2013 年 9 月　　外语教学与研究(外国语文双月刊)　　Sept. 2013
第 45 卷　第 5 期　Foreign Language Teaching and Research (bimonthly)　Vol. 45 No. 5

读后续写题型研究*

广东外语外贸大学　王初明　亓鲁霞

提要：本文报道一项开发考试新题型的研究，探讨促学优势明显的“读后续写”任务能否用于外语水平考试。调查在高中生中取样，运用 Rasch 模式等统计方法分析。结果显示：从效度方面看，读后续写分数与高考难度相当的英语阅读理解和书面表达分数显著相关，还与教师给学生英语水平的排名显著相关，说明该题型能够有效测量学生的阅读与写作水平；从信度方面看，续写题型的可靠性在很大程度上取决于评分工具的质量、评分员的培训以及评分的操作，而非题型本身。依据本次调查的评分量表打分，能够较好地将各能力段的学生区分开来。

关键词：读后续写、阅读、写作、效度、信度

[中图分类号] H319.6 [文献标识码] A [文章编号] 1000-0429(2013)05-0707-12

1. 引言

当今语言测试设计把反拨效应视为首要考量因素(参阅 Bachman & Palmer 2010)，格外在意测试是否促教促学。影响反拨效应的一个决定因素是测试题型，题型的选择和使用影响教法应用和学习活动，牵涉到教法是否有效促学，学习是否遵循科学规律。因此，选好用对题型是测试产生正面反拨效应的一项基本保证。

语言测试的历史可以说是一部题型变更发展史。某时期某一题型的使用总有时髦理论的影子，反映人们对语言、语言学习、心理测量、教育心理等理论的当

* 本文获得国家社科基金项目“汉语二语学习的认知过程与高效率教学模式研究”(12&ZD224；主持人王初明)和教育部人文社科重点研究基地项目“互动和语境与第二语言发展”(10JJD740009；主持人王初明)的资助。衷心感谢参与本次调查的师生和评分员朱其韵、晏盛兰等，特别感谢蔡宏文和徐柳在统计分析上提供的帮助。

下认识，认识的深化通常带来题型的更替。例如，在行为主义理论和结构主义语言学的影响下，分离式的语法多项选择题一度成为主流题型之一（参阅 Bloomfield 1933；Lado 1961），其后随着单一能力假说（the unitary competence hypothesis）的提出，综合题型如完形填空、听写等得以普及（参阅 Oller 1979）。题型的改变自然带动反拨效应的变化，时间更是反拨效应的催化剂，各种相关因素随着时间的推移而发酵，倒逼题型变换。有的题型因久用而钝化，或因有更佳选择而遭淘汰。题型变化表明：有考试就有对新题型的需求，需求的存在决定了新题型的研发必须未雨绸缪，走在应用前面。

基于开发新题型的需要，更是基于对反拨效应重要性的认识，本文从语言测试的视角，针对促学优势显著的“读后续写”任务，开展效度和信度取证分析，探究其应用于外语水平考试的可行性。本研究在高中生中取样，运用 Rasch 模式等统计方法进行分析，展示调查结果，了解题型利弊，为题型百宝箱贡献一名新成员，供相关人士选用或进行更深入探讨。

2. 读后续写的考试应用潜力

读后续写是一种将语言输出与输入紧密结合、旨在加速提高学生语言运用能力的方法。此法的任务设计简单易行，只需将一篇读物的结尾抹去，让学生阅读截留部分，在理解的基础上续写，补全内容。读物或是故事，或是科普，或是对话，或是论说，只要内容连贯有趣，利于发挥想象力，切合学生的语言水平，并有一定长度（如 500 词以上），均可用于续写。王初明（2012）从理论上对读后续写的促学效应进行了论证，并开展了一系列研究，证明此任务用于二语教学益处多多。读后续写的主要促学优势是：（1）释放创造力和想象力；（2）融入读物提供的语境学习语言；（3）语言可模仿，内容则要创新；（4）顺着语篇续写，学习连贯表达；（5）补全的是语篇，而非单词或短语；（6）语言理解与产出结合紧密，显著提高外语学习效率；（7）对外语教学和学习几无负面影响。

迄今，开发读后续写的测试用途，国内外尚未见有报道。若将它应用于大规模考试，必须对其信度和效度进行检验。为此，我们开展了一项语言测试实证研究，研究的主要问题是：读后续写用于外语水平考试是否具备良好的信度和效度？为了回答这个问题，此次研究从四个方面检验读后续写的可行性：评分信度、评分量表、题型难度和共时效度（concurrent validity）。总目标是为后续深入研究奠定基础。

