解析PISA: 理念、技术与结果

黄晓婷

2016，12

国际上几个主要的大规模考试

- PISA (Programme for International Student Assessment)
  - OECD
  - http://www.oecd.org/pisa/
- TIMSS (Trends in International Mathematics and Science Study)
  - IEA (the International Association for the Evaluation of Educational Achievement, 国际教育成就评价协会)
  - http://www.iea.nl/timss
- PIRLS (Progress in International Reading Literacy Study)
  - IEA
  - http://www.iea.nl/pirls

PISA与其他国际考试的不同

- 最广泛的参与
  - 2015年，72个国家和地区
- 测试对象不同
  - 不针对具体的年级
  - 针对年龄：15岁3个月到16岁2个月的学生
- 提供不同的教育质量评价标准
  - 测试内容不针对课程，而是以“素养（literacy）”为核心，
  - 科学素养、阅读素养、数学素养
- 通过收集背景数据，提供教育公平的国际视角

主要内容

- 以“素养”为核心的质量观
- 测评学生“素养”用了哪些专业技术
- 对PISA结果的使用

“素养”的定义和特点

- PISA测试的三种核心素养
  - **科学素养**：作为公民参与科技相关的事务的能力，解决社会生活中与科学技术相关问题，并有科学的意识，会科学反思
  - **阅读素养**：为实现个人目标，发展个人的知识和潜能，有效地参与社会生活，而理解、运用和反思书面材料的能力
  - **数学素养**：人们在各种情境下提出数学问题、运用数学知识和解释数学结果的能力，这些能力能够帮助个体理解数学在社会生活中的作用，并且做出好的决策和判断，成为一个具有建设性、 参与性、反思能力的公民
- 特点
  - 教育与职业生涯的发展，终身学习所需要的能力
  - 强调学习、思考与研究的方法

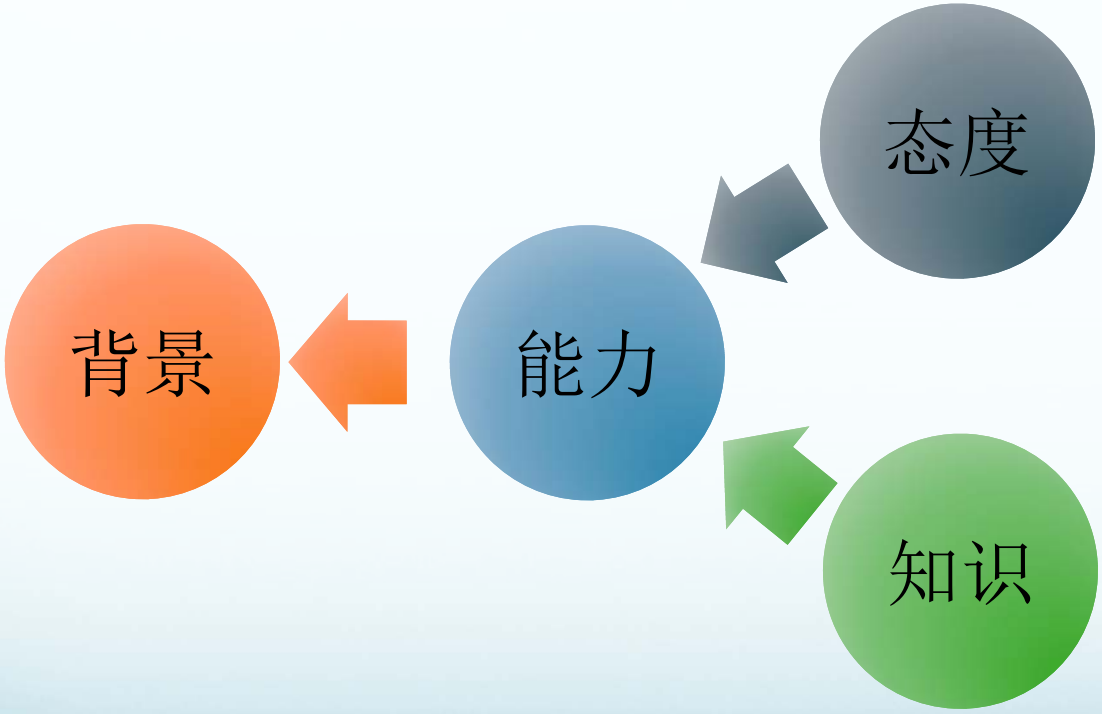
PISA如何测量学生“素养”

- 测试框架
- 评价工具的开发
  - 命题和组卷
  - 翻译修改
  - 试测与最终定稿
- 评价结果的分析
  - 测量学分析 （IRT scaling）
  - 国际可比性检验 (DIF分析)
  - 学生素养的发展变化 （等值）
  - 能力等级的划分

测试框架：以科学素养为例

科学素养

- 在社会生活背景下的科学能力、知识和态度



科学素养：背景

	个人	地方/国家	全球
健康与疾病	营养，健康食品，突发事件	疾控，社区健康	流行病，传染病研究
自然资源	个人能源消耗	食品、能源的生产和分配	新能源，人口增长与可持续发展
环境质量			
自然灾害			
科技发展前沿	个人生活、音乐、体育活动的中的科技	地方公共交通、医疗卫生中的科技	太空探索，物种灭绝

科学素养：能力（competency）

- 科学地解释现象 (Explain phenomena scientifically)
- 设计和评估科学研究的方法 (Evaluate and design scientific enquiry)
  - Identify the question explored in a given scientific study
  - Distinguish questions that could be investigated scientifically
  - Propose a way of exploring a given question scientifically
  - Evaluate ways of exploring a given question scientifically
  - Describe and evaluate how scientists ensure the reliability of data, and the objectivity and generalizability of explanations
- 解释数据和论据 (Interpret data and evidence scientifically)

