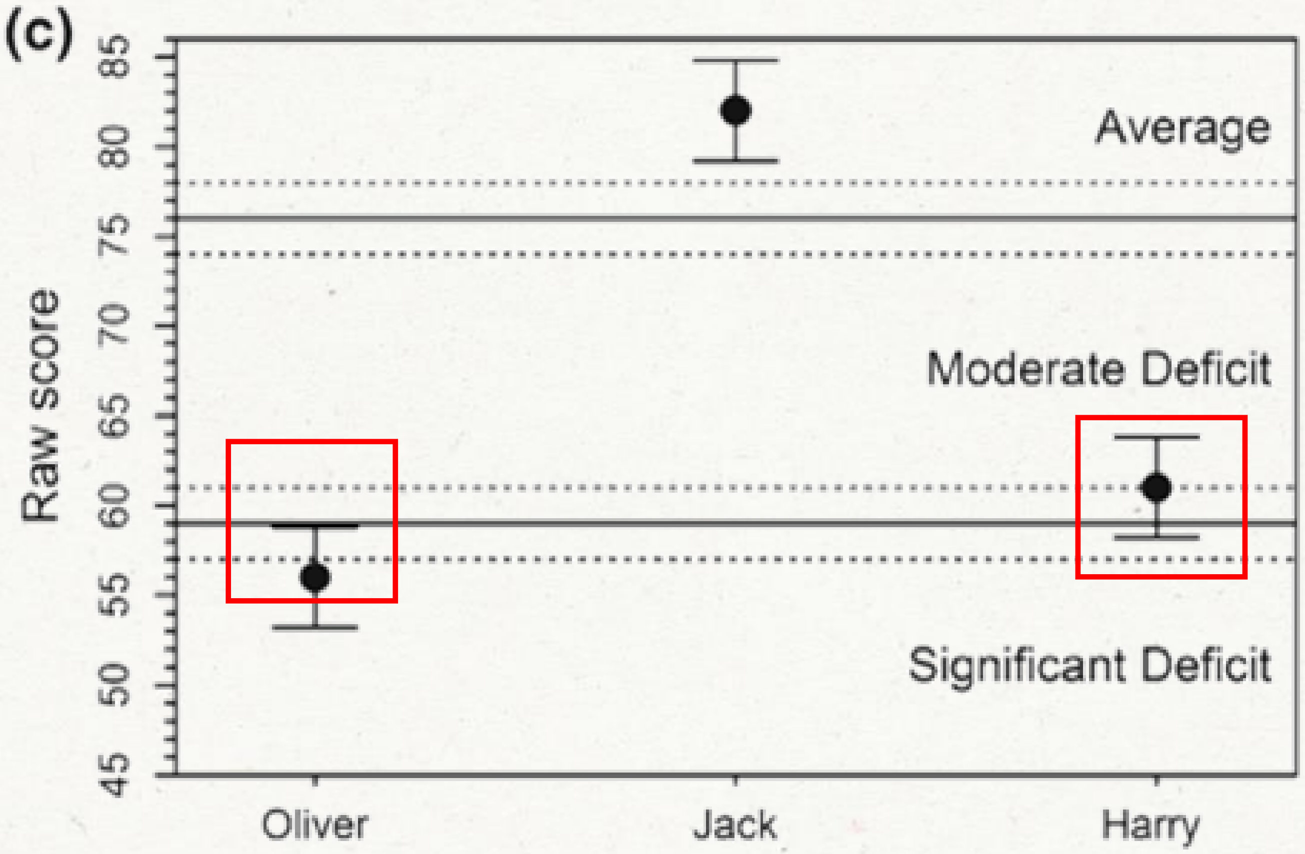

(c)
Raw score
85
80
75
70
65
60
55
50
45
Average
Moderate Deficit
Significant Deficit
Oliver
Jack
Harry

一、问题提出

根据启发式规则(heuristic rule)，两个统计量之间重叠的区域>25%时，两者差异不显著(Van Belle, 2003, Sect. 2.6)。



一、问题提出



J属于平均水平；不确定O与H属于中度缺陷还是显著缺陷

一、问题提出

- 误差对结果的解释存在影响
- 在临界值附近的值受到的影响很大

什么因素会影响估计精度？
- 样本量(Crawford & Howell,1998)
- 统计量在常模样本分布中的位置 (Brennan&Lee,1999;Leeetal.,2000)

如何量化抽样的误差？——与标准误(SE)直接相关的Wald-based置信区间(CI)

一、问题提出

推导SEs & CIs的难点

- SE难以推导，常用软件包的缺乏
- 对各种常模统计量的SEs & CIs估计不全
- 测验分数不满足前提假设（数据离散、分布偏态）

一、问题提出

能否在温和的假设条件下推导出SE？

分布、数据类型

① 服从多项式分布、近似正态分布

② 可用于离散、连续型数据

二、推导流程

一般框架

两步法(Two-Step Procedure)

①常模统计量 = f (原始分频率)

②计算常模统计量的方差近似解

广义指对数表示法(Generalized exp-log Notation)

求出向量函数$g(\hat{m})$与雅可比矩阵G

二、推导流程

- 测验分数均值：$S_{\bar{X}} = \sqrt{\sum_{j=1}^{k} \widehat{m}_j[(r_j - \bar{X})/N]^2}$

- 标准差：$S_{s_X} \approx 0.5 s_X \sqrt{\sum_i \sum_j \left(\frac{d_i^2}{SS} - e\right)\left(\frac{d_j^2}{SS} - e\right)\left(\delta_{ij}\hat{m}_i - \frac{\hat{m}_i \hat{m}_j}{N}\right)}$（27）

Ahn & Fessler (2003) 假设数据服从正态分布下得出：

$\dot{S}_{s_X} = \frac{s_X}{\sqrt{2(N-1)}}$（29）

二、推导流程

- 百分等级分数：$S_{PR_x} \approx \frac{50}{N}\sqrt{\sum_i \hat{m}_i \left(\gamma_{xi} - P(X < x) - P(X \le x)\right)^2}$

- 标准九分：

$$S_{St_b} \approx \sqrt{\sum_{j=1}^{k}\sum_{i=1}^{k}\left[\delta_{ij}\hat{m}_j - \frac{\hat{m}_i\hat{m}_j}{N}\right] \cdot \left[\frac{f_b s_X}{2} \cdot (d_i^* - e) + \frac{d_i}{N}\right] \cdot \left[\frac{f_b s_X}{2} \cdot \left(d_j^* - e\right) + \frac{d_j}{N}\right]}$$

- Z分数：

$$S_{Z_h} \approx \sqrt{\sum_{j=1}^{k}\sum_{i=1}^{k}{}^{t}\left[-\bar{X} \cdot \left(\frac{r_i}{\sum X} - \frac{1}{N} - u_i\right) - r_h u_i\right] \cdot \left[-\bar{X}\left(\frac{r_j}{\sum X} - \frac{1}{N} - u_j\right) - r_h u_j\right]}$$

三、模拟研究

➢ 数据产生：

αj = 0.85,0.95,1.05, … ,1.75

- 两参logistic模型(2PLM)，通过斜率参数$\boxed{\alpha_j}$和位置参数$\boxed{\beta_j}$描述对应每个项目j (0,1计分)的作答概率；

β_j = -2.25,-1.75,-1.25, … ,2.25　从标准正态分布中随机抽取

生成项目分数向量

$$\boxed{P(}X_j = \boxed{1|\theta)} = \frac{\exp[\alpha_j(\theta-\beta_j)]}{1+\exp[\alpha_j(\theta-\beta_j)]}$$

- 重复 Q = 10,000次

三、模拟研究

- 模拟条件：
  - 自变量：题目数量 J = 10,30,50 (Oosterhuis et al., 2016)

