英语口语考试与中国英语能力等级量表对接研究*

——以 CET－SET 4 为例

揭　薇

提要： 本研究尝试对接大学英语四级口语考试和中国英语能力等级量表的“口语量表”，采用标准设定方法建立口语量表和考试的关联，划定针对口语量表级别的考试分界分数。研究发现，专家组能够通过考试任务分析选择合适的考生表现描述语，经过系统的培训后专家组在标准设定的各个阶段体现出较好的一致性和准确性。研究结果对于未来口语量表在英语考试中的应用研究具有参考意义。

关键词： 中国英语能力等级量表；CET-SET 4；标准设定；对接研究

Abstract: This study attempts to relate CET-SET 4 to the speaking scales of China's Standards of English (CSE). Using standard setting methods to establish the relationship between CET-SET 4 and the speaking scales, the study calculates the cut-scores for the corresponding CSE levels. It finds that the panelists are able to select appropriate descriptors through the test task analysis. After systematic training, the panelists show good consistency and accuracy at each stage of standard setting. The study results are of reference significance for the future research on application of the speaking scales to English tests.

Key words: China's Standards of English; CET-SET 4; standard setting; relating study

中图分类号： H319　**文献标识码：** A　**文章编号：** 1004-5112(2019)01-0071-10

1. 引言

中国英语能力等级量表于 2018 年 4 月由教育部正式发布。这一量表立足于中国国情和现状，结合时代发展的新需求，注重科学性、实用性和可操作性(刘建达，彭川 2017)。从实用性来看，量表要对英语学习、教学和测试等具有切实的参照意义和促进作用；从可操作性来看，量表必须可在当前条件下实施，便于英语教学和测评机构、学习者和其他用户理解、接受和使用(刘建达 2017)。

中国英语能力等级量表口语量表(以下简称口语量表)能为我国各阶段的英语口语教学和测试提供参照标准。目前，我国口语能力量表的相关研究有的讨论口语量表的建设原则和方法(金艳，揭薇 2017)，有的探讨口语能力描述语(杨惠中等 2011；王佶旻 2013；揭薇，金艳 2017)，但尚未出现考试和口语量表的对接研究。本研究旨在探讨口语量表与大学英语四级口语考试(CET-SET 4)的关系，促使两者有效衔接，从而为中国英语能力等级量表在我国英语教学、学习和测评之间架起连通的桥梁提供实证依据。

* 本研究得到 2015 年教育部哲学社会科学研究重大课题攻关项目“中国英语能力等级量表建设研究”(编号 15JZD049)、上海对外经贸大学国际商务外语学院 2018 年高原(培育)学科项目的资助。衷心感谢上海交通大学金艳教授的悉心指导。

2. 文献回顾

2.1 考试与量表的对接研究

考试对接量表是一个复杂的过程,需要研究者建立论证为对接的结果提供充分证据,这一过程也可称为论证对接主张的过程(Council of Europe 2009) 。有效的对接论证框架是联系考试和量表的基础(Jones & Saville 2009) 。对接研究为建立考试和量表的关系提供途径,对接过程中的证据收集能使基于量表的考试分数解释更为合理,同时也是考试和量表效度验证的重要组成部分(Tannenbaum & Baron 2010) 。

为了便于研究者开展量表对接研究,提高对接研究的可行性和可操作性,欧洲理事会于 2009 年推出了《语言考试和欧洲语言共同参考框架对接手册》(以下简称《手册》) (Council of Europe 2009) 。Martyniuk(2010) 汇编了相关领域多位专家遵循《手册》的四步骤,对接多个国家的主要考试和《欧洲语言共同参考框架》(以下简称《欧框》) (Council of Europe 2001) 各能力级别的案例研究,为之后的考试与量表的对接研究提供了重要参考。考试和《欧框》的对接研究已经成为测试研究领域的热点(Bechger *et al.* 2009; Figueras & Noijons 2009; Freunberger *et al.* 2013) 。

