外语评测

评分经验对 CET–4 作文评分人差异的影响研究[1]

徐 鹰
华南理工大学

© 2015 中国外语教育（3），74–84 页

提 要：本研究采用混合研究法分析了评分经验对评分人差异的影响。27 位 CET–4 作文评分人按照评分经验平均分为老手、中手和新手三组。通过严格培训后，他们对 30 篇 CET–4 模拟作文评分，并按重要性依次排列提供三条评分理由。分数的多层面 Rasch 模型（MFRM）结果表明：1）三组在严厉度上不存在显著差异；2）老手评分内在一致性最好，新手次之，中手最差；3）除一位新手外，三组都没有出现明显的随机效应。评分标准相关理由编码百分比的混合多元方差分析（MANOVA）结果表明：评分经验的主效应、评分经验和评分理由重要性的交互效应都不显著，但是评分理由重要性的主效应显著。这一结果说明经过培训后，评分经验产生的评分差异可以得到有效控制。

关键词：评分经验；CET–4 作文；多层面 Rasch 模型

1. 引言

写作测试的评分涉及到评分人对作文质量的判断和决策，主观性强，一直是语言测试界的研究重点。评分经验被认为是影响评分的重要因素，不同经验评分人在严厉度、关注的作文特征以及评分策略等方面均存在差异。Cumming（1990）发现评分经验丰富的评分人（下面简称老手）对作文内容的心理表征更全面，并且采用了大量不同的标准和自我控制策略；评分经验少的评分人（下面简称新手）则采用了相对比较少的标准并过度依赖于阅读能力。Huot（1993）的研究揭示新手和老手都关注内容和组织，同时也兼顾风格和文本外观，但老手对文本的反应呈现更强的个性化趋势。在评分过程中，老手和新手都能突破评分人的角色从而构建意义，但老手更能扮演教师和导师的双重角色。在阅读作文过程中，老手阅读更流畅，且倾向于在阅读完整篇作文后进行评价；而新手则喜欢边读边作评论，打断阅读的次数更多。尽管两组评分人都能根据评分标准评分，但是新手内化的评分标准不稳定，从而影响了他们和作文文本的互动；而老手则更娴熟地使用评分标准，并且对使用评分标准形成了一套内化的评分策略。Pula & Huot（1993）的研究也证实评分经验有助于评分人构建一套内化的评分标准。大量阅读作文比使用评分标准更能增加评分经验。新、老手的区别主要在于个人背景（阅读作文的经验）、专业培训和评分经验等三方面。Weigle（1998）的研究发现，和老手相比，新手给分严厉度更极端，而且自身一致性更差。Weigle（1999）进一步对 8 位新手和 8 位老手对两种不同作文题所给分数进行了研究，多层面

1 本文为广东省教育科学“十二五”规划项目“大规模考试中评分人决策风格的诊断研究”（2013JK013）、广州市社科规划青年项目“大规模语言测试评分人反馈信息有效性的实证研究——以 CET–4 为例”（14Q11）、广东省高等学校教学质量与教学改革工程项目“大学英语课程多元评估体系建设与实践”（X2WY/N913078a）和广东教育教学成果奖（高等教育）培育项目“以评促学——大学英语国家精品课程写作自我评估系统的支架作用研究与实践”的阶段性研究成果。

Rasch模型（MFRM）结果显示新手相对更严格，但是培训可以消除组间的严厉度差异；对评分人的有声思维的分析结果发现两组评分人在评分标准的使用难易度上存在差异，且评分人对作文题的合适度态度不同。Cumming *et al.*（2002）在开发TOEFL 2000考试时对作文评分人的决策过程进行了系统研究，结果发现培训过的ESL/EFL评分人关注的文本特征更广泛，且对语言的关注度高于修辞和内容，而EMT（英语为母语）评分人关注的文本特征更平衡。大部分评分人都认为评分经验和教学经验影响了他们的评分行为和评分标准。Erdosy（2004）采用有声思维对4位不同经验TOEFL作文评分人在无评分标准条件下对60篇TOEFL作文评分情况进行了深入研究，结果发现他们不仅给出的分数差异较大，而且评分过程也完全不同。Lumley（2005：107）采用了6个变量描述接受过培训且经验丰富的评分人：STEP考试的评分经验、参加STEP考试评分培训的经验、其他标准化考试的评分经验、教育背景、教学经验和内在一致性，其中前三个变量和评分经验直接相关。Suto *et al.*（2011）认为评分经验和专业培训都属于背景特征，有助于提高评分人的评分技能。Lim（2011）对MELAB考试作文评分的新手和老手的历时研究发现，新手评分人开始表现很差，但是很快能提高评分质量；老手能始终把严厉度保持在可接受范围内；评分人所评作文数量和他们的评分质量之间可能存在一定关系。

