基于文本特征的校本写作考试与《中国英语能力等级量表》对接效度研究

何莲珍[1]，阮吉飞[2]，闵尚超[1]

（1. 浙江大学 外国语言文化与国际交流学院 浙江 杭州 310058；2. 浙江省春晖中学 英语组 浙江 绍兴 312300）

摘　要：本研究以某高校校本英语考试为研究对象，采用分析性判断法作为标准设定方法，建立校本考试与《中国英语能力等级量表》之间的对接关系。结果表明，该校的英语水平考试写作卷及格线对应量表 5 级，符合考试定位。外部效度验证结果表明：总体而言，对应的量表级别越高，文章长度越长，词汇更加丰富多样，句法更加复杂，但与语篇连贯性无直接关系。本研究是校本考试对接量表的一次有益探索，为对标研究提供方法参考，对于量表的修订完善具有一定的借鉴意义。

关键词：校本考试；中国英语能力等级量表；标准设定；分析性判断法；文本分析

中图分类号：H319　　**文献标识码**：A　　**文章编号**：1000-5544(2021)03-0052-06

Abstract: The present study attempts to align an in-house English proficiency test to *China's Standards of English Language Ability* (CSE) using the analytical judgment method. The results show that the writing subtest is aligned to CSE Level 5. Meanwhile, with the aim to externally validate the alignment results, differences in textual features between different levels of writing samples are examined, indicating that writing samples at higher CSE levels are of greater length, lexical diversity and syntactic complexity, but writing samples at different levels differed little in terms of textual cohesion. As an exploratory attempt at test-standard alignment, this study not only provides a methodological reference for future alignment studies, but also sheds light on the revision and refinement of CSE.

Key words: in-house test; China's Standards of English Language Ability (CSE); standard setting; analytical judgment method; text analysis

DOI:10.16362/j.cnki.cn61-1023/h.2021.03.009

1. 引言

校本考试是由学校自主命题，在学校内部使用，为了考查本校学生的某项能力而设计和实施的考试。通过对接校本英语考试与其他语言考试，能为教师和学生提供更加丰富的语言能力信息和决策依据。但国内英语考试种类繁多，不同考试对于语言能力的定义和划分各有不同，逐一建立校本英语考试与其他语言考试的对应关系既不现实也无必要。2018 年 2 月，《中国英语能力等级量表》（以下简称《量表》）正式发布，为我国外语能力测评体系建设提供统一标准（刘建达、彭川 2017）。有了量表作为统一的“度量衡”，就能连接校本考试与量表，有利于校本考试明确能力定位，提高校本考试的透明度，为教师与学生提供更加全面的语言能力参考（Council of Europe 2009；Tannenbaum & Cho 2014）。本研究以某高校英语水平考试（以下简称“水平考试”）写作卷为例，建立校本考试与量表之间的对应关系，为校本考试分数解释提供新思路，为后续考试与量表的对接研究提供参考。

对接研究主要包括四个阶段：框架熟悉、试题检视、标准设定、效度验证（Council of Europe 2009）。Kaftandjieva（2004）曾指出，在对接过程中，标准设定处于核心地位，因为这一环节决定了考试对接到量表各个级别的临界分数线。因此，标准设定方法对于最终的对接结果至关重要。学界通常将标准设定方法分为两类：试题中心法和考生中心法。前者以试题为评判基础，适用于选择题等客观题型，常用于接受型技能考试；后者对考生表现进行直接评判，适用于写作题等主观题型，常用于产出型技能考试。其中，分析性判断法（analytical judgement method，Plake & Hambleton 2001）颇受研究者青睐，被广泛运用于高风险、大规模产出型语言考试与语言标准对接研究，如雅思考试对接《欧洲语言共同参考框架》（Lim *et al.* 2013）。该方法要求专家基于考生真实的考试表现作出评判，具有操作简便直观、省时等特点，且不依赖于标准卷或参照卷（Abbott 2006），因此本研究选取该方法作为标准设定方法。