考试对接量表的方法主要包含调整一致(alignment) 和标准设定。调整一致是重要的教育测量学概念,指内容标准、表现标准、课程和考试评价必须相互调整、协同一致、形成整体(Davis-Becker & Buckendahl 2013; 雷新勇 2011) 。收集考试和量表调整一致的证据是标准设定的必要前提(Council of Europe 2009) 。标准设定指将考生分为若干层次或类别的决策过程。只有对考试和量表进行匹配,在匹配标准下开展统一的分数解释和能力鉴定,并收集内容效度证据之后,才可使用标准设定方法划定分界分数。

国外围绕《欧框》的对接研究反映出一些问题。首先,《欧框》的语境效度和评分效度有待验证,《欧框》在考试开发中的有用性及其是否能和考试有效对接仍然遭到质疑(Weir 2005) 。尽管荷兰分析框架(Dutch Grid) 完善了《欧框》内容,提供了考试内容分析框架,有助于接受性技能考试的对接研究(Alderson *et al.* 2006) ,然而这一框架模型对于产出性技能考试的对接研究操作并不适用。其次,和《欧框》描述语无关的因素,如专家的评判经历、考试评分标准、边缘(borderline) 考生的概念理解等,会影响专家的试题难度判断(Papageorgiou 2010) 。专家组依据《欧框》的能力级别对考生口语表现的评判出现较大分歧(Maris *et al.* 2009) 。对于某个能力级别对应试题的分界特征、级别之间的级差等,《欧框》只给出很少的指导意见(Moe 2009) 。

我国有关考试与量表的对接研究主要是国内考试与《欧框》的匹配或对接研究。鹿士义(2011) 运用标准设定方法,参照《手册》中的对接程序,将商务汉语考试(Business Chinese Test) 的 4 项能力与《欧框》对接,探讨考试在《欧框》中对应的能力等级; 黄婷和贾国栋(2012) 从内容评定角度对接大学英语四、六级考试和《欧框》,以实现大学英语四、六级考试与《欧框》在内容上的匹配; 罗莲(2017) 并未依循《手册》中的对接程序,而是结合教师评价、测试分数和学生基于《欧框》的自评数据,使用以学生为中心的方法,将汉语分级测试与《欧框》对接。总体来说,我国有关考试与量表的对接研究相对较少,而量表对接需将量表的能力描述和分级转化运用于具体的教学和评估,因此这一工作需要研究者持续探讨和实践。

2.2 标准设定方法

标准设定是教育测量的重要方法。相邻两个级别或类别考生之间的分界一般称为分界值或分界分数(cut-score)(Cizek & Bunch 2007; Zieky *et al.* 2008)。标准设定涉及诸多工作步骤,包括选定标准设定方法、遴选专家组和设计研究、准备能力水平描述、专家培训、收集试题评分、提供专家判断反馈、组织专家组讨论、专家评分统计、收集效度验证证据等(Hambleton & Pitoniak 2006; AERA *et al.* 2014)。

由于目的和用途不同,标准设定的方法各有差异,目前已知的标准设定方法超过一百多种。常见的标准设定方法分为两类:(1)以测试为中心(test-centered)的方法,包括 Angoff 法、Ebel 法、Nedelsky 法等;(2)以考生为中心(examinee-centered)的方法,包括对照组法(contrasting group)、边缘组法(borderline group)等(Cizek & Bunch 2007; Bechger *et al.* 2009; Council of Europe 2009)。量表对接研究中,产出性技能考试和接受性技能考试对接量表的标准设定方法有所不同,侧重点也不一样。

标准设定方法中,专家的主观判断发挥重要作用,但为避免标准设定结果的争议性,应以现代测量理论为依据,严格验证专家判断的效度和信度(AERA *et al.* 2014)。标准设定的信度指标有很多,主要包括决策一致性系数(decision consistency),如 P 系数和 kappa 系数、基于概化理论的等级线决策信度、基于项目反应理论的分界点信息量等。常见的估计方法还包括 Huynh 法、Subkoviak 法、LL 法等(Livingston & Lewis 1995; 温红博等 2017)。此外,多层面 Rasch模型分析是广泛运用的标准设定质量评价和改进工具(Hsieh 2013),能够检验标准设定过程中的可变性,识别小组成员的异常决策,为标准设定的组织者和专家组成员提供有效反馈。Kecker & Eckes(2010)在考察《欧框》能力级别和德福考试关系的研究中,使用 Rasch 模型分析软件 FACETS 去除了专家在试题评分中的判断不一致性。