3. 研究方法和步骤

调查在两所高中开展。参加者是四个班 203 名高三学生，其中两个班来自

一所普通中学，另两个班来自一所较好中学。他们处于高三第一学期，通常花一定时间针对高考操练阅读理解和书面表达题，成绩可跟读后续写比较。调查要求他们完成一个英语测试，内容含英语阅读与写作。阅读部分由三篇短文和 12 道多项选择阅读题组成。写作部分有两项任务：一是读后续写，一是高考英语全国卷采用的书面表达。读后续写的短文选编自一本英语读物，长度为 331 词[1]，叙述一位名叫 Arthur 的人在上班途中遭遇的离奇事件，故事无结尾。该任务要求考生读后补全故事，续写长度为 150 词左右[2]。此外，该任务分两种形式：一种有段首语，另一种无段首语。有段首语的提供续写第一段和第二段的第一句话，旨在对续写内容有所控制，提高评分信度；无段首语的要求学生根据所提供的短文直接续写。两种形式的试卷按照电脑生成的随机抽样号码分类，测试时随机分派给学生。最终数据显示，105 名考生做了无段首语的试卷，98 名考生做了有段首语的试卷。阅读理解和书面表达题取自全国英语等级证书二级考试过期真题，与高考英语全国卷的题型相同，难度相当，但从未公开过。

测试分别在两所学校实施。第一所学校的学生先做阅读理解和书面表达题，后做读后续写；第二所学校学生的做题顺序相反。阅读理解与书面表达历时 55 分钟。读后续写历时 1 小时，完成后立即收卷。测试全程由英语任课教师和研究人员共同监督完成，以确保学生认真作答。

阅卷分为两部分。阅读理解部分采用的是客观题，评分快捷。书面表达和读后续写答卷的评分由两位大学英语教师承担，其中一位大学教龄 6 年，另一位 8 年；她们拥有语言测试硕士学位，均参加过高考英语评卷。评分之前她们接受了培训，先熟悉考题和评分标准，然后试评样卷，讨论存在的问题并达成共识，最后分别评阅 203 位学生的 406 份写作答卷，独立打分。

为了更可靠地检验题型的有效性，我们抽出一个班的学生（共 51 人），请他们的英语任课教师兼班主任为其英语水平排名，用作参考数据，与读后续写分数进行相关分析，以了解读后续写的共时效度。

4. 数据分析与结果

收集到的数据包括测试分数和教师排名，用多层面 Rasch 模型和 SPSS 进行统计分析。统计结果围绕前面提出的 4 个研究重点开展分析和讨论，先归纳测试结果的总体情况，然后集中分析读后续写的信度和共时效度。

1　作为促学练习，文章最好长一些，如不低于 500 词，但作为试题，受限于考试时间，不得已而缩短。

2　有兴趣的读者，可向我们索阅读物材料。

4.1 测试结果总体情况

表 1 和图 1 归纳了此次测试结果的总体情况。表 1 所列的基本数据显示，在总分为 15 分的两种写作题型中，两位评分员所给分数的平均值未超过一半(7.5)，表明书面表达和读后续写均偏难。但是，该结果能够较好地反映学生的实际英语水平，他们刚进入高三年级，此时距真考时间尚有 10 个月，做高考难度的试题，分数稍微偏低属于正常。

