

大学英语四级考试写作测评量表效度研究

邹绍艳¹ 范劲松²

(1. 青岛农业大学 外国语学院, 山东 青岛 266109; 2. 墨尔本大学 语言测试研究中心, 澳大利亚 墨尔本 3010)

摘要:本文以大学英语四级写作测试(简称 CET-4 写作测试)中采用的整体评分量表为例,采用混合研究设计方案,借助问卷和访谈调查了评分员对该量表的意见。研究结果表明:现有的 CET-4 写作评分量表在清晰度和可操作性方面基本令人满意,但在量表的完整性、提供的反馈信息以及量表在评分员培训中的作用这三个方面仍有较大的提升空间。根据研究结果,本文建议进一步运用实证研究方法为 CET-4 写作测试开发分项评分量表,以便为大学英语写作教学提供更加丰富的诊断性信息,确保考试对教学产生积极的后效和影响。

关键词:CET-4 写作测试;整体评分量表;分项评分量表;大学英语写作教学

中图分类号:H319.3

文献标志码:A

文章编号:1674-6414(2019)03-0148-09

0 引言

写作测试在大规模外语测试中被广泛应用,但与此同时,写作测试的评分却经常遭到批判。Weigle (2002)指出,写作测试的分数不仅体现了考生和测试之间的交互作用,而且涉及其他诸多因素,如测试任务、写作文本本身、评分员和评分量表等。这些因素之间的复杂交互作用导致写作测试所测量的语言能力构念往往不够清晰,因而影响外界对写作测试分数的理解(Cumming *et al.*, 2001),即影响写作测试的效度。根据美国三大研究机构即美国教育研究协会、美国心理学协会和全美测量公会联合颁布的《教育与心理测试标准》(AERA *et al.*, 2014: 9),“效度是指根据测试的分数做出恰当、有意义、有用的推论”。近年来,随着考试用户对考试效度的要求不断提高,如何提供更加合理、有效的分数解释,以便鼓励恰当的分数使用成为语言测试开发者面临的艰巨任务之一”(Chapelle *et al.*, 2008)。

在这种背景下,写作测试的评分量表成为语言测试领域关注的焦点。研究者们一致认为,评分量表能够体现写作测试实际测量的构念(McNamara, 1996; Turner, 2000; Weigle, 2002; Shaw *et al.*, 2007)。但是, Knoch (2009)指出,现有的大规模外语测试中使用的评分量表往往都存在以下问题:(1)评分量表的制定过程主要基于专家的直觉判断,因此无法体现写作文本的真实特征;(2)评分量表中经常使用一些印象化的术语,容易导致主观性的解读;(3)量表并未给每个等级的表现提供准确、详细的描述,不同等级之间的区分主要是依靠一些相对性的措辞。Knoch (2011)进一步指出,目前关于这些评分量表的开发过程基本都无从得知,这无疑加剧了学界对这些量表效度问题的关注。

以国内最大规模的外语考试——大学英语四级考试(以下简称 CET-4)为例,其写作部分评分量表的效度近年来引起了不少研究者的关注(蔡基刚, 2002; 费茜等, 2008; 简庆闽等, 2005)。费茜和赵毓琴(2008)指出, CET-4 写作测试中采用的评分量表比较笼统,评分标准也不够全面,因此量表的效度值得探讨。鉴于此,本文以 CET-4 写作测试中目前使用的评分量表为研究对象,调查评分员对该评分量表的意见,初步论证该量表的效度,抑或发现影响量表效度的因素,为进一步完善或提升 CET-4 写作测试评分量表的效度提供依据。

