大学英语四、六级考试三十年之回顾

2016年4月17日，北京

引言

Introduction

CET的缘起和基础：

- 社会对英语能力的需求
- 英语教学的改革和发展
- 对学生英语能力的检测
- 语言测试的理论与实践

提要

第一部分 语言测试效度理论的发展

第二部分 CET考试三十年的回顾

第三部分 对考试发展方向的展望

1

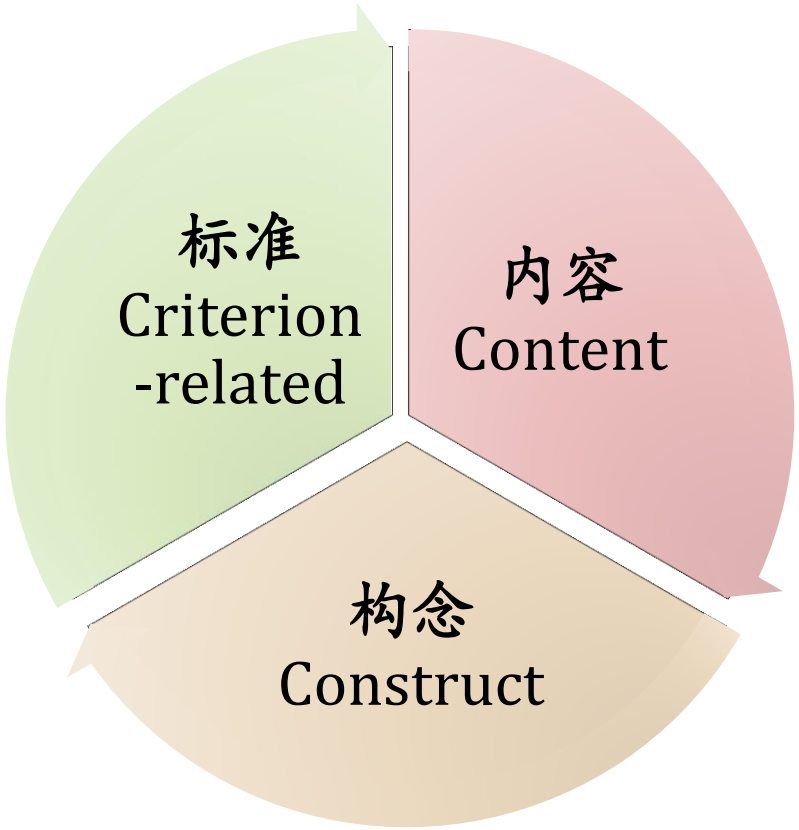
语言测试效度理论的发展

重要文献

- S. Messick (1989)
  Validity
- S. Messick (1995)
  Validity of Psychological Assessment
- C. A. Chapelle (1999)
  Validity in language assessment
- AERA/APA/NCME (2014)
  Standards for Educational and Psychological Testing

早期语言测试效度理论

The earlier trio of validities（60-70年代）



Robert Lado (1961): Language Testing
John Oller (1979): Language Tests at School

中期语言测试效度理论

Expansion on the earlier trio（80年代）

答题效度 Response validity	同期效度 Concurrent validity	预测效度 Predictive validity
情感 Affect	后效 Washback	伦理 Ethics

Grant Henning (1987): A Guide to Language Testing
Arthur Hughes (1989): Testing for Language Teachers
Michael Canale (1987): The Measurement of Communicative Competence

传统效度理论的主要问题

- 概念的割裂（fragmented）
- 概念不完整（incomplete）

- 未考虑：
  - 考试分数所包含的价值
    （value implications of score meaning）
  - 基于分数所做决定的社会后果
    （social consequences of score use）