标准设定方法直接决定对接结果，而效度验证则是检验对接结果的重要衡量标准，通常包括程序效度、内

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

部效度和外部效度。过往对接研究或仅提供程序效度与内部效度证据(O'Sullivan 2009),或通过常规方法提供外部效度证据(Papageorgiou 2007;Dunlea 2015),如对比同批考生在其他考试中的成绩或不同考生在同一考试中的成绩、对比同批或不同批专家使用不同标准设定方法得出的结果等进行交叉验证(Council of Europe 2009)。迄今为止,尚未有研究从作文文本特征的角度为对接结果提供外部效度证据。与阅读、听力等接受型技能测试不同,写作测试中可以直接观察到二语学习者的语言产出,文本特征是衡量写作能力的重要标准之一(Crossley & McNamara 2012; Vo 2019; 张煜杰、蒋景阳 2020)。因此,探究不同级别作文之间的文本差异是检验对接结果效度的有效途径之一。本研究在对接结果的基础上,将水平考试写作样本分成不同级别,分析不同级别之间的文本差异,进行外部效度验证,同时也为量表的进一步修订提供参考。

综上,本研究拟回答以下两个问题:1)水平考试写作卷对接到量表哪一个级别?2)不同量表级别的作文在文本特征上有何差异?

2. 研究方法

2.1 参与专家

专家小组由14人组成,包括语言测试领域的专家学者和具有丰富大学英语教学经验的一线教师。其中7人曾参与量表研制,对量表非常熟悉。专家组成员对水平考试都很了解。此外,有一人担任对接工作的组织协调者,负责介绍对接流程、组织和引导讨论、收集评分表、掌握对接进度等。

2.2 研究工具和材料

本研究以水平考试写作卷为工具,选用2018年4月的一套水平考试写作卷和写作量表作为对接素材,以组构知识量表作为辅助和参考资料。水平考试写作卷要求考生根据所给话题或图表,在30分钟内撰写一篇不少于160词的议论文。评阅采用整体评分法,从语言和内容两方面进行评分,人工评阅和机器评分相结合。写作卷满分20分,及格分为12分。本次研究共选取考生作文796篇。从中选取5篇用于标准设定培训,50篇用于标准设定。选取方法如下:根据评分标准中的5个分数档(1~4分,5~8分,9~12分,13~16分,17~20分)对考生作文进行分组,再从每个分数档中随机抽取相应比例(与作文原始分数的分布比例相同)的作文。

2.3 研究步骤

对接会前,项目组召开工作会议,对水平考试写作卷考查能力和量表描述语进行仔细分析,确定本研究聚焦水平考试写作卷与量表3~7级的对应关系。同时,项目负责人编制了《英语水平考试与量表对接工作手册》(以下简称《工作手册》),帮助专家充分了解对接流程。工作手册详细列出写作量表、组构知识量表、水平考试写作卷的考试大纲与评分标准、标准设定方法、量表熟悉任务和考试内容分析任务等,便于专家随时参阅。

对接会流程包括框架熟悉、试题检视、标准设定与效度验证。针对框架熟悉与试题检视,本研究设计了三个量表熟悉任务和一个考试内容分析任务,帮助专家充分了解写作量表和水平考试写作卷。就标准设定而言,主要采用两轮分析性判断法[①],两轮判断之间有反馈与讨论。针对效度验证,本研究主要收集程序效度、内部效度和外部效度证据。程序效度证据来自于问卷调查,内部效度证据来自于专家内部一致性与专家间一致性检验结果,外部效度证据来自于不同级别作文的文本特征分析。

2.4 数据收集与分析

针对研究问题一,专家们首先对5篇培训卷进行评判与讨论,就评判标准达成共识,建立基准线。然后对50篇标准设定卷进行第一次独立评判,给出等级。具体而言,研究者将量表3~7级的每个级别细分为三级,共15级(如图1所示),然后随机选取评判结果进行小组讨论,各成员分享评判依据并做相应调整,随后再对50篇标准设定卷进行第二次独立评判。临界分的计算基于第二次独立评判结果。