收稿日期:2019-01-22

作者简介:邹绍艳,女,青岛农业大学外国语学院副教授,博士,主要从事语言测试与评估、外语教学研究。

范劲松,男,墨尔本大学语言测试研究中心副教授,研究员,博士,主要从事语言测试与评估研究。

1 文献回顾

1.1 CET-4 写作测试及其评分

CET-4 写作测试要求考生根据所给的题目、提纲、情景、图片或图表,在 30 分钟内写一篇不少于 120 词的作文。自 1987 年首次施考以来,CET-4 写作测试经历了一系列变革,以适应教育部 2014 年颁布的《大学英语课程教学要求》(以下简称《教学要求》)的变化。变革的内容包括:作文的长度、最低及格分的设置以及写作顺序的调整等(辜向东等,2009)。自 2006 年 1 月以来,围绕 CET-4 写作测试实施的一项重要变革便是其评分系统的变化,CET-4 写作测试的评分由传统的纸笔评分转变为网上评分。目前已有不少学者针对 CET 写作测试的评分开展了实证研究(黄燕,2007;王跃武,2004;王跃武等,2006;张森等,2010),结果表明 CET-4 写作测试网上评分的效率和信度均高于比纸笔评分。但值得一提的是,这些研究关注的焦点基本都是 CET-4 写作测试的评分信度。相比之下,CET-4 写作测试评分的效度鲜少被关注。如前所述,评分量表能够体现写作测试实际测量的构念,只有对 CET-4 写作测试评分的效度进行充分研究,我们才能理解 CET-4 写作测试分数的意义。

根据杨惠中 & Weir (1998)的观点,CET-4 作文的评分主要采用整体评分法,评分中关注的方面包括:内容的相关度、语言的质量和语篇的连贯性。在评分过程中,评分员需要综合考虑这三个方面,然后给出一个总分。根据这些原则,大学英语四、六级考试委员会制定了一个包含五个等级的评分量表(见表 1)。

表 1 大学英语写作评分量表(杨惠中等,1998:134)

等级分数	等级描述
14 分	切题。表达思想清楚,文字通顺,连贯性较好,基本上无语言错误,仅有个别小错
11 分	切题。表达思想清楚,文字连贯,但有少量语言错误
8 分	基本切题。有些地方表达思想不够清楚,文字勉强连贯,语言错误相当多,其中有一些是严重错误
5 分	基本切题。表达思想不清楚,连贯性差,有较多的严重语言错误
2 分	条理不清,思路紊乱,语言支离破碎或大部分句子均有错误,且多数为严重错误

1.2 CET-4 写作评分量表的研究现状

目前,关于 CET-4 写作测试评分量表的研究主要分为两大类:(1) 通过与其他大规模外语写作测试的评分量表进行对比,阐述现有的 CET-4 写作评分量表的不足之处。例如,费茜、赵毓琴(2008)对比了 CET-4 写作和托福写作的评分量表,认为与托福写作的评分量表相比,CET-4 写作测试的评分量表较为笼统,过分关注语言使用的准确性,而忽略对写作内容和结构方面的描述,因此认为该量表在评分中的效果可能不太理想。(2) 分析评分员的评分过程和结果,论证评分量表的效度。高怀勇(2011)采用有声思维法研究评分员的评分过程,发现评分员在实际评分中关注的文本特征略多于现有的 CET-4 写作评分标准。例如,语言得体性和句子结构这两项评分标准在评分中被运用的频次较高,但这两项标准并未在现有的 CET-4 写作整体评分量表中得到体现。李航(2015)通过实证研究对比了现有的 CET 写作整体评分量表和一项自主开发的分项评分量表在 CET-6 写作评分中的作用,指出分项评分量表能帮助评分员更加细致、准确地区分考生的英语写作能力。尽管该研究初步探索了分项评分量表在 CET 写作评分中的应用前景,但由于该研究中使用的分项评分量表没有遵循严格的量表开发步骤,其效度有待论证。

综上,我们认为上述研究存在以下不足之处:(1) 目前的研究大都是基于主观论述得出结论,鲜有研究采用实证方法检验量表效度。(2) 现有的研究基本都缺乏相关理论框架的指导,因此即便运用了实证方法,也无法为评分量表的效度提供充分的证据。考虑到写作测试在大规模外语考试中的广泛应用,而且 CET 属于高风险语言考试,我们认为有必要研究目前使用的 CET 写作评分量表究竟能否有效地测量考试设计者想要考察的语言能力。鉴于此,本研究借鉴 Bachman & Palmer (1996)提出的测试有用性理

论框架以及 Knoch (2009)对评分量表效度指标的阐述 (见表 2),探讨评分员对 CET-4 写作评分量表的看法和意见。根据 Bachman & Palmer (1996)的观点,测试有用性一般通过六项指标来体现:信度、构念效度、真实度、互动性、考试影响和实用性。Weigle (2002)认为,互动性这一指标在评分量表的效度中很难体现。Knoch (2009)进一步提出了检验评分量表效度的具体论据,这些论据成为本文设计研究工具的依据。

