北京大学 PEKING UNIVERSITY
eLearning Forum Asia
Next Generation Learning
Visions • Innovations • Possibilities
eLearning Forum Asia 2012
24-27 April 2012 @ Peking University, China

自动化汉语口语考试研发与思考

Research and Thoughts of Automated Test of Spoken Chinese

北京大学　　李晓琪

Peking University　Li Xiaoqi

2012年4月

April 2012

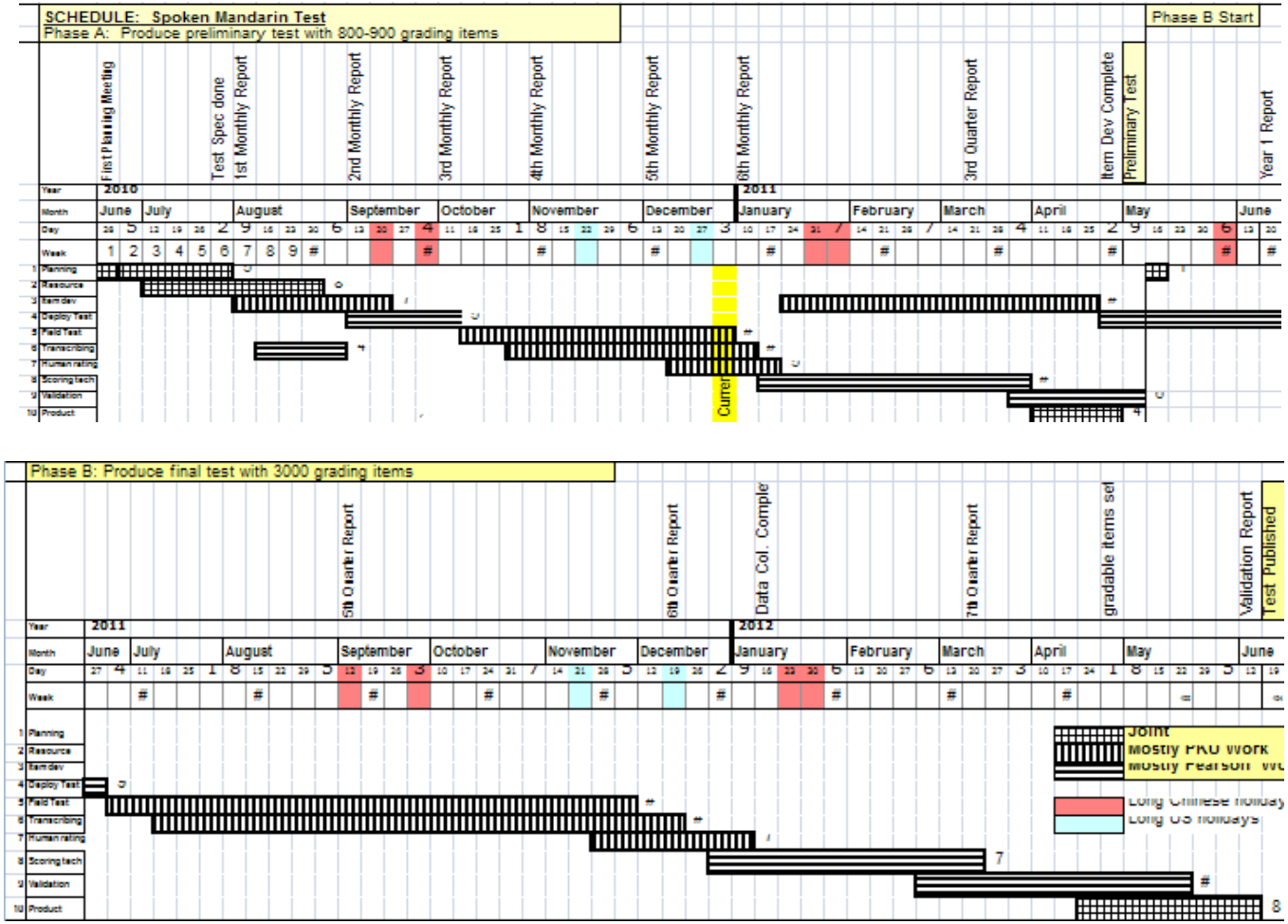
项目概述 *Project Overview*

- **简介(*Brief Introduction*)：** 自动化汉语口语能力测试项目由北京大学和培生（PEARSON）公司合作开发。该项目旨在开发一个面向汉语学习者（以中文作为第二语言）的汉语口语考试，通过建立起一个成熟的语音识别系统，实现计算机自动评分，在考生完成考试2分钟后即可给出成绩。该项目从2010年7月正式启动研发，计划在2012年8月推出考试产品。

- **阶段安排(*Two Phases*)：** 项目研发分为A、B 两个阶段。A阶段从2010年7月到2011年4月，进行小规模预测，初步建立起自动评分模型。B阶段从2011年5月到2012年8月，该阶段展开大规模预测，进一步完善自动评分系统。

研发日程表

Development Schedule



项 目 概 述

Project Overview

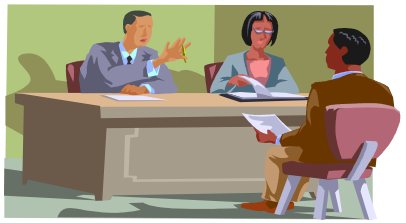
- **项目特点***(The Characteristic)*:

项目的显著特点是考试由Versant的系统自动实施。所谓自动实施包括：自动派发和自动评分。

因此该考试可以在任何时间、任何地点通过电话或计算机进行，自动评分系统通过语音处理技术即时生成客观、可靠的分数。自动化汉语口语能力测试是一个全自动化的口语考试。

口语考试的方法
Methods of Spoken Language Test

- 直接考试 *Interview*
  - *面试*
  - *成对或分组测试*

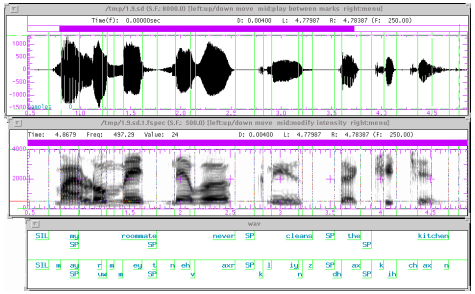


考官:人
评分:人

- 半直接考试 *Semi-direct test*
  - *以计算机为媒介*
  - *在磁带上录制考试过程*

考官:技术
评分: 人

- 自动化考试 *Automation test*
  - *以计算机为媒介*
  - *语音处理*



考官: 技术
评分: 技术

Versant 考试

Versant Test

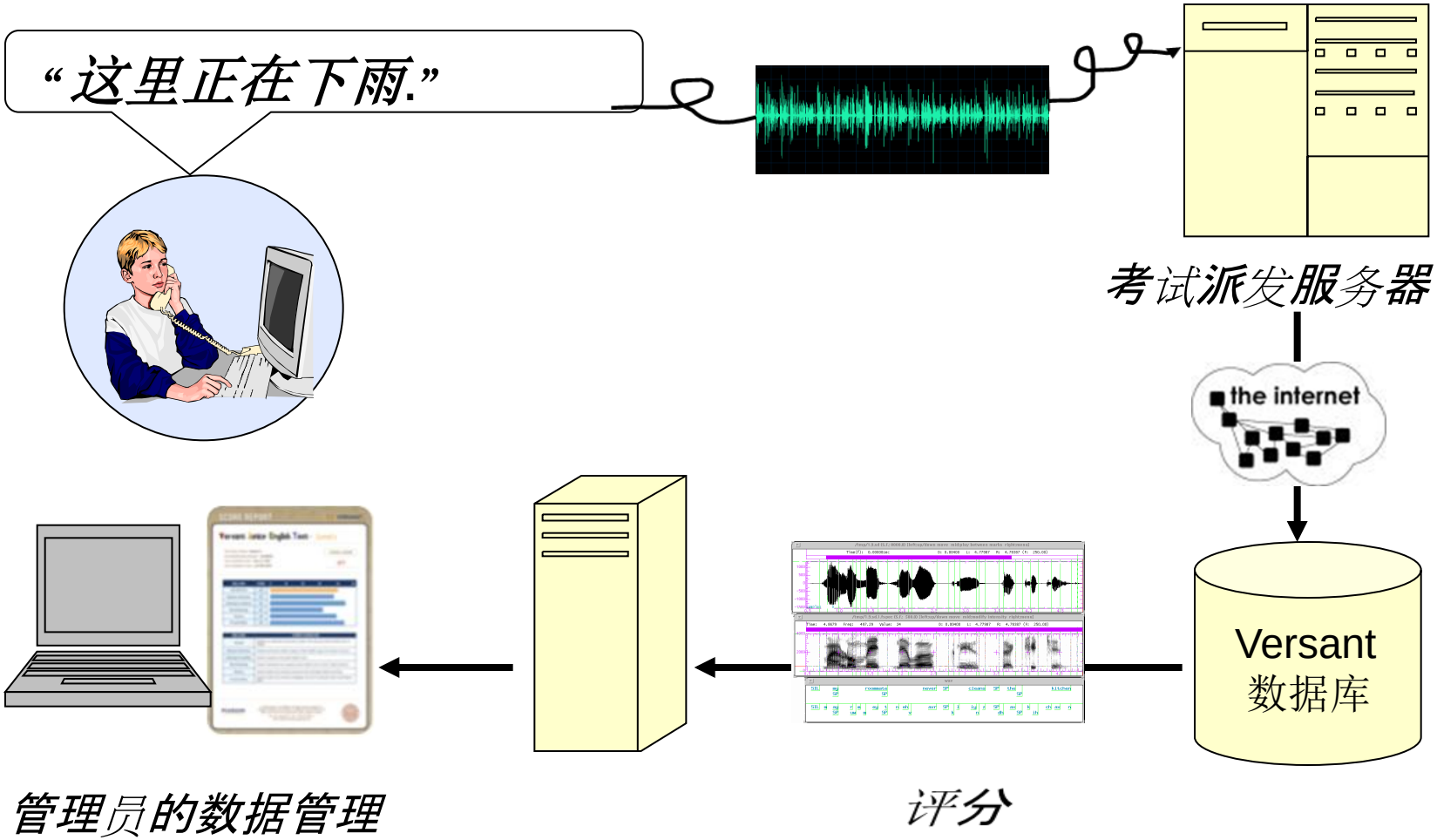
- 全自动口语考试
- 自动派发

  通过电话或电脑进行考试
- 自动评分

  通过语音处理技术和计算机评分系统来为考试评分

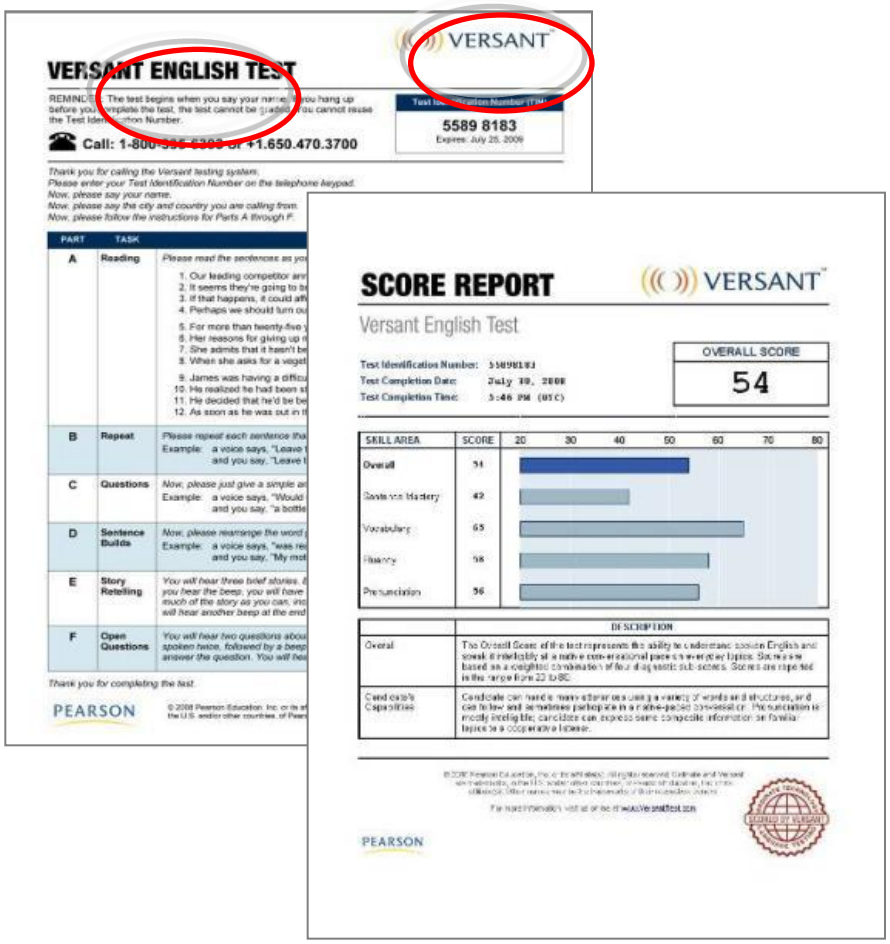
Versant 方法

Versant's Operating Method



Versant 试卷和分数报告

Test Paper and Score Report of Versant Test



- *试卷*
  - *朗读句子 Read Sentences*
  - *重复句子 Repeat Sentences*
  - *回答问题 Short answer questions*
  - *重组句子 Sentence builds*
  - *短文重述 Passage retellings*

- *整体分数 Overall*
- *句子的掌握 Sentence*
- *词汇 Vocabulary*
- *流利度 Fluency*
- *发音 Pronunciation*

汉语口语考试
Spoken Chinese Test

汉语口语考试

SPOKEN CHINESE TEST

REMINDER: The test begins when you say your name. If you hang up before you complete the test, the test cannot be graded. You cannot reuse the Test Identification Number.

Call: +86.800.810.0278

Test Identification Number (TIN)
5379 3440
Expires: December 8, 2012

Thank you for calling the Versant testing system.
Please enter your Test Identification Number on the telephone keypad.
Now, please say your name.Now, please say the city and country in which you are taking this test.
Now, please follow the instructions for Parts A through I.

PART	TASK	TEST DETAILS
A	Tone Phrases	Please read aloud each word or phrase as requested. 1. 借 jiè 2. 停 tíng 3. 代表 dài biǎo 4. 依靠 yī kào 5. 好天气 hǎo tiān qì 6. 差不多 chà bu duō 7. 性格开朗 xìng gé kāi lǎng 8. 游戏规则 yóu xì guī zé
B	Read Aloud	Please read aloud the sentences as you are instructed. 1. 今天上午小王去公司开会。 Jīn tiān shàng wǔ Xiǎo Wáng qù gōng sī kāi huì. 2. 开会的时候他把手机的声音关了。 Kāi huì de shí hou tā bǎ shǒu jī de shēng yīn guān le. 3. 小李来电话，小王没有听到，差点误了事。 Xiǎo LǏ lái diàn huà，Xiǎo Wáng méi yǒu tīng dào，chà diǎn wù le shì. 4. 小红太喜欢吃巧克力了。 Xiǎo hóng tài xǐ huan chī qiǎo kè lì le. 5. 她总是拿着一大块巧克力，一边吃一边看电视。 Tā zǒng shì ná zhe yí dà kuài qiǎo kè lì，yì biān chī yì biān kàn diàn shì. 6. 妈妈非常担心她的牙齿会越来越不好。 Mā ma fēi cháng dān xīn tā de yá chǐ huì yuè lái yuè bù hǎo.
C	Repeat	Please repeat each sentence that you hear. Example: When you hear: "这个周末我们去看电影吧。" You say: "这个周末我们去看电影吧。"

