


The Second International Conference on Language Testing and Assessment

第二届语言测试与评价国际研讨会



主办单位 (Hosted by):

浙江大学 Zhejiang University

教育部考试中心 National Education Examinations Authority

协办单位 (Co-organized by):

高等教育出版社 Higher Education Press

外语教学与研究出版社 Foreign Language Teaching and Research Press

《中国考试》杂志社 China Examinations

赞助单位 (Sponsored by):

英国文化委员会 British Council

北京外研在线教育科技有限公司 Beijing Waiyan Online Education Technology Co. Ltd.

科大讯飞股份有限公司 IFLYTEK Co. Ltd.

流利说 Shanghai Liulishuo Information Technology Ltd.

批改网 (www.pigai.org) Pigai (en.pigai.org)

深圳市海云天科技股份有限公司 Shenzhen Seaskyland Technology Co. Ltd.

Conference Schedule

Friday, Nov.11th 11月11日, 周五	
14:00-21:00	Registration 报到注册 (Zijingang Hotel 紫金港大酒店)
17:00-19:00	Conference Dinner 晚餐 (Buffet at Zijingang International Hotel 紫金港国际饭店自助餐)
Saturday, Nov. 12th 11月12日, 周六	
08:30-08:45	Opening Ceremony 开幕式 (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
08:45-09:05	Invited Speech 特邀讲话 (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
09:05-09:50	Keynote Speech (1) : Prof. Liu Jianda and Prof. He Lianzhen (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
09:50-10:10	Group Photo 集体照
10:10-10:30	Tea Break 茶歇
10:30-11:15	Keynote Speech (2) : Dr. Nick Saville (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
11:15-12:00	Keynote Speech (3) : Prof. Alister Cumming (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
12:00-14:00	Lunch 午餐 (Linhu Dining Hall 临湖餐厅套餐)
14:00-17:20	Parallel Sessions 分会场 (Teaching Building East No.6 东六教学楼)
18:00-20:00	Conference Dinner 晚餐 (Zijingang International Hotel 紫金港国际饭店桌餐)
Sunday, Nov. 13th 11月13日, 周日	
08:30-11:50	Parallel Sessions 分会场 (Teaching Building East No.6 东六教学楼)
12:00-14:00	Lunch 午餐 (Linhu Dining Hall 临湖餐厅套餐)
14:00-14:45	Keynote Speech (4) : Dr. Xi Xiaoming (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
14:45-15:05	Tea Break 茶歇
15:05-15:50	Keynote Speech (5): Dr. Vivien Berry (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
15:50-16:10	Closing Ceremony 闭幕式 (139 Lecture Theatre, Mengminwei Building 蒙民伟楼 139)
18:00-20:00	Conference Dinner 晚餐 (Junting Restaurant 杭州君庭大酒店桌餐)

Contents

Conference Program	1
Abstracts of Keynote Speeches	12
Abstracts of Parallel Sessions	18
► Saturday Sessions.....	18
Room 101.....	18
Room 102.....	21
Room 301.....	26
Room 303.....	34
Room 304.....	37
Room 305.....	41
Room 306.....	45
Room 307.....	50
Room 309.....	55
Room 310.....	61
► Sunday Sessions.....	65
Room 101.....	65
Room 102.....	70
Room 301.....	76
Room 303.....	81
Room 304.....	84
Room 305.....	88
Room 306.....	94
Room 307.....	99
Room 309.....	102
Room 310.....	105

Conference Program

Friday, Nov. 11th, 2016 2016 年 11 月 11 日, 周五		
14:00-21:00	Registration (报到注册)	Zijingang Hotel (紫金港大酒店)
17:00-19:00	Conference Dinner (晚餐)	Buffet at Zijingang International Hotel (紫金港国际饭店自助餐)

Saturday, Nov. 12th, 2016 2016 年 11 月 12 日，周六			
08:30-08:45	Opening Ceremony Chair: Prof. He Lianzhen 开幕式 主持人：何莲珍教授	Opening Address by Prof. Wu Zhaohui, President of Zhejing University 浙江大学校长吴朝晖教授致辞	139 Lecture Theatre, Mengminwei Building (蒙民伟楼 139)
8:45-9:05	Invited Speech Chair: Prof. He Lianzhen 特邀讲话 主持人：何莲珍教授	Invited Speech by Dr. Lin Huiqing, Vice Minister of Education 教育部副部长林蕙青讲话	
09:05-9:50	Keynote Speech (1) Chair: Prof. Zou Shen (主旨发言 1) 主持人：邹申教授	Prof. Liu Jianda & Prof. He Lianzhen China's Standards of English and Their Applications 刘建达教授 & 何莲珍教授 从量表到考试	
9:50-10:10	Group Photo (集体照)		
10:10-10:30	Tea Break (茶歇)		
10:30-11:15	Keynote Speech (2) Chair: Prof. Zeng Yongqiang (主旨发言2) 主持人：曾用强教授	Dr. Nick Saville Developing and Validating Test Materials Within a Common Framework of Reference 在《欧洲语言共同参考框架》内开发验证 语言测试	139 Lecture Theatre, Mengminwei Building (蒙民伟楼 139)
11:15-12:00	Keynote Speech (3) Chair: Prof. Han Baocheng (主旨发言3) 主持人：韩宝成教授	Prof. Alister Cumming Connecting Writing Assessments to Teaching and Learning: Distinguishing Alternative Purposes 衔接写作测试与教学：明确测试的目的	
12:00-14:00	Lunch (Linhu Dining Hall) 午餐（临湖餐厅套餐）		
14:00-17:20	Parallel Sessions (Teaching Building East No.6) 分会场（东六教学楼）		
18:00-19:30	Conference Dinner (Zijingang International Hotel) 晚餐：紫金港国际饭店桌餐		



Sunday, Nov 13th, 2016 2016 年 11 月 13 日，周日			
08:30-11:50	Parallel Sessions (Teaching Building East No.6) 分会场 (东六教学楼)		
12:00-14:00	Lunch (Linhu Dining Hall) 午餐 (临湖餐厅套餐)		
14:00-14:45	Keynote Speech (4) Chair: Prof. Wu Zunmin (主旨发言 4) 主持人：武尊民教授	Dr. Xi Xiaoming Transforming Language Learning and Assessment Experience with Technology: Outlook and Challenges 运用技术手段改善语言学习与测试体验：展望 与挑战	139 Lecture Theatre, Mengminwei Building (蒙民伟楼139)
14:45-15:05	Tea Break (茶歇)		
15:05-15:50	Keynote Speech (5) Chair: Prof. Zhang Wenxia (主旨发言 5) 主持人：张文霞教授	Dr. Vivien Berry Innovation in Foreign Language Testing in China: Researching a Face-to-face and Video- conferencing Delivered Speaking Test 外语考试在中国的创新：面对面口语测试与视 频会议口语测试对比研究	139 Lecture Theatre, Mengminwei Building (蒙民伟楼139)
15:50-16:10	Closing Ceremony (闭幕式)	Prof. Liang Junying 梁君英教授	
18:00-19:30	Conference Dinner (Junting Restaurant) 晚餐：杭州君庭大酒店桌餐		

Saturday Sessions

Room 101 (101 教室)

THEME (主题)

Innovation in Foreign Language Testing in China (中国外语考试改革研究)

Chair: 于涵

14:00-14:40	教育部考试中心 乔辉 章建石 高考外语科一年两考改革研究与实践 高考外语科一年两考改革如何科学落地? ——考试评价视角下的技术保障
14:40-15:10	广东省教育考试院 黄友文 广东省英语听说考试的改革探索
15:10-15:40	上海市教育考试院 褚劲风 语言测试与评价视阈下质量标准相关问题研究
15:40-16:00	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
16:00-16:50	北京教育考试院 臧铁军 徐加永 关于北京市中考英语改革的几点思考
16:50-17:20	清华大学 张浩 对于高考英语一年两考改革的态度调查——一项基于外语能力测评现状及需求分析的研究

Room 102 (102 教室)**Symposium 1**

Discussant: 张文霞

Symposium 2

Discussant: 邹申

14:00-15:30	张文霞 张浩 魏兴 程蒙蒙 吴莎 颜奕 中国外语测试改革现状与需求的调查 Survey of Status Quo and Reform Needs of Foreign Language Testing in China
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-17:20	Pan Mingwei, Zou Shen, Chen Jianlin, Deng Jie, Li Qinghua, Zhang Wenxing 迈向统一的《中国英语写作能力等级量表》: 原因、方法与指向 Towards a Unified Writing Proficiency Scale of China's Standards of English (CSE): Why, How and Where

Room 301 (301 教室)**THEME (主题)**

Language Proficiency Scales (外语能力量表研究)

Chair: 冯莉

14:00-14:30	Wu Zunmin, Luo Shaoqian, Lin Dunlai, Qian Xiaofang, Xu Yun, Liu Liping, Gao Miao, Yang Lvna, Zhao Haiyong, Jia Yidong 中国英语能力等级量表研究——以语言组织能力为例 Working on the China's Standards of English (CSE)—Organizational Competence Specified
14:30-15:00	冯莉 论语言能力的描述 Description of Language Proficiency
15:00-15:30	揭薇 中国英语口语能力量表的临界值研究——IRT 理论的应用 Research on the Cut-off Point of China's Standards of English-Speaking Scale (CSES)—The Adoption of IRT
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	王隽 刘畅 中国英语能力等级量表建设——英语口语典型活动调查 China's Standards of English (CSE): Typicality Investigation of English Speaking Activities
16:20-16:50	周建华 中国大学生英语写作元认知策略能力等级量表的构建研究
16:50-17:20	彭川 链接中国英语能力等级量表与《欧洲语言共同参考框架》 Aligning China's Standards of English (CSE) with the CEFR

**Room 303 (303 教室)****THEME (主题)**

Framework for Language Testing and Assessment (外语能力测评体系建设研究)

Chair: 王巍巍

14:00-14:30	袁靖 关于中国大学生外语能力评价的思考
14:30-15:00	吕生禄 Strategic Vision and Approaches to Constructing the Foreign Language Ability Assessment System in China
15:00-15:30	郑群 李悦 中国英语学习者会话含意理解能力测评
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	王巍巍 口译教学中的质量测评参数 Quality Parameters in Interpretation Teaching
16:20-16:50	刘森 基于国家英语能力等级量表的考试说明制订研究

Room 304 (304 教室)**THEME (主题)**

Innovation in Foreign Language Testing in China (中国外语考试改革研究)

Chair: 刘宝权

14:00-14:30	刘婧 大学英语 A、B 级考试对高职英语教学反拨效应的研究
14:30-15:00	刘宝权 范劲松 侯艳萍 高风险测试改革教师态度研究——以英语专业八级考试为例
15:00-15:30	张放 大学英语四级考试新闻听力测试的反拨效应研究——聚焦考生
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	杨志红 Aptitude for Consecutive Interpreting and Its Testing
16:20-16:50	韩潮 Searching for an Optimal Design for a Bi-directional English/Chinese Interpreting Test: Using Pooled Variance Components from Multiple Generalizability Studies

Room 305 (305 教室)

THEME (主题)

Formative Assessment in Foreign Language Teaching (外语教学中的形成性评价研究)

Chair: 张荔

14:00-14:30	黄静 陈文存 Helping Learners Take Control of Their Own Writing Within the Framework of Assessment for Learning Framework
14:30-15:00	Lance Knowles Measuring the Learning Process: Theory and Practices
15:00-15:30	李少兰 詹全旺 Formative Evaluation of College English Based on Output-Driven Input-Enabled Hypothesis
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	张荔 陈德凤 A Study of Formative Assessment for Academic English Writing of Chinese EFL Learners
16:20-16:50	徐国柱 On the Multi-feedback Assessment of College English Writing

Room 306 (306 教室)

THEME (主题)

Rating Scale (评分量表)

Chair: 吴雪峰

14:00-14:30	邹绍艳 张晓艺 关晓仙 大学英语四级写作评分标准探析——评分员的视角 Exploring the Rating Criteria of CET-4 Writing from the Raters' Perspective
14:30-15:00	吴雪峰 英语写作测试评分标准模型的建构及其效度研究——以概要写作评分标准为例 The Construction and Validation of a Model for English Writing Rating Scale—Take Summary Writing as an Example
15:00-15:30	纪小凌 A Comparative Study of Holistic Scoring and Analytic Scoring in Writing-Scoring Reliability, Teachers' and Students' Perceptions
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	杨志强 全冬 A Longitudinal Study of PRETCO-Oral Rater Effects
16:20-16:50	Ellen Head Rating Fluency: An Inquiry into How Far Perceptions of Fluency Align with Quantifiable Measures

**Room 307 (307 教室)**

THEME (主题)

Prompt Characteristics (提示特征)

Chair: 赵冠芳

14:00-14:30	赵冠芳 吕云鹤 刘子仪 英语专业本科生对 EFL 学术写作的构念认知困难解读 The Construct of EFL Academic Writing Ability and Students' Difficulties with Academic Writing: From the Perspective of Chinese Undergraduates of English
14:30-15:00	葛晓华 An Empirical study on the Effect of Writing Prompts on Tertiary Students' Writing Process and Performance
15:00-15:30	孙悠夏 An Investigation of Rater Bias Patterns in a Large-scale EFL Writing Assessment
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	何琼 A Comparative Study on the Use of Pictures in Designing Picture-based Writing Test Tasks
16:20-16:50	施雅俐 Linguistic Features of Picture-prompted Writing-Differences by Three Caption Types
16:50-17:20	吕洲洋 The Influence of Interlocutor Proficiency in a Paired Oral Test

Room 309 (309 教室)

THEME (主题)

Application of Artificial Intelligence in Language Testing (人工智能技术在语言测试中的应用)

Chair: 辜向东

14:00-14:30	汪张龙 人工智能技术及其在外语测试领域中的应用
14:30-15:00	刘洋 纸笔到计算机化考试转变过程中数据库与计算机技术的应用
15:00-15:30	Lance Knowles Innovations in Online Language Testing: Adaptivity and Speech Recognition
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	杨宏波 虞程远 辜向东 BEC 中级口语考试考官和考生问卷调查对比分析
16:20-16:50	钟瑜 辜向东 肖巍 剑桥商务英语考试的反拨效应机制——基于结构方程建模的研究及其启示
16:50-17:20	游忠惠 陈光斌 One Billion Customers: Lessons on Application of Artificial Intelligence from the Front Lines of Language Testing in China

Room 310 (310 教室)

THEME (主题)

ESP (特殊用途英语)

Chair: 张聪

14:00-14:30	葛诗利 贾清 基于语料库的商务术语使用与商务写作质量的相关性研究
14:30-15:00	高霄 高媛 学术英语阅读测试的信效度检验
15:00-15:30	方秀才 信息化时代专门用途英语测试与效验框架研究
15:30-15:50	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
15:50-16:20	周淑莉 基于语言能力发展的 ESP 课程测试研究
16:20-16:50	张聪 Multi-modal Analysis of Interviewers' Pragmatic Identity in English Oral Testing

Sunday Sessions

Room 101 (101 教室)

THEME (主题)

MHK (中国少数民族汉语水平等级考试)

Chair: Zhiming Yang

08:30-09:00	凌纾宇 陈辉 Using Rasch Modeling to Analyze MHK Examinees Data of Different Years
09:00-09:30	Zhiming Yang Practical Considerations for Developing Item Banks in China
09:30-10:00	王婧 魏立艳 Automated Scoring of Retelling Proficiency in the Oral Test of MHK (the Fourth Level)-Practicability and Reliability Study
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	王妍 Research on the Automatic Scoring Validity of Subjective Question in MHK (level III)
10:50-11:20	张健 周成林 任杰 洪润 Criterion Evidence of Toulmin's Argument Model for MHK Level 3 Oral Test
11:20-11:50	郭明明 The Research of Chinese Proficiency and Evaluation



Room 102 (102 教室)

Symposium 3

Discussant: 何莲珍

Symposium 4

Discussant: 武尊民

08:30-10:00	何莲珍 何佳文 闵尚超 陈大建 赵亮 张洁 《中国英语能力等级量表》之听力子量表的构建 Development of the Listening Proficiency Subscale of China's Standards of English (CSE)
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼118教室
10:20-11:50	武尊民 柳丽萍 何晓阳 张春青 李久亮 周红 董连忠 高中生英语成长诊断系统的建设及相关研究

Room 301 (301 教室)

THEME (主题)

Language Proficiency Scales (外语能力量表研究)

Chair: 彭康洲

08:30-09:00	严明 《中国英语能力等级量表》研究项目下的笔译能力量表研发报告: 构念、方法、过程与进展 Development Report of the Translating Scale of China's Standards of English (CSE): Construct, Method, Process and Progress
09:00-09:30	彭康洲 彭之尧 基于语料库的高级英语学习者听力能力建构研究 Exploring the Construct of Listening Competence of Advanced English Learners: An Corpus-based Study
09:30-10:00	周艳琼 中国大专院校EFL学习者英语阅读策略量表的开发与效验 Developing and Validating a Reading Strategy Scale for Chinese Tertiary EFL Learners
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼118教室
10:20-10:50	王海萍 Toward a Framework of CEFR-referenced Legal English Proficiency Scales
10:50-11:20	王爽 王佑旻 A Study on the Descriptor Database of the Reading Ability of Undergraduate Foreign Students in China in Preparatory Education

Room 303 (303 教室)

THEME (主题)

Validity (效度研究)

Chair: 孔祥

08:30-09:00	孔祥 张玄 Prediction Research of Item Difficulty of Verbal Comprehension and Expression
09:00-09:30	Philip Horne 'The Rubber Ruler.' Using Proficiency Scales as an Accurate Measurement of Classroom Achievement
09:30-10:00	刘蓓蓓 赵琪凤 A Study on the Language Ability of the Employees in the Window Industry in China
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	高淑玲 Ethics of Language Testing from the Perspective of Validity

Room 304 (304 教室)

THEME (主题)

Formative Assessment in Foreign Language Teaching (外语教学中的形成性评价研究)

Chair: 王薇

08:30-09:00	王飞宇 Formative Assessment Implementation in India: A New Reform on English Curriculum in Elementary Schools
09:00-09:30	辜向东 高晓莹 李玉龙 高考英语四十年内容效度历史研究
09:30-10:00	梁丽 Measuring and Understanding Self-regulated EFL Learning Within an Online Formative Assessment Module
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	王薇 形成性评价在大学商务英语课程中的应用研究
10:50-11:20	李廉 Use Argument Against Scores of College English Course Assessment: A Regional Study of Jiangsu Province

Room 305 (305 教室)

THEME (主题)

Diagnostic Assessment (诊断测试)

Chair: 陈慧麟

08:30-09:00	刘书慧 A Diagnostic Assessment on the Receptive and Productive Vocabulary Size of Advanced Chinese Learners
09:00-09:30	王华 听力诊断性测试效度研究: 基于考生口陈报告的证据
09:30-10:00	杜文博 马晓梅 A TBR-based Cognitive Diagnostic Modeling for EFL Reading Test
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	孟亚茹 马晓梅 赵宁宁 晏艺赫 基于英语听力认知诊断试题的动态干预模式
10:50-11:20	陈慧麟 CDM Selection for English Reading Test and the Implications on Teaching and Learning

**Room 306 (306 教室)**

THEME (主题)

Peer Assessment (同伴评价)

Chair: 周季鸣

08:30-09:00	周季鸣 蒋燕 Peer Assessment as an Innovation Strategy: Students' Perceptions and Practices
09:00-09:30	李雪莲 Perception of Self, Peer and Teacher Feedback: The Perspective of Receivers and Givers
09:30-10:00	杜玉霞 基于同伴互评的专业英语演讲教学模式研究
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	张聪 Xun Yan Reconsidering Language Assessment Training Under the Framework of Teacher Education: Focusing on Assessment Contexts, Practices and Teachers
10:50-11:20	李雪平 How do Teachers and Students Perceive an Enhanced Score Report of EFL Reading Test

Room 307 (307 教室)

THEME (主题)

Classroom Assessment (课堂评价)

Chair: 唐雄英

08:30-09:00	Qiaozhen Yan Lawrence Jun Zhang Helen Ramsey Dixon Exploring Teachers' Conceptions and Practices in Assessing Young EFL Learners in the Classroom
09:00-09:30	唐雄英 Classroom Assessment Oriented to Self-regulated Learning
09:30-10:00	孙杭 Seeking Alternatives: Incorporating Teacher-based Assessment in a Spoken English Program
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	任玲玲 大学英语口语教学动态评估研究
10:50-11:20	印蕾 大学英语课程中“人际测评+人机测评”模式的构建与实证研究

Room 309 (309 教室)**THEME (主题)**

Application of Artificial Intelligence in Language Testing (人工智能技术在语言测试中的应用)

Chair: 潘之欣

08:30-09:00	徐莎莎 The Application of Automated Scoring System in the Teaching of ESL Writing
09:00-09:30	王海军 Huang Qian An Empirical Research into Reliability and Validity of China's AES Pigai and iWrite
09:30-10:00	潘之欣 Researching Task Difficulty of English News Listening-Based on Automatized Text Characteristic Analytical Tools
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	张荔 A Study on the Evaluation, Interpretation and Extrapolation of Pigai Automated Writing Evaluation System
10:50-11:20	孔菊芳 The Impacts of Response Format on Test-takers' Reading Comprehension Test-taking Process—Based on Eye-tracking Evidence

Room 310 (310 教室)**THEME (主题)**

Text Complexity (文本复杂度)

Chair: 李航

08:30-09:00	李航 Linguistic Complexity as Indicator of EFL Writing Quality
09:00-09:30	徐李荣 Is Syntactic Complexity Predicative of L2 Writing Quality? A Preliminary Study into Mean Dependency Distance as a Measure of Syntactic Complexity
09:30-10:00	雷蕾 A Corpus-based Analysis of Mandative Subjunctive Triggers In Chinese Learners' English Speaking and Writing
10:00-10:20	Tea Break (Room 118) 茶歇: 东六教学楼 118 教室
10:20-10:50	张玉美 罗少茜 Connecting English Test Performance with Writing Teaching and Learning
10:50-11:20	周珊珊 A State-of-the-art Review of International Language Testing (2011-2015)



Abstracts of Keynote Speeches

Keynote Speech 1

China's Standards of English and Their Applications



LIU Jianda



HE Lianzhen

With a view to creating a coherent framework of foreign language learning, teaching and assessment in China, efforts have been made in developing a national foreign language assessment system which includes the establishment of testing and assessment standards, foreign language tests reform, Gaokao reform, the use of formative assessment to facilitate learning, and the development of China's Standards of English (CSE). This presentation, following a justification for the development of CSE, introduces the theoretical underpinnings of CSE, the procedures followed in developing CSE such as descriptor collection, quality insurance, etc. The presentation ends with one application of CSE, focusing on the development of a National English Test System (NETS) including test design, intended test use, score report and test validation.

Keynote Speech 2

Developing and Validating Test Materials Within a Common Framework of Reference



Dr. Nick Saville

In relation to the development of China's Standards of English (CSE), this talk will reflect on the processes and procedures that are needed to develop language assessment systems based on a common framework using reference levels with multiple proficiency scales.

The speaker will use examples from the work carried out by the Association of Language Testers in Europe (ALTE) to illustrate some of the key issues. ALTE was formed in 1991, and over a period of 25 years, the members in 23 countries have engaged in the challenge of relating their own examinations to a common framework of levels for 26 languages.

In 1992 ALTE published its own framework and then worked closely with the Council of Europe (CoE, Strasbourg) on developing the CEFR – the Common Framework of Reference for learning, teaching and assessment that has now become widely used across Europe and around the world.

Given the diversity of membership, guidance has been needed to help develop and validate the assessment of language skills that serve multiple purposes across

the reference levels – from tests for younger children to high stakes examinations for migration purposes. A key challenge has been to communicate key principles to ALTE stakeholders in many different contexts so that they can understand and implement the necessary processes in consistent ways.

In 1996, the CoE commissioned ALTE to produce a guide for language test development and examining to accompany the first draft of the CoE, and then in 2009, the CoE requested an updated version to add to the 'toolkit' of resources to accompany the published version of the CEFR (2001).

The revised version of the guide, now known as the Manual for Language Test Development and Examining (2011) will be the main focus of the talk. It is a non-prescriptive document that seeks to highlight the main principles and approaches to test development that users can refer to when developing tests in their own educational contexts. It will be argued that many of these principles could be adapted for the Chinese context too.



Keynote Speech 3

Connecting Writing Assessments to Teaching and Learning: Distinguishing Alternative Purposes



Prof. Alister Cumming

Four conceptually-distinct options exist for relating assessments of writing to teaching and learning in programs of language education, each fulfilling different but interrelated purposes. Assessment purposes may either be normative, diagnostic, formative, or summative. Eight conventional assessment practices that realize these options are: (a) proficiency tests and curriculum standards, each based on different kinds of normative principles and data, (b) diagnostic and dynamic assessments focused on individual learners and their learning potential within a specific educational context, (c) responding to students' written drafts for formative purposes of informing their improvement, and (d) grades or local tests of summative achievements in a particular course. I argue that alternative options and intersecting relationships exist for connecting assessment, teaching, and learning, which can lead to mixing and confusions of purpose, conceptualization, and responsibility. I try to clarify these options and relationships by examining the premises, issues, and challenges associated with each of them.

Keynote Speech 4

Transforming Language Learning and Assessment Experience with Technology: Outlook and Challenges



Dr. Xi Xiaoming

In an era when technology is transforming education globally, how can research on technology build on robust theories and practice to motivate the design of innovative language learning materials and assessments? In this talk, I will provide an overview of current research on technology innovations in English language learning and assessment, and discuss ways in which emerging technologies continue to generate fresh impetus for innovations. While technology innovations hold tremendous promise for facilitating the design of new learning and assessment experience, these endeavors need to be driven by a few fundamental questions: what technologies to use in specific learning & assessment contexts? Why use them in these contexts? How to use them responsibly and effectively in each specific context? I will emphasize how inquiries into these areas naturally stimulate deeper thinking into theories and practices that provide the foundation for technology-enabled applications in language learning and assessment.

Construct and assessment design theories are at the core of designing language assessment. Technology can significantly enhance each aspect of assessment design, including assessment models, task design, psychometric

models, scoring, and score reporting. I will provide a few illustrative examples where technology plays a central role in expanding test constructs, enabling more efficient and accurate measurement, improving the efficiency and reliability of scoring, and providing fine-grained feedback for the purpose of improving learning. I will call special attention to some rapidly evolving technological capabilities such as games, simulations, and human-computer spoken dialogues, which have been extensively researched and put into use in language learning tools but have seen limited applications in assessments, with most work centering on research investigations. I will discuss how these new technologies can potentially challenge our traditional way of assessing language skills and enable more authentic, meaningful and engaging assessment experience. I will focus on laying out the opportunities and challenges for applying these technologies in an assessment context.

