Overseas English Testing: Pedagogy and Research 国外英语考试教学与研究, 2020, 2(4), 184-194
Published Online November 2020 in Hans. http://www.hanspub.org/journal/oetpr
https://doi.org/10.12677/oetpr.2020.24018

Hans汉斯

《传承性与创新性：基于证据的六级、雅思、托福考试效度对比研究》引介

辜向东

重庆大学，重庆
Email: xdgu@cqu.edu.cn

收稿日期：2020年11月8日；录用日期：2020年11月16日；发布日期：2020年11月30日

摘　要

本文系国家社科基金重点项目结题成果《传承性与创新性：基于证据的六级、雅思、托福考试效度对比研究》(14AYY010)的引介，由引言和结语构成，是项目研究的整体设计、具体实施与完成的全局性概览。该研究以“社会－认知效度验证框架”为理论基础，使用了效度研究一些传统的研究方法，如问卷调查、有声思维等，也尝试了一些创新性的研究方法，如数据挖掘、眼动追踪等。该研究成果不仅丰富了六级、雅思、托福三项考试的效度证据，为教育、人事部门及广大利益相关者提供入学、就业、人才流动等决策依据，而且为其他语言测试效度对比研究提供了思路与方法上的借鉴。

关键词

效度，基于证据的“社会－认知”语言测试效度验证框架，六级，雅思，托福

An Introduction to *Continuity and Innovation: An Evidence-Based Comparative Validation Study of CET-6, IELTS and TOEFL iBT*

Xiangdong Gu

Chongqing University, Chongqing
Email: xdgu@cqu.edu.cn

Received: Nov. 8th, 2020; accepted: Nov. 16th, 2020; published: Nov. 30th, 2020

Abstract

This paper is an introduction to the accomplishment of the key research project of the National Philosophy and Social Sciences Foundation of China *Continuity and Innovation*: *An Evidence-Based*

文章引用: 辜向东. 《传承性与创新性: 基于证据的六级、雅思、托福考试效度对比研究》引介[J]. 国外英语考试教学与研究, 2020, 2(4): 184-194. DOI: 10.12677/oetpr.2020.24018

Comparative Validation Study of CET-6, IELTS and TOEFL iBT **(Fund No. 14AYY010). It consists of the introduction and the conclusion parts, which provides an overall glimpse of the whole research design and its implementation processes. This study, utilizing the evidence-based "Socio-cognitive" Framework for test validation as its theoretical basis, employs some traditional validation study methodologies, such as questionnaire survey and verbal protocol, and explores some innovative technologies such as data-mining and eye-tracking. The research findings not only enrich the validation evidence of the three tests, CET-6, IELTS and TOEFL iBT, for decision-making references for study, employment and migration for education, human resources and other relevant stakeholders, but also offer references for comparative validation studies of other language assessments in methodologies and ways of thinking.**

Keywords

Validity, Evidence-Based "Socio-Cognitive" Framework for Test Validation, CET-6, IELTS, TOEFL iBT

1. 引言

国家社科基金重点项目“基于证据的四六级、雅思、托福考试效度对比研究”(14AYY010)于 2014 年 6 月立项。课题组随即开展了长达五年六个月系统深入的调查与研究。本专著为该项目的结题成果。

本专著涉及的大学英语四六级、雅思、托福三项考试是中国乃至全球极具代表性的语言考试。三项考试规模大、风险高、影响广。四六级考试包含四级和六级两个难度级别；雅思分学术类和培训类两种考试用途；托福有纸笔考、机考和网考三种考试形式。在本研究中，我们选择的分别是六级、雅思学术类、托福网考三项考试。因为六级与四级相比在难度上更加接近雅思学术类和托福网考；相较于培训类，雅思学术类更多用于升学考试用途，这与托福考试用途更接近；而网考是托福三种答题形式中最普及、使用最多的考试形式，所以我们选择六级、雅思学术类、托福网考三项考试进行对比。为了表达简洁，在本专著中，六级、雅思学术类和托福网考三项考试分别简称为：六级、雅思、托福。

本引言将首先简要概述国内外效度研究的现状，包括三项考试的研究及存在的不足；然后简要介绍本选题的价值和意义、研究的基本观点及主要内容；最后对结题成果进行概述性说明。

1.1. 国内外研究现状述评

1.1.1. 效度理论

效度(validity)是测试评价中最重要的考虑因素[1]。较早的效度定义为“一项测试是否测量了它所要测量的东西”[2]。自 20 世纪 60 年代以来，语言测试与评价的研究一直围绕效度展开[3]，效度理论取得了从“分类效度观”到“整体效度观”的重大发展。分类效度观[4]认为效度可分为效标关联效度、内容效度、构念效度等多种类型[5]。该效度观证操作性强，但比较零散，且未考虑分数使用和解释等方面的证据。整体效度观给出了具有突破意义的效度定义，即“对经验证据和理论依据在多大程度上支持分数的解释与使用进行的综合评价就是效度”[6]。这种“一元多维”的效度观确定了构念的核心地位[7]，明确了效度验证的对象是测试结果的解释和使用[8]。