和国外研究相比，国内语言测试界专门针对不同经验评分人差异的研究尚属空白。笔者以“评分经验”为关键词或篇名搜索中国知网，得到的结果为零。但是，不少学者的研究对评分经验都有论述。首先，评分经验存在迁移现象。王海贞（2007，2008）在对TEM–4口试的效度研究中发现，以往的评分经验一方面可以帮助评分人快速掌握评分标准，另一方面也会造成对当前评分的干扰，如3位评分人会将以往的BEC评分经验无意识地迁移到TEM–4口试评分中。因此，评分经验可能导致分数产生系统性的差异。其次，评分经验和评分质量之间存在一定关系。在张洁（2009）对较好和较差CET–4作文评分人的对比中，尽管没有将评分经验作为区分两种评分人的标准，但是结果显示，较好评分人（60%的评分人的评分次数大于5次，40%的评分人在3—5次之间）参加CET–4作文评分的次数普遍多于较差评分人（40%的评分人的评分次数在3—5次之间，60%的评分人小于3次）。最后，评分培训效果也因评分人的经验不同而有差别。徐鹰（2014）对CET–4作文评分人个性化反馈信息有效性的实证研究发现，结合MFRM分析结果和评分理由编码分析结果的反馈信息对不同经验的评分人产生了不同影响：帮助经验丰富的评分人（评分次数为10次）调整评分模式，帮助缺乏经验的评分人（评分次数为1次）建立稳定的评分模式，帮助经验适中的评分人（评分次数为5次）固化已有的评分模式。

在研究设计上，前人研究手段主要是采用MFRM从定量角度对评分人分数差异进行比较，或采用有声思维手段探讨评分人决策的差异。由于有声思维研究工作量大，文本转写和编码耗时较长，因此存在样本量相对较小的问题，尽管该方法能对评分人的决策过程得到较深刻的认识，但本质上属于探索性研究，研究结果的概括性和外推力都存在一定局限性。

目前，我国各种大规模考试（如高考、CET）都设有写作任务，作文评分存在时间紧、任务重的特点。由于评分人经验往往参差不齐，因此不同经验评分人是否存在明显差异是影响考试信、效度的关键问题。

鉴于此，本研究以CET–4模拟作文为语料，结合MFRM和评分人评分理由编码的混合多元方差分析（MANOVA），旨在对不同经验评分人的评分差异进行全面研究，试图回答以下两个问题：

1) 不同经验评分人所给分数在严厉度、内在一致性和随机效应上是否存在差异？

2) 不同经验评分人对所给分数的解释理由是否存在差异？

2. 研究方法

2.1 作文

参加本研究的学生来自广州某大学2011级非英语专业的1个班和该校二级学院的3个班，共计200人。首先，学生就2012年6月CET–4作文题目On Excessive Packaging写一篇随堂作文。然后，笔者对全部作文按照CET–4评分标准（满分15分）进行了初评，并根据初评分采用分层随机抽样的方法抽取了30篇作文作为研究材料，这30篇作文涵盖了2分档（5篇）、5分档（6篇）、8分档（11篇）、11分档（5篇）和14分档（3篇）等5个等距的分数档。

2.2 评分人

共有来自广东省9所高校的27位评分人在2012年7月CET–4作文评分期间参与了本研究。他们背景类似，都讲授大学英语课程，拥有硕士学位，同时都通过了当次CET–4作文评分培训。在本研究中，评分经验的操作化定义是评分人以往参加CET–4作文评分的次数。鉴于徐鹰（2014）研究中的三位不同经验评分人的CET–4、作文评分次数分别为1、5、10，因此，我们将评分人按照以往参加CET–4作文评分次数平均分为三组，每组各9人：新手（Novice，简写为N）评分次数在1—5次之间；中手（Inhand，简写为I）在6—10次之间；老手（Veteran，简写为V）在11—15次之间。为方便标识，评分人代号按照“评分经验代号+评分人序号”方式编排。单因素方差（ANOVA）分析结果（$F(2, 24) = 64.508$，$p < .001$）说明三组评分人的评分次数存在显著差异，Scheffe事后检验发现三组评分人两两之间存在显著差异（$p < .001$）。

2.3 评分标准

CET–4作文评分采用总体评分法，从内容和语言两方面对作文进行综合评判，满分15分，包括5个等级：2分档（1—3分）、5分档（4—6分）、8分档（7—9分）、11分档（10—12分）和14分档（13—15分），每个等级对作文的内容和语言提出了具体的要求，评分标准细则包括切题、表达思想清晰程度、连贯和语言错误等4种文本特征（杨惠中、Weir 1998）。CET–4作文评分标准明确规定，评分人评分时应首先判断分数档，然后通过在每个档内加减一分的方法将5个分数档扩展为15个连续分数；此外，对于字数不足的作文，应酌情扣分。