    样本规模 N = 500,1000,1500,2000,2500

    共包括3×5 = 15种组合
  - 因变量：Bias，均方根误差(RMSE)，95%CI覆盖率

三、模拟研究

- Bias

$\bar{\theta} = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\theta}_q$ （常模统计量θ的均值）

$$s_{\hat{\theta}} = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}\left(\hat{\theta}_q - \bar{\theta}\right)^2}$$ （Q次重复的θ的标准差）⟶ SE真值

$$bias.se = \frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{S}_{\hat{\theta}_q} - s_{\hat{\theta}}\right)$$ （$\hat{S}_{\hat{\theta}q}$为第q次θ的标准误估计）

- RMSE

$$RMSE_{\hat{S}_{\hat{\theta}}} = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{S}_{\hat{\theta}_q} - s_{\hat{\theta}}\right)^2}$$

三、模拟研究

标准差与标准九分的尺度依赖性

↓

不同样本和测验特征情况下不可比

$$\bar{s}_Y = \frac{1}{Q}\sum_{q=1}^{Q} \boxed{s_{Y_q}}$$（模拟分数的平均标准差）

第q次重复中原始分数的标准差

↓

校正bias与RMSE，使其可比

三、模拟研究

- 95%置信区间覆盖率

覆盖率越接近0.95, 表明结果越可靠。

计算95%CI覆盖率的方法是: 通过判断参数是否落入0.25%到97.5%两分位点对应的方差分量之间, 如果某次成功, 则包含次数加1, 最后计算落入的总次数, 并除以10000, 即为最后的覆盖率。

三、模拟研究

- 结果

先对标准差及标准九分的Bias与RMSE进行校对

即计算出题目数量为10,30,50时

$\bar{s}_Y$分别等于2.076（10）,5.530（30）,8.966（50）

三、模拟研究

- 标准差

SE(27)在任何组合条件下均未表现出偏差；

并且CIs覆盖率均接近0.95。

TABLE 1.
Standardized bias, standardized RMSE, and coverage probability of 95% CIs of the estimated SEs of the standard deviation.

Items	N	Bias		RMSE		Coverage	
		S_{s_X}	$\dot{S}_{s_X}$*	S_{s_X}	$\dot{S}_{s_X}$*	S_{s_X}	$\dot{S}_{s_X}$*
10	500	−0.0001	0.0037	0.0010	0.0038	0.949	0.973
	1000	−0.0001	0.0026	0.0005	0.0026	0.951	0.974
	1500	−0.0001	0.0021	0.0003	0.0021	0.947	0.971
	2000	0.0000	0.0019	0.0002	0.0019	0.950	0.97
	2500	0.0000	0.0017	0.0002	0.0017	0.951	0.976
30	500	−0.0004	−0.0033	0.0012	0.0034	0.946	0.970
	1000	−0.0002	0.0024	0.0006	0.0024	0.949	0.972
	1500	0.0000	0.0021	0.0004	0.0021	0.944	0.971
	2000	−0.0001	0.0016	0.0003	0.0017	0.948	0.971
	2500	0.0000	0.0016	0.0002	0.0016	0.949	0.973
50	500	−0.0002	0.0035	0.0011	0.0036	0.951	0.975
	1000	−0.0001	0.0024	0.0006	0.0025	0.952	0.973
	1500	−0.0001	0.0020	0.0004	0.0020	0.952	0.975
	2000	0.0001	0.0019	0.0003	0.0019	0.950	0.973
	2500	−0.0001	0.0014	0.0003	0.0015	0.95	0.972

* SE based on Ahn & Fessler (2003, Eq. 29).

三、模拟研究

- 标准差

SE(29) 表现出较小的正向偏差；

并且CIs覆盖率均大于0.95。

TABLE 1.
Standardized bias, standardized RMSE, and coverage probability of 95% CIs of the estimated SEs of the standard deviation.

Items	N	Bias		RMSE		Coverage	
		S_{s_X}	$\dot{S}_{s_X}*$	S_{s_X}	$\dot{S}_{s_X}*$	S_{s_X}	$\dot{S}_{s_X}*$
10	500	−0.0001	0.0037	0.0010	0.0038	0.949	0.973
	1000	−0.0001	0.0026	0.0005	0.0026	0.951	0.974
	1500	−0.0001	0.0021	0.0003	0.0021	0.947	0.971
	2000	0.0000	0.0019	0.0002	0.0019	0.950	0.97
	2500	0.0000	0.0017	0.0002	0.0017	0.951	0.976
30	500	−0.0004	−0.0033	0.0012	0.0034	0.946	0.970
	1000	−0.0002	0.0024	0.0006	0.0024	0.949	0.972
	1500	0.0000	0.0021	0.0004	0.0021	0.944	0.971
	2000	−0.0001	0.0016	0.0003	0.0017	0.948	0.971
	2500	0.0000	0.0016	0.0002	0.0016	0.949	0.973
50	500	−0.0002	0.0035	0.0011	0.0036	0.951	0.975
	1000	−0.0001	0.0024	0.0006	0.0025	0.952	0.973
	1500	−0.0001	0.0020	0.0004	0.0020	0.952	0.975
	2000	0.0001	0.0019	0.0003	0.0019	0.950	0.973
	2500	−0.0001	0.0014	0.0003	0.0015	0.95	0.972

* SE based on Ahn & Fessler (2003, Eq. 29).

三、模拟研究

- 标准差

精度&覆盖率：

SE(27) > SE(29)

TABLE 1.
Standardized bias, standardized RMSE, and coverage probability of 95% CIs of the estimated SEs of the standard deviation.

Items	N	Bias		RMSE		Coverage	
		S_{s_X}	$\dot{S}_{s_X}*$	S_{s_X}	$\dot{S}_{s_X}*$	S_{s_X}	$\dot{S}_{s_X}*$
10	500	−0.0001	0.0037	0.0010	0.0038	0.949	0.973
	1000	−0.0001	0.0026	0.0005	0.0026	0.951	0.974
	1500	−0.0001	0.0021	0.0003	0.0021	0.947	0.971
	2000	0.0000	0.0019	0.0002	0.0019	0.950	0.97
	2500	0.0000	0.0017	0.0002	0.0017	0.951	0.976
30	500	−0.0004	−0.0033	0.0012	0.0034	0.946	0.970
	1000	−0.0002	0.0024	0.0006	0.0024	0.949	0.972
	1500	0.0000	0.0021	0.0004	0.0021	0.944	0.971
	2000	−0.0001	0.0016	0.0003	0.0017	0.948	0.971
	2500	0.0000	0.0016	0.0002	0.0016	0.949	0.973
50	500	−0.0002	0.0035	0.0011	0.0036	0.951	0.975
	1000	−0.0001	0.0024	0.0006	0.0025	0.952	0.973
	1500	−0.0001	0.0020	0.0004	0.0020	0.952	0.975
	2000	0.0001	0.0019	0.0003	0.0019	0.950	0.973
	2500	−0.0001	0.0014	0.0003	0.0015	0.95	0.972

* SE based on Ahn & Fessler (2003, Eq. 29).

三、模拟研究

- 标准差

随着样本量的增加，两个SE的偏差均减小；

但不受测验长度的影响。

TABLE 1.
Standardized bias, standardized RMSE, and coverage probability of 95% CIs of the estimated SEs of the standard deviation.

Items	N	Bias		RMSE		Coverage	
		S_{sx}	$\dot{S}_{sx}*$	S_{sx}	$\dot{S}_{sx}*$	S_{sx}	$\dot{S}_{sx}*$
10	500	−0.0001	0.0037	0.0010	0.0038	0.949	0.973
	1000	−0.0001	0.0026	0.0005	0.0026	0.951	0.974
	1500	−0.0001	0.0021	0.0003	0.0021	0.947	0.971
	2000	0.0000	0.0019	0.0002	0.0019	0.950	0.97
	2500	0.0000	0.0017	0.0002	0.0017	0.951	0.976
30	500	−0.0004	−0.0033	0.0012	0.0034	0.946	0.970
	1000	−0.0002	0.0024	0.0006	0.0024	0.949	0.972
	1500	0.0000	0.0021	0.0004	0.0021	0.944	0.971
	2000	−0.0001	0.0016	0.0003	0.0017	0.948	0.971
	2500	0.0000	0.0016	0.0002	0.0016	0.949	0.973
50	500	−0.0002	0.0035	0.0011	0.0036	0.951	0.975
	1000	−0.0001	0.0024	0.0006	0.0025	0.952	0.973
	1500	−0.0001	0.0020	0.0004	0.0020	0.952	0.975
	2000	0.0001	0.0019	0.0003	0.0019	0.950	0.973
	2500	−0.0001	0.0014	0.0003	0.0015	0.95	0.972

* SE based on Ahn & Fessler (2003, Eq. 29).