1.1.2. 效度验证模式

整体效度观给语言测试的开发与研究带来了重大变革，但由于该理论高度概括且过于抽象，使效度

验证缺乏可操作性。近年来更多的语言测试学家根据该理论提出了一些具体的效度验证框架。其中较有影响的效度验证框架如下：

“交际语言能力模型”和“测试方法层面框架”[9]为效度验证开启了新视角。Bachman *et al.* [10]运用该框架对剑桥熟练英语证书考试、第一英语证书考试和托福考试三项考试所考查的能力和测试任务特征做了分析，并对该框架进行了完善。

“测试有用性框架”[11]涵盖信度、构念效度、真实性、交互性、考试影响和可行性六个质量属性，进一步阐释了[6]的效度理论。该框架可操作性强，但质量属性之间的关联不甚明确[12]。

“基于论证的效验模式”[13]与整体效度观一脉相承，包括两个步骤：提出效验观点、收集有关证据。Chapelle *et al.* [14]运用该模式论证了托福的效度。

“测试使用论证框架”[15]发展了 Kane [13]的效度论证观。该框架遵循“事实→主张”的推理机制，包含构建与评价两个过程[16]。不过其架构(后果、决策、解释、测试记录)比较抽象，能否成为指导测试开发与使用的新范式尚需检验。

“基于证据的效度验证框架”[17]从社会－认知视角出发，涵盖多个方面的效验证据，具有很强的可操作性，在剑桥五级主体英语证书考试的效度对比研究中得到丰富和完善[18] [19] [20]。

目前关于效度理论和验证模式的研究主要集中在国外，国内类似的研究还处于起步阶段，主要是对国外相关领域的发展进行引介和评述[12] [21]。

1.1.3. 六级、雅思、托福三项考试的研究及存在的不足

三项考试相关研究比较丰富，主要涵盖以下方面(括号中的文献仅为部分举例)：

六级的整体效度研究[22] [23]、各单项技能及题型研究[24] [25] [26]、评分与网考研究[27] [28] [29]、反拨效应及考试影响研究[30] [31] [32]。

雅思的开发及效度验证[33] [34]、考官与评分[35] [36]、反拨效应及考试影响[37] [38]。

托福的效度论证[14] [39] [40]、网考设计[41] [42]、公平性与可及性[43] [44]、评分与技术应用[45] [46]、信度与可推广性[47] [48]、分数解释[49] [50]。

尽管关于三项考试研究的文献比较丰富，但能够将这些研究组织起来并形成有关联和强有力的论证文献较为缺乏，而且涉及三项考试中任何两项的考试效度对比研究，尤其是实证研究也相当少。现有的对比研究多集中在分数等值方面[51]，但事实上还有其他很多方面需要对比，如受试的认知过程、考试的影响等。此外，几乎没有文献将国内的考试与国际权威考试进行较全面的效度对比研究，现有的文献只是就两项或三项考试的某一技能、题型或考试媒介等做初步探讨[52] [53] [54] [55]，因此，针对三项考试全面系统的效度对比研究亟待开展。

1.2. 选题的价值和意义

学科理论与实践价值：理论上，验证“基于证据的社会－认知效度验证框架”在考试效度对比研究中的可行性，并进一步构建更加科学合理的语言测试效度对比研究模型。实践上，通过对比三项考试的效度，形成将三项考试关联起来的论证。这不仅可以丰富考试对比研究领域的文献类型，而且能为类似的研究提供思路和方法上的借鉴。

社会和现实意义：一方面，本研究有助于推动我国语言测试开发与研究的国际化水平，有利于提升我国自行开发的英语考试在国际上的认可度，为教育、人事部门及广大利益相关者提供入学、就业、人才流动等决策依据；另一方面，本研究中的雅思和托福考试均已实现与国际公认的语言能力标准 CEFR (Common European Framework of Reference for Languages: Learning, Teaching, Assessment) [56]和我国的

《中国英语能力等级量表》的对接[57]，其开发与使用遵循了国际公认的语言测试标准。因此，三项考试的效度对比研究有望为《中国英语能力等级量表》的应用与推广提供参考数据。

1.3. 研究的基本观点

尽管六级、雅思、托福三项考试的目的、性质、构念、分数解释和结果使用等诸多方面存在不同，但三项考试都是以英语为外语或二语的大规模、高风险语言考试，受试将接受或正在接受高等教育，三项考试的效度应该具有可比性，三者的效度应该既有较大的相似性，也有一定的差异。而实际情况是否如此，有待进行全面深入的实证研究。