Compared to assessment contexts, we have seen much more prevalent use of technology in language learning contexts, ranging from formal classroom teaching contexts to informal learning environments where technology-based solutions are used. There is one



notable trend in language learning that is most closely intertwined with and propelled by technology advances – the emerging trend towards adaptive learning solutions. Increasingly, learners expect learning materials to be customized to their proficiency levels, needs, learning styles and preferences. In the same vein, ESL/EFL teachers are expected to provide differentiated instruction to promote learner autonomy in group instructed settings. They need to develop skills to effectively harness the power of technology to structure learning experiences in a way that accommodates individual learning trajectories and promotes deep, meaningful, student-centered learning in a classroom context. While the “holy grail” of truly adaptive language learning solutions has not been

found yet, there are research endeavors and practices that attempt to get us closer to succeeding in that quest. I will talk about ways in which current technologies can be utilized to put us on a productive path to designing customized, engaging language learning experiences, and note some gaps where active research and development is needed.

I will offer some concluding thoughts about the future landscape of technology innovations in the area of language learning and assessment, and discuss the need to continually leverage and adapt to ever-shifting technologies while remaining grounded in theories and best practices.

Keynote Speech 5

Innovation in Foreign Language Testing in China: Researching a Face-to-face and Video-conferencing Delivered Speaking Test



Dr. Vivien Berry

This presentation will report on a major study, carried out in China, into the impact on test construct of the mode of delivery of an existing speaking test, online video-conferencing and face-to-face. An initial, small-scale pilot study, designed to explore how new technologies could be harnessed to deliver the face-to-face version of a standardized speaking test, had investigated what similarities and differences, in scores, linguistic output and test-taker and examiner behaviour, could be discerned between the two formats.

The Shanghai study was a larger-scale follow-up investigation which used a convergent parallel mixed-methods design to allow for collection of an in-depth, comprehensive set of findings derived from multiple sources. 99 test-takers each took two speaking tests under face-to-face and computer-delivered conditions. Performances were rated by 10 trained examiners. The data collected included an analysis of feedback interviews

with test-takers as well as their linguistic output during the tests (especially types of language functions) and score ratings awarded under the two conditions. Examiners responded to two feedback questionnaires and participated in focus group discussions relating to their behaviour as interlocutors and raters and to the effectiveness of the examiner training. Eight observers also took field notes from the test sessions.

I will describe the examiner and test-taker materials developed for the training aspect of the study and present comments from the observers, in addition to presenting both quantitative and qualitative findings. The presentation will conclude with a discussion of the comparability of the construct(s) measured by the two delivery modes, which may have important implications for the future assessment of oral language proficiency and for the interpretation of scores derived from the different modes of delivery.



Abstracts of Parallel Sessions

Saturday Sessions

Room 101

高考外语科一年两考改革研究与实践

乔 辉
教育部考试中心

【摘要】 高考外语科开展一年两次考试是落实《国家教育中长期改革和发展规划纲要》和十八届三中全会精神，进一步深化高考内容和形式改革的重要举措。一年两考的实施方案从2012年底由教育部考试中心开始研制，经过多轮修改和论证才形成包括英语、日语、俄语、德语、法语和西班牙语6个语种的最终方案，并于2015年8月出版了针对一年两考英语科的考试大纲。此次改革将会对中学外语教学产生较为积极的影响。一年两考降低了考试的利害程度，将会减轻应试教学的负面影响。新设计的外语考试能更好地反映课程标准的要求，有助于学生综合语言运用能力的发展。

一年两考的英语科试卷全面考查学生的语言运用能力。新试卷结构的设计过程经过了新题型多轮试测和统计分析以及问卷调查结果分析等长达两年多的研究。一年两考英语科试卷保留了现行高考英语试卷的基本结构，替换掉其中的两个题型（单句填空和短文改错），形成了既科学合理又承前启后的试卷结构。一年两考的命题将加大试卷和试题积累，在保证试题质量的前提下，建立基于高考立场的现代化题库。

2016年10月16日，高考外语科一年两考各语种在浙江省首次实施，考前和考后的调研结果表明，一年两考的设计方案已经得到了广大师生和社会的认可，正日益在新一轮高考改革中发挥着重要的作用。

高考外语科一年两考改革如何科学落地？ ——考试评价视角下的技术保障

章建石
教育部考试中心

【摘要】 国务院颁布的“关于深化考试招生制度改革的实施意见”明确要求“外语科目提供两次考试机会”。目前，试点省份的方案已经公布，引起了一些争论和担忧。从微观层面来看，高考外语科实施一年两考，对教育测量技术在大规模教育考试中的应用提出了更高的要求。两次考试的试卷如何尽可能的“平行”以保证科学？两次考试成绩如何处理以确保公平？成绩如何报告以更好的导向教学？这些难题的破解，国内没有现成的经验，国际考试行业的规范也并不适用。立足于我国高考改革的现实土壤，我们需要对来自西方的现代考试评价技术进行创新性的改造和使用。在这方面，教育部考试中心近年来开展的高考考试评价改革，进行了一些积极的探索。

广东省英语听说考试的改革探索

黄友文

广东省教育考试院

【摘要】2004年开始，广东省在高考英语（2）的考试科目引入了“人机对话”形式，借助计算机配以口语计算机考试系统进行英语口语考试。本文从广东省英语听说考试改革的背景、改革的实践及成效三个方面总结了广东省英语听说考试改革的探索路径。从广东省英语听说考试改革实践的情况看，经过十二年的不懈探索，英语听说考试改革在缓解组考工作压力、促进学生健康成长和维护公平公正等方面取得了显著的成效。

【关键词】高考 计算机考试 口语 英语听说

语言测试与评价视阈下质量标准相关问题研究

褚劲风

上海教育考试院

【摘要】考试质量标准是指一套成体系的、能够对测试设计、试题命制、考务实施到成绩使用的整个测试流程的各个环节进行规范和检测的标准和准则，使考试各环节步骤有章可循、有据可依，从而保证命题的科学性、考试组织的严密性、评分的一致性和考试结果的可比性。论文以上海高考外语考试为案例，借鉴语言测试与评价理论，通过回顾上海外语高考改革的历程，考量现有做法，以期对未来建立国家外语能力测评体系提供实证依据和理论思考。

关于北京市中考英语改革的几点思考

臧铁军 徐加永

北京教育考试院

【摘要】教育部《关于进一步推进高中阶段学校考试招生制度改革的指导意见》从国家层面对初中学业水平考试提出了明确要求，这表明高中阶段的学校考试招生制度改革被正式提上日程。本文结合北京市中考改革的具体实际，简要介绍了北京市关于中考英语改革的方案，分析了改革的背景和必要性，梳理了改革的思路，粗线条地描绘了中考英语考试改革的蓝图，明确了考试内容和形式改革的方向和路径。

【关键词】中考；英语测试；内容改革；形式改革



高考英语北京卷计算机辅助听说考试研究

肖立宏
北京教育考试院

【摘要】为贯彻落实党的十八届三中全会精神及《国务院关于深化考试招生制度改革的实施意见》、《国家中长期教育改革和发展规划纲要》（2010-2020年）要求，北京自2010年起，依托“北京高考英语“听说机考”研究（2010年-2013年）”、“2014年北京英语能力测试暨北京市高考英语科目考试改革研究”两个项目，探索高考英语考试改革，进行了高考英语考试内容与考试形式改革研究。《北京市深化考试招生制度改革实施方案》公布之后，依托“北京市中高考英语机考系统建设（2016年）”项目，重点进行计算机辅助下的高考英语听说考试研究。

研究采用定性研究与定量研究相结合的方法。通过研究，明确了计算机辅助下高考英语北京卷听说部分的考试结构、考核目标、考试内容与范围、评分原则与标准，验证了计算机辅助听说考试的信度、效度、难度、区分度和可操作性，了解了目标考生、一线教师和评卷教师对计算机辅助听说考试的认识。在研究过程中，还进行了计算机人工智能评卷系统与人工阅卷的对比研究。

研究表明，计算机辅助下的高考英语听说改革方案可行，对中学教学有良好的导向作用。在口语列为必考问题上，研究表明，郊区校学生在口语考试上的得分并不低于同等水平的城区校，并且郊区一线教师和教研员对口语考试大多表示赞同。这些研究，为《北京市深化考试招生制度改革实施方案》的实施奠定了理论和实践基础。

对于高考英语一年两考改革的态度调查

——一项基于外语能力测评现状及需求分析的研究

张 浩
清华大学

【摘要】“一年两考”这一高考英语考试制度的重大变革对高中英语教学和学习必将产生深远影响。在这一背景下，本研究基于“中国外语能力测评现状及需求调查”课题高中学段的调查结果，讨论和分析了教师和学生群体对于高考英语一年两考改革所持的态度，以期能为一年两考政策和方案的制定及实施提供参考。通过对高中英语教师、本科一年级学生和高职高专一年级学生三个群体所反馈结果的分析，本研究发现：1）对一年两考政策持积极态度的受访者在三个群体中均占多数；2）高中英语教师和本科一年级学生对于一年两考可能带来的负面影响有着不同的认识；3）部分背景因素显著影响着调查对象对这一问题的观点和看法。

【关键词】一年两考；高考英语改革；态度调查；外语能力测评现状及需求分析

Room 102

Symposium 1

Discussant: Zhang Wenxia

Presenters: Zhang Hao, Wei Xing, Cheng Mengmeng, Wu Sha, Yan Yi

Survey of Status Quo and Reform Needs of Foreign Language Testing in China

Abstract: With the approval of the Ministry of Education, the National Education Examinations Authority conducted the largest ever survey of the status quo and reform needs of foreign language testing in China. The survey was launched in order to lay a foundation for the construction of the National Assessment System of Foreign Language Proficiency and to provide guidance on foreign language teaching in China. It covered 16 representative provinces and was administered mainly in the form of questionnaires (for ten different respondent groups). Each questionnaire is made up of three major parts: current status of English testing, English proficiency requirements, and reform needs. The survey ended up with an effective total sample of almost 80,000 respondents, including teachers and students from different stages of education (i.e., high schools, colleges and universities) in China and admissions officers from overseas colleges and universities.

This symposium is composed of five papers. The first paper concentrates on National Matriculation English Test (NMET). With responses from high school English teachers and freshman students from both colleges and higher vocational schools, the paper attempts to examine the impact of holding NMET twice a year on the parties involved. The second paper is devoted to four-year colleges and investigates how college senior students and college English teachers perceive the College English Test and its effect on college English teaching and learning, with an eye to possible assessment reform. The third paper explores to what extent existing English tests can satisfy the need of vocational colleges, for admissions and for graduation. Based on a test satisfaction model, the fifth paper will compare teachers and students and admission officers' different opinions on current National Entrance Examination for Postgraduate Schools, intending to provide some suggestions for its reform. The last paper explores the needs of overseas educational institutions for prospective Chinese students' English language proficiency and its testing, and also seeks their suggestions for further improvement on English language teaching and testing in China.

Prof. Bonnie Wenxia Zhang from Tsinghua University will chair the symposium and organize the discussion.



主 持 人：张文霞

工作单位：清华大学

联系方式：wxzhang@mail.tsinghua.edu.cn

分报告题目一：对于高考英语一年两考改革的态度调查——一项基于外语能力测评现状及需求分析的研究

A Survey on Attitudes on NMET's Two Test Administrations per Year -A Study Based on Large-scale Surveys

作 者：张浩

工作单位：清华大学

联系方式：hao-zhang14@mails.tsinghua.edu.cn

分报告题目二：以测试改革支持教学改变：基于本科阶段外语能力测评现状及需求调查

To support education reform through assessment reform: Based on a large-scale survey of college students and English teachers

作 者：魏兴

工作单位：清华大学

联系方式：wei-x14@mails.tsinghua.edu.cn

分报告题目三：关于高等职业教育分类英语考试必要性的调查及思考

A survey on the Need for Classified English Examinations for Higher Vocational Colleges

作 者：程蒙蒙

工作单位：教育部考试中心

联系方式：chengmm@mail.neea.edu.cn

分报告题目四：硕士研究生招生英语考试改革方向初探——基于考试满意度模型的构建与分析

Suggestions on the Reform of National English Entrance Examination for Postgraduate Schools: Based on Test Satisfaction Model Construction and Analysis

作 者：吴莎

工作单位：教育部考试中心

联系方式：wus@mail.neea.edu.cn

分报告题目五：国外教育机构对中国学生外语能力及其测评的需求分析

Analysis of Overseas Educational Institutions' Needs for Chinese Students' English Language Proficiency and its Testing

作 者：颜奕

工作单位：清华大学

联系方式：Yanyee@mail.tsinghua.edu.cn

总 联 系 人：吴莎 wus@mail.neea.edu.cn, 010-82520162

邮编及地址：北京市海淀区清华科技园立业大厦外语测评处 100084

Symposium 2

Discussant: Zou Shen

Presenters: Pan Mingwei, Zou Shen, Chen Jianlin, Deng Jie, Li Qinghua, Zhang Wenxing

Presentation 1

Towards a Unified Writing Proficiency Scale of China Standards of English (CSE): Why, How and Where

Li Qinghua, Southern Medical University

Abstract: This paper makes an attempt to explore the models of ELF writing ability. The theoretical models regarding ELF writing were reviewed and discussed critically and a tentative socio-cognitive model of EFL model is proposed, in which social, linguistic and cognitive variables are involved and interacted. This model offers the theoretical basis for the descriptors of EFL writing abilities of Chinese learners.

Keywords: EFL writing, ability model, cognitive perspective, social perspective

Presentation 2

Development of the Writing Strategy Sub-scale of China Standards of English

Deng Jie, Hunan Normal University

Abstract: The present research primarily focuses on the development and preliminary validation of writing strategy descriptors that form a sub-scale of China Standards of English (CSE). The descriptors were first collected and categorized in accordance with a three-dimensional model of writing strategy use, put forward on the basis of literature review of cognitive studies of writing processes and then edited and validated by way of intuitive, qualitative and quantitative analyses. A total of 238 initial descriptors were first collected from a wide range of source materials officially published both at home and abroad and then sorted into three main categories and six sub-categories, with each sub-category covering nine levels of strategy competence. After a series of workshops, cross-validations and questionnaire investigations, 166 descriptors remained and were put together with other CSE descriptors in nation-wide, large-scale investigations. The extracted descriptors with writing strategies are intended for an enhancement of English writing teaching and materials development.

Keywords: writing strategies, writing process, strategy competence



Presentation 3

Towards Exemplary Writing Activities for the China's Scales of English Writing Proficiency: A Systemic—Functional-Linguistics Text Typology Perspective

Pan Mingwei, Guangdong University of Foreign Studies

Abstract: Written production plays a crucial part in the China Scales of English Language Proficiency. Whether, if so, how language learners or users can perform exemplary writing tasks or activities becomes significant indicators, against which their language proficiency can be described and measured. With the text typology from a Systemic Functional Linguistics (SFL) perspective as a point of departure, this paper, in congruence with the SFL text typology, elucidates on how to extract exemplary writing activities from a pool of existing writing ability descriptors so that texts with a repertoire of functions can be implanted into the descriptor construction embedding exemplary writing activities of different domains. It is also asserted that the construction of language proficiency descriptors should take into account the text quality as well as the conditions under which texts are produced in order that learners or users across different proficiencies can be stratified. More quantitative studies need to be furthered to triangulate the results of stratification.

Keywords: exemplary writing activities, systemic functional linguistics, text typology

Presentation 4

Are the CEFR Adaptable in the Construction of a Writing Ability Scale for English Majors in Chinese Universities? —A Case Study

Zhang Wenxing, Jingdezhen Ceramic Institute/National Educational Examinations Authority
Zou Shen, Shanghai International Studies University

Abstract: The CEFR, ever since its inception, has had profound impact on language teaching, learning and assessment not only in Europe but also in other parts of the world. This study focuses on the adaptability of CEFR writing descriptors in the context of English majors in Chinese universities. First, we constructed a questionnaire based on the descriptors collected from various sources in order to elicit university teachers' views on the importance of these descriptors. A revised version was produced based on the feedback from the initial questionnaire survey. In order to further investigate what level or levels these remaining descriptors would fall into, 35 university teachers of English were invited to complete the revised questionnaire while rating 36 TEM (Test for English Majors) writing scripts. Then, band-setting of the descriptors was initially determined on the basis of the questionnaire data, the result of which was the draft scale of writing ability. In order to collect further evidence for our calibration of the descriptors, 8 university teachers of English were interviewed. Based on the interview data, some descriptors were fine-tuned before the scale was finalized. The results have showed that CEFR writing descriptors can be used in the description of the writing ability of our English majors, but most of the CEFR descriptors surveyed have had their original level altered in our ability scale. The study results have implications for further research in the CEFR's adaptability in other contexts, and also for the construction of China Standards of English (CSE).

Keywords: CEFR, writing scale, Test for English Majors (TEM), descriptors

Presentation 5

The Interface Between China Standard of English Writing and Second Language Writing Teaching

Chen Jianlin, Lanzhou University

Abstract: China has a large English learning population so that the development of China Standard of English (CSE) must serve, among others, for English learning and teaching. Language scale is naturally related to language teaching. It in the first place theoretical and scientifically divides language proficiency into a series of continuously and systematically related levels which will constitute a framework for language teaching of various purposes and at all stages. It also serves as the guideline for what to teach as indicated by the contents it describes at all levels. It will, therefore, make language teaching transparent for stakeholders to know their roles in the process. The development of China Standards of English Writing (CSEW) follows the principle of serving English writing teaching practice. First, teaching was taken full consideration in establishing the description framework that was decided to be composed of communication purpose, text features and activities which are expected to provide specific guidelines of content and activities for English writing learning and teaching. Second, writing teaching was also weaved into the process of the development that includes descriptor collection, validation and classification. Specifically, descriptors were to a large extent collected from teaching standards, syllabuses, and textbooks; teachers in all levels were invited to be involved in the validation process and a large number of learners and teachers were employed in questionnaire data collection.

Keywords: China Standards of English Writing, second language writing teaching, interface

Working on the China's Standards of English (CSE) —Organizational Competence Specified

Research Team: Wu Zunmin, Luo Shaoqian, Lin Dunlai, Qian Xiaofang, Xu Yun,
Liu Liping, Gao Miao, Yang Lvna, Zhao Haiyong, Jia Yidong
Beijing Normal University Minzu University of China; Foreign Language Teaching and
Research Press; Central University of Finance and Economics; Shandong University of
Finance and Economics

Organizational competence (Bachman 1990, Bachman & Palmer 2010) is the basis for the use of language skills in communication. It is therefore an indispensable part of the China Standards of English (CSE). The proposed presentation reports the steps taken and research done in the process of the specification of Chinese English learners' organizational competence. While ELT in China has a long tradition of stressing grammar and vocabulary instruction, CSE decided to expand the scope to include pronunciation and graphology, sentence structure, vocabulary, cohesion and coherence. The first efforts were to collect descriptions of learner ability from a vast range of resources, CEFR, CBM, curricula standards, to name but a few. Then it is the generation of descriptors. It involved adaptation, combination, complementation or reduction to shape the desired descriptors for the Chinese EFL/ELT context. Investigations to verify the usefulness of the descriptors was the most important part of the job.

It is believed that the coming into being of the organizational competence descriptors for the CSE will be of great importance to English teaching, learning and assessment in China across formal educational stages, for informal training institutions, learning materials publication and with various stakeholders.

论语言能力的描述

冯 莉

黑龙江大学

全球化的时代使语言学习活动的重要性日益突显，因此各国都积极致力于语言力量表的开发，以期能够对语言教学与测试提供统一的指导框架。语言力量表开发成功与否很大程度上取决于描述语的质量，但关于语言能力的描述问题还没有在理论上得以解决。本文从“能做”描述语的语义结构分析入手，结合国际上现有语言力量表的描述语实例，试图探讨语言能力描述的理论基础与方法范式。本文还探讨了语言能力描述的难点和提高描述质量的策略。具体内容如下：

一、语言能力描述的需求与语言力量表研究回顾

二、作为事件描写的“能做”描述框架

三、“能做”描述的语义结构分析

1. 事件

2. 行动（+方式）

3. 事物（+属性）

4. 结果

四、语言能力描述的难点

1. 指称的不确定性

2. 分类问题

3. 难度问题

五、提高描述准确性的策略

1. 确保可观测性

2. 确保典型性

3. 把握共性与个性



Description of language proficiency

Since language communication activities have been increasingly important in the era of globalization, many countries are developing language proficiency scales in order to provide a unified framework for language teaching and testing in their education systems. The success of language proficiency scales is highly dependent on the quality of descriptors, but the description of language proficiency has remained an unsettled theoretical problem. This paper strives to establish an event description framework for language proficiency based on the semantic analysis of “can-do” descriptors in the scales widely used worldwide. Difficulties in describing language proficiency and strategies to cope with them are also discussed. The outline is as follows:

1. Research background
 - 1.1 Needs of description of language proficiency
 - 1.2 A literature review of the studies of language proficiency scales
2. An event-description “can-do” framework for the description of language proficiency
3. A semantic analysis of “can-do” descriptors in current scales
 - 3.1 event
 - 3.2 action (+modes)
 - 3.3 object(+properties)
 - 3.4 result
4. Difficulties in describing language proficiency
 - 4.1 uncertainties of referential functions of metalanguage
 - 4.2 diverse classifications
 - 4.3 obscure level of difficulties
5. Strategies to improve the quality of descriptors
 - 5.1 Striving for observeability
 - 5.2 Striving for typicality
 - 5.3 Balancing universality and individuality
6. Summary

中国英语口语能力量表的临界值研究——IRT理论的应用

揭薇

上海交通大学/上海对外经贸大学

最佳临界值的确定是量表研制中的一项重要工作，科学的临界值界定对量表的灵敏度、稳定性至关重要。本研究采用Rasch模型以及IRT两参数模型分析中国英语口语能力量表（CSES）在大学阶段B级和C级的最佳临界值。通过分层抽样的方法对11个省市89所高校的部分英语教师和学生根据CSES、CSES自评量表编写13份问卷，将问卷等值后展开调查。对收集的数据进行分析，绘制学习者能力特征曲线，确定口语能力量表中20个分量表各自的最佳临界值并判断其价值，为CSES量表的级别划定提供可靠证据。

The determination of optimal cut-off point is an important work of scale development. Scientific investigation of the cut-off point is of critical value for the sensitivity and stability of the scale. This research adopts the Rasch model and IRT two parameter model to analyze China Standards of English-Speaking scale (CSES), focusing on the Level B and Level C descriptors. We investigated 89 universities in 11 provinces in China by the method of stratified sampling, teachers of English and students participate in this investigation. We constructed questionnaires based on the provisional scale, 13 questionnaires were equated across different levels. The learner ability characteristic curves were drawn according to the data collected. The best cut-off points were determined and discussed in 20 speaking subscales respectively; therefore provide reliable evidence for the defining of consecutive CSES scale levels.



中国英语能力等级量表建设——英语口语典型活动调查

王隽 刘畅
上海交通大学

当前，中国英语能力等级量表正处于开发的重要阶段。其中，典型口语活动量表（后简“活动量表”）的制定是口语量表开发中的重要一环。活动量表将囊括英语口语教学和日常生活中代表性较高的口语活动，以及在不同能力级别中口语表现的详细描述。对于口语活动的典型性调查则是建立活动量表的前提，更是其内容效度（Content Validity）的保证。由此，为了使活动量表能最大化贴近英语口语活动的实际使用情况，同时保证较高程度的内容效度，本研究以使用频率、对应级别、任务难度和教学适用度为切入点，为英语口语活动的使用现状做一次系统的调查和分析。

在第一阶段的研究中，为了能够最大程度地网罗现有的口语活动，本研究先后进行了9次小规模的教师工作坊和针对各类口语资料的文献分析。工作坊在全国范围内展开，通过问卷和访谈收集了从A1到C2不同级别口语教学中常见的22类口语活动。文献分析主要通过对口语考试（36个）、考试大纲（13个）、国内外教材（35本）等文献的系统分析，收集了来自5类场景（考试场景、教学场景、工作场景等）中的共15大类、45小类口语活动。以上活动均应用于后续的问卷调查中。

在第二阶段的研究中，本研究将进行两次大规模的问卷调查，分别针对教育领域和非教育领域的典型口语活动，调查对象包括教师和社会人士，预计收集约450份问卷，将运用描述统计和推论统计的分析方法，较为全面地呈现各类口语活动在不同场景中的典型程度，为量表的内容效度提供有力保证，也为活动量表的制定提供宝贵依据。

China's Standard of English: Typicality Investigation of English Speaking Activities

In the project of developing China's Standard of English, within the speaking sector, Scales for Speaking Activities (SSA) is rather an important part. SSA will include not only the typical types of speaking activities in various contexts of teaching, learning and daily life, but detailed descriptions of how well the tasks are accomplished by learners from different proficiency levels using illustrative performances descriptors. To fulfill such purpose, a need analysis collecting the existing speaking activities and evaluating the typicality of them is somewhat an inevitable step in the early developmental stage of SSA, which may be also considered an indispensable assurance for maintaining SSA's Content Validity. Thus, for SSA to maximally reflect the real-life applications of English oral activities and the maintaining of high-level Content Validity, this study aims to investigate the current status of speaking activities in China's English teaching, learning and daily life contexts from the aspects of frequency, associated proficiency levels, task difficulty and teaching adoptability.

In the first phrase of research, this study has conducted 9 small-scale nation-wide teacher workshops as well as a series of document analyses to collect the existing English-speaking activities. The teacher workshops applied both the methods of questionnaires and interviews and successfully collected 22 most-commonly used English-speaking activities in the classroom ranging from A1 to C2 level. On the other hand, document analyses involved materials such as 36 speaking tests, 13 test syllabus, 35 English teaching textbooks and so forth gathering activities from 5 different contexts, namely testing, teaching, personal, transactional and occupational contexts. The analyses resulted in 15 macro-taxonomical and 45 micro-taxonomical speaking activities, which will all applied into the following survey questionnaire design.

In the second research phrase, this study will conduct two large-scale survey investigations targeting English-speaking activities in the educational and non-educational domains respectively. The participants will include English teachers and the general public. Overall the study may collect approximately 450 valid questionnaires. Both descriptive and inferential statistics will be used in data analyses. The results will unveil considerably the typicality of English-speaking activities used in diverse communicative contexts providing strong empirical support for the Content Validation of SSA as well as referential information for the development of SSA.



中国大学生英语写作元认知策略能力等级量表的构建研究

周建华

长沙环境保护职业技术学院

【摘要】本文是在《中国英语能力等级量表》研制这一背景下，以中国大学生为目标群体，探讨如何建构英语写作元认知策略能力等级量表。本研究以Bachman的交际语言能力理论、Flavell的元认知理论和邓杰的话语信息认知处理为理论基础，借鉴CEFR框架对策略的分类，将英语写作元认知交际策略能力等级量表分为规划、执行、评估、补救四个维度，并采用定性与定量相结合的方法，对学生的写作元认知策略能力等级量表的建构进行研究。本文主要探讨以下两个问题：1) 学生英语写作元认知策略能力等级量表可分为多少维度？2) 学生英语写作元认知策略能力等级量表可分为多少个等级？针对以上两个研究问题本文进行了如下具体研究：描述语的采集与编制、描述语的分类与定级。描述语的编制主要是运用翻译、改写、拆分、合并、自编等方法对描述语进行加工处理，以“能做”的形式呈现。描述语的分类主要采用因子分析进行验证，描述语的分级主要采用多层面Rasch Facets模型，将收集来的描述语进行定量分析，完成初始量表的验证。分类和定级是构建等级量表构建的关键步骤，分类应有坚实的理论基础，定级应有定性和定量证据的支持，从而保证量表的信度和效度。元认知策略语言力量表的建立有利于学习者明确目标的阶段，激发和提高学生英语写作的积极性。另一方面，语言能力标准也能改变外语考试种类繁杂，教学互不相接的局面，同时为《中国英语能力等级量表》研制提供借鉴。

【关键词】英语写作；元认知策略；等级量表；分类；定级

Aligning China's Standards of English (CSE) with the CEFR

Peng Chuan

Guangdong University of Foreign Studies

Abstract: The CEFR has played a dominant role in language and education policy, language teaching, learning and assessment practice within and beyond Europe since its publication in 2001. China Standards of English (CSE) is a transparent, coherent and comprehensive scale for Chinese English learners, a common framework of reference for foreign language learning, teaching, and assessment in China. Although the CSE is developed specific to the Chinese context rather than by adapting the CEFR to the Chinese context, it is still important and inevitable to align the CSE to the CEFR in the light of international communication, the CEFR's influence worldwide, internalization of CSE and the improvement of Chinese students' English proficiency.

