


## 第三章 项目分析

- ①第一节 项目难度
- ①第二节 项目区分度
- ①第三节 选择题反应模式的分析
- ①第四节 项目功能差异与测量偏差
- ①第五节 项目反应理论 (IRT)


Li, J. School of Psychology, Beijing Normal University



## 第一节 项目难度

- ①1.1 难度的定义
- ①1.2 计算方法
- ①1.3 难度水平的确定
- ①1.4 难度的转换

Li, J. School of Psychology, Beijing Normal University




## 1.1 定义

①难度 (difficulty): 即项目的难易程度。

📌 本节内容主要针对能力测验。

Li, J. School of Psychology, Beijing Normal University



## 1.2 难度的计算

1) 二分法记分的项目:

$$P = \frac{R}{N}$$


📌 R: 答对该题的人数  
📌 N: 参加测验的总人数

📌 对选择题的解答, 由于受测者可猜测, 故需对难度系数加以校正:

$$C_p = \frac{KP - 1}{K - 1}$$

📌 C<sub>p</sub>: 校正后的真实难度值  
📌 P: 实际数据得到的通过率  
📌 K: 选项数目

Li, J. School of Psychology, Beijing Normal University




## ①当受测者人数较多时可用极端分组法

$$P = \frac{P_H + P_L}{2}$$

- ①按测验总分由高到低排序
- ①从高分段向下选出全部试卷的27%作为高分组
- ①从低分段向上选出全部试卷的27%作为低分组
- ①分别求出两组受测者在该项目的平均通过率
- ①按照上述公式计算项目难度值

Li, J. School of Psychology, Beijing Normal University



## 2) 非二分法记分的项目:

$$P = \frac{\bar{X}}{X_{\max}} \times 100\%$$

$\bar{X}$ : 全体受测者在该题上的平均分  
 $X_{\max}$ : 该题的满分

例: 一次语文考试的作文题, 总分为40分, 1000名考生的平均分为25分, 请问该作文题的难度是多少?

Li, J. School of Psychology, Beijing Normal University

### 1.3 难度水平的确定

#### 1) 从测量的目的考虑

- 测量都希望能准确测量个体间的差异。如果某题受测者全对或全错，则不能把不同人区别开来。
- 应使试题的平均难度接近0.50，而各题难度在  $0.50 \pm 0.20$  之间。

#### 2) 从测量的作用考虑

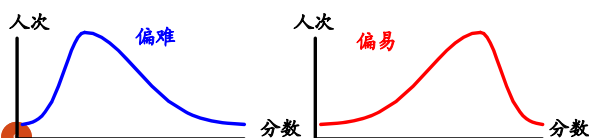
- 对于选拔用的人事测量，应尽量使难度值接近录取率。

#### 3) 从题目的形式考虑

- 选择题的实际P值应大于概率水平

#### 4) 从测验分数的分布考虑整体难度

- 结果为正偏态分布时，大多数得分集中在低分，说明测验对于所研究的样本团体来说偏难。
- 结果为负偏态分布时，说明测验过易。
- 标准参照测验出现偏态分布是允许的。

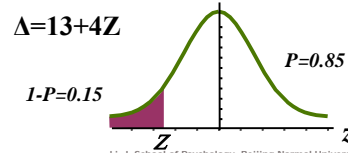


### 1.4 难度的转换

- 由于P是百分比，为了便于比较，我们可以将其转换为等距量表。
- 方法：根据正态曲线表，将试题难度（概率）转换成对应的Z分数（Z值对应的概率是1-P）。
- 常用美国教育测验服务机构采用的难度指标：