2.4 研究过程

由于CET–4作文评分第一天主要安排评分人培训，30篇模拟作文复印后随机排序并在2012年7月评分开始后第二天结束时发给评分人，评分人利用当天晚上对这些作文按照他们在评CET–4正式作文时的标准评分，同时对每篇作文的分数按重要性顺序依次提供三条评分理由（最重要的是第一条评分理由，其次是第二条，最不重要的是第三条），材料在第三天评分工作开始前交回。采用三条评分理由是借鉴了Shi（2001）研究英语为母语和英语为非母语的EFL教师评估中国学生英语作文的做法。所有材料收回后，我们对材料进行了复查，从而确保了每位评分人的评分理由和具体分数的一致性。

2.5 数据分析

2.5.1 定量数据分析

反馈前、后评分人所评分数用FACETS 3.58（Linacre 2005）进行分析。多层面Rasch模型一共包括评分人、考生两个层面（评分经验作为虚拟层面），其数学模型如下：

$$\log(P_{ijk}/P_{ijk}\text{-}1) = B_i - C_j - F_k$$

P_{ijk}表示评分人j给考生i打k分数的概率；P_{ijk-1}表示评分人j给考生i打k-1分数的概率；B_i是考生i的能力；C_j是评分人j的严厉度；F_k是k分数相对于k-1分数的难度。

2.5.2 定性数据分析

27位评分人对30篇作文共提出了2,196条评分理由（有部分评分人没有给足3条评分理由）。在Shi（2001）评分理由编码基础上，笔者对评分理由进行了编码，并经过三轮次反复修改，最后确定了评分理由编码框架。为保证编码信度，一

位语言测试方向的博士研究生对7位评分人（占评分人总数的26%）的评分理由进行了编码，在所编码的606条评分理由中，不同编码人之间的信度（inter-coder reliability）达到了95.71%，从而保证了编码框架的可靠性。编码完成后，对评分标准相关理由的编码频数占全部评分理由的百分比进行混合多元方差（MANOVA）分析，包括两个自变量（IV）和5个因变量（DV），第一个IV是评分经验（组间变量），第二个IV是评分理由重要性（组内变量）；5个DV分别为切题、表达思想清楚程度、连贯、语言错误以及篇幅等编码频数占全部评分理由编码频数的百分比，前4个DV对应了CET–4作文评分标准规定的写作构念，最后一个DV（篇幅）和分数相关且评分标准有明确规定，因此也加以分析。这一做法的依据在于MANOVA功能强大，能同时处理多个相互之间有一定相关关系的DV，同时能有效控制犯第一类错误的概率。由于MANOVA要求DV是连续性变量（Tabachnick & Fidell 2013：12），因此可以适用于本研究中的评分理由编码百分比。在前人研究中，Cai（2012）采用MANOVA对TEM–4口试评分人有声思维的编码主题百分比进行了分析。因此，本研究采用MANOVA分析评分理由编码百分比，而不采用卡方检验分析评分理由频数。

3．结果

3.1 FACETS分析结果

表1是部分评分人（最严厉和最宽松的5位评分人）评分结果统计。

表1　部分评分人评分结果统计

评分人	严厉度	模型标准误	加权均方拟合统计量	标准Z值	点二列相关系数
V1	.91	.15	.77	-.8	.65
N8	.66	.14	.58	-1.7	.66
N6	.52	.15	1.21	.8	.61
V8	.45	.15	1.08	.3	.66
V4	.35	.16	1.32	1.2	.60
…	…	…	…		…
I7	-.26	.14	.40	-2.8	.67
N2	-.32	.14	.83	-.5	.64
V2	-.37	.14	1.10	.4	.63
N5	-.61	.16	1.40	1.4	.63
N3	-1.00	.17	1.47	1.6	.59
平均值	.00	.14	.92		.64
标准差	.38	.01	.27		.02

分隔比率：2.50；分隔信度：.86；卡方值：176.8；自由度：26；显著性：.00

Rasch模型分析结果显示27位评分人的严厉度存在显著差异，表1第二列的measure值显示最严的评分人V1（.91 logits）和最松的评分人N3（-1.00 logits）之间相差1.91 logits，平均严厉度为.00 logits，标准差为.38。12位评分人严厉度大于.00 logits，15位评分人严厉度小于.00 logits。按照Knoch（2011）的标准，严厉度在全部评分人平均值 ±.50 logits之外可以认为显著偏严或偏松，因此V1、N8和N6显著偏严；V2、N3、N5显著偏松，其他21位评分人评分比较合适。从评分人严厉度人数统计来看，新手评分偏严或偏松情况最多。