考试与量表对接工作的主要目的是通过标准设定获得稳定可靠的分界分数,逻辑回归(logistic regression)(Council of Europe 2009)和中点分析法(midpoint analysis)(Cizek & Bunch 2007; Freunberger *et al.* 2013)是计算分界分数的主要方法。分界分数应根据考试目的、教育环境等实际情况进行调整应用。标准设定专家组的主要任务是提供分界分数建议而不是做出最终的分类决策,决策机构需要综合考虑各种因素,基于实际情况调整确定分界分数(雷新勇 2011)。

本研究的主要目的是为建立口语考试与口语量表的关系提出具体、详实的方案,并通过实证研究验证方案的科学性和有效性,以使口语量表得到更为合理、恰当的使用。研究将运用决策一致性系数、多层面 Rasch 模型分析等方法评估标准设定质量,运用逻辑回归和中点分析法计算分界分数。考试与口语量表对接研究是对考试透明度和专业化承诺的实践检验,能够帮助量表使用者分析量表级别对应的考试任务,从而更好地了解量表及其级别和考试之间的关系。对接结果也可帮助试题编写者编写与口语量表特定级别相对应的试题,开发基于中国英语能力等级量表体系的口语考试任务。

3. 研究方法

3.1 研究问题

本研究主要回答以下问题:

(1) 专家组在熟悉口语量表后是否能够判断量表描述语的级别?判断的准确性和一致

性如何?

(2) 专家组是否能够运用口语量表判断 CET-SET 4 考生的量表能力级别? 判断的准确性和一致性如何?

(3) CET-SET 4 对应于口语量表级别的分界值是多少?

3.2 研究步骤和方法

本研究主要参考考试与量表对接研究领域的《手册》、Martyniuk(2010) 、Bärenfänger & Tschirner(2012) 等的步骤和方法,具体分为口语量表熟悉阶段、考试内容评定阶段和标准设定阶段。

口语量表熟悉阶段。专家组熟悉口语量表的级别、口头表达能力总表和各个分量表、描述语三要素构成等,确保在整个标准设定过程中充分、详细地了解口语量表。对接有效性建立在这一步骤有效性的基础之上。

专家组需要熟悉的口语量表主要内容是: (1) 口语量表的级别,包括各个级别的目标学习者; (2) 口语量表的各个分量表,包括口语量表的理论框架和分类框架; (3) 口语量表的描述语,包括口语能力描述语举例和口语描述语表现、标准和条件三要素讲解; (4) 口语量表的能力总表、策略总表和各个级别描述语的典型特征。

这一阶段的主要活动是: (1) 口语量表级别之间的差异(level differentiation) 和内容连续性(content coherence) 问答活动,比如: 口语量表中高一级别和低一级别的能力差异体现在哪里? (2) 专家组对口语能力描述语进行分级,将描述语与相应级别相匹配。

考试内容评定阶段。这一阶段向专家组介绍 CET-SET 4,包括考试任务分析、分界分数、分数解读,旨在分析 CET-SET 4 的考试内容和能力目标在多大程度上覆盖口语量表的范围和水平。

这一阶段的主要活动是: (1) 选出和口语描述语相关的 CET-SET 4 考试任务; (2) 将 CET-SET 4 考试任务和相关口语描述语的级别、分类匹配。

标准设定阶段。这一阶段是专家组将口语量表能力级别对应于考试的标准化过程,要确保专家在表现评级(performance rating) 中的判断反映口语量表描述语的能力构念和级别,并且确保判断结果建立在充分的证据之上。