76 - 63482 - 10

PEARSON

- *试卷 Test paper*
  - *声调词语 Tone phrases*
  - *朗读 Read aloud*
  - *重复 Repeat*
  - *反义词 Opposites*
  - *问答 Questions*
  - *声调识别 – 词语 Recognize tone–word*
  - *声调识别- 句子 Recognize tone-sentence*
  - *组句 Sentence builds*
  - *短文重述 Passage retellings*
- *整体分数 Overall*
  - *句子的掌握 Sentence*
  - *词汇 Vocabulary*
  - *流利度 Fluency*
  - *发音 Pronunciation*
  - *声调 Tone*

一、前期准备 *Preparation*

- 项目职责分配 *Allocation of responsibilities*
  - 北大—PKU
    - 命题
    - 词表研制
    - 组织、实施预测
    - 转写
    - 人工评分
  - 培生集团—PKT
    - 辅助北大的命题、转写和评分工作
    - 建立考试系统
    - 建立语音识别模型
- 双方交流 *Comunication*
  - 每天邮件往来
  - 每周一次视频会议
  - 每月互发月度报告

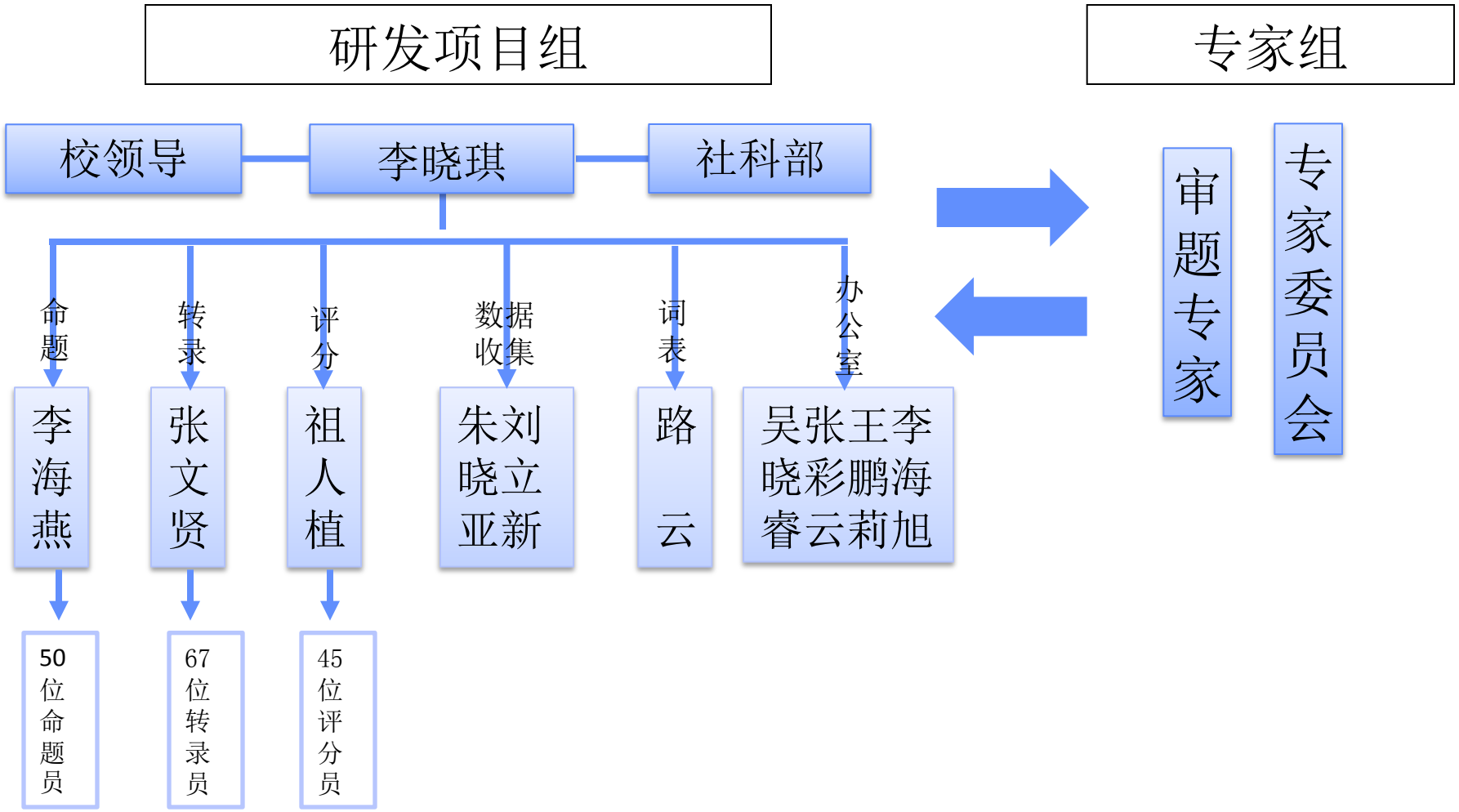
PKU人员组织结构

Organization Structure of PKU Team

李晓琪	• 项目总负责人
李海燕	• 命题总负责人、参与命题
张文贤	• 转写总负责人、参与命题
祖人植	• 评分总负责人、参与命题
路 云	• 负责词表的研制、参与命题
朱晓亚	• 负责数据采集、参与命题
刘立新	• 负责数据采集、参与命题
张彩云	• 参与命题、转写和数据收集
王鹏莉	• 办公室事务性工作，辅助评分工作
吴晓睿	• 负责与美方联络，参与转写和数据收集
李海旭	• 负责管理数据库