表 2 评分量表效度指标 (Knoch,2009:65)

构念效度
量表提供了符合测试目的和测试背景的测试结果,评分员认为量表充分体现了测试的构念
量表能够有效地区分受试者,评分员认为量表能够发挥预期的作用
量表的描述语体现了当代应用语言学的理论和研究
信度
评分员能够稳定可靠地使用评分量表进行评分
真实度 (内容效度)
量表尽可能地反映目标语使用语境下读者对写作的看法
量表的影响 (后效)
受试者得到的反馈是相关的、完整的、有意义的
其他考试利益相关者 (stakeholders)认为考试的分数和得到的反馈是相关的、完整的、有意义的
量表对评分员的影响是正面的、积极的。
实用性
量表在评分中是实用的
量表的开发是实际可行的

2 研究设计

2.1 研究问题

本研究旨在解决以下研究问题:

- (1)评分员对现有的 CET-4 写作评分量表的总体看法如何?
- (2)评分员的评分经验在多大程度上影响他们对评分量表的看法?
- (3)现有的 CET-4 写作评分量表在哪些方面需要进一步改进?

为解决以上研究问题,本研究采用 Creswell & Clark (2011)提出的解释性混合研究设计方案,首先开展定量研究,然后在定量分析的基础上开展定性分析。混合研究设计的优势就在于其能够将定量研究与定性研究有机地结合起来,更加全面、深刻地解答研究问题。

2.2 研究工具

(1)调查问卷

本研究采用的问卷分为两部分:第一部分主要调查受访者的背景信息,如性别、教学经历、职称、教育背景和 CET-4 的写作评分经验。第二部分包括六道题目。根据 Knoch (2009) 的量表效度理论框架,问卷第二部分的六道题目设计如表 3 所示:

六道题目均采用李克特五级量表的形式(0~4),其中“0”代表“完全否定”,“4”代表“完全肯定”。问卷题目的作答形式之所以采用“0”和“4”作为量表的两个极端,而非传统的“1”和“5”,是为了促使受访者做出真正的选择,防止过多使用中间项“3”(Dörnyei,2003)。问卷调查首先于 2016 年 6 月在 11 名 CET-4 写作评分员中进行了试测,根据试测结果对问卷内容作了微调。

另外,本研究还设计了用于访谈的半结构式访谈大纲,包含两个问题:(1)如何看待 CET-4 写作评分

量表的描述语和等级划分;(2) CET-4 写作评分量表的哪些方面还需进一步改进。

表 3 问卷的题目以及涉及的效度指标

题目编号	题目内容	效度指标
1	现有的 CET-4 写作评分量表的等级描述的清晰度如何?	构念效度
2	现有的 CET-4 写作评分量表的等级描述的完整性如何?	真实度
3	现有的 CET-4 写作评分量表的等级划分的可操作性如何?	实用性
4	现有的 CET-4 写作评分量表为大学英语教学提供的反馈信息如何?	后效
5	阅卷点提供的样本作文对于 CET-4 写作评分的作用如何?	信度
6	正式阅卷前的评分培训的作用如何?	信度

2.3 研究对象

问卷调查的对象是来自上海、山东、吉林、四川等几所 CET-4 阅卷点的 179 名作文评分员,他们的背景信息如表 4 所示:

表 4 评分员背景信息

性别	教学经验	学位	四级作文评分经历
男:64 位 (35.8%)	5 年或以上 (85.5%)	博士 (10.6%)	3 次及以上 (49.7%)
女:115 位 (64.2%)	5 年以下 (14.5%)	硕士及以下 (89.4%)	3 次以下 (50.3%)

问卷数据被分析之后,12 位评分员受邀参与了后续访谈,其中六位评分员拥有三次及以上 CET-4 写作评分经历,而另外六位评分员参与 CET-4 写作评分的次数不足三次。

2.4 数据收集和分析

问卷数据于 2016 年 7 月四级作文评分期间被收集。为了解决研究问题一,本研究运用 SPSS 软件对问卷调查的数据进行描述统计分析,了解评分员对现有的 CET-4 写作评分量表的总体看法;为了解决研究问题二,本研究运用 SPSS 软件对问卷调查的数据进行独立样本 *t* 检验,探讨拥有不同评分经验的评分员对于 CET-4 写作评分量表的看法是否存在显著性差异。问卷数据分析完成后,研究者根据数据分析的结果邀请 12 位评分员进行了一对一的深入访谈。访谈的内容被录音并转写成文字,然后利用 Given (2008) 提出的归纳法对访谈中反复提到的主题进行归纳和分析。