这一阶段的主要活动是将口语量表的能力级别和 CET-SET 4 考生表现进行匹配。

3.3 数据收集

专家组。专家甄选是标准设定过程中非常重要的环节,专家的数量和质量直接影响标准设定的结果和效果。《手册》建议对接活动的专家组至少由 10 名专家组成。本研究综合考虑专家的学科背景、教学经验和研究领域,确定 12 名专家组成专家组,其中包括 8 位 10 年以上教学经历的大学英语教师、2 位经验丰富的 CET-SET 4 评分员和 2 位口语量表开发者。

CET-SET 4 考生。本研究选取 2016 年的一次 CET-SET 4 考试,按照性别、地区、分数段采用分层抽样从总样本53 287名考生中随机抽取 30 名考生的考试信息。考生包括男生和女生各 15 名,覆盖北京、江苏、广东、辽宁、四川、上海、天津等省市,在 A、B+、B、C+、C、D 各分数段以每个等级 5 名均匀分布。

CET-SET 4 考试任务。研究者和专家组通过对应口语量表和 CET-SET 4 试题,挑选出 5 个具有代表性的考试任务话题,具体如表 1 所示。

表 1　抽取的 CET-SET 4 考试话题

话题	话题领域
Environment	Social Issues
Visiting Park	Leisure Activities
Travel	Everyday Life
Super Cities	Modern World
Smartphones	Technology

4. 结果分析

4.1 口语量表熟悉阶段

通过培训活动和讨论，专家组总结出口语量表三级至八级的口语典型特征（见表 2），这一过程可以帮助专家组熟悉口语量表各个级别的学习者能力水平描述。

表 2　口语量表三级至八级口语典型特征

三级	行为表现：相对事实的（relatively factual）、围绕日常生活和学习（如问候他人） 话题：熟悉、普通、日常话题 语言特征：简单、简短、基本正确（如时态、词汇）
四级	行为表现：叙述、描述、回应他人 话题：个人、熟悉话题 语言特征：简单、简短、适当、连贯、详细
五级	行为表现：描述、解释、比较（比四级更具挑战性） 话题：日常、社会热点问题 语言特征：简短、清楚、有条理（总体上比四级要求更高）
六级	行为表现：与五级相似但更多样化 话题：个人生活、社会热点事件（范围更广） 语言特征：清楚、有效、恰当（简单活动生动详细，复杂、正式活动简要）
七级	行为表现：各种形式的活动、正式学术报告 话题：各种熟悉的个人、抽象话题 语言特征：调整讲话内容和表达方式，保持发言权
八级	行为表现：类型广泛的各种活动 话题：各种专业、学术话题 语言特征：在正式和非正式场合即兴、简明、灵活、有条理地表达，准确性提高

专家准确性检验。培训之后，专家组成员依据自己的判断给一组没有标明级别的描述语定级。专家判断准确率是级别判断准确的描述语数量与所判断描述语总数之比。计算结果显示，专家总体的判断准确率为 0.67。研究者将计算结果反馈给专家组，因为提供有效的反馈是必要的，这能促进他们审核和修改判断（Reckase 2001）。

专家一致性检验。P 系数方法通过计算两次平行测量项目判断中相同判断的比例来验证

一致性,分类一致的项目比例越高则信度越高; kappa 系数是在排除机遇一致性后,测验真实分类一致性与最大可能分类一致性的比值,对 P 系数进行校正(陈平等 2011) 。一致性检验结果显示,口语量表四级至六级 3 个级别的专家判断一致性 kappa 系数范围是 0.5840—0.7465。考虑到本研究所选的口语量表描述语具有同质性,上述范围内的 kappa 系数显示了较好的一致性。因此,专家组的判断可被视为具有较高的一致性,专家对于描述语的级别判断较为统一。

4.2 考试内容评定阶段

专家组依据抽取出来的 5 个 CET-SET 4 考试任务组成的试题册,从口语量表中挑选出符合考试任务或和考试任务相关的描述语。然后,研究者汇集 5 个考试任务的相关描述语,以便后续的表现评级。