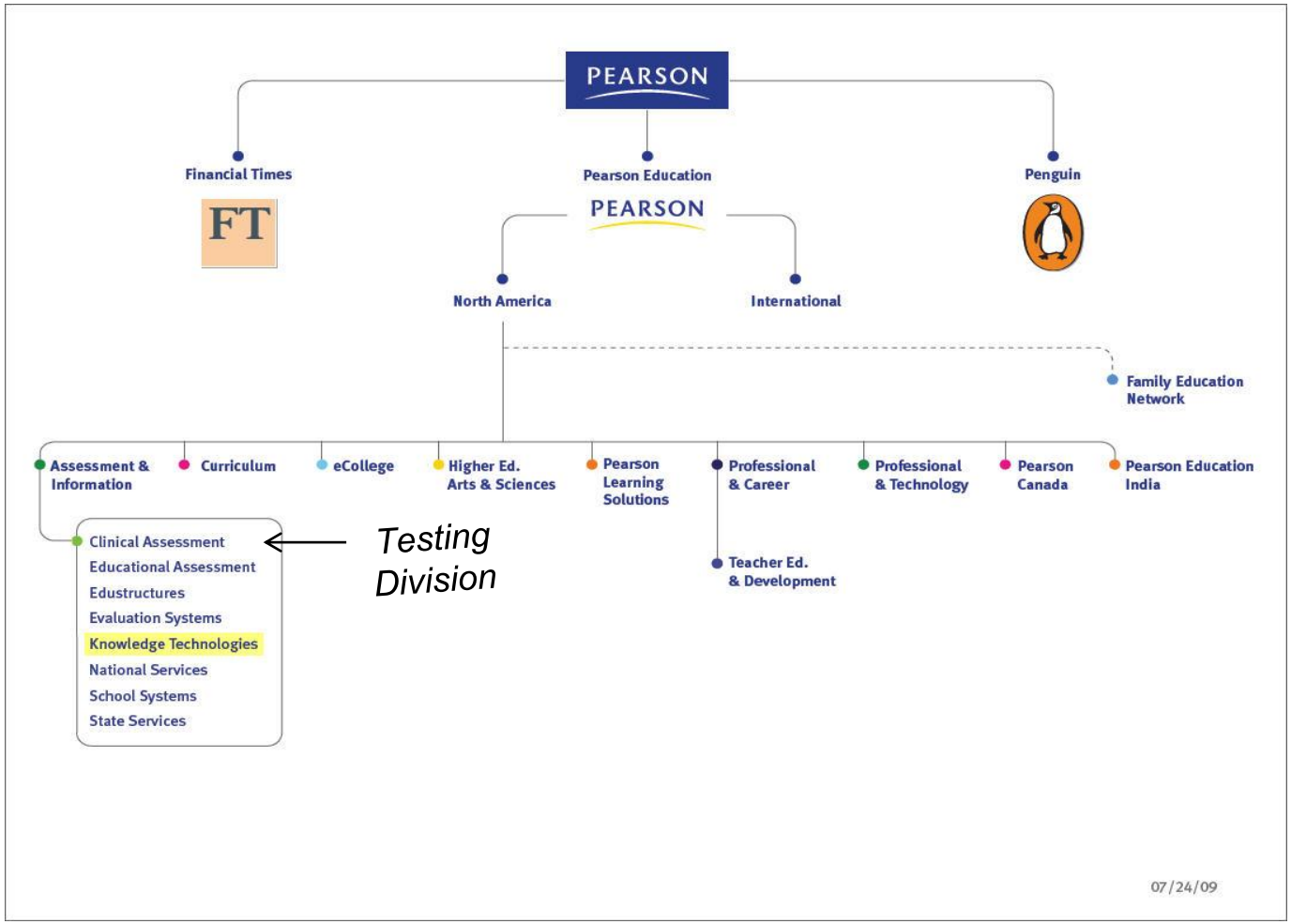
PKU人员组织结构

Organization Structure of PKU Team



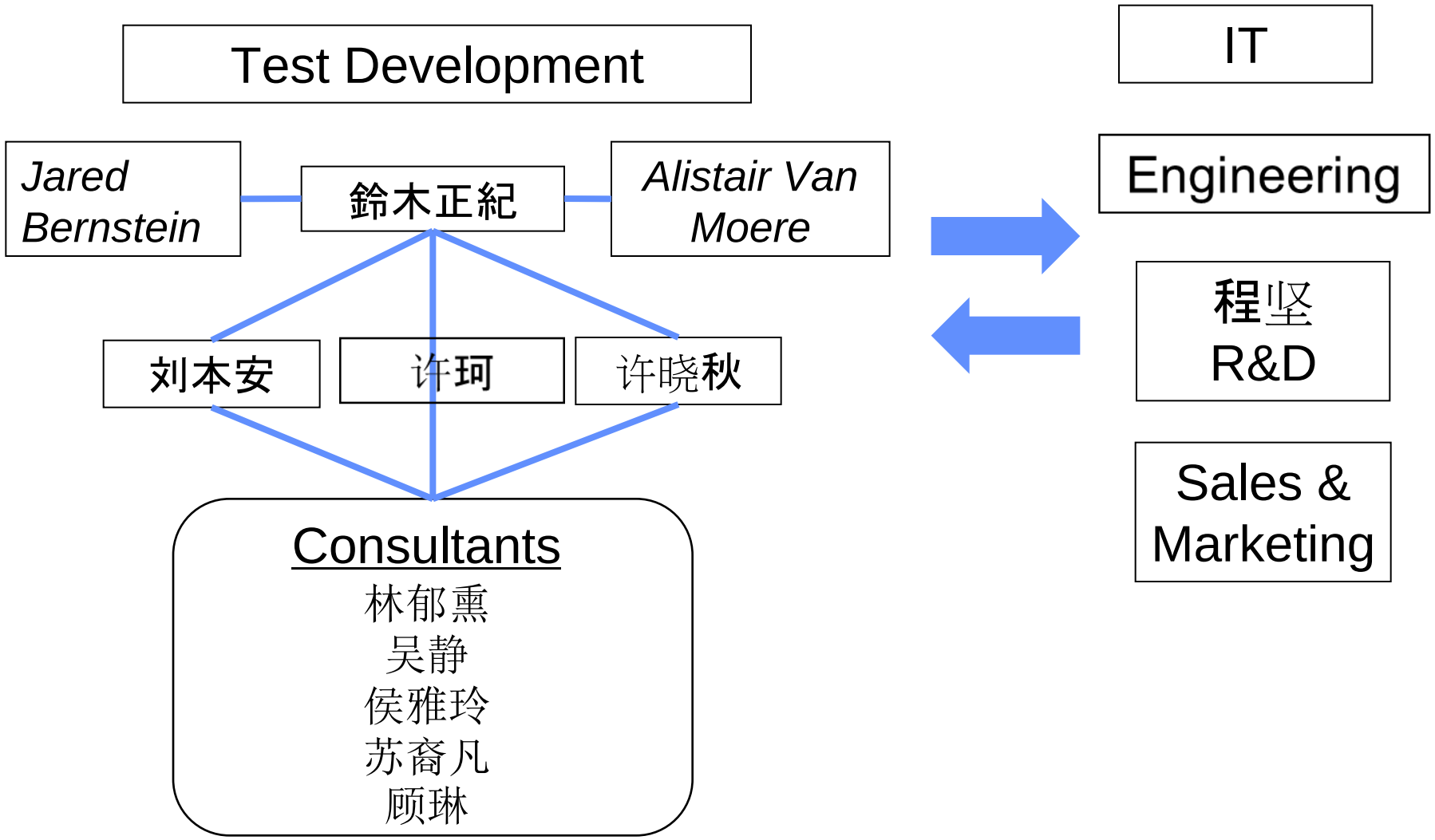
PEARSON组织结构

Organization Structure of Pearson



PKT 人员组织结构

Organization Structure of PKT Team



二、项目研发

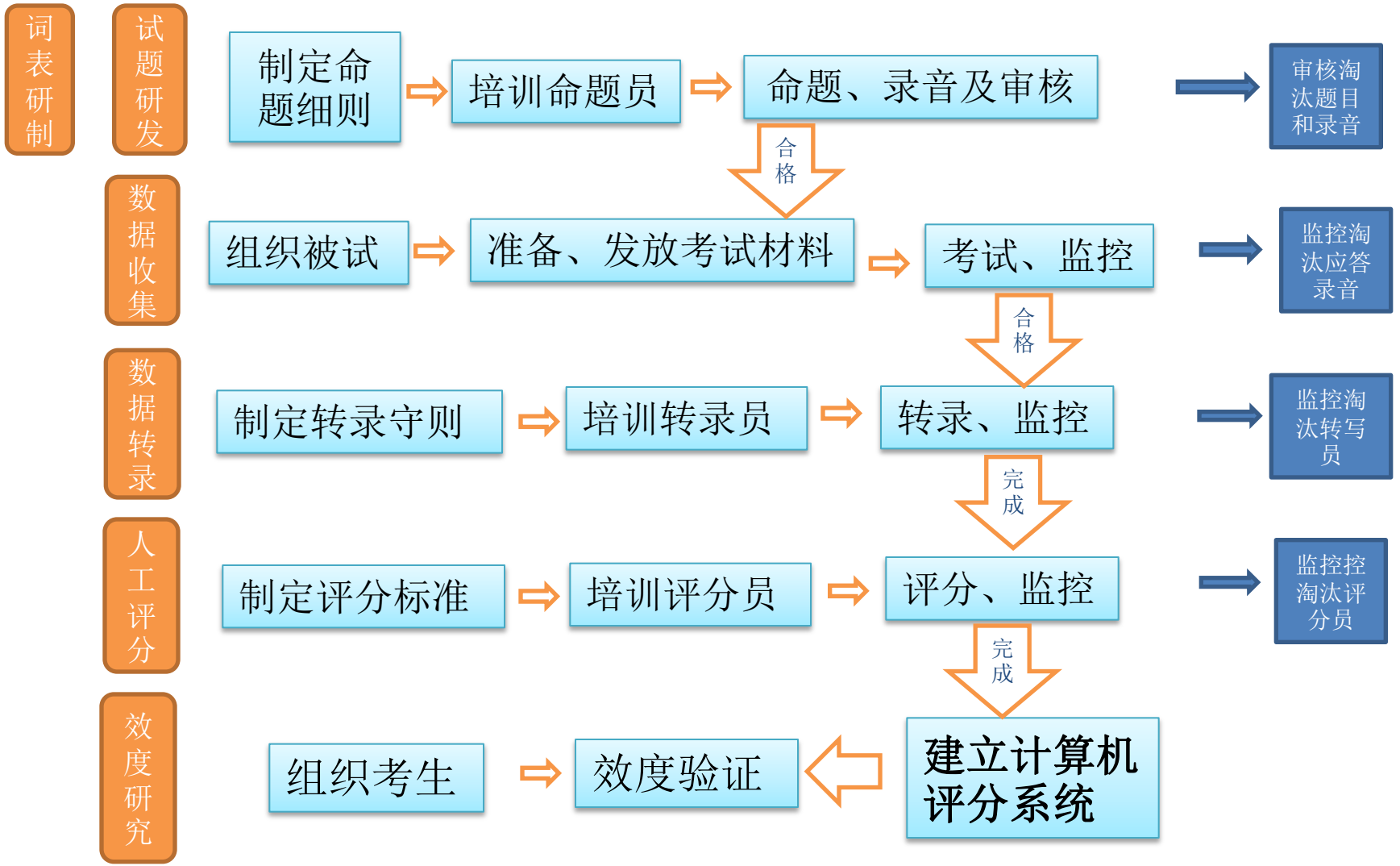
Research and Development of The Project

- 词表研制 *Vocabulary development*
- 试题研发 *Item Development*
- 数据收集 *Data collection*
- 数据转录 *Transcription*
- 人工评分 *Human Rating*
- 计算机评分系统建立

  The establishment of computer score system
- 效度研究 *Validation*

研发流程

Research & Development Process



词表研制
Vocabulary Development

- **研制人*(Teams)*：** PKU&PKT
- **语料来源*(The source of materials)*:**
  1）北语汉语口语语料库（PKU）
  2）16套重要国内口语教材（PKU）
  3）6套重要国外口语教材（PKU 4套，PKT 2套）
  4）汉语频率词典（PKT）
  5）电话录音Callhome（PKT）
- **研制方式*(Development method)*:**
  软件分析处理+专家人工干预
- **词表规模*(Vocabulary size)*：** 5186个

试题研发
Item Development

	题型	A阶段数量	B阶段数量	AB两阶段数量
A	声调词语	104	325	429
B	朗读句子	81	237	318
C	重复句子	455	1665	2120
D	反义词	74	118	192
E	简短问答	256	482	738
F	声调识别-词语	75	117	192
	声调识别-句子	75	111	186
G	组句	124	577	701
H	短文复述	51	242	193
	总计	**1295**	**3874**	**5169**

数据收集
Data Collection

- **目的*(Aim)*：**采集足够多的母语考生和非母语考生的汉语语音样本和应答，以培训和优化自动语音处理系统，并开发汉语口语自动评分模型，从而实现自动评分。
- **实际完成考试次数*(Completed tests)*：**

  A阶段2043次； B阶段3446次；

  AB两阶段共5489次。
- **收集方式*(Collection Method)*：**考生通过打电话或上机进行考试，每次考试约30分钟。为了验证本考试的效度，每位考生参加两次考试，以便查看两次考试分数的差值是否在正常范围内。
- **考生来源*(Sources of test-takers)*：**

  母语考生：来自近20个各方言代表城市与地区；

  非母语考生：1/3来自北大，2/3来自十余所各地大学或其他院校。

数据收集

Data Collection

北京大学 PEKING UNIVERSITY

母语考生取样表

Native test-takers sample distribution

							标准要求									
									年 龄							
							性别		15-24		25-34		35-44		45-59		≥60	
各区比例		方言划分	区	总比例	主要城市	人数	男	女	男	女	男	女	男	女	男	女	男	女
45%	7%	北方官话	东北官话	65%	长春	16	8	8	3	3	2	2	1	1	1	1	1	1