科学素养：知识（knowledge）

- 内容（content）：科学事实和理论
- 过程（procedural）：科学研究的方法，特别是数据采集、分析、解读
- 认知（epistemic）：科学的认知方法，假设、观察、模型、理论、事实

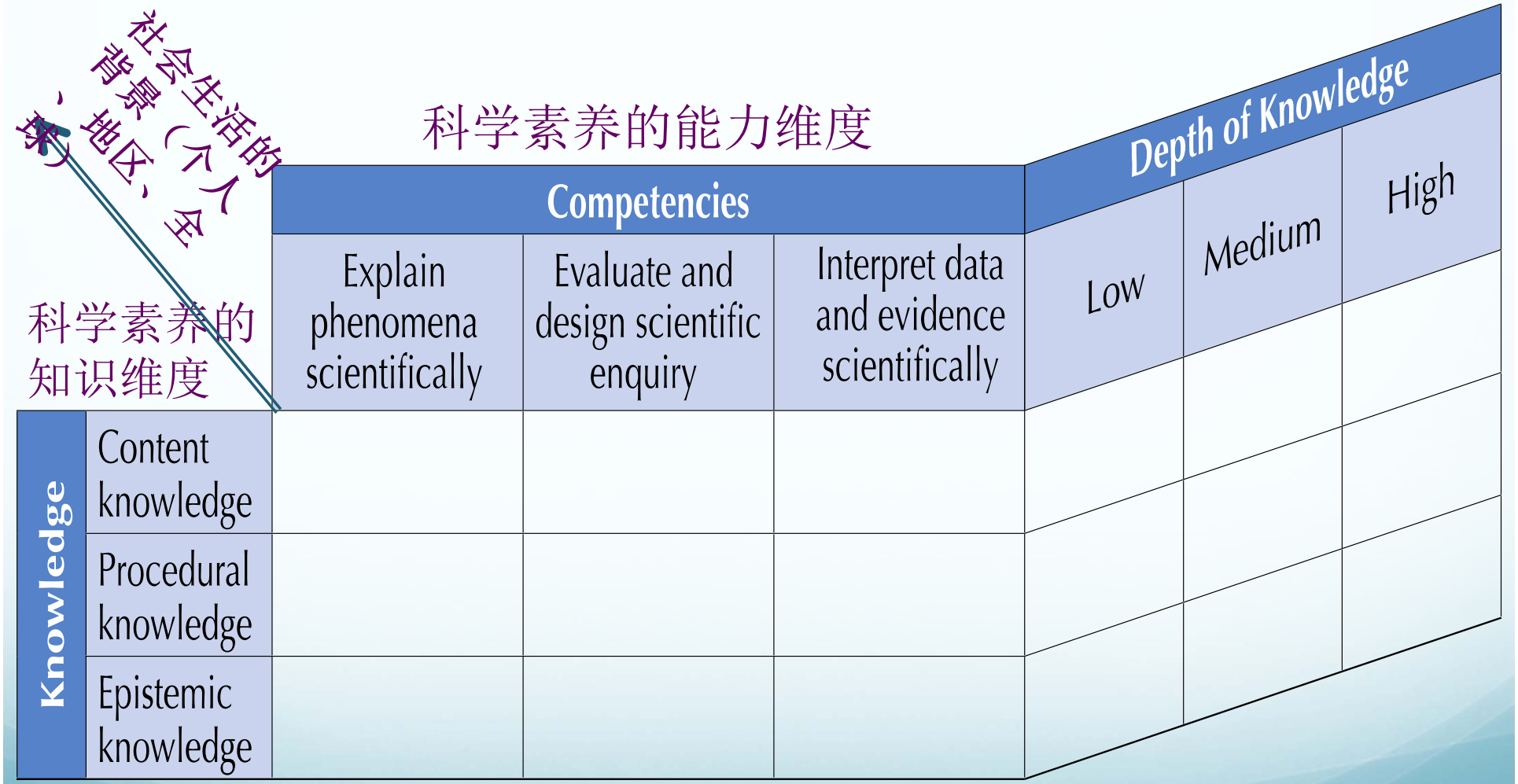
	Systems			
Knowledge types	**Physical**	**Living**	**Earth and space**	**Total over systems**
Content	20-24	20-24	14-18	54-66
Procedural	7-11	7-11	5-9	19- 31
Epistemic	4-8	4-8	2-6	10-22
Total over knowledge types	36	36	28	100

科学素养：态度

- 包含三部分内容
  - 对科学技术的兴趣
  - 环境意识
  - 对科学的研究方法的价值认同
- *通过问卷进行调查，结果不计入总分*
- 部分国家出现能力高分，态度低分