This alignment study is exploited based on three research questions: 1) In what way are Chinese students related to the CEFR? 2) To what extent is the CEFR applicable to the Chinese context? 3) How is the nine levels in the CSE related to the six levels in the CEFR? To answer the three research questions, firstly, linked questionnaires made up of CEFR descriptors will be distributed to the representative sample based on Rasch scaling, and the relationship between CEFR and Chinese students' English proficiency is figured out. And then some anchor items (descriptors) in CSE are anchored to the CEFR covering all levels. At last all the descriptors in CSE and CEFR are on the same scale based on Rasch analysis, the relative relationship between CSE and the CEFR can be figured out.

Key words: China Standards of English (CSE); CEFR; alignment

关于中国大学生外语能力评价的思考

袁 靖
黑龙江大学

中国大学生外语能力评价是我国外语教育领域的核心研究问题之一。2014年,《国务院关于深化考试招生制度改革的实施意见》也提出要加强“外语能力测评体系建设”。可见,构建我国的外语能力评价体系十分迫切。

当前的中国外语教学倾向于将测试分数作为评价学生外语能力的重要指标,这种终结性评价不能完全展现中国大学生外语能力的全貌。外语能力评价应该从外语教育的各个阶段收集证据,偏重过程,将评价的理念渗透到外语教学的所有环节之中,将形成性评价和终结性评价有机结合,建立一套合理的中国大学生外语能力综合评价体系。本文基于对丹麦大学理科课程改革的研究,阐释其分析和评价学生能力发展的方法,思考如何将该法加以改造,并应用于我国大学生外语能力的评价上,为我国外语能力评价体系建设提供思路。本文的主要研究问题包括:

1. 当前国内外外语能力评价的方法主要有哪些,存在的问题主要是什么;
2. 基于标准的评价:介绍丹麦使用SOLO taxonomy来分析学习者的能力发展(competence progression);
3. 研究反思:思考上述方法的理据是否充分,操作是否合理,如何利用或者改造利用以建构我国大学生外语能力评价综合体系。

Strategic Vision and Approaches to Constructing the Foreign Language Ability Assessment System in China

Lv Shenglu
Beijing Foreign Studies University

Abstract: The ideology of constructing the foreign language ability assessment system in China has been the hot issue in the academic and social circles since it was put forward, which reflects the new ideas, national strategies and reality appeals of foreign language education and assessment in our country. This paper puts forward a strategic vision of constructing the foreign language ability assessment system in China, and makes a comprehensive analysis on it from eight different aspects of assessment need analysis, test code, professional assessment institute, item bank building, proficiency scale, collaboration mechanism, test quality evaluation system and tester's assessment literacy, and accordingly points out some reflections on China foreign language assessment.

Key Words: foreign language proficiency scales; assessment system; construction vision; specific paths

中国英语学习者会话含意理解能力测评

郑群 李悦

中国科学院大学

【摘要】语用理解是成功的语用表达的前提（李清华、邹润 2015：47）。以往的会话含意理解研究（Bouton 1992, 1994, 1999; Roever 2006; Taguchi 2005, 2007）主要涉及学习者水平、反应时与理解不同会话含意的关系，但是对会话含意难度界定和造成学习者理解差异的具体原因并无深究。本研究拟从社会认知视角出发，探讨界定会话含意难度的标准，并以此标准分析学习者的理解差异，旨在为会话含意的测试和教学提供参考依据。我们选取了16道会话含意理解题，由三位外教判断难易度，而后对38位非英语专业的大一学生进行了上机测试，通过软件记录答题时长。测试结果显示：（1）会话含意的理解与英语水平并不呈现相关关系，但是与答题时长有关；（2）理解较为困难的会话含意规约性较弱，而会话含意理解相对容易的规约性较强。一周后我们选取其中6位同学进行访谈，对会话含意理解进行口头回顾。我们发现，学生的前知识与现实情境知识相互作用，影响会话含意理解。规约性较强的会话含意调动的母语经验和个人体验显著多于规约性较弱的；而规约性较弱的会话含意调动的现实情境知识显著多于规约性较强的会话含意。这一点说明，规约性可以作为界定会话含意难度的标准，而且学习者的个人体验和前知识对于理解会话含意，尤其是规约性较强的会话含意，至关重要。在教学和跨文化交流中，鼓励学习者注意有差异的经验知识，并积累个人体验，对于语言学习将大有裨益。

【关键词】会话含意；理解；个人前知识；现实情境知识；相互作用

口译教学中的质量测评参数

王巍巍

广东外语外贸大学

【摘要】在职业译员的培养过程中，口译质量测评是翻译教育院校检验教学效果的重要手段，对译员口译质量观的形成起着至关重要的作用。测评的关键在于测评参数的确立。本文采用文献法考察行业规范文件、口译资格认证考试及翻译院校毕业考试中所涉及的相关口译产品质量内容，通过梳理分析找出三方面均认可的口译质量参数集。在此基础上，结合“广外模式”的教学设计理念，描述面向专业译员能力培养的口译教学测评参数建构。

【主题词】口译教学；测评参数；广外模式

Abstract: Quality evaluation, as an important element of interpreting teaching, exerts great influence on student interpreters. And the key for valid quality evaluation lies in specified quality parameters. Upon such consideration, this paper reviewed quality related documents from professional standards, accreditation tests, and schools of translation and interpreting to establish quality parameter framework in “GDUFS Approach”.

Key Words: interpretation teaching; evaluation parameters; GDUFS Approach



基于国家英语能力等级量表的考试说明制订研究

刘 森

北京师范大学/黑龙江中医药大学

【摘 要】语言力量表 (Language proficiency scales) 描述了语言使用者运用某种语言的能力, 不同级别描述语对应该级别语言能力水平, 如欧洲语言共同框架 (CEFR)。语言力量表应用范围广泛, 其对语言学习、教学和测评具有指导作用。涵盖不同语言发展阶段的量表体系可以为语言测试开发提供参照。语言力量表相关研究更多地关注已有测试与语言力量表的连接问题, 以改进测试或者验证测试效度。如2010年《连接语言测试与欧洲语言共同参考框架的手册》收录了12项不同国家和地区语言测试与CEFR的连接研究, 并指出语言测试与CEFR连接需要四步骤: 熟悉 (familiarization)、考试说明 (specification)、标准设定 (standard setting) 和效度验证 (validation)。但很少研究涉及如何利用语言力量表开发新测试, 并制订测试的考试说明。本文将参考语言测试与语言力量表的相关研究, 试图探讨基于语言力量表制订考试说明的可能性。

Abstract: Language proficiency scales are a series of descriptions about language users' language proficiency, corresponding to different language proficiency levels, such as the Common European Framework of Reference (CEFR). The scales can be applied in a wide range, as a guideline for language learning, teaching and assessment. The scale system covering different stages of language development can provide reference for developing language tests. Research in language proficiency scales focuses more on the alignment between well-developed tests and language proficiency scales in order to improve or validate the tests. For example, *Aligning Tests with the CEFR* introduced 12 studies on the alignment between tests from different countries or regions with CEFR, and put forward the four steps for alignment: familiarization, specification, standard setting and validation. But little research focused on how to use the language proficiency scales to develop new tests, and formulate specifications. Drawing on some related research, this article aims to explore the possibility of formulating specifications based on language proficiency scales.

大学英语A、B级考试对高职英语教学反拨效应的研究

刘 婧

安徽三联学院

【摘 要】语言测试是伴随着外语教学的发展而产生的，是外语教学中不可缺少的一个环节。语言教学离不开语言测试，两者之间是相辅相成，不可分割的。根据Alderson&Wall (1993) 的研究，测试既影响教学也影响学习，这种影响被称为反拨效应 (back wash or wash back)，它会对教师的教学内容、方法、速度、顺序、程度、深度和态度以及对学生的学习内容、方法、速度、顺序、程度、深度和态度、学习动机等方面产生正面和负面的反拨效应。大学英语A、B级考试属于我国英语语言测试的一种，全名为全国高等学校英语应用能力考试，简称PRETCO, 分为A和B两个级别，A级考试为高职高专学生应该达到的标准要求，B级略低于A级，AB级的能力要求相当于大学英语三级水平，以大专院校在校生为对象。通过次考试者，国家统一颁发证书。本文就是以Alderson&Wall (1993)的反拨效应理论为背景，研究AB级应用能力考试的反拨效应对于教师教学以及学生英语学习的影响。研究以本学院非英语专业的50名大学生和教授实用英语3名老师为研究对象，基于对学生和老师的问卷调查，通过对AB级考试是否影响以上所阐述的各方面进行调查研究，并对教师和高职类学生进行后效调查，分析其反拨效应。结果表明，AB级考试对高职类学生的学习动机、态度、学习内容和方式以及学习的广度和深度有一定的影响，并且在不同类型的学生身上表现出了一定的差异性；同时对于教师的教学内容和方式也有一定的影响。其原因是由于语言测试的反拨效应引起的。

【关键词】大学英语AB级考试；英语教学；反拨效应



高风险测试改革教师态度研究 ——以英语专业八级考试为例

刘宝权 范劲松 侯艳萍
上海财经大学 复旦大学 上海外国语大学

【摘要】英语专业八级考试（TEM8）是专为中国大学英语专业本科生设计的一项高风险英语测试。该测试用来衡量英语专业学生在第八学期的英语水平，并评估这些学生是否符合全国英语专业教学大纲中规定的英语能力要求。2016年，英语专八实施历史上的第三次变革。该变革将对接全国本科英语专业教学质量标准的颁布。英语测试改革会对所有利益相关者带来重大影响，因此在教育中发挥着重要的作用。新测试必须要得到利益相关者的认可。本研究的目的是调查教师对英语专八改革的意见和看法，并确定其报告意见的积极性和消极性的来源。本研究采用混合方法设计，包括两个阶段：问卷调查（ $n = 192$ ）和面对面访谈（ $n = 25$ ）。问卷调查中，作者对大量数据进行了有效分析，包括信度分析、因子分析、相关和方差分析。定性数据使用NVivo进行编码处理。研究结果表明，教师普遍对TEM8的改革进行了积极的评价，特别是诸如听力小型讲座，阅读和写作，他们相信这些改革措施会产生积极的反拨作用，且会对他们的教学和学生的学习产生积极的影响。但是对于新闻听力、人文知识和英汉翻译等部分的取消也存在不同的意见。这项研究的结果对英语专八未来的发展及改革，以及其它情况下的测试改革有启示作用。

【关键词】英语专业八级；改革；教师态度

大学英语四级考试新闻听力测试的反拨效应研究 ——聚焦考生

张 放

东莞理工学院

【摘 要】2016年6月，全国大学英语四级考试取消短对话和短文听写，新增短篇新闻听力理解。新闻听力涉及现实广泛的题材和大量专有词汇，属于较高难度的听力测试，曾经仅作为高等院校英语专业四八级考试的测试项目。近年来，新闻听力测试在四六级网考中实施过，但调查显示，考生普遍反应该题目难度大，应考有困难，不知如何展开备考。目前，新闻听力测试首次对全国数千万考生统一施考，这一最新的考试改革将在大学生中产生何种反响，后者如何应对这项新的测试，考生的备考情况与命题者的初衷又是否一致，结果尚不可知。本研究主要调查四级考试新闻听力测试对考生的英语学习产生何种影响。主要研究问题包括：新闻听力题型的命题初衷是什么？考生如何看待新闻听力测试，包括试题难度和分值（占全卷7%权重）等特征？考生投入多少时间和精力、利用何种材料、采取何种方法进行新闻听力备考？本研究旨在对大学英语四级考试命题工作提供反馈，从应试者角度检验新闻听力测试的后效效度。调查发现，考生的备考表现并不统一，对于考试难度和分值的不同理解促使考生实施不同的备考行为，部分考生由于新闻测试难度大、分值低而投入很少的时间和精力，甚至放弃备考该题型，使得命题的初衷难以得到实现。本研究结果为大学英语教学提供了建设性意见，有利于帮助教师思考如何在新的教学要求和考试要求下有效辅助学生提高英语听力技能。

Aptitude for Consecutive Interpreting and Its Testing

Yang Zhihong

Soochow University

Abstract: Aptitude for interpreting includes bilingual proficiency, interpreting-related skills, cognitive abilities, encyclopedic knowledge, affect, motivation and personality traits, etc. The first four components are often known as hard skills while affect, motivation and personality traits are known as soft skills. There has been an increasing amount of research on soft skills over recent years while hard skills of interpreting, which are believed to be the core skills, still remain to be further empirically investigated. In this context, this study attempts to conduct an empirical research on aptitude for interpreting. Given that consecutive interpreting and simultaneous interpreting differ in the skill sets required, this study only focuses on aptitude for consecutive interpreting. To be more specific, it probes into the hard skills and their testing. Following a discussion on what constitutes aptitude for consecutive interpreting and based on a literature review and a survey on the testing practices of some translation programs, an empirical study targeting undergraduate and postgraduate students is carried out to investigate the relationship between subjects' language proficiency, interpreting-related skills, public-speaking skills and their performances in consecutive interpreting.

Key words: aptitude, consecutive interpreting, testing



Searching for an Optimal Design(s) for a Bi-directional English/Chinese Interpreting Test: Using Pooled Variance Components from Multiple Generalizability Studies

Han Chao
Southwest University

Abstract: Summative assessment of student interpreters' performance by the end of a training course/program is widely and routinely practiced at tertiary-level educational institutions in mainland China. Usually, numeric scores obtained from assessment are used to inform such decisions as granting academic credit, conferring a degree, and determining the effectiveness of curriculum design. These score-based decisions have consequential effects on students, teachers, curriculum developers, and other relevant stakeholders. Despite the relatively high-stakes nature of summative assessment, little research has been conducted to identify an optimal assessment design(s) so that the generalizability of outcomes can be achieved for interpretation testing in the context of interpreter education.

Against this background, we report an empirical study to find out an optimal combination of tasks and raters to achieve an acceptable level of score generalizability in summative assessment of English/Chinese consecutive interpretation (CI), using generalizability (G) theory. Particularly, we draw upon pooled variance components from three fully-crossed generalizability studies to inform future assessment design.

In the study, 38 undergraduate students participated in three assessments for the Advanced Interpreting course on three occasions. On each occasion, each student performed CI in three English (E)-to-Chinese (C) tasks and three C-to-E; each task featured a two-minute generalist speech; and all students' performance was audio-recorded. Six raters assessed all students' CIs, based on three rating scales of information completeness (InfoCom), delivery of fluency (FluDel), and target language quality (TLQual). Ultimately, three fully-crossed data sets (i.e. 38 students \times 3 tasks \times 6 raters \times 3 criteria for each direction) were subjected to G-theory analysis, using the software of EduG 6.1e.

The analyses show that: 1) based on the current design (i.e. three tasks and six raters) the InfoCom scores were, on average, less generalizable than those of FluDel and TLQual for both interpreting directions; 2) although sampling more tasks and/or using more raters would generally improve score generalizability, the marginal effect became increasingly diminished with more tasks and/or raters; 3) using more raters was more effective in raising score generalizability than adding more tasks in C-to-E interpreting for all three criteria, while such consistent effect was not observed in the other direction; and 4) based on the analyses, several potentially cost-effective designs that produced a G coefficient larger than 0.80 were identified.

These results were discussed, highlighting the complexity of ensuring score generalizability for summative assessment of bi-directional interpreting, and of identifying an optimal measurement design for high-stakes educational use under budget constraint.

Room 305

Helping Learners Take Control of Their Own Writing within the Framework of Assessment for Learning Framework

Huang Jing Chen Wencun
China West Normal University

Abstract: How to help learners take control of their own writing has become one of the many hot research issues. In order to overcome the limitations of summative assessment and feedback approach, a formative integrated assessment and feedback approach was proposed as the result of synthesizing social learning theories and Assessment for learning (AFL) theory, as well as taking into consideration current educational technology development and the researcher's teaching experience. This presentation intends to explore how to design an EFL writing course and manage to activate students' autonomy in and out of the classroom and its impact on the students. The presenter also will talk about the function of technology and students' voice in course design and management.

Key words: Writing course, Assessment for Learning, students' voice, technology



Measuring the Learning Process: Theory and Practices

Lance Knowles

Liulishuo

A major benefit of the technology revolution in education is the ability to monitor student learning activities in much greater detail than previously possible. Not only can we measure progress, but we can, for the first time, measure the quality and efficiency of the learning process itself.

However, assessing the quality of a learning program requires a learning theory and a learning sequence to measure against. Data alone can be misleading. Without a learning theory and a set of standards defined by the theory, data correlations may lead to false conclusions. In other words, there must first be a theory about how learning takes place in the brain and how to design and sequence learning activities to work with it.

In our learning programs, the goal is to facilitate and optimize language skill acquisition. This requires practice, which is fundamental to skill acquisition. We view practice as having four dimensions: (1) amount of practice, (2) frequency of practice, (3) quality of learning activities within the practice and (4) level and sequence of learning activities relative to proficiency level.

By monitoring and analyzing these dimensions, we can estimate the effectiveness of practice over time and make predictions about learning outcomes. We can also use this information to coach learners in how to modify and improve their practice.

To do this, we use a learning theory, Recursive Hierarchical Recognition (RHR), to decide whether a pattern of learning activities facilitates or impedes learning. While developing listening comprehension, for example, inappropriate use of text, which is spatial, can interfere with the development of language chunking skills. Accordingly, the presence of text can be a distraction which desensitizes the neural pathways and subconscious processors necessary to search out, recognize and employ patterns in spoken language.

In our learning theory, language chunking, which is a subconscious process, is essential for fluency, so anything that impedes its development has a negative value. We therefore suppress the initial use of text and encourage students to engage with the spoken form of English first. To track this, we have designed our program's user interface to facilitate that behavior and to collect the necessary data. In other words, the learning theory defines and shapes the data required to make a judgment.

Once data is obtained and analyzed, we check to see whether or not it makes sense according to the learning theory. If the data doesn't make sense, it doesn't necessarily mean that the learning theory is false. Rather, it may indicate that the data itself is skewed or incomplete. This happens when inappropriate tests are used to measure progress, which is a frequent problem. Tests themselves are problematic, especially when they are more the achievement type rather than a measure of skill proficiency.

We will present an example of data that leads to false conclusions and show how the learning theory reveals fundamental problems in the data and suggests remedies.

Formative Evaluation of College English Based on Output-Driven Input-Enabled Hypothesis

Li Shaolan

Zhan Quanwang

West Anhui Medical College Anhui University

Abstract: According to Output-Driven Input-Enabled Hypothesis, this essay attempts to construct a formative evaluation system of College English. Based on the Input inside and outside the classroom, the learners are required to accomplish targeted output tasks or projects. Meanwhile, their output performance is evaluated and included as a part of course achievement. The Input-Output-Evaluation teaching process encourages students to learn actively, improves learning efficiency, maximizes the effectiveness of classroom teaching, puts College English into application, adjusts talent cultivation to social development.

A Study of Formative Assessment for Academic English Writing of Chinese EFL Learners

Zhang Li Chen Defeng

Shanghai Jiao Tong University

Abstract: Formative assessment(FA), as a newly emergent method of assessment compared with summative assessment(SA), has enjoyed much popularity in education assessment these years, especially for the teaching of writing. However, the application of formative assessment for academic English writing (AEW) has rarely been studied. This paper, employing both quantitative and qualitative methods, is designed to explore the effectiveness of formative assessment in the process approach for academic English writing of EFL learners in China. Experimental group and control group were included in this study with the difference that formative assessment was carried out in the experimental group while summative assessment was used in the control group during their writing process. And to be specific, the quantitative method consists of pretest, post-test of both groups and scoring of the three drafts of the experimental group during the implementation of formative assessment. And the qualitative method refers to the employment of classroom observation, interview and analysis of learners' writing data before and after formative feedback. Results from quantitative analysis of T-test show that there is no significant difference in the pretest of both the experimental group and the control group whereas there is significant difference between the post-tests of both groups. Meanwhile, the results of repeated measures ANOVA demonstrate significant difference among the three drafts of the participants in the experimental group. In addition, qualitative studies also show that formative feedback provided by either peers or the teacher is treated seriously by most learners and almost all learners hold more positive attitude towards formative assessment than summative assessment. Therefore, from both quantitative and qualitative analysis, it can be concluded that formative assessment is more effective than summative assessment for the improvement of learners' academic English writing ability and their interest and learning autonomy as well.

Keywords: formative assessment; summative assessment; academic English writing



On the Multi-feedback Assessment of College English Writing

Xu Guozhu

Foreign Languages College of Northwest Normal University

Product approach is commonly used in the instruction of College English Writing in China. This approach, valuing result over process, leads to single feedback as well as summative assessment in writing assessment. Accordingly, this research aims at changing product approach into process approach, and the restricted teacher feedback into the multi-feedback, which includes self-feedback, peer feedback, teacher feedback and computer feedback. Through our study, we found that these forms of feedback play an undeniable role in promoting learners' writing competence, but there are really some defects in the assessment. For example, in the implementation process of self-feedback, teachers offered lots of instruction, but it was still difficult for learners to master the standards of writing assessment, not to mention rethinking of their writing and progress. When it comes to peer feedback, learners are reluctant to accept or even doubt their criticism, which result in a low acceptance rate. Teacher feedback is the most popular method for writing assessment, but it takes teachers much time to do it, and their suggestions and criticism are too general for learners to follow. Besides, they often focus on grammar rather than the techniques of arrangement, and feedback slowly. The computer works effectively and gives quick feedback, but it is hard for it to recognize content inaccuracies. Furthermore, the function of recognizing writing logic and discourse structure should be improved. On the basis of formative assessment, this research use multi-feedback in the instruction of English writing in college in order to leverage their unique capabilities to form a new teaching and assessment model.

大学英语四级写作评分标准探析——评分员的视角

邹绍艳

张晓艺

关晓仙

上海交通大学

上海交通大学

华东师范大学

【摘要】近年来，随着语言测试领域对测试效度和问责制度的密切关注（Chapelle, Enright & Jamieson, 2008），大规模英语考试中使用的评分量表也愈发引起了研究者的兴趣（如 Banerjee, Yan, Chapman, & Elliott, 2015; Hawkey & Shaw, 2005; Knoch, 2009, 2011; 李清华, 2014）。这种“新兴”的研究趋势究其原因主要在于，语言测试领域的学者们广泛认为行为测试中使用的评分量表实质上体现了测试的构念（McNamara, 1996, 2002; Turner, 2000; North, 2003）。因而，检验评分量表的使用属于语言测试构念效度研究的范畴。Bachman (2000)曾指出，考试效度研究是语言测试领域永恒的主题之一。

在这种背景下，本研究聚焦大学英语四级写作测试（以下简称四级写作）的评分标准，通过调查评分员的意见，验证现有的四级写作整体评分量表的效度，探索四级写作评分真正采用的评分标准，以便为下一步构建四级写作分项评分量表奠定基础。与此同时，我们也期望本研究能够从方法上对其它大规模英语考试评分标准的界定产生启示。

本研究要解决的问题如下：第一、四级写作评分员对现有的整体评分量表的意见如何？第二、四级写作的评分标准究竟有哪些？本研究采用问卷调查的方式，调查了179位四级写作评分员的意见。问卷调查收集的数据运用SPSS 19.0软件进行分析，结果表明：第一、涉及到整体评分量表的六道题目的得分都较高，说明总体而言评分员对该量表持比较肯定的态度。但是在六道题目中，关于“现有的整体评分量表对大学英语教学提供的反馈信息”这道题目的得分最低。这说明现有的量表在这方面还有待于改进。第二、评分经验比较丰富和不太丰富的评分员在“评分培训的作用”这道题目上，意见差异比较显著。评分经验不太丰富的评分员对评分培训的作用更加肯定，这也间接反映了现有量表可能在评分培训中的发挥的作用不甚理想。第三、评分员总体上比较认同问卷中十项评分标准的重要性，尤其是语言的准确性、任务完成度、写作规范、内容和思想以及作文长度这五项标准得到的认同度最高。但是，评分经验比较丰富和不太丰富的评分员在语言得体性、任务完成度、写作规范和作文长度这四项标准上也表现出了显著的差异。相比较而言，评分经验比较丰富的评分员对这四项目标准的认同度较低。

总之，上述研究结果初步表明构建四级写作分项评分量表的合理性和必要性。前人的研究已经证明，分项评分量表的优势在于能够为受试者提供更加丰富的反馈信息（Hamp-Lyons, 1986, 1991a; Shaw & Weir, 2007; Weigle, 2002），而且在评分培训中能够帮助评分员快速掌握评分标准（Weigle, 2002; Weir, 1990）。更为重要的是，通过实证研究确定的四级写作评分标准，为下一步分项评分量表的构建奠定了坚实的基础。

【关键词】四级写作评分标准；整体评分量表；分项评分量表；测试效度



Exploring the Rating Criteria of CET-4 Writing from the Raters' Perspective

Zou Shaoyan Zhang Xiaoyi Guan Xiaoxiao
Shanghai Jiao Tong University East China Normal University

Abstract: In recent years, as a response to the growing concerns over test validity and accountability (Chapelle, Enright & Jamieson, 2008), increasing research attention has been focused on the rating scales used in large-scale EFL tests (e.g., Banerjee, Yan, Chapman, & Elliott, 2015; Hawkey & Shaw, 2005; Li, 2014; Knoch, 2009; 2011). Such a seemingly new research trend is attributed largely to the widely-recognized view that rating scales adopted in performance tests represent the de-facto test construct (McNamara, 1996, 2002; Turner, 2000; North, 2003). In this sense, to examine the use of a rating scale falls into the research domain of test validity which, according to Bachman (2000), is one of the eternal themes of language testing field.

Situated in such a context, the present study set its research scope on the rating criteria of CET-4 writing through surveying the raters' opinions. The study aimed at both validating the existing holistic rating scale and exploring the actual rating criteria favored by the raters, thus laying foundations for the construction of an analytic rating scale for CET-4 writing in the next step. Meanwhile, the study was expected to generate methodological implications for the defining of the rating criteria adopted by other large-scale EFL tests.

The research questions were therefore: 1) how well do the raters of CET-4 writing perceive the existing holistic rating scale? 2) what are the de facto rating criteria of CET-4 writing? Survey data were collected from a valid sample of 179 raters. The raters were required to respond to a questionnaire which consists of three parts. In addition to the biodata in the first part, the second part concerns the perception of the existing holistic rating scale adopted by CET-4 writing, whilst the third part mainly taps into the raters' opinions over the importance of some tentatively-defined rating criteria.