$$\Delta = 13 + 4Z$$

$$1-P=0.15$$



例：

$$P=.9987 \quad Z=-3 \quad \Delta = 13 + 4(-3) = 1$$

$$P=.8413 \quad Z=-2 \quad \Delta = 13 + 4(-2) = 5$$

$$P=.5000 \quad Z=0 \quad \Delta = 13 + 4(0) = 13$$

$$P=.1587 \quad Z=2 \quad \Delta = 13 + 4(2) = 21$$

$$P=.0013 \quad Z=3 \quad \Delta = 13 + 4(3) = 25$$

$\Delta$ 值介于1与25之间，平均难度值为13；越大，表试题越困难；反之则越容易。

## 第二节 项目区分度

### 02.1 区分度的定义

### 02.2 计算方法

### 02.3 区分度与难度的关系

## 2.1 什么是区分度

①项目区分度 (item discrimination), 也叫鉴别力, 是指测验项目对受测者的心理特性的区分能力。

✎区分度高的项目, 能将不同水平的受测者区分开来; 区分度低的项目, 则不能很好地鉴别受测者水平, 水平高和水平低的受测者差不多。

✎评价项目质量、筛选项目的主要指标

✎通常用D表示, 取值范围为 +1.00至 -1.00

13

Li, J. School of Psychology, Beijing Normal University

## 2.2 计算方法

### 1) 极端分组法 (鉴别指数法)

✎比较测验总分高和总分低的两组受测者在项目通过率上的差别:  $D = P_H - P_L$

①按测验总分从高到低排序

①确定测验总分最高的27%的受测者作为高分组, 最低的27%的受测者为低分组

①分别求出这两组受测者的通过率, 代入公式计算

14

Li, J. School of Psychology, Beijing Normal University

### ①鉴别指数的评价标准

伊贝尔 (L.Ebel, 1965)

D	评价
.40以上	非常优良
.30-.39	良好, 如能改进更好
.20-.29	尚可, 需要修改
.19以下	差, 淘汰或改进

15

Li, J. School of Psychology, Beijing Normal University

### 2) 相关法

✎以项目得分与测验总分的相关作为区分度的指标

①前提: 测验总分可以看作是受测者能力/人格水平的评价指标

✎使用所有数据, 相关越高区分能力越好, 原则:

① keep the best  $m$  items with  $m$  largest  $r_{IT}$

① keep an item if  $r_{IT} > \text{criterion}$  (e.g. 0.4)

✎具体方法: 积差相关法、点二列相关、二列相关、 $\Phi$ 相关

16

Li, J. School of Psychology, Beijing Normal University

### ①区分度具有相对性

✎采用不同的计算方法, 区分度的值不同。因此在分析一个测验时, 各个项目的区分度值要采用同一种指标, 否则无法比较。

✎用相关法计算的区分度值受样本大小影响。

✎用两个极端组通过率的差异作为区分度的指标, 其值受分组标准的影响。

✎区分度值的大小与样本的同质性有关。

17

Li, J. School of Psychology, Beijing Normal University

## 2.3 难度与区分度的关系

难度 (P)	区分度 (D的 <del>最大值</del> )
1.00	0.00
0.90	0.20
0.70	0.60
0.50	1.00
0.30	0.60
0.10	0.20
0.00	0.00

18

Li, J. School of Psychology, Beijing Normal University



- ④ 较难的项目对高水平受测者区分度高，较易的项目对水平低的受测者区分度高，中等难度的项目对中等水平的受测者区分度高。
- ④ 由于人的多数心理特性呈正态分布，所以项目难度的分布也以正态为好，即特别难的与特别容易的题目较少，越接近中等难度的题目越多，而所有项目的平均难度为0.50。

19

Li, J. School of Psychology, Beijing Normal University

### 第三节 选择题反应模式的分析



#### ④ 关于猜测行为

##### ④ 两种观点:

- ④ 猜测行为带来测量误差，应加以控制。
- ④ 猜测并非盲目进行，也会反映能力，无需控制。

##### ④ 如何控制猜测行为

- ④ 增加选项的个数（一般不少于四个）。
- ④ 如有特殊需要，可采用警告、扣分等方式控制。

20

Li, J. School of Psychology, Beijing Normal University

### 选项分析的过程



- 1) 根据受测者的测验总分，对受测者进行排序。
- 2) 确定高分组和低分组受测者。
- 3) 分别登记高分组受测者和低分组受测者在每个备选项的人数及未作答人数（也可以是人数百分比），最终将数据资料整理成表格形式。
- 4) 根据第3步整理好的数据资料，进行具体分析。

题号	组别	a	b	c	d	未作答	正解	难度	区分度
1	高分组	36	0	39	23	2	d	.04	.04
	低分组	32	0	46	18	4			
2	...						...	...	...