表1第4列是加权均方拟合统计量（Infit MnSq），也即评分人内在一致性。由于评分人通过了CET–4作文评分培训，因此取值范围相对较严，设定在0.7到1.3之间（McNamara 1996），> 1.3的评分人出现了不拟合（misfit），即评分人的内在一致性很差；< 0.7的评分人评分出现了过度拟合（overfit），即评分人的分数没有区分考生的差异。中手出现问题的人数最多，有4人（I9，I7，I3和I6），除I6外其他3人属于过度拟合；新手有3人（N8，N5和N3），其中N5和N3不拟合；老手表现最好，只有2人（V9和V4），不拟合和过度拟合各1人。I7加权均方拟合统计量最低（.40），N3值最高（1.47）。从人数比例来看，老手评分内在一致性最好，新手次之，中手最差。综合来看，新手评分随意性较大，中手评分倾向于打保险分，老手评分则较平衡。

表1最后一列显示27位评分人的点二列相关系数（PtBis）在.59到.67之间，平均值为.64，标准差为.02，没有出现典型的随机效应，在可接受的范围内（Myford & Wolfe 2004）。但N3的值（.59）低于平均值两个标准差，说明该评分人的评分具有随机性，在使用某些分数段时其评分有明显不一致的地方，导致部分考生的排序与其他评分人有显著差别，该结果和加权拟合分析结果类似。

此外，分隔信度（.86）和卡方分析结果（χ^2 = 176.8，df = 26，p = .00）说明评分人的严厉度有显著差异。评分人分隔比率（2.50）说明评分人严厉度的差异比测量误差大2倍多，按照Myford & Wolfe（2004）的分隔指数计算公式（4G+1）/ 3可算出分隔指数为3.67，说明评分人的严厉度大约可分为4个不同层次。

表2是评分人按经验分组的分析结果，三组评分人的严厉度都是.00 logits，无显著差异。加权均方拟合统计量都在0.7—1.3之间。点二列相关系数都在平均值 ± 两个标准差的合理范围内。

表 2　不同经验评分人分组结果统计

评分人分组	严厉度	模型标准误	加权均方拟合统计量	标准 Z 值	点二列相关系数
新手	.00	.04	.95	-.6	.61
中手	.00	.05	.81	-2.2	.62
老手	.00	.05	1.02	.2	.61
平均值	.00	.05	.93	-.9	.62
标准差	.00	.00	.09	1.1	.01

分隔比率：.00；分隔信度：1.00；卡方值：0；自由度：2；显著性：1.00

以上分析说明，三组评分人在个体层面上表现出明显不同的严厉度，但是在小组整体上没有出现明显不同；他们的内在一致性存在差异，相对来说老手的内在一致性最好，新手次之，中手最差；绝大多数评分人没有出现随机效应。

3.2 评分理由编码分析结果

3.2.1　评分理由编码对比

表3是评分人评分理由编码框架，包括6个一级编码和15个二级编码。

表 3　评分理由编码框架

一级编码	二级编码	定义
总体评价	总体评价	对写作质量的概括性评价或对交际有效性的评价
内容 *	总体评价	对内容的总体评价
	论点 *	对思想、观点的总体或详细评价
	论证 *	对论证的各个方面（如条理性、相关性、逻辑性、创新性等）的总体或详细评价
组织 *	总体评价	对作文组织结构的概括性评价
	对结构的具体评价	对结构的完整性、合理性、清晰度的评价
	段落	对作文导入、展开、结尾各部分的评价
	衔接和过渡 *	对句子和段落间的衔接和过渡情况的评价
语言 *	总体评价	对语言的总体评价
	可理解程度	对作文是否清晰易读的评价
	准确度 *	对作文总体和各方面（如语法、写作规范、词汇使用、句子使用）准确性的评价
	流利度	对语言是否流畅、意思是否连贯的评价
	套用模板	有没有使用准备好的模板
长度 *	作文篇幅 *	作文有没有达到字数要求
其他因素	书写、卷面、同情分	是否包括书写和卷面等文本特征不相关因素

注：* 代表该编码和评分标准相关

新手、中手、老手分别提供了716、745、735条理由，共计2,196条。所有评分人都知道对每篇作文要提供三条理由，但是有一些评分人认为用两条理由能对少数作文（尤其是篇幅不够的作文）提供足够理据，无须再添加第三条理由。因此，接下来分析比较各评分理由的百分比。