表 3 呈现了一个考试任务的描述语挑选示例。这次口语考试的话题是“Environment”,话题领域是“Social Issues”,要求考生在简短的自我介绍后朗读一篇 119 词的短文,回答两个问题:(1) What can people do to help improve the environment? (2) What should college students do to save natural resources? 考生回答完问题后,稍作准备就话题“The measures China has taken to protect the environment”做 1 分钟个人陈述,陈述结束之后经计算机随机配对和另一名考生围绕相似话题展开 3 分钟讨论。专家组熟悉口语考试任务后,从口语量表中挑选出和考试任务所需的能力水平相当的描述语。

表 3 对照口语量表的 CET-SET 4 考试任务分析

场景/话题	Environment	
话题领域	Social Issues	
活动/行为	口头描述、口头说明、口头论述、口头互动、执行、评估和补救	
相关描述语		
四级	口头描述	能简单描述自己的学校或工作场所,如地点、人员特征等
四级	口头互动	能与他人简单讨论家庭、学校等方面的话题
五级	口头说明	能清楚、连贯地解释自己的计划或方案,如做什么、怎么做等
五级	口头互动	能与他人就社会热点问题展开口头讨论
五级	评估和补救	能通过提问确认对方是否理解自己的谈话内容
五级	执行	能通过提问确保他人理解自己的意思
六级	口头说明	能在讨论时解释问题产生的原因并陈述解决问题的意见
六级	口头论述	能就社会热点话题发表意见、表明立场并给出充分理由
六级	执行	能在恰当的时机自然地接话、插话和结束交谈

4.3 标准设定阶段

专家判断分析。标准设定阶段需要评判专家组成员判断的差异。分析结果表明,专家组的评分者信度系数 Cronbach’ Alpha 为 0.93,Pearson 相关系数均值为 0.75。组内相关系数

(ICC)计算显示,专家判断的平均一致性系数为0.98。

专家组的评分一致性也可通过多层面Rasch模型分析。图1显示了专家评分所有层面的总体分布情况,大部分专家的评判相差不大(除了J5和J3)。这表明,标准设定阶段12位专家的评分一致性较好,尽管专家评分严厉度的洛基值(logits)在-0.47—0.88区间之内,各位专家之间的严厉度仍有差异。12位专家评分的加权均方拟合统计量(InfitMnsq)在0.63—1.46之间,均在可接受的拟合值范围之内。专家的统计量分隔系数(separation)为2.58,分隔信度(reliability)为0.87,卡方检验显著。这表明12位专家的严厉度整体上差异不大,均能较好地把握严厉度尺度。

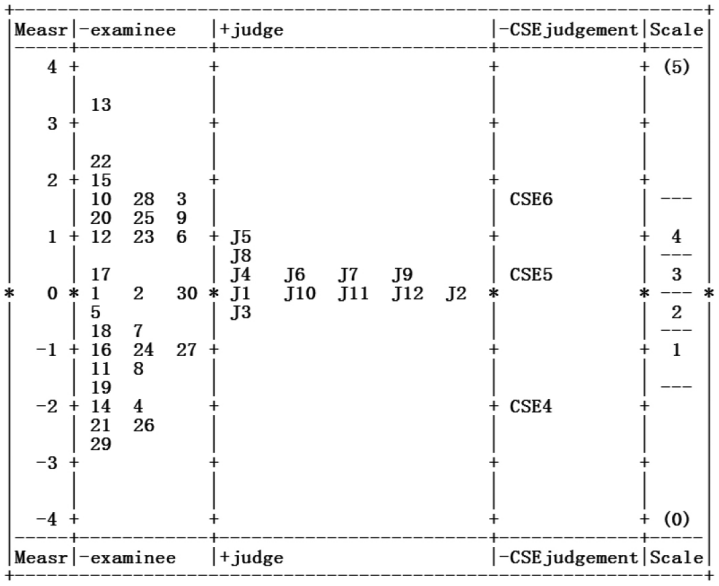

图1 专家评分总层面图

逻辑回归。逻辑回归计算与量表级别对应的考试分界分数(Cizek & Bunch 2007: 109; Council of Europe 2009: 72)。本研究中,专家评分如果达到某种标准为1,否则为0,因而使用线性回归方式检验这种二元选择因变量可以得到相较而言更加直观的结果。这种广义的逻辑回归方法能够更为清晰地反映专家评分与判断标准概率之间的关系,逻辑回归的概率线性表达式可以写为:

$$\ln \frac{p}{1-p} = a + \sum_i b_i scores$$

这一表达式中ln为自然对数,scores为各专家评分,a和b分别表示对应的待估回归系数,p为达到标准的概率。尽管概率不能直接观测得到,但是我们可以通过专家评分对判断标准进行渐进估计。根据专家评分数据,我们分别估计了专家评分系数和概率。基于此,概率系数的分界分数为达到标准的概率与不能达到标准概率之间的比值。

逻辑回归除了可以直观描述逻辑概率(纵轴)与得分之间的关系,也得到一个参照"截断"得分标准(纵轴虚线),两种计算方法的结果如图2a和图2b所示。由图可知,低级别的分界分数估计差距较大,分别是13.558和12.103;高级别的分界分数估计相近,分别是17.350和17.324。造成这种结果的原因可能是专家在高级别考生表现上的判断具有更高的一致性,或

者说是高级别考生表现具有专家可以把握的典型特征。Papageorgiou(2010) 发现,在判断考生表现对应量表级别的过程中,专家主要通过量表的措辞(如描述语的词句) 来解读表现。这一过程就像评分员使用评分量表打分,描述语的作用是为专家提供级别判断的依据。因此,如要获得更为准确的低级别分界分数,研究者需为专家组提供更为明晰的口语量表四级和五级划分标准。

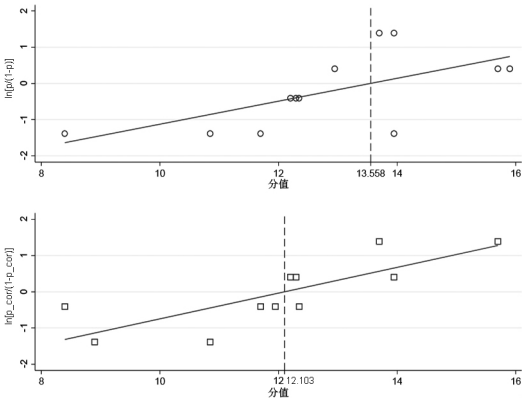

图 2a　两种计算方法的分界分数(低级别)

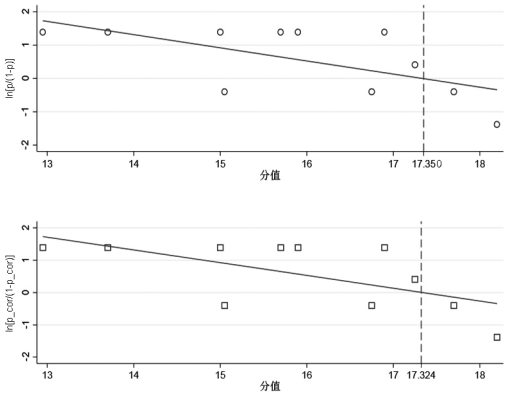

图 2b　两种计算方法的分界分数(高级别)

中点分析法。中点分析法用于计算口语量表四级和五级、五级和六级的分界分数。中点分析法借助两个级别学生的考试分数均值来计算级别之间的中点(Cizek & Bunch 2007) 。由各个级别学生考试分数的均值可知,四级和五级的中点为 "(11.13593 + 14.31639) /2 = 12.72616",五级和六级的中点为 "(17.28152+14.31639) /2 = 15.79896"。