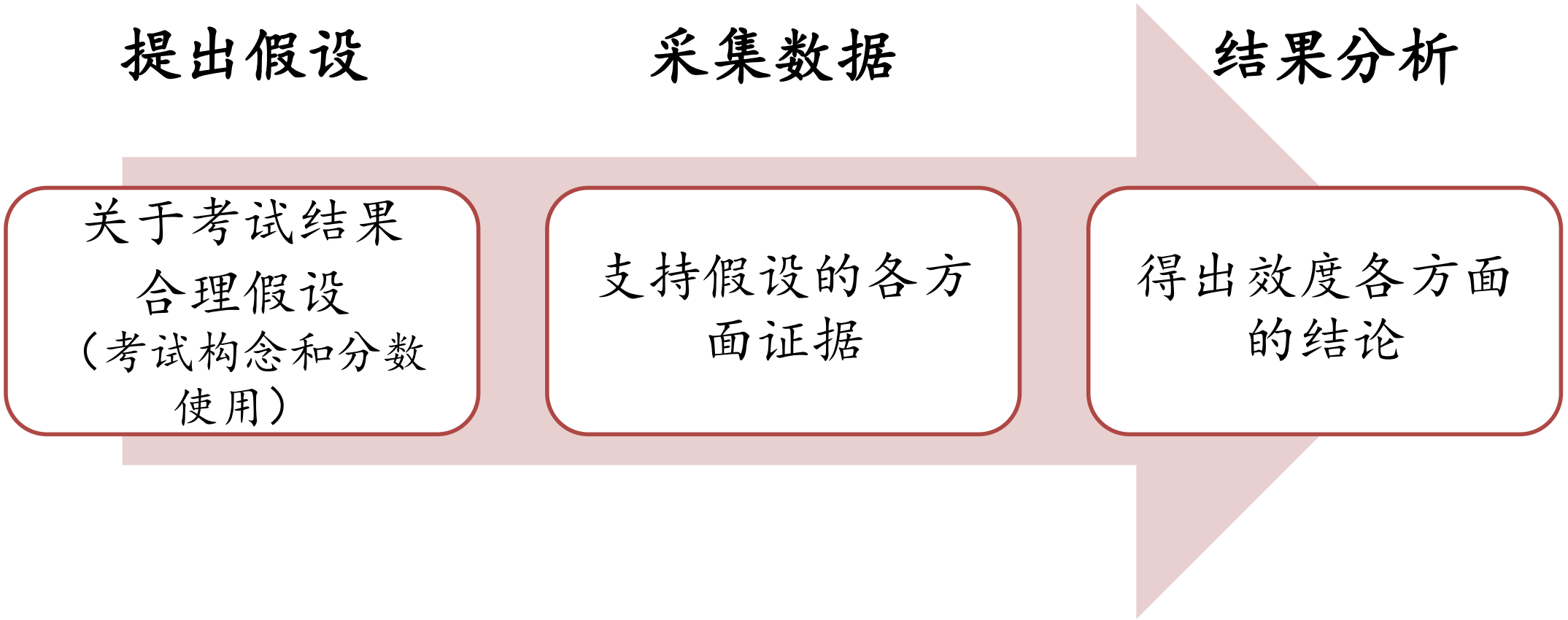
现代语言测试效度理论

90年代以后：Messick（1989, 1995）整体效度理论

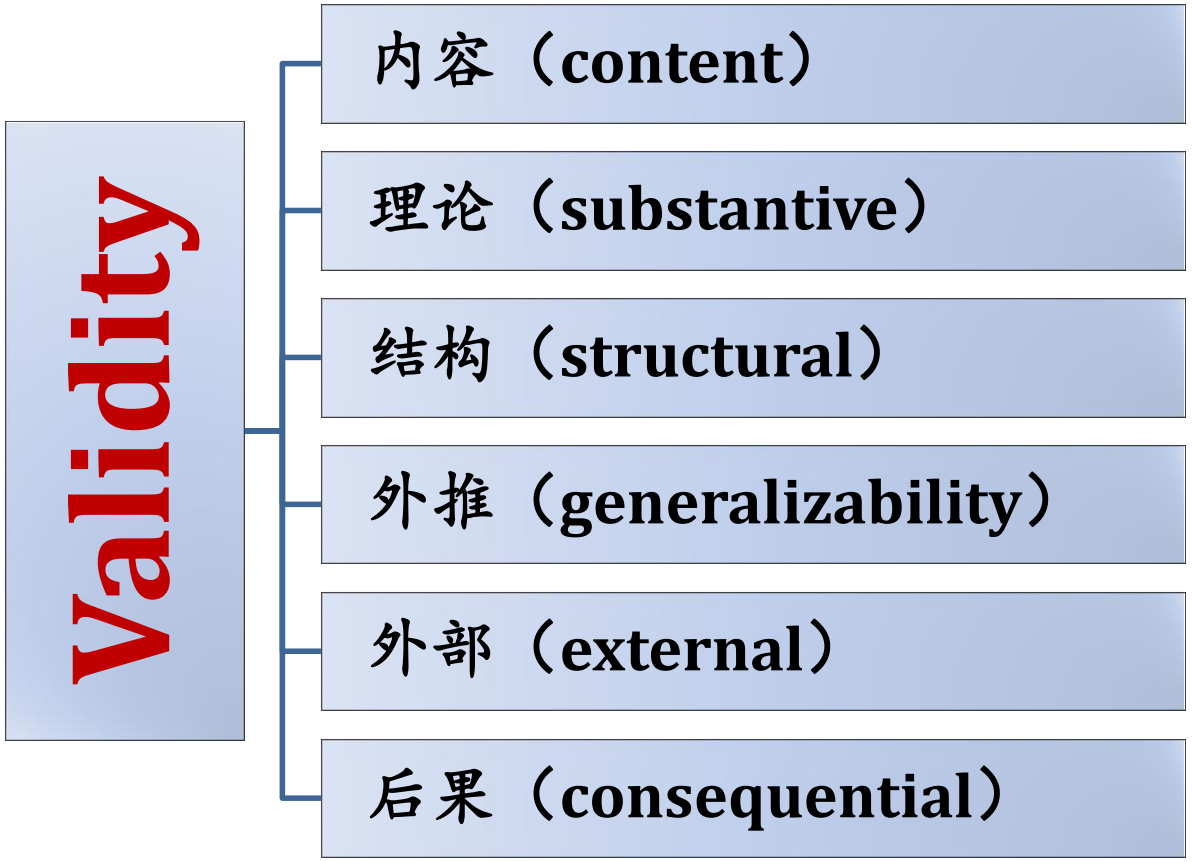
- 整体的概念（unified）
- 具有多层面（multifaceted）
- 构念效度为核心（construct validity）

- 效度论证需要实证研究和理据
- 效度论证关注考试分数的含义和使用
- 效度论证是一个持续进行的过程

效度论证的方法（1990s-）



效度的六个不同方面（distinguishable aspects）



多层面效度的递进矩阵

Facets of validity as a progressive matrix

	考试的功能和后果	
论证依据	分数解释	考试使用
效度的证据	构念效度(CV)	CV + 相关性/有用性(R/U)
使用的后果	CV + 价值含义(VI)	CV + R/U + VI + 社会后果

多层面效度的递进矩阵

若考试分数被用于某个用途，必须：

① 收集支持考试分数解释的证据（CV: construct validity）
② 论证考试分数所含的价值（VI: value implications）
③ 论证考试使用的相关性和有用性（R/U: relevance/utility)
④ 论证考试被用于该目的社会后果

2

大学英语四、六级考试：1987-2016

大学英语四、六级考试		
	四级	
	六级	
	口试	面试
		机考
	网考	四级
		六级

大学英语四、六级考试

CET-4	1987年
CET-6	1989年
IB-CET4	2008年
IB-CET6	2008年
CET-SET	1999年
CB CET-SET	2013年
CB CET-SET4	2015年
CB CET-SET6	2015年

- 标准化
- 标准参考
- 常模参照
- 高风险
- 大规模

考试内容和题型

考试评分和计分

成绩报道和使用

考试运行机制

1987-1996

1997-2006

2007-now

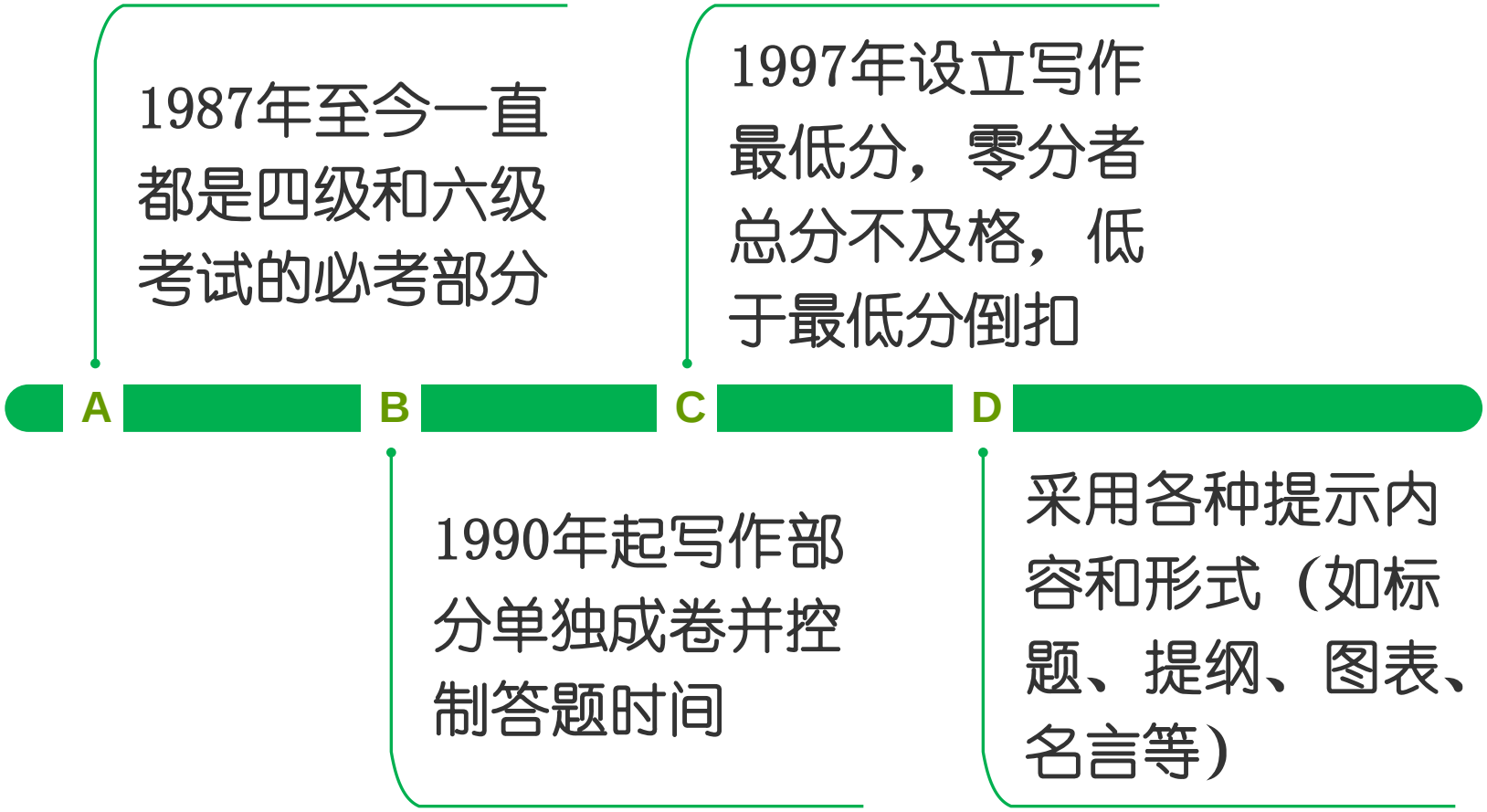
大学英语四、六级考试内容和题型（1987-1996）

听力（20%）	短对话和短文理解	多项选择题
阅读（40%）	仔细阅读理解	
知识（15%）	词汇语法知识	
综合（10%）	综合能力	完型填空或改错
写作（15%）	写作能力	短文写作
口语	不考核	