表4对评分标准相关理由百分比进行了统计。

表 4　评分标准相关理由百分比统计

		切题（%）	表达思想清晰程度（%）	连贯（%）	语言错误（%）	篇幅（%）	评分标准相关理由总计（%）
理由一	新手	55.85	15.09	2.26	13.96	2.26	89.43
	中手	58.52	6.67	0.74	21.11	3.70	90.74
	老手	48.88	10.45	0.00	11.94	10.82	82.09
理由二	新手	14.52	7.66	8.87	31.05	3.63	65.73
	中手	18.94	6.82	10.98	38.64	1.52	76.89
	老手	11.45	20.99	6.49	25.57	1.91	66.41

（待续）

（续表）

		切题（%）	表达思想清晰程度（%）	连贯（%）	语言错误（%）	篇幅（%）	评分标准相关理由总计（%）
理由三	新手	11.33	5.42	1.97	56.16	7.39	82.27
	中手	4.27	1.90	4.74	57.35	7.11	75.36
	老手	6.83	7.32	1.46	51.71	8.78	76.10
全部	新手	28.91	9.78	4.47	31.84	4.19	79.19
	中手	29.13	5.37	5.50	37.58	3.89	81.48
	老手	23.81	13.33	2.72	27.89	7.07	74.83

如表4所示，评分人在三条理由上都能提供和评分标准相关的评分理由，但在理由二上新手评分标准相关理由的百分比（65.73%）和老手的百分比（66.41%）相对偏少。在总体上评分人的评分理由基本上和评分标准相关，其中中手的评分标准相关理由百分比最高（81.48%），新手次之（79.19%），老手最少（74.83%）。这个结果也和其他研究（Cumming 1990；Milanovic *et al.* 1996；Sakyi 2000）利用有声思维得到的分析结果一致。老手评分经验丰富，已经建构了一套内化的评分标准，且能有效区分不同考生，所以他们敢于采用一些评分标准不相关的理由；而新手和中手内化的评分标准尚不稳定，因而更多采用评分标准相关的理由。如果新手采用多种评分标准之外的理由，则可能出现混乱，导致随机效应。N3就是一个典型例证。她是最宽松的评分人（严厉度为-1.00 logits），加权均方拟合统计量最高（1.47），且是唯一出现随机效应的评分人（PtBis值为.59），她所给的90条评分理由中只有57条和CET–4评分标准相关，占63.33%，不仅远低于新手平均值（79.19%），也低于老手平均值（74.83%）。这说明在经验不够的前提下，评分人还没有建立一套内化的评分标准，因此如果采用评分标准之外的评分理由很有可能产生评分严厉度、内在一致性以及随机性等问题。

在5个DV中，切题和语言错误最受评分人重视，所占百分比最大，而连贯和篇幅的百分比最小。主要原因在于CET–4作文评分采用总体评分法，容易产生较高的评分信度，但也容易产生评分还原主义（reductionism）倾向：即评分人要对学习者认知和语言维度的复杂性给出一个分数，他们必然倾向于关注某个典型特征。而二语学习者写作最大的特点是能力发展不均衡（Hamp-Lyons 1991；Weigle 2002），其中语言准确性特征最为明显，且容易辨识。此外，在实际评分过程中，评分人往往只用不到1分钟的时间给出分数，因此会出现这种现象。尽管这种做法有违于评分规定，但是却能有效地区分学生。这是因为中国二语学习者的写作能力普遍不高，所以评分人只需采用某种容易辨识且具有较强区分能力的文本特征（如语言准确性）就可以完成评分任务，因而他们也就不会采用其他要求更多认知努力判断的文本特征（如连贯）。这种做法会造成评分标准表征不足（rubric under-representation）的问题，因此需要加强对分数效度的培训。

3.2.2 评分经验和评分理由重要性的混合多元方差分析

对5个DV所作的相关分析发现，5个DV在低相关（.084）到中度相关（-.484）之间。Box协方差矩阵齐性检验发现Box’ M统计量为395.588，对应的F值为3.192（$p < .05$），说明各DV在不同经验评分人组间的协方差矩阵不一致，需要查看Pillai’s Trace指标。Barlett球形检验结果符合混合多元方差分析条件（$p < .05$）。此外，Levene方差齐性检验结果表明切题和语言错误在.05水平上没有达到显著意义（$p > .05$）。以上结果说明可以对评分标准相关理由百分比进行混合多元方差分析。分析结果见表5。

表 5　评分经验和评分理由重要性对评分标准相关理由百分比的多变量检验

效应	Pillai's Trace	F	p	Partial Eta Squared
评分经验（组间）	0.14	1.02	0.43	0.069
评分理由重要性（组内）	0.62	6.25	.000***	0.312
评分经验 * 评分理由重要性	0.13	0.47	0.98	0.032

*p < .05,**p < .01,***p < .001

如表5所示，评分经验主效应不显著（p > .05），评分经验和评分理由重要性的交互效应也不显著（p > .05），但是评分理由重要性的主效应显著（p < .01），且评分理由重要性对于评分标准相关理由百分比有较高的解释力（31.2%）。然后，可以对评分理由重要性在5个DV上进行简单比较，结果见表6。