必须注意的是,不同计算方法会产生不同的分界分数,因此我们需要谨慎对待标准设定结果(Freunberger *et al.* 2013) 。为使分界分数更加可信,《手册》建议研究者在对接研究后收集更多的证据来验证分界分数。验证证据主要来自 3 方面: (1) 交叉验证,重新选择一组专家,重复标准设定过程; (2) 标准设定方法验证,选择另一种适用的标准设定方法,重新计算分界分数; (3) 外部验证,将标准设定的结果用于一个外部标准,外部标准可以是已经和量表可靠对接的考试。

5. 结语

本研究通过英语口语考试与口语量表的对接发现,专家们能够通过考试任务分析选择合适的考生表现描述语,描述语判断具有较好的一致性和准确性。根据标准设定不同阶段的培训反馈,研究者总结出对接研究的相关经验: 首先,研究者需要解释和改写口语描述语,增强描述语和考生表现的相关性。其次,部分口语描述语缺乏明晰的语言特征描述,研究者可以探明专家分级或打分的显著特征。最后,口语描述语缺乏相关情景的详细描述时,应尽量避免专家的考生表现评判过于依赖想象。

不同于考试研究中的标准设定目的,考试对接量表的标准设定是为了确定与考试机构期望的量表级别对应的分界分数,并允许根据考试结果将考生能力划分为不同的量表级别(Hambleton & Pitoniak 2006; Fulcher 2010: 248) 。高风险考试的量表对接研究中,基于标准设定所得的不恰当的分界分数可能会产生严重后果,分界分数过高或过低都会造成考试误用和

不良社会后果。本研究中,尽管研究者仔细遴选专家组成员,严格按照抽样框抽取考生,并在标准设定过程中尽可能避免测量误差,但因受限于考生和专家样本,研究结论仍然不能作为对接研究有力的主张,不过考试对接量表的过程可为系统的考试和量表改进机制建设提供参考。此外,考试与量表的对接关系会随着教育资源、教育政策而变化,对接研究应是持续发展的动态过程。

参 考 文 献

[1] AERA, APA & NCME. *Standards for Educational and Psychological Testing* [M]. Washington, D.C: American Educational Research Association, 2014.

[2] Alderson J C *et al.* Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project [J]. *Language Assessment Quarterly*, 2006, 3 (1): 3-30.

[3] Bärenfänger O & Tschirner E. *Assessing Evidence of Validity of Assigning CEFR Ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by Computer (OPIc)* [R]. Leipzig: Institute for Test Research and Test Development, 2012.

[4] Bechger T M, Kuijper H & Maris G. Standard setting in relation to the Common European Framework of Reference for Languages: The case of the state examination of Dutch as a second language [J]. *Language Assessment Quarterly*, 2009, 6(2): 126-150.

[5] Cizek G J & Bunch M B. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests* [M]. London: Sage Publications, 2007.

[6] Council of Europe. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment* [M]. Cambridge: Cambridge University Press, 2001.

[7] Council of Europe. *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. A Manual* [M]. Strasbourg: Language Policy Division, Council of Europe, 2009.

[8] Davis-Becker S L & Buckendahl C W. A proposed framework for evaluating alignment studies [J]. *Educational Measurement: Issues and Practice*, 2013, 32(1): 23-33.

[9] Figueras N & Noijons J (eds). *Linking to the CEFR Levels: Research Perspectives* [C]. Arnhem: CITO, Council of Europe & EALTA, 2009.

[10] Freunberger R, Bazinger C & Itzlinger-Bruneforth U. *Linking the Austrian Standards-Based Test for English to the CEFR: The Standard-Setting Process* [R]. Salzburg: BIFIE, 2013.

[11] Fulcher G. *Practical Language Testing* [M]. New York: Routledge, 2010.

[12] Hambleton R K & Pitoniak M J. Setting performance standards [A]. In Brennan R L (ed). *Educational Measurement* (4th Ed.) [C]. Westport, CT: American Council on Education & Praeger, 2006. 433-470.

[13] Hsieh M. An application of Multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure [J]. *Language Testing*, 2013, 30(4): 491-512.

[14] Jones N & Saville N. European language policy: Assessment, learning and the CEFR [J]. *Annual Review of Applied Linguistics*, 2009, (29): 51-63.