大学英语四、六级考试内容和题型（2016–）

部分	测试内容		测试题型	比例	时间
写作	写作		短文写作	15%	30分钟
听力理解	四级新闻	六级长对话	多项选择	15%	25-30分钟
	四级长对话	六级篇章			
	四级篇章	六级讲座或报道		20%	
阅读理解	词汇理解		选词填空	5%	40分钟
	长篇阅读		匹配	10%	
	仔细阅读		多项选择	20%	
翻译	汉译英		段落翻译	15%	30分钟
口语	四级口语、六级口语		报告A、B、C、D等级		

四、六级写作能力的考核


1987年至今一直都是四级和六级考试的必考部分
1997年设立写作最低分，零分者总分不及格，低于最低分倒扣
A
B
C
D
1990年起写作部分单独成卷并控制答题时间
采用各种提示内容和形式（如标题、提纲、图表、名言等）

四、六级翻译技能的考核

- 1998.6-2000.6
  单句英译汉，重点考核阅读理解能力
- 2006.6-2013.6
  单句汉译英，考核词和词组的用法以及句子结构
- 2013.12-
  段落汉译英，考核语言表达能力和初步翻译技能
  - 中国文化、历史、地理、经济、社会发展等
  - 四级140-160个汉字；六级180-200个汉字

面试型大学英语四、六级口语考试（1999-2013）

评分标准

- Accuracy & Range
- Size & Discourse Management
- Flexibility & Appropriacy

成绩报道：

A+, A, B+, B, C+, C, D

大学英语四、六级口语考试

1999-2012

2013-now

从面试向机考过渡

面试型口试的问题

- 耗时、效率低
- 投入大量人力
- 采用题量多
- 考官要求高
- 评分误差的后期调整难以实施

机考的优势

- 扩大规模，提高效
- 考务实施成本降低
- 采用的题量有所减少
- 去除考官的影响因素
- 监控评分过程，及时调整评分误差

大学英语六级口语考试（机考）

部分	时间	任务	任务描述
Part 1	3分钟	简短问答	▪ 自我介绍 ▪ 回答一个话题相关宽泛问题
Part 2	10分钟	个人陈述和 **两人讨论**	▪ 准备1分钟后根据提示作个人陈述（1.5分钟） ▪ **两位考生就指定话题讨论**（4.5分钟）
Part 3	2分钟	进一步提问	▪ 回答一个话题相关具体问题

大学英语四级口语考试（机考）

任务名称		任务描述	时间
0	自我介绍 Self introduction	考生作简短自我介绍	20秒/人
1	短文朗读 Reading aloud	考生朗读一篇120字左右的短文	准备：45秒 答题：1分钟/人
2	简短回答 Question & answer	考生回答2个与朗读短文有关的问题	答题：20秒/题
3	个人陈述 Individual presentation	考生经过准备根据所给提示作1分钟发言	准备：45秒 答题：1分钟/人
4	小组互动 Pair work	两位考生根据设定的情景和任务进行交谈	准备：1分钟 答题：3分钟(2人)
总计			约12分钟

鼓励学生参加四、六级口语考试

2016年5月起：

- 降低报名资格线：四级笔试成绩达到425分可参加大学英语四级口语考试；六级笔试成绩达到425分可参加大学英语六级口语考试；
- 合并成绩报告单：大学英语四、六级笔试成绩单与大学英语四、六级口语考试成绩单合二为一。

2015年6月和12月考试（笔试）

- 考生群体包括研究生、本科生和专科生，本科生为主体。

2015年6月和12月本科生的人数和得分情况如下：

年次	级别	人数	听力(249)	阅读(249)	翻译和写作(212)	总分(710)
2015年6月	四级	3703709	130	136	125	392
	六级	2578459	118	145	109	373
2015年12月	四级	4165758	130	138	126	395
	六级	2443694	121	149	109	381

四级考试本科院校分层随机抽样

〈试题难度 〉	满分	均值	标准差	平均难度
听力理解	35	20.22	7.19	0.58
阅读理解	35	24.94	7.05	0.71
翻译和写作	30	17.68	4.75	0.59
整卷	100	62.84	16.98	0.63