23

Li, J. School of Psychology, Beijing Normal University

### 选项分析的原则



#### ④ 总体原则

- ④ 对于一套试卷，大部分选择题的正解的选择率应该在0.4到0.6之间，各诱答均分剩余选择率。
- ④ 对于正解，考察高分组和低分组在正确答案上的选择率是否是正差，及这一差距是否足够大。
- ④ 对于诱答，考察高分组和低分组在干扰项上的选择率是否是负差，及这一差距是否足够大。

22

Li, J. School of Psychology, Beijing Normal University

#### ④ 其它情况:

- ④ 如果正确的备选答案被所有受测者选择，说明该题目太容易或题目中有某种暗示。
- ④ 如果一个题目未作答人数过多或选择各个备选答案人数相等，则说明题目过难或题意不清，使得受测者无法作答或凭猜测作答。

23

Li, J. School of Psychology, Beijing Normal University

- ④ 如果高分组受测者的选择集中在两个答案上，而且选择率接近，说明该题可能有两个正确答案。
- ④ 如果高分组对正确答案的选择与低分组相等或低于后者，说明该题所考察的东西与整个测验的测量目标无关。

24

Li, J. School of Psychology, Beijing Normal University



- 如果某个诱答没有任何受测者选择，说明该诱答不具迷惑性，错得过于明显，一般来讲，除非有2%以上的人选择，否则这个备选答案就应该删除。
- 如果所有受测者都选择了同一个诱答，可能是答案设置错误，也可能是教学中的错误。

25

Li, J. School of Psychology, Beijing Normal University



#### 例：四个题目的项目分析

题号	组别	a*	b	c	d	未答	P	A	r <sub>b</sub>	D
1	高	36	0	40	24	0	.34	14.7	.04	.04
	低	32	0	46	20	2				
2	高	22	12	10	48	8	.27	15.6	-.12	-.10
	低	32	25	11	23	9				
3	高	62	16	7	15	0	.44	13.6	.37	.36
	低	26	36	7	28	3				
4	高	95	2	1	2	0	.78	10.0	.55	.41
	低	54	18	12	16	0				

注：a为正确选项

26

Li, J. School of Psychology, Beijing Normal University



#### 第四节 项目功能差异与测量偏差

- 项目功能差异 (differential item function, DIF) 是指不同群体对同一项目的答对概率 (或得分率) 不同，即不同群体在同一项目上得分存在差异。
- 原因可能是群体间本身水平存在差异 (良性DIF)，也可能其它原因 (如文化、性别、地域、种族、职业、年级、SES) 所致 (不良DIF)。
- 良性DIF不存在测量偏差，不良DIF存在测量偏差 (test bias)。

27

Li, J. School of Psychology, Beijing Normal University



#### ① 一致性DIF & 不一致性DIF

- 一致性DIF是指受测者的能力水平与其组别之间不存在交互作用，即在所有的能力水平上，一组受测者回答某一项目的正确率都大于另一组。
- 当受测者的能力水平与其组别之间存在交互作用时，则表现为非一致性DIF。

28

Li, J. School of Psychology, Beijing Normal University



#### DIF及测验偏差侦查的一般步骤

##### (1) 受测者群体的确定及数据资料的准备

- 根据测验特征，确定哪两类受测者群体可能会存在DIF。然后再随机抽取这两类受测者群体的测验反应数据。所抽取的受测者总量应足够多 (一般不少于1000人)，但也不能太多 (一般不多于2000)。

Why?

29

Li, J. School of Psychology, Beijing Normal University



##### (2) 目标组和参照组的确定

- 根据项目特征，确定项目对以上确定的哪类群体有利，若第(1)步确定为男生和女生两个群体，且认为项目可能对男生有利，那么一般把男生样本作为参照组，女生样本作为目标组。

30

Li, J. School of Psychology, Beijing Normal University

### (3) 匹配变量的确定

一般以测验总分作为匹配变量，将两个群体中相同测验总分的受测者一一匹配，采用适当的**DIF侦查方法**进行分析，找出并剔除存在DIF的项目，从而组成一个不含DIF的子测验。然后再以该子测验的总分作为匹配变量，再对子测验进行DIF分析，找出并剔除存在DIF的项目，如此反复，直到找到一个不含DIF项目的子测验，并以受测者在该子测验上的得分作为最终的匹配变量。

31

Li, J. School of Psychology, Beijing Normal University

### (4) DIF的探索性分析

根据第(3)步确定的最终的匹配变量对测验的所有项目进行正式的DIF分析。

### (5) DIF成因分析

成立专家小组对项目产生DIF的原因进行分析，并确定哪些项目是真正存在偏差(bias)，即存在不良DIF。这需要学科专家、所测特质研究专家、测量学专家多方面人员共同参与完成。