[15] Kecker G & Eckes T. Putting the Manual to the test: The TestDaF-CEFR linking project [A]. In Martyniuk W (ed). *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual* [C]. Cambridge: Cambridge University Press, 2010. 50-79.

[16] Livingston S A & Lewis C. Estimating the consistency and accuracy of classifications based on test scores [J].

Journal of Educational Measurement, 1995, 32(2) : 179-197.

[17] Maris G, Noijons J & Reichard E. Benchmarking of videotaped oral performances in terms of the CEFR [A]. In Figueras N & Noijons J (eds) . *Linking to the CEFR Levels: Research Perspectives* [C]. Arnhem: CITO, Council of Europe & EALTA, 2009. 81-85.

[18] Martyniuk W (ed) . *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual* [C]. Cambridge: Cambridge University Press, 2010.

[19] Moe E. Jack of more trades? Could standard setting serve several functions? [A]. In Figueras N & Noijons J (eds) . *Linking to the CEFR Levels: Research Perspectives* [C]. Arnhem: CITO, Council of Europe & EALTA, 2009. 131-138.

[20] Papageorgiou S. Investigating the decision making process of standard setting participants [J]. *Language Testing*, 2010, 27(2) : 261-282.

[21] Reckase M D. Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact [A]. In Cizek G J (ed) . *Setting Performance Standards: Concepts, Methods, and Perspectives* [C]. Mahwah, NJ: Lawrence Erlbaum, 2001. 159-174.

[22] Tannenbaum R J & Baron P A. *Mapping TOEIC Test Scores to the STANAG 6001 Language Proficiency Levels* [R]. Princeton, NJ: Educational Testing Service, 2010.

[23] Weir C J. Limitations of the Common European Framework for developing comparable examinations and tests [J]. *Language Testing*, 2005, 22(3) : 281-300.

[24] Zieky M J, Perie M & Livingston S A. *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests* [M]. Princeton, NJ: Educational Testing Service, 2008.

[25] 陈平等. 标准参照测验决策一致性指标研究的总结与展望 [J]. 心理发展与教育,2011,(2) : 210-215.

[26] 黄婷,贾国栋. 语言测试与《欧洲语言共同参考框架》匹配的可行性研究——以大学英语四、六级考试为例 [J]. 外语测试与教学,2012,(1) : 38-49.

[27] 揭薇,金艳. 口语能力描述语的语体分析: 基于中国英语能力等级量表的研究 [J]. 外语界,2017,(2) : 20-28.

[28] 金艳,揭薇. 中国英语能力等级量表的"口语量表"制定原则和方法 [J]. 外语界,2017,(2) : 10-19.

[29] 雷新勇. 基于标准的教育考试——命题、标准设置和学业评价 [M]. 上海: 上海科学技术出版社,2011.

[30] 刘建达. 中国英语能力等级量表与英语学习 [J]. 中国外语,2017,(6) : 4-11.

[31] 刘建达,彭川. 构建科学的中国英语能力等级量表 [J]. 外语界,2017,(2) : 2-9.

[32] 鹿士义. 商务汉语考试(BCT) 与欧洲语言共同参考框架(CEFR) 的等级标准关系研究 [J]. 华文教学与研究,2011,(2) : 56-63.

[33] 罗莲. 汉语分级测试与 CEFR 等级的连接研究 [J]. 语言文字应用,2017,(2) : 110-118.

[34] 王佶旻. 汉语能力标准的描述语任务难度研究——以中级口语能力量表为例 [J]. 世界汉语教学,2013,(3) : 413-423.

[35] 温红博,卜文娟,刘先伟. 初中学业水平考试中固定比例法标准设定的信度分析 [J]. 考试研究,2017,(5) : 55-63.

[36] 杨惠中,朱正才,方绪军. 英语口语能力描述语因子分析及能力等级划分——制定语言能力等级量表实证研究 [J]. 现代外语,2011,(2) : 151-161.

作者单位: 上海交通大学,上海 200240; 上海对外经贸大学,上海 201620