三、模拟研究

- 标准差

两个SE的CIs覆盖率均不受样本量与测验长度的影响。

TABLE 1.
Standardized bias, standardized RMSE, and coverage probability of 95% CIs of the estimated SEs of the standard deviation.

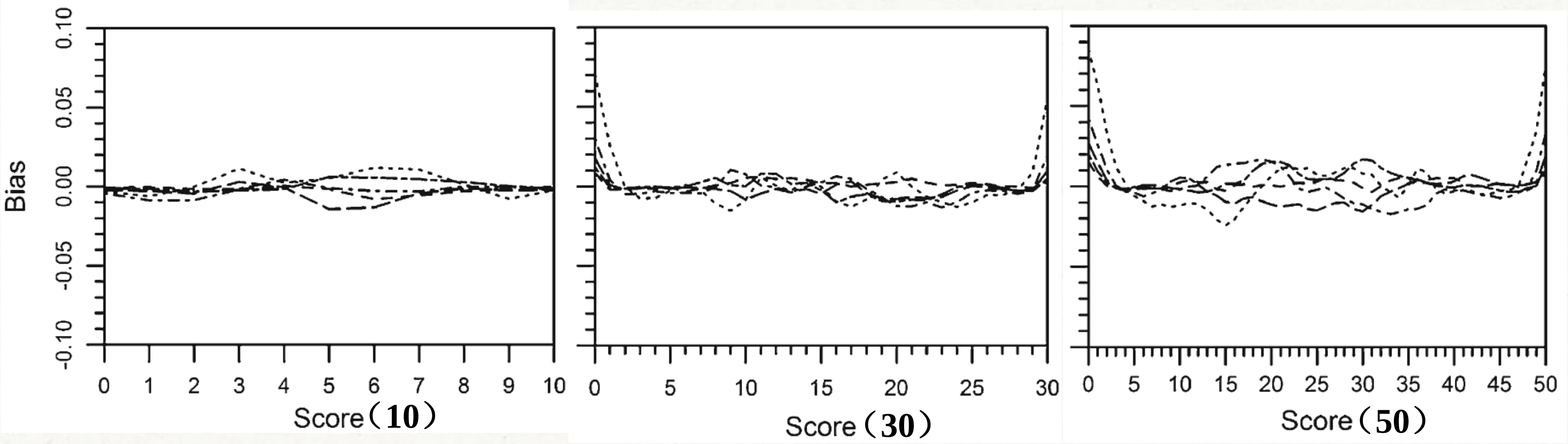
Items	N	Bias		RMSE		Coverage	
		S_{s_X}	$\dot{S}_{s_X}*$	S_{s_X}	$\dot{S}_{s_X}*$	S_{s_X}	$\dot{S}_{s_X}*$
10	500	−0.0001	0.0037	0.0010	0.0038	0.949	0.973
	1000	−0.0001	0.0026	0.0005	0.0026	0.951	0.974
	1500	−0.0001	0.0021	0.0003	0.0021	0.947	0.971
	2000	0.0000	0.0019	0.0002	0.0019	0.950	0.97
	2500	0.0000	0.0017	0.0002	0.0017	0.951	0.976
30	500	−0.0004	−0.0033	0.0012	0.0034	0.946	0.970
	1000	−0.0002	0.0024	0.0006	0.0024	0.949	0.972
	1500	0.0000	0.0021	0.0004	0.0021	0.944	0.971
	2000	−0.0001	0.0016	0.0003	0.0017	0.948	0.971
	2500	0.0000	0.0016	0.0002	0.0016	0.949	0.973
50	500	−0.0002	0.0035	0.0011	0.0036	0.951	0.975
	1000	−0.0001	0.0024	0.0006	0.0025	0.952	0.973
	1500	−0.0001	0.0020	0.0004	0.0020	0.952	0.975
	2000	0.0001	0.0019	0.0003	0.0019	0.950	0.973
	2500	−0.0001	0.0014	0.0003	0.0015	0.95	0.972

* SE based on Ahn & Fessler (2003, Eq. 29).