32

Li, J. School of Psychology, Beijing Normal University

## DIF侦查的常用方法

分为实际得分方法和潜在特质方法。

实际得分方法包括散点图法、卡方方法、MH法、标准化方法(STND)、SIBTEST方法、逻辑斯蒂回归方法等。

潜在特质方法包括验证性因素分析和项目反应理论两个领域的检验方法。

33

Li, J. School of Psychology, Beijing Normal University

## 例：MH方法 (Mantel - Haenszel Procedure)

用于侦查0、1记分项目的DIF，以测验总分作为匹配变量。MH方法统计量的计算建立在一张 $S \times 2 \times 2$ 的列联表中，其中S是测验总分的水平数（由研究者根据需要自行确定），对于其中任一水平K，可得出两子群体得/失分的 $2 \times 2$ 列联表。

群体	题目得分		总计
	1	0	
参照组	$f_{1rk}$	$f_{0rk}$	$n_{rk}$
目标组	$f_{1k}$	$f_{0k}$	$n_k$
总计	$n_{1k}$	$n_{0k}$	$n_k$

34

Li, J. School of Psychology, Beijing Normal University

根据S个列联表，计算如下统计量：

$$\alpha_{MH} = \frac{\sum_{k=1}^S \frac{f_{1rk} \cdot f_{0rk}}{n_k}}{\sum_{k=1}^S \frac{f_{0rk} \cdot f_{1rk}}{n_k}}$$

群体	题目得分		总计
	1	0	
参照组	$f_{1rk}$	$f_{0rk}$	$n_{rk}$
目标组	$f_{1k}$	$f_{0k}$	$n_k$
总计	$n_{1k}$	$n_{0k}$	$n_k$

$\alpha_{MH}$ 的取值介于0至正无穷之间， $\alpha_{MH}=1.0$ 时，表示该研究项目无DIF； $\alpha_{MH}<1.0$ 时，表示研究项目对目标组有较低难度； $\alpha_{MH}>1.0$ 时，表示所研究项目对参照组有较低难度。

35

Li, J. School of Psychology, Beijing Normal University

但是由于 $\alpha_{MH}$ 的计算来自于样本数据，因此对其值是否等于1.0必须进行统计检验。

检验统计量是 $MH\chi^2$ ，其计算公式为：

$$MH\chi^2 = [ |\sum_{k=1}^S f_{1rk} - \sum_{k=1}^S E(f_{1rk})| - 0.5 ]^2 / \sum_{k=1}^S Var(f_{1rk})$$

其中：

$$E(f_{1rk}) = n_{1k} \cdot n_{rk} / n_k$$

$$Var(f_{1rk}) = n_{1k} \cdot n_{0k} \cdot n_{rk} \cdot n_{rk} / [ n_k^2 (n_k - 1) ]$$

36

Li, J. School of Psychology, Beijing Normal University

### 例：性别DIF（SIBTEST法）

① Schutte于1998年编制的Emotional Intelligence Scale (EIS) 自陈问卷共有33个项目，采用5点记分，包括感知情绪、调控自我情绪、调控他人情绪、运用情绪等4个维度。使用华南师范大学王才康修订中文版，比较**男性**和**女性**的DIF。

03. 我期望能够做好自己想做的绝大多数事情

05. 我难以理解别人的肢体语言(\*)

14. 我会去寻找一些让自己开心的活动

30. 当别人消沉时，我能够进行帮助，使他感觉好些

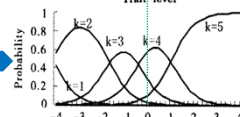
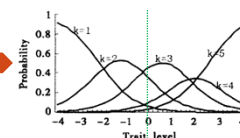
37

Li, J. School of Psychology, Beijing Normal University

### 例：文化DIF（IRT方法）

① SHIBA 简易人格量表中“环境敏感性”维度，回答形式为“1-完全不介意”~“5-非常介意”，共5个等级，比较**中日**受测者。

1.	对周围声音、臭味等
2.	身边有不洁之物时
3.	对于食品、食具的卫生
4.	把手、拉手被弄脏时
5.	生活中身边的安全性
6.	衣服等个人卫生
7.	自己的手被弄脏时
8.	住宿处的被褥卫生



38

Li, J. School of Psychology, Beijing Normal University

### 例：性别DIF（IRT方法）

① 《新生适应性量表》中的“情绪适应性”分量表，共5个项目。回答形式为“1-非常符合”~“6-非常不符合”6个等级，得分越高表示被试的情绪适应水平也就越高，比较**男生**和**女生**DIF。

