“信效度还得更高！”


“多好一孩子怎么只有B-？”

# 不同最大化信息量位置下的分类准确性与一致性

**黄颖诗**[1] 陈平[1] 张敏强[2]

[1]北京师范大学中国基础教育质量监测协同创新中心
[2]华南师范大学心理学院

# 我们想要些什么？



- 考察学生是否达到某一能力水平标准

CET6 College English Test → 通过 > **425** > 不通过

- 分类准确性 (Classification Accuracy)

| | | 观测 | |
|---|---|---|---|
| | | pass | fail |
| 真实 | pass | $p_{11}$ | $p_{10}$ |
| | fail | $p_{01}$ | $p_{00}$ |

- 分类一致性 (Classification Consistency)

| | | B卷 | |
|---|---|---|---|
| | | pass | fail |
| A卷 | pass | $p_{11}$ | $p_{10}$ |
| | fail | $p_{01}$ | $p_{00}$ |

(Rudner, 2001, 2005; 陈平, 2011; Lee, 2008)

# 存在什么困境?



## 如何获得较高的分类准确性与一致性?

- **分界分数**（cut score）(Birnbaum, 1968; Spray & Reckase, 1994; Lord, 1980; van der Linden, 2005)

VS

- 不同的测验情景**应该选择不同的位置**(Nydick, 2014; Reckase, 1983; Wyse & Babcock, 2016; Jones, Kopp, & Ong, 2019)

→

不同的指标计算方法对各类误差的敏感性存在差异（Lathrop & Cheng, 2013)

一直使用高区分度的题目并不经济（Chang & Ying, 1999)

## 研究目的



- 探讨两种主要的IRT分类准确性与一致性计算方法（即Rudner方法与Lee方法）在特定测验情景下，**如何选择不同的最大化信息量位置**以获得高分类准确性与一致性；
- 并且**考虑题目区分度**参数以及其他可能因素的作用。

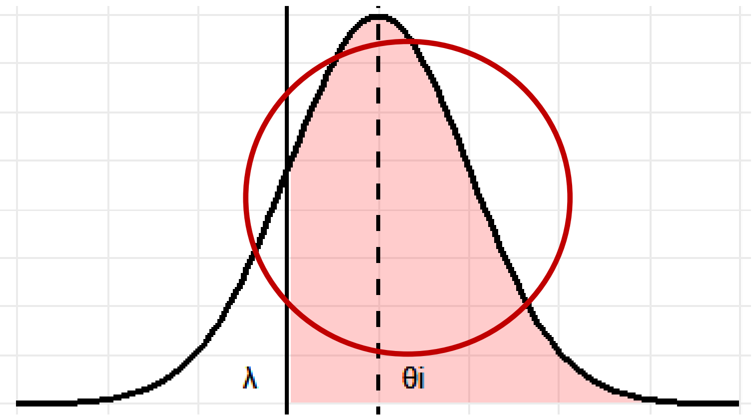
➡ 服务于分类考试题库建设

# Rudner方法



- 假定测量误差服从正态分布，使用极大似然对考生能力进行估计，

令$\lambda$为分界分数，且有$\lambda_0 = -\infty$和$\lambda_\infty = \infty$




$$\gamma_{Rud} = \sum_{i=1}^{N_e} \frac{\phi(\lambda_\infty, \theta_i, \sigma_{\theta_i}) - \phi(\lambda, \theta_i, \sigma_{\theta_i})}{N_e}$$

$$\tau_{Rud} = \sum_{i=1}^{N_e} \frac{[\phi(\lambda, \theta_i, \sigma_{\theta_i}) - \phi(\lambda_0, \theta_i, \sigma_{\theta_i})]^2 + [\phi(\lambda_\infty, \theta_i, \sigma_{\theta_i}) - \phi(\lambda, \theta_i, \sigma_{\theta_i})]^2}{N_e}$$

(Rudner, 2001, 2005)