表 6　评分理由重要性在 5 个 DV 上的简单比较

Source	DV	F	p	Partial Eta Squared
评分理由重要性	切题	26.134	.000***	0.421
	表达思想清晰程度	1.553	0.219	0.041
	连贯	1.477	0.235	0.039
	语言错误	10.455	.000***	0.225
	篇幅	2.275	0.110	0.059

*p < .05,**p < .01,***p < .001

如表6所示，评分理由重要性仅在切题和语言错误上存在显著差异。Scheffe事后检验结果表明，切题在第一条评分理由上的百分比显著高于其在第二条和第三条评分理由上的百分比（p < .01），而其在第二条和第三条的评分理由上的百分比之间没有显著差异（p > .05）；语言错误在第一条评分理由上的百分比显著低于其在第二条和第三条评分理由上的百分比（p < .01），而其在第二条和第三条上的评分理由上的百分比之间没有显著差异（p > .05）。该结果说明，不同经验的评分人都会在最重要的第一条评分理由上评判作文是否切题，在第二条和第三条评分理由上观察作文的语言错误，从而作出评分决策。

4．讨论

通过对不同经验评分人所给分数和评分理由进行对比分析，本研究主要有以下发现：

1）三组不同经验评分人在严厉度上没有明显差异，但是他们在内在一致性上存在一定差异，相对来说老手的内在一致性最好，新手次之，中手最差；此外，绝大多数评分人没有出现随机效应。首先，这一结果说明了评分培训的重要作用。由于本实验是在评分人通过严格的CET–4作文评分培训之后进行，因此评分培训能够有效降低不同经验评分人的评分差异。其次，这一结果反映了评分经验在评分实践中的关键作用。新手由于缺乏评分经验，只有尽可能地依赖评分标准，同时不拟合评分人数量最多，评分随意性较明显。中手虽然有一定的评分经验，但是其内化的评分标准还不成熟，还只能依赖评分标准，同时不拟合人数减少，但是过度拟合人数增加，给“保险分”趋势明显。而老手丰富的评分经验帮助他们构建了一套较成熟且有效的内化评分标准，能够关注更全面的文本特征，形成对文本连贯、合理的解读，达到最好的内在一致性。

2）不同经验评分人基本上能按照评分标准规定给出评分理由，但也有一小部分评分理由和评分标准无关，这就意味着评分人的评分决策包含评分标准不相关（rubric-irrelevant）因素；同时，在评分标准相关的5个DV中，切题和语言错误的百分比最大，连贯和篇幅的百分比最小，其中连贯百分比最低的老手只有2.72%，最高的中手有5.50%，因此在分数效度上也存在评分标准表征不足的问题，直接影响分数的可解释性及基于分数所作的推理。同时，评分经验和评分理由重要性对评分标准相关理由百分比的混合多元方差分析发现，评分经验主效应、评分经验和评分理由重要性的交互效应都不显著，但是评分理由重要性在切题和语言错误上存在显著差异。

尽管评分标准被认为是保证评分质量的关键因素（Bacha 2001；Knoch 2011；McNamara 1996），但本研究发现不同经验评分人都没有严格按照评分标准进行评分，从而威胁了分数的可解释性（即分数的效度），其原因主要在于以下三点。

首先，CET–4作文评分标准属于一种基于直觉性的量表（Fulcher & Davidson 2007），是由相关专家基于直觉制定（李清华、孔文 2011），尽管和《大学英语课程教学要求》（教育部高等教育司 2007）相一致，但存在理论框架不明晰、经验证据不足等问题。因此，应当从评分人对考生作文的评价中归纳作为评判依据的语言行为特征，从而进行修改和完善。

其次，CET–4作文评分采用总体评分法，容易产生较高的评分信度，但也容易产生评分人过分关注文本表层特征（如拼写、长度、用词和书写）的问题，从而忽略评分标准中的深层次特征，比如说连贯。语言错误就是一种明显的表层特征，也是评分人所给理由频数最多的特征。有10位评分人的评分理由中语言错误的百分比超过60%（N4和I1甚至超过90%），其中老手2人，中手5人，新手3人。中手表现最为明显，不仅过度拟合人数比例最高，而且评分理由相对单一。但是表1显示他们的严厉度、加权均方拟合度等都在正常范围内。这说明不少评分人存在评分还原主义倾向，即将多个评分标准简化成某个标准。总体评分法容易产生还原主义，因为评分人要对考生认知和语言维度的复杂性给出成一个分数，必然倾向于关注某个典型特征。