评价工具的开发

基于“科学素养”理论框架的测试蓝图

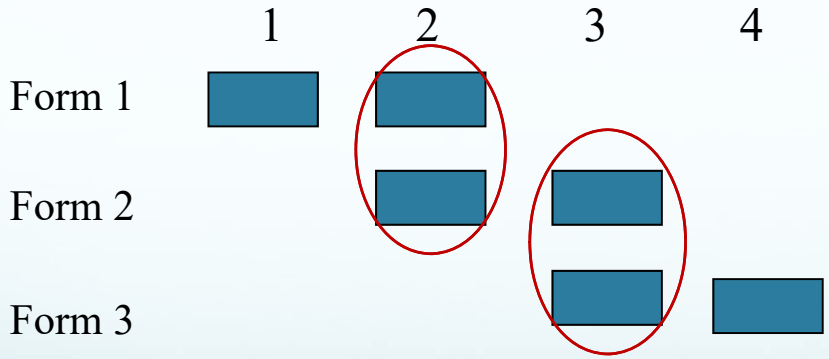


PISA的命题和组卷过程

- 由专业命题人员和部分参与国的专家根据测试框架编题
  1. 题目的总量约为实际测试的4倍
  2. 第一轮审核修改（item paneling)
  3. 小规模试测，测试后保留1/2左右在难度和内容上符合要求的题目
  4. 参与国家和地区进行翻译、本土化
  5. 通过第一轮试测的题目进行大规模试测，在所有参与的国家和地区抽样1000名左右，依据这次试测数据，技术试题的各项指数，审核试题的国际可比性
  6. 根据测试蓝图和试测结果，选择内容和难度适宜、具有国际可比性的性能良好的题目，组成最终测试的试卷

AB卷设计

- 题目多，而测试时间有限
- 2015年之前，组成13个题本（testlet）；2015年计算机考试，共产生396种组合
- 为了实现不同题本间可比，每个题目均包含一定数量的与其他题目相同的题目



- 每个题本力求内容、难度、性别和地域偏向的平衡

翻译与本土化

- 约翻译成44种语言
  - 英语和法语两个母版
  - 分别翻译为本国语言后，对比两个翻译版本
  - 修改和本土化
  - 交给第三方审核，确保在翻译和本土化过程中，题目语义和难度没有产生变化
  - 反馈和修改
  - 最后眼检

全球试测和最终定稿

- 试测目的
  - 检验试题的测量学性能（难度、区分度、单题与总分的相关度、答题数据与统计模型的拟合度、难度的跨国可比性等）
  - 检验测试组织工作的能力（抽样、监考、阅卷等）
  - 提供试题及评分指南等方面的修改意见