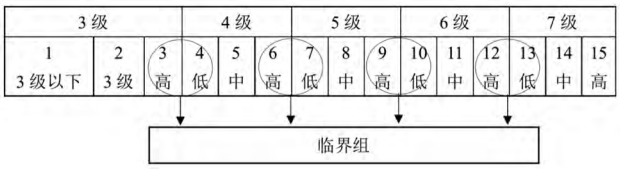

图1. 分析性判断法等级分类

针对研究问题二,本研究根据临界分将796篇作文分成3级及以下、4级、5级、6级、7级及以上五个组。然后,从前三组中各随机抽取50篇作文,从6级中抽取40篇进行文本分析(6级作文总量未达到50;7级作文数量极少)。随后,使用文本词汇分析工具VocabProfile(Cobb 2007)和自然语言处理工具Coh-Metrix 3.0(Graesser *et al.* 2004; McNamara *et al.* 2014)对作文进行文本分析。借助VocabProfile工具获取8个文本指标:文章长度、形符比、词汇密度、以及文本词汇在英国国家语料库(British National Corpus)与美国当代英语语料库(Corpus of Contemporary American English)中的K1-K5词频。Coh-Metrix 3.0可以对11个类别共106个变量进行自动分析,这些类别依次为:描述性统计量、文本易读性主成分得分、指称衔接、潜在语义分析、词汇多样性、连词、情景模式、句法复杂度、句法组构密度、词汇信息和可读性(McNamara *et al.* 2014;何莲珍、孙悠夏 2015;江进林、韩宝成 2018)。本研究将以上指标进行整合归类,从文本基本信息、词汇多样性、句法复杂度和语篇连贯性四个维度,使用单因素方差分析方法对不同级别的考生作文文本特征进行比较。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3. 结果

3.1 考试对接量表的临界分

考试对接量表的本质是将考试成绩关联到量表的不同级别。在计算各级别临界分前,需首先保证对接效度。因此,本部分首先呈现对接的内部效度和程序效度证据,然后计算对接后的各级别临界分。

内部效度指的是标准设定结果的准确性和一致性(Council of Europe 2009)。本研究主要收集了专家内部一致性与专家间一致性效度证据。专家的评判是定序变量,因此使用 Kendall's tau-b 等级相关系数(Council of Europe 2009)。共有 14 位专家参与第一轮评分,13 位专家参与第二轮评分。专家的平均信度达到 0.944,最高为 1,最低为 0.867,表明专家内部一致性较高。

表 1 呈现的是两轮评判结果的标准差和专家间一致性结果,第二轮评判结果标准差(2.99)小于第一轮评判结果标准差(3.06),说明第二轮评判的专家间一致性更高;两轮评判的 Kendall's tau-b 信度系数均值都在 0.55 以上,且第二轮均值大于第一轮,说明专家间整体一致性较好。综上,本次对接实验取得了较高的内部效度。

表 1. 专家间一致性结果

	人数	标准差	Kendall's tau-b			
			最小值	最大值	平均值	标准差
第一轮	14	3.06	.392**	.722**	.567	0.068
第二轮	13	2.99	.403**	.751**	.592	0.074

(注:** $p < 0.01$)

程序效度是指对接过程中所用方法合理且实施质量较高,主要包括以下五个方面:清晰度、实用性、实施系统性、专家反馈和文案记录(Council of Europe 2009)。本研究通过问卷调查收集程序效度证据。基于 Dunlea(2015)的调查问卷,使用李克特四级评分量表(1 到 4 分别代表完全不同意,不同意,同意,完全同意),主要调查对接过程实施、讨论有效性及时间充分性、熟悉任务有效性、反馈作用等方面。问卷结果显示,每道题目的最高分为 4 分(满分),最低分为 3 分,平均分范围为(3.4, 3.9),说明所有专家对于以上各方面都给予了肯定评价。因此,本次对接实验取得了较好的程序效度。