The collected data was analyzed using SPSS 19.0. The results showed that: 1) the raters' opinions over the existing rating scale were positive in general with the mean scores achieved through the six questions all above the medium level. Nevertheless, when it comes to the feedback that the existing rating scale offers to the College English teaching, the raters' views were not that positive as can be reflected by the lowest score on this question. 2) there were differences in the perception of the function of rater training in advance to formal rating between more experienced raters and less experienced raters. Compared with more experienced raters, the views of less experienced raters were more positive toward rater training, implying that the existing scale might not have functioned as desired in the rater training process. 3) raters basically agreed on the importance of the ten rating criteria, specifically the five criteria concerning Language accuracy, Task fulfillment, Convention of writing, Content and idea, and Text length. However, more experienced raters and less experienced rater again showed different attitudes towards the importance of Linguistic appropriacy, Task fulfillment, Convention of writing and Text length, with more experienced raters agreeing less willingly on these four criteria than the less experienced raters.

In conclusion, the above findings rationalized the construction of an analytic rating scale for CET-4 writing. Previous studies have confirmed that analytic rating scales are more advantageous in providing more detailed diagnostic information about a test taker's performance (Hamp-Lyons, 1986, 1991a; Shaw & Weir, 2007; Weigle, 2002), as well as in facilitating rater training (Weigle, 2002; Weir, 1990). More importantly, the rating criteria thus determined on an empirical basis has laid a solid foundation for the construction of the analytic rating scale for CET-4 writing.

Key words: the rating criteria of CET-4 writing, holistic rating scale, analytic rating scale, test validity

英语写作测试评分标准模型的建构及其效度研究

——以概要写作评分标准为例

吴雪峰

南京林业大学/上海外国语大学

写作评分标准是评分员判断受试文本质量的主要依据,其效度好坏直接关系到评分质量的高低乃至考试的公平性。根据写作任务输入语及写作要求差异可将写作测试分为两大类别:“独立型写作”与“综合型写作”。本文总结了国内外各类英语考试中上述两大类别写作测试评分标准,探讨评分标准中各评分维度构建的一般规律。在此基础上,本文尝试构建了英语写作能力测试评分模型:独立型写作测试评分标准可分为5个维度,其中内容、结构、语言为构建评分标准时的必选维度,是无论何种形式、体裁的写作测试评分标准均应囊括的一般性维度;交际效果与格式、语境恰当为备选维度,可根据不同写作形式灵活选用。备选维度的参与使评分标准的制定更具动态性、灵活性,考试开发者可根据不同写作任务在必选、备选维度中灵活选择,搭配使用;此外,综合型写作测试评分标准既包含上述“必选”维度及“备选”维度中的有关维度,还应包含与自身相适应的独特评分维度。根据该模型,本文以“概要写作”题型为例,设计了相应的评分标准,并依据多层面Rasch模型对评分标准进行了效度验证。结果表明,总体而言该评分标准具有较好的区分度和效度,但在“语言措辞”评分维度上,个别分数段的使用存在非拟合现象。个别评分员与评分标准各维度存在显著偏性交互作用。在数据分析的基础上,本文对评分标准进行了适当修改,增加了评分档次,调整和修改了相关描述语,并强调应对评分员加强评分前的培训和指导。综上所述,本文提出的写作能力测试评分模型具有较好的效度和一定的推广价值。

Rating scales for testing English writing are the main criteria for the evaluation of the quality of test-takers' compositions. Therefore, the validity of rating scales is directly related to rating quality or even test fairness. Based on the input of writing tasks and various test requirements, we could divide writing tests into 2 kinds, namely independent writing & integrated writing. This paper summarizes and analyzes rating scales of writing tests from English language tests home and abroad, investigating the regularity of the construction of rating dimensions in rating scales. On this basis, this paper tentatively constructs a rating model for tests of English writing: rating scales of independent tests involve 5 dimensions, among which content, structure and language are compulsory dimensions, which must be included as common dimensions in rating scales regardless of the forms or genres of the writing tests. Communicative effects together with considerations of appropriate formats and contexts are optional dimensions and could be chosen flexibly according to various writing patterns. The participation of optional dimensions makes the construction of rating scales more dynamic and flexible. Tests developers could smartly choose among compulsory or optional dimensions and make perfect matches. In addition, rating scales of integrated writing tests not only include the above compulsory dimensions and certain optional dimensions, but also their own unique dimensions that are typical of themselves. According to the proposed rating model, this paper designed rating scales for summary writing as an integrated writing test and validated the scales through application of a Many-Facet Rasch Model. Results indicate that the rating scale designed has good discrimination power and validity but there were bias interactions between raters and rating dimensions. Misfits appeared in terms of the dimension of "language use" with the use of a few scores. Based on data analysis, as modification of the rating scale, more levels were added to the original scale and descriptors were also polished, emphasizing in the meantime that training and guidance for raters should be strengthened. On the whole, the rating model for writing tests has relatively good validity and value of promotion.



A Comparative Study of Holistic Scoring and Analytic Scoring in Writing—Scoring Reliability, Teachers’ and Students’ Perceptions

Ji Xiaoling
Shanghai Jiao Tong University

Weigle (2002) gives a detailed comparison of the pros and cons of holistic and analytic scoring. Despite the apparent merits of the latter, it is seldom employed in large-scale proficiency tests. In two editorials of this year’s *Assessing Writing*, Liz Hamp-Lyons calls for “the end of holistic scoring” and advocates what she calls “multi trait scoring”, believing that the latter “can make sense to teachers and to test agencies”. The present study compares two analytic scoring and one holistic scoring, examining their scoring reliability and teachers’ and students’ perceptions of them. The two analytic rating methods are Jacobs, Zingraf, Warmuth, Hartfiel & Hughey’s (1981) five-dimension scale and a three-dimension scale (content, organization and language) employed by the researcher in writing instruction, and the holistic approach is the rating scale of CET writing. Three raters will each rate 30 essays with these three different methods, and their scoring reliability will be examined. A questionnaire survey will also be administered to the three raters and students regarding their perceptions of the three methods.

A Longitudinal Study of PRETCO-Oral Rater Effects

Yang Zhiqiang Quan Dong
Chongqing University of Science & Technology

Abstract: This paper analyzes the recent 5 times’ rating results of PRETCO-Oral over the last three years. Raters’ effects have been investigated longitudinally in terms of such 5 perspectives as leniency/severity, central tendency, randomness, halo effects and differential leniency /severity. The results show that there are significant differences in raters’ leniency/severity in all occasions of rating, which, however, exerts little influence upon the overall rating quality; No sign of evident central tendency, randomness, halo effects and differential leniency /severity has been detected on the whole; while the middle rating category of “Reading” has been overused; Several raters show randomness and halo effects while adopting certain rating categories; There is significant rater bias in the interaction of raters and the four traits of tasks. At last, some implications on the rating of PRETCO-Oral and raters’ training are put forward.

Key words: PRETCO-oral; rater effects; longitudinal study

Rating Fluency: An Inquiry into How far Perceptions of Fluency Align with Quantifiable Measures

Ellen Head
British Council Shanghai

Fluency is a widely accepted component of rating the performance of candidates in oral tests of foreign language ability, but it is a composite and to some extent ambiguous term. In this paper the researcher will report on a preliminary study of perceptions of fluency based on piloting a replication of a research project by Cecilia Dore. The latter study starts from the analytical model which divides fluency into utterance fluency, cognitive fluency and perceived fluency, made by Segalowic (2010). In Dore's study, participants were asked to rate three speech samples and identify the factors which they took into consideration when rating. As a follow up they also ranked twenty variables which were possible components of fluency, derived from a focus group discussion.

Dore's study raises interesting questions about the relative significance of temporal and non-temporal variables in relation to fluency. There are also issues of nomenclature. For example, pronunciation seems to play some role in perceived fluency, as do cohesion, lexical and grammatical complexity. This raises questions in terms of rating scales which may duplicate and thus over-emphasize the aforementioned elements.

The current researcher, Head is at the pilot stage of what is hoped to be a year-long collaborative study with academic partners in Asia, which may eventually look at threshold levels in terms of perceived fluency, temporal variables and complexity. The current paper will be limited in scope, reporting on the first stages of replicating Dore's study with China-based language teachers including some with experience of rating high-stakes exams such as FCE and IELTS. We will look at data derived from written answers to the same open-ended questions posed by Dore; what is fluency? What are the causes of disfluency? To what extent is fluency subjective?

Earlier studies including Derwing et al. (2006), who investigated untrained raters listening to low-level students, will be drawn on to add insight into quantifiable variables of fluency such as speech-rate and pitch. The impact of coherence on fluency ratings, studied by Iwashita et al. will also form part of the context for analysing how raters reacted to the questions both with and without priming by listening to speech samples. Finally, rating scales such as the CEFR and IELTS will be compared with the components ranked as important by Dore's respondents and those in China, and the implications discussed.



英语专业本科生对EFL学术写作的构念认知困难解读

赵冠芳 吕云鹤 刘子仪

上海外国语大学

随着学术交流的国际化,英语已经成为学术界的通用交流语言(Tang, 2012)。正是因此,随着我国高等教育英语教学改革的不深入,学术英语教学与实践逐渐开始出现在高等教育的各阶段。在这一系列教学与实践,写作教学与训练无疑是重要的一环,它对学生的英语能力和学术能力均提出了更高的要求。尽管学界对中国学生EFL学术写作问题已有所研究,但鲜有研究关注学生对学术写作这一构念的认知以及这种认知对写作实践的影响。同时,现存文献中的学术写作相关研究几乎都停留在硕士和博士层面,忽视了本科生的学术写作发展研究。然而,已有学者指出,我国学生本科阶段的学术写作能力低下已经“直接影响了硕士甚至博士阶段的教学和研究水平”(张冲2010:311)。因此,本研究根据Snow & Uccelli (2009)的学术英语模型,从语言技能、体裁知识、论证思辨策略和专业知识四个方面,对178名英语专业本科生进行了问卷调查,并对6名本科生进行了半开放式的深度访谈。通过对问卷调查中衡量认知程度和困难程度的两个里克特量表的定量分析以及对访谈和开放式问题的定性分析,本研究总结了英语专业本科生对于英语学术写作这一构念的认知,并探讨了这种认知与学生写作实践中遇到的困难间的关系,具体分析了哪些困难是由于本科生对学术写作本质的认知不足造成的,哪些困难是由于英语语言能力不足造成的。研究结果也对本科学术写作课程和教学提供了借鉴和指导。

An Empirical Study on the Effect of Writing Prompts on Tertiary Students' Writing Process and Performance

Ge Xiaohua

Renmin University of China

Various variables of the task that have been empirically found to affect a writer's test score. However, the effect of tasks with different prompts on the writing performance of the test takers has not been widely conducted yet. Situation-based writing task and outline-based writing task are chosen in this study to examine the effects of different tasks on the tertiary non-English majors' writing performance. Both quantitative and qualitative research methods are used: the verbal protocol, the analysis of the textual features of the writing products and rater's holistic rating and assessment on 12 indicators of the product quality. The study shows that different writing tasks do have impact on students' writing process and product. As for the writing process, when students work on the situation-based writing task, they usually form their own opinion in the preparatory stage and organize their essay closely related to the main theme; while they work on the outline-based task, they are declined to extend the Chinese outline without a clear theme in their minds and can't express their own idea adequately and convincingly. As for the writing products, different writing tasks have significant effects on the place of the theme and the manner of ending the essays; and also some indicators of the product quality, such as "whether the theme is adequately addressed", "conscious of the reader", and "the conclusion is based on the previous discussion", etc. However, different tasks have no significant effect upon the test takers' holistic scores as well as the linguistic features such as "cohesive devices", "lexical and grammar" and "varieties of expressions". This study has significance in both writing tests and teaching by providing us with some insights of choosing appropriate writing tasks.

An Investigation of Rater Bias Patterns in a Large-scale EFL Writing Assessment

Sun Youxia

Zhejiang University

The present study employed multi-faceted Rasch measurement (MFRM) to explore the rater bias patterns of EFL raters when they rate EFL essays. A total of 30 raters from different areas of China were involved, and a total of 846 essays of four different prompts were rated with each prompt being used in certain areas. The essays were assessed using the analytic rating criteria (Content, Organization, and Language). MFRM suggested several recurring bias patterns among rater subgroups. In rating difficulty measurement, Content was generally considered more difficult or harshly-scored, Organization tended to be easier or leniently-scored and Language almost leveled off in difficulty or severity. In rater-examinee bias interactions, there did emerge some clear tendencies in the overall bias patterns: a) there were more rater-by-examinee bias interactions involving high-ability test takers than low-ability test takers; b) severe raters tended to show more bias interactions with test takers than lenient raters; c) severe raters were more likely to have severe bias and lenient raters tend to have lenient bias toward high- and low-ability test takers; d) there was also a tendency for severe raters to show more severe bias and for lenient raters to show more lenient bias toward test takers at the ends of the ability scale; and e) higher- and lower-ability test takers attracted more bias interactions with severe raters than those with lenient raters. This study has implications for issues of rater training in EFL writing assessment.



A Comparative Study on the Use Of Pictures in Designing Picture-Based Writing Test Tasks

He Qiong
Shanghai Jiao Tong University

The advocacy and development of multimodality and multiliteracies in China entail wide application of various visuals, pictures included. Indeed, pictures have been in wide use in language pedagogy and assessment, from giving young learners pictures to tell a story in Spoken English practice to requiring test-takers to describe what's in pictures. Like language, pictures as a social semiotic have meaning potentials as well. Yet, pictures in second language writing are put under the dimension of stimulus in task design, receiving little attention from researchers.

This author propounds the concept of picture-based writing, intending to examine the comparability of pictures as the sole focus of a writing task with those playing a supportive role in a writing task. Given a wide variety of pictures employed in writing different modes of discourse, this study focuses on cartoons used in argumentative writing wherein distinct messages can be conveyed and received unequivocally.

Drawing on both an authentic CET Band 6 writing task and the task format of Graduate Admission English Test, two writing tasks using the same picture are designed, one requiring students to write based on their interpretation of a picture (Task 1) and the other to write based on a specified rubric accompanied by a picture (Task 3). A pre/post-writing questionnaire and one in-class writing test (using either Task 1 or Task 3) are used with a sample of 247 college freshmen. It found that most students with different English proficiency perceived picture-based writing tasks favorably although overtly concerned about the picture interpretation. Four major benefits of using pictures have been identified. Students' performance on picture-based writing task is also found to be closely related to that on non-picture writing tasks. To both high-achieving and low-achieving students, their performances on the two writing tasks where pictures play different roles are not significantly different. It means, despite Task 1 incurring greater worries over deviation from the topic than Task 3, the two tasks are not as differentiating in terms of difficulty as expected. It is thus concluded that a clear picture with unambiguous messages and a clear rubric with a specified idea are equally important for designing picture-based writing tasks. Due to the limited time, this exploratory study is merely quantitative examining the effects of the two picture-based writing tasks on students' writing product. The author believes that their impacts on writing process can make a promising direction for future research.

Linguistic Features of Picture-prompted Writing—Differences by Three Caption Types

Shi Yali
Zhejiang University

Abstract: The study investigates the influence of caption types (absent, content and context) in the picture prompt on the linguistic features of writing quality in terms of Lexical Richness and Syntactic Complexity, and the relationship of those linguistic features to the writing quality. Altogether 154 participants took part in the study and the data was analyzed by means of web-based L2 Syntactic and Lexical Complexity Analyzer (Lu, 2010). Results show that: 1) 9 among 14 indices of Syntactic Complexity (Mean length of sentences, Mean length of T-unit, Mean length of clause, Clause per sentence, Verb phrase per t-unit, Clause per t-unit, Coordinate phrase per t-unit, Complex nominal per t-unit and Complex nominal per clause) exhibit significantly higher values in writing products with context caption than with other two caption types respectively; 2) 7 among 25 indices of Lexical Richness (Verb sophistication II, Corrected verb sophistication I, Number of different words, Corrected TTR, Root TTR, Squared verb variation I, Corrected verb variation I) exhibit significantly higher values in writing products with context caption than with other two caption types respectively; 3) there is no correlation between writing scores with any caption types and any index of Syntactic Complexity; 4) Lexical Richness indices such as Number of different words and NDW in first 50 words account for 29.2% of writing score with content caption; Number of different words accounts for 35.6% of writing score with no caption; Number of different words and Lexical density accounts for 28.6% of writing score with context caption. Finally, implications for writing instruction and assessment are discussed.

Key words: picture-prompted writing, caption, syntactic complexity, lexical richness



The Influence of Interlocutor Proficiency in a Paired Oral Test

Lv Zhouyang
Zhejiang University

Performance-based second-language oral tests have become increasingly prevalent in recent decades, due to the enhanced interpretability of test scores, greater theoretical and construct validity (Kane, et al., 1999) and positive washback of such assessment tools (Bonk & Ockey, 2003). The paired or small group oral testing, in which two or more test takers interact with one another but are observed and assessed as individuals, has emerged as an alternative to the traditional one-on-one oral proficiency interview (OPI). A number of advantages have been reported, which make it a potentially viable and attractive solution to the problem of the previous OPI format. First, it is time-efficient and cost-effective, since it reduces the assessment workload of the raters (Folland and Robertson, 1976; Hilsdon, 1995). Besides, it avoids several of the criticisms associated with the OPI, such as the asymmetric discourse structure of the interaction between the interviewer and test takers (Van Lier, 1989), and the marked effect of the interviewer on test-taker performance (O'Sullivan, 2000).

However, despite the potential benefits, the paired oral test is in fact easily subject to the plausible threats to its validity due to the complexity of interaction between paired test takers. The issue of interlocutor effect has received much attention in the past several years both from a conceptual standpoint as well as in empirical studies (Davis, 2009). A couple of studies have examined the interlocutor effect of language proficiency, although the research findings are mixed and somewhat contradictory. While the previous research puts great emphasis on the interlocutor effects on the score (Iwashita, 1996), the interaction pattern (Davis, 2009; Galaczi, 2008), the identity constructed in the talk (Lazaraton & Davis, 2008), the current study is more focused on the individual discourse.

The purpose of this article is to explore whether interlocutor proficiency level influences the performance of the test takers. Specifically, the article aims to investigate whether the test-takers' scores and linguistic performance as indicated by the levels of complexity, accuracy and fluency (CAF) are affected by the interlocutor proficiency. The study will include a total of 40 EFL learners, who are sophomores in one of the major universities in China. The students are divided into groups of relatively high and low English proficiency according to their self-reported TOFEL iBT speaking scores. They are tested once with a partner of similar proficiency and once with a partner of higher or lower proficiency on an oral discussion task. Two prompts adopted in this study are selected from the CET-SET and largely comparable. Participants are scored by four raters on the rating scale of CET-SET, a scale of 2–5 in the sub-categories of accuracy and range, size and discourse management, flexibility and appropriacy. Multi-faceted Rasch analysis is used to examine the influence of interlocutor proficiency on scores, while linguistic analysis is adopted to investigate the influence of interlocutor on the discourse elicited. The implications of the findings are addressed in terms of recommendations for test developers and users who desire to change procedures for grouping test takers.

人工智能技术及其在外语测试领域中的应用

汪张龙

科大讯飞股份有限公司教育产品事业部副总经理

2016年初，一场全球瞩目的人机围棋大战，引爆了“人工智能”话题。AlphaGo（“阿尔法围棋”）的获胜，反映出人工智能技术迅猛发展的强劲势头。“人工智能”概念自1956达特茅斯会议首次提出，距今已走过六十一年发展历程。在经历了两次起伏之后，随着计算机图像识别、语音识别等技术的日趋成熟，人工智能技术已进入真正爆发的前夜，全球互联网巨头也纷纷加大在人工智能领域布局。有专家预言，人工智能技术的崛起必将引领新一轮的产业创新与变革。“人工智能”也因此上升为各个国家的国家战略。日前，中共中央、国务院印发的《国家创新驱动发展战略纲要》中，“发展新一代信息技术”在各项战略任务中居于首位，要把数字化、网络化、智能化、绿色化作为提升产业竞争力的技术基点，加强类人智能等技术研究。科大讯飞在主要人工智能技术领域具有长期技术积累，中英文语音合成、语音识别、手写识别和评测技术处于全球领先，并在语言测试领域取得多项重大应用：

1. 基于手写识别技术和中英文作文智能评分技术，在大学英语、高考作文阅卷研究中，智能阅卷系统阅卷水平已经达到或超过阅卷专家水平，并且对抄袭、套作等情况进行反馈，对改进现有人工阅卷效果、提升阅卷效率、保障考试公平公正具有重要意义。

2. 基于语音识别和中英文口语自动化评分技术，在中高考、大学英语（CET）、中高考、普通话测试（PSC）、民族汉考（MHK）等各类口语考试中进行研究和应用，年测试考生量超过700万人次，解决了语言听说考试组织难、阅卷难的问题。为科学、公正开展语言能力测评提供了技术保障。

3. 通过大数据挖掘和机器学习的方法，通过大量试题数据的训练，与人工的参数标注相结合，训练出一套语言测试题难度预测系统，在阅读理解等题型上准确度超过专家预测。这一技术为国家题库建设和英语一年多考，解决试测、等值等实践难题。



纸笔到计算机化考试转变过程中数据库与 计算机技术的应用

刘 洋

教育部考试中心外语测评处

【摘 要】本文梳理了纸笔考试与计算机化考试的从命题到考试的基本流程，在此基础上分析了计算机数据库的应用对考试流程的影响。原有的纸笔考试从命题端到成绩报告端是线性的串行过程，而以数据库为核心的计算机化考试则使考试报名、命题和成绩报告等流程可以依赖统一的标准，并使流程并行立体化，从而提高考试数据的分析报告以及考试实施的效率。与此同时纸笔到机考的转变也面临着需要克服的困难，而且相较纸笔考试，机考也存在着相应的不足。

【关键词】纸笔考试 计算机化考试 数据库 成绩报告

Innovations in Online Language Testing: Adaptivity and Speech Recognition

Lance Knowles

Liulishuo

A major benefit of the technology revolution in English Language teaching is the ability to employ adaptivity and a variety of interactive tasks, including speech recognition, in tests. This presentation shows how these innovations have been used to develop an online placement test.

With regard to adaptivity, we have developed a variable length test that has cut the time necessary to assess student proficiency level with little or no reduction in accuracy, especially considering time requirements. The basic principle is obvious: low-level students require less time to assess than higher level students. In constructing the test, item banks and testing paths are determined by algorithms that decide when enough items have been completed to end the test.

Guiding students to the appropriate item banks to determine proficiency level is the main challenge because of the need to construct test items that differentiate proficiency levels. It's important to note that a "Catch 22" situation arises when one uses existing test results as a standard for defining proficiency levels and test items. From our experience, one should not assume that existing tests are accurate or have sufficient validity. Therefore, an incremental approach to item bank construction and testing path design is employed.

Furthermore, a learning theory that defines the relationships between various language patterns and language skills is an important tool. Otherwise, test items are constructed based on a language framework that may not be appropriate. For example, a framework based on concepts and language chunking is quite different from a framework based on situations, grammar rules or vocabulary. The interplay between these different dimensions is important and needs to be optimized in a well-constructed language test.

Another important innovation is the ability to use speech recognition technology to assess pronunciation and oral fluency. Once again, we use a variable-length test to score a student's ability to pronounce key phonemic patterns as well as their ability to chunk language in working memory as a measure of fluency. We believe that chunking ability is a necessary, though not sufficient, condition for fluency, and is absent from most tests.

Designing effective test items requires an understanding of chunking and the recursive nature of language. The length and complexity of language chunks which learners can process is a good indicator of proficiency. Of course, it is also important to recognize the interplay between pronunciation levels and the efficacy of items designed to assess chunking ability.

The speech recognition technology that we are using has proven to have an excellent ability to score pronunciation while also being robust with regard to technical issues such as microphone and ambient noise levels. This is essential for practical use.



BEC中级口语考试考官和考生问卷调查对比分析

杨宏波 虞程远 辜向东
重庆大学

【摘要】本研究就BEC中级口语考试的反拨效应对35名考官和99名考生进行问卷了调查，并辅以访谈。主要从考官及考生的背景信息、考官和考生对BEC中级口语考试的认识和考试的影响三方面进行分析。研究表明，考官任职资历较高，年龄结构多样化，考生以在读本科生为主，专业涵盖面广。关于考试的认识，考官和考生都认为BEC口试整体时长及各部分时长比较合理，BEC中级口试题型，口试内容和口试效果总体上都具有真实性。但是，考官与考生对考试时长和考试题型的真实性的看法存在明显差异。与考生相比，考官认为BEC口试第二部分Mini-presentation的时长更合理。与考官相比，更多考生认为BEC口试第一部分Conversation这一题型能考察出职场英语水平，具有真实性。此外，本研究还分别调查了考官对口试评分标准的认识以及考生对考试难度和考试熟悉程度的认识。考官认为BEC中级口试的整体评分标准和分项评分标准总体比较合理。考生认为BEC中级口试难度总体上比较合理。考生对口试流程比较熟悉，对BEC口试评分标准的了解程度并不高。关于考试影响，总体上，BEC中级考试对学生口语学习和应用以及对考官个人都产生了比较正面的影响。此外，本研究还从考官、考生、考试组织实施、语言及非语言因素方面来探讨考试的影响因素。分析中也发现了一些问题，如考生在BEC中级口语备考的时间投入不足。学生备考活动以个人陈述为主，互动式的备考活动如与老师或同伴进行会话练习的频率较低。考生在BEC口试备考内容，备考方法等方面缺乏有效的指导等。

【关键词】BEC中级，口语考试，反拨效应，考官，考生

剑桥商务英语考试的反拨效应机制 ——基于结构方程建模的研究及其启示

钟瑜 辜向东 肖巍
重庆大学

【摘要】反拨效应指考试对教和学的影响。现有的反拨效应研究多集中于通用英语考试（如高考、CET），而对商务英语类考试的研究不多。对商务英语考试的反拨效应研究有助于推动商务英语教和学的优化改革，也利于拓展反拨效应模型的应用范围，检验现有模型在专门用途英语考试中的普遍适用性，具有重要意义。

为此，本文通过问卷调查，在四个考点搜集了829名剑桥商务英语考试中级考生的数据，并采用结构方程建模分析考试设计和考试使用方式对备考行为的影响，以探究剑桥商务英语考试的反拨效应机制。

基于问卷数据，本文发现：（1）考试设计对应试技巧训练有较弱的正面影响，但对语言技能发展几乎没有影响；（2）内在型的考试使用方式对语言技能发展和应试技巧训练有微弱的正面影响；（3）工具型的考试使用方式对语言技能发展和应试技巧训练有较弱的正面影响，但比内在型的影响更大。

基于研究结果，本文对剑桥商务英语考试反拨效应机制内部各因素之间的关系进行了深入讨论，并对商务英语的教和学提出了一些建议。同时，本文对于使用新技术挖掘反拨效应研究数据也提供了一定启示。

【关键词】剑桥商务英语考试，反拨效应机制，结构方程建模



One Billion Customers: Lessons on Application of Artificial Intelligence from the Front Lines of Language Testing in China

You Zhonghui Chen Guangbin

Imagine you are the Program Director for a college entrance exam called “SATACT.” As you are reviewing plans one day, your boss enters your office to let you know you must add a section to evaluate Chinese language skills. This new section must include scoring multiple key-entered constructed responses as well as spoken responses. As the boss is leaving the room, she mentions that the volumes will triple from approximately 3 million to more than 10 million candidates annually.