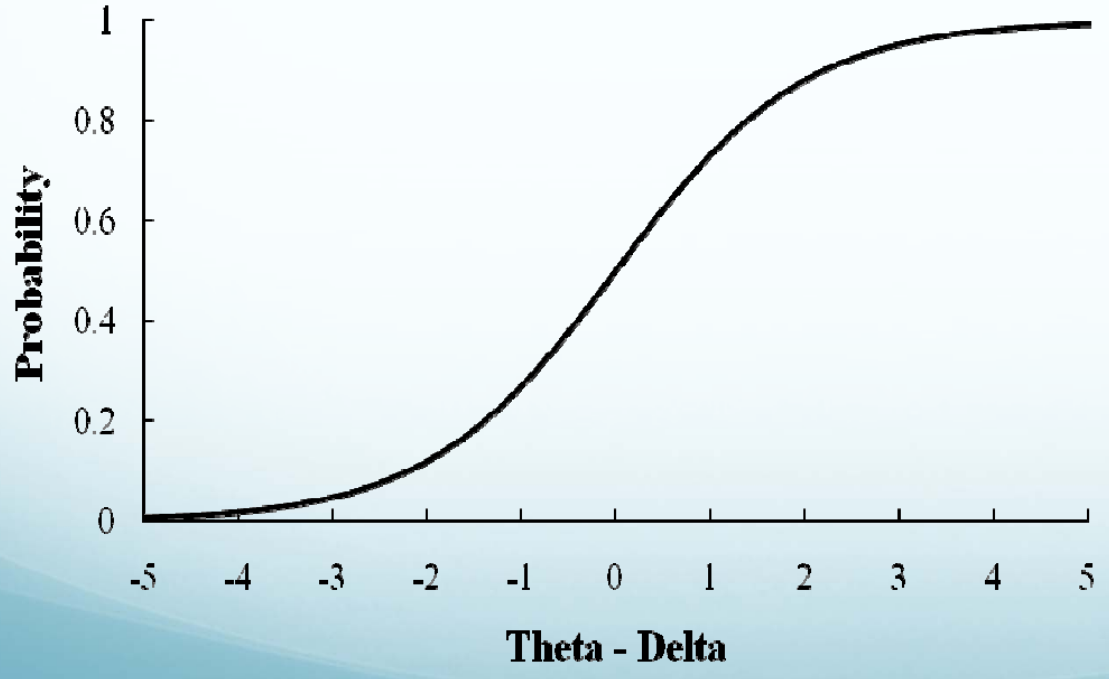
评价结果的分析

PISA的测量学分析

- 用项目反应理论(item response theory)估算学生能力

$$P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

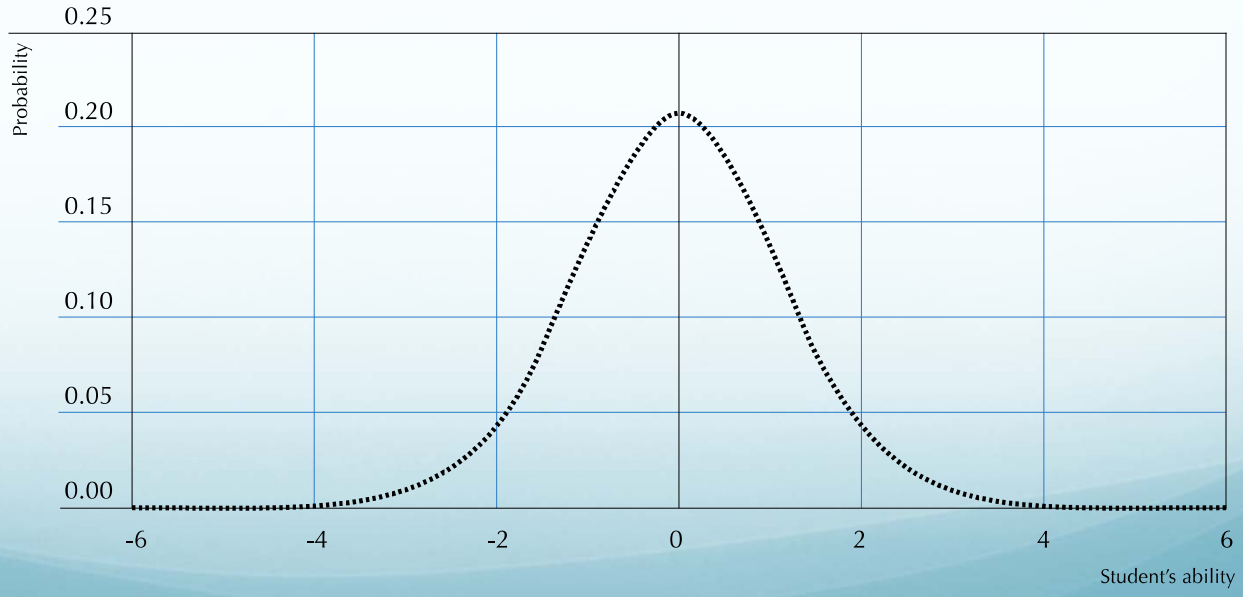
$$\exp = e^1 = 2.718...$$



Theta-Delta	exp(theta-delta)	P
-4	0.02	0.02
-3	0.05	0.05
-2	0.14	0.12
-1	0.37	0.27
0	1.00	0.50
1	2.72	0.73
2	7.39	0.88
3	20.09	0.95
4	54.60	0.98

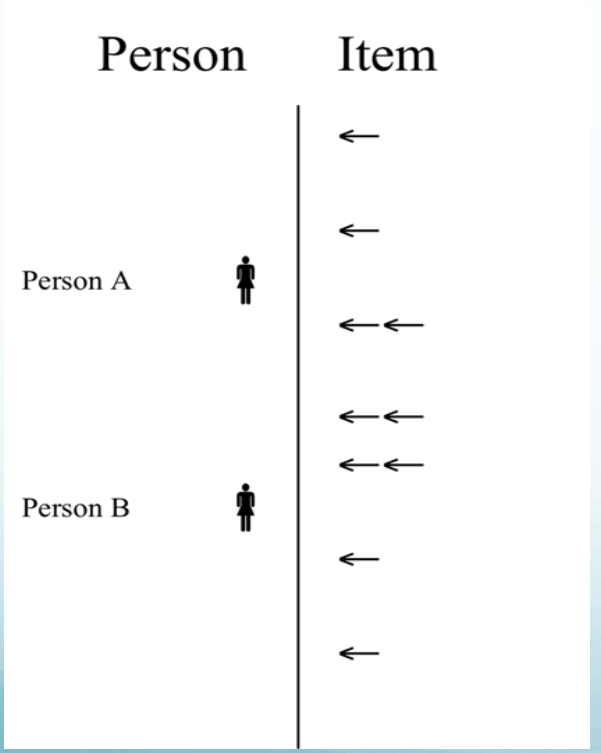
IRT模型中，能力如何估算

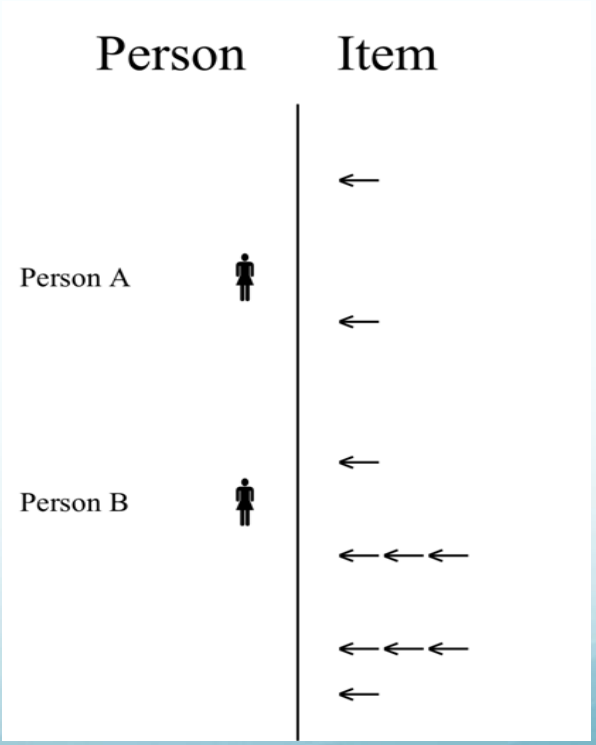
- 假设有4道题，难度分别为(-1，-0.5，0.5，1)
- 某考生的得分为（1，1，0，0）
- 根据 $$L(U_i \mid \theta_j, b_i) = \prod_{j=1}^{k} \prod_{i=1}^{n} P_i^{U_i} Q_i^{1-U_i}$$
- 代入不同的θ值（试用-4到4），该考生出现这样的表现的几率为



为什么不能使用原始分

- 试题难度＝答对试题的受试者百分比 （sample-dependent）
- 能力＝测试的原始得分 （test-dependent）
- 不同测试之间成绩不可比！





学生素养的发展变化

- IRT分值转换到平均分为500，方差为100的标准分
  - 后期数据分析发现，相差一个年级平均相差40分
- 跨年度等值
  - 连续两轮考试中，使用将近一半的共同题
  - 使用IRT的等值方法
  - 最后从IRT分值换算到标准分