1.4. 研究的主要内容

本课题的理论基础为“基于证据的社会－认知效度验证框架”(Evidence-based Socio-cognitive Framework for test validation) [15]，该框架最初认为效度验证需要收集五个方面的效度证据：基于理论的效度、情景效度、评分效度、效标关联效度和后果效度。在剑桥五级主体英语证书考试的效度对比研究中，该框架得到丰富和完善，基于理论的效度更名为认知效度，受试特征也成为效度验证证据的一个重要方面。因此，最新的“基于证据的社会－认知效度验证框架”认为效度验证应该收集六个方面的效度证据：受试特征(test taker characteristics)、情景效度(context validity)、认知效度(cognitive validity、评分效度(scoring validity)、校标关联效度(criterion-related validity)和后果效度(consequential validity) [18] [19] [20]。

受试特征指受试生理、心理和经历特征。情景效度取代的是传统意义上的内容效度，指测试任务在多大程度上代表了该任务所取样的全域(universe)。认知效度指测试任务在多大程度上引发了考生在真实语言使用中相似的认知过程。评分效度被纳入效度整体概念的一部分，代替的是传统的信度，它回答的问题是测试分数在多大程度上是可靠的。校标关联效度指测试本身以外的效度证据，如一项考试与其他测量相同构念且已得到公认的有效测试或测量的相关程度。后果效度指测试过程及测试结果对所有相关人员产生了什么影响，包括宏观的后果(如对机构、社会的影响)和微观的后果(如对考生、教师的影响)。由于本课题主体是基于考生的证据，而且是关于三项考试的效度对比研究，即多方面的效度证据收集将包含受试特征和校标关联效度，因此，受试特征和校标关联效度在本研究中没有单独列出。

本课题研究从“基于证据的社会－认知效度验证框架”出发，从情景效度、认知效度、评分效度和后果效度四个方面对六级、雅思、托福考试进行了全面深入的效度对比研究。这些研究回答了一个总的研究问题：六级、雅思、托福三项考试的效度有何异同？图 1 为本课题的研究概览。

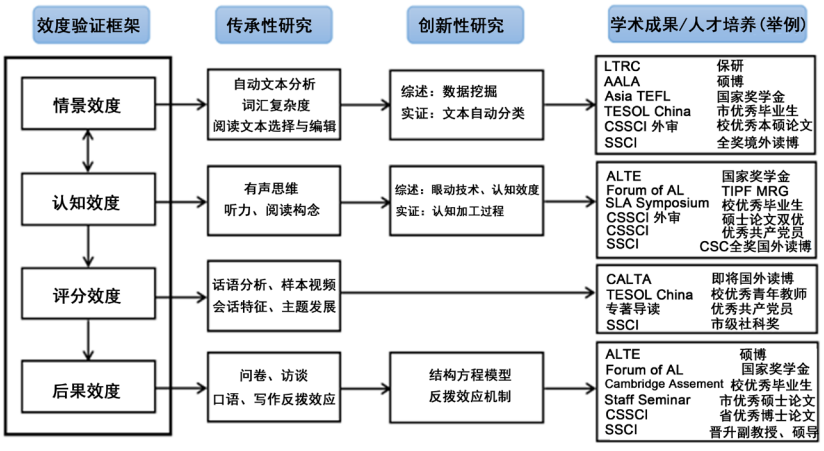

Figure 1. An overview of the evidence-based comparative validation study of CET-6, IELTS and TOEFL iBT
图 1. 基于证据的六级、雅思、托福考试效度对比研究概览

1.5. 结题成果目录

2. 结语

本课题以“基于证据的社会－认知效度验证框架”[15]为理论指导，从情景效度、认知效度、评分效度和后果效度四个方面，开展了基于证据的六级、雅思、托福考试效度对比研究。一方面，本课题运用语言测试效度研究普遍使用的研究方法(自动文本分析、有声思维、话语分析、问卷调查和半结构式访谈)做了八项传承性研究，主题涉及三项考试阅读文本词汇复杂度、阅读文本选择与改编、听力长对话和阅读测试受试有声思维认知过程、口语测试样本视频考官和考生会话特征和主题发展、三项考试的口语和写作测试对考生的反拨效应；另一方面，又尝试使用语言测试领域近年较新的跨学科研究方法(数据挖掘、眼动技术、结构方程模型)做了三项创新性研究，主题涉及文本自动分类、受试认知加工过程、写作反拨效应机制。

这些研究使用了丰富的数据：测试文本 900 余篇，受试作答题目 777 项(听力和阅读)，口语测试样本

视频 7 段；参与问卷调查、访谈、有声思维、眼动的考生和受试共 1000 余人次。这些丰富的数据充分体现了本课题的宗旨：基于证据的六级、雅思、托福考试效度对比研究。所有这些实证研究旨在回答一个总的研究问题：六级、雅思、托福考试的效度有何异同？