随后,本研究根据两轮分析性判断数据,计算各级别临界分。由于第二轮判断结果比第一轮的一致性更高,最终临界分以第二轮评判为准。3 级和 4 级、4 级和 5 级、5 级和 6 级、6 级和 7 级之间的临界分分别是 11 分、12 分、15 分和 16 分。根据考试大纲要求,水平考试写作卷及格分为 12 分。该结果表明,水平考试写作卷及格分对接量表 5 级,符合考试的整体定位。

3.2 不同量表级别的作文文本差异

本研究通过 VocabProfile 和 Coh-Metrix 3.0 分析获得四组文本(3 级及以下、4 级、5 级和 6 级)的文本特征数据,包括文本基本信息、词汇多样性、句法复杂度和语篇连贯性等四个维度,然后使用 SPSS 20.0 软件进行单因素方差分析。

3.2.1 文本基本信息

单因素方差分析结果显示:不同级别写作样本在文章长度上差异显著,$F(3, 186) = 16.72$, $p = .000$, $\eta^2 = 0.212$。图 2 呈现的是不同级别写作样本在文章长度上的变化趋势,可以看出,级别越高,文章越长。

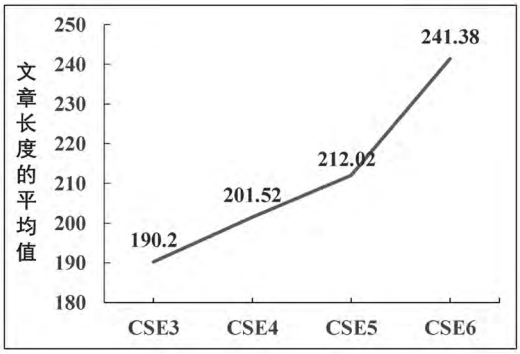

图 2. 文章长度随写作样本级别变化趋势

3.2.2 词汇多样性

单因素方差分析结果显示:不同级别写作样本在 11 个词汇多样性指标上有显著差异,包括形符比(TTR)、K1-K4 词频、单词长度(音节)、词汇多样性(MTLD[②])、词汇多样性(vocd[③])、第一人称复数代词、实词(CELEX[④])、名词上下义。表 2 呈现的是 11 个词汇多样性指标在不同级别的变化趋势。

表 2. 不同级别写作样本的词汇多样性指标

指标(平均值)	CSE3	CSE4	CSE5	CSE6
形符比	0.530	0.556	0.561	0.557
K1	86.77	85.50	85.50	83.43
K2	4.21	6.35	6.98	8.09
K3	1.79	1.91	2.74	3.12
K4	0.40	0.59	0.81	0.90
词汇长度(音节)	1.45	1.49	1.50	1.51
词汇多样性(MLTD)	64.06	76.80	77.26	87.82
词汇多样性(vocd)	70.52	82.19	79.55	88.76
第一人称复数代词	44.82	35.57	34.20	32.79
实词(CELEX)	2.69	2.61	2.54	2.51
名词上下义	5.71	5.95	6.32	6.27

形符比(TTR)是衡量文本词汇多样性的常用指标。形符比越大,说明文本词汇越多样化。不同级别作文在形符比上有显著差异,$F(3, 186) = 2.87$, $p = .038$, $\eta^2 = 0.044$。从表 2 中可以看出,3~5 级写作样本的形符比呈递增趋势,级别越高,形符比越大,说明文本词汇越丰富;而 6 级写作样本形符比平均值略低于 5 级。事后 LSD 检验表明,5 级与 6 级写作样本在形符比指标上无显著差异($p = .738$)。

在词频方面,不同级别作文在 K1-K4 词频上有显著差异,$F_{K1}(3, 186) = 4.39$, $p_{K1} = .005$, $\eta_{K1}^2 = 0.066$; $F_{K2}(3, 186) = 25.65$, $p_{K2} = .000$, $\eta_{K2}^2 = 0.293$; $F_{K3}(3,$