试题难度的国际可比性的检验

- 项目差异反应（differential item functioning）
  - 能力相同的受试人，得分的几率是否相同
- 例子：

STRAWBERRY : RED

- (A) peach: ripe
- (B) leather: brown
- (C) grass: green
- (D) orange: round
- (E) lemon: yellow

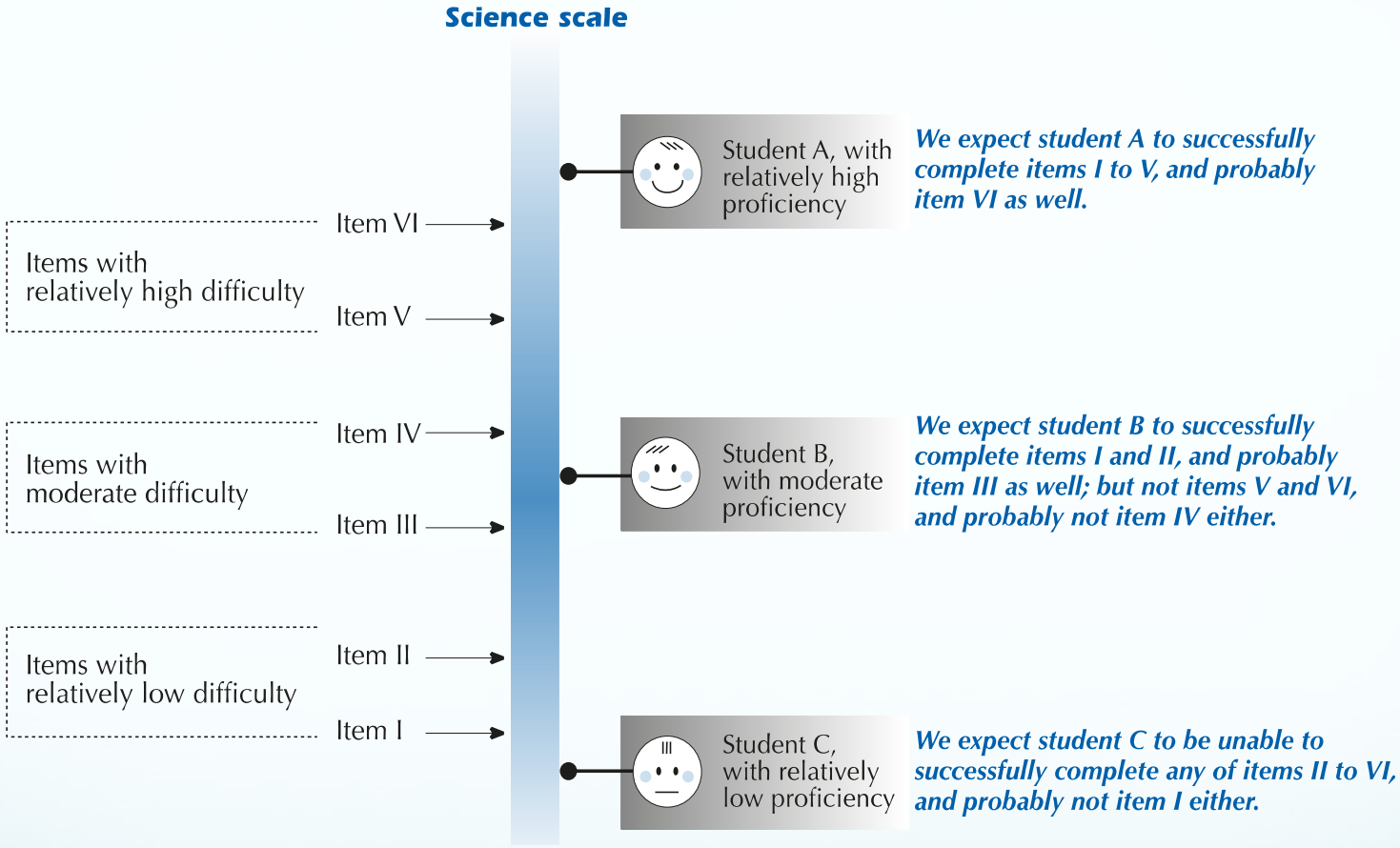
能力等级的划分

- 分值的含义是什么？
  - 通过试测数据分析和专家组对题目内容的质性分析，设定合格、优秀的分数标准和能力标准
- 例：科学素养的等级划分
  - 优秀：5-6级

Level	Lower score limit	Characteristics of tasks
6	708	At Level 6, students can draw on a range of interrelated scientific ideas and concepts from the physical, life and earth and space sciences and use content, procedural and epistemic knowledge in order to offer explanatory hypotheses of novel scientific phenomena, events and processes or to make predictions. In interpreting data and evidence, they are able to discriminate between relevant and irrelevant information and can draw on knowledge external to the normal school curriculum. They can distinguish between arguments that are based on scientific evidence and theory and those based on other considerations. Level 6 students can evaluate competing designs of complex experiments, field studies or simulations and justify their choices.
5	633	At Level 5, students can use abstract scientific ideas or concepts to explain unfamiliar and more complex phenomena, events and processes involving multiple causal links. They are able to apply more sophisticated epistemic knowledge to evaluate alternative experimental designs and justify their choices and use theoretical knowledge to interpret information or make predictions. Level 5 students can evaluate ways of exploring a given question scientifically and identify limitations in interpretations of data sets including sources and the effects of uncertainty in scientific data.