3 研究结果

3.1 问卷调查的结果

3.1.1 描述统计分析

对问卷数据进行信度分析的结果表明,六道题目作答数据的内部一致性系数为 0.781,说明评分员对这些题目的作答都比较可靠(Barrett,2001)。对问卷数据的描述统计分析如表 5 所示。

表 5 显示,问卷中六道题目的得分均值都在 2.50~3.18 之间。由于问卷设计是采用 0~4 的五级量表,所以这六道题目的得分均值都可视为高于中等水平。这一结果表明评分员对现有的 CET-4 写作评分量表基本上持肯定态度。但从表 5 中也可以看出,第 4 道题目的得分均值(2.50)在所有题目中最低,仅略高于中等水平,说明评分员对于现有的 CET-4 写作评分量表为大学英语教学提供的反馈信息这一方面的认同度略低。而且,从六道题目作答数据的标准差来看,第四、五、六道题目的标准差较大,说明评分员在这几道题目上的意见分歧较大。

表 5 问卷题目的描述统计分析 (n = 179)

题目编号及内容	平均值	标准差	偏度值	峰度值
Q1 量表等级描述的清晰度	3.11	0.736	-0.619	0.458
Q2 量表等级描述的完整性	3.18	0.688	-0.667	0.810
Q3 量表等级划分的可操作性	2.93	0.790	-0.432	-0.149
Q4 量表为教学提供的反馈信息	2.50	0.932	-0.029	-0.666
Q5 样本作文的作用	3.06	0.846	-0.613	-0.237
Q6 评分培训的作用	3.01	0.874	-0.430	-0.720

3.1.2 独立样本 T 检验

按照评分员在 CET-4 写作评分中的经历,179 位评分员被分为两组:资深组(即参与评分三次及以上的评分员,n = 90)和新手组(即参与评分三次以下的评分员,n = 89)。对两组评分员在六道题目上的作答数据进行独立样本 T 检验,结果如表 6 所示。

首先,对两组评分员在六道题目上的作答数据进行 F 检验,结果表明两组评分员的数据差异符合方差齐性假设,因此满足进行独立样本 T 检验的前提条件。T 检验的结果显示,在看待现有的 CET-4 写作评分量表的清晰度(Q1)、完整性(Q2)、可操作性(Q3)、提供的反馈信息(Q4)、样本作文(Q5)这五个方面时,两组评分员的意见的 Sig 值都大于 0.05,说明两组评分员在这些方面没有显著性差异。但在看待评分培训的作用上,两组评分员意见的 Sig 值为 0.00,小于 0.01,说明两组评分员的意见之间存在显著性差异。具体而言,新手组评分员对评分培训的看法不如资深组评分员肯定。

表 6 两组评分员数据的 T 检验结果

题目编号及内容	组别	人数	均值	标准差	F 值	Sig 值
Q1 量表等级描述的清晰度	资深组	90	3.08	0.75	0.11	0.53
	新手组	89	3.15	0.70		
Q2 量表等级描述的完整性	资深组	90	3.17	0.72	0.18	0.81
	新手组	89	3.19	0.66		
Q3 量表等级划分的可操作性	资深组	90	2.89	0.83	2.08	0.46
	新手组	89	2.98	0.75		
Q4 量表为教学提供的反馈信息	资深组	90	2.51	0.99	2.39	0.91
	新手组	89	2.49	0.87		
Q5 样本作文的作用	资深组	90	3.09	0.86	0.15	0.60
	新手组	89	3.02	0.84		
Q6 评分培训的作用	资深组	90	3.27	0.83	0.00	0.00
	新手组	89	2.75	0.84		

3.2 访谈结果

为了进一步理解问卷调查的结果,研究者邀请部分评分员进行了访谈。研究者反复阅读了基于评分员访谈转写的文本材料,归纳出四个评论最为最多的主题:

第一,量表描述语的清晰度

受访的 12 位评分员中,67% 的评分员(五位资深评分员、三位新手评分员)认为,量表的描述语比较清晰易懂。但也有 33% 的评分员(三位资深评分员、一位新手评分员)指出,量表描述语的有些措辞比较