注：2013年12月数据，N=3427

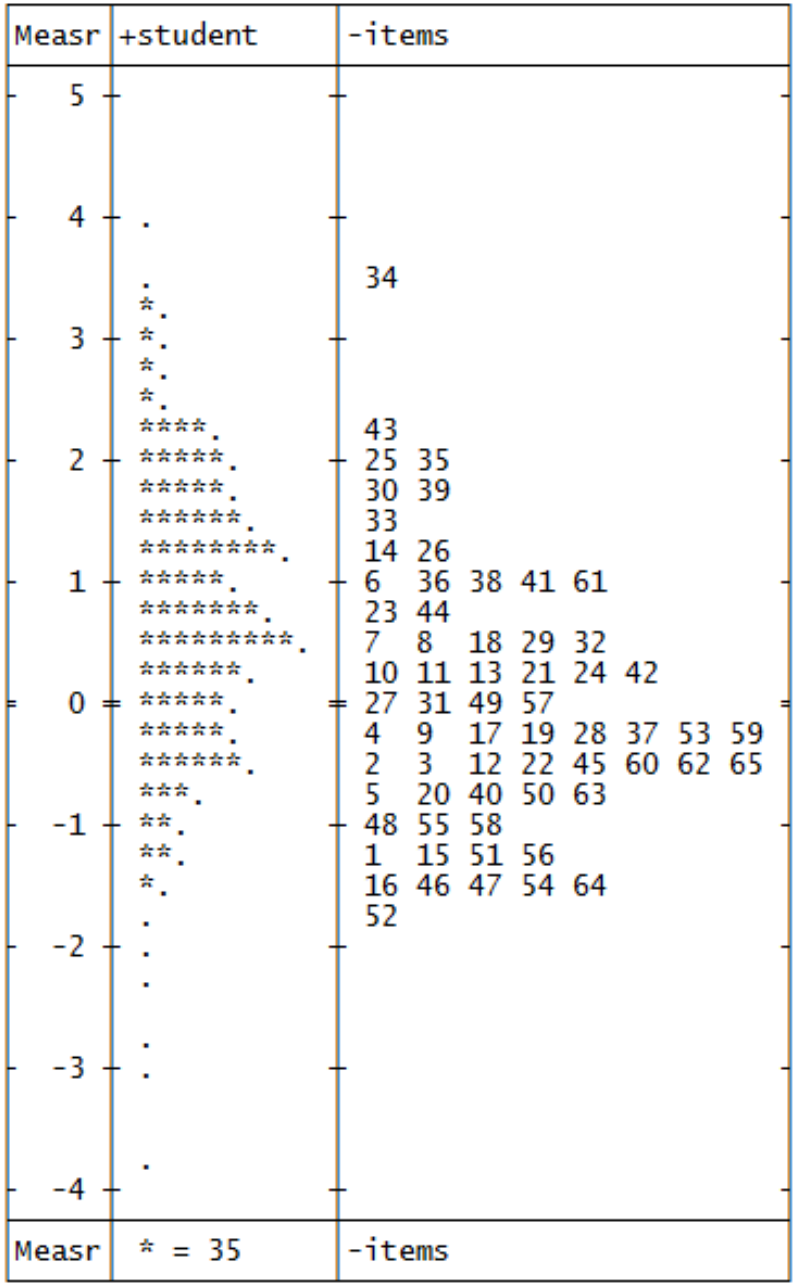
四级考试本科院校分层随机抽样

	听力理解	阅读理解	翻译和写作	整卷
听力理解	1	.72	.67	.91
阅读理解		1	.67	.91
翻译和写作			1	.84
整卷				1

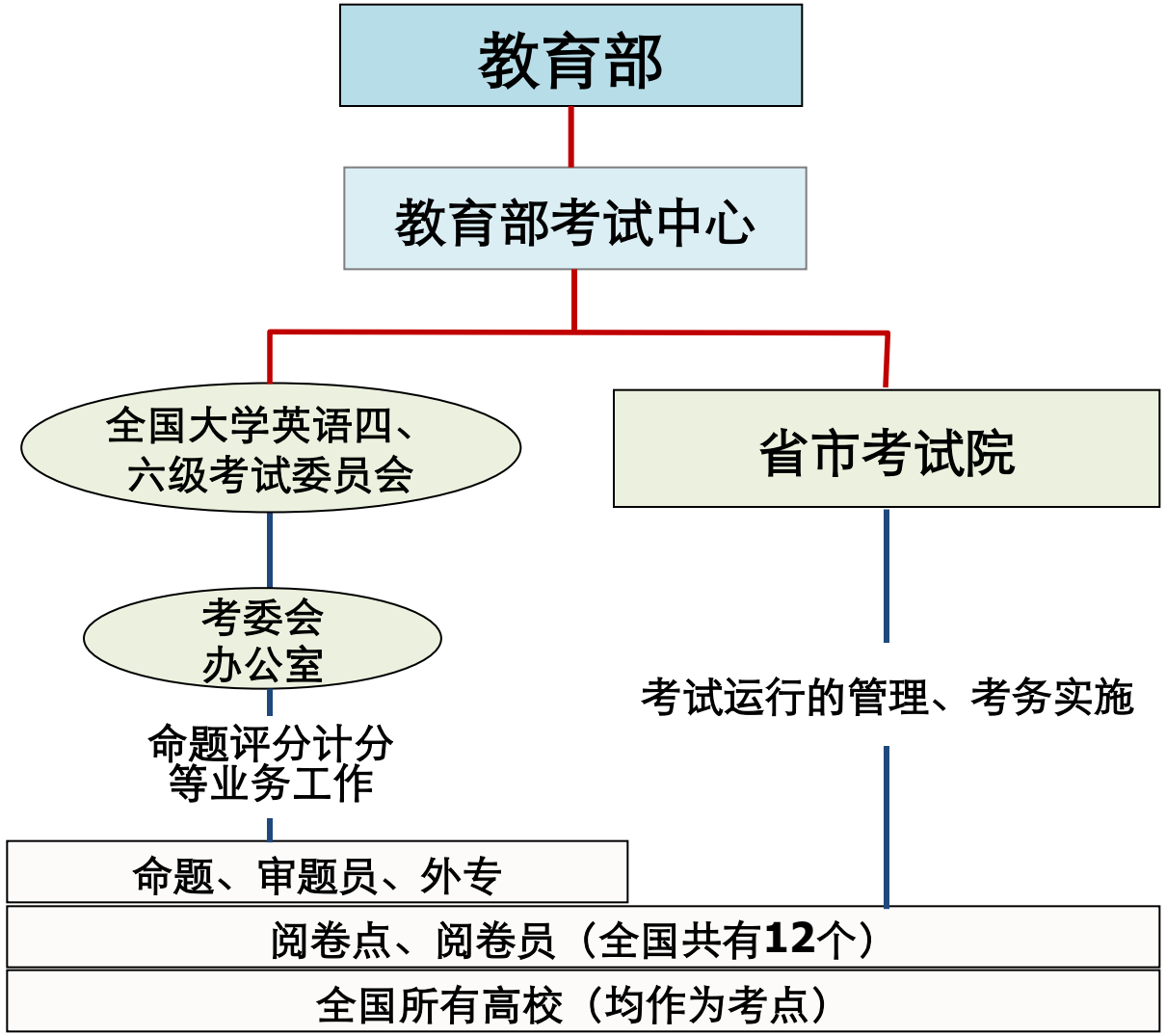
注：2013年12月数据，N=3427

考生能力与试题难度的对应：

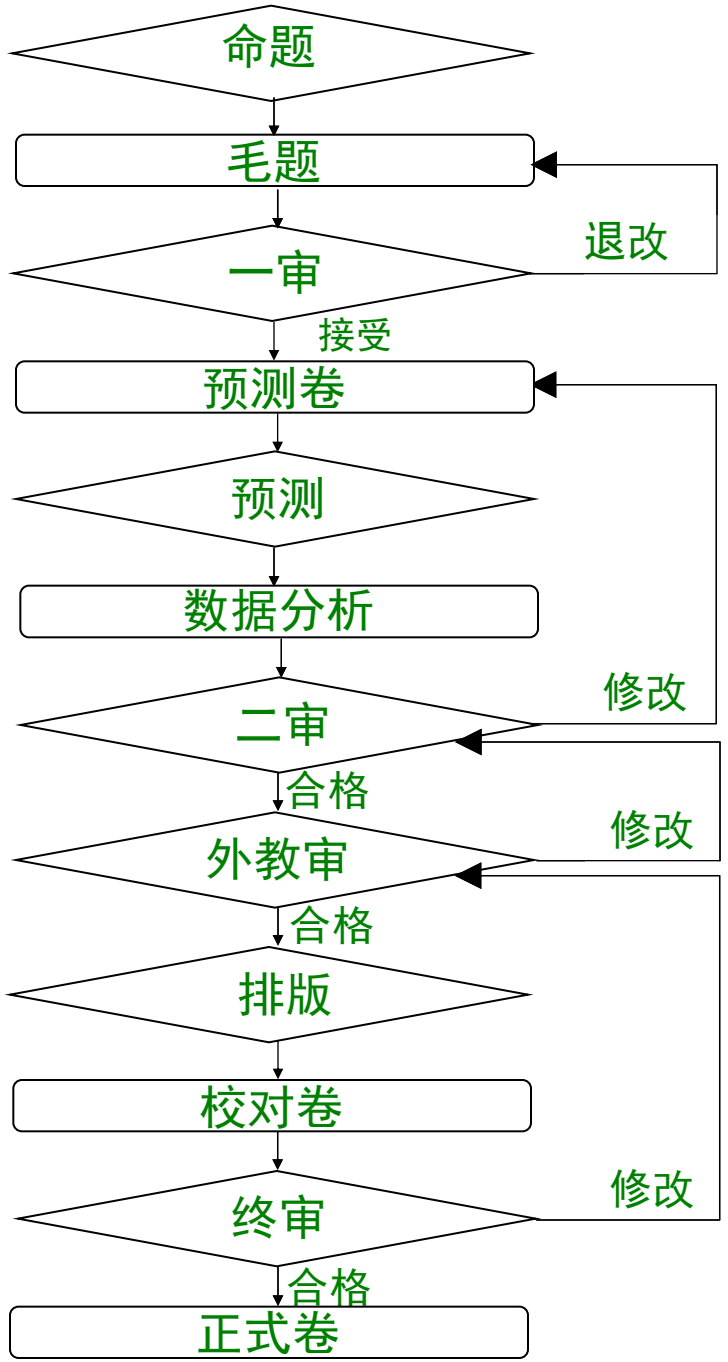
1. 考生能力基本呈正态分布，且分布较分散。
2. 试题难度覆盖了绝大多数考生语言能力，分布较均匀，考生的水平与试题分布基本匹配，试卷可对考生能力水平做出比较精确的估计。
3. 第34题（听写）最难，与其他题目难度的差异较大。
4. 绝大多数试题集中分布在±2个logits范围内，试题总体难度分布合理。