三、模拟研究

N = 500：点状
N = 1000：点虚线
N = 1500：长虚线
N = 2000：长短虚线
N = 2500：虚线

➢ 百分等级

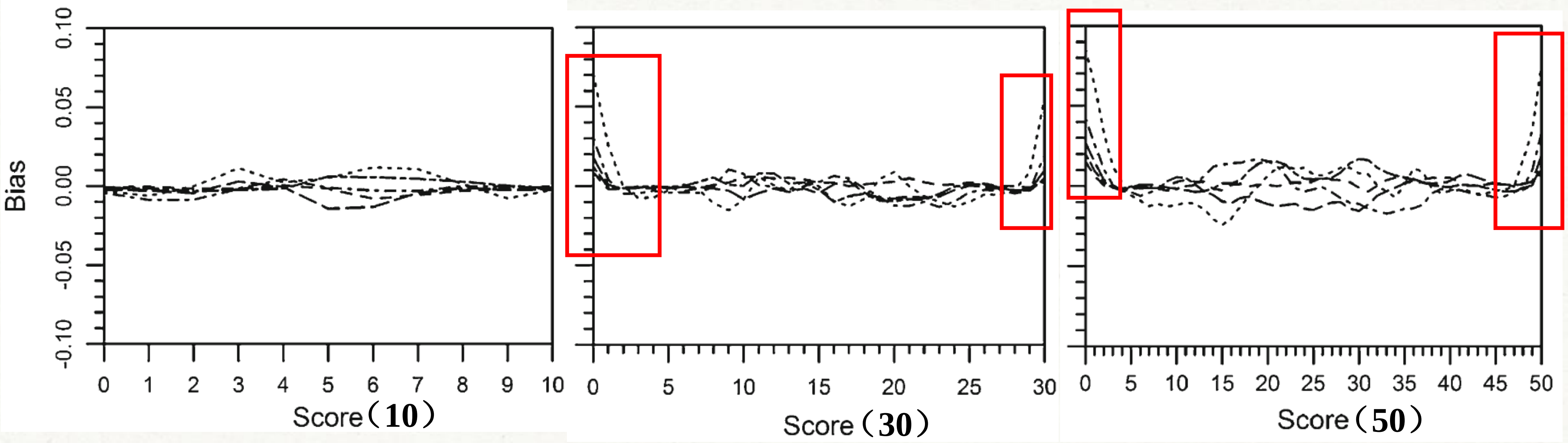


题目数为10在所有样本量组合条件下，均未表现出偏差。

三、模拟研究

N = 500：　点状
N = 1000：点虚线
N = 1500：长虚线
N = 2000：长短虚线
N = 2500：虚线

➢ 百分等级

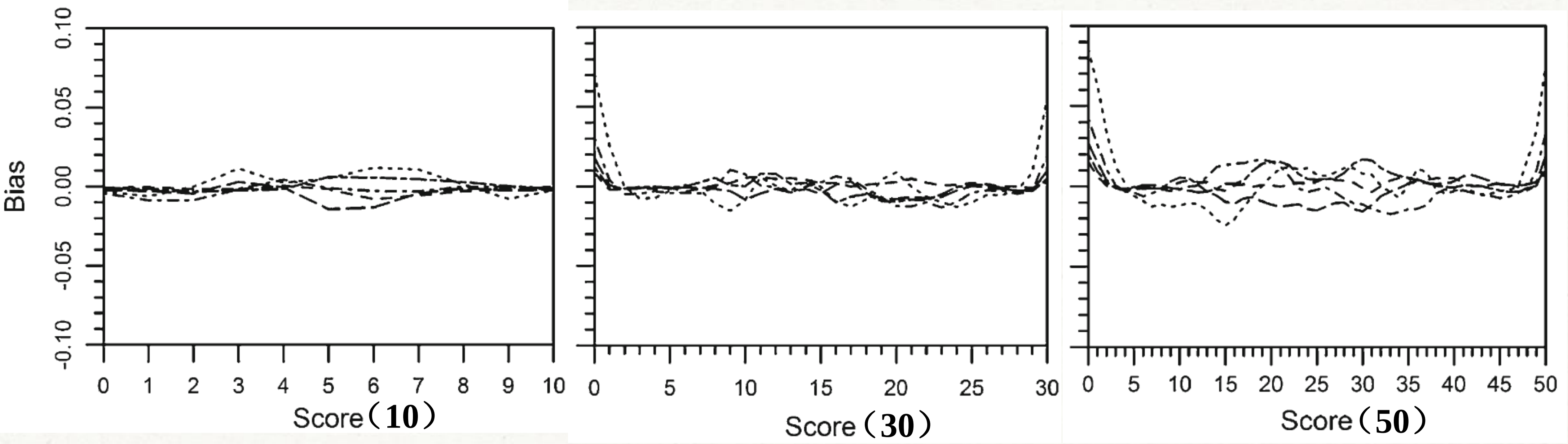


题目数≥30时，原始分最高与最低处表现出较小的正向偏差

三、模拟研究

N = 500：　点状
N = 1000：点虚线
N = 1500：长虚线
N = 2000：长短虚线
N = 2500：虚线

- 百分等级

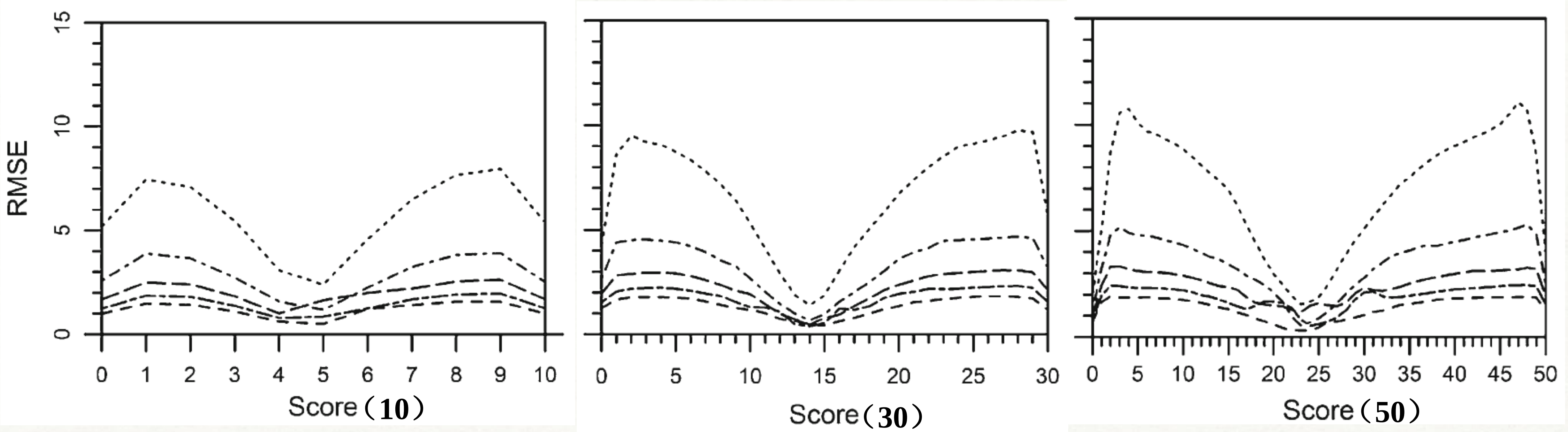