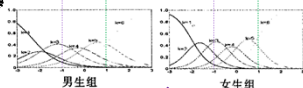
1. 进入大学后我有失落感

2. 我的心情很浮躁

3. 在大学里我有些自卑

4. 进入大学后我有种莫名的恐惧感

5. 进入大学后我的心情愉快充实(\*)



39

Li, J. School of Psychology, Beijing Normal University

### 不良DIF的控制

① 保证题目所测心理品质与全卷所测心理品质完全一致，避免题目测量了测验之外的第二种构想。

① 题目语言规范、无歧义，避免使用方言、俚语等。

① 编制好的题目应广泛征求不同群体的意见。

① 在正式使用前，最好先进行小规模试测，并修改或删除存有不良DIF的试题。

40

Li, J. School of Psychology, Beijing Normal University

## 第五节 项目反应理论

### ① 经典测验理论

✎ 试题难度和区分度因样本不同而不同，因此同一份试卷很难获得一致的难度、区分度。

✎ 对高、低能力两极端组的受测者而言，他们的能力估计不合理且不准确。

✎ 忽视受测者的项目反应模式 (item response pattern)：即使原始得分相同的受测者，其反应组型亦不见得一致，其能力估计值应该会有所不同。

41

Li, J. School of Psychology, Beijing Normal University

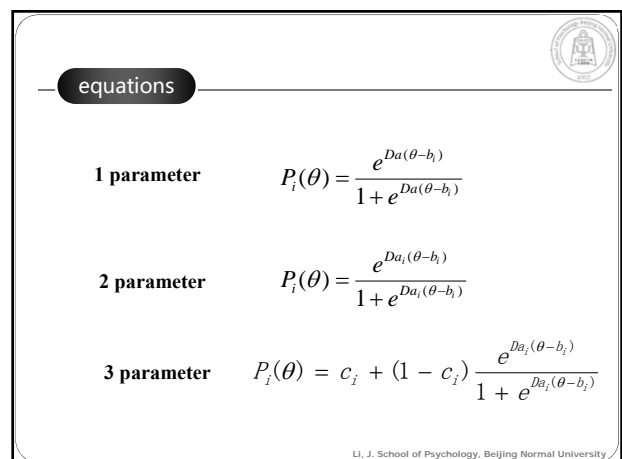
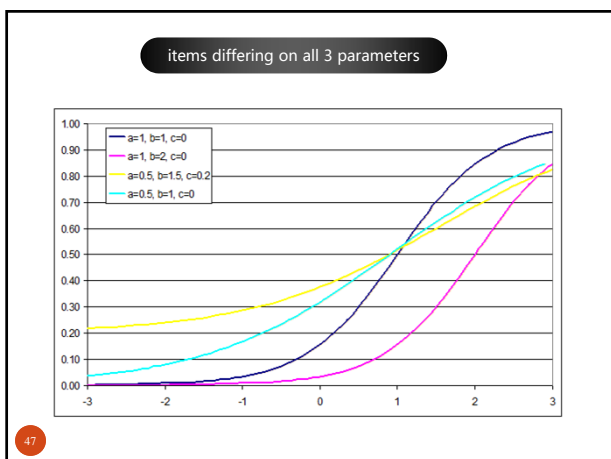
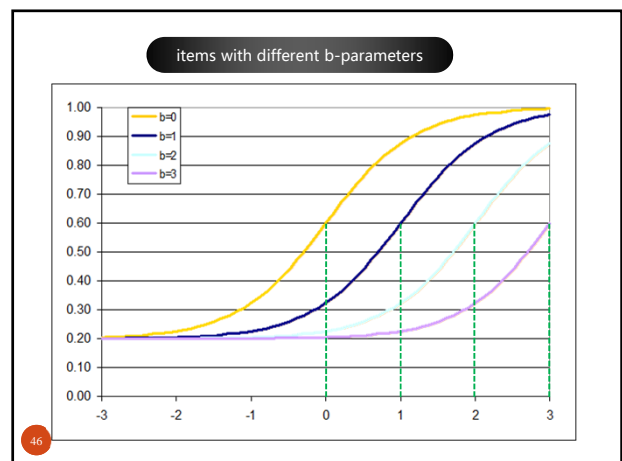
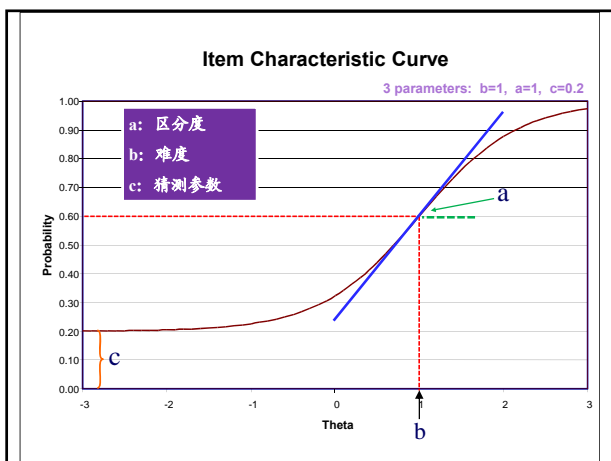
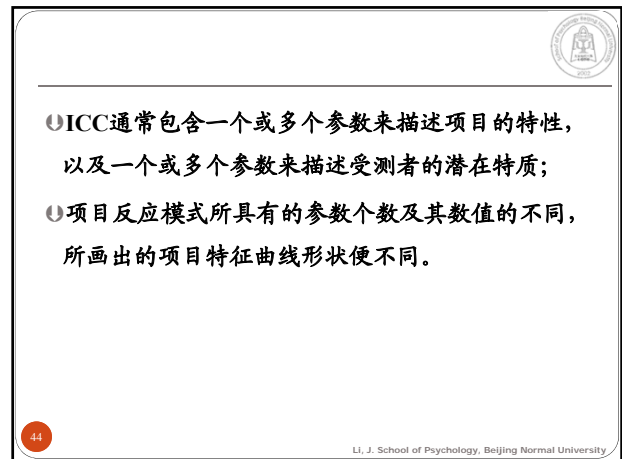
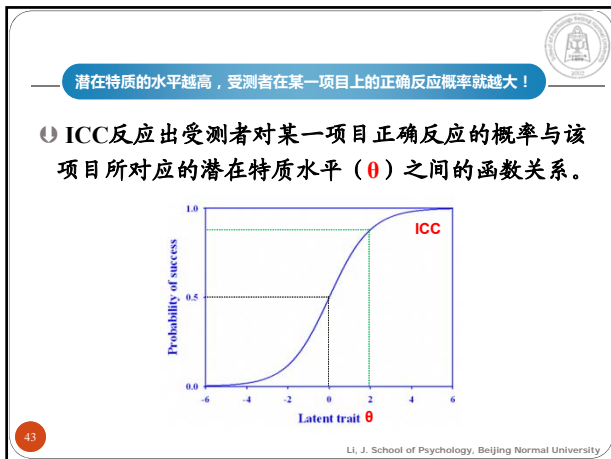
### ① IRT的基本思想