大学英语四、六级考试运行模式



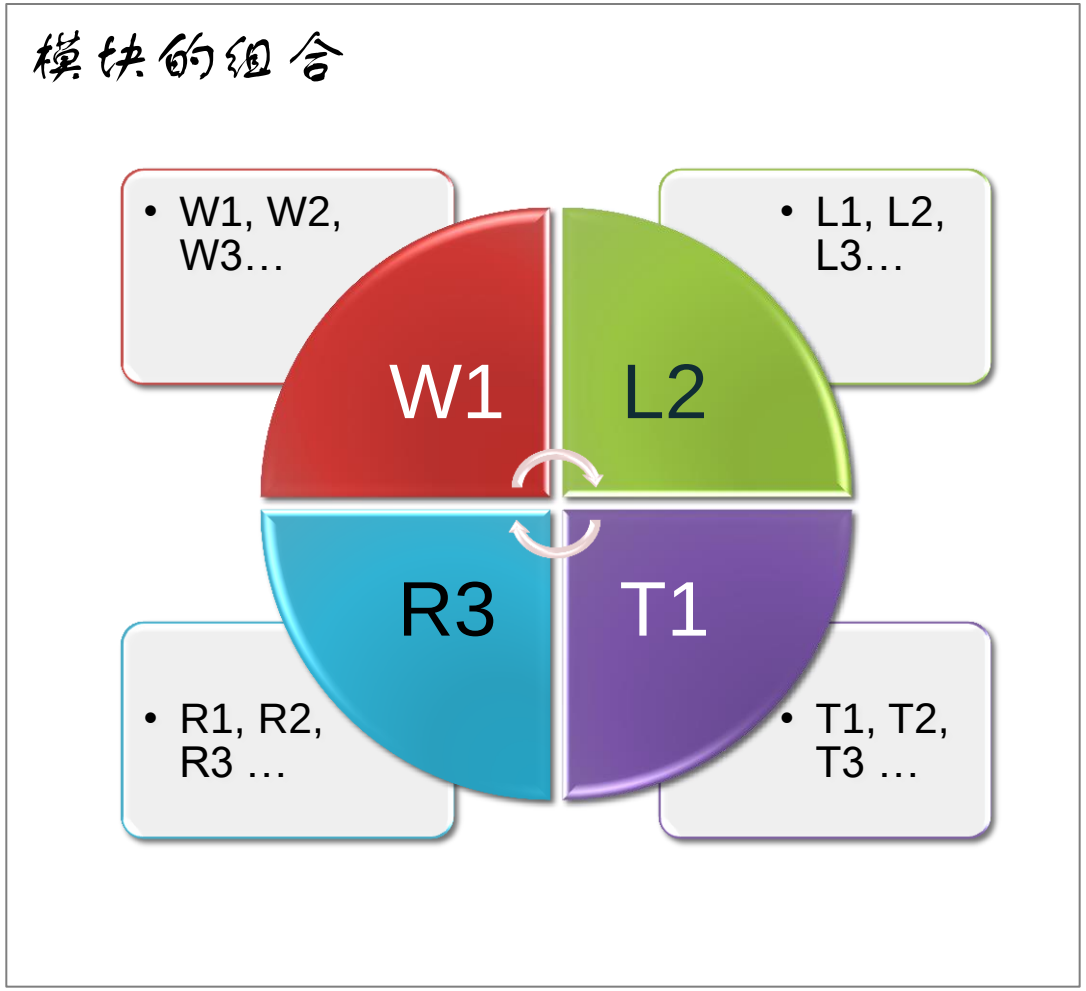
四、六级考试命题和审题流程



四、六级考试实施：多题多卷模式

实施方式：

- 不同内容的试题+模块组合：
  - W1+L2+R3+T1
  - W2+L3+R2+T2
  - W3+L1+R1+T3
  - …
- 每个考场内有n套不同内容(除听力)的试题



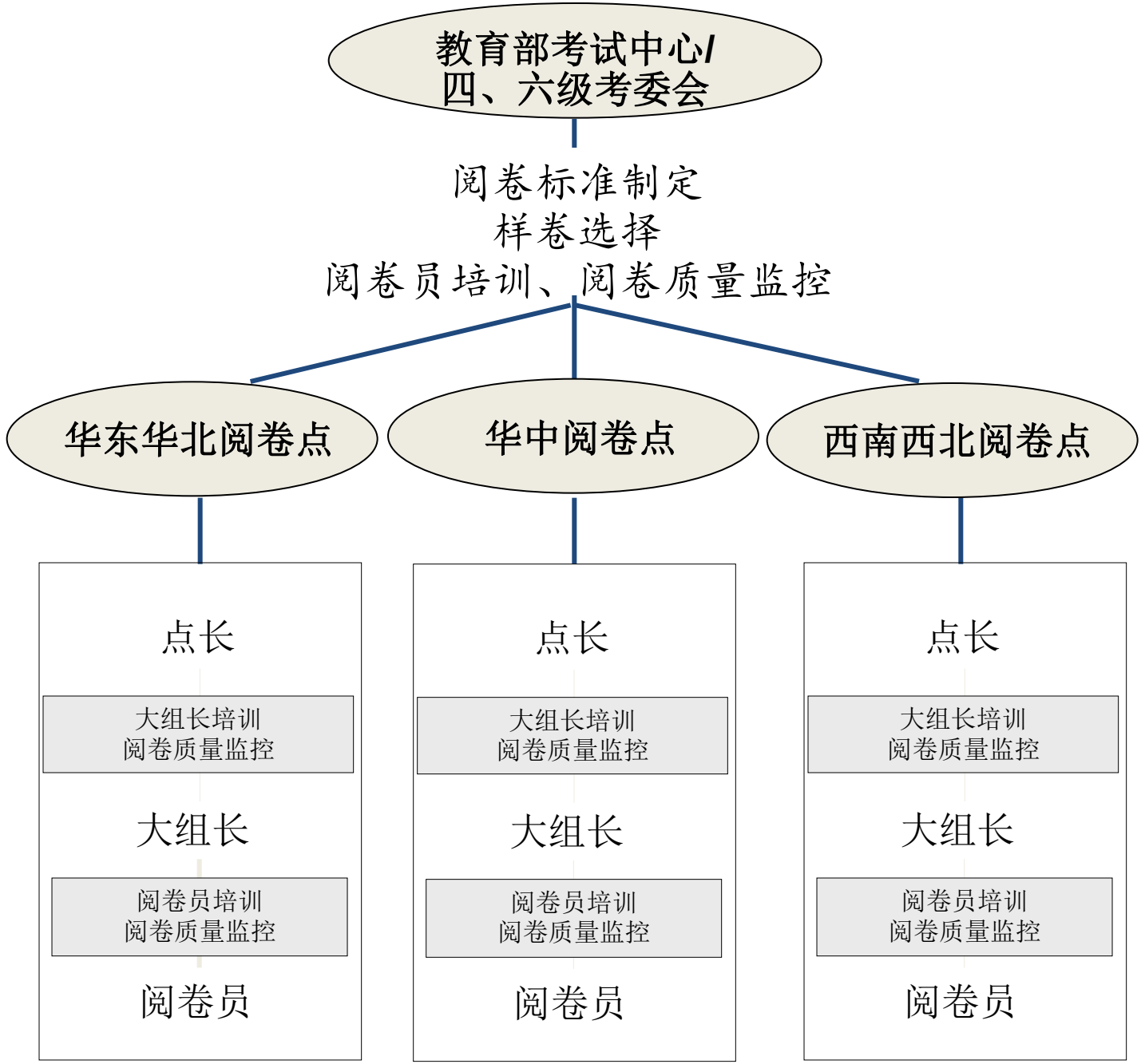
四、六级考试实施：多题多卷模式

该模式对考试各环节提出挑战，包括：

1. 命题、试题预测、审题
2. 制卷和印卷、考试实施
3. 主观题的阅卷评分
4. 试题难度等值和计分

作用：有效预防了高科技手段的群体性作弊

主观题阅卷


教育部考试中心/
四、六级考委会
阅卷标准制定
样卷选择
阅卷员培训、阅卷质量监控
华东华北阅卷点
华中阅卷点
西南西北阅卷点
点长
大组长培训
阅卷质量监控
大组长
阅卷员培训
阅卷质量监控
阅卷员
点长
大组长培训
阅卷质量监控
大组长
阅卷员培训
阅卷质量监控
阅卷员
点长
大组长培训
阅卷质量监控
大组长
阅卷员培训
阅卷质量监控
阅卷员

四、六级考试的计分体制

<table>
<tr><th></th><th>均值
(标准差)</th><th colspan="2">分数报道</th><th>分数解释</th></tr>
<tr><td rowspan="2">1987/89
~2004</td><td rowspan="2">72
(12)</td><td>及格</td><td>优秀</td><td rowspan="3">六所常模院校中的
百分位位置</td></tr>
<tr><td>60</td><td>85</td></tr>
<tr><td>2005.06</td><td rowspan="2">500
(70)</td><td colspan="2" rowspan="2">不设及格线
报道单项和总分</td></tr>
<tr><td>2006.12~</td><td>十六所常模院校中的
百分位位置</td></tr>
</table>