含糊,容易引起歧义。例如,三位评分员都提到,量表中的“严重语言错误”“语言错误相当多”“少量语言错误”之类的描述语有点笼统,不太容易理解。资深评分员 R2 指出:“有相当多的人认为句法结构错误是严重的语言错误,有人则认为不符合英文表达习惯的错误是严重的语言错误,也有人认为只有影响语义理解的错误才是严重的语言错误。我一般都是自己反复研读阅卷点提供的样本作文,理解什么是严重的语言错误,什么是少量的语言错误。”另外,新手评分员 R7 认为,11 分档和 14 分档的作文在“切题”方面没有做出明确的区分,而五分档和八分档的作文则都是要求“基本切题”,不明白其中有何差异。

访谈中涉及的这一主题与问卷的第一道题目基本吻合。该题目在问卷调查中的得分均值为 3.11,表明总体而言,评分员对于 CET-4 写作评分量表的清晰度基本满意。而对访谈的分析则显示出,1/3 的评分员对量表的清晰度仍有一定的质疑,尤其是涉及“语言错误”和“切题”方面的描述语不够具体、清晰。

第二,量表描述语的完整性

50% 的评分员(三位资深评分员、三位新手评分员)在访谈时提到了量表描述语的完整性。他们认为,量表的描述语不够完整,不足以涵盖四级作文的特点。例如,某资深评分员 R4 指出:“量表中对于内容和思想的描述基本就是围绕切不切题。实际上,有些作文尽管切题,但是内容空洞,缺乏严谨的论证,是不是应该考虑增加写作内容方面的描述?”资深评分员 R5 提到,量表的等级描述中没有涉及语言使用的得体性,“我在阅卷时发现,有的考生根本没有读者意识,比如在给老师写信时,连称呼都不得体。我认为评分量表应该对语言使用的得体性进行描述,引导学生在写作中树立读者意识,注意语言使用的得体性”。此外,新手评分员 R11 表示:“从 11 分档和 14 分档的描述语来看,好像这两个等级的四级作文就是在错误数量上有所区别。实际上,14 分档的作文一般用词上要比 11 分档的作文丰富、精准,句型结构也更加丰富,但是 CET-4 的写作评分量表并没有体现这些特点。”

访谈的这一主题与问卷调查的第二道题目内容基本一致。这道题目在问卷调查中的得分均值为 3.18,表明评分员对量表的完整性比较满意。而在访谈中,却有 50% 的评分员认为量表的完整性存在一定的问题。当被问及在问卷调查中是否忠实地表达了自己的意见时,评分员 R4 说:“问卷调查的时间比较短暂,没有仔细思考‘完整性’的含义,所以在问卷中还是选择了比较肯定的选项。”该评分员的说法在一定程度上能够解释这道题目在问卷调查时得分均值较高的原因。

第三,量表为大学英语教学提供的反馈信息

由于受访的 12 位评分员都是一线大学英语教师,他们无一例外地都表示非常关注 CET-4 写作评分量表为大学英语教学提供的反馈信息。其中 67% 的评分员认为,目前的量表无法为大学英语写作教学提供充分、有用的信息。例如,资深评分员 R3 指出:“无论从大学英语教学中还是从 CET-4 评分中都能看出,写作在大学生的听、说、读、写四项技能中是最弱的一项。但是不同水平的学生在写作中究竟存在哪些问题,有哪些地方还需要改进和提高,仅仅依靠课堂教学中教师反馈的信息是不够的。如果 CET-4 写作的评分量表能够提供比较详细的反馈信息,让考生明白自己写作中的问题所在,就能帮助他们有针对性地克服自己的不足,提高写作水平。”新手评分员 R9 则提到:“在 CET-4 写作正式评分前,培训员对不同水平的样本作文的特点讲解得比较详细,我在评分过程中也会不时地回顾这些作文的特点。但是,考生和大学英语教师无从得知不同层次四级作文的具体特征,如果能把这些信息补充到量表中,并在报道考试成绩时附带报道学生的作文分数和等级,那么教师和学生都将从中受益。”