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

186) = 7.26, p_{K3} = .000, η_{K3}^2 = 0.105; F_{K4}(3, 186) = 4.39, p_{K4} = .005, η_{K4}^2 = 0.081。随着写作样本级别的提高,K1 词频比例逐级下降,K2-K4 词频比例逐级上升,说明高级别写作样本使用词汇难度较高,词汇较多样。

在单词长度(音节)方面,级别间有显著差异,级别越高,单词越长,F(3, 186) = 4.91, p = .003, η^2 = 0.073。同时,在 MTLD 与 vocd 两个词汇多样性指标上,不同级别样本之间差异显著,整体呈现出级别越高多样性指标值越大的趋势, F_{MTLD}(3, 186) = 7.352, p_{MTLD} = .000, η_{MTLD}^2 = 0.106; F_{vocd}(3, 186) = 5.796, p_{vocd} = .000, η_{vocd}^2 = 0.085。尽管 5 级写作样本词汇多样性(vocd)指标略低于 4 级,但事后 LSD 检验表明,两者无显著差异(p = .534)。

在第一人称复数代词、实词(CELEX)和名词上下义这三个词汇指标上,不同级别的写作样本之间具有显著差异。写作样本级别越高,文本中使用的第一人称复数代词(如"we""us")越少,F(3, 186) = 3.032, p = .031, η^2 = 0.047,说明低级别写作样本使用的人称代词比较单一。在实词(CELEX)指标上,写作样本级别越高,文本包含实词比例越低,F(3, 186) = 15.689, p = .000, η^2 = 0.202,说明高级别写作文本中词汇重复率较低,使用词汇更多样化。在名词上下义指标上,不同级别的写作样本有显著差异,F(3, 186) = 13.724, p = .000, η^2 = 0.181。3~5 级写作样本的名词上下义指标呈递增趋势,级别越高,上下义指标越高;而 6 级写作样本的名词上下义指标略低于 5 级写作样本,但事后 LSD 检验表明,两者无显著差异(p = .703)。

3.2.3 句法复杂度

不同级别写作样本在 4 个句法复杂度指标上有显著差异,包括最小编辑距离、介词短语密度、动名词密度、Coh-Metrix 二语可读性指标。表 3 呈现这 4 个句法复杂度指标随写作样本级别的变化趋势。

表 3. 不同级别写作样本的句法复杂度指标

指标(平均值)	CSE3	CSE4	CSE5	CSE6
最小编辑距离(全部词汇)	0.87	0.88	0.88	0.89
介词短语密度	105.03	117.70	124.78	119.50
动名词密度	18.40	24.48	28.22	25.60
Coh-Metrix 二语可读性	28.66	25.69	24.87	23.26

表 3 表明,随着写作样本级别的提高,最小编辑距离(全部词汇)指标整体呈上升趋势,级别之间差异显著,F(3, 186) = 4.4, p = .005, η^2 = 0.066。但在该指标上,5 级样本的平均值略小于 4 级样本的平均值,但事后 LSD 检验表明,两者无显著差异(p = .478)。不同级别写作样本在介词短语密度(F(3, 186) = 7.28, p = .000, η^2 = 0.105)与动名词密度指标(F(3, 186) = 4.44, p = .005, η^2 = 0.067)方面有显著差异。3~5 级文本中使用的介词短语与动名词结构的密度平均值都呈递增趋势,说明级别越高,使用的介词短语和动名词结构越复杂;但从 5 级至 6 级呈下降趋势。事后 LSD 检验表明,两个级别在介词短语指标上无显著差异(p = .257),在动名词结构密度指标上同样无显著差异(p = .374)。不同级别写作样本在 Coh-Metrix 二语可读性指标上有显著差异(F(3, 186) = 7.22, p = .000, η^2 = 0.104),样本级别越高,该指标平均值越小,说明文本复杂程度越高。