	7%		北京官话		北京	16	8	8	3	3	2	2	1	1	1	1	1	1
	7%		冀鲁官话		济南	16	8	8	3	3	2	2	1	1	1	1	1	1
	7%		胶辽官话		大连	16	8	8	3	3	2	2	1	1	1	1	1	1
	7%		中原官话		郑州	16	8	8	3	3	2	2	1	1	1	1	1	1
	10%	西北官话	兰银官话		兰州	22	11	11	4	4	3	3	2	2	1	1	1	1
20%	10%	西南官话	西南官话		武汉、重庆	22	11	11	4	4	3	3	2	2	1	1	1	1
	10%	南方官话	江淮官话		南京	22	11	11	4	4	3	3	2	2	1	1	1	1
35%	10%	方言	吴方言	35%	上海	22	11	11	4	4	3	3	2	2	1	1	1	1
	3%		赣方言		南昌	8	4	4	1	1	1	1	1	1	1	1		
	3%		湘方言		长沙	8	4	4	1	1	1	1	1	1	1	1		
	8%		粤方言		广州	18	9	9	3	3	3	3	1	1	1	1	1	1
	8%		闽方言		福州、厦门?	18	9	9	3	3	3	3	1	1	1	1	1	1
	3%		客家方言		潮州、汕头	8	4	4	1	1	1	1	1	1	1	1		

数据收集
Data Collection

北京大学 PEKING UNIVERSITY

Immersion≠0 的非母语考生取样表
Sample distribution of non-native test-takers : Immersion≠0

语言背景	总人数	男女分布		各年龄段分布		
		男	女	15~24岁	25~34岁	≥35岁
英语	64	32	32	26	20	18
西班牙语	56	28	28	22	17	17
印地语	24	12	12	10	7	7
阿拉伯语	36	18	18	14	12	10
欧洲语言	36	18	18	14	12	10
东南亚国家语言	50	25	25	20	15	15
孟加拉语	10	5	5	4	3	3
日语	30	15	15	12	9	9
韩语	18	9	9	7	5	6
俄语	20	10	10	8	6	6
非洲国家语言	12	6	6	5	4	3

数据收集
Data Collection

北京大学 PEKING UNIVERSITY

Immersion=0 的非母语考生取样表
Sample distribution of non-native test-takers : Immersion=0

语言背景	总人数	男女分布		各年龄段分布		
		男	女	15~24岁	25~34岁	≥35岁
英语	12	6	6	5	4	4
西班牙语	11	6	5	5	3	3
印地语	5	3	2	2	1	1
阿拉伯语	7	4	3	3	2	2
欧洲语言	7	4	3	3	2	2
东南亚国家语言	10	5	5	4	2	3
孟加拉语	2	1	1	1	3	1
日语	6	3	3	2	1	1
韩语	4	2	2	2	2	1
俄语	4	2	2	2	1	1
非洲国家语言	3	2	1	1	1	1

数据转录
Transcription

- **目的*(Aim)*：**将考生的口头应答以书面的文字形式呈现出来。转写稿要准确地记录说话人所说的，甚至试图所说的内容，同时用特殊的符号记录下考试时外界环境发出的声音，以及考生说话中任何不流利的地方。
- **总工作量*(The amount of work)*：**A阶段转写249,053条，B阶段转写635,743条，共884,796条；B阶段效度验证64,287条。转写全部条数为949,083条。
- **人员培训*(Transcriber training)*：**招募了4批共67名转写员，组织了7次培训，主要讲解转写细则，并针对转写中常见的错误进行重点练习。
- **质量监控*(Quality control)*：**同一应答至少由两名转写员转录，如同一应答的两个转录结果不一样，应答将会交给第三个转写员裁定。此外，还实行两级监控。一级是抽调出4名优秀转写员专门进行质量监控，每周发一次针对每个转写员的转写反馈。二级是项目组设立两位专职监控人员，对一级监控结果进行再次监控。