在N = 500条件下，SE估计精度随题目数增加而提高，
对于其他样本量大小，测验长度不影响估计精度。

三、模拟研究

N = 500： 点状
N = 1000：点虚线
N = 1500：长虚线
N = 2000：长短虚线
N = 2500：虚线

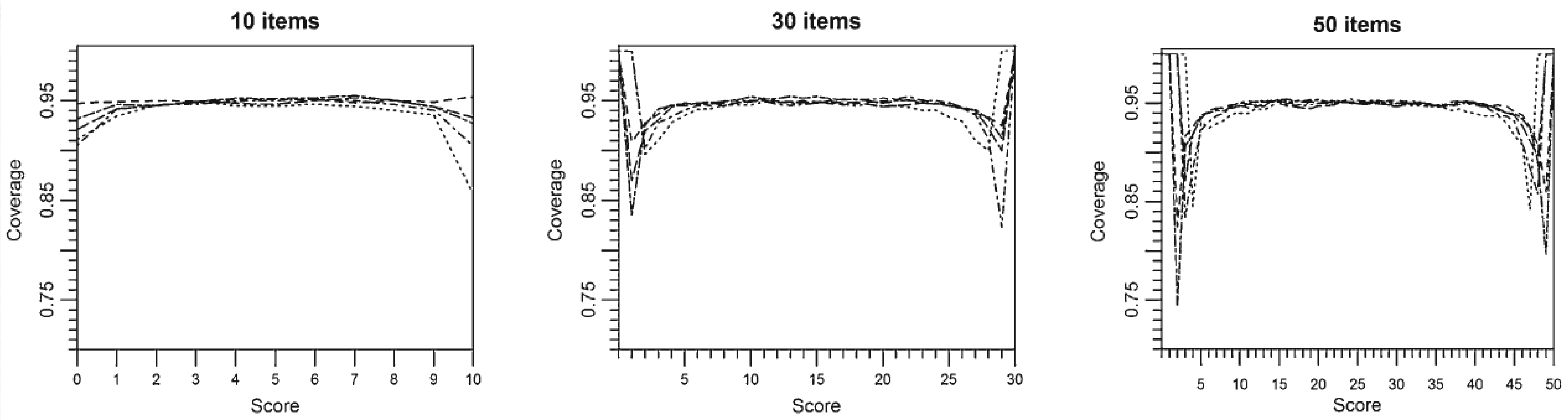
- 百分等级



RMSE先随着分数逐渐极端而增加，后在最极端处减小。

三、模拟研究

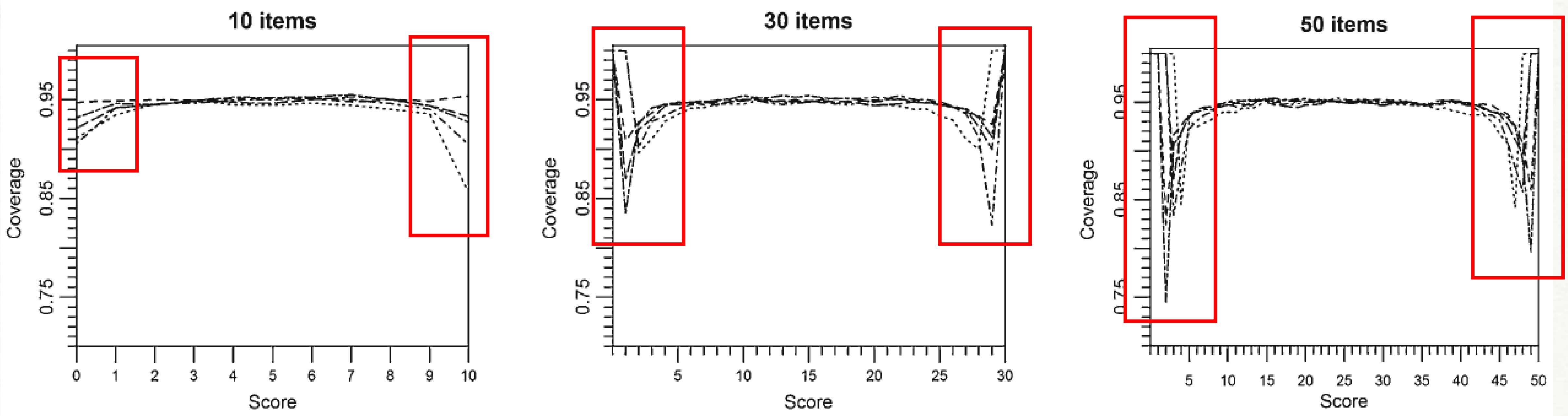
➤ 百分等级



在所有组合条件下，除非原始分数接近极端值，CI覆盖率均接近0.95。

三、模拟研究

➢ 百分等级



当原始分数逐渐接近极端值，CI覆盖率下降；

对于最极端分数，CI覆盖率急剧上升。

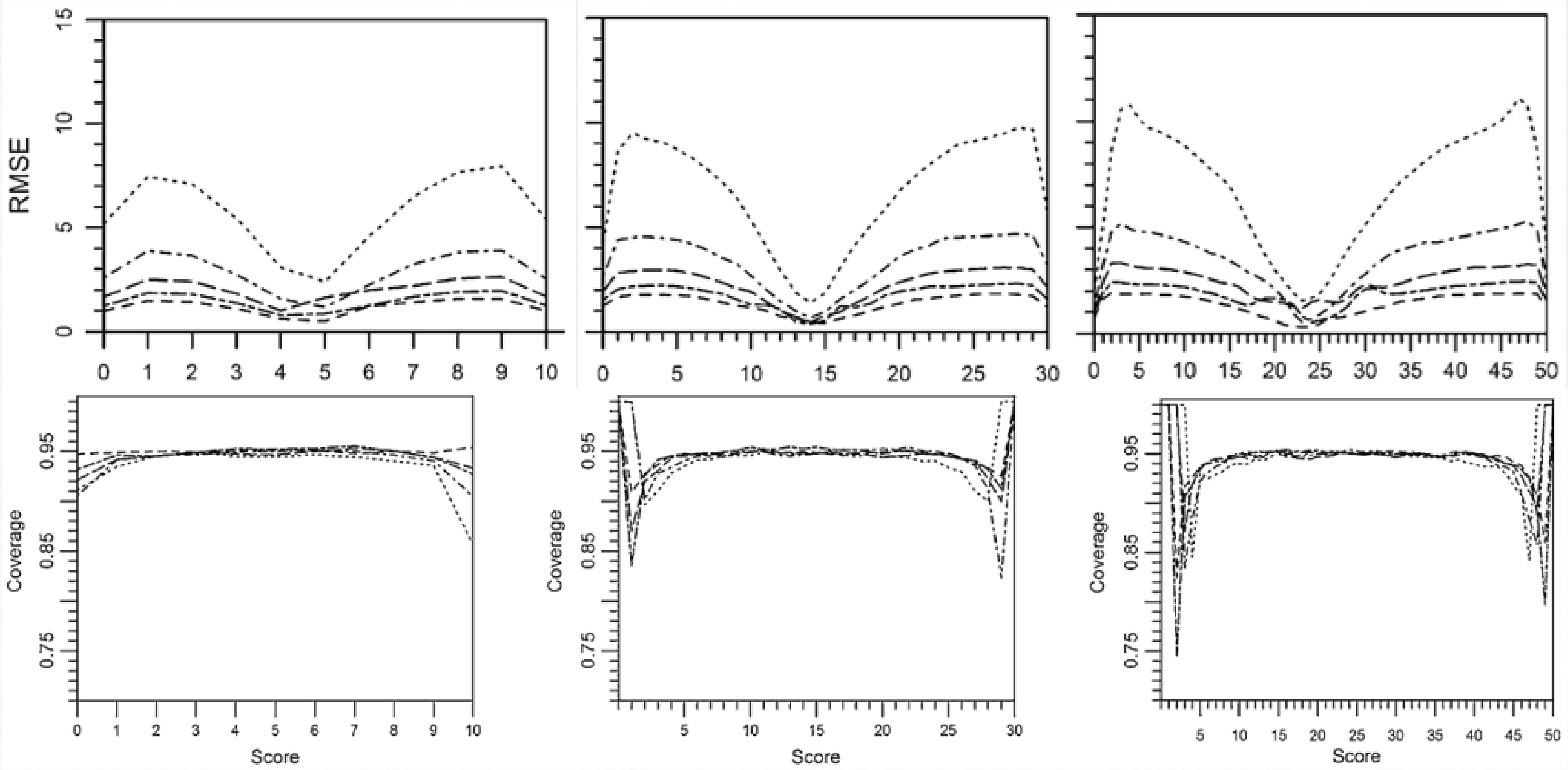