2.1. 主要研究结论

本课题研究结果表明，六级、雅思、托福三项考试均具有良好的效度。换言之，根据三项考试的成绩对考生英语语言能力做出的推论是可靠合理的。总体而言，雅思与托福的效度在各方面相同点较多，而六级的效度与二者相比差异性更大。三项考试效度的异同具体见表 1。

Table 1. Similarities and differences of the validity of CET-6, IELTS and TOEFL
表 1. 六级、雅思、托福考试效度的异同

效度研究	相同点	不同点
词汇复杂度		• 三项考试词汇复杂度各项指标均有显著差异(BNC 型符、类符除外)，托福词汇难度最高，雅思其次，六级最低；相反，六级词汇多样性最高，雅思其次，托福最低。
文本选择与改编	• 六级、雅思阅读测试选材话题覆盖面较广，且都选自原版杂志、报刊、学术书籍，语言真实性强。 • 六级、雅思阅读文本改编方式多样。两项测试文本改编前后在 BNC1000 词、范围外词、隐性衔接指标、二语可读性指标改革前后均有显著差异。	• 六级选材更青睐英语国家讨论国际时事的主流杂志与报纸，而雅思阅读文本大多选自关注科学技术发展或学术研究问题的书籍与学术性杂志。 • 雅思改编前后多项文本特征指标没有显著性差异，而六级改编后很多文本性特征(词汇、句法、显性衔接、文本抽象性、Flesh 易读度)有显著差异。
听力长对话构念	• 三项考试听力长对话主要测量了语法知识和认知策略及部分元认知策略，说明三项考试的构念效度较好。 • 三项考试听力长对话都未测量语用知识和社会语言知识，且都有构念无关知识和构念无关策略使用，对三项听力长对话的构念都形成了一定威胁。	• 托福构念的无关知识比例最高，六级其次，雅思最低。 • 托福听力长对话构念无关因素还涉及阅读策略。
阅读构念	• 三项考试均考查了单句、句间和段落三个信息层面的知识，且考查比例均按单句、句间和段落依次递减；三项考试都考查了受试理解细节、理解大意和推断的能力，表明三项考试均注重考查考生是否理解阅读材料并能读出言外之意。 • 三项考试受试的答题过程与预期答题操作的拟合度高，说明三项阅读测试均有较好的构念效度。 • 三项考试受试都有不同程度的不符合答题预期操作却选对正确答案、或理解错误却选对答案、或推理解释不清却选对答案的情况，说明存在对构念无关信息或技能的使用，这对三项考试的构念形成了一定威胁。	• 托福受试答题过程使用排除法的比例最高，六级理解错误但答对题的比例最高，雅思推理解释不清但答对题的比例最高。
会话特征	• 三项口语测试考官和考生都表现出较为丰富的话轮转换关联位置特征，均能使用多样化的话轮保持方式，且都能提供话轮。 • 三项口语测试都基本按照一问一答的序列结构展开，且大都按照毗邻语对的形式展现。 • 三项口语测试考官和考生都没有体现他人修正和他人启动－自我修正。	• 雅思考官和考生表现出的话轮转换关联位置特征和比邻语对最丰富，六级、托福其次。 • 雅思考官会发起自我启动－自我修正；雅思考生和托福考生同样会发起自我启动－自我修正，但是六级考生未发起自我启动－自我修正。

Continued

主题发展	• 三项口语测试的考官都会根据事先给定的题目发起提问，并使用明确指示标志来转移话题。 • 考生都会按照保持主题发展时长和发音准确性、词汇和语法使用多样性和准确性的要求发展主题。	• 雅思考官和考生管理主题发展的策略最丰富，六级、托福其次。
口语反拨效应	• 三项口语测试的考生都有中等偏强的成就性和工具性测试使用认识；有中等偏肯定的测试设计评价；有中等偏高的自我效能；有中等偏积极的反拨效应和中偏负面的反拨效应。 • 都倾向于认为语言技能和测试技能在三项口语测试中都很必要；考生语言技能发展在三项口语测试之间没有显著性差异。	• 六级与雅思、六级与托福在多项上存在显著差异，而雅思与托福之间相似程度很高。 • 六级与托福在工具性使用和对测试设计的评价方面没有显著性差异。
写作反拨效应	• 考生有较强的成就性工具使用和中等的工具性测试使用；对写作测试任务设计比较认可；有比较积极的反拨效应和中等偏弱的消极反拨效应；技能提升和记忆两项备考活动频率均为中等。 • 上述这些方面三项写作的反拨效应没有显著差异。	• 考生在非写作能力和备考管理方面均有显著差异，均为雅思最高，托福、六级其次。 • 六级与雅思、六级与托福在自我能力认知、自我效能、主观任务价值、备考过程投入多方面存在显著差异，但雅思和托福差异不显著。