转录符号与方法

Transcription Symbols and Methods

北京大学 PEKING UNIVERSITY

转录符号	说明	例句
-() or ()-	漏字：一个词未说出的部分要放在括号中，并在括号的前面或后面紧加短划线来注明漏掉的字是词的前面，或是后面。	1.-(学)校 2.上(课)-
=	声音文件中的最后一个词因录音结束被从中切断，其后加“=”。	3.来 中国 以后=
*	声韵错误或声韵及声调错误，加在有错误的词前面。	4.*习惯
%	纯声调错误，加在声调出错的字前面。	5.“骑车”读成“qì chè”：%骑车
@	不可识别的或非汉语普通话的语言，如方言、外语。	6.这 是 @ 的 电脑
uh	考生发出的口头犹豫，如“啊”、“哦”等，都记做“uh”。	7.这个 周末 uh 我们 uh 去 看 电影
#	考生发出的口齿杂音，如舌头敲打声、笑声等。	8.# 要 下雨 了 #
:	音（元音或是辅音）拖长时，将“：”加在被拖长音的字后面。	9.一 个 苹：果
+	用来记录多音节的词内部的停顿现象。	10. 每个 周+末 他 都 和 朋友 一起
[N]	记录背景杂音。	11. 我喜欢 [N] 饺子。 12.[N>] 苹果 [<N]
[S]	记录背景语言。	13. 他 不要 [S] 一起 去
[R]	记录通讯录音杂音。	14.我 不 喜欢 [R] 我 的 新 车
[!]	记录罕见的或是有问题的录音。	
[A]	外国口音，在转录页面的对话窗下的相应选项上打勾。	
[D]	方言口音，在转录页面的对话窗下的相应选项上打勾。	
[U]	作弊等不正常应答，在转录页面的对话窗下的相应选项上打勾。	

转录实例

Examples of The Transcription

北京大学 PEKING UNIVERSITY

例1、4.2

参考答案： 他 决定 下 星期 找 个 时间 去 一 趟 大使馆

转录稿： 他 决+定 %下 星期 *找 个 时间 去 一 %趟 %大%使馆 [N] （R）（A）

例2、3.8

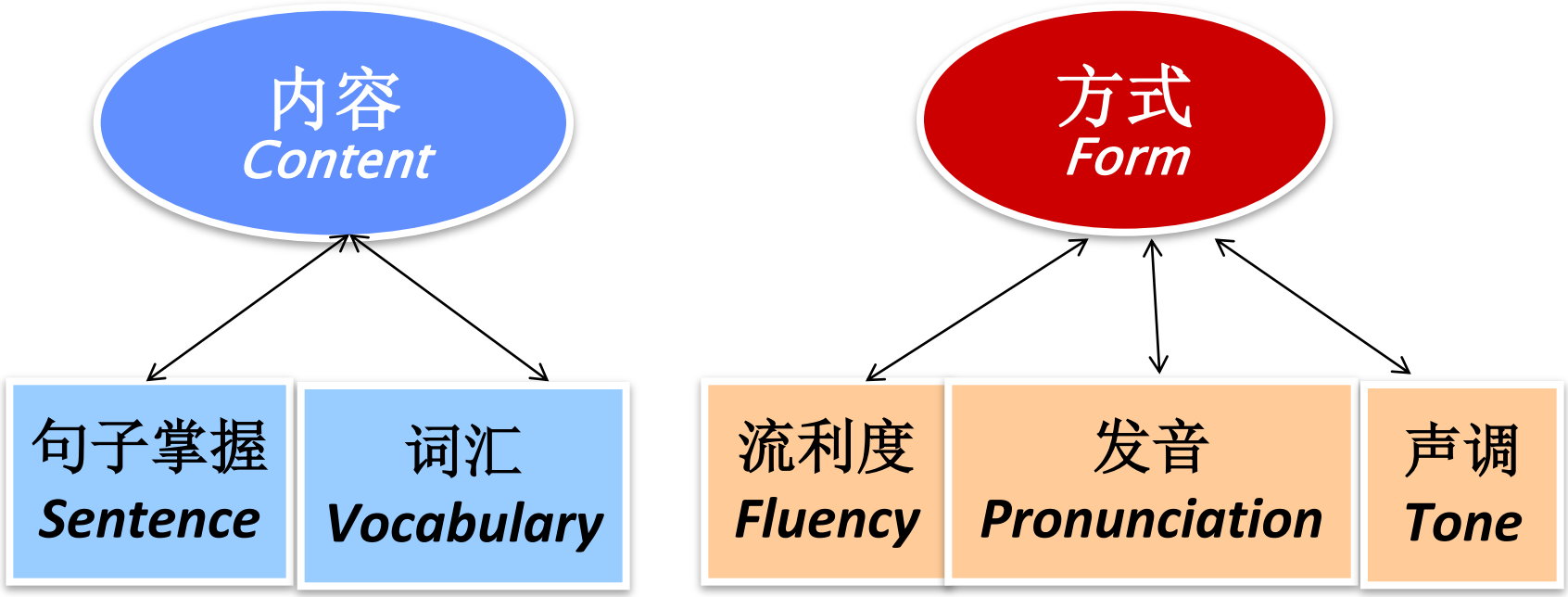
参考答案： 他 每天 从 早 忙到 晚 因为 要 学 的 东西 太 多 了

转录稿： [R] [S] 他 每天 [N] %从 早 # %忙 到 晚 [N>] %因为 [<N] 要 *学 的 东西 [N>] [S] [<N] 太 多 了 [N] [S] [N] （A）

人工评分
Human Rating·

- **目的*(Aim)*：**产生足够数量的人工评分，并用来开发和验证自动评分模型。
- **任务*(Task)*：**利用培生的评分界面，根据一系列的评分标准，聆听每个应答，并给予评分。所有进入考试题库的试题都经过人工评分。
- **评分员*(Raters)*：**共培训评分员40名，分为对比组、声韵组、声调组、流利度组展开工作。