3.2.4 语篇连贯性

就语篇连贯性而言,不同级别写作样本在 4 个指标上有显著差异,包括叙事性、词汇具体性、指称衔接、实词重叠指标。在叙事性指标上,样本级别越高,叙事性指标越低,F(3, 186) = 7.80, p = .000, η^2 = 0.112。一般而言,叙事文本更接近故事,文本口语化程度越高(Biber 1998)。而水平考试写作卷要求考生独立完成一篇议论文写作,所以文本的叙事性越低,文本越正式,语言越书面化。

表 4 表明,3~5 级文本的词汇具体性指标随级别提高而显著递增,F(3, 186) = 3.32, p = .021, η^2 = 0.051,说明级别越高,语篇越连贯;但 5~6 级呈下降趋势。事后 LSD 检验表明,两者无显著差异(p = .077)。3 级写作样本的指称衔接和实词重叠指标最高,说明 3 级样本语篇连贯性最好,4 级写作样本语言连贯性指标呈下降趋势,5 级样本连贯性指标有增长趋势,6 级写作样本的指标最低。事后 LSD 检验表明,在指称衔接指标上,3 级与 4 级、5 级与 6 级有显著差异(p = .012; p = .037),4 级与 5 级无显著差异(p = .434);在实词重叠上,3 级与 4 级有显著差异(p = .002),但 4 级与 5 级、5 级与 6 级无显著差异(p = .104; p = .065)。

表 4. 不同级别写作样本的语篇连贯性指标

指标(平均值)	CSE3	CSE4	CSE5	CSE6
叙事性	0.83	0.55	0.40	0.37
词汇具体度	-0.46	-0.30	-0.02	-0.27
指称衔接	0.65	0.12	0.28	-0.18
实词重叠(全部)	0.14	0.11	0.12	0.10

4. 讨论

4.1 水平考试写作卷与量表的对接

本研究是将校本考试与量表对接的一次有益尝试。通过收集内部效度证据和程序效度证据,验证了对接的效度,且发现写作卷的及格线对标量表 5 级,说明该考试定位准确。对接的成功在很大程度上取决于两个因素:一是具有专业性与代表性的专家小组;二是适合研究设计的标准设定方法。

标准设定本质上是专家小组的主观判断,专家的代表性与可靠性是对接是否具有效度的关键因素(Cizek & Bunch 2007; Dunlea 2015)。写作考试的对接结果通常直接建立在专家对于写作样本的评判之上,这对专家的专业化程度、专家对于考试与考生的认知和了解以及对量表的熟悉程度有很高的要求。本研究中的专家组成员包括考试命题人员、测试领域专家、具有丰富教学经

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

验的一线教师等,具有很强的专业性与代表性,为对接结果的科学性提供保障。其次,标准设定方法在整个对接过程中处于核心地位。不同的标准设定方法所得到的对接结果可能不同,因此选择适合研究设计的标准设定方法极为重要。分析性判断法适合写作测试的标准设定,可操作性强(Lim *et al.* 2013; Dunlea 2015),但专家需要对写作样本进行两轮独立评判,工作量大,对专家组成员要求较高。

项目组在对接过程中也遇到一些问题。首先,写作量表主要依据写作活动、文本体裁的不同对写作能力进行划分与定级。而水平考试要求考生完成一篇议论文写作,写作量表中关于议论文的描述语数量有限,专家在对接过程中参考信息不足。因此本研究一方面引入了组构知识量表,为专家提供不同级别的文本特征描述语;另一方面,在标准设定过程中,专家们以量表总表、写作量表与组构知识量表的区别性特征为依据,坚持以整体评分法进行评判。

4.2 不同级别写作样本的文本特征差异

在文本基本信息方面,本研究发现写作样本级别越高,文本篇幅越长,与前人研究结果一致,即高水平的二语学习者可以在规定时间内有更多的语言产出,语言也更流畅(Grant & Ginther 2000)。在词汇多样性和句法复杂度方面,整体而言,写作样本级别越高,使用词汇越丰富,句法越复杂,但在语篇连贯性维度上结论不一。