2.2. 对策建议

除了上述主要研究结论，本课题为三项考试不同利益相关群体提出了以下针对性对策建议。

1) 考生：大力加强真实语言材料输入，阅读六级、雅思、托福考试文本来源报刊杂志、新闻网站、广播电台、学术教材、著作等，如《时代周刊》、《经济学人》、《卫报》、《纽约时报》、《新科学家》、《国家地理》等。加强实践性练习，切实提升语言综合运用能力。要深信“语言学好了，考试没问题”。促进成就性测试使用，增强能力自我认知，提升自我效能，加强社会情感策略和备考管理，加大学习投入等。

2) 教师：通过语言教学，培养学生人生胜任力(life competencies)。充分利用现有教学资源，更新教学内容。关注学生情感因素，如鼓励他们建立学好语言的信心，降低考试焦虑度，实现“三全育人”。

3) 考试设计者/考试机构/决策者：提升命题质量，全面测量受试的语言能力和策略能力，如在听力测试中加强语用知识和社会语言知识的考查，在阅读中加强语篇层次的考查，丰富考试题型。避免构念无关因素影响，如字面匹配、随机猜题、背景知识运用等。改善测试环境，提供高质量、有代表性、连贯性、完整性的样本视频、样题、备考材料等。提供明晰的评分标准，做到标准化与人性化的统一，确保考试的公平性、公信力和透明度。

2.3. 本课题的价值和意义

2.3.1. 学术价值

本研究探索、验证、丰富和发展了效度研究理论，充实、完善和建立了新的效度研究模型，尤其是三项考试效度对比研究的多个子框架或模型，比如听力长对话测试构念描述框架，三项阅读测试分析框架，三项口语考试考生反拨效应理论框架，写作考试考生反拨效应理论框架等，为今后其他大规模、高风险考试的效度研究，尤其是效度对比研究，提供了理论和方法上的借鉴。

2.3.2. 应用价值

本课题为考生的学习、教师的教学提供了富有建设性的意见和建议，为三项考试设计者/考试机构/决策者进一步提高命题质量、施测环境、评分标准等提供了具有针对性和可操作性的方案与决策依据，为语言测试研究者和工作者以及对此感兴趣的广大读者提供了思路和方法上的参考。

2.3.3. 社会影响

由于三项考试涉及的考生人数达数千万，对其效度的对比研究的社会价值难以估量。本研究有理论、有实践，有数据、有分析，论点鲜明、论据充分、论证有力。研究成果具有启迪性、说服力和实用价值，部分成果在国内高校、出版社、科研机构，如上海交通大学、武汉大学、四川大学、大连理工大学、东北师范大学、西安交通大学、外语教学与研究出版社、高教出版社等上百场学术讲座、工作坊上分享，而且在中小学国培计划、中小学外语教学与研究中得以应用，如词汇复杂度、阅读文本选择与改编等的研究思路与方法已经应用于高考、高中英语教材、报刊等的研究与中小学教学实践，并通过个人和机构的微信公众号、出版社网课等得以广泛传播，受益人次达数十万。

2.3.4. 社会效益

人才培养的国际化及可持续发展可能是本课题最大的社会效益。依托该课题研究，我们带动和培养了一批国际化的语言测试工作者和研究者，正在实现可持续发展。在研期间，课题负责人受聘世界一流测试机构(剑桥大学英语考评部)高级学术研究顾问。一位课题组成员应聘到国外高水平大学任专职研究员，四位团队成员获国家留学基金委全额奖学金境外、国外攻读博士学位，两位获得国际英语教育研究基金会首届硕士研究奖。课题部分成果在国际高水平会议上宣读，如国际语言测试界最高级别会议语言测试国际研讨会年会(LTRC)，欧洲语言测试者协会年会(ALTE)，亚洲语言测试者协会年会(AALA)，国际英语教师协会年会(IATEFL)等；在世界一流大学做专题研讨，如剑桥大学、纽约大学、香港理工大学等；部分成果在国际高水平期刊发表，如 *Applied Linguistics*，*TESOL Quarterly*，*Language Testing*，*Language Assessment Quarterly*，*Assessment & Evaluation in Higher Education*，*Innovations in Education and Teaching International*，*System* 等。

另外，课题组有三位成员从硕士生导师晋升为博士生导师，一位课题组成员博士论文获省优秀博士论文，两人硕士论文被评为市优秀硕士论文，五名团队成员获国家奖学金，多人次被评为优秀青年教师、优秀研究生、优秀毕业生、优秀共产党员。我们因此建立起了一支成长型和研究型语言测试团队，领衔了中国大百科全书语言测试词条的编制，成为了雅思、托福、普思、剑桥英语系列考试与《中国英语能力等级量表》对接的专家组成员，并申请获得国际合作语言测试研究基金项目、教育部人文社科项目、中央高校基金重大项目和跨学科项目等。因为该项目而建立的跨学科导师团队最近被评为市级研究生导师团队，正在实现可持续发展。课题负责人因其在科学研究和人才培养中的突出成绩，于今年获得国家留学基金委高级访问学者奖，将再赴剑桥大学访问与合作研究。课题组和团队期望通过人才培养的国际化和可持续发展，为我国语言测试与研究的国际化贡献力量。