三、模拟研究

- 百分等级

对于RMSE的增加与CI覆盖率的降低

↓

远离平均数的观测分数的下降



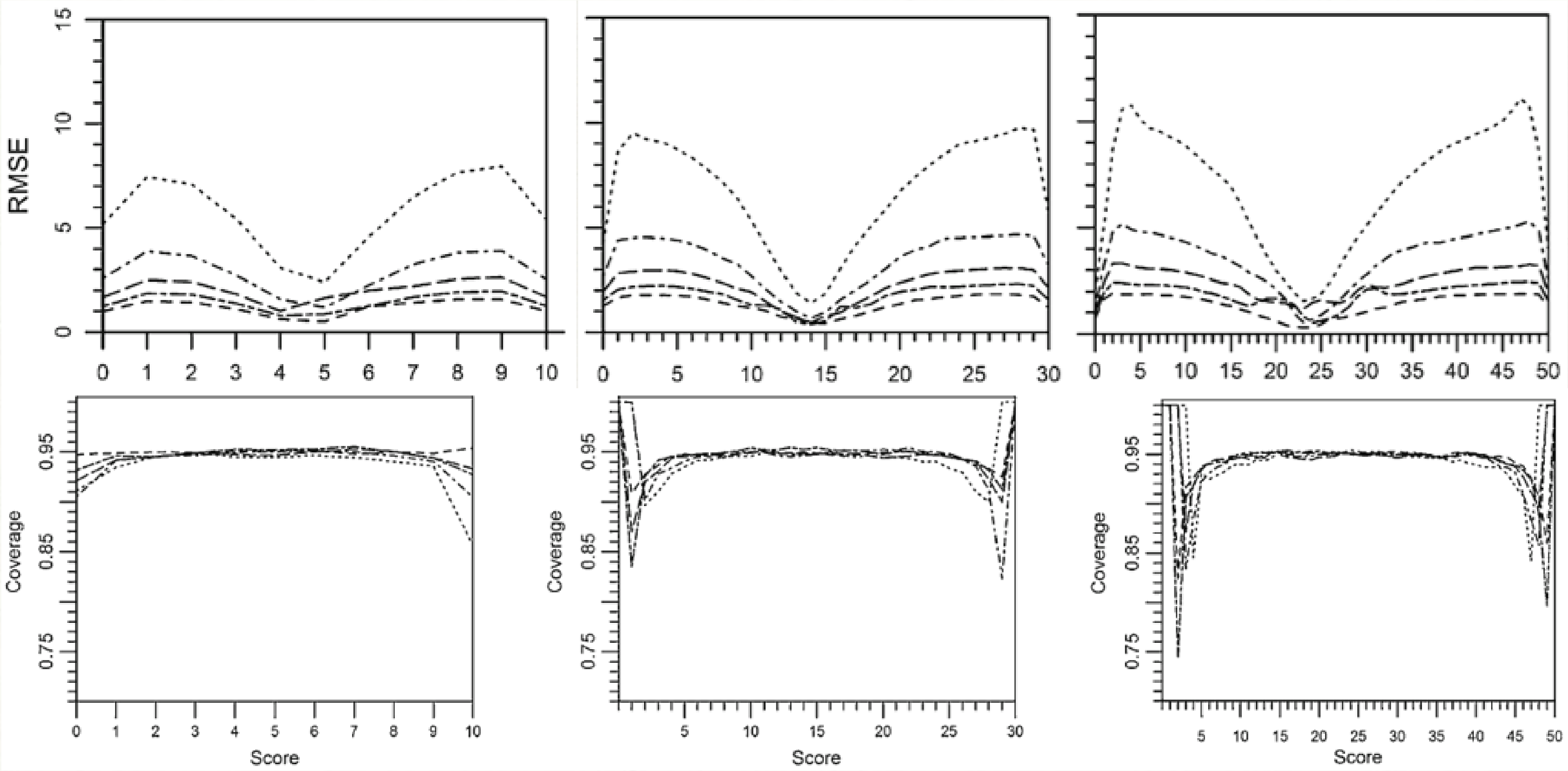
三、模拟研究

- 百分等级

对于RMSE的减小与CI覆盖率的增加

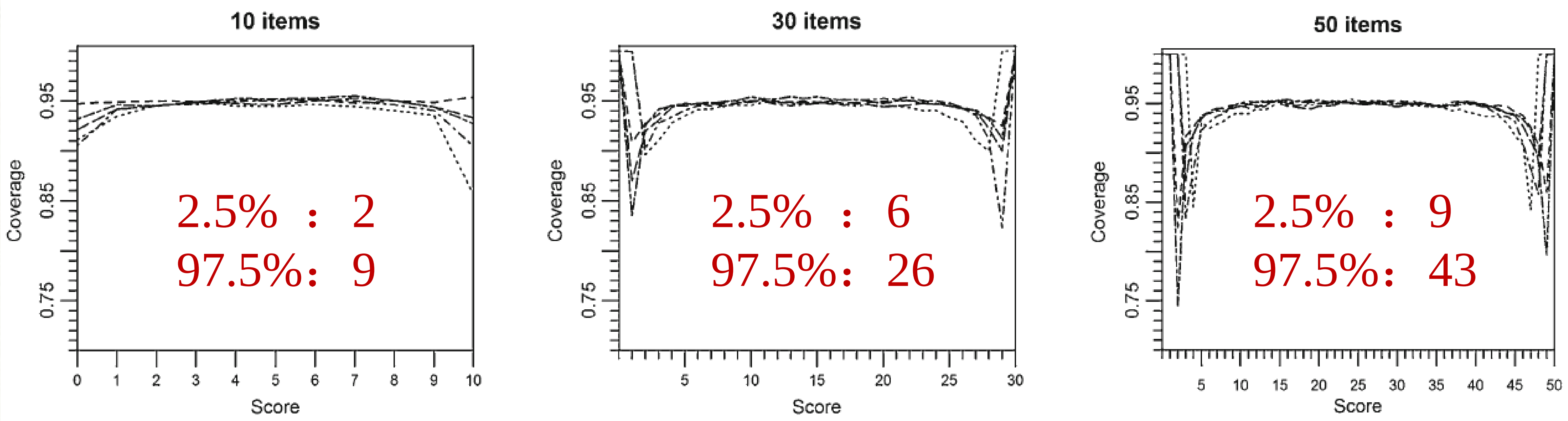
↓

百分等级分别接近0和100



三、模拟研究

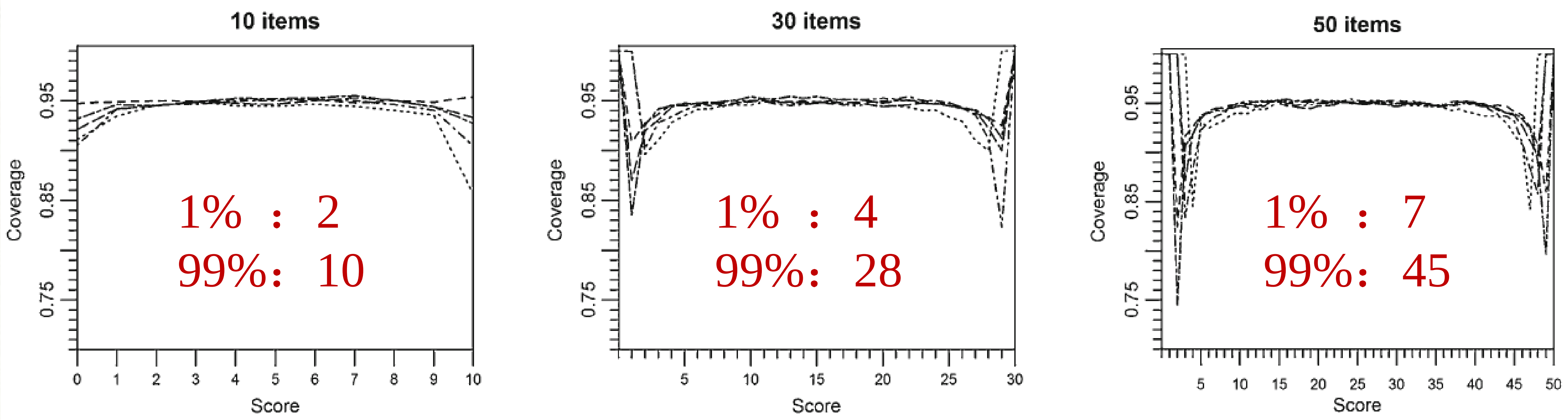
➢ 百分等级



当N = 500，2.5% ≤ 总体百分等级分数 ≤ 97.5%时，CI覆盖率接近0.95。

三、模拟研究

- 百分等级



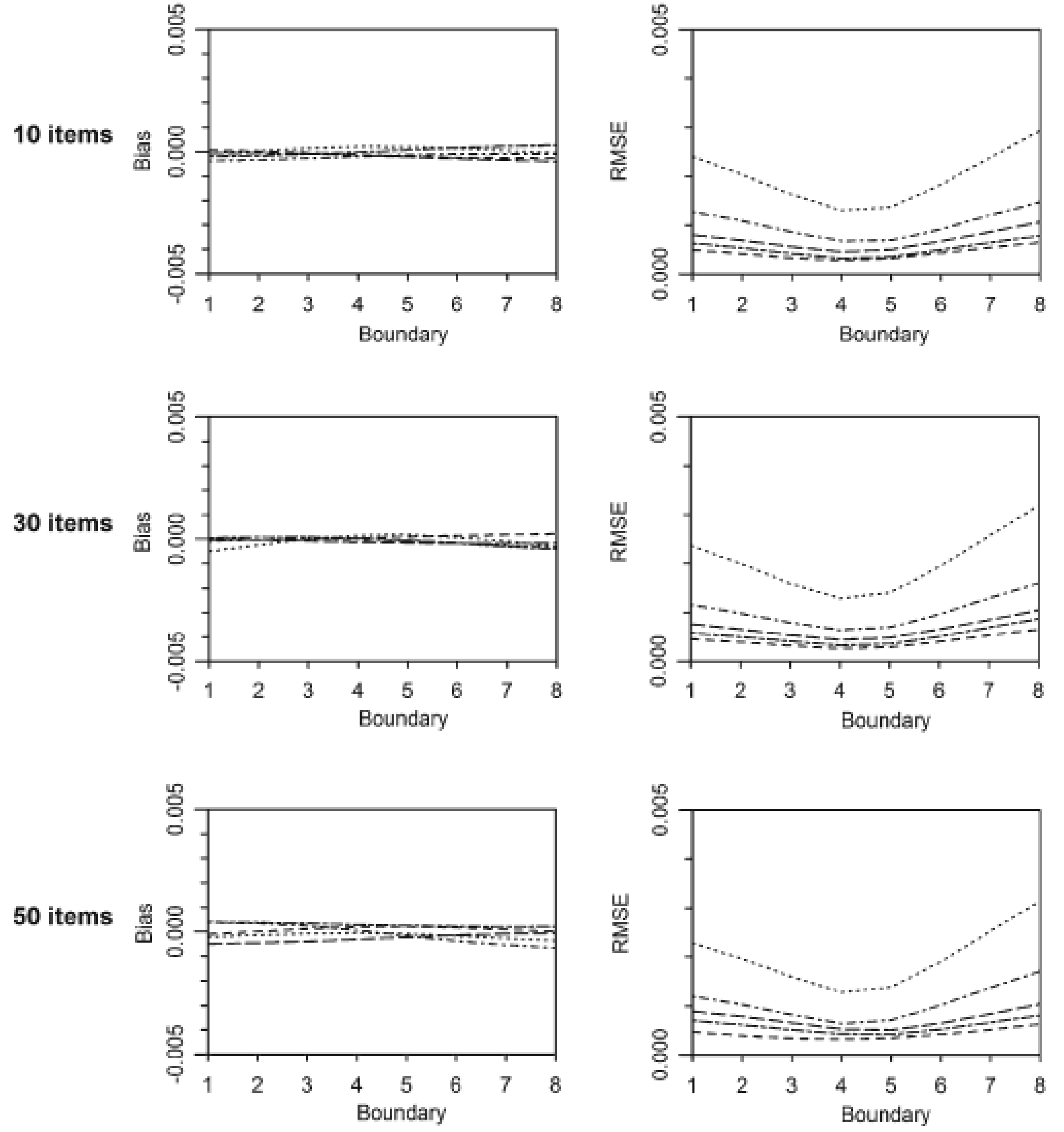
当N > 500，1% ≤ 总体百分等级分数 ≤ 99%时，CI覆盖率接近0.95。