访谈的这一主题与问卷调查的第四道题目内容一致。该题目在问卷调查的六道题目中得分均值最低(2.50),而访谈的结果恰好能够解释这道题目得分较低的原因。从上述分析可以看出,由于所有评分员都是大学英语教师,他们迫切希望 CET-4 写作评分量表能为大学英语教学提供更加具体、详细的反馈信息,从而促进大学英语教学和学习。

第四,评分培训的作用

访谈中,有 42% 的评分员(三位资深评分员、两位新手评分员)提到了阅卷点组织的评分培训的作

用。其中资深评分员 R2 提到:“我参加四级写作评分很多次了,已经基本掌握了评分标准。每次我参加评分培训都会重点看一下阅卷点提供的样本作文,熟悉一下作文的话题。不太需要专门去看评分量表,因为标准已经记在我心里了。”而新手评分员 R8 指出:“评分培训时,专家对样本作文的特征和得分点讲解得比较仔细。但一到实际阅卷时,还是经常不明白该怎么评分,因为评阅的作文与样本作文之间还是有一定的差异。所以,只能时不时找出样本作文进行对比,看看评阅的作文与哪个分数段的样本作文比较接近,然后再给分。”

从这一访谈结果不难看出,问卷调查中资深评分员之所以对评分培训的作用比较肯定,可能主要因为他们已经熟悉、内化了 CET-4 写作测试的评分标准。而相比之下,新手评分员对评分培训的作用认可度偏低则是因为评分培训未能帮助他们快速掌握有效的评分标准。

4 讨论

在考试的开发和效度验证过程中,征求考试利益相关者的意见十分必要(如 AERA *et al.*, 2014; Alderson *et al.*, 1995; Chun, 2008; Karelitz, 2013; Messick, 1989; Fan, 2014)。参与本研究问卷调查和访谈的都是大学英语教师,他们既是评分量表的使用者,也是 CET-4 考试利益的相关者,因此他们的意见对于检验和完善 CET-4 写作评分量表的效度而言十分重要。从研究结果来看,尽管评分员对目前使用的 CET-4 写作评分量表总体上持肯定态度,但是该量表在清晰度、完整性、为大学英语教学提供的反馈信息量这几个方面,仍有待完善和提升。另外,在看待 CET-4 写作评分培训的作用时,资深评分员和新手评分员的意见出现了显著性差异,新手评分员的看法不如资深评分员肯定,说明现有的评分量表在评分培训中的作用也有待于进一步加强。根据 Knoch (2009) 提出的评分量表效度框架,这几个方面是体现量表效度的重要指标,因此对量表的这几个方面进行修订或完善有助于提高 CET-4 写作评分量表的总体效度。

邹绍艳、高秀雪 (2015) 提到他们在大学英语教学过程中发现学生的语言表达空洞,没有实质性内容,思辨缺席现象严重。如果 CET-4 写作测试的评分量表能针对不同水平的写作表现提供详细的诊断性信息,必将对大学英语教学产生积极的后效,从而促进教学的提高。但由于 CET-4 写作测试目前采用的是整体性评分量表,很难克服这种量表自身的桎梏。根据 Weigle (2002) 的观点,整体评分量表无法捕捉考生写作中的具体不足和优势所在,因而影响考官区分写作文本在不同方面的特征,例如词汇的使用、修辞组织、语法结构的控制和准确性,更无法为考生提供有用的诊断性信息。Lee *et al.* (2008:1) 也强调,整体评分量表的这一缺陷对于二语学习者来说尤为明显,因为这类学习者的写作水平仍在发展中,他们可能会在写作的不同方面展现出不均衡的特点。虽然 CET-4 的考生都是处于大学阶段的外语学习者,但他们的写作能力发展并不均衡(刘建达等,2011; 唐锦兰等,2012)。鉴于此,我们建议运用实证方法为 CET-4 写作测试开发分项评分量表。现有文献表明,分项评分量表能够针对受试者的表现,提供诊断性信息,反映受试者的进步(如 Hamp-Lyons, 1986, 1991; Shaw, 2007; Weigle, 2002)。而且,分项评分量表在评分员培训中的作用更加明显,因为经验不足的评分员更容易理解并应用量表中的评分标准(Weigle, 2002; Weir, 1990)。