2.4. 课题成果存在的不足

1) 文献综述和研究方法分布在各章

本课题设计的重心是基于证据的系列实证研究，课题组在设计的时候认为每项实证研究都会涉及相关的文献综述和研究方法，为了避免重复和方便专家审读，专著中没有单独设文献综述和研究方法两章。不过这样的谋篇布局可能给人印象两个部分不够集中或凸显。

2) 差异成因阐释不够深入

课题实证研究数据十分丰富，发现了三个考试效度的异同，尤其是差异，但对差异的成因阐释不够深入。三个考试的目的、构念、考试形式、考试题型、受试的水平等有诸多不同，因此对于考试对比发现的差异，其成因本身很难阐释。

3) 创新性研究还处于起步阶段

创新性研究在本课题中属于延伸性研究，尚处于起步阶段。因为主客观原因，创新性研究推进相当

艰难。主观原因在于跨学科难度大，耗时长；客观原因在于有些资源我们无法获取，比如数据挖掘文本分类中三个考试的样本太小。另外，有些设备迟迟不能到位，比如眼动设备因行政、财务管理缺乏灵活性和经费使用严格受限，至今没有到位，课题中的眼动实验设备是2019年暑假课题组临时租借的。

4) 三项考试的社会性影响研究不足

关于三项考试的社会性影响我们也进行了相关的研究，比如考试结果的使用，2015年我们调查了中国财富500强企业对三项考试结果的使用，发现只有12家企业对雅思和托福成绩有要求，而且要求参差不齐，其余全部是对四六级考试成绩的要求。我们很希望跟踪调查在“一带一路”倡议下，在雅思和托福与《中国英语能力等级量表》对接结果公布之后，中国财富500强企业对三项考试结果的使用是否有变化，如果有，变化的成因是什么。另外，我们对知乎问答社区上三个考试的考生上百万词的经验帖进行了初步分析，非常意外地发现三项考试考生的备考模式非常相似。只是这两项研究比较类似于调研报告而非学术研究，尚需更深入的调研，所以未将其收入此专著。

5) 研究受试的代表性不足

虽然课题组尽了最大的努力通过多种渠道(纸质、网络、考试、实验等)收集考生数据，但三项考试考生的样本量和代表性仍然很不足，因此未能将受试分为高中低不同水平组进行研究。此外，本课题研究对象未涉及三项考试其他利益相关者群体，比如教师、培训机构、教育主管部门等。

6) 平行研究成果呈现未能高度统一

本课题研究远比申报时预计的庞大和复杂，需要多个强有力的团队通力合作，而我们只有一个团队，且在成长中。团队由本科、硕士、博士、青年教师、研究员、副教授、教授组成，尽管我们考虑了效度对比研究在每个方面都尽可能有平行研究，以获取充分的证据并实现数据的三角验证，但平行研究相对独立，最后很难保证平行研究的逻辑、内容、思路、方法和结果的呈现高度统一。

2.5. 对未来研究的启示与建议

根据上述研究中存在的不足，我们对未来的考试效度对比研究提供如下启示与建议：

1) 高水平、高质量文献输入，穷尽性文献收集、系统性文献综述是效度对比研究必须完成的第一步，是做好顶层研究设计的基石。

2) 借鉴其他学科的理论、实证研究的方法，挖掘、分析、阐释效度对比研究发现的异同背后的成因。

3) 加强考试的社会性、公平性、考试结果的使用等研究，充分利用大数据方法，挖掘更多类型的语言测试数据，如论坛或贴吧上关于备考和考试的讨论以及经验帖、相关的考试政策和文件、历年考试真题、考生关于备考的日志或学习记录等。

4) 要从不同利益相关者收集证据，尤其要关注不同利益相关者的过程性证据。

5) 坚持不懈提升团队能力，尤其是跨学科团队建设与合作。效度研究是一个持续收集多方面效度证据的过程，一项考试的效度验证已经是一个系统工程，而考试效度对比研究涉及多项考试，其工作量更为庞大，需要强大和强有力的团队，尤其是跨学科团队的通力合作与支持。课题研究是高等教育，尤其是硕士、博士研究生培养必不可少的一部分，是培养学生创新思维、实践能力、团队精神和社会责任感的重要途径；坚持不懈加强团队建设，通过传、帮、带，既要提升团队的整体科研能力，又要培养团队不断创新、无尽求索的科研精神。团队、团队精神和科研精神是出色完成任何考试效度对比研究，甚至任何学术研究的前提条件。

基金项目

国家社科基金重点项目“基于证据的四六级、雅思、托福考试效度对比研究”(14AYY010)。

参考文献

[1] APA, AERA, and NCME (2014) Standards for Educational and Psychological Testing. Revised Version, American Educational Research Association, Washington DC.