最后，二语学习者写作能力突出体现为语言准确性。由于实际评分时的工作量大，评分人对于水平参差不齐的考生，必须要给出一个比较恰当、精准的分数，同时评分人工作表现和报酬还取决于评分数量，因此评分人会形成某种更高效的评分策略：即在保证一定的评分准确度前提下最大化评分数量或用最小认知努力完成评分工作，此时从语言错误入手来判断考生能力无疑是一种比较高效且准确的方式。所以不同经验的评分人最终都选择在第一条评分理由上评判作文是否切题，在第二条和第三条评分理由上观察语言错误的方法。

综上所述，三组不同经验评分人整体严厉度没有显著差异，给出的大部分评分理由和评分标准相关。其中，老手的内在一致性最好，说明老手内化的评分标准最稳定。这一结论同Cumming（1990）、Huot（1993）、Huot & Pula（1993）的研究结果类似，有力地说明了评分经验对分数差异的重要影响。从解释学（Moss 1994；Petruzzi 2008）视角来看，评分是一种复杂的心理认知和决策过程，评分人在综合考虑文本印象、文本特征以及评分标准后最终给出一个总体分，并且会对所给分数提供一个连贯且整体的解释。因此，评分人的经验对分数解释产生重要影响。由于评分标准不可能覆盖所有的文本可能性，因此，评分人必须自己建构一套内化的评分标准和策略，从而形成了评分标准和主观印象之间的一种张力和不确定性。

5．结束语

本研究结果说明，经过培训后，不同经验评分人在整体上不存在严厉度区别，但在内在一致性上存在差异。虽然不同经验的评分人关注的文本特征不尽相同，但大部分的评分理由都和评分标准相关。上述结果说明培训可以帮助不同经验

评分人形成一个评分共同体，减少评分人差异。

本研究有两点启示：第一，只要经过严格培训，不同经验的评分人都可以胜任评分工作，但在同等条件下应优先选择多次参加评分的评分人。第二，现行的CET–4作文评分培训主要是利用评分人分数和专家分数的相关系数来控制评分质量，对于评分人的评分理由监控力度不大，因此这种过分关注分数信度的培训方式可能导致通过培训的评分人相关系数较高，但却容易产生评分标准表征不足和评分标准不相关两种威胁分数效度的情况，需要进一步改进。

本研究局限性在于评分人的选取是一种便利抽样，因而外推力有限；其次，研究样本是模拟作文，且评分人在一天评分工作之余参与研究，疲劳因素也会影响结论的准确性。下一步研究应在克服上述缺点的同时，采用有声思维、访谈等质性研究方法多边验证上述结论。

参考文献

Bacha, N. 2001. Writing evaluation: What can analytic versus holistic essay scoring tell us? [J]. *System* (3): 371-383.

Cai, H. 2012. Weighting Patterns and Rater Variability in an English as a Foreign Language Speaking Test [D]. Ph. D. dissertation. Los Angeles: University of California.

Cumming, A. 1990. Expertise in evaluating second language compositions [J]. *Language Testing* (1): 31-51.

Cumming, A., R. Kantor & D. E. Powers. 2002. Decision making while rating ESL/EFL writing tasks: A descriptive framework [J]. *The Modern Language Journal* (1): 67-96.

Erdosy, M. U. 2004. Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions (TOEFL® research report No. RR-70) [R]. Princeton, NJ: Educational Testing Service.

Fulcher, G. & F. Davidson. 2007. *Language Testing and Assessment: An Advanced Resource Book* [M]. London: Routledge.

Hamp-Lyons, L. 1991. Scoring procedures for ESL contexts [A]. In L. Hamp-Lyons (ed.). *Assessing Second Language Writing in Academic Contexts* [C]. Norwood, NJ: Ablex Publishing Corporation. 241-276.

Huot, B. 1993. The influence of holistic scoring procedures on reading and rating student essays [A]. In M. M. Williamson & B. A. Huot (eds.). *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* [C]. Cresskill, NJ: Hampton Press. 206-236.

Knoch, U. 2011. Investigating the effectiveness of individualized feedback to rating behavior—A longitudinal study [J]. *Language Testing* (2): 179-200.

Lim, G. S. 2011. The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters [J]. *Language Testing* (4): 543-560.

Linacre, J. M. 2005. *A User's Guide to FACETS: Rasch-model Computer Programs* [M]. Chicago: MESA Press.

Lumley, T. 2005. *Assessing Second Language Writing* [M]. Berlin: Peter Lang.

McNamara, T. F. 1996. *Measuring Second Language Performance* [M]. London: Longman.

Milanovic, M., N. Saville & S. Shuhong. 1996. A study of the decision-making behaviour of composition markers [A]. In M. Milanovic & N. Saville (eds.). *Performance Testing, Cognition and Assessment* [C]. Cambridge: Cambridge University Press. 92-111.