This sounds like the ultimate challenge for an assessment expert, right? What program administrator wouldn’t want to have to come up with human raters for spoken and written response for a foreign language where volumes of responses will exceed 100 million! The reality is, this is not just a hypothetical challenge. This is a real, live, operational, assessment program. This is the College English Test (CET), China’s college level English language examination. This is among the largest assessment programs in the world and brings with it challenges of scale not seen elsewhere.

This session will discuss the current state of testing in China. We will place a particular emphasis on the CET and how the application of Artificial Intelligence will one day soon change the delivery of these exams as well as discussing the challenges and rewards faced by two partners, one Chinese and one American, working together to address the opportunity. We will demonstrate how innovations in Artificial Intelligence we automate the processes used to score the CET and the opportunities for the big data collected in the process to drive educational outcomes for students.

The People's Republic of China (PRC), is the world's most populous nation, with a population of nearly 1.4 Billion. The country is governed by its vanguard party based in the capital of Beijing and the CET is uniformly administered by the 22 provinces, five autonomous regions, four direct-controlled municipalities (Beijing, Tianjin, Shanghai, and Chongqing), and two special administrative regions (Hong Kong and Macau). This centralized control presents a substantial opportunity for efficiency in the administration of the assessment. Similarly the distributed administration creates the challenge for the provincial governments to scale-up sufficient human-raters for the spoken and written responses in English.

We will review several challenges related to the cultural, operational and business perspectives of bringing together two companies. We will review the application of automation using Artificial Intelligence for the constructed response scoring in both English and Chinese and discuss the automation of spoken-response scoring. Additionally, the opportunities presented by the scale and big-data collection will be reviewed to understand the impact on teaching and learning. The lessons learned and the opportunities gained or lost along the way will highlight the tremendous power of AI in the China market and the tremendous benefits for Chinese education looking to leverage this innovative opportunity.

基于语料库的商务术语使用与商务写作质量的相关性研究

葛诗利 贾清
广东外语外贸大学

【摘 要】自动作文评分中重要的是特征选取。商务英语写作中术语的使用是必不可少的，那么术语与写作质量或成绩的相关度高低，就决定了术语可否成为商务英语写作自动评价的特征。我们基于词典构建了商务英语术语列表，对商务英语写作用文本语料库进行自动标注，分析了商务英语写作用文本成绩与其中术语使用的相关程度。统计结果表明，商务英语写作成绩与术语使用总体数量呈中等强度的相关关系，不同类型与领域的商务英语写作与相应类型的术语使用数量存在更强的相关关系。这说明了商务术语应该成为商务英语写作自动评分的一个特征，尤其是特定领域术语对相应类型的商务写作成绩预测力更强。



学术英语阅读测试的信效度检验

高霄 高媛
河北经贸大学

学术英语（English for Academic Purposes）相关研究日益引起关注。相关研究紧紧围绕“学术”内涵、课程开发、教学范式、教材建设、师资发展及测试评估等方面做着有意义的尝试，取得丰硕成果。

本研究聚焦学术英语阅读，旨在验证阅读测试的信度与效度。研究从学术语言概念入手，综合考量文献关于学术英语概念的界定，参考John Slaght（2012，2014），根据学术英语阅读课程培养目标，提出学术英语阅读课程学术语言的操作化定义。学术英语阅读语言水平指学生用英语搜索相关资料和文献的能力，分析、综合、评价和运用各种信息的能力，提出问题、并通过思考和推理以解决问题的能力等（Slaght，2012，2016）。上述能力的高低表征为以下学术技能水平的高低：识别文本思想、总结主旨大意、辨别文本细节、基于语境的语义推理、分析文本结构、阐明逻辑推理过程、创建思维导图、区别观点与事实、区别论点与论据、识别段落作用和鉴赏学术语言。以上技能也就成为旨在测评学术阅读语言水平的观测维度或指标。

被试共168名，包括81名硕士研究生和87名本科生，来自某省属财经类骨干院校。被试开设《学术英语阅读》课程，持续一个学期。测试文本长度约为1000词，题目为“Has China Outgrown the One-Child Policy?”，选自国外主流网络文本。要求完成6项任务：概括文本大意（10分）、分析文本结构或阐明文本逻辑推理过程（10分）、创建思维导图（20分）、区分分论点与支撑性细节（20分）、思辨性阅读（要求被试从内容与形式两个层面分析文本的优点与不足）（20分）、分析书面语的正式性（formality）、弱化语表达（hedging expressions）和路标语使用（signposting expression）等学术语言的运用及其效果（20分）。

信度检验包括评分员信度与内在一致性，效度检验从内容效度、构念效度、效标效度、结构效度和反拨效度五个维度进行分析。研究结果发现，阅读测试所考查的六个任务能够反映有效测量被试学术英语阅读水平，各项统计指标显示，试卷具有可靠的信度与效度。

信息化时代专门用途英语测试与效验框架研究

方秀才

淮南师范学院

交际语言能力框架是语言测试进入第三个阶段即“整体-社会语言学阶段”(Spolsky, 1978)之后的主要指导框架,代表性的有Bachman (1990)的交际语言能力模型和Bachman & Palmer (2010)评测使用论证框架。对于专门用途英语能力测试,早在1987年,Hutchinson & Waters (1987)就认为这类测试非常重要,因为它评价的是考生完成某项任务或者工作的能力,而不是普通的英语能力;Douglas (2000)认为专门用途英语测试存在的理据有三:语言使用随语境变化;专门用途英语的表达更为精确;专门用途英语包含专业知识。他主张通过分析目标语使用(TLU)情形来界定测试内容、方法和测试行为评价标准。但是,无论是语言习得还是语言测试研究领域在最近三十年都没有系统论证专门用途英语能力内涵,也没有为该能力的测试和效验提供明确的可操作框架。

本研究基于信息化时代专门用途英语现状分析,提出专门用途英语能力模型主要包含以下维度:英语能力、专门领域知识、跨文化交际能力、批判性思维能力、自主学习能力和信息素养;专门用途英语能力测试需从内容维度(即外语能力和专门领域知识)、情境维度(教育信息化、英语国际化、经济全球化)、测试属性(社会水平考试和教学检查类考试)和评估方式(形成性vs. 终结性、全国统一考试vs. “个性化、自主化、信息化、多样化”(蔡基刚, 2015))等四个方面细化界定;效度验证借鉴Bachman & Palmer (2010)评测使用论证(AUA)框架的根本理念“依赖证据、通过论证”分步骤有序开展,并在每个环节重视信息化时代特征对于证据类型和收集方式的影响。



基于语言能力发展的ESP课程测试研究

周淑莉
南京林业大学

【摘要】语言测试伴随着语言教学而出现，并伴随着语言教学的发展而发展。随着ESP（English for Specific Purpose）教学的深入开展，ESP测试越来越重要。如何测试ESP课程中学生的英语语言能力已成为国内语言测试界非常关注的问题。本文基于需求分析和ESP测试理论，从学生语言能力发展的维度探讨分析以下几个问题：（1）ESP中语言能力是如何界定的？ESP语言能力与通用英语能力的区别有哪些？（2）专门用途英语试卷设计中如何体现ESP特点的语言能力的各个语言技能？（3）专门用途英语测试的受试者语言能力与测试任务之间如何实现互动？（4）ESP测试的分数怎么用来解释受试者在真实语言环境中使用语言的能力？本文通过对以上问题的探讨，以期对ESP测试有较为深入的了解。但相对于ESP教学而言，ESP测试在国内的研究才刚刚起步，因此，这方面的研究与探索则显得格外重要。作者希望通过本次探讨能为后来者探索ESP测试提供参考性价值。

Multi-modal Analysis of Interviewers' Pragmatic Identity in English Oral Testing

Zhang Cong
Hohai University

Abstract: Based on the Pragmatic Identity Theory, this paper discusses, by analyzing the video data of BEC, the identity construction of the interviewers in the interactive process, i.e. different identities created by the interviewer, multi-modal means of identity construction, and the motivation to create different identities in BEC oral tests. It is found that, in order to encourage the candidates to put themselves in the best possible conditions and to obey the principle of fairness and objectivity in the interview, the interviewers not only construct the identities of professionally authoritative interviewers but also the identities of servants, coordinators, monitors, managers, motivators, etc. via a multi-modal communication with candidates (verbal and non-verbal) in an institutional discourse context of English Oral Tests. This study sheds light on the interviewer training..

Key words: pragmatic identity; multi-modal analysis ; interviewer ; oral testing

Sunday Sessions

Room 101

Using Rasch Modeling to Analyze MHK Examinees Data of Different Years

Ling Yuyu Chen Hui
Beijing Language and Culture University

Abstract: Rasch Modeling is a latent trait model which has drawn international interest among researchers. By logarithmic transformation, this model put the person and item parameters on the same scale to make it convenient for parameter comparison between the person and person, item and item, person and item.

Taking the data of MHK examinees in different years as samples, in this research, through equivalent design of common problems, with the application of Rasch Modeling to transfer scores of examination to a same system to analyze the variety of academic performance of examinees, the ground was provided support for assessing academic performance.

Practical Considerations for Developing Item Banks in China

Zhiming Yang
National Education Examination Authority (NEEA)

Abstract: An item bank is a collection of test items accessible for use during exam preparation (Ward & Murray-Ward, 1994). It is a very powerful tool for item entry and storage, which allows items to be easily retrieved for review, edit, revision, and test form assembly. With item banks, item parameter estimates derived from any field testing or operational work can be easily maintained, the quality of any new test form is assured in terms of content validity, reliability, form difficulty and discrimination, and the working efficiency can be improved dramatically.

Developing Item banks in China, however, presents some special challenges. For example, for any high-stake tests in China, pilot or field testing for new items may result in an item exposure issue. The standard ways of pre-testing new items in the West, such as embedding a few new items into operational forms, are not doable in China. This is because all items are expected to be released after administration. In addition, scaled scores are typically not used in most testing programs. Both experts and the public prefer constructed-response items to multiple-choice items, resulting in relatively large equating errors and scoring errors. In fact, most item banks in China have no item parameter estimates, equating design, or equating. This paper discusses some practical considerations for developing item banks in China in an effort to begin solving these issues. Some guidelines on improving item bank quality and efficiency are provided.

References: Ward, A. W. & Murray-Ward, M. (1994). Guidelines for the Development of Item Banks. Educational Measurement: Issues and Practice, Spring



Automated Scoring of Retelling Proficiency in the Oral Test of MHK (the Fourth Level)-Practicability and Reliability Study

Wang Jing Wei Liyan
Beijing Language and Culture University

Abstract: The retelling proficiency mainly examined the students' comprehensive ability, including obtain key information, language, logic and etc. As the first oral test in MHK(the fourth level) is mainly used to examine the language ability of minority college graduates. Currently, the test mentioned above use of two teachers and one expert scoring online that we called "2+1" mode. Although the score quality has been guaranteed, it still be affected by costing a great deal of time, low efficiency and raters factor.

Along with the computer technology developing rapidly and the artificial intelligence breaking a new ground. It's possible to solve the difficult questions of the retelling proficiency of oral test in MHK(the fourth level) and to achieve the automated scoring. Here are the reasons: 1) the retelling proficiency of oral test belongs to the semi-open tests, with clear themes, certain contents and easy extract Eigen values. 2) the rapid development of computer technology for automated scoring provide a strong support. 3) technical support team developed a multi-lingual intelligent oral evaluation system, relying on speech recognition, artificial intelligence and etc have already accumulated abundant experience in multilingual assessment exams.

In this study, we use the measured data form MHK(Xinxiang) which tested in May 2016 as a sample, use the retelling proficiency of oral test in MHK(the fourth level) automated scoring system did the research and analyzed the results. The results showed that five experts' correlation between average 0.55 or more; five experts and the system correlation score in a minimum of 0.633, up to 0.682; the average score of experts and system reaches 0.859. In addition, we verify the reliability and stability of automated scoring system by experts manual scoring, huge differences and etc. The study shows that the indexes of automated scoring system are exceeded ordinary rates and achieved our desired results. If the system can optimize in a further step, get more reasonable score distribute, to replace one rate with automated scoring system and to achieve human-machine combination just around the corner.

Key Words: MHK Retelling proficiency Automated scoring Practicability Reliability

Research on the automatic scoring validity of subjective question in MHK (level III)

Wang Yan

Beijing language and culture university/student

Abstract: Subjective questions are widely used in the oral test because of its high validity. At present, with the rapid development of computer technology and measurement technology, the score of the oral test technology has improved significantly. There are three types of test in the MHK(level III), reading test has been completely implemented in computer automatic marking, closeness question has been realized by combining computer automatic scoring and human scoring and the scoring reliability exceeds the humans. However, the automatic scoring technology for subjective question is still in the stage of exploratory.

Subjective question in MHK oral tests of scale have for now been graded by computers two times, which lays foundation for justifying the validity of automatic grading of subjective questions in oral tests by computers. Guided by the assessment framework for automatic grading by computers proposed by Xi (2012), the validity of automatic grading by computers will be verified and illustrates in five steps, namely explanation, assessment, extrapolation, generalization and application; representation on constructs in oral tests of indexes in automatic grading system by computers will be analyzed through comparison between machine grading process and human grading process and statistical analysis of relevance, consistency, rigor and dispersion between machine grading results and human grading results based on classic measurement theory. Comparison on objective questions between machine grading results and human grading results will be conducted as well. Test results of students from a school in Xinjiang taking MHK oral tests (Level III) will be selected as samples for comparison between oral test results and final exam test results. To lower errors of machine grading results and reduce differences between machine grading results and human grading results, data and texts of voice records from re-evaluated students will be analyzed to explore the relation between difficulty level of questions and re-evaluation rate. Different types of untypical answers (such as answers irrelevant to questions or reciting of model essays) by automatic grading system by computers will be particularly verified as well. Limitations of automatic grading system by computer will be thus concluded for further improvement.

Keywords: Subjective question of oral test Automatic scoring Validity



Criterion Evidence of Toulmin's Argument Model for MHK Level 3 Oral Test

Zhang Jian Zhou Chenglin Ren Jie Hong Run
Beijing Language and Culture University

Abstract: Test validation of Toulmin's Argument Model is a systematized process which is from examinee's performance and to test use. A series of arguments are included and the terminal point of the former argument is the start of next argument, which makes the explanation of scores based on test validation more reasonable. The test's external criterion evidence is mainly applied on the extrapolation state. This article, first of all, based on the argument of test validation of Toulmin's Argument Model, combined with New TOFEL, introduces the framework of test validation of Toulmin's Argument Model.

In the second step, a small descriptive corpus is built based on 10 pieces of language ability rating scale. Depended on the standard scoring of MHK level 3 oral test and intuitive method, we will make a descriptive analysis of "what can they do" the examinees whose mandarin ability reach to MHK Level 3. That analysis include situation & model, language performance and communication strategy. On that condition, the elementary scoring scale for oral test of MHK level 3 is established.

After adopting experts' suggestion of revision, the scoring scale is converted to a questionnaire (55 applied descriptions are included). The scoring formula for this questionnaire is the Likert scale which has 5 ranks.

In the third step, we will train the front- line teachers from a certain school of the scoring method and standard. After that, the trained teachers are required to give a rank (both whole ability rank and each single item rank) to every student in their classes of their oral ability based on the questionnaire. The assessment from teachers will be the criterion for calculating the correlation between examinees' score of MHK oral test and teacher's assessment.

In the last step, the test validation of extrapolation state of MHK from empirical perspective would be done with the correlation result for backing and Toulmin's Argument Model for framework. The results reveal that the test validation of extrapolation state of MHK level 3 is basically reasonable, the scores of examinees' in MHK level 3 oral test can reflect their oral performance in practical mandarin communication.

Key words: test validation; MHK; oral test; criterion evidence; Toulmin's Argument Model

【基金项目】本课题为北京语言大学学院级科研项目（中央高校基本科研业务专项资金资助），项目编号为“16YJ050005”

The Research of Chinese Proficiency and Evaluation

Guo Mingming

Beijing Language and Culture University

Language proficiency is an important part of the quality of citizens, and it is significant for the country's political, economic and cultural development . In recent years, due to excessive emphasis on foreign language teaching, flooding of network language and neglect of mother tongue nurturing, the misuse of language becomes more and more, such as a misnomer, grammatical errors, sentences are not fluent, misprint, and various vulgar language, language proficiency level of the citizens tend to decline. It is time to establish a comprehensive, scientific evaluation system of Chinese ability, which can do language proficiency test and provide feedback to guide mother-tongue education, so as to effectively prevent the decline of language proficiency.

At present, there are a number of studies in theory of the ability of Chinese construction assessment system in domestic, but the phased research results are different. The study reference the research results which are related to capacity evaluation system at home and abroad, at the same time, combined with the Chinese syllabus, in order to investigate what Chinese native capacity assessment should test and how to evaluate.

The article used descriptive research methods, through literature research to learn history and current situation of Chinese mother tongue related issues. On this basis, we attempted to analyze the existing research on the Chinese capacity evaluation system, summarize the status, characteristics, and we proposed that the evaluation of Chinese mother tongue should cover linguistic knowledge, language skills, learning strategies and social competence.



Room 102

Symposium 3

Discussant: He Lianzhen

Presenters: He Jiawen, Min Shangchao, Chen Dajian, Zhao Liang, Zhang Jie

Development of the Listening Proficiency Subscale of China's Standards of English (CSE)

Presentation 1

Development of the Listening Proficiency Subscale of China Standards of English

Developing China Standards of English (CSE) is the highest priority and the most fundamental work in the construction of the National Foreign Language Testing and Assessment System. This paper introduces the process of developing the Listening Proficiency Subscale, which is an indivisible part of CSE, including the theoretical and empirical foundation, the development of a descriptive scheme and a descriptors pool as well as scaling and validation. Also the innovations of the scale are discussed in this paper.

Presentation 2

Validation of listening descriptors of China Standards of English: An analysis of self-assessment data using polytomous IRT models

In order to address the problem of overlaps and inconsistencies in the EFL curriculum requirements at different educational stages in China, the project of developing China Standards of English (CSE) was launched in September 2014, aiming to establish a national framework of reference for English language education. This study mainly focuses on the development of listening sub-scale of CSE and preliminary validation of the listening descriptors. In terms of development, the major breakthroughs of the listening subscale of CSE, as compared to the Common European Framework of Reference for Languages (CEFR) and other international standards and frameworks, will be presented. In terms of validation, we will report a relatively large-scale empirical study in which three different polytomous IRT models are used to scale 1,315 university students' self-assessment data on 40 descriptors. The results show that, generally, the descriptors have good discrimination and medium difficulty parameters, indicating that they work well in describing learners' English listening proficiency at that level. However, some descriptors are too difficult and thus relocated to other levels by expert review, and 10 out of the 40 descriptors cannot fit the models and are further modified in terms of language and content by experts. This study confirmed the feasibility of scaling descriptors by using polytomous IRT models to analyze self-assessment data and pointed out the need to triangulate the results by using data from teacher assessment and expert judgment.

Presentation 3

Investigating the inner structure of listening ability descriptors: analysis of “salient features” based on students’ self-assessment

Descriptors are basic elements of the English ability scales (China Standards of English), and its structure and content bear both the language ability construct and distinguishing features between levels. It’s a vital part of the scale developing process to systematically analyze and interpret the inner structure of the descriptors. This step benefits not only for ordinary scale users to have a better understanding of the levels, but also for test developers and textbook writers to conveniently link their work to the scales. This study takes the descriptors of the cognitive ability scales of listening comprehension as example, and focus on the sketching of salient features between descriptors, then tries to validate the appropriateness and consistency of these salient features through the data of student’s self-evaluation.

Presentation 4

Analyzing the Different Understanding of Listening Ability Descriptors: Students’ Perspective

In order to verify the descriptors of the China Standards of English (CSE), large-scale online surveys on these descriptors have been conducted. Through the analysis of the data, certain significant differences of the understanding on the rating of listening subscale descriptors have been found between teachers and students. This study first analyzes the data from level B2 and B3, then 26 descriptors that may lead to such differences are chosen to make a survey. After the survey, 10 representative students are selected to do the follow-up interview exploring the potential reasons behind those different understandings from the students’ perspective.



Symposium 4

Discussant: 武尊民

Presenters: 武尊民 柳丽萍 何晓阳 张春青 李久亮 周红 董连忠

高中生英语成长诊断系统的建设及相关研究

高中英语教学将面临深刻变革：《高中英语课程标准》的修订、高考考核方式的转变、国家外语测评体系的建设等都将对高中英语教学产生深远影响。在这样的变革中，评价是不容忽视的重要组成部分。同时，在高中英语教学实践中，课堂评价体系、教师评价素养等的欠缺已经成为制约英语教学发展的因素之一，因此诊断测评作为与教学联系最为紧密的测试形式备受学界关注，也为解决上述需求提供了可能。本次会议拟采用专题讨论（Symposia）的形式，介绍“高中生英语成长诊断测评系统”的建设及相关研究课题。

“高中生英语成长诊断测评系统”简称“优诊学”，是外研社研发的在线诊断学习系统，采用“实施诊断——发现问题——提出建议——专项练习”的诊学模式，为高中三个年级的教师和学生提供在英语教学和学习方面的诊断报告、针对性的建议和补救性练习。诊断首先要清楚诊什么，高中英语能力及其构成主要以国家对高中英语教学的相关要求以及教学实际为出发点。诊断报告既反映学生实际英语水平离目标水平的差距，同时也体现阅读、听力、语言知识运用和写作四个技能及其微技能上的优势和不足。针对诊断出的问题，教师和学生将获得针对性的学习建议和相应的补救练习。

该课题由外语教学与研究出版社于2015年正式发起，由北京师范大学外语测试与评价研究所提供学术指导、河北省教育学会中小学外语教学专业委员会协作开展。目前“优诊学”已经在4所学校、20多个班级、1200多名学生开始试用，课题组正在围绕该系统在一线教学中的应用展开相关研究。

此次研讨拟从1）该诊断测试的意义和贡献；2）该诊断测试的效度验证；3）该诊断测试的后效研究；4）诊断测试在高中英语课堂教学中的应用模式四个方面进行（各部分具体【摘要】见下）。所提专题与大会主题密切相关，诊断性测试是在当今考试改革与发展的新形势下蓬勃发展的一种测试形式，希望本研究为诊断测试在中国英语教学的应用、研究和发展提供更多启示。

基于AUA框架的高中英语“优诊学” 诊断测试系统的效度验证

浙江外国语学院 张春青

本研究依据Bachman& Palmer(2010)“评价使用论证”框架收集外研社研发的“高中英语优诊学诊断测试系统”的效度证据。除了分别验证听力、阅读、语言知识运用和写作分技能的效度，本研究还收集整卷的效度证据。在测评记录一致性方面，研究除了计算 α 系数，写作评分的评分员间和评分员自身信度，还分析分类一致性比率。在分数解释部分，一方面，研究使用验证性因子分析考察整卷的因子结构来验证试卷各个技能试题所考查的能力结构是否与拟测能力结构相同，使用IRT分析每个技能的考查是否存在构念体现不足和构念无关因素，使用有声思维来探查考生是否使用了测试者拟测构念来答题；另一方面，研究还分析试题对掌握者和未掌握者的区分能力，采用学生掌握微技能数量与学生自评的相关分析来确定测试的校标关联效度，采用群组分析来获得构念效度证据。另外，研究还通过问卷和访谈形式来进一步验证诊断反馈报告的准确度和充分性。



高中生英语成长诊断测评系统的后效研究

北京服装学院 李久亮

考试后效(consequence)，即考试所产生的影响，是语言测试领域的一个重要课题。语言测试专家Bachman和Palmer（2010）把考试后效放在考试开发步骤最重要的位置，认为考试设计首先要考虑的，是考试结果的使用对涉考者所产生的影响和后果及其受益情况(beneficial consequence)。英语诊断性测评旨在测量学生语言技能方面的优势和不足，确保师生从教学中受益。由此看来，诊断性测评应该会产生积极后效，但目前很少有学者专注于此类考试的后效研究。本项目计划以此为切入点，观察与分析优诊学 – 一项新近开发出来的诊断测评系统 – 所产生的后效，补充学界在此领域内的研究与发现。拟采用Bachman & Palmer(2010)的AUA框架来指导优诊学的后效研究。AUA虽然不是一个考试后效模型，但是它强大的解释力为后效研究提供了理论依据。研究的一个主要问题是：优诊学系统在何种程度上产生积极后效？初始的研究设计见下表：

	研究问题	研究工具	参与者
1	优诊学系统的期望后效与实际后效是否一致？	课堂观察、访谈、问卷	高中一年级师生
2	系统在何种程度上有益于教师的教学和学生的学习？	访谈、问卷、测试成绩	同上
3	使用系统后学生学习能力（策略/方法）和思维能力是否发生了变化？	MSLQ问卷、前测后测反馈报告	同上
4	教师和学生如何评价优诊学反馈报告？	访谈、反思	高中二年级师生
5	使用优诊学系统的师生和未使用该系统的师生的课堂有何不同？	课堂观察、访谈、	同上
6	学生的学习动机、价值取向、优诊学系统的使用、和高考之间有何种关系？	MSLQ问卷、优诊学成绩、高考（模拟）题成绩	高中一年级师生

高中英语诊断测评与课堂教学相结合的模式研究

石家庄市教育科学科学研究所 周红

北京劳动关系学院 董连忠

【摘要】诊断测试是基于特定语言发展理论、借助二语习得研究的低利害或无利害考试；而诊断测评是指为准确把握学习者的学业发展或不足，通过对学习者学习过程中的多种因素，如学习者的学习动机、学能、心智成熟度、学习经验、个体记忆、认知等进行广泛的调查而实施的诊断性评价（diagnosis assessment）。本研究通过采用基于计算机在线诊断测试与高中英语教学深度融合的课堂模式，培养学生英语学科核心素养。同时，通过课题研究提高教师反思意识，不断改进教学，促进其专业化的不断发展。

课题研究采用定性与定量相结合的方法。研究工具包括学生诊断测试成绩，教师和学生调查问卷和访谈（开放式访谈、半结构和结构式访谈）、课堂观察、学生日志/周记等。研究对象为北京和河北省4所高中实验校实验班的英语教师及所在班级学生。每所实验校5名教师和高一、高二各两个实验班学生。各实验校在兼顾总体（听、读、写、语言知识运用等）诊断测评的情况下，分别针对其中一个子课题进行深入研究。四所实验校子课题研究分别为：优诊学测评体系与高中英语1）词汇和语法教学、2）听力教学、3）阅读教学、4）写作教学相结合的课堂教学模式研究。四所实验校课题研究旨在通过采用优诊学系统，有针对性地探究提升高中生英语学习技能与综合运用能力的有效途径，从而使“以学论教，以教促学”的课堂教学模式得以呈现。

【关键词】诊断测评；高中英语；课堂教学；教学模式

《中国英语能力等级量表》研究项目下的笔译能力量表 研发报告：构念、方法、过程与进展

严 明
黑龙江大学

研发笔译能力等级量表是中国英语能力等级量表项目研究工作的一部分，也是我国未来外语测评体系改革和建设的重要组成部分，目的是为中国语境下的翻译能力测评体系提供理论基础，为各阶段翻译教学提供参照框架，以促进各阶段教学衔接，提高翻译测试质量和功用，进而提升全国翻译教学水平和学习者翻译能力。目前国际上采用“能做”描述方法描写笔译能力的成果几乎为零，更缺少对中国学习者笔译能力的“能做”描述。本课题在这方面进行了创新性研究。

本文将汇报中国英语能力等级量表框架下的（汉译英）笔译能力等级量表子课题的构念、原则、方法、流程和最新进展。具体包括：

1. 在国内外翻译能力研究相关文献研究基础上建立的翻译能力构念；
2. 以能力可观测性为基本描述原则，设计的适用于本量表框架下的笔译能力量表参数体系；
3. 采用“能做”描述法所设计的笔译能力描述语的语义与句法规范；
4. 设计并实施的各种研究方法，包括语料库方法、调查法、定性研究法、定量研究法、直觉法以及描述语收集与修订的相关方法，以及确立的典型笔译活动；
5. 对笔译能力量表方案和描述语进行的各种验证。已经对收集的笔译能力描述语首先进行了多轮合格性判断、审核和修订，目前在全国本科与硕士相关院校进行大规模分类与分级信度、效度验证；
6. 笔译能力描述语收集的最新统计结果；
7. 对量表研究与验证过程的一些思考。

基于语料库的高级英语学习者听力能力建构研究

彭康洲 彭之尧

重庆师范大学

本研究旨在借助语料库的手段对所收集的国外英语能力量表的描述语进行分析,以便探讨高级英语学习者听力能力的衡量指标。采用语料库软件AntConc3.4.1对国外高段的听力能力描述语中的词频进行统计,一是找出频率较高的单词及所在的描述语段,二是对词簇(包括2个单词的短语)进行统计,提炼出听力能力的描述指标。我们的初步分析发现,国外英语能力量表中描述高级英语学习者听力能力的关键词有:多样性(如场景、话语参与者、话语风格和口音、交际任务),文学性(如手法、文本、表达),复杂性(如场景、语言结构、认知活动)。基于以上发现,本研究尝试提出了中国高级英语学习者听力能力的建构要素和测评启示。

The present study intends to find out the indicators of listening competence for advanced English learners by analyzing the description of English language proficiency scales with a corpus-based approach. To this end, description of listening proficiency scales in the advanced levels were analyzed with AntConc 3.4.1 in terms of the following features: the frequency of the words, the clusters (2-Grams). The initial findings showed that, for advanced level of listening proficiency scales, the concepts of Variety, Literature, and Complexity are highlighted. Variety is used to describe the voices, styles of delivery, native-speaker accents, context, and communicative tasks. Literature is used to describe the devices of text, text, and expression. And Complexity is used to describe the context, language structure and thinking skills (like interpreting, analyzing, evaluating and responding, etc.). Implications for constructing and assessing listening competence of advanced English learners are discussed.