[2] Kelly, T.L. (1927) Interpretation of Educational Measurements. New World Book Company, New York.

[3] Kunnan, A.J. (1998) An Introduction to Structural Equation Modelling for Language Assessment Research. *Language Testing*, **15**, 295-332. https://doi.org/10.1177/026553229801500302

[4] Lado, R. (1961) Language Testing. McGraw-Hill, New York.

[5] APA, AERA, and NCME (1966) Standards for Educational and Psychological Tests and Manuals. American Psychological Association, Washington DC.

[6] Messick, S. (1989) Validity. In: Linn, R.L., Ed., *Educational Measurement*, 3rd Edition, Macmillan, New York, 13-103.

[7] Alderson, J.C. and Banerjee, J. (2001) Language Testing and Assessment. *Language Teaching*, **35**, 79-113. https://doi.org/10.1017/S0261444800014464

[8] Davies, A., Hamp-Lyons, L. and Kemp, C. (2003) Whose Norms? International Proficiency Tests in English. *World Englishes*, **22**, 571-584. https://doi.org/10.1111/j.1467-971X.2003.00324.x

[9] Bachman, L.F. (1990) Fundamental Considerations in Language Testing. Oxford University Press, Oxford.

[10] Bachman, L.F., Davidson, F., Ryan, K. and Choi, I. (1995) An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study. Cambridge University Press, Cambridge.

[11] Bachman, L. and Palmer, A. (1996) Language Testing in Practice. Oxford University, Oxford.

[12] 韩宝成, 罗凯洲. 语言测试效度及其验证模式的嬗变[J]. 外语教学与研究, 2013, 45(2): 411-425.

[13] Kane, M.T. (1992) An Argument-Based Approach to Validity. *Phycological Bulletin*, **112**, 537-535. https://doi.org/10.1037/0033-2909.112.3.527

[14] Chapelle, C.A., Enright, M.K. and Jamieson, J.M. (2008) Building a Validity Argument for the Test of English as a Foreign Language. Routledge, New York.

[15] Bachman, L. (2005) Building and Supporting a Case for Test Use. *Language Assessment Quarterly*, **2**, 1-34. https://doi.org/10.1207/s15434311laq0201_1

[16] Bachman, L. and Palmer, A. (2010) Language Assessment in Practice. Oxford University Press, Oxford.

[17] Weir, C. (2005) Language Testing and Validation. Prentice Hall, London. https://doi.org/10.1057/9780230514577

[18] Cheung, K.Y.F. and Emery, J. (2017) Applying the Socio-Cognitive Framework to the Bio-Medical Admissions Test (BMAT). Cambridge University Press, Cambridge.

[19] Papp, S. and Rixon, S. (2018) Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners. Cambridge University Press, Cambridge.

[20] Shaw, S.D. and Weir, C.J. (2007) Examining Writing: Research and Practice in Assessing Second Language Writing. Cambridge University Press, Cambridge.

[21] 李清华. 语言测试之效度理论发展五十年[J]. 现代外语, 2006, 29(1): 214-217.

[22] Jin, Y. and Yang, H. (2006). The English Proficiency of College and University Students in China: As Reflected in the CET. *Language, Culture & Curriculum*, **19**, 21-36. https://doi.org/10.1080/07908310608668752

[23] 杨惠中, Weir, C. 大学英语四、六级考试效度研究[M]. 上海: 上海外语教育出版社, 1998.

[24] He, L.Z. and Dai, Y. (2006) A Corpus-Based Investigation into the Validity of the CET-SET Group Discussion. *Language Testing*, **23**, 370-401. https://doi.org/10.1191/0265532206lt333oa

[25] 贾国栋. 大学英语口语测试的预期反拨效应——以全国大学英语四、六级口语测试为例[J]. 外语测试与教学, 2016(4): 1-9.

[26] 金艳, 吴江. 以"内省法"检验 CET 阅读理解测试的效度[J]. 外语界, 1998(2): 47-52.

[27] 金艳. 计算机化语言测试的效度研究——浅析计算机能力与测试构念的关系[J]. 外语电化教学, 2012(1): 11-15.

[28] 王跃武, 朱正才, 杨惠中. 作文网上评分信度的多面 Rasch 测量分析[J]. 外语界, 2006, 27(1): 69-76.

[29] 朱正才. 大学英语四、六级考试分数等值研究——一个基于铆题和两参数 IRT 模型的解决方案[J]. 心理学报, 2005, 37(2): 280-284.

[30] 辜向东. 正面的还是负面的——大学英语四六级考试反拨效应实证研究[M]. 重庆: 重庆大学出版社, 2007.

[31] 辜向东. 大学英语四六级考试反拨效应历时研究(上、下卷)[M]. 成都: 四川大学出版社, 2013.