评 分
Rating

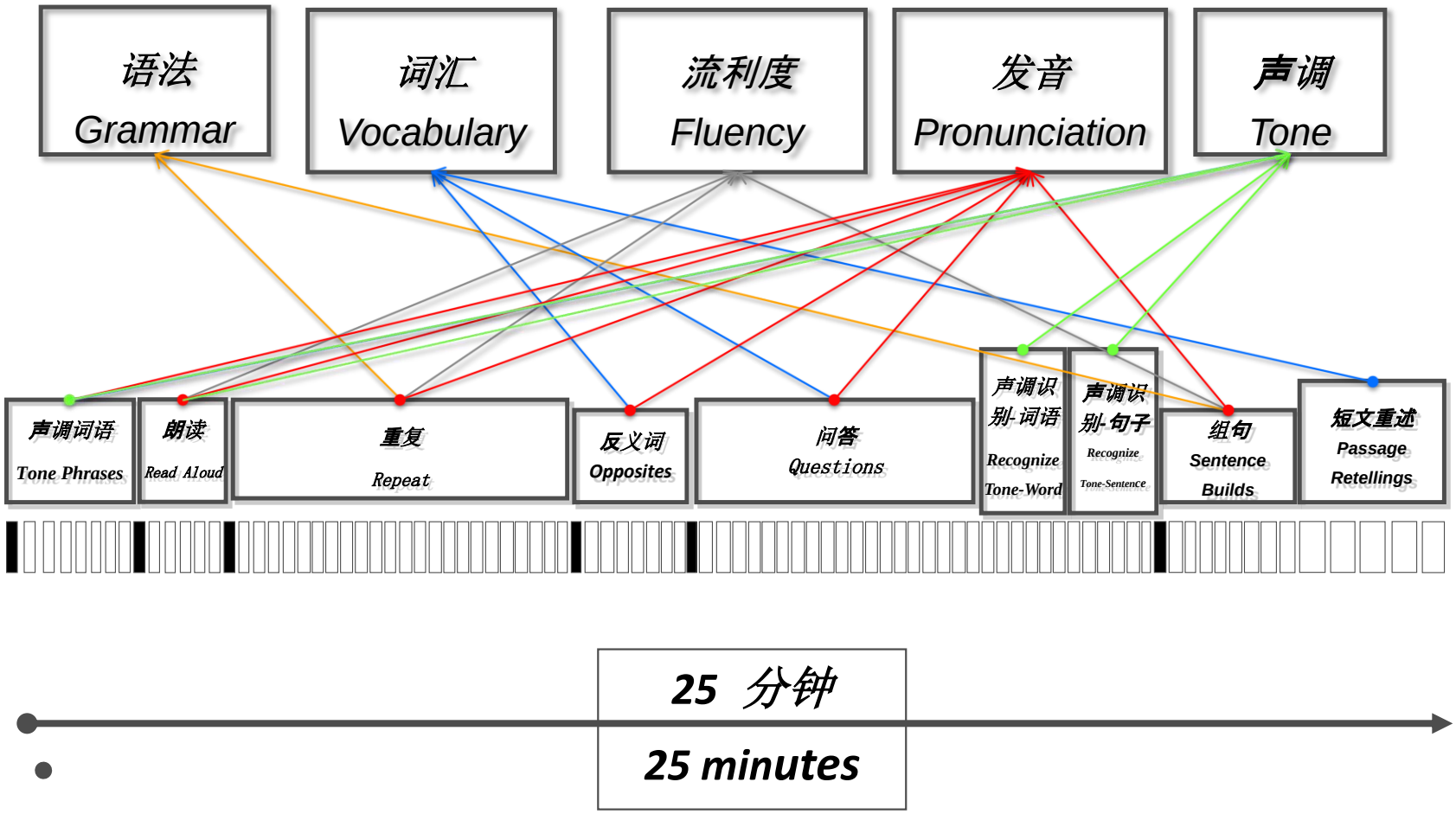

内容
Content
方式
Form
句子掌握
Sentence
词汇
Vocabulary
流利度
Fluency
发音
Pronunciation
声调
Tone

计算机评分系统建立

The Establishment of Computer Score System

北京大学 PEKING UNIVERSITY

试用版考试评分逻辑 *Preliminary SCT score logic*



评分权重研究

Rating Scale Research

- 制定评分标准*(Establish grading standard)*

级别	分值	总体水平描述	语法	词汇	流利度	发音	声调
主要题型			重复句子、组句	反义词、问答、短文复述	朗读、重复句子、组句	声调词语、朗读、重复句子、反义词、问答、组句	声调词语、声调识别
最高级	10	回答了所有试题；内容正确； 语音自然，表达流畅，符合母语语感； 相当于母语者水平。	结构、语序、虚词正确。	词汇丰富，表达清楚，完整，准确。	非常流畅，自然。	自然，清楚。没有外国口音。	自然，调型、调值准确。
高级	9	回答了大部分试题，只有个别遗漏； 内容基本正确，有个别不完整，不准确之处； 语音比较自然，表达比较流畅，很少有影响交际的发音错误； 接近母语者水平。	结构、语序、虚词基本正确，有个别遗漏和错误，但不妨碍意义理解。	词汇比较丰富，词语使用基本正确，偶尔不够准确，完整。	能流利表达，不恰当的停顿、重复很少。	比较清楚，有个别的音有错误，但基本不妨碍理解。有一些外国口音。	调型、调值基本正确，偶尔表现出外国人特征，但不妨碍理解。
	8						
	7						
中级	6	大约只回答了一半试题； 内容正确与错误参半，很多时候不完整不准确； 有些洋腔洋调，表达不很流畅，有些不恰当停顿，不常接触外国人的普通中国人听起来有些吃力；	结构、语序、虚词有一些明显的错误和遗漏，有时候影响意义的理解。	有一定的词汇量，词语使用大部分正确，但有时有明显错误，影响意义理解。	多数语句说得比较流利，但表达复杂内容时，流利度不足，有一些不恰当的停顿或明显的迟疑。	有些明显的发音错误，有时候影响意义理解。 有较明显的外国口音。	大部分声调正确，但不自然不准确的情况较明显。带有一些规律性的外国人特征。
	5						
	4						
初级	3	只回答了少部分试题； 内容大部分不完整不准确； 洋腔洋调比较严重，表达不流利，不恰当停顿很多，听起来比较吃力，常常听不懂。	结构、语序、虚词有比较多的明显的错误和遗漏，常常影响意义的理解。	词汇量有限，词语使用常常有明显错误，难以听懂。	不流利，不恰当的语音停顿比较多，而且停顿时间比较长。	发音错误多且明显，有时候难以听懂。 外国口音比较严重。	缺乏正确的声调变化，有时候难以听懂。 听起来很生硬、吃力。
	2						
	1						
	0	基本上没有回答试题。					

评分权重研究
Rating Scale Research

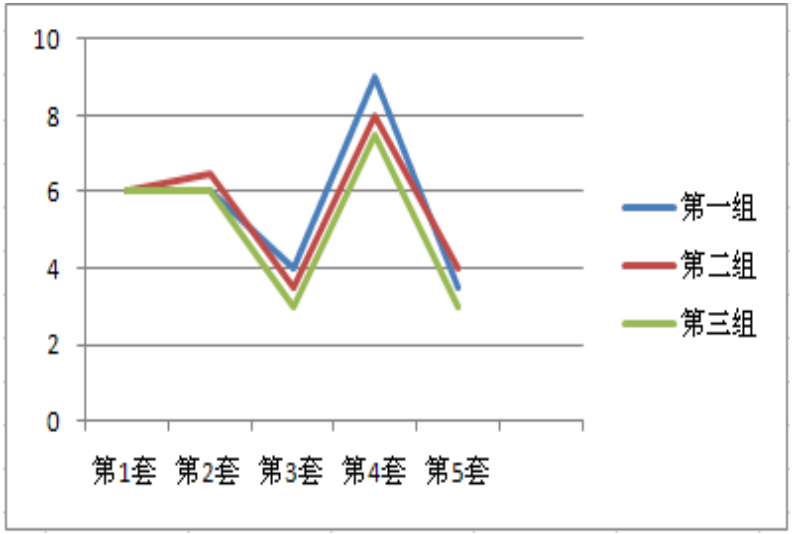
- **评分员信度验证*(Inter-rater reliability verification)***