- 不合格：1级或以下

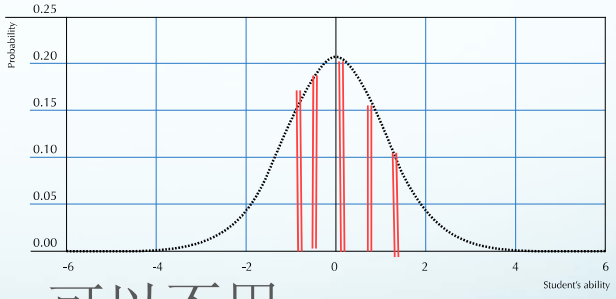
1a	335	At Level 1a, students are able to use basic or everyday content and procedural knowledge to recognise or identify explanations of simple scientific phenomenon. With support, they can undertake structured scientific enquiries with no more than two variables. They are able to identify simple causal or correlational relationships and interpret graphical and visual data that require a low level of cognitive demand. Level 1a students can select the best scientific explanation for given data in familiar personal, local and global contexts.
1b	261	At Level 1b, students can use basic or everyday scientific knowledge to recognise aspects of familiar or simple phenomenon. They are able to identify simple patterns in data, recognise basic scientific terms and follow explicit instructions to carry out a scientific procedure.



- 2015年，各国/地区平均约有7.7%的学生达到优秀；新加坡的优秀率为24.2%，台湾15.4%，芬兰14.3%。

其他技术细节

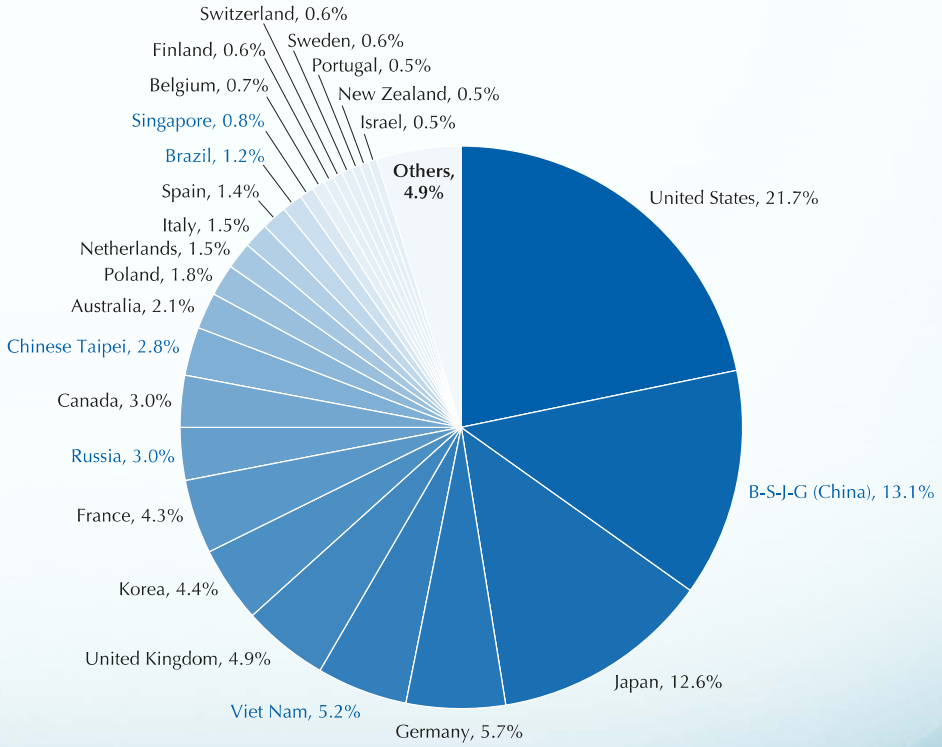
- 权重（weights）
  - 分层随机抽样
  - 考虑到部分群体要满足一定的样本量，会故意多抽
  - 每个学校和学生被抽到的几率不同
  - 在所有的分析中，都应该加权计算
- 似真值（Plausible value）
  - 给出5个似真值
  - 似真值的作用：更精确的误差估算
  - 在统计分析中，样本量较大的时候，可以不用