三、模拟研究

- 标准九分

  在任何组合条件下均未表现出偏差；

  并且估计精度随样本量增加而提高。



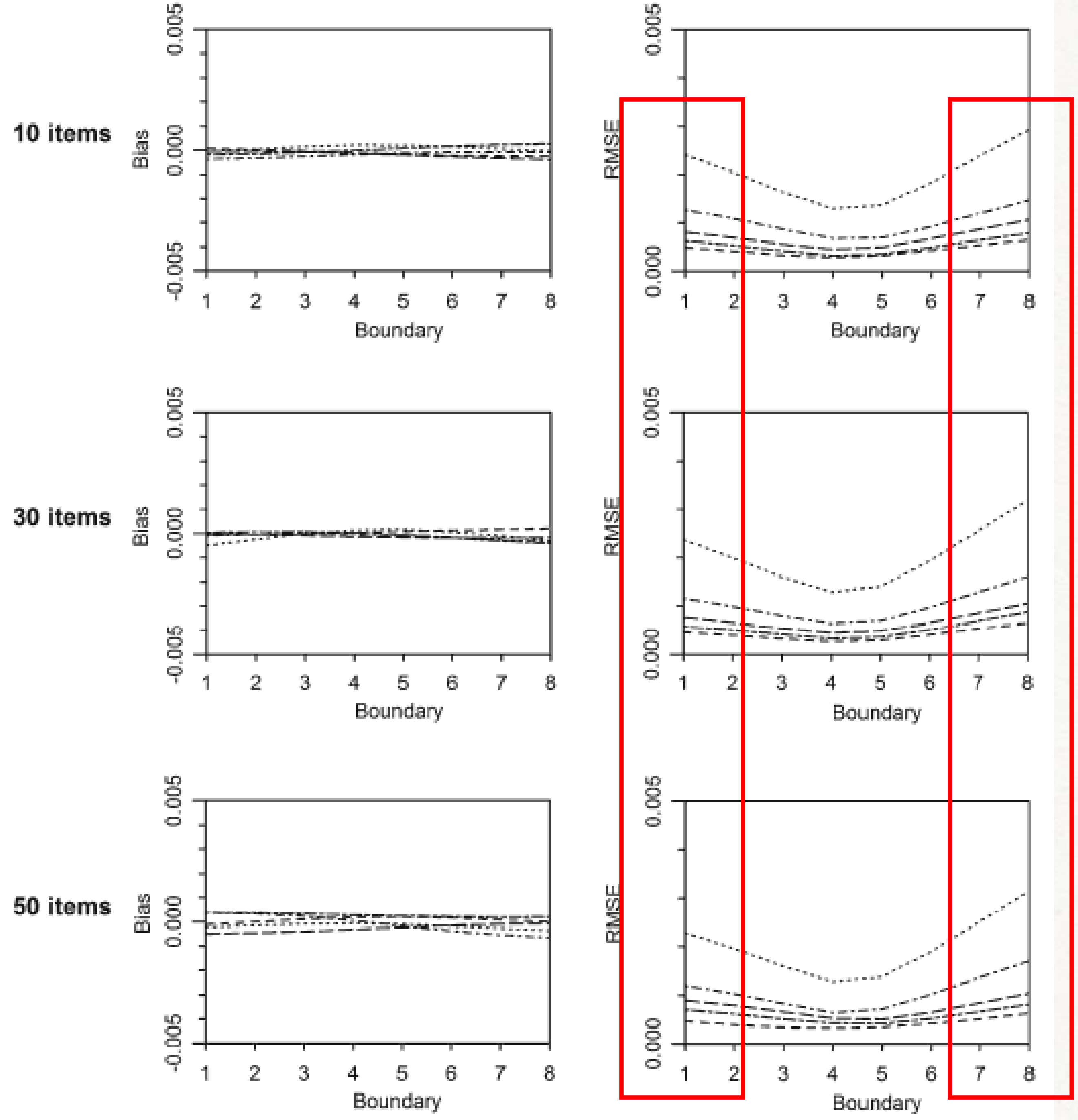
三、模拟研究

- 标准九分

在最低与最高的标准九分边界，精度较低。

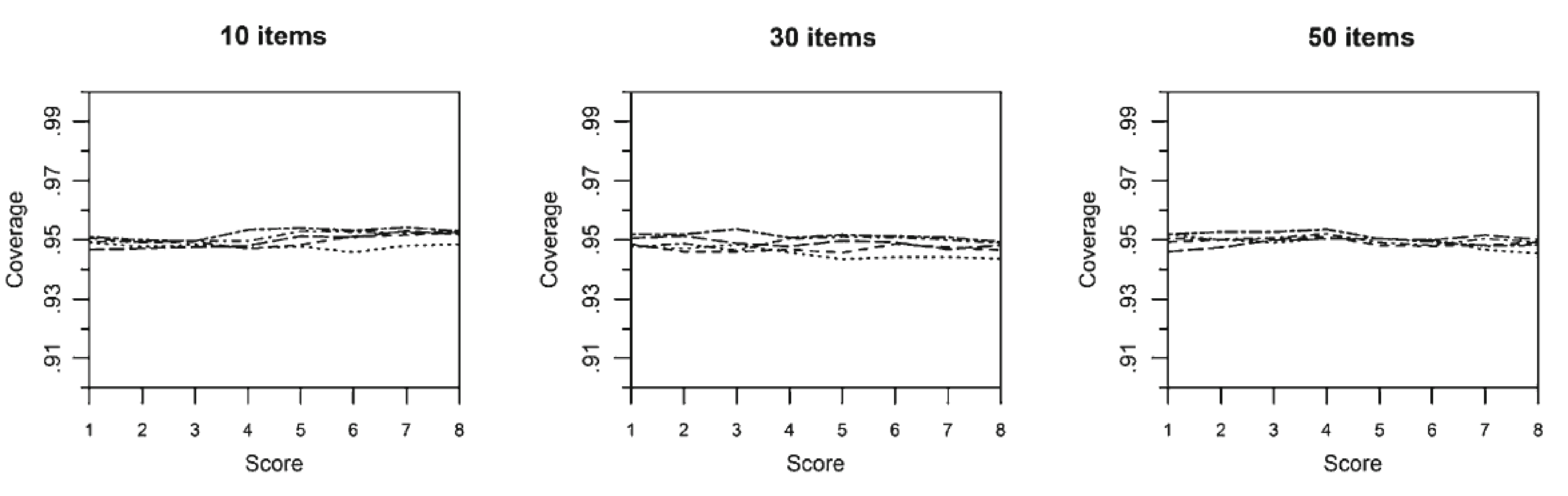
↓

少数观测值远离均值



三、模拟研究

- 标准九分



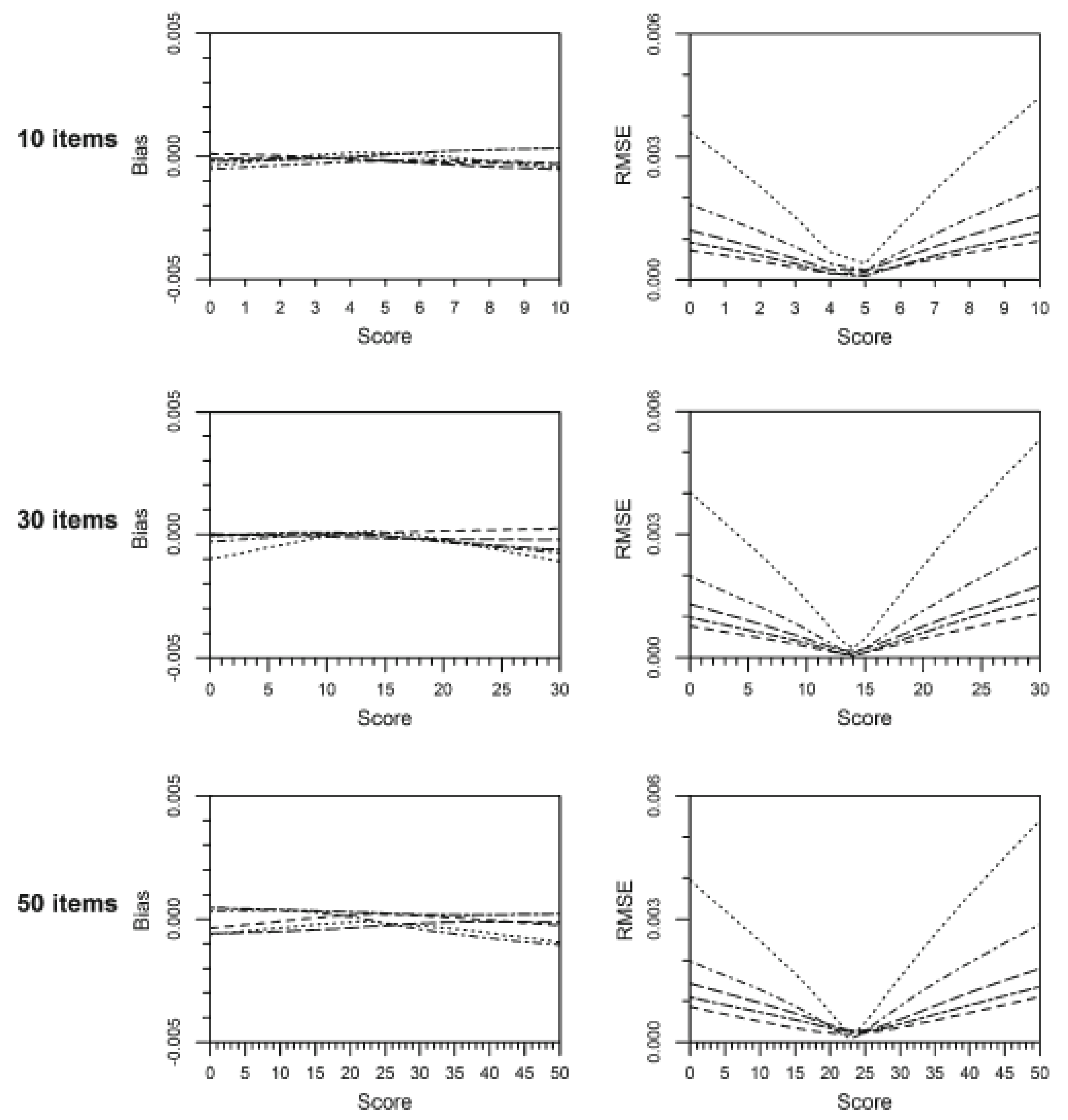
所有组合条件下，CI覆盖率均接近0.95。

三、模拟研究

- Z分数

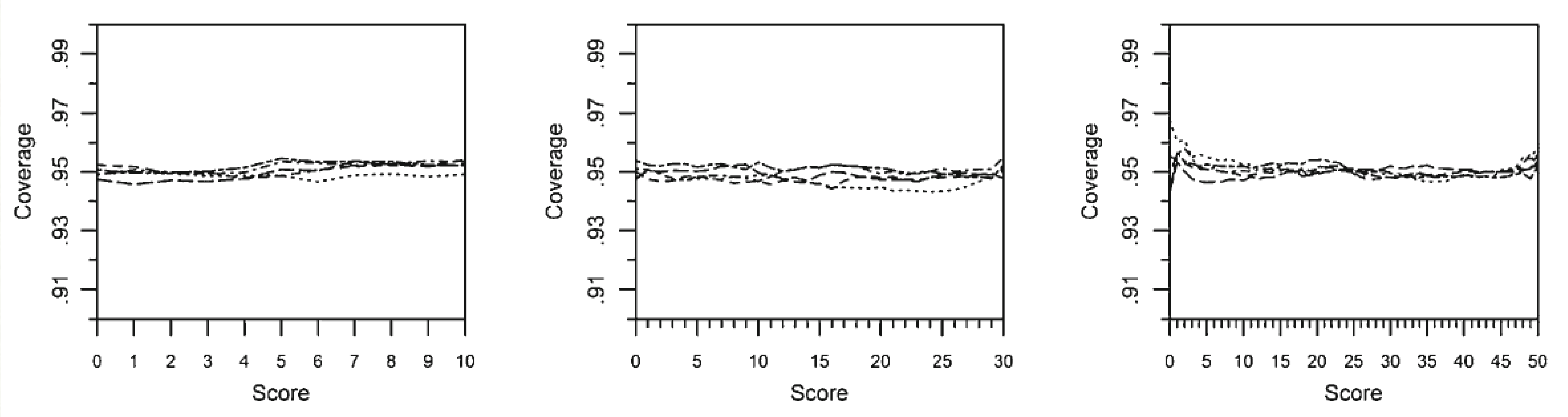
  在所有组合条件下均未表现出偏差；

  并且估计精度随样本量增加而提高。



三、模拟研究

- Z分数



所有组合条件下，CI覆盖率均接近0.95。

四、总结思考

➢ 讨论

泰勒一阶展开 + 广义指对数表示法

$\big\downarrow$ 弱假设条件下

标准差、百分等级、标准九分、Z分数的SE

四、总结思考

① 标准差、标准九分、Z分数的SE在所有条件的组合下均无偏，Wald-based的CI具有良好的覆盖率；

② 百分等级对于小样本(N < 1000)，
在[2.5%，97.5%]范围SE无偏，且CI覆盖率良好；
对于大样本(N > 1000)，
在[1%，99%]范围SE无偏，且CI覆盖率良好。