四、六级考试的成绩计算

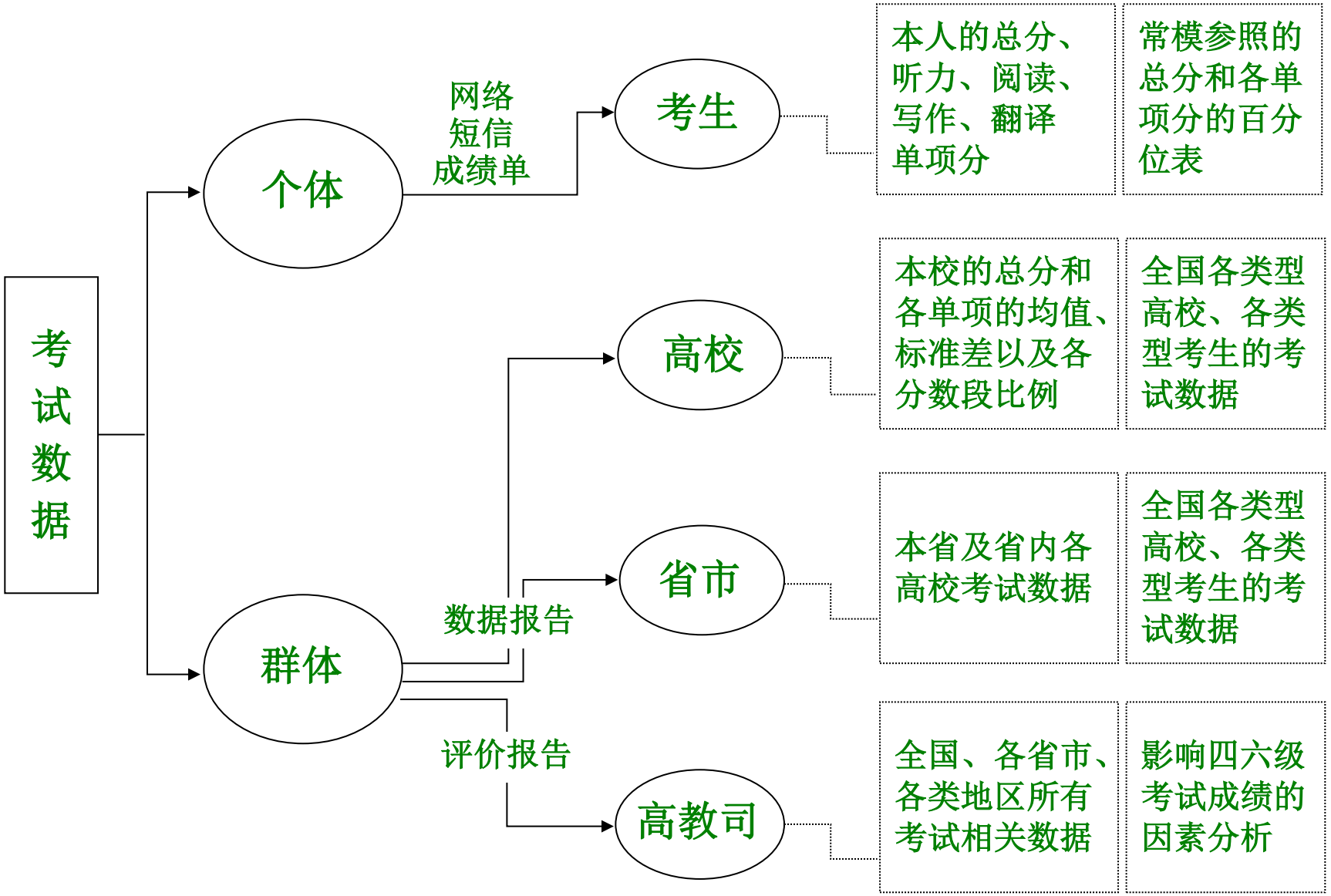
各部分试题的原始得分

各部分试题难度等值

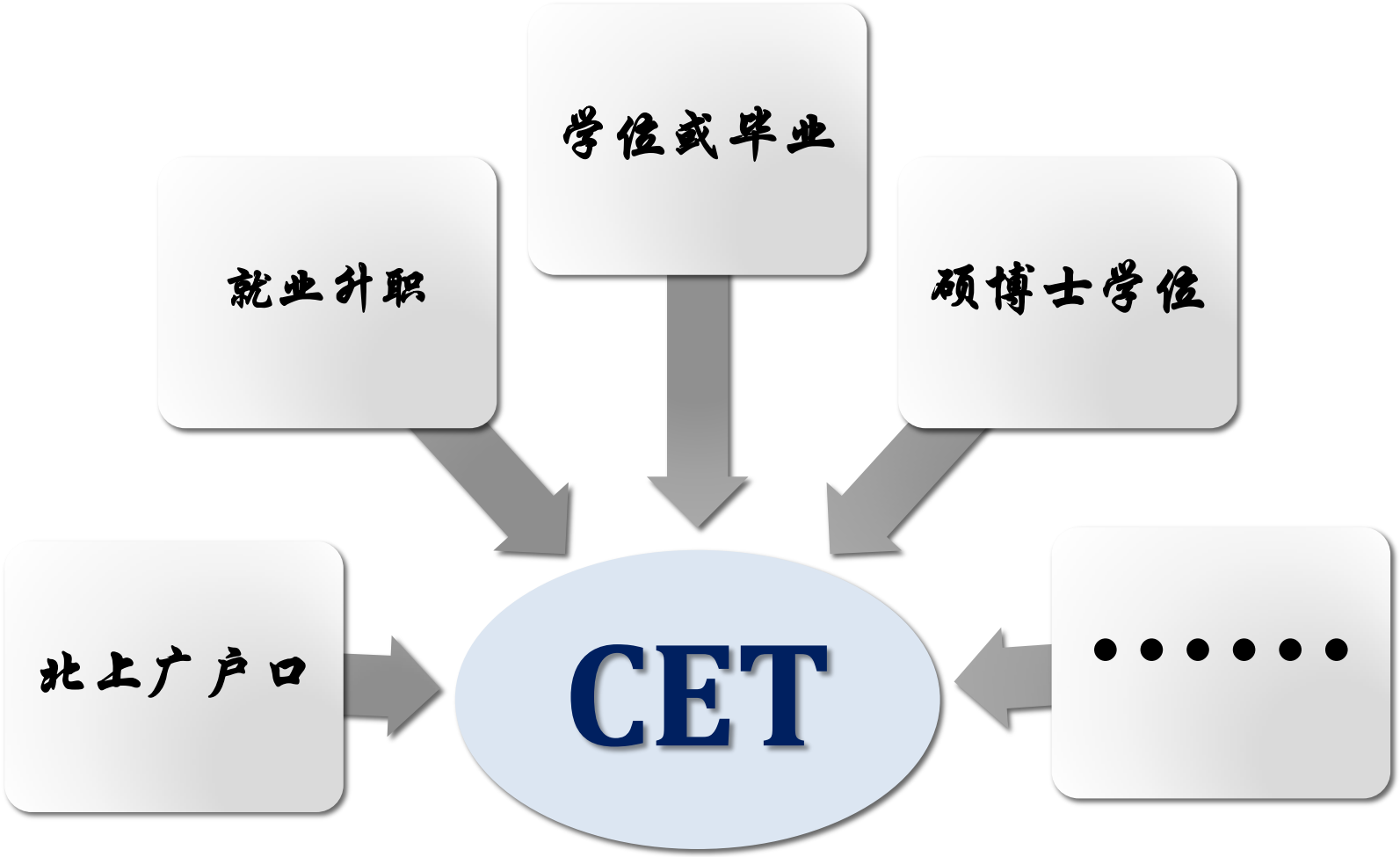
常模转换

报道：总分以及各单项分

四、六级考试的成绩报道


考试数据
个体
网络
短信
成绩单
考生
本人的总分、听力、阅读、写作、翻译单项分
常模参照的总分和各单项分的百分位表
群体
数据报告
高校
本校的总分和各单项的均值、标准差以及各分数段比例
全国各类型高校、各类型考生的考试数据
省市
本省及省内各高校考试数据
全国各类型高校、各类型考生的考试数据
评价报告
高教司
全国、各省市、各类地区所有考试相关数据
影响四六级考试成绩的因素分析