Moss, P. A. 1994. Can there be validity without reliability? [J]. *Educational Researcher* (2): 5-12.

Myford, C. M. & E. W. Wolfe. 2004. Detecting and measuring rater effects using Many-Facet Rasch measurement: Part II [J]. *Journal of Applied Measurement* (2): 189-227.

Petruzzi, A. 2008. Articulating a hermeneutic theory of writing assessment [J]. *Assessing Writing* (3): 219-242.

Pula, J. J. B. A Huot. 1993. A model of background influences on holistic raters [A]. In M. M. Williamson & B. A. Huot (eds.). *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* [C]. Cresskill, NJ: Hampton Press. 237-265.

Sakyi, A. A. 2000. Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions [A]. In A. J. Kunnan (ed.). *Fairness and Validation in Language Assessment: Selected Papers from the*

19th Language Testing Research Colloquium [C]. Cambridge: Cambridge University Press. 129-152.

Shi, L. 2001. Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing [J]. *Language Testing* (3): 303-325.

Suto, I., R. Nádas & J. Bell. 2011. Who should mark what? A study of factors affecting marking accuracy in a biology examination [J]. *Research Papers in Education* (1): 21-51.

Tabachnich, B. G. & L. S. Fidell. 2013. *Using Multivariate Statistics* (6th ed.) [M]. New York: Pearson.

Weigle, S. C. 1998. Using FACETS to model rater training effects [J]. *Language Testing* (2): 263-287.

Weigle, S. C. 1999. Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches [J]. *Assessing Writing* (2): 145-178.

Weigle, S. C. 2002. *Assessing Writing* [M]. Cambridge: Cambridge University Press.

教育部高等教育司，2007，《大学英语课程教学要求》[M]。北京：外语教学与研究出版社。

李清华、孔文，2011，二/外语写作测试评分研究综述[J]，《外语测试与教学》（4）：18-26。

王海贞，2007，基于评分过程证据的英语专业四级口试效度研究 [J]，《解放军外国语学院学报》(4)：49-53。

王海贞，2008，全国英语专业四级口试评分员对评分标准的理解和使用 [J]，《外语教学理论与实践》（2）：33-39。

徐鹰，2014，评分人个性化反馈信息有效性实证研究[J]，《天津外国语大学学报》（1）：62-69。

杨惠中、C. J. 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。

张洁，2009，评分过程与评分员信念——评分员差异的内在因素研究[D]。博士学位论文。广州：广东外语外贸大学。

作者简介

徐鹰（1979—），华南理工大学外国语学院副教授。主要研究领域：语言测试。电子邮箱：xuying@scut.edu.cn

Politeness strategies in written comments for college English writing

CHEN Meisong.....54

Within Brown & Levinson's framework of politeness strategies, this study analyzes teachers' written comments for students' writings, and finds that the most frequently used one is on-record politeness strategy while the least employed is the off-record, which signifies that teachers tend to ignore students' emotional needs and take abundant face-threatening acts with different levels of consciousness when correcting and criticizing their work. It is suggested that face value and politeness strategies should be well calculated when teachers are writing those comments, and teachers should be cautious in choosing appropriate politeness strategies and take students' face needs into consideration.

Detecting TEM-4 essay rater effects with a Many-Facet Rasch model

ZHANG Wenxing & ZOU Shen.....61

In order to explore TEM-4 essay rater effects, we apply a Many-Facet Rasch model of the item response theory to analyze severity/leniency, central tendency, randomness effect, halo effect, and rater bias. The findings show that (1) there is a significant difference in severity/leniency; (2) good consistency exists within the raters; (3) there is no significant central tendency; (4) some randomness and halo effect have been detected; (5) there is significant rater bias in the interaction of raters and examinees, and raters and the three traits of the rating scale. The findings have implications for rating quality control.

An empirical study of the impact of rating experience on the CET-4 essay raters' variability

XU Ying.....74

This paper studies the impact of rating experience on raters' variability following a mixed-methods approach. The 27 CET-4 essay raters with different experience were evenly distributed into three groups: Novice, In-hand and Veteran. After training, they were asked to mark 30 CET-4 mock essays and then write and rank 3 reasons for their ratings. Many-Facet Rasch Model (MFRM) results suggested that: (1) three groups were not significantly different regarding severity; (2) Veteran raters seemed to be the most internally consistent, while In-hand the least; (3) no randomness in ratings was displayed except a Novice. MANOVA on percentages of coded rubric-relevant reasons for the ratings showed that the main effect of experience and the interaction effect between experience and reason importance were not significant, but the main effect of reason importance was significant. The results proves that after training, raters' variability caused by experience can be effectively brought under control.