Developing and Validating a Reading Strategy Scale for Chinese Tertiary EFL Learners

Zhou Yanqiong

Guangdong University of Foreign Studies

The China State Council issued the document The Implementation Opinions in September 2014 which stipulated the construction of the China standards of English (CSE). According to the theoretical foundation of CSE, strategic competence is one component of language ability. It is a consensus that strategic competence is an indispensable part of language ability. However, language learning strategy is seldom included in the current language proficiency scales and the categories of strategy are not complete either.

The aim of this study was to develop a theoretically-based and empirically-developed reading strategy scale and to evaluate whether such a scale functions reliably and validly. Three questions were addressed: (1) What strategies do the Chinese tertiary EFL learners adopt in reading comprehension? (2) Are the reading strategies gradable? (3) How reliable and valid is the scale of EFL reading strategies? The study was undertaken in three phases. During Phase 1, the strategy items were elicited from EFL learners through verbal report and written diaries, generated from EFL instructors through exemplar generation and collected from literature survey as well, then all the items were transformed into descriptors and these descriptors were compiled into questionnaires and administered to a sample of 1200 EFL tertiary learners to address the first question; Nvivo 7 is used for the qualitative analysis. In phrase 2, the results of phase one were used as the basis for the development of a reading strategy scale through 6 experts' judgment and an analysis of the judgments of 2000 EFL learners using many-facet rasch model . Phase 3 involved the validation of this empirically-developed scale, the Rasch-measurement software package Facets (Linacre, 2010) were used for the analysis of 600 EFL learners' judgment. Interviews were also administered to EFL learners and instructors to elicit the raters' perceptions of the efficacy of the scale.

This study is supposed to answer the three research questions; one overall reading strategy scale and one self-assessment reading strategy scale are expected.

Toward a Framework of CEFR-referenced Legal English Proficiency Scales

Wang Haiping

East China University of Political Science and Law

Abstract: In order to design scientific and valid legal English proficiency scales, we must clarify the target of law talent training in the context of globalization. In face of the challenge to become international law talents with competitive edge, law majors are required to develop the ability to communicate effectively in international legal contexts. Yet up till now, standardized legal English proficiency scales are underdeveloped. The only legal English test in mainland China that can be used to assess test-takers' comprehensive legal English proficiency is LEC (Legal English Certificate) test, with content construct including basic knowledge of American law and legal English reading, writing, translation and other aspects of knowledge and skills (LEC syllabus, 2013). But the oral communication is excluded from the scope of the LEC test and there is no sign that signals LEC is related to any well-accepted international English test or language proficiency scales.

CEFR resorts to “can do” statements to describe language competence, which is an “action-oriented” standard. CEFR provides a framework on how to describe language proficiency on a scale in an authentic context and hence it can be used as a reference that contributes to the formulation of legal English proficiency scales. College English Syllabus provides a scale of general language proficiency descriptions mainly derived from the education domain. Legal English belongs to ESP that targets language use situations. Test tasks should be authentically representative of tasks in the target situation and such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain. Therefore, legal English proficiency scales must be defined in the occupational domain, in which the person concerned is engaged in his or her job or profession. CEFR has detailed description of external contexts of language use including occupational domain that can serve as valid contexts where scales of legal English proficiency are developed.

Finally, the detailed occupational domain of CEFL provides a tool for needs analysis. CEFR can help us make clear what the key characteristics of the target language use domain are and help us analyze learners' needs and define specific characteristics of ESP use. Formulation of legal English proficiency scales without profound needs analysis from the occupational domain as well as learners will lead to assessments that are neither authentic nor accepted.

Key Words: Framework; CEFR; Legal English; Proficiency Scales



A Study on the Descriptor Database of the Reading Ability of Undergraduate Foreign Students in China in Preparatory Education

Wang Shuang Wang Jimin
Beijing language and Culture University

Descriptors of language proficiency is a specific concept of language proficiency scales, it is described in the real life, the users in different levels could complete the tasks and the features they have.

The preparatory education of undergraduate foreign students in china is an important part of the teaching of Chinese as a second language. Chinese reading ability is the most important way to acquire knowledge for the foreign students. Reading proficiency scales for preparatory students can make the teaching, learning and testing with a unified reference framework, and a critical study is the descriptor database. This database will be a detailed description of the level of Chinese language proficiency and its characteristics.

The descriptor database of the reading ability of this preparatory students in china includes the descriptors themselves and the parameters of the descriptors. The parameter database system consists of two parts, namely the ability attribute of each descriptor (such as reading process) and implementation difficulty (mathematical index marked the difficulty value, illustrating that the difficulty of completing the task description language or the language features).

We used literature method to collect and determine the descriptors of reading ability, compiling the reading ability of descriptor questionnaire, and asked the teachers of the students who participated in the test of 2015 Chinese government scholarship program for undergraduate education in china answer the questionnaire. 844 questionnaires is valid, statistical analysis showed that the questionnaire and examination of the relevant 0.491.

In the process of parameter estimation, we will use the Rasch model to calculate the logit value of the difficulty of the descriptors, and will use the multidimensional scaling method to analyze the capability of the descriptors, perfecting the descriptor database.

Prediction Research of Item Difficulty of Verbal Comprehension and Expression

Kong Xiang Zhang Xuan
Beijing Language and Culture University

Abstract: With the continuous development of computerized test and large scale test, the item bank construction is increasing more and more attention. Construction high quality item bank can guarantee the security of the test, raise the scientific nature of the test. Item difficulty as a subject of some important parameters is the foundation of the construction of high quality item bank, therefore predicting the item difficulty precise before the unused item put in bank is of great significance.

Verbal comprehension and expression is a part of the provinces and cities enroll in course of the basic knowledge of the comprehensive exam, mainly inspects examinee's ability to comprehensive analysis of language. the item difficulty of the predicting method of the verbal comprehension and expression was the expert subjective evaluation method, from the data in recent years, a large part of items existed deviation between predicted difficulty and measured difficulty. With the development of information technology, especially the big data technology, we see the possibilities of using scientific means to improve the prediction accuracy. This study is based on the basic knowledge of the comprehensive exam of six years during 2010-2015 verbal comprehension and expression items and measured data as the research sample, try to use the related algorithm based on Bayesian method to evaluate the item difficulty of the verbal comprehension and expression. Verbal comprehension and expression specific divided into two parts: reading comprehension and blanking filling . Finding out the factors that influence the item difficulty of reading comprehension and blanking filling respectively, and building the difficulty factors framework. Combined with the actual situation of the verbal comprehension and expression, selecting a certain proportion of the test set and prediction set from all kinds of subjects , Using Bayesian method to predict the item difficulty, and comparing with the expert subjective evaluation method and the actual difficulty ,assessment the feasibility of Bayesian method to the item difficulty estimates.

Key words: Construction of the Item Bank, Prediction of Item Difficulty, Verbal Comprehension and Expression, Bayesian Method



‘The Rubber Ruler.’ Using Proficiency Scales as an Accurate Measurement of Classroom Achievement

Philip Horne
British Council China

Providing feedback to students can be a tricky affair. Teachers are often able to make qualitative statements such as ‘Jane has good grammar, but struggles a little with pronunciation’ or ‘Peter is able to write very well, but is shy to speak.’ However, making informed quantitative judgements (e.g. B+ or 7/10) is often either too simplistic, or lacking in understanding of standardized testing. Measurement of cognitive language ability is often referred to as a ‘rubber ruler,’ because unlike precise physical measurements, language is an abstract concept and is subject to many variables. Giving a precise grade is, therefore, something that requires a little more expertise, and an appreciation of principled, routine assessment.

The purpose of this workshop is to look at and analyze various methods of quantitative assessment, in order to help teachers make more informed judgements, and provide students with more accurate feedback. We will analyze theories behind rating scales and test constructs, and then look at turning this understanding into practical classroom application. By the end of the session, it is hoped that participants will have a better understanding of testing theory, and how it can inform their own teaching.

A study on the Language Ability of the Employees in the Window Industry in China

Liu Beibei Zhao Qifeng
Beijing Language and Culture University

In the new period of Reform and Opening up, the state has paid more attention to language use in the window industry. Policy reports are mentioned to play a good role in the role of the public service industry.

This study first reviewed the relevant domestic research, it mainly on the development of the window industry policy guidance, the definition of the concept of language ability, as well as the language use survey of specific tourism, medical, civil aviation, banking and other services industry. Secondly, this paper briefly presents the language skills and related testing methods of foreign business, leisure tourism and hotel service industry. On the basis of this, this article use the foreign Telephone Bureau to recruit the employee, s test and the hotel staff recruitment test, try to put forward the evaluation of the language ability of the hotel industry in our country.

The purpose of this test is to test and evaluate the language ability of the hotel industry practitioners, in order to choose the right staff to work in the hotel lobby. The target group of the test is a young man with a professional education in the hotel, fluent in Mandarin and a certain foreign language. The language ability is defined as two parts: language knowledge and professional knowledge. Test method has two forms of oral and writing. Both speaking and writing form the syntax, vocabulary, information organization, text cohesion, industry terms five parts to score, each part is based on a scale of five, from two aspects of range and accuracy to score.

Ethics of Language Testing from the Perspective of Validity

Gao Shuling
Northwest University

Abstract: The ethics of language testing is the main problem and one of the difficulties in the field of language testing in the 21st century. Ethics refers to order and criterion among people, people and society, country and their behavior. Any group behavior or professional one which have impact on society has its own built-in ethical exception. Since Messick expanded the concept of validity, ethical issues have become the frontier of this field a little further. With the development of language testing, language testers become more and more concerned with the study of ethics. In 2000, the International Language Testing Association published the Code of Ethics in which ethics is embodied in every principle. Then according to the specific implementation of Code of Ethics, it published Draft Code of Practice in 2007 which serves as the guidelines for practice, specifying the responsibilities and obligations of both language testers and examinees. The ethical issues, in its essence, belong to the category of the validity. This paper, taking the initial stage of China's basic study condition of ethics whether it is about theory or practice into consideration, using the methods of literature research, descriptive method, qualitative analysis, employing the validity as its theoretical basis, discusses the ethical problems through each part of the test, specifically, the production of test, the preparation and implementation of examination, the organization of grading and scoring, the analysis of the results and the feedback and summary; based on Code of Ethics and Draft Code of Practice, it also summarizes the causes of ethical issues and puts forward the countermeasures. Finally, it proposes some suggestions.

Key words: validity; language testing; ethics.



Formative Assessment Implementation in India: A New Reform on English Curriculum in Elementary Schools

Wang Feiyu
Yanshan University

Since 2011, a new reform on English curriculum and the teaching practice of six-step teaching cycle was implemented in selected elementary schools in Mumbai, India. The six-step teaching cycle (task starter, literature circle, writer's workshop, grammar, speaking & listening and weekly assessment) was designed based on the concept of formative assessment to promote students' continuous improvement and to inform teaching decisions on a day-to-day basis. Therefore, it's of great necessity to get to know how teachers understood the concept of formative assessment and how they implemented formative assessments into their English curriculum in order to monitor students' learning and provide information for further instruction.

The framework for exploring formative assessment in this study is based on the idea that informal formative assessment can take place at any level of student–teacher interaction in class such as whole class and small groups (Black & Wiliam, 1998), and can help teachers collect information constantly about students' level of understanding.

The case study of India elementary school teachers' formative assessment provides a detailed description on teachers' knowledge on formative assessment as well as how they implement formative assessment into the English curriculum. By using the initiation– response–evaluation-utilization (IREU) coding system, the study analyzes teachers' formative assessment implementations and their assessment conversations with students, meanwhile visualizing the benefits as well as the problems of current formative assessment implementation such as the inconsistency between teaching objectives and assessments practice.

高考英语四十年内容效度历史研究

辜向东

高晓莹

李玉龙

重庆大学

解放军后勤工程学院

东华理工大学

本文为国家社科基金青年项目《高考英语试卷内容效度历时研究》(项目号: 08CYY013) 结题报告的研究成果汇报, 结题等级为良好。该研究针对高考英语科试卷进行历时研究, 探索高考英语1977年恢复实施以来的发展趋势和总体特征, 为高考政策决策者和高考命题工作者提供一定参考。该研究不仅涉及全国卷而且涉及各省市自主命题卷, 不仅有对全国整体情况的宏观研究, 也有对单卷、单题型或单考点的微观分析, 其研究范围广, 时间跨度大, 全国尚属首例。

通过对高考英语科试卷的历时分析, 我们得出英语科试卷总体来看具有较高的内容效度; 随着教育部相继出台恢复高考、高考标准化测试、分省市自主命题等重大改革政策, 我国高考英语试卷在命题方面做出了相应的调整与完善, 符合当时改革的政策和大纲要求, 所测内容是教学大纲和考纲的代表性抽样, 符合该时期的欲测目标。此外, 高考英语在改革历程中也呈现出一定的变化发展趋势: 其试题选材日趋丰富多样, 话题日渐真实、新颖、富有时代特色; 考点抽样量逐渐增大, 逐步从语言知识型测试转向语言运用能力型测试, 逐步淡化单纯语法测试, 开始加强交际能力测试; 题项设计日趋科学、规范, 并且体现出一定自主性和创新性。虽然总体上讲, 高考英语试题的命题质量与水平较高, 但是分析中也发现仍有部分试题在设计上存在问题, 特别是2004年各省市自主命题以后, 各地命题难于统一监督与管理, 个别省市的个别试题在命题质量上有待提高。



Measuring and Understanding Self-regulated EFL Learning within an Online Formative Assessment Module

Liang Li

Guangdong University of Foreign Studies

Abstract: Emerging since 1980s, self-regulated learning(SRL) and formative assessment have been relatively new and heated topics in the fields of learning strategy and assessment (eg. Berry, 2011; Li, 2012). Educational testing should not merely be the assessment of learning but the assessment for learning (Jin, 2000; Liu, 2013) and formative assessment is acknowledged to possesses huge potential in promoting an assessment for learning. Generally speaking, in a formative assessment environment, students are self-regulated learners if they are motivationally, meta-cognitively and behaviorally active participants in their own learning process, instead of relying on teachers or other resources (Zimmerman, 1986; 1989). From this definition, it is argued that SRL is one of the central parts of formative assessment and well-worth an in-depth study. But it is noted that some essential issues of formative assessment remain rarely touched yet, such as the question of how formative assessment promotes learners' motivation, self-efficacy beliefs, and regulation of learning, and so on. Given the above question mentioned, this paper reports a case study of SRL in an online blended formative assessment module in the context of non-English-major college English instruction in GDUFS. The intent of the present research is to exemplify how SRL, an active area in educational psychology, can help to illustrate and understand the mechanism and process of SRL in the online formative assessment environment.

形成性评价在大学商务英语课程中的应用研究

王 薇

浙江海洋大学

【摘 要】形成性评价是一种贯穿于整个学习过程的评价方式。学生可以根据反馈的信息调整学习策略，教师也可通过及时的反馈改善教学方法。本文旨在研究形成性评价在大学商务英语课程中的应用过程与实践效果。研究团队采用“项目式学习”的教学模式，确立明确的学习目标，鼓励学生进行协作学习，并以“学习档案”的形式要求学生进行自我记录与自我评价，同时对协作学习过程与项目成果进行同伴互评与教师评价。通过课堂观察、问卷调查与访谈等方式，研究人员发现形成性评增强了学生的学习动力，促进了学生的自主学习能力，并提高了学生的协作学习技能。

Use Argument Against Scores of College English Course Assessment: A Regional Study of Jiangsu Province

Li Lian

China University of Mining and Technology

Abstract: Under the unified concept, the validity of an assessment is in the interpretation of its scores. A language assessment, to be valid, must justify in the statistical sense how its scores are computed and, testify that the coverage of its scores' interpretation conforms to test purpose. The present study is a use argument against the scores College English course assessment of 7 universities of Jiangsu Province, China, composed of three modules, final examination, class evaluation, and on-screen exercises. The argument provides claim, data, warrant and rebuttals for the scores and their computation along three moves in the assessment, i.e., design, execution and decision.

The argument finds apparent rebuttals against the scores of sample assessments, along with statistical backings for the rebuttals. Results of homogeneity show that scores of the three major modules are not fit in meaning, with significant deviation among each other. Spoken test in the final exam is also deviated from the other sections. Factor analysis reveals that the construct validity is impaired by the missing scores and redundant scores on certain constructs.

Computation of the surveyed scores is also found to weaken validity of the assessment. Scores of different modules or sections, not correlated enough, are added up to total scores, which are even less correlated to sectional score sets and undermine their interpretation. The assessment's consequential validity is thus reduced. Furthermore, such composite scores are significantly biased, compared to the standard synthetizing, in terms of interpretation for their weighted compositions are inhomogeneous. Inconsistent changes of task difficulties and raters are found as significant influencing factors, causing the shortage of sufficient reliability evidence.

In response to the rebuttals and their backings, the present study proposes new claims on the computation and interpretation of college English assessment scores, as criterion-based construct range checklists, local course assessment norms, report form of percentile rank plus standard score, and equation of reliability evidence as moderators in score computation.

Keywords: College English Course Assessment, Test score, Use Argument, Unified Concept of Validity.

A diagnostic Assessment on the Receptive and Productive Vocabulary Size of Advanced Chinese Learners

Liu Shuhui

South China Normal University

International Curriculum for Chinese Language Education (2014) describes clearly the linguistic knowledge, skills, strategies and cultural awareness in six levels which Chinese learners are supposed to possess. It is imperative for both Chinese teachers and learners to understand whether the learners have acquired the necessary vocabulary they are supposed to at each level.

This study intends to investigate learners' receptive and productive vocabulary size. Receptive vocabulary means that the learners recognize the word when you hear or see it. Productive vocabulary means that the learners are able to use it in speech and writing (Read 2000:26). The research instrument is a 60-item vocabulary test battery composed of a receptive test (Test One) and a productive test (Test Two). The subjects are 39 juniors in SCNU with 14 males and 25 females, who come from Indonesia, Thailand, Russia, Vietnam, Tajikistan and other countries. The test battery is validated to be effective based on the following evidence: (1) Cronbach's alpha is 0.832; (2) Test One and Test Two are correlated significantly and positively, $0.559, p < 0.01$, while Test One and Test Two are correlated significantly with the total respectively with alpha 0.832 and 0.932, $P < 0.01$; (3) factor analysis concludes that only one component stands out, i.e. Vocabulary ability; (4) regression analysis finds that the test battery could predict 14.4% of the final writing test score; (5) 60 Chinese words are sample with equal interval from HSK 5 level vocabulary list (1300), while the texts of the test battery are examined with the software "Chinese Test Compass" and are marked with HSK 5 level. All the evidences about the reliability, construct validity, predictive validity and content validity of the test battery enable the test battery to be an effective evaluation instrument.

The study finds out that advanced learners achieve 91% correctness on receptive vocabulary test and 39% correctness on productive vocabulary test. Thus, junior Chinese learners at SCNU master 1183 receptive vocabulary and only 507 productive vocabulary out of 1300 words. The paper also forwards some suggestions on vocabulary teaching.

本文为广东省教育厅教育科学研究项目“汉语国际教育视阈下留学生汉语阅读能力评价机制研究”(项目编号: 2015GXJK023)阶段性成果之一。

听力诊断性测试效度研究：基于考生口陈报告的证据

王 华

山西大学

诊断性测试目的是发现语言学习者在某一目标语言技能存在的不足和优势，并提供有针对性的诊断反馈和补救性学习方法（Lee 2015）。Alderson et al (2015)指出，诊断性测试成功的关键在于诊断信息的准确性，因此诊断性测试效度验证的首要任务是验证基于诊断性测试所做出的解释在多大程度上可准确反映学习者掌握某一技能的强弱。本研究使用即时回忆方法，对外语教学与研究出版社研发的高中生英语成长诊断测评（简称优诊学）的听力理解模块进行效度验证。优诊学听力诊断性测试通过单选题和简单题考查考生获取特定信息、听懂指令和步骤、理解所听材料的主要观点和细节信息、了解所听材料的主旨大意和推测说话者的观点、态度、意图、关系和所在场所5个微技能。

对考生即时回忆的口陈报告分析表明，考生在完成优诊学听力诊断性测试时，均使用了优诊学所测的微技能。每个考生口陈报告所呈现的微技能强弱基本上符合系统预测的微技能的强弱。除此之外，本研究还发现考生在考试中使用了其它微技能，这一研究结果对进一步改善优诊学听力测试提供了参考意见。



A TBR-based Cognitive Diagnostic Modeling for EFL Reading Test

Du Wenbo Ma Xiaomei
Xi'an Jiaotong University

Abstract: Diagnostic language assessment (DLA) has recently gained a lot of attention from teachers, language testers and second language acquisition researchers. It seeks to promote further learning designed to address the learners' weaknesses and increase their overall growth potential with "learning-oriented assessment" as its rationale. With the empowerment of Cognitive Diagnostic Approach (CDA), the latest theory in psychometrics, DLA is possible to be achieved not only theoretically but also methodologically and practically. However, the results of most CDA-based research have classified learners' mastery of a set of tested skills into a dichotomous-score pattern (0/1 pattern), lacking accuracy in that the critical value of mastery is vague.

Aiming at providing students with finer-grained diagnostic feedback, this paper addresses how an EFL reading diagnostic assessment model is constructed based on CDA along with Tree-Based Regression (TBR) approach. The study starts with building hypothetical diagnostic model including cognitive reading attributes, diagnostic test design, and Q-matrix models, followed by the validation of reading attributes via TBR analysis and Q-matrix via Fleiss Kappa. Then Ox software is adopted to generate Group-level and Individual-level mastery probability of each identified reading attributes and LOWESS approach is applied to gain a group-level mastery tendency model. Finally, all the diagnostic information is synthesized into a well-designed diagnostic feedback including 1. Learners' relative placement in group-level diagnostic result; 2. The individual reading attribute mastery probabilities; 3. The interpretation of learners' performance in the diagnostic test. The findings are significant in achieving the goal of personalized assessment so as to fulfill the purpose of "tailored" L2 learning through DLA framework—integration of diagnosis, feedback, and remedy.