四、六级考试的使用→高风险考试



3

大学英语四、六级考试的发展方向

效度理论的挑战

C. A. Chapelle (1999) Validity in language assessment

1. **构念定义：**
   支持构念定义的理论；
   构念与考试使用结合；
2. **支持考试使用的论据：**
   采集所有能支持考试使用的论据；
   整合各方面的论据以论证考试效度；
3. **效度理论的实际运用：**
   考试使用者需论证考试的合理使用；
   教育并帮助考试使用者合理使用考试。

Modern Language Testing at the Turn of the Century: Assuring What We Count Counts

开展效度研究，关注考试使用

"a strong program of test validation that includes considerations of ethical test use"

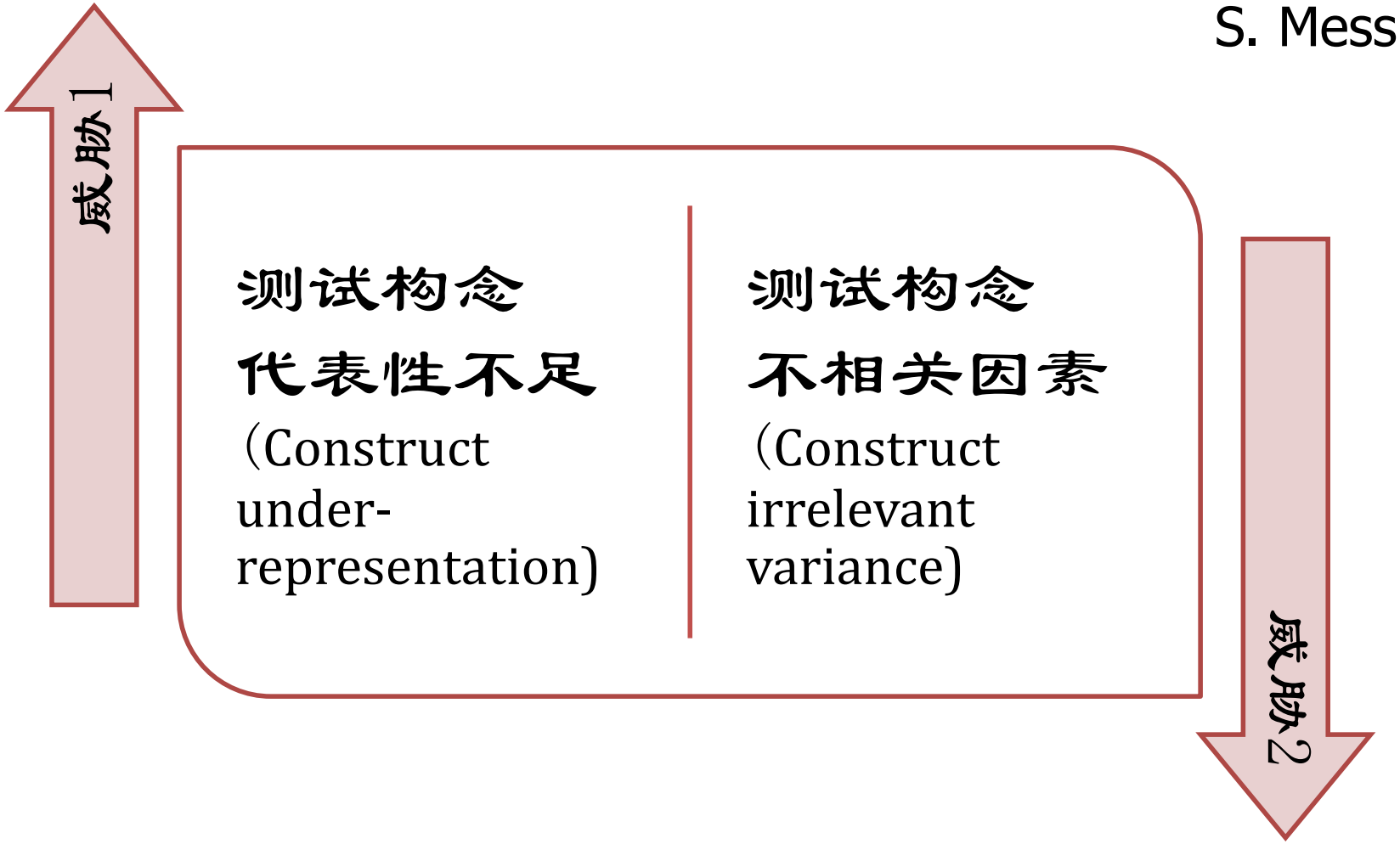
Bachman (2000, p. 1)

New Directions

考试构念

威胁测试构念的两种方式

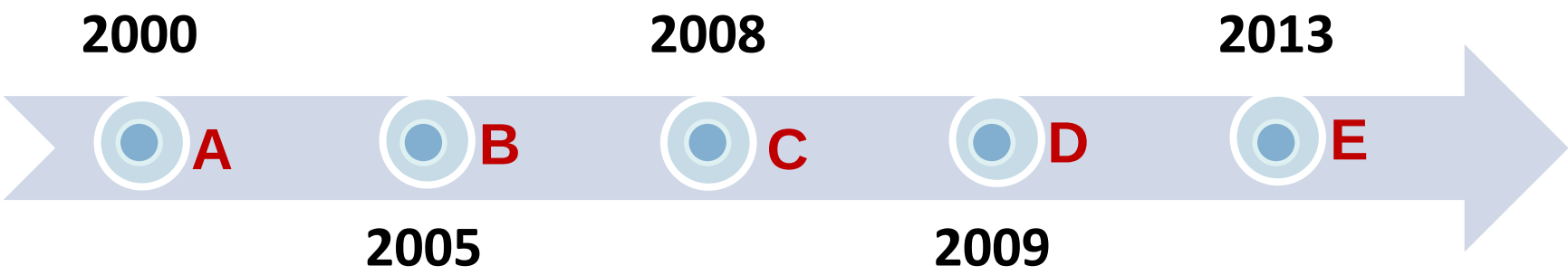
S. Messick (1996)



母语者标准（native speaker norm）

- McNamara & Roever (2006)：更多研究母语者标准对英语教学和测试的适用性
- Wang & Hill (2011)：英语越来越多地成为非母语者之间交流工具，亚洲地区的英语教学应充分考虑这一现实背景
- J.D. Brown (2012)：提出与World Englishes研究团队合作的七项建议，建议充分借鉴其研究成果

大规模计算机化语言测试



A.雅思机考（CB-IELTS）
B.托福网考（TOEFL iBT）
C.四、六级网考（IB-CET）
D.培生考试（PTE-Academic）
E.四、六级口试（CB-CET SET）

基于计算机和网络的语言测试构念定义

计算机能力/数字化能力(computer/digital literacy)是21世纪语言测试构念的干扰因素吗(Construct-Irrelevant Variance)?