  评分员*(Raters)*： 9人，三组

  评分差*(Score variance)*： 各小组之间分差基本在0.5-1分之间。（见曲线图）

  相关系数*(Correlation)*： 三个评分组间的相关系数均达0.94以上。（见表格）

  结论*(Conclusion)*： 各小组的评分信度很高，一致性很好，评分可靠。



	第1组	第2组	第3组
第1组	1		
第2组	0.96	1	
第3组	0.94	0.98	1

评分权重研究
Rating Scale Research

- 评分员评分与PKT计算机系统评分相比较

  Human rating score vs. PKT computer rating score

样本数量***(Sample size)***：115个（+210个）

权重分布示例***(One example of various score weight distributions)***

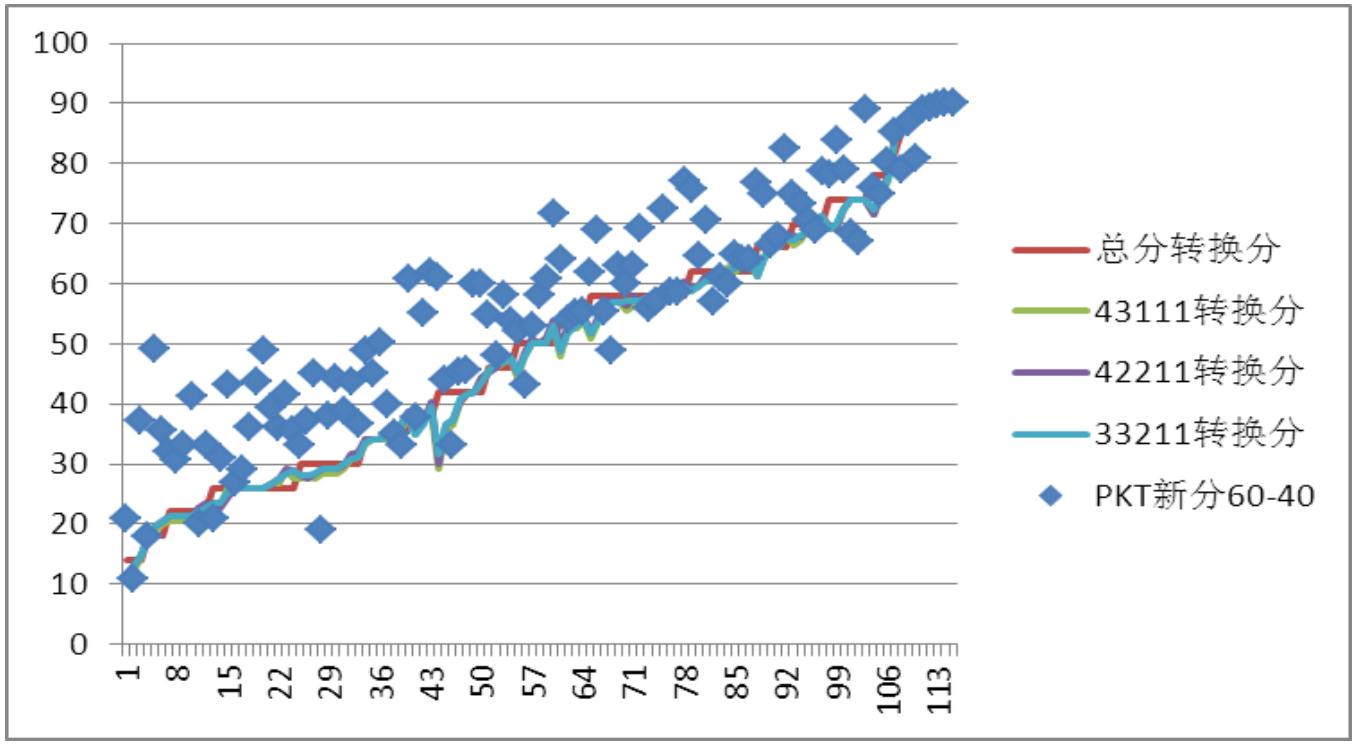
考号	总分	语法	词汇	流利度	发音	声调	43111 总分	42211 总分	33211 总分	总 分
3474748	3	3	2	4	4	4	30	32	31	30

评分权重研究

Rating Scale Research

- 结论*(Conclusion)*：

PKT分数与PKU的评分转换分的相关系数为0.92。说明PKT评分模型的区分度较好，能够较好地区分出不同水平的考生，只是整体分数偏高。



评分权重研究
Rating Scale Research

- 建议*(Suggests)*：

1.关于各分项分数在总分中所占的权重问题；

2.关于对未答题的减分比率问题；

3.关于高分段区分度问题。

效度研究
Validation

- **研究内容*(Content)*:**

  1）考试信度研究；

  2）自动评分和人工评分的相关度研究；

  3）汉语口语考试和其他同期口语考试对比研究。

- **研究对象*(Objects)*:**

  37 个研究对象 （21 女, 16 男）

  21 名在中国, 16 名在美国

  不参与评分系统建立的数据收集

考试信度研究
Test Reliability

Score	分半信度 Split-half Reliability (N=37)	重测信度 Test –Retest Reliability (N=31)
总分 Overall	**0.98**	**0.95**
语法 Grammar	0.97	0.96
词汇 Vocabulary	0.94	0.87
流利度 Fluency	0.96	0.93
发音 Pronunciation	0.91	0.86
声调 Tone	0.87	0.74

其他语言(Other languages)：0.94--0.97

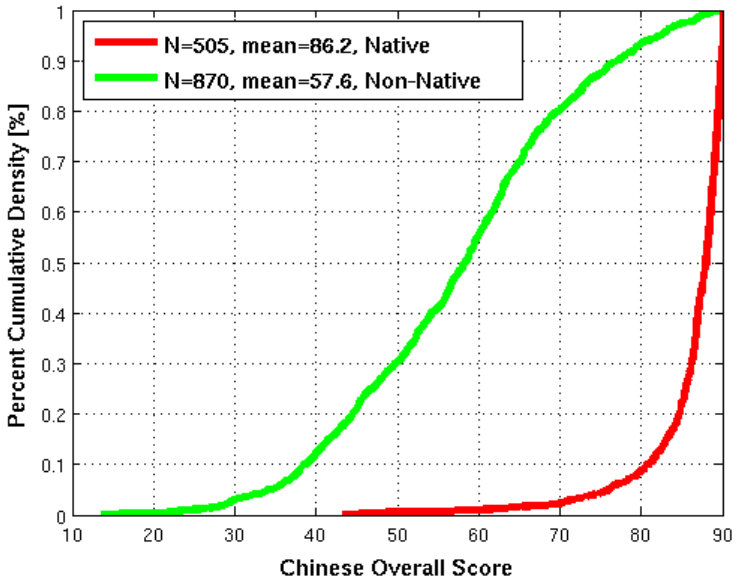
自动评分——人工评分的相关度

Machine – Human Correlation

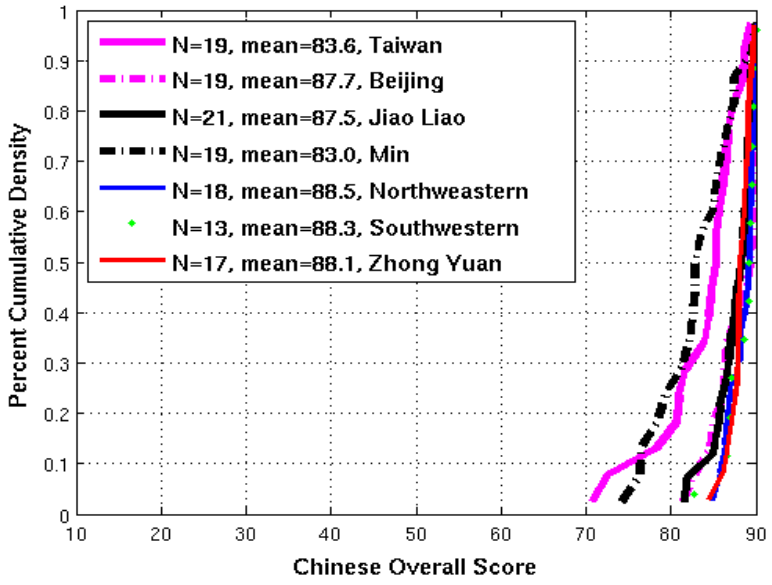
Score	Correlation
总分 Overall	0.98
语法 Grammar	0.99
词汇 Vocabulary	0.94
流利度 Fluency	0.89
发音 Pronunciation	0.90
声调 Tone	0.90

其他语言(Other languages)：0.93--0.98

母语与非母语对比*(Native vs. Non-native)*



母语内部相关度*(Native Dialect Groups)*



与其他同期口语考试的对比
Relation to Concurrent Tests

- 对比对象*(Objects)*：

  ILR Oral Proficiency Interview

  HSK Oral Test

- 对比结果*(Conclusions)*：

	相关度 Relation	***样本人数 Sample size***
vs. ILR-OPI	0.79	37
vs. HSK 中级口语考试	0.86	23

其他语言(Other languages): 0.81 - 0.92

北京大学
PEKING UNIVERSITY

效度研究结论
The Conclusion of Validation

- 试用版考试的研究得到了令人满意的结果。
- Versant 方法同样适用于汉语口语考试。

三、研发思考

Sevral Thoughts

- 汉语口语考试的目的是什么？汉语口语考试的分数应该代表了什么？
- 如何定义并限定考题中可以接受的汉语口语范围？

名 称	非标准形式
平舌音和翘舌音的合并	{zh,ch,sh} 和 {z,c,s}的合并
/n/ 变成 /l/	声母 /n/ 被发成 /l/; 例如：南 nán = 蓝 lán
/r/ 变成 /l/	声母 /r/被发成 /r/; 例如： 热 rè = 乐 lè
鼻音韵母的变化	前鼻音和后鼻音的合并 {n, ng} ； 例如：应yīng = 音yīn
儿化现象	在词尾用儿化; 例如： 花 读成 花儿.
轻声缺失	轻声的缺失; 父亲 fùqin 读成 父亲 fùqīn
声调的变化	声调的自由转换;例如：卧室 wò shì 读成 卧室 wò shǐ

- 在样本数据采集中，如何判断一个被试的样本为可接受的“母语者”？
- 如何为汉语的发音评分？

北京大学
PEKING UNIVERSITY

谢 谢！

Thank you!