5 结语

金艳 (2005) 指出, CET-4 的考试目的一方面是客观地测量大学生的总体英语水平,另一方面期望对中国的大学英语教学产生积极的后效和影响。目前,相关研究已经证明 CET-4 写作测试的评分信度比较令人满意。这说明 CET-4 写作测试能够比较客观地测量考生的英语写作水平。但就考试为教学带来的后效或产生的影响这一点来说, CET-4 写作测试仍有较大的提升空间。因此,本研究建议进一步运用实证方法为 CET-4 写作测试开发分项评分量表,以便为大学英语教学提供更加丰富的诊断性信息,确保考试能够对教学产生积极的后效和影响。

参考文献:

- AERA, APA & NCME. 2014. *Standards for Educational and Psychological Testing* [M]. Washington D. C: AERA.
- Alderson, J. C., C. Clapham & D. Wall. 1995. *Language Test Construction and Evaluation* [M]. Cambridge: Cambridge University Press.
- Bachman, L. F. & A. S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests* [M]. Oxford University Press.
- Barrett, S. 2001. The Impact of Training on Rater Variability [J]. *International Education Journal* (1): 49-58.
- Chapelle, C. A., Enright, M. K. & J. M. Jamieson. 2008. *Building a Validity Argument for the Test of English as a Foreign Language* [M]. New York: Routledge.
- Chun, C. 2008. Comments on “Evaluation of the Usefulness of the Versant for English Test: a Response”: the Author Responds [J]. *Language Assessment Quarterly* (2): 168-172.
- Creswell, J. W. & V. L. P. Clark. 2011. *Designing and Conducting Mixed Methods Research* [M]. Sage publications.
- Cumming, A., Kantor, R. & D. Powers. 2001. An Investigation into Raters’ Decision-making, and Development of a Preliminary Analytic Framework for Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks [G] // *TOEFL Monograph Series*. Princeton, NJ: Educational Testing Service.
- Dörnyei, Z. 2003. *Questionnaires in Second Language Research: Construction, Administration, and Processing* [M]. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Fan, J. 2014. Chinese Test Takers’ Attitudes towards the Versant English Test: a Mixed-methods Approach [J]. *Language Testing in Asia* (1): 6.
- Given, L. M. 2008. *The SAGE Encyclopedia of Qualitative Research Methods* (Volume 1 & 2) [M]. London: Sage Publications Ltd.
- Hamp-Lyons, L. 1986. No New Lamps for Old Yet, Please [J]. *TESOL Quarterly* (4): 790-796.
- Hamp-Lyons, L. 1991. Scoring Procedures [G] // L., Hamp-Lyons. *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex: 241-276.
- Karelitz, T. M. 2013. Using Public Opinion to Inform the Validation of Test Scores. Research Report [EB/OL]. [Http://www.nite.org.il/files/reports/e387.pdf](http://www.nite.org.il/files/reports/e387.pdf).
- Knoch, U. 2009. *Diagnostic Writing Assessment: The Development and Validation of a Rating Scale* [M]. Frankfurt, Germany: Peter Lang.
- Lee, Y. W., Gentile, C. & R. Kantor. 2008. Analytic Scoring of TOEFL CBT Essays: Scores from Humans and E-rater [J]. *ETS Research Report Series* (1): 22-31.
- McNamara, T. F. 1996. *Measuring Second Language Performance* [M]. London and New York Longman.
- Messick, S. 1989. Validity [G] // R. Linn. *Educational Measurement* (3rd ed.). Macmillan: American Council on Education, 13-103.
- Shaw, S. D. & Weir, C. J. 2007. *Examining Writing: Research and Practice in Assessing Second Language Writing* [M]. Cambridge: Cambridge University Press.
- Turner, C. E. 2000. Listening to the Voices of Rating Scale Developers: Identifying Salient Features for Second Language Performance Assessment [J]. *Canadian Modern Language Review* (5): 555-584.
- Weigle, S. 2002. *Assessing Writing* [M]. Cambridge: Cambridge University Press.
- Weir, C. J. 1990. *Communicative Language Testing* [M]. Mahwah: Prentice Hall Regents.
- 蔡基刚. 2002. 大学英语四、六级写作要求和评分标准对中国学生写作的影响 [J]. *解放军外国语学院学报* (5): 49-53.
- 费茜, 赵毓琴. 2008. 大学英语 CET-4 写作评分标准中的问题分析 [J]. *外语教学理论与实践* (4): 45-52, 93.
- 高怀勇. 2011. 大学英语 CET-4 写作测试构念效度研究 —— 一项基于试题内部结构和评分过程的尝试 [J]. *外语测试与教学* (4): 33-41.
- 辜向东, 杨志强. 2009. CET 二十年写作试题分析与研究 [J]. *外语与外语教学* (6): 21-26.
- 黄燕. 2007. 大学英语四、六级考试网上阅卷情况调查 [J]. *外语界* (2): 82-88.
- 简庆闽, 陆建平. 2005. 2004 年 6 月大学英语四级考试写作部分的命题探讨 [J]. *外语与外语教学* (1): 32-33.