# Lee方法



- 条件总分分布：能力值为$\theta$ 的考生在测验中获得某个总分$x$的概率

$$P(X = x|\theta)$$

- 分界分数：期望总分

$$\boldsymbol{\varepsilon} = E(x|\theta = \theta^*) = \sum_{j=1}^{J}\sum_{m=0}^{M_j} mP(m|\theta^*)$$

- 能力为$\theta$ 的考生被归到类别$k$ 的概率

$$p_\theta(k) = \sum_{x=\varepsilon_{(k-1)}}^{\varepsilon_k - 1} P(X = x|\theta)$$

$$\gamma_{Lee} = \sum_{i=1}^{N_e} \frac{p_{\hat{\theta}_i}(k)}{N_e}$$

$$\tau_{Lee} = \frac{1}{N_e}\sum_{i=1}^{N_e}\sum_{k=1}^{k}[p_{\hat{\theta}_i}(k)]^2$$

(Lee, 2008, 2010)

# 研究一：各种因素如何影响位置的选择？



1

**操纵变量**

- 最大化题目信息量位置　[-2, 2]的范围里按0.01的增量取能力点$\theta_l$
- 区分度参数$a$　$a \sim N(\mathbf{0.3}, 0.04^2)$；$a \sim N(\mathbf{0.7}, 0.04^2)$
- 能力均值（$\theta_{cut} = 0.58$）　高能力组：$\theta_{high} \sim N(\mathbf{1.49}, 0.46)$
  低能力组：$\theta_{low} \sim N(\mathbf{-0.33}, 0.46)$
- 题目数量　20题、50题、70题和100题

2

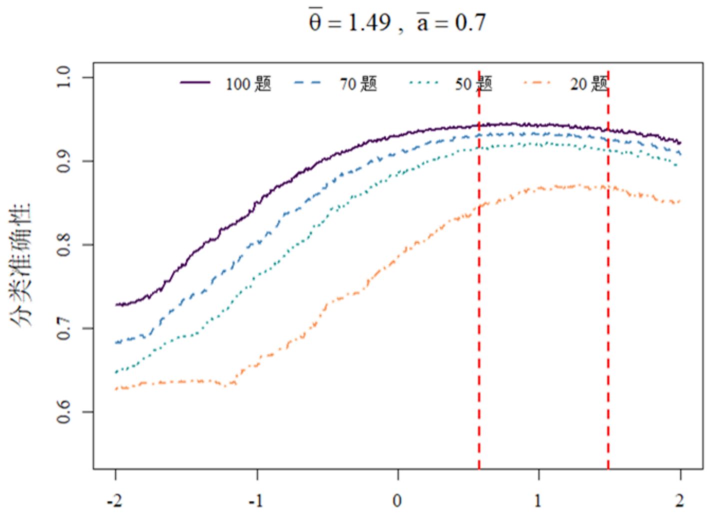
**固定变量**

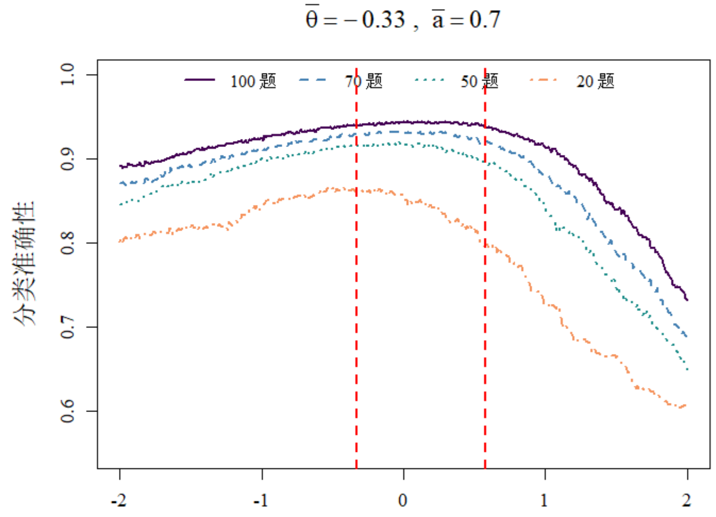
- 难度$b \sim U(-2.5, 2.5)$
- 猜测系数$c = 0.25$
- 量尺常数$D = 1.7$
- 考生5000人
- 题库大小1000题