对PISA结果的使用

B_S_J_G_China

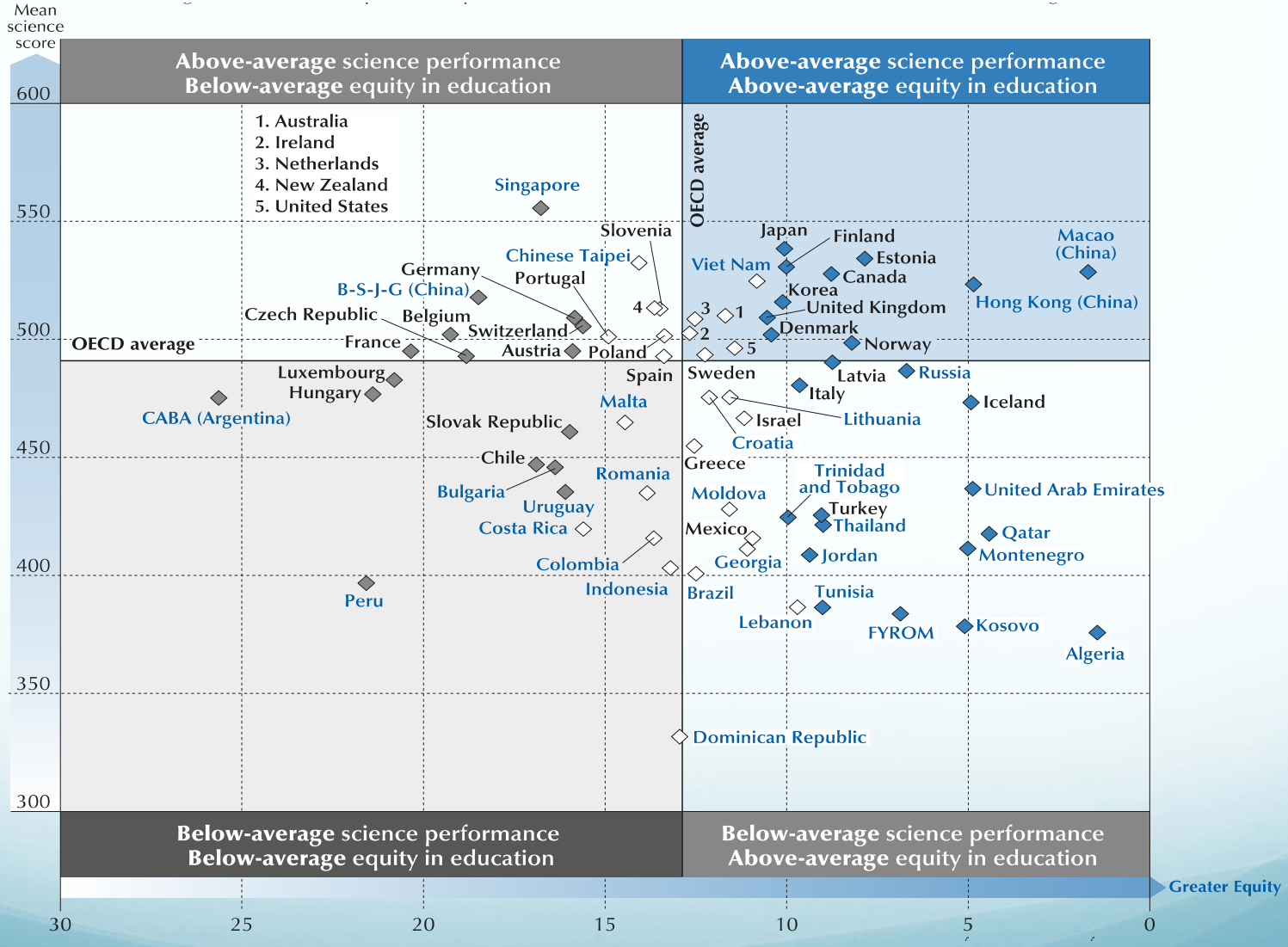
- 不是全国代表性抽样
  - 不是第一名在情理之中
- 三科排名
  - 科学：第10名
  - 数学：第6名
  - 阅读：接近OECD平均
- 全球优秀学生分布



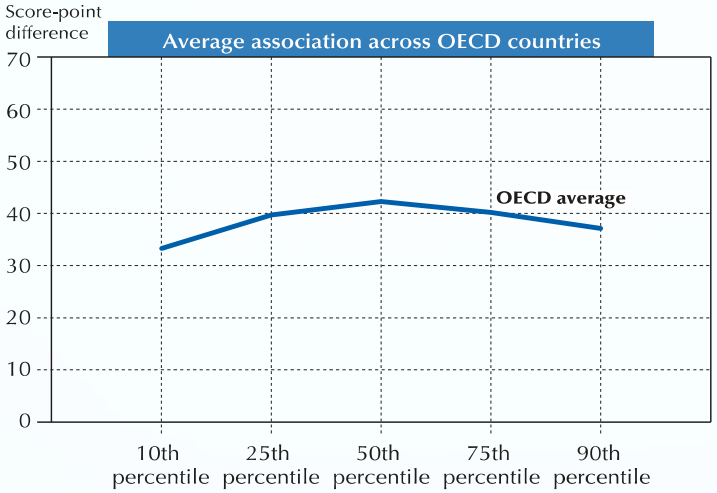
PISA结果：质量与公平

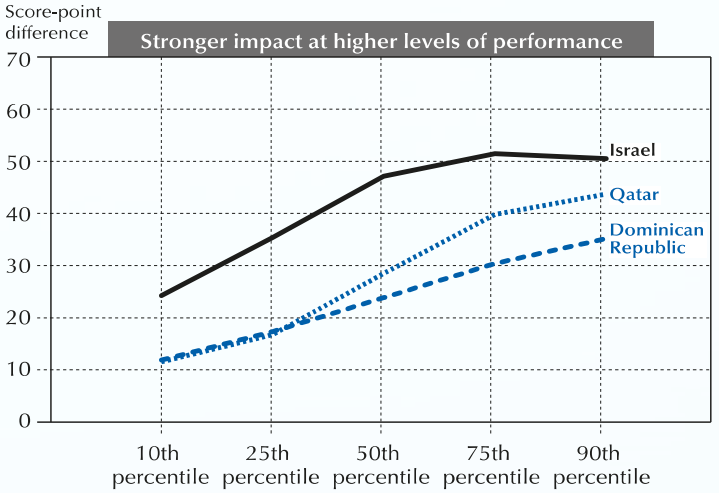
- 对质量的审视和反思
  - 高水平学生的比例
  - 不合格学生的比例
- 对教育公平的分析
  - 全纳
    - 是否获得教育机会和教育资源
    - 核心素养是否都合格
  - 公平
    - 受学生家庭经济文化背景的影响多大（ESCS指数，问卷中关于家庭背景的相关信息，通过principal factor analysis合成）
    - 处于经济文化背景弱势家庭的学生，有多大的机会“逆袭”：抗逆指数