就语篇连贯性而言,在叙事性维度上,写作样本级别越高,叙事性指标越低,这与 Nelson *et al.* (2012)的发现基本吻合。该研究发现文本叙事性与学生的年级高低相关,年级越低,产出文本叙事性越强。但从词汇具体性、指称衔接和实词重叠这三个指标上来看,级别高的文本语篇连贯性未必更好。Crossley & McNamara (2012)曾发现高水平的二语学习者产出的文本语言更加复杂,但并非更加连贯;McNamara *et al.* (2010)也发现连贯性指标与作文得分之间不相关。上述发现都说明语篇连贯性与写作样本级别的高低无直接关联。

另外,本研究发现在多个指标上,5 级写作样本的平均值高于 6 级写作样本,如词汇多样性维度的形符比和名词上下义指标、句法复杂度维度的介词短语密度与动名词短语密度指标。虽然事后 LSD 检验表明 5 级和 6 级写作样本在以上指标上无显著差异,但从平均值来看,5 级写作样本在词汇多样性和句法复杂度方面高于 6 级写作样本。笔者认为这与水平考试写作部分的评分标准有关。评分标准中明确提出采用整体评分法,从语言和内容两方面进行评分。Coh-Metrix 仅从语言维度对写作样本进行分析,而忽视了文本的内容。从本研究对接结果来看,水平考试的及格线对接到量表 5 级。因此 5 级与 6 级的写作样本皆已达到水平考试要求,二者在词汇和句法的运用上水平相当,但 6 级写作样本在文章内容上更胜一筹。

基于此,笔者认为,当二语学习者语言水平较低时,词汇量的提升与句型的训练仍然是英语写作教学与学习的重点;但当二语学习者的英语能力达到一定水平时,想要实现写作能力的提升,则需更多关注文章内容。

5. 结语

本研究成功建立了校本考试与量表之间的对应关系,并从文本分析的角度就对接结果进行了效度验证。通过考试与量表的对接,可以使考试成绩报告更加丰富。结合量表中的"能做"描述语,教师和学生能够获得更加直观、全面的语言能力信息,指导教学活动的有效开展。此外,文本分析一方面提供了外部效度证据,另一方面提供了不同级别写作样本在词汇、句法和语篇层面的文本特征信息,为量表的进一步修订与完善提供了实证证据。诚然,本研究也存在不足之处。出于可操作性及试题库保密的考虑,本研究仅选取一套试题用于对接。其次,在外部效度验证阶段,本研究以定量手段探究不同级别写作样本的文本特征差异,其结果具有一定的局限性。未来研究可采用其他数据收集方法,开展交叉验证。

注释:

① 在分析性判断法中,专家结合评分标准与写作子量表对写作样本进行独立评判。样本的原始分仅用于计算临界分。

② MTLD,即 Measure of Textual Lexical Diversity,是词汇多样性指标(详见 McCarthy & Jarvis 2010)。

③ vocd 是词汇多样性测量工具(详见 McCarthy & Jarvis 2007)。

④ CELEX 是 Dutch National Expertise Centre CELEX (Centre for Lexical Information) 研发的语料库,包含荷兰语、英语和德语三种,约 1790 万词汇(Baayen *et al.* 1995)。

参考文献

[1] Abbott. M. L. Setting cut-scores for complex performance assessments: A critical examination of the analytic judgment method[J]. *Alberta Journal of Educational Research*, 2006(1): 25-35.

[2] Baayen, R. H., Piepenbrock, R. & L. Gulikers. *The CELEX Lexical Database (CD-ROM)*[M]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania, 1995.

[3] Biber, D. *Variation across Speech and Writing*[M]. Cambridge: Cambridge University Press, 1998.

[4] Cizek, G. & M. Bunch. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*[M]. Thousand Oaks, CA: Sage Publications, 2007.

[5] Cobb, T. Computing the vocabulary demands of L2 reading[J]. *Language Learning & Technology*, 2007(3): 38-63.