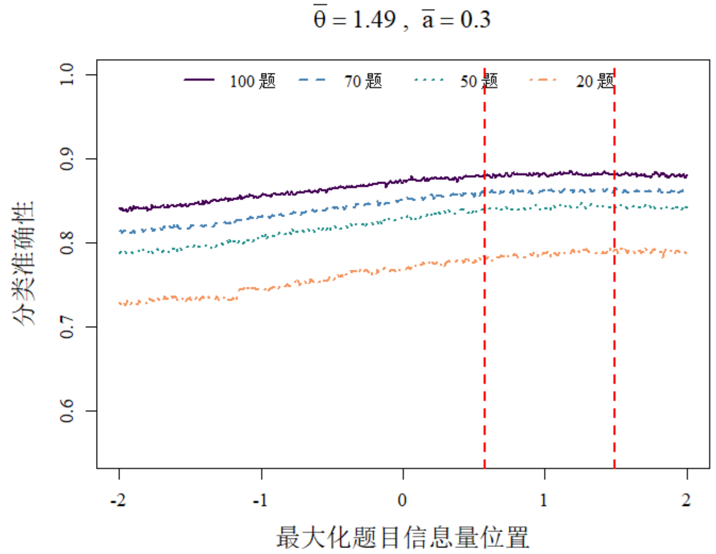
(王孝玲, 2005; Thompson & Ro, 2007; Thompson, 2011; Wyse & Babcock, 2016)