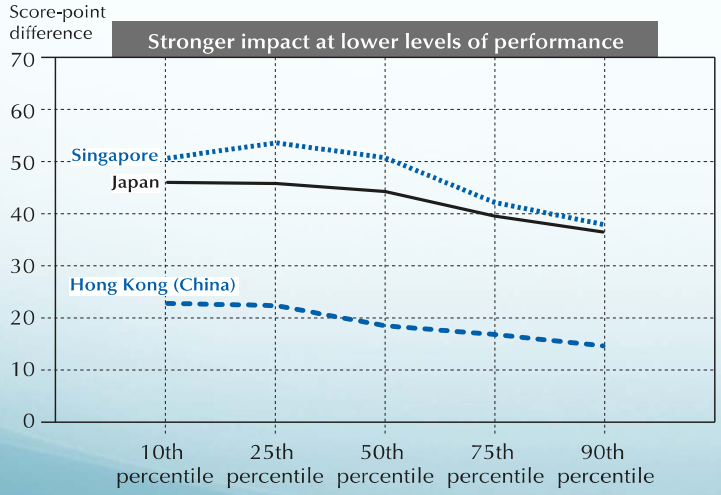
质量与公平能否兼得

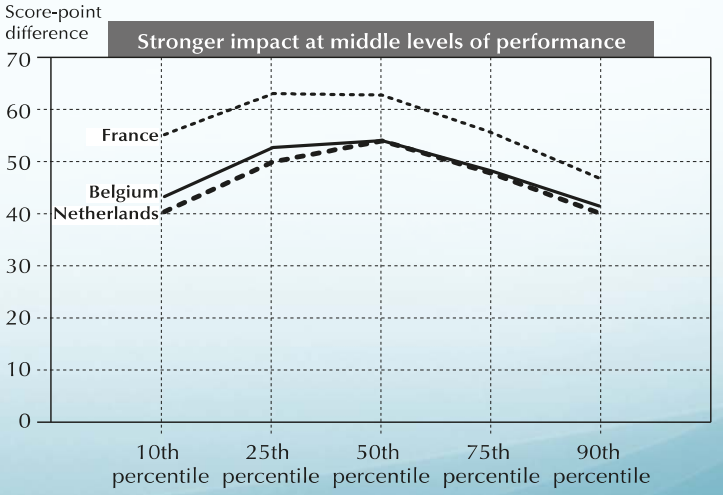

Mean science score
Above-average science performance
Below-average equity in education
Above-average science performance
Above-average equity in education
1. Australia
2. Ireland
3. Netherlands
4. New Zealand
5. United States
OECD average
Singapore
Slovenia
Chinese Taipei
Germany
Portugal
B-S-J-G (China)
Czech Republic
Belgium
Switzerland
France
Austria
Poland
Spain
Luxembourg
Hungary
CABA (Argentina)
Slovak Republic
Malta
Chile
Romania
Bulgaria
Uruguay
Costa Rica
Colombia
Indonesia
Peru
Japan
Finland
Estonia
Viet Nam
Canada
Korea
United Kingdom
Denmark
Norway
Macao (China)
Hong Kong (China)
Sweden
Latvia
Russia
Italy
Iceland
Israel
Lithuania
Croatia
Greece
Trinidad and Tobago
Moldova
Turkey
Thailand
Mexico
Georgia
Jordan
United Arab Emirates
Qatar
Montenegro
Brazil
Tunisia
Lebanon
FYROM
Kosovo
Algeria
Dominican Republic
Below-average science performance
Below-average equity in education
Below-average science performance
Above-average equity in education
Greater Equity
600
550
500
450
400
350
300
30
25
20
15
10
5
0

- 不同国家/地区，ESCS对不同水平的学生影响也不同









PISA结果：影响因素

- 对影响因素的分析
  - 通过宏观政策、学校、班级、学生个体四个层面的背景信息，了解教育资源分配和教学过程等相关信息
- PISA可以提供的国际比较
  - 学生成绩与个体因素（性别、是否为移民、ECSC、学习动力等）的关系
  - 学生成绩与学校因素（经费、规模、地区、鼓励学生学习的措施等）的关系
  - 校际vs.校内差异
  - 学校在个体因素与成绩之间的中介作用
  - 各种影响因素的历史变化（多次参与的国家和地区）
- PISA不能证明因果关系

各国/地区对PISA结果的使用

- PISA的政策导向
  - 课程改革
  - 对核心素养不合格的学生的帮扶政策
  - 经济文化背景弱势家庭的学生的帮扶政策
- 在部分国家/地区引起高度的重视
  - 德国、挪威：PISA shock
- 部分国家并不重视，没有后续的改革举措
  - 少数成绩优秀的国家（如：新西兰）
  - 少数没有政策意愿、资源或专业力量的国家

PISA面临的争议和挑战

- PISA能否定义全球的核心素养?
  - 到底考哪些素养
- PISA结果的全球可比性到底有多高?
  - 试题的全球可比性
    - *社会生活的背景，各地不同*
  - 兴趣、态度问卷的全球可比性
    - *社会倾向性难以避免*
    - *能力高分国家，兴趣与态度分数却偏低*
- PISA不能证明因果关系，影响因素分析有多可靠?
    - *部分高分国家，班额和成绩成正比*

没有一个测试是完美的…

- PISA为我们提供了一个全球视野
  - “素养”的理念
    - 对我国学生是否重要？能否在课程和教学中体现
    - 其他重要的核心素养
  - 现代的测评技术
    - 中、高考改革
    - 基础教育乃至高等教育阶段，核心素养如何测评？
  - 对教育公平的审视
    - 不合格的学生有哪些特征？是否聚集在部分学校？
    - 哪些政策能最有效的促进教育公平？

感谢聆听！

预祝大家新年快乐！

xthuang@pku.edu.cn