四、总结思考

- 建议

小样本 & 接近参数空间边界CI覆盖率差

(Agresti&Min,2001)

大多情况表现良好

列联表的分数与剖面似然CI(Score and profile likelihood)可保值域(Lang,2008; also see Agresti, 2012, Sect. 2.3.3)

Wald-based的CI

软件包中不可用、计算复杂

增大样本量

四、总结思考

➢ 建议

随相应原始分数的观测数量增加，百分等级、标准九分、Z分数的SE估计精度提高

↓

提高样本量

↓

极端百分等级 / Z分数，大样本规模标准差，较小规模也可实现精确估计

四、总结思考

- 不足：

  泰勒一阶展开基于线性假设，意味着对于非线性函数(如，标准差、标准九分、Z分数)可能无法得出近似值；

  泰勒一阶展开基于中心极限定理，不适用于小样本。

- 展望：

  其他函数的使用情况研究(α系数、相联测量、拟合优度)。

四、总结思考

- 疑问：

  ① 泰勒一阶展开的线性假设条件，意味着相对于非线性函数，对于百分等级线性函数的近似值估计应该更优，但结果表明其RMSE最大；

  ② 在百分等级RMSE结果中，随着测验长度增加，出现估计精度下降的现象。

- 思考：

  ① 与概化理论方差分量的置信区间估计结合；

  ② 基于弱假设的推导过程中涉及的方法，如大数定律、泰勒展开、克罗内克函数。

欢迎批评指正！