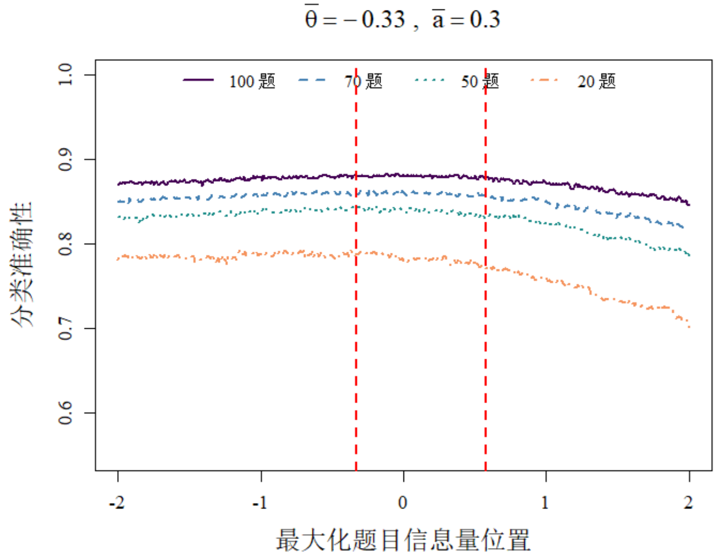
# 结果1：Rudner方法





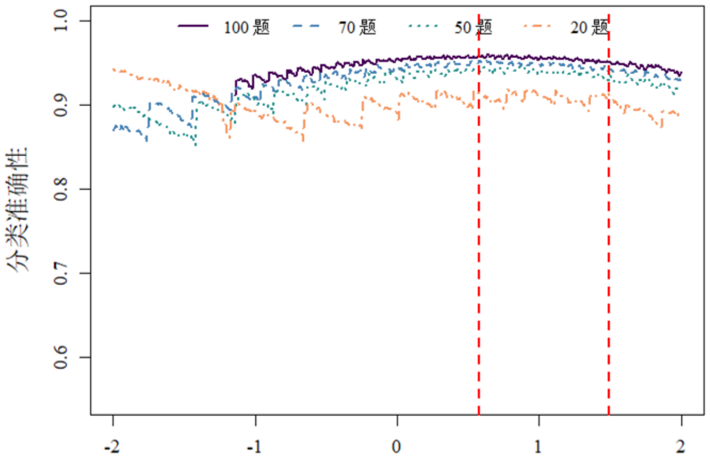



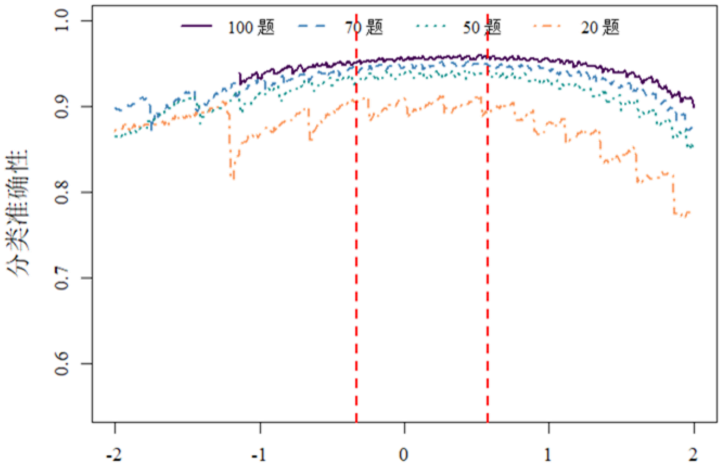
# 结果2：Lee方法



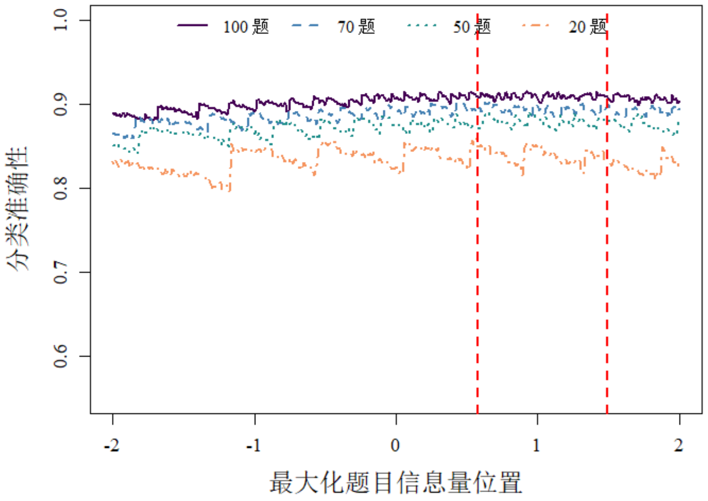
$\bar{\theta} = 1.49$ , $\bar{a} = 0.7$



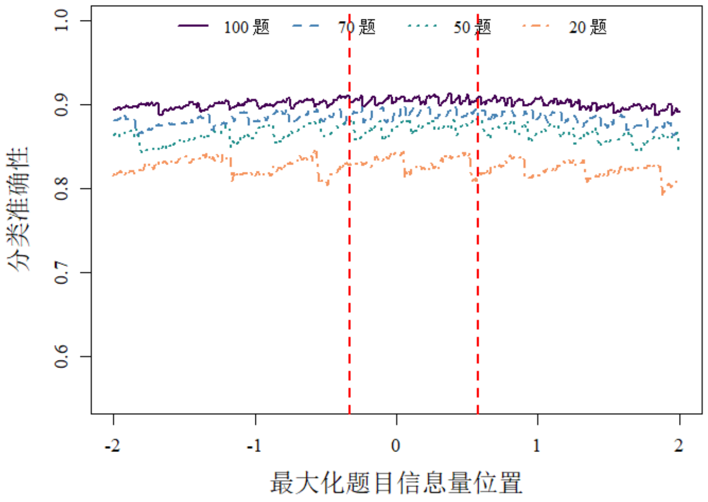

$\bar{\theta} = -0.33$ , $\bar{a} = 0.7$




$\bar{\theta} = 1.49$ , $\bar{a} = 0.3$




$\bar{\theta} = -0.33$ , $\bar{a} = 0.3$



# 研究二：多分类情景下如何选择？



1

**操纵变量**

- 最大化题目信息量位置
  - 当前能力估计值
  - 分界分数 (Weighting Methods)： $max\sum_{c=1}^{C}\frac{1}{|\hat{\theta}-\theta_c|}I(\theta_c)$
  - 随机选题
- **分界分数个数**
  - **2个：33rd、66th**
  - **3个：25th、50th、75th**