Key Words: Diagnostic language assessment; EFL reading; Cognitive diagnosis approach; Tree-based regression

基于英语听力认知诊断试题的动态干预模式

孟亚茹 马晓梅 赵宁宁 晏艺赫

西安交通大学

传统的听力测试只能为受试者提供终结性成绩,对相同分数背后的个体之间存在的知识状态与认知结构的差异却不得而知,当然要达到个性化、即时性和动态化的干预就更无从谈起。诊断性测试作为形成性评估因更加紧密联系教和学(Nitko, 1989)而成为语言测试研究新的关注点,其特点是“诊断-反馈-指导/干预”三位一体。

认知诊断评估cognitive diagnosis assessment(简称CDA)把现代教育测量方法与认知心理学的结合,通过获得被试在特定测验题目上(可观察)的作答反应模式而推知其不可观察的知识状态(Leighton & Gierl 2007),将传统的单一考试分数转化为学生对题目所涉及到的认知过程、认知结构与技能的掌握情况(戴海崎, 2010),据此才能针对性地提供补救和干预。

认知诊断的思想符合二语习得认知派的视角。认知语言派主要研究对象的是储存在大脑内的二语知识体系和习得二语知识的过程。前者从静态的角度分析储存在个体大脑中二语知识体系的特征,后者从动态的角度考察个体学会某种二语的过程(文秋芳, 2008)。

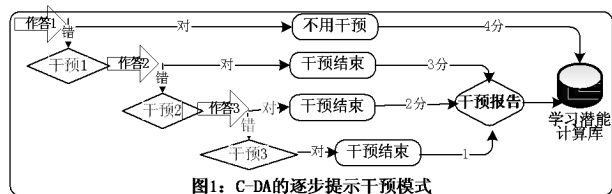
国内已有的诊断性研究也主要以英语阅读为主(李美娟 2011; 蔡艳等 2011),也有对海外汉语水平考试HSK听力(徐式婧2007; 周霞2009)和阅读(张宠, 2009)进行认知诊断研究、同时陈慧麟等(2013, 2015)对该方法的综述性介绍及对国外PISA测试的认知诊断分析。这些研究表明认知诊断方法能挖掘个体被试的认知加工等细颗粒度的信息,并能对被试进行成功归类,使后期的针对性干预“有法可依,有章可循”。

动态评估DA(Dynamic Assessment)是最近二、三十年国外兴起的一种新的能力评估取向,是社会文化理论(sociocultural theory)与教学结合的范

例,它指把测量和干预结合起来,通过交互方法以提示、指导和反馈等手段让被评估者积极参与到测验活动之中,对其思维、认知、学习和解决问题的能力进行评价的过程,因此评估与教学是一个以发展为导向的辩证统一体(Vygotsky, 1989; Lantolf & Poehner, 2011)。教师不再是学习行为的观察者,而是致力于和学习者共同解决问题、促进学习者进步的中介者,根据学习者水平和实际需求不断调整介入支持手段,学生未来的发展是依靠师生共同努力使学习者跨越最近发展区(ZPD)实现的(Poehner & Lantolf, 2005; Poehner, 2007; Lantolf & Poehner, 2008)。正因为ZPD能为学习者能力和问题的诊断提供框架,在上世纪中后期就引起教育诊断方面的研究热点(Leontiev, Luria, & Smirnov, 1968; Lidz, 1997; Lidz & Elliott, 2000; Kozulin & Gindis, 2007),所以说动态评估是具有强大诊断力的工具。国内动态评估的探索研究如火如荼,韩宝成(2009)认为DA适当的指导和干预使其学习潜能被进一步激发,从而达到能力的进步和认知水平的提高;张艳红(2008, 2010)和孔文等(2013)用实证研究的方法讨论了动态评估在写作过程的干预作用。李清华和李迪(2015)分享了近年来二语动态评估的25项实证研究,研究发现包括:个案小规模比较多、很多研究把焦点放在语法和阅读上面、计算机可以对介入起到辅助作用。图1以4个选项为例,如果被试第一次就选对,他不需要干预,即得分4分;选错第一次,他可以通过该选项的链接得到一次干预提示,做对后系统会给出3分评估分;以此类推,如果第三次选错,他获得三次干预,得分为1。该系统还借鉴了学习潜能得分(Learning Potential Score, LPS)这一概念试图量化动态评估。LPS是由Sternberg and Grigorenko (2002)提出的用于量化学习者初始状态和最终能够达到的能



力之间的差异。该系统把费时费力的干预从教室的限制中解放出来,为学生个性化自主性学习提供有力的工具。C-DA的诊断试题构建过程依据前期学生一一作答后的反馈设计,同时据此设计选项和相对干预提示的内容。虽然动态评估的介入式干预有无与伦比的优势,但他们对该试题和选项作为诊断量具的测量学属性如构念效度等并未全面考虑,其诊断结果的有效性也受到质疑。国内把DA用到语言诊断方面的研究寥寥无几,利用计算机网络技术实现动态诊断更鲜有问津。



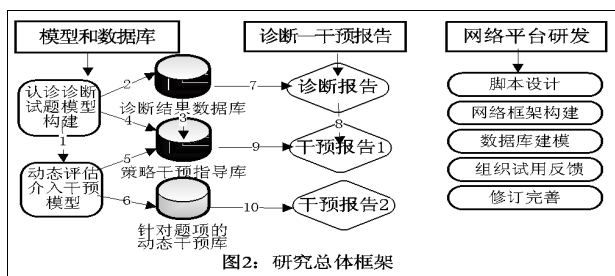
纵观现状,听力诊断测试研究相对薄弱,本研究结合认知诊断科学的诊断和精细化反馈的优势,利用动态评估的有效干预,扬长避短,构建一个从认知诊断到动态干预的英语听力平台。

总体框架 本研究涉及到的四个主要理论:听力理解的认知过程理论、语言测试理论、认知诊断理论和动态评估理论。研究总体框架如图2所示,首先要构建两个模型:“认知诊断试题模型”和“动态评估介入式干预模型”,包括它们生成的“诊断结果数据库”、“策略干预指导库”和“针对题项的动态干预库”。这个模块中各部分关系如下:“认知诊断试题模型”是“动态评估介入干预模型”的基础(箭头1),因为后者是基于前者的题项和及其选项建立的;前者生成“诊断结果数据库”(箭头2),据此建立“策略干预指导库”(箭头3),“策略干预指导库”的部分数据同时也可直接基于左边两个模型产生的数据(箭头4和5),最下方的“针对题项的动态干预库”直接来自其左侧的“动态评估介入干预模型”

(箭头6)。

其次生成“诊断-干预报告”,包括认知诊断的“诊断报告”(箭头7)和基于它和“策略干预指导库”的“干预报告1”(箭头8和9),“干预报告2”是基于左侧的“针对题项的动态干预库”(箭头10)。

最后是网络平台的研发,主要涉及到脚本设计、网络框架构建、数据库建模以及后期的组织使用分析反馈和修订完善。具体见图2。



价值和意义

学术价值 认知诊断作为新一代心理测量理论在EFL测试领域仍属尖端前沿,屈指可数的研究中介绍和综述居多,实证研究鲜有耳闻,外语学术界对它的巨大优势和潜力仍不甚了解;其次应用认知诊断技术开发英语诊断试题在外语界更少人问津。本课题将这一理论应用到EFL听力诊断试题构建中,为诊断性语言测试提供新的理论支持,拓宽了语言测试界的研究范围和方法。此外,动态评估理论虽然已经引入国内十多年,但大多数局限于课堂评估,因操作性受限而费时低效,难以展开应用性研究。本研究利用网络技术使其从课堂束缚中解放出来,为国内动态评估打开一个新的突破口。

更重要的是本研究摒弃学习与运用二元对立的观点,既遵循认知派注重研究二语习得者语言能力和加工过程等,也遵循社会派的互动干预理论,是把两者引入诊断测试,进行具体化、操作化、实施化的典范,为外语教学和研究提供了一条新思路,同时架起了不同学派互相学习借鉴的交流桥梁。

CDM Selection for English Reading Test and the Implications on Teaching and Learning

Chen Huilin

Shanghai International Studies University

According to Leighton and Gierl (2007), cognitive diagnostic approach is to measure unobservable knowledge structures and processing skills in students through their observable performance in tests. Cognitive diagnostic approach deepens and expands the traditional research on cognitive ability since traditional research on cognitive ability only aims to find out the ability level of students while cognitive diagnostic approach also tries to disclose the skill structure and skill characteristics possessed by a particular student or a group of students. Cognitive diagnosis is realized through cognitive diagnosis models (CDMs). CDMs can be classified into non-compensatory and compensatory models according to whether the probability of mastering a specific skill will contribute to the probability of answering the item correctly. CDMs can also be classified into reduced and saturated models according to whether the interactions among skills are considered.

This research adopted six CDMs (DINA, DINO, A-CDM, LLM, R-RUM, and G-DINA) to analyze the test results of PISA English reading test in order to get answers to the two questions below:

1. What kinds of CDMs are suitable to diagnose reading skills?
2. What implications will CDM selection have on reading testing, teaching, and learning?

The research found that the compensatory and saturated G-DINA model has the best model fit for the reading test data, which also demonstrated that reading skills are hierarchical and multidimensional.

Based on both statistical and substantive analysis, the research further discussed what implications CDM selection has on reading testing, teaching, and learning.

1. Appropriate CDMs may help to analyze the construct validity of reading tests.
2. The reading skill structure disclosed by appropriate CDMs may help to improve teaching plans so as to enhance teaching effect.
3. Appropriate CDMs may help to realize individualized teaching based on individual information of diagnosis.

Appropriate CDMs may help to establish a skill score reporting system which will make students realize their specific weaknesses and find ways to get rid of them.

Peer Assessment as an Innovation Strategy: Students' Perceptions and Practices

Zhou Jiming Jiang Yan
Fudan University Guangdong Teachers College of Foreign Language

Peer assessment has been positioned on the innovation agenda of Chinese universities. In the Chinese exam-oriented context, a traditional value is that only teachers have the expertise and agency to be assessors. There is a need for examining students' perceptions and practices in relation to peer assessment as an innovation strategy.

The three research questions are related to students' perceptions of peer assessment, their practices in a variety of peer assessment activities, and the interplay between their perceptions and practices. A university which redesigned its college English curriculum to incorporate peer assessment was selected as an information-rich case. Forty-five students taught by three English tutors were interviewed in eight groups at two different times over one academic year. Forty lessons were observed and audio-recorded in two classroom observation cycles. Students were also invited to write reflective journals at the end of the academic year. Data were analyzed following an inductive coding procedure adapted from the qualitative analysis protocols established by Miles and Huberman (1994).

Findings indicate that using peer assessment for summative assessment was perceived by students as "totally nonsense and unfair". In the second semester, when students were encouraged to give peer feedback to inform their peers of the learning gaps, students began to recognize the learning potential of peer assessment. Their understandings of assessment criteria and self-beliefs in their evaluative abilities were gradually enhanced during their participation processes.

This study illuminates students' perceptions and practices when they encounter assessment tasks that represent a learning culture different from that in their secondary school. These findings carry practical implications for implementing peer assessment in culturally appropriate ways in the non-western context. With the rapid influx of Chinese students into Western universities, the learning difficulties that students experience in their transitional year pose a challenge to tutors and universities administrators alike. Thus, the study also carries pedagogical implications for working with Chinese students in Western universities.

Perception of Self, Peer and Teacher Feedback: The Perspective of Receivers and Givers

Li Xuelian

Liuzhou Vocational & Technical College

Feedback, in present study, refers to the descriptive information posed by teachers, peers or selves, delivering what was done well, what need to be improved, how to be improved and so on. It plays a vital role in Assessment for Learning context (Black et al. 2003), and it is a crucial process which connects teaching and learning, for it addresses three central questions, where am I going, how am I going and where to next. Additionally, the focus of feedback lies in four levels: task, process, self-regulation and self (Hattie & Timperley 2007).

According to the principles of Assessment for learning, students are assigned to take part in self- and peer assessment to enhance their involvement and autonomy. Their learning independence, however, can be achieved only after they can pose effective feedback (Huang 2016) and value the received feedback (Hattie et al. 2016). Few studies, up to date, have examined how EFL students provide and receive feedback as independent and cooperative learners under Hattie and Timperley's framework. The current study, therefore, intends to answer the following research questions:

1. Are there any significant differences in students' perception of self-feedback, peer feedback and teacher feedback of the same assessment?
2. Are there any significant differences between students' perception of peer feedback received and provided?

Data from self-designed questionnaire, interview and classroom observation will be collected to answer the above questions. A questionnaire, including a 6-point Likert scale will be designed to investigate how students perceive the task, process, self-regulation, and self-level feedback provided by others and themselves. To probe the deep reasons of possible misalignment/alignment, interview and classroom observation will be carried out afterwards. The specific research procedures will be introduced in the following paragraph.

About two natural classes (n=80) will take part in the study. Teachers and students create harmonious Assessment for Learning environment. Each student is supposed to set learning objectives, establish achievement criteria, work toward learning objectives, take part in self- and peer assessment, receive teacher assessment, reflect on their behavior, propose suggestions, regulate their learning and then restart another assessment cycle. They will learn to talk about some career topics, such as career paths and career principles. Every two weeks, students need to have a short discussion on the unit topic in groups of four to five. The discussion will be videoed, and then each student assesses three group members' work and their own work as well online. Assessment in present study is comment-only, for marks are of little value in improving student learning (Black & Wiliam 2006). That is, each student will receive three peers' feedback and one teacher feedback. Attaching to the feedback, a questionnaire is designed for students to fill in. Then four students will be invited to take part in my interview to investigate the reasons for their perception of feedback and their classroom performance will be observed. After collecting the above data, I will analyze the quantitative data by SPSS and qualitative data by Winmax 98 to find out the answers to research questions.



基于同伴互评的专业英语演讲教学模式研究

杜玉霞

广州大学

【摘要】同伴互评是充分发挥学生主体性、以促进学习为目的形成性评价方式，同伴互评在外语写作教学等方面得到了有效应用，专业英语教学对这方面的探索还很少。在动态评估理论(Dynamic Assessment)和社会建构理论(Social Constructivism)指导下，根据《教育技术学专业英语》课程对学生能够准确运用英语进行专业问题交流能力培养的课程目标和教学内容，设计基于同伴互评的教育技术学专业本科生课堂英语演讲学习活动，开发同伴互评量规，发挥同伴互评的激励与导向作用，设计以评论对话和书面评价相结合的多元互评机制。通过三轮教学实践和行动研究的修正完善，形成了基于同伴互评的五阶段多维能力培养的专业英语演讲教学模式。研究表明，此模式激发了学生运用英语开展专业交流的兴趣与积极性，提高了学生运用英语对专业问题协商的深度和意义建构的质量，增强了学生应用专业英语的理解能力、思维能力、综合表达能力和评鉴能力等能力。

【关键词】同伴互评 专业英语 演讲 教学模式

Reconsidering Language Assessment Training under the Framework of Teacher Education: Focusing on Assessment Contexts, Practices and Teachers

Zhang Cong
Shandong University

Yan Xun
University of Illinois at Urbana-Champaign

Recent literature in language testing and teaching witnesses burgeoning interests in landscaping the construct of language assessment literacy and developing assessment training for language teachers. However, current assessment training tends to follow a top-down applied-science model, targeting language testers and teachers indistinguishably and focusing heavily on measurement theories and statistical skills, with less attention to teachers' practical needs. This case study, situating language assessment training under a teacher education framework, examines assessment practices and training needs of EFL teachers in an urban secondary school in China.

This case study contextualized AL/LAL and examined assessment practices of language teachers from five data sources:

- 1) Top-down assessment policies and their impact on assessment practices at the school;
- 2) EFL teachers' experience of assessment development and use;
- 3) Content and psychometric quality of a teacher-developed English final exam;
- 4) EFL teachers' experiences with teacher education and assessment training; and
- 5) Perceived needs in assessment training by the school principal, head of the English teaching unit, and regular EFL teachers.

Triangulated data across multiple sources revealed unique characteristics in the (development of) assessment literacy skills among language teachers, which differ from those of prospective language testers in fundamental ways. Based on teachers' interests, strengths, and constraints in language assessment, a reflective approach to assessment training is proposed, which situates training in the local assessment context, creates opportunities for language teachers to reflect upon their assessment practices, and encourages teachers to develop theoretical knowledge of language assessment in an inductive manner. We argue that, due to the differences in professional preparation between language testers and teachers, language teachers are more likely to benefit from a reflective approach to assessment training than an applied-science approach.



How do Teachers and Students Perceive an Enhanced Score Report of EFL Reading Test

Li Xueping

Guangdong University of Foreign Studies

As an important medium to communicate test results to teachers and students, score report has been expected to provide detailed information about students' strengths and weaknesses in the context of learning-oriented testing and assessment, so that remedial actions can be taken to help students improve in a particular domain area. However, previous studies have shown the ineffectiveness of current score report practices in multiple aspects. This study proposes to apply Fusion Model, one of the cognitive diagnostic models, to the development of enhanced score reports of a standardized EFL reading comprehension test, hoping to facilitate teaching and learning by means of the detailed diagnostic information about students' EFL reading ability that can be obtained from the process of cognitive modeling. Specifically, this study will focus on the perception of teachers and students towards the enhanced score report in terms of clarity and usefulness. Firstly, several teachers and students will be provided with a sample student score report that contains diagnostic information obtained from fusion modeling and is adapted by the author from previous studies, together with a traditional score report that contains only total score and subscores; then, interviews will be conducted on a one-by-one basis to probe how each teacher and student think about the format and information of the sample report; and thirdly, opinions of teachers and students will be analyzed and reported both qualitatively and quantitatively. It is expected that despite of some concerns, both teachers and students hold positive attitude toward the enhanced score report.

Exploring Teachers' Conceptions and Practices in Assessing Young EFL Learners in the Classroom

Qiaozhen Yan Lawrence Jun Zhang Helen Ramsey Dixon
The University of Auckland

Classroom assessment for a long time has only been an offshoot of external standardized testing in the fields of both general education and second and foreign language (L2) assessment. It is not until around the 1990s that there has been a shift in the focus of attention away from the psychometrics of standardized testing and to the interactions between assessment and learning in the classroom. It is argued that the preliminary purpose of classroom assessment should be the improvement of student learning, being formative in nature (Crooks, 1988; Black & Wiliam, 1998). Since the landmark review by Black and Wiliam (1998), researchers have sought to address the characteristics of formative classroom assessment (cf., Brookhart and Durkin 2003, 27-54; Ingersoll 1999, 26-37; Stiggins and Conklin 1992; McMillan 2001, 20-32; McMillan 2007, 1-7; McMillan, Myran, and Workman 2002, 203-213; Saefurrohman and Balinas 2016, 82-92). These studies have addressed the complexity of classroom assessment. Nonetheless, comprehensive studies on the characteristics of classroom assessment and on the alignment of teachers' conceptions of classroom assessment and their practices are still lacking. Our literature review shows that most studies are concerned with students in secondary schools and tertiary institutions. Young EFL learners as a special learner group have received little attention. Given the increasing number of young learners learning an EFL worldwide and the promotion of classroom assessment, an understanding of how young EFL learners are assessed in the classroom is essential. This study intends to fill this research gap by exploring teachers' conceptions and practices in classroom assessment of young EFL learners within the Chinese testing culture context.

Specifically, this study aims to look into how EFL teachers conceptualize classroom assessment and implement assessment practices for young EFL learners. It also aims to examine how teachers' conceptions play out and align with their assessment practices by adopting a mixed methods research design, incorporating two phases. In the first phase, quantitative data will be collected through a teacher questionnaire with 200 primary school teachers in Guangzhou and Chongqing respectively. In the second phase, qualitative data will be collected through a case study involving three EFL teachers selected purposively from the 200 teachers who have answered the teacher questionnaire. Data collection methods used in this phase include classroom observation and post-observation semi-structure interviews.

The preliminary findings of Phase one survey study will be discussed and implications for teaching, learning, and assessment for young EFL learners in the Chinese context and beyond will be considered. It is hoped that valuable insights will be gained into the nature of teachers' conceptions of classroom assessment, the characteristics of their assessment practices, and the manner in which teachers' conceptions play out in their practices.



Classroom Assessment Oriented to Self-regulated Learning

Tang Xiongying
Jiangxi Normal University

Abstract: After decades of verifying the effect of new assessment tools on learning and understanding features of assessment for learning, research in classroom assessment currently relates itself much to students' self-regulated learning (SRL). Drawing on studies in educational psychology, this presentation reviews SRL on its definitions, affecting factors and developing process and discusses the implications for classroom assessment. It also looks at the existing preliminary studies of classroom assessment oriented to SRL and sees the opportunities of practicing such classroom assessment at the levels of curriculum, lesson, and minute-to-minute feedback to students' performance in class as well. It concludes that classroom assessment enhancing students' SRL could become an important and worthy area to explore in future studies of assessment for learning.

Seeking Alternatives: Incorporating Teacher-based Assessment in a Spoken English Program

Sun Hang
Shanghai Jiao Tong University

Recent years have witnessed an increasing popularity of formative assessment. Contrasted with summative assessment, formative assessment is used to evaluate students' language proficiency in the process of learning and to foster students' continuous learning in terms of language assessment. Teacher-based assessment (TBA), one of the most important alternatives in formative assessment, emphasizes the assessment is done by the actual teacher, who involves in the first place in students' learning and can be defined as non-standardized assessment carried out by teachers in the classroom. Engaged in the classroom, teachers have a better grasp of students' learning strengths and weaknesses. Thus, they can combine multiple assessment choices with instruction within everyday classroom context. Many researchers argue that language teaching will be more effective if assessment is integral to instruction and the activities structured in class. This research paper aims to report the author's teaching practice throughout the semester within the context of the classroom, where teacher-based assessment and instruction were woven together. The teaching site was in a non-profit organization – International House – in Philadelphia, USA. Nine students, aged from 19 to 37, coming from eight different countries were enrolled in the spoken English program. Task-based language teaching approach was used during the semester. In this paper, the author will discuss the current issues in TBA, including reliability, validity, practicality, authenticity and washback of TBA, and the teaching project. In doing so, She will elaborate on how she implemented TBA in class by providing examples from her field journals which included lesson plans, observation and reflection, and the outcome and limitations of the research, as well as the pros and cons of alternative assessments.

大学英语口语教学动态评估研究

任玲玲

洛阳师范学院

【摘要】本文以Vygostky心智社会文化理论为依据，在参考国内外评估理论最新成果的基础上构建了一个口语教学动态评估框架。依据此框架展开班级及个案研究。主要采用实验验证、量化分析和微变化分析等方法，通过课堂教学实践、观察、访谈、调研等形式实施，证明集体教学干预和对话式的教学使教师与学生、学生与学生之间形成良好的互动，不但有助于提高学生口语兴趣，而且能提高学生的口语能力。该模式的实施将有助于改变当前我国大学英语口语教学终结性测试为主导的局面，减少目前评价方式对学生英语学习的负面影响，使评价朝过程性、发展性、多元化方向发展，重视针对性的学习策略干预及介入资源建设，及时为学生提供教学补救措施，从根本上促进学生英语口语学习能力的发展。其主要创新之处在于：、把DA理论引入了以汉语为母语的外语学习者的大学英语口语教学实践；构建大学英语口语教学的DA理论框架及应用模式；设计系统的外语口语学习支持性介入手段，检验DA在大学英语口语教学领域中的可行性。总之，本研究将为大学英语口语教学评价体系的改革提供更多的选择，也将为整体外语教学改革探索新方法和新途径。

大学英语课程中“人际测评+人机测评”模式的构建与实证研究

印 蕾

江南大学

在当前大学英语课程评价体系中，形成性评估+终结性考试模式广泛被纳入高校大学英语测试与评价改革体系中并被赞誉，但却缺乏对其有用性或质量进行相关的实证研究。本研究基于Bachman & Palmer的“测试有用性框架”中的六项属性，通过定性与定量相结合的方法，对大学英语课程中“人际测评+人机测评”这一模式进行六个方面的分析：（1）分析“人际测评”+“人机测评”模式中测评结果的稳定性；（2）分析该模式在多大程度上对测试分数所做的解释是有意义和恰当的；（3）分析对该模式在语言测试任务特征和目标语言使用特征之间的吻合程度；（4）分析在该模式运用过程中，学生的哪些个体特质参与其中及被激活的程度如何；（5）分析对该模式在学生思维、语言和技能上产生的后果；（6）分析对该模式在设计和使用过程中所需的资源和可用资源的关系。本研究一方面，将在实践中运用并发展Bachman & Palmer的“测试有用性框架”理论；另一方面，检测这个模式在教学测评中的优势与不足，从而提高大学英语评价体系的设计与质量。

The Application of Automated Scoring System in the Teaching of ESL Writing

Xu Shasha

Zhejiang University of Finance & Economics

Based on blended learning theory and modern educational technology, automated scoring system has made it possible to investigate the influence of ESL learners' self-corrections on writing quality and writing scores. Through SPSS analysis of data of 58 argumentative writing collected from Pigai network and in-depth interview, the paper concludes that self-corrections significantly increase the writing scores and the writing quality in terms of composition length, word variety and number of clauses. The application of automated scoring system in the teaching of ESL writing has certain limitations, for instance, self-corrections motivated by error type feedback are largely focused at single-word and inter-word level rather than sentential level. The results have implications for future research as well as for further development of automated scoring system.

An Empirical Research into Reliability and Validity of China's AES Pigai and iWrite

Wang Hai-jun

Huang Qian

Zhijiang College

Zhejiang University of Technology

In recent years, AES has gradually sprung up in China, but with little empirical research on its reliability and validity. The research, a case study of Pigai and iWrite, carried out a thorough research into their reliability and validity, aiming to provide data support and constructive suggestions for AES developers and users such as teachers, students and writing test developers.

149 written essays were stratified sampled, with narrative, expository and argumentative essays being 50, 50 and 49 respectively. They were submitted through Pigai and iWrite and were automated scored by these two systems and 4 experienced human raters. Reliability in this research is shown in such indicators as facility, standard deviation, discrimination and inter-rater reliability whereas validity is operationalized in such indicators as construct validity and divergent/convergent validity. The statistics used to analyze the data concerned include Pearson correlation, regression analysis and agreement, etc.

It is concluded that the overall reliability and validity of both iWrite and Pigai AES are too low to satisfy the needs of large-scale English writing tests, with the latter being a little better than the former. But the facility value and agreement between human raters and Pigai are acceptable. Therefore, Pigai can be used to assist the low-stake EFL English Writing classroom teaching.

Researching Task Difficulty of English News Listening-Based on Automatized Text Characteristic Analytical Tools

Pan Zhixin

Shanghai Jiao Tong University

Factors affecting listening task difficulty include characteristics of the input, activity and learner. Among them, input characteristics play the most decisive role. Following principles of task-based language teaching and testing, this study aims to investigate the input characteristics that affect the difficulty level of authentic listening comprehension tasks by adopting authentic listening materials and utilizing automatized text analysis tools online. The study focuses on the task of listening to English broadcast news, which is divided into two subtypes, short reports and longer features. The researchers downloaded 84 reports and 84 features from the Internet, and compiled 28 short-answer question listening tests. 900 undergraduate non-English majors took these tests and rated the difficulty level of these news materials, from which the difficulty measures of these tasks are obtained. Simultaneously, three online analytical tools – VocabProfile, Coh-Metrix 和 L2 SLA - are used to measure the lexical, syntactic, cohesive and semantic characteristics of these texts. Finally, correlation and multiple regression analyses are operated to identify the principal input characteristics that have significant influence on task difficulty. The results show that about 20 input characteristics affect news listening task difficulty. Among them, the difficulty level of short news reports is mainly determined by figure density, Flesch Reading Ease, noun phrase density, speech rate and Academic Word List percentage, while that of longer news features by Coh-Metrix L2 Readability, proportion of anaphor overlap in adjacent sentences and negative density. This result can guide material selection in listening class instruction, textbook compilation, and test development. It also sets a new practical research paradigm for authentic listening task difficulty research.



A Study on the Evaluation, Interpretation and Extrapolation of Pigai Automated Writing Evaluation System

Zhang Li
Shanghai Jiao Tong University

Abstract: A lot of research has been done on the validity of Automated Writing Evaluation (AWE) systems such as PEG, IEA, e-rater, IntelliMetric, etc. However, research on the validity of Pigai AWE system is rather scarce. This study is going to verify its validity in terms of evaluation, interpretation and extrapolation. Correlation and agreement were measured between scores given by the system and those by human graders to testify the validity of evaluation. Then, generalization was verified by measuring the correlations between three writing tasks and by conducting One-Way ANOVA and Post Hoc to compare the means of the scores for the three writing tasks. Finally, extrapolation was verified by measuring the correlation between the writing scores given by Pigai and scores for listening, reading, speaking and students' portfolios. It is found that there is a strong correlation between scores by Pigai and those by human graders, and the exact and adjacent agreement of scores (in terms of the five grading levels of CET writing) by humans and Pigai is high, showing the validity of evaluation. The correlation is also high between different writing tasks, showing a certain validity of generalization, but lower than that reported with the similar AWE systems abroad. Writing scores by Pigai are also correlated to scores for listening, reading and portfolios, with higher correlations than most other AWE systems abroad can display, which demonstrates the validity of extrapolation. But writing scores by Pigai are not correlated to scores for speaking, about which the researcher gives the explanation. The study provides different perspectives of testing the validity of Pigai AWE systems, which has a significance to its development, application and improvement.