- Jin & Wu, 2009 四、六级纸笔与网考成绩对比
- Jin & Zhang, X., 2013 四、六级网考与培生考试对比
- Jin & Yan, under review 四、六级纸笔与网考的写作测试对比
- Jin & Zhang, L., 2015) 四、六级机考与面试型口试中交际策略运用对比
- Taylor et al. 1998 计算机熟悉程度对托福考试的影响
- Sawaki 2001 L2阅读机考和纸笔考试对比
- Choi et al. 2003 韩国TEPS机考与纸笔考试对比
- Brown 2003 雅思手写与打字写作测试对比
- Breland et al. 2004　托福手写与打字写作测试对比
- Wolfe & Manalo 2005 测试模式对托福写作考试的影响
- Weir et al. 2007 雅思写作机考和纸笔考试对比

局部构念定义(local construct definition)

语境中的语言运用能力：

(abilities – in language users – in contexts)

语言测试的构念定义应该将语言使用者的能力与语境结合起来，探索两者之间的关联以及语言使用者如何运用相关知识和能力完成各类交际任务。

(Chalhoub-Deville, 2003: 378)

New Directions

分数解释

解释考试分数的含义

- 完善常模参照体系（norm-referenced）
  - 常模的调整或重建
  - 百分位表的制定和使用

- 加强标准相关的解释（criterion-related）
  - 采用能做描述语解释分数的含义
  - 与**大学英语教学指南**的对接
  - 与**中国英语能力量表**的对接
  - 与**国外语言标准体系**的对接

考试使用

高风险考试可能产生的后效

→ **事实上的课程标准**（the *De Facto* curriculum）

- 导致应试教学和学习

→ **考试作弊、伪造或使用证书**

- 导致考试丧失公平公正性，学生丧失道德

→ **考试的科学性和决策的可信度**

- 当定量的工具被用于社会决策后，可能会既破坏了工具的科学性也影响了决策的可信度

(Campbell, 1975, in Madaus et al., 2009: 155)

考试的使用和误用

(B. Spolsky, 1995: 358)

- 更多研究、关注考试的预期用途：考试必须有明确的考试目的（即预期的用途），这决定对考试各方面的要求（如考试的信度）。

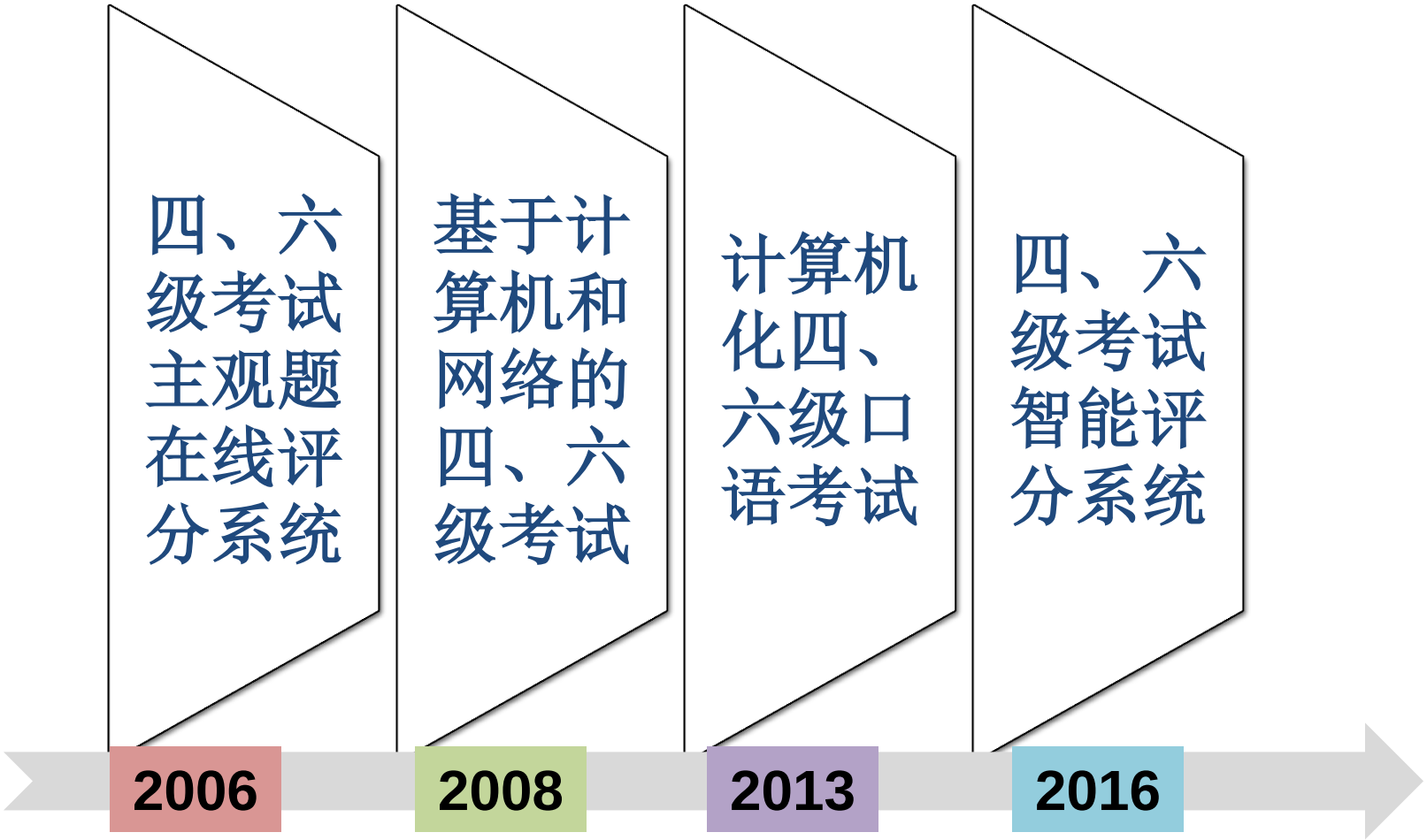
(Bachman & Palmer, 2010, p. 429)

- 预防考试过度使用或误用：一旦考试分数被用于预期的用途并产生影响之后，考试就开始掌控自己的命运，考试使用者利用考试的影响力，将其用于其他用途，而这些命运并非设计者所预期的。

New Directions

信息技术

信息技术在四、六级考试中的运用



结语

语言测试关注的三个主要问题

Three concerns have dominated language testing since the 1960s （A. Davies, 2014）

1. 怎么考?（How to test）
2. 考什么?（What to test）
3. 测试者的职责?（Who are the testers）

四、六级考试三十年

1987-1996：
考试内容和形式的设计以及考试体系建设

1997-2006：
考试题型改革以及口语考试的实施和推广

2007-now：
机考的试点和实施以及考试的社会学研究

考什么？
考试内容

怎么考？
考试方法

为何考？
考试使用

考试设计者

明确考试目的

解释分数含义

研究考试后效

考试使用者

知晓考试目的

理解分数含义

制定合理政策

(Bachman and Palmer, 2010; Davies, 1997, 2004)

沟通交流、合作互赢，推动考试结果的合理使用以及考试职业和伦理道德的发展

参考文献

杨惠中、C. Weir. 1998.《大学英语四六级考试效度研究》上海外语教育出版社

杨惠中、金艳“大学英语四、六级考试分数解释”《外语界》2001年第一期

杨惠中“大学英语四、六级考试十五年回顾”《外国语》2003年第三期

杨惠中、桂诗春“语言测试的社会学思考”《现代外语》2007年第四期

杨惠中、桂诗春（编）《语言测试社会学》2015年. 上海外语教育出版社

张琳、陈琳丽“大学英语四级考试质量评估：基于经典测量理论和Rasch模型的数据分析”《当代外语研究》2015年第十期

Jin, Y., & H. Yang. 2006. The English Proficiency of College and University Students in China: As Reflected in the CET. *Language, Culture and Curriculum*, 19 (1), 21-36.

Jin, Y. 2010. The National College English Testing Committee. In L. Cheng & A. Curtis (Eds.). English Language Assessment and the Chinese Learner. Routledge, Taylor & Francis Group (pp44-59).

Jin, Y. 2014. The Limits of Language Tests and Language Testing: Challenges and Opportunities Facing the College English Test. In Coniam, D. (ed.). *English Language Education and Assessment: Recent Developments in Hong Kong and the Chinese Mainland* (pp. 155-169). Springer Singapore.

上海交通大学外国语学院

语言测试与评估研修班

2016年8月1日-6日

联系人：罗老师（139 0190 3652）

校本考试；考试设计；命题；统计分析；

ESP测试；大规模考试改革

谢谢！ 欢迎意见和建议！

yjin@sjtu.edu.cn