2

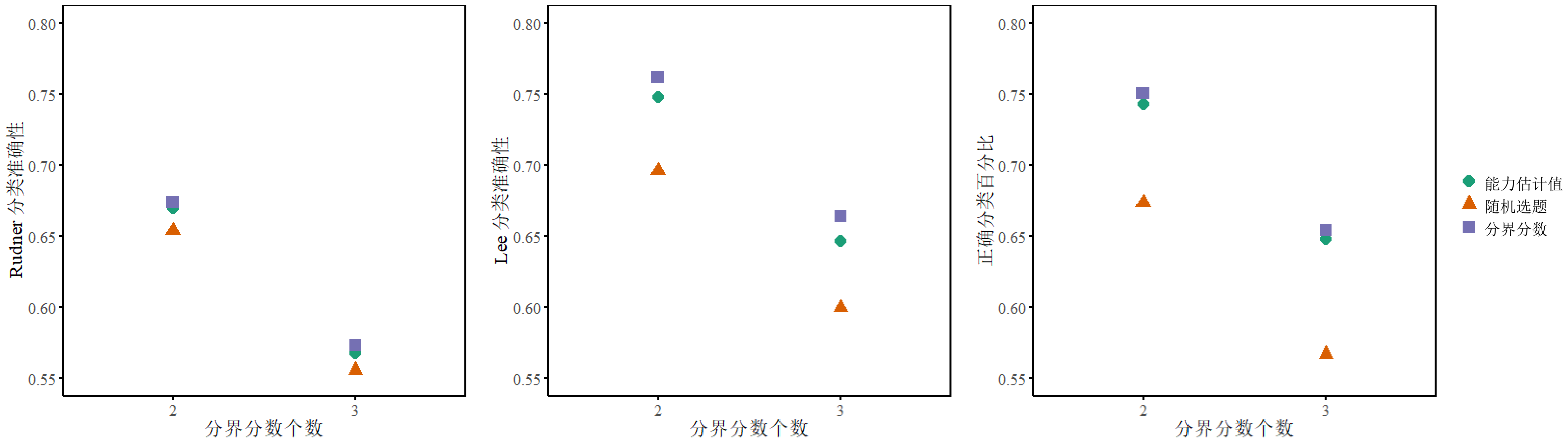
**固定变量**

- 区分度$a$ ~ $N$ (**0.7**, $0.04^2$)
- 题目数量**20题**

(van Groen, Eggen, & Veldkamp, 2013; Veldkamp, 1999)

# 结果3：多分类情景




0.80
0.75
0.70
0.65
0.60
0.55
Rudner 分类准确性
Lee 分类准确性
正确分类百分比
2
3
分界分数个数
能力估计值
随机选题
分界分数

# 讨论与展望



## 1 对于Rudner方法

- 二分类：当使用**高区分度题目或者较短测验**时：靠近考生**能力均值位置**
  而当**题目数量较多**时（比如，70题和100题）：分界分数位置/能力均值位置
- 多分类：考虑分界分数的加权方法

## 2 对于Lee方法

- 二分类：在各种最大题目信息量位置上的表现均相似
- 多分类：考虑分界分数的加权方法

## 3 局限与展望

- 探索不同的**IRT 模型**对结果的影响、开发适用于**多维能力**的计算指标、利用**反应时**数据

# 欢迎各位专家批评指正！

黄颖诗　h_yingshi@163.com