Key words: Automated Writing Evaluation, validity, evaluation, interpretation, extrapolation

The Impacts of Response Format On Test-takers' Reading Comprehension Test-taking Process—Based on Eye-tracking Evidence

Kong Jufang
Zhejiang Normal University

Reading comprehension competence is an indispensable component covered in many large-scale English tests both at home and abroad. The assessment of reading comprehension competence is realized through the design of specific tasks for test-takers to solve since reading comprehension cannot be observed directly. This study adopts an intra-subject response format×question type design and intends to explore the potential impacts of response format on test-takers' reading comprehension test-taking process based on the eye movement data recorded by Eyelink 2000. The findings are expected to have implications for reading comprehension test design and reading classroom teaching.

Linguistic Complexity as Indicator of EFL Writing Quality

Li Hang
Zhejiang University

Complexity, as a basic descriptor of L2 performance and an indicator of L2 proficiency, has been extensively researched in the fields of second language acquisition and language testing. However, due to the various ways complexity is defined and operationalized, studies on complexity have produced mixed and sometimes contradictory results. Based on Bulté and Housen's (2012) taxonomy of complexity constructs, the present study defines linguistic complexity, i.e., lexical and syntactic complexity, as both a dynamic property of the learner's L2 system at large, and a more stable property of the individual linguistic items, structures or rules that make up the learner's L2 system. To find measures for assessing syntactic and lexical complexity that are sensitive and non-overlapping with respect to writing by EFL learners at intermediate level, a total of 210 argumentative essays on two different topics by Chinese non-English major students are analyzed. On the basis of previous theoretical discussions and empirical studies, a number of syntactic and lexical complexity measures gauging both diversity and sophistication of L2 writing performance are selected. These include measures provided by such automatic analysis tools as Coh-Metrix (3.0), Lexical Complexity Analyzer (LCA), and L2 Syntactic Complexity Analyzer (L2SCA). Meanwhile, since subordination is the dominant means of syntactic complexity at intermediate stages (Norris & Ortega, 2009), subjective analysis and annotation of finite and non-finite clauses is also conducted to gauge syntactic diversity at the clause-level. A factor analysis is run on these indices to estimate which of the complexity measures might actually group together as underlying variables of syntactic and lexical complexity. Furthermore, measures that correlate strongly with human raters' judgment of L2 argumentative writing are identified. A stepwise regression analysis using measures that correlate most strongly with human judgment as independent variables is then carried out to predict the human ratings of essay quality. Findings of the study provide useful insight into the defining features of L2 writing by learners at intermediate level and the nature of various complexity measures; and carry practical implications for automated essay scoring.



Is syntactic Complexity Predicative of L2 Writing Quality? A Preliminary Study into Mean Dependency Distance as a Measure of Syntactic Complexity

Xu Lirong
Zhejiang University

Abstract: The study aims to explore the relationship between syntactic complexity and overall quality of L2 writings produced by college-level students in an English writing contest. More specifically, MDD (Mean Dependency Distance) in the framework of dependency grammar was adopted as a measure of syntactic complexity of L2 writings. Forty-one English essays (totaling 23404 words) written by college students from various majors and different universities were sampled and entered into Stanford Parser for part-of-speech tagging and dependency annotation. The treebank was built on the basis of Stanford Parser output first and then perfected through human error-detection and revision. Results showed a varying pattern concerning the relationship between MDD and writing quality (as represented by scores). Generally, MDD of all essays examined ranged from 1.74 to 2.76 and correlation was not found between the two variables in question. Possible reasons for the results were presented and discussed with reference to both second language acquisition and language assessment. Further study is needed to provide more empirical data concerning MDD across various proficiency levels due to the small sample size used in this study.

Key words: Syntactic complexity, MDD, dependency grammar, writing quality

A Corpus-based Analysis of Mandative Subjunctive Triggers In Chinese Learners' English Speaking and Writing

Lei Lei

South China University of Technology

This paper is devoted to the corpus-based study on the governing words (triggers) of Mandative Subjunctives and their elicited subordinate clauses including non-inflected forms, should-periphrasis, modal periphrastic alternants and the elicited infinitives. Data for this investigation is retrieved from the Corpus of Contemporary American English (COCA), the Spoken and Written English Corpus of Chinese Learners (SWECCCL) and the small-scale corpus of written English constructed by the author. The paper firstly attempts to compare the Raw Frequencies of the triggers and the elicited forms, and secondly assess the differences between the corpuses by SPSS software statistics, and thirdly design a questionnaire survey to perform a qualitative analysis. Research findings show that Chinese College English learners are able to cognize the functions of the triggers and their elicited forms to a certain extent and have formed their own usage patterns and habits that are not fully consistent with the native speakers'. The Chinese College English learners are more likely to use should-periphrasis than the non-inflected forms. For some triggers, the should-periphrasis are overused. Compared with non-inflected forms, should-periphrasis and modal periphrastic alternants, the infinitive forms are more preferred and are overused for some triggers while underused for some others. Malpractice of monotonous forms and the acquisition deficiency lead to low sensitivity to the subtle changes of contexts and pragmatic information. In most cases, students fail to develop a well-rounded view of the triggers and in consequence can only resort to stiff expressing ways, which is a more prominent problem in speaking than writing. Some syntactic constructions can be recognized and understood by students in questionnaire, but cases of these structures such as passive non-inflected forms are still very rare in students' output in both speaking and writing. Further study is needed to explore the cause of this phenomenon. The deficiency of this research lies in the situation that it is not all-inclusive. The indicative periphrasis and the gerund are not covered in this research, hence this paper cannot present a complete picture of the governing words. And more corpuses, such as BNC, LOB are not included in this research, which limits the research findings to American English as a reference system. The findings such as more frequent uses of should-periphrasis, which happen to conform to the tendency of British English, need to be compared against British English as a reference system. The research findings can also be instructive to teaching practice. The present input practices prove to be inefficient to cover as more branches of the language tree as possible, students repeatedly use certain forms and fail to attain to the native speakers' realm of richness, appropriateness, versatility and accuracy.



Connecting English Test Performance with Writing Teaching and Learning

Zhang Yumei Luo Shaoqian
Beijing Normal University

Abstract: Writing ability is one of the key abilities in language learning, and writing is also a difficult area for language teaching and learning. Scholars, especially in China, have realized students' writing problems in content and logic, and have been investigating the reasons or searching for solutions. Also, researchers have been studied on writing strategies and self-efficacy that are believed to influence writing performance. This research employs, firstly, a Likert-scale questionnaire to probe into junior high school students' English writing difficulties, self-efficacy and strategy application. Then the researchers design some writing tasks and analyze the potential relationships between writing performance and the results of the questionnaire. Thereby, implications for writing teaching and learning are elicited.

Key words: writing performance, writing difficulties, writing strategies, self-efficacy

A State-of-the Art Review of International Language Testing (2011-2015)

Zhou Shanshan
Beijing Information Science and Technology University

Abstract: The domestically published literature review of international language testing is subject-specific rather than comprehensive. In order to gain an insight into the state-of-the art of international language testing, this paper reviews all the research articles published from 2011 to 2015 on Language Testing and Language Assessment Quarterly, two leading journals in the area. The articles are reviewed from three aspects, i.e. research methodology, research subjects and research concerns. Research indicates that quantitative research is still the dominant methodology. However, qualitative and mixed methods are on the increase. EFL or ESL tests and adult EFL or ESL learners are mostly researched. There has been a variety of research concerns in the past five years. However, most issues which have long been the concerns of language testing professionals remain, such as validity and validation, construct of tests. Six emerging concerns are surveyed in detail, i.e. rater's behavior, integrated test, classroom-based test and school-based test, diagnostic test, assessment of pronunciation and involvement of stakeholders. The following trends are revealed and discussed. The concerns of language testers are changing from test development to test use, from test of learning to test for learning. The practice of involving stakeholders in the research of language testing indicates a growing concern for the social dimension of language tests. The concept of fairness of tests deserves further research. This paper contributes to a better understanding of status-quo of language testing for the domestic language testers.

Key words: language testing; literature review; emerging concerns; trends

基于测试方式的高中英语阅读理解

邢文骏¹ 何冰²

【摘要】本文从目前高考阅读检测方式的变化为依据，归纳高考阅读测试方式分类，分析了主观性和客观性阅读检测题型的利弊。阅读测试题型对语言水平的高的和低的学生影响都不大，区分度最大的是中等语言水平的学生。从影响英语阅读的变量入手，结合自己的教学实践，提出高考阅读检测是一个沟通的过程，是考生和作者、命题者对话的过程。考生的阅读表现固然由其语言图式（语言知识）决定，也和命题者的设题方式有关。文章怎么写很重要，怎么读也很重要。所以，要提高学生的阅读能力，我们固然要加强训练。但同时也需要注意测试方式的分类和层次研究，

【关键词】高考阅读、测试方式、阅读变量、教学思考

稳中求新的英语高考内容改革

马燕¹ 胡小力² 何冰³

【摘要】2016年全国普通高等学校统一考试英语试卷充分体现了考试改革的精神，坚持立德树人，加强社会主义核心价值观、依法治国和创新精神的考查，除北京、天津、上海、浙江、江苏等考试院命制本省市考生使用的试卷外，其他省市均使用教育部考试中心命制的三份试卷。每份试卷的整体难度适中，符合高校选拔人才的要求。试题内容有传承也有创新，听力理解的话题广泛且很生活化；阅读理解题材体裁广泛，均考查了《考试大纲》里规定的六项微技能；完形填空题材均积极向上，有励志激励的作用；短文填空和短文改错题材贴近考生生活，旨在弘扬中华优秀传统文化。2016年高考英语各份试卷在考生、家长以及社会中的反馈是积极的，同时也对中学今后的教育教学起到了良好的导向作用。

【关键词】英语高考、内容改革、导向作用



Index

Alister Cumming
CHEN Dajian
CHEN Defeng
CHEN Guangbin
CHEN Hui
CHEN Huilin
CHEN Jianlin
CHEN Wencun
CHENG Mengmeng
CHENG Yujing
CHU Jinfeng
DENG Jie
DONG Lianzhong
DU Wenbo
DU Yuxia
Ellen Head
FAN Jinsong
FANG Xiucai
FENG Li
FENG Meiqing
GAO Miao
GAO Shuling
GAO Xiao
GAO Xiaoying
GAO Yuan
GE Shili
GE Xiaohua
GU Xiangdong
GUAN Xiaoxian
GUO Mingming
HAN Chao
HAN Jiaxun
HE Bing
HE Jiawen
HE Lianzhen
HE Qiong
HE Xiaoyang
Helen Ramsey Dixon
HONG Run
HOU Yanping

HU Xiaoli
HUANG Jing
HUANG Qian
HUANG Ye
HUANG Youwen
JI Xiaoling
JIA Qing
JIA Yidong
JIANG Yan
JIANG Yizheng
JIE Wei
JIN Mu
KONG Jufang
KONG Xiang
Lance Knowles
Lawrence Jun Zhang
LEI Lei
LI Hang
LI Jiuliang
LI Lian
LI Qinghua
LI Shaolan
LI Xuelian
LI Xueping
LI Yue
LI Yulong
LIANG Junying
LIANG Li
LIN Dunlai
LING Yuyu
LIU Baoquan
LIU Beibei
LIU Chang
LIU Fan
LIU Jianda
LIU Jing
LIU Liping
LIU Sen
LIU Shuhui
LIU Yang

LIU Yiguang
LIU Ziyi
LU Lingwei
LUO Shaoqian
LV Shenglu
LV Yunhe
LV Zhouyang
MA Xiaomei
MA Yan
MENG Yaru
MIN Shangchao
Nick Saville
NIU Jia
PAN Mingwei
PAN Zhixin
PENG Chuan
PENG Kangzhou
PENG Zhiyao
Philip Horne
QI Xin
QIAN Xiaofang
QIAO Hui
Qiaozhen Yan
QIU Siqi
QUAN Dong
REN Jie
REN Lingling
RUAN Jifei
SHEN Mengting
SHI Yali
SUN Hang
SUN Youxia
TANG Xiongying
TIAN Wanning
Vivien Berry
WANG Feiyu
WANG Haijun
WANG Haiping
WANG Hua
WANG Jimin

WANG Jing
WANG Jun
WANG Shuang
WANG Wei
WANG Weiwei
WANG Yan
WANG Yizhen
WANG Zhanglong
WEI Liyan
WU Jiao
WU Sha
WU Xuefeng
WU Zhaohui
WU Zunmin
XI Xiaoming
XIAO Lihong
XIAO Wei
XING Wenjun
XIONG Lidi
XU Guozhu
XU Jiayong
XU Lirong
XU Shasha
XU Yun
Xun Yan
YAN Ming

YAN Yi
YANG Hongbo
YANG Lvna
YANG Zhihong
YANG Zhiqiang
YE Xiaoqing
YIN Lei
YOU Zhonghui
YU Chengyuan
YU Han
YU Jie
YU Wuzhe
YUAN Jing
ZANG Tiejun
ZHAN Quanwang
ZHANG Chunqing
ZHANG Cong
ZHANG Cong
ZHANG Di
ZHANG Fang
ZHANG Hao
ZHANG Hongxin
ZHANG Jian
ZHANG Jianshi
ZHANG Jie
ZHANG Jingjing

ZHANG Li
ZHANG Wenxia
ZHANG Wenxing
ZHANG Xiaoyi
ZHANG Xuan
ZHANG Yi
ZHANG Yujie
ZHANG Yumei
ZHAO Guanfang
ZHAO Haiyong
ZHAO Liang
ZHAO Ningning
ZHAO Qifeng
ZHENG Qun
Zhiming Yang
ZHONG Yu
ZHOU Chenglin
ZHOU Hong
ZHOU Jianhua
ZHOU Jiming
ZHOU Shanshan
ZOU Shaoyan
ZHOU Shuli
ZHOU Yanqiong
ZHU Jieqiang
ZOU Shen

Everyone thinks; it is our nature to do so. But much of our thinking, left to itself, is biased, distorted, partial, uninformed or down-right prejudiced. Yet the quality of our life and that of what we produce, make, or build depends precisely on the quality of our thought. Shoddy thinking is costly, both in money and in quality of life. Excellence in thought, however, must be systematically cultivated.

— Richard Paul & Linda Elder



THINKER'S GUIDE LIBRARY 思想者指南系列丛书

(美) Richard Paul (美) Linda Elder 等著



扫描二维码，立即购买

专业权威

深入浅出

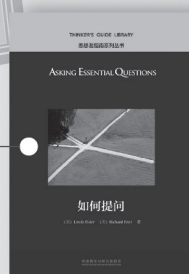
易于阅读

便于携带

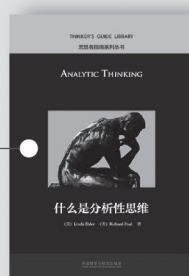
教学篇



大众篇



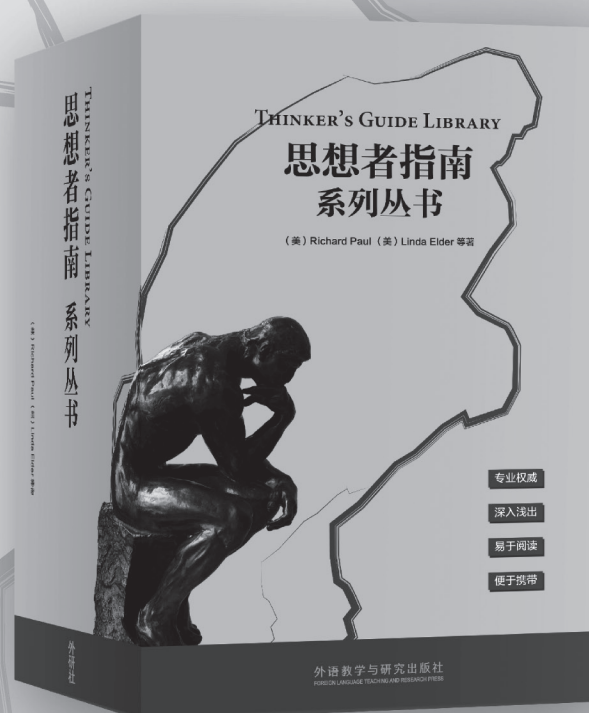
基础篇



套装版

定价：257.90 元

当当、亚马逊、京东均有售



“思想者指南系列丛书”由外研社原版引进，供读者培养、提升思辨能力（批判性思维能力）使用。作者 Richard Paul 和 Linda Elder 是两位专门从事思辨能力研究的专家，他们创办的专门研究和培训思辨能力的机构 Foundation for Critical Thinking 享誉全球。Richard Paul 和 Linda Elder 认为，思辨能力并不是玄虚的存在，而是有方法可依、有规律可循，他们将长期研究发现并总结的方法与规律凝聚在了“思想者指南系列丛书”当中。“思想者指南系列丛书”共 21 本，分为基础篇、大众篇、教学篇，分别针对入门基础学习者、社会大众读者、广大教师及学生阅读学习。

THINKER'S GUIDE LIBRARY

思想者指南系列丛书

(美) Richard Paul (美) Linda Elder 等著

思辨能力的高下将决定一个人学业的优劣、事业的成败乃至一个民族的兴衰。在此意义上，我向全国中小学教师、高等学校教师和学生以及社会大众郑重推荐“思想者指南系列丛书”。相信该套丛书的普及阅读和学习运用，必将有利于促进教育改革，提高人才培养质量，提升大众思辨能力，为创新型国家建设和社会文明进步作出深远的贡献。

——孙有中
北京外国语大学

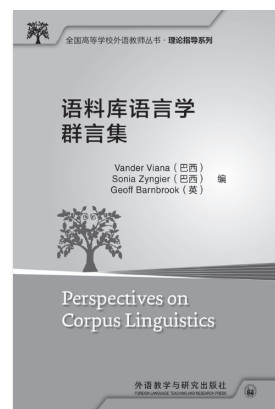
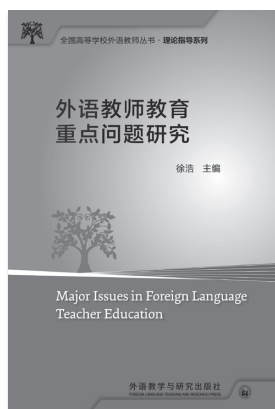
中 文 书 名	英 文 书 名	书 号	定 价 (元)
基础篇			
批判性思维术语手册	A Glossary of Critical Thinking Terms and Concepts	978-7-5135-7531-7	15.9
批判性思维概念与方法手册 (第7版)	Critical Thinking Concepts & Tools	978-7-5135-7832-5	7.9
大脑的奥秘	Taking Charge of the Human Mind	978-7-5135-7469-3	9.9
批判性思维与创造性思维 (第3版)	The Nature and Function of Critical & Creative Thinking	978-7-5135-7836-3	9.9
什么是批判性思维	The Guide to Critical Thinking	978-7-5135-7534-8	9.9
什么是分析性思维	Analytic Thinking	978-7-5135-7470-9	13.9
大众篇			
识别逻辑谬误	Fallacies: The Art of Mental Trickery and Manipulation	978-7-5135-7833-2	13.9
思维的标准	Intellectual Standards	978-7-5135-7535-5	13.9
如何提问	Asking Essential Questions	978-7-5135-7471-6	9.9
像苏格拉底一样提问	The Art of Socratic Questioning	978-7-5135-7530-0	15.9
什么是伦理推理	Ethical Reasoning	978-7-5135-7533-1	13.9
什么是工科推理 (第2版)	Engineering Reasoning	978-7-5135-7528-7	13.9
什么是科学思维 (第3版)	Scientific Thinking	978-7-5135-7834-9	13.9
教学篇			
透视教育时尚	Educational Fads	978-7-5135-7529-4	15.9
思辨能力评价标准	Critical Thinking Competency Standards	978-7-5135-7532-4	13.9
思辨阅读与写作测评 (第2版)	The International Critical Thinking Reading & Writing Test	978-7-5135-7835-6	13.9
如何促进主动学习与合作学习	Practical Ways for Promoting Active & Cooperative Learning	978-7-5135-7472-3	7.9
如何提升学生的学习能力	How to Improve Student Learning: 30 Practical Ideas	978-7-5135-7467-9	9.9
如何通过思辨学好一门学科	How to Study & Learn a Discipline: Using Critical Thinking Concepts & Tools	978-7-5135-7473-0	9.9
如何进行思辨性阅读	How to Read a Paragraph: The Art of Close Reading	978-7-5135-7466-2	13.9
如何进行思辨性写作	How to Write a Paragraph: The Art of Substantive Writing	978-7-5135-7468-6	9.9



全国高等学校外语教师丛书

“全国高等学校外语教师丛书”包括理论指导、科研方法、教学研究和课堂活动四个子系列。这是一套开放性的丛书，既有精心挑选的国外引进著作，又有特邀国内学者编写的专题论述，为教师教学与科研提供切实、全面、前沿的引导与支持。本套丛书的特色为：突出科学性、可读性和操作性，做到举重若轻，条理清晰，例证丰富，深入浅出。

最新出版



书名	作者	书名	作者
理论指导系列			
二语习得重点问题研究	文秋芳	英语教学中的行动研究方法	Anne Burns
英语文体学重点问题研究	张德禄 贾晓庆 雷 茜	应用语言学中的个案研究方法	Patricia A. Duff
外语教师教育重点问题研究	徐 浩	应用语言学中的质性研究与分析	杨鲁新 王素娥 常海潮 盛 静
外语学与教的心理学原理	张庆宗	应用语言学中的微变化研究方法	周丹丹
认知语言学与二语教学	文秋芳等	应用语言学论文写作指导：实证研究报告的撰写	John Bitchener
语用学与外语教学	陈新仁	第二语言研究中的统计案例分析	许宏晨
词汇研究	Norbert Schmitt	有声思维在外语教学研究中的应用（第二版）	郭纯洁
语言测评实践：现实世界中的测试开发与使用论证	Lyle Bachman Adrian Palmer	外语教学定量研究方法及数据分析	秦晓晴 毕 劲
语料库语言学群言集	Vander Viana Sonia Zyngier Geoff Barnbrook	应用语言学中的复制研究方法	Graeme Porte
科研方法系列			
语料库应用教程	梁茂成 李文中 许家金	教学研究系列	
第二语言研究中的数据收集方法	Susan M. Gass Alison Mackey	英语教学中的学习策略培训：阅读与写作	顾永琦等
第二语言研究中的问卷调查方法（第二版）	Zoltán Dörnyei Tatsuya Taguchi	英语写作教学与研究	徐 昉
第二语言研究中的启动研究方法	Kim McDonough Pavel Trofimovich	英语听力教学与研究	王 艳
		英语词汇教学与研究	马广惠
		英语阅读教学与研究	陈则航
		英语语言教学材料：理论与实践	Nigel Harwood
		反思性实践：重燃你的教学热情	Thomas Farrell
		课堂活动系列	
		演讲的艺术课堂活动教师手册	田朝霞 周红兵

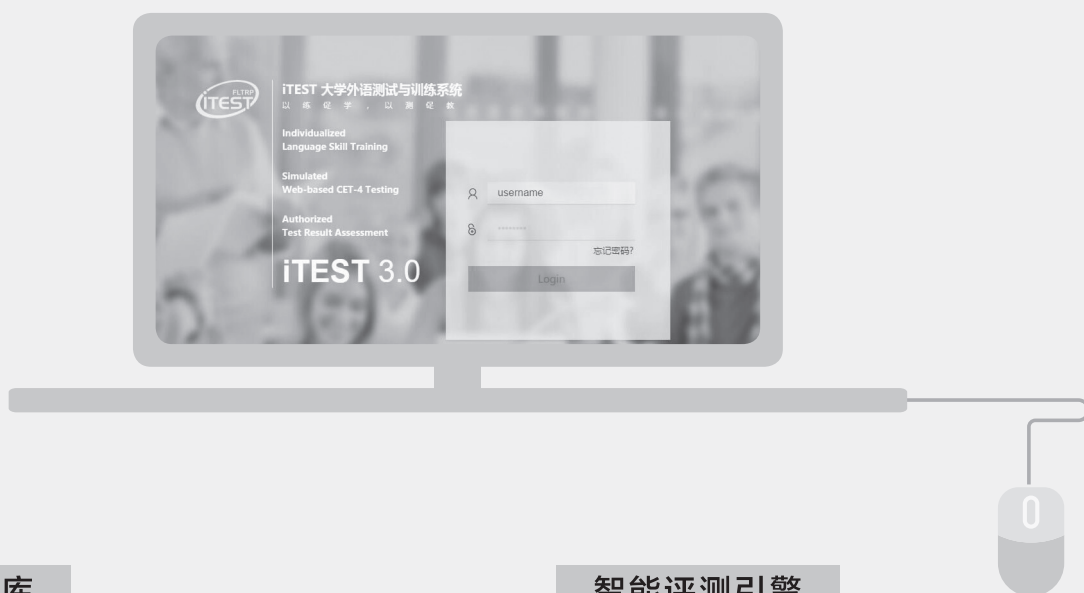
iTEST 3.0 大学外语测试与训练系统

“以练促学，以测促教”

iTEST. unipus.cn



iTEST 3.0 基于云服务的基础架构和大数据分析技术搭建，支持院校组织多种规模、各类教学场景下的考试与教学，支持学生随时随地进行基础训练与各类模考训练，为院校进行数字化教学评估提供专业的解决方案。



高质量试题库

- 外研社独家授权教材配套题库
- 引进柯林斯、多乐园雅思、托福模拟试题库
- “Unicomm 试题库联合共建项目”源源不断开发专业试题库

智能评测引擎

- 成绩统计与试卷分析双维度数据 智能生成 一键下载
- 写作、翻译、口语机器智能评测
- 写作：调用 iWrite2.0 智能批改引擎
- 口语：调用驰声科技语音智能评测引擎

在线测评管理

- 浏览器端（B/S）班级测试轻松发布
- 客户端（C/S）高利害考试 防作弊 数据双存储
- 公网模式：支持百万人同时在线考试
- 局域网模式：支持 5000 人同时在线考试

自建题库功能

- 三秒智能组卷
- 30 多套标签快速筛选试题
- 自动生成 AB 卷
- 自建题库，即建即用



外研社

客服电话：400-898-7008

客服邮箱：service@unipus.cn



外研社



高中生英语成长诊断学习系统

优诊学是由北京师范大学外语测试与评价研究所学术指导，由外语教学与研究出版社自主研发的一款英语在线诊学系统，包括诊断测试、智能练习两大部分，通过实施诊断→发现问题→提出建议→专项练习→稳步提高的测试模式，帮助高中生告别题海战术，有效提高英语水平。

系统特点：

- ✓ 微技能诊断，发现问题
- ✓ 练习推送，强化训练
- ✓ 原创试题，真实严谨
- ✓ 即时报告，科学高效
- ✓ 灵活组卷，个性辅导
- ✓ 测评算法，精准可靠

诊断对象：高中学生

诊断内容：阅读、听力、语言知识运用（语法、词汇）、写作



优师生之忧，
诊学习之感！

uzx.iceshi.org



咨询更多信息请联系：高老师 010-88819486

App下载 (Android)