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

[7] Crossley, S. A. & D. S. McNamara. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[J]. *Journal of Research in Reading*, 2012(2): 115-135.

[8] Dunlea, J. Validating a set of Japanese EFL proficiency tests: Demonstrating locally designed tests meet international standards[D]. University of Bedfordshire, 2015.

[9] Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Z. Cai. Coh-Metrix: Analysis of text on cohesion and language[J]. *Behavior Research Methods, Instruments & Computers*, 2004(2): 193-202.

[10] Grant, L. & A. Ginther. Using computer-tagged linguistic features to describe L2 writing differences[J]. *Journal of Second Language Writing*, 2000(2): 123-145.

[11] Kaftandjieva, F. *Standard-setting. Section B of the Reference Supplement to the Preliminary Version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment*[M]. Strasbourg: Council of Europe, 2004.

[12] Lim, G. S., Geranpayeh, A., Khalifa, H. & C. W. Buckendahl. Standard setting to an international reference framework: Implications for theory and practice[J]. *International Journal of Testing*, 2013(1): 32-49.

[13] McCarthy, P. M. & S. Jarvis. Vocd: A theoretical and empirical evaluation[J]. *Language Testing*, 2007(4): 459-488.

[14] McCarthy, P. M. & S. Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment[J]. *Behavior Research Methods*, 2010(2): 381-392.

[15] McNamara, D. S., Crossley, S. A. & P. M. McCarthy. Linguistic features of writing quality[J]. *Written Communication*, 2010(3): 57-86.

[16] McNamara, D. S., Graesser, A. C., McCarthy, P. M. & Z. Cai. *Automated Evaluation of Text and Discourse with Coh-Metrix*[M]. Cambridge: Cambridge University Press, 2014.

[17] Nelson, J., Perfetti, C., Liben, D. & M. Liben. *Measures of Text Difficulty: Testing Their Predictive Value for Grade Levels and Student Performance*[M]. New York: Student Achievement Partners, 2012.

[18] O'Sullivan, B. City & Guilds Communicator Level IESOL Examination (B2) CEFR linking project case study report[A]. City & Guilds Research Report[R]. London: British Council, 2009:73-91. Available online at: www.britishcouncil.org/exam/aptis/resea.

[19] Papageorgiou, S. *Relating the Trinity College London GESE and ISE Examinations to the Common European Framework of Reference*[M]. London: Trinity College London, 2007.

[20] Plake, B. S. & R. K. Hambleton. The analytic judgment method for setting standards on complex performance assessments[A]. In G. Cizek(ed.). *Setting Performance Standards: Concepts, Methods, and Perspectives*[C]. Mahwah, NJ: Erlbaum, 2001: 283-312.

[21] Tannenbaum, R. J. & Y. Cho. Criteria for evaluating standard-setting approaches to map English language test scores to frameworks of English language proficiency[J]. *Language Assessment Quarterly*, 2014(3): 233-249.

[22] Vo, S. Use of lexical features in non-native academic writing[J]. *Journal of Second Language Writing*, 2019(2): 1-12.

[23] 何莲珍,孙悠夏. 提示特征对中国学生综合写作任务的影响研究[J]. 外语教学与研究,2015(2): 237-250.

[24] 江进林,韩宝成. 基于 Coh-Metrix 的大学英语六级与托福、雅思阅读语篇难度研究[J]. 中国外语,2018(5): 86-95.

[25] 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): 4-11.

[26] 张煜杰,蒋景阳. 任务复杂度对二语写作复杂度和准确度的影响[J]. 西安外国语大学学报, 2020(4): 49-54.

作者简介: 何莲珍,浙江大学外国语言文化与国际交流学院教授,博士,博士生导师,研究方向:应用语言学、语言测试。

阮吉飞,浙江省春晖中学英语组教师,硕士,研究方向:语言测试。

闵尚超,浙江大学外国语言文化与国际交流学院教授,博士,博士生导师,研究方向:语言测试。

责任编校　　**薛旭辉**

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net