[32] 辜向东, 张正川, 刘晓华. 改革后的 CET 对学生课外英语学习过程的反拨效应实证研究——基于学生的学习日志[J]. 解放军外国语学院学报, 2014, 37(5): 44-164.

[33] Davies, A. (2008) Assessing Academic English: Testing English Proficiency 1950-1989—The IELTS Solution. Cambridge University Press, Cambridge.

[34] Taylor, L. and Weir, C. (2012) IELTS Collected Papers 2: Research in Reading and Listening Assessment. Cambridge University Press, Cambridge.

[35] Annie, B. (2003) An Examination of the Rating Process in the Revised IELTS Speaking Test. *IELTS Research Report*, Vol. 6, 11-30.

[36] Yates, L., Zielinski, B. and Pryor, E. (2011) The Assessment of Pronunciation and the New IELTS Pronunciation Scale. *IELTS Research Report*, Vol. 12, 1-44.

[37] Merrifield, G. (2012) An Impact Study into the Use of IELTS by Professional Associations in the United Kingdom, Canada, Australia and New Zealand. *IELTS Research Report*, Vol. 13, 1-53.

[38] Read, J. and Hayes, B. (2003) The Impact of IELTS on Preparation for Academic Study in New Zealand. *IELTS Research Report*, Vol. 4, 153-191.

[39] Biber, D. and Gray, B. (2013) Discourse Characteristics of Writing and Speaking Task Types on the TOEFL iBT Test: A Lexico-Grammatical Analysis. TOEFL iBT Research Report Series, 2013, i-128.

[40] Stricker, L.J. and Attali, Y. (2010) Test Takers' Attitudes about the TOEFL iBT. TOEFL iBT Research Report Series, 2010, i-16.

[41] Powers, D.E., Roever, C., Huff, K.L. and Trapani, C.S. (2003) Validating Language? Courseware Scores against Faculty Ratings and Student Self-Assessments. TOEFL iBT Research Report Series, 2003, i-25.

[42] Sawaki, Y., Lawrence, J., Stricker, H.O. and Andreas, H.O. (2009) Factor Structure of the TOEFL Internet-Based Test. *Language Testing*, **26**, 5-30. https://doi.org/10.1177/0265532208097335

[43] Wolfe, E.W. and Manalo, J.R. (2005) An Investigation of the Impact of Composition Medium on the Quality of TOEFL Writing Scores. TOEFL iBT Research Report Series, 2005, i-58.

[44] Rahimi, F., Bagheri, M.S., Sadighi, F. and Yarmoh, A. (2014) Using an Argument-Based Approach to Ensure Fairness of High-Stakes Tests' Score-Based Consequence. *Procedia—Social and Behavioral Sciences*, **98**, 1461-1468. https://doi.org/10.1016/j.sbspro.2014.03.566

[45] Weigle, S.C. (2011) Validation of Automated Scores of TOEFL iBT Tasks against Non-Test Indicators of Writing Ability. TOEFL iBT Research Report Series, 2011, i-63.

[46] Xi, X., Higgins, D., Zechner, K. and Williamson, D. (2012) A Comparison of Two Scoring Methods for an Automated Speech Scoring System. *Language Testing*, **29**, 371-394. https://doi.org/10.1177/0265532211425673

[47] ETS (2011) Reliability and Comparability of TOEFL iBT Scores. TOEFL iBT Research Report Series, 2011, i-12.

[48] Zhang, Y. (2008) Repeater Analyses for the TOEFL iBT Test. ETS Research Memorandum, i-12.

[49] Jamieson, J. and Poonpon, K. (2013) Developing Analytic Rating Guides for TOEFL iBT Integrated Speaking Tasks. TOEFL iBT Research Report Series, 2011, i-93.

[50] Richard, J., Tannenbaum, R.J. and Wylie, E.C. (2008) Linking English-Language Test Scores onto the Common European Framework of Reference: An Application of Standard-Setting Methodology. TOEFL iBT Research Report Series, 2011, i-93.

[51] Taylor, L. (2004) Issues of Test Comparability. Research Notes 15, 2-12.

[52] 仇茵晴, 张艳莉. 新老大学英语四级和雅思听力试题的对比研究——改革后新四级成效初探[J]. 外语测试与教学, 2011(3): 29-38.

[53] 金艳, 张晓艺. 技能综合对语言测试构念效度的影响——培生英语考试与大学英语六级网考的对比研究[J]. 外语电化教学, 2013(6): 3-10.

[54] 李鑫, 修旭东. 雅思和我国高考英语阅读测试题型的对比[J]. 解放军外国语学院学报, 2009, 32(5): 60-71.

[55] 王丽. 三种大规模标准化英语考试听力测试部分之比较——一项基于语篇、任务、说话人相关因素的研究[J]. 外语电化教学, 2007(2): 67-72.

[56] Council of Europe (2001) Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press, Cambridge.

[57] 教育部考试中心. 中国英语能力等级量表[Z]. 北京: 高等教育出版社, 2019.