教育部高等教育司. 2004. 大学英语课程教学要求(试行)[M]. 北京: 高等教育出版社.

金艳. 2005. 大学英语四、六级考试改革思路与未来展望——解读《全国大学英语四、六级考试改革方案(试行)》[J]. 中国大学教学(5): 49-53.

李航. 2015. 整体与分项量表的使用对 EFL 作文评分信度的影响[J]. 外语与外语教学(2): 45-51.

刘建达, 黄亚萍. 2011. 词汇特征量对 CET 作文得分的影响[J]. 外语电化教学(3): 11-15.

唐锦兰, 吴一安. 2012. 写作自动评价系统在大学英语教学中的应用研究[J]. 外语与外语教学(4): 53-59.

王跃武. 2004. 大学英语四、六级考试作文网上阅卷实验研究[J]. 外语界(5): 74-79.

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

杨惠中, C. Weir. 1998. 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.

张森, 于朋. 2010. 大学英语四级考试作文网上评阅信度保障研究[J]. 外语界(5): 79-86.

邹绍艳, 高秀雪. 2015. 大学英语教学中思辨能力培养的现状调查——一项基于反思性问卷的研究[J]. 外语研究(3): 61-65.

A Preliminary Examination of the Rating Scale
Validity of the CET-4 Writing

ZOU Shaoyan FAN Jingsong

Abstract: Taking the holistic rating scale currently adopted in the CET-4 writing assessment as an example, this study investigates how CET-4 raters perceives the scale through questionnaire survey and interviews. The results indicate that the rating scale adopted by the CET-4 writing is generally satisfactory in terms of its clarity and practicality; however, there remains much room for improvement with respect to its completeness, feedback, and usefulness in rater training. Therefore, the study proposes to develop an analytic rating scale for the CET-4 writing based on empirical studies so that it can provide more useful diagnostic information for college English teaching and learning, and exercise more positive washback.

Key words: the CET-4 writing assessment; holistic rating scale; analytic rating scale; college English teaching and learning

责任编辑: 朱晓云

《中国莎士比亚研究》征稿启事

《中国莎士比亚研究》由中国外国文学学会莎士比亚研究会/四川外国语大学莎士比亚研究所主办, 为我国唯一的莎士比亚研究学术刊物。现特向海内外莎学专家、莎学爱好者征稿。

莎士比亚研究在世界文学艺术领域具有广泛而重要的影响。在一个多世纪里, 几代中国学者为莎士比亚在中国传播付出了毕生的心血。为了进一步推动中国莎士比亚研究的发展, 四川外国语大学决定在原有《中国莎士比亚研究通讯》(2011 年创刊) 的基础上, 编辑出版学术集刊《中国莎士比亚研究》(由出版社公开出版), 暂定为半年刊。

《中国莎士比亚研究》主要集中于莎士比亚作品文本、翻译、演出、改编及批评史的研究, 以及莎学书评、莎学名人访谈等, 文艺复兴时期戏剧研究、莎士比亚与同时代作家研究等原创性论文均可投稿。除上述议题外, 研究者也可就自己感兴趣的其他莎学论题撰写论文, 反映中国莎学在 21 世纪对莎士比亚的最新研究成果。

本刊不向作者收取任何费用, 论文发表后即向作者赠送两本样刊。

投稿地址: 文稿撰写完成后请发送电子邮件(用附件形式)到: zzsbyyj@163.com; 同时从邮局用挂号寄送一份打印稿到: 重庆市沙坪坝区壮志路 33 号 四川外国语大学莎士比亚研究所《中国莎士比亚研究》编辑部, 邮政编码: 400031

《中国莎士比亚研究》编辑部