✎ 假设受测者对测验的反应受某种心理特质（潜在特质/latent traits）的支配。根据受测者的特质水平可以预测其对项目和测验的反应。

✎ 受测者的表现情形与潜在特质间的关系，可透过一条连续递增的函数来描述，这个函数被称为项目特征曲线 (item characteristic curve / ICC)。

42

Li, J. School of Psychology, Beijing Normal University





## 3 parameter equation

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

$\theta$  = trait measured

$a$  = discrimination

$b$  = difficulty

$c$  = pseudo-chance

$i$  = index for item

$D$  = scaling constant (1.7)

$P$  = probability of answering correctly

a: none	0
very low	.01 - .34
low	.35 - .64
moderate	.65 - 1.34
high	1.35 - 1.69
very high	> 1.70
perfect	infinity
b: range	- 3 ~ +3
b对应的P	( 1+c ) / 2
c: range	0 <= c <= 1.0
unacceptable	> 0.35

Li, J. School of Psychology, Beijing Normal University

## Test assumptions

Unidimensional latent trait

But, there are some multidimensional models

Speed would be an additional dimension

Local independence

Responses to items are statistically independent

## Data requirements

1 parameter - n=200

2 parameter - n=600

3 parameter - n=1000

Li, J. School of Psychology, Beijing Normal University

## IRT 的优点

- 能力参数估计的不变性：受测者能力参数的估计与所使用测验包含哪些测验项目无关；
- 项目参数估计的不变性：项目参数的估计与所使用的样本无关；
- 参数设计科学：a与b独立，c通过数据计算得出；
- 受测者能力和项目难度在同一个量表上，为测验的编制、测验分数的报告和解释提供便利；
- 提供受测者能力估计值的误差指标——项目信息函数（Item Information Functions）和测验信息函数（Test Information Function）：测验前就可以确定各项目及整个测验对于受测者能力估计的精确度。

38

Li, J. School of Psychology, Beijing Normal University