表 1. 测试分数

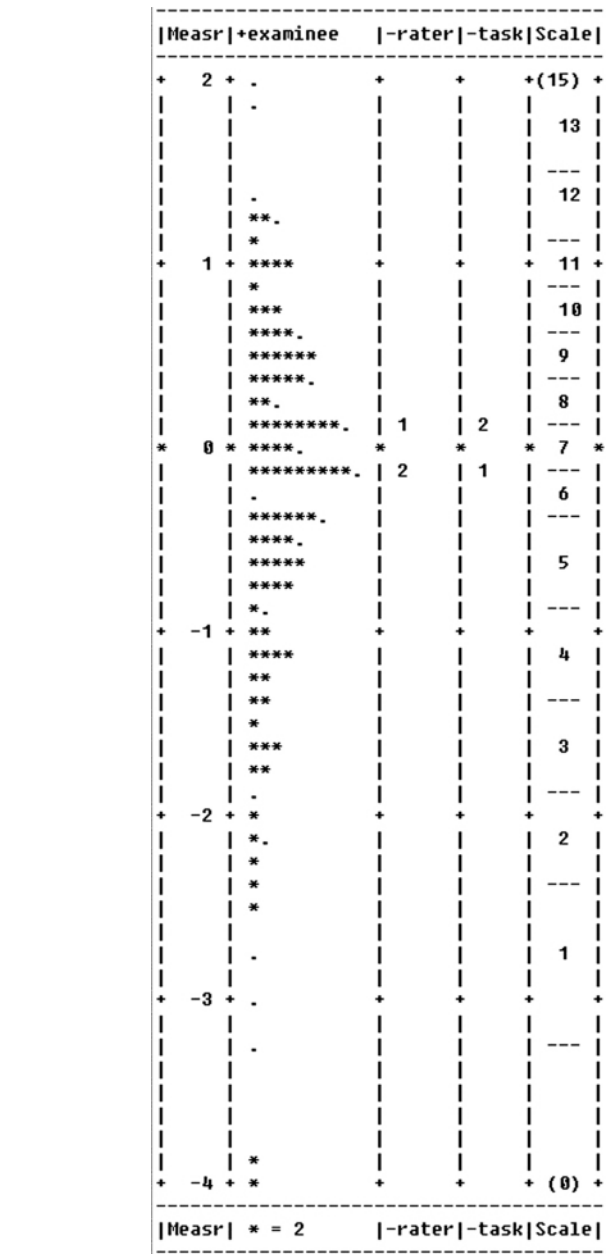
试卷结构	人数	最低分	最高分	平均分	标准差
阅读理解	203	0	12	7.12	2.13
书面表达(评分员 1)	203	0	15	6.06	3.37
书面表达(评分员 2)	203	0	15	7.23	3.26
读后续写(评分员 1)	203	0	15	5.89	3.57
读后续写(评分员 2)	203	0	15	6.16	3.18

接下来使用多层面 Rasch 模型(FACETS 3.58，Linacre 2005)对数据进行分析，该模型包括考生、评分员和任务层面，如下所示：

$$\log (P_{nijk}/P_{nijk-1}) = B_n - T_i - R_j - F_k$$

这里 P_{nijk} 表示评分员 j 在任务 i 下给考生 n 评分值 k 的概率，P_{nijk-1} 表示评分员 j 在任务 i 下给考生 n 评分值 k-1 的概率。B_n 指考生 n 的写作能力，T_i 为任务 i 的难度，R_j 为评分员 j 的严厉程度，F_k 指分值 K 相对于分值 K-1 的难度(Engelhard 1992)。

图 1 是书面表达和读后续写结果的一个总层面图，呈现 Rasch 模型分析所有层面的总体分布情况。从第 2 列可以看出，考生的能力度量值基本呈正态分布，跨度较大，约为 6 logits(不计最高和最低的 7 个度量值)，说明此次测试能够较好地区分考生的写作能力。此外，FACETS 考生层面分析结果报告所提供的总体统计量显示：分隔系数为 3.06，分隔信度为 0.90，卡方检验统计量为 1701.6(df＝202，p＝.00)，说明此次测试分数的差异具有统计上的显著意义，这些差异源自考生能力的差异而非测量误差，因此测试总体信度较高，可以区分考生的英语写作能力。FACETS 结果还显示，在总共 812 个估算中，标准化残差大于±2 的数量为 30，占估算总数的 3.7%，根据 Linacre(2005：104)，标准化残差大于±2 的数量约占或少于 5%时，表明数据与模型拟合。此外，在第 3 列中，1 号评分员列于 2 号评分员之上，说明 1 号比 2 号评分稍严厉，这与 SPSS 分析结果吻合(见表 1)。第 4 列显示，读后续写(任务 2)比书面表达(任务 1)稍难。



注：Measr＝能力度量值；examinee＝考生；rater＝评分员；task＝任务；scale＝分数等级

图 1. 书面表达和读后续写结果总层面图

4.2　总体评分信度

用 SPSS 对两位评分员的打分情况进行分析，结果显示两者评分的内部一致性较高（书面表达：$\alpha=0.88$；读后续写：$\alpha=0.86$），两人的评分显著相关（书面表

达：r＝0.79；读后续写：r＝0.76；p＜.01）。

从 Rasch 模型分析结果来看（见表 2），两位评分员的严厉程度不一，相差 0.24 logits，差异具有显著意义（分隔系数 5.85，分隔信度 0.97，卡方值 35.3，df＝1，p＝.00）。但是，她们的评分与模型的拟合度较好，加权均方拟合值分别为 1.20 和 0.80，未加权均方拟合值分别为 1.12 和 0.83（参看表 2），均在 0.70－1.30 的可接受范围之内，说明每位评分员打分的内部一致性较好（McNamara 1996）。因此，SPSS 与 Rasch 分析结果均表明，本研究的评分总体可信，可以用于进一步统计分析。

表 2.　评分员层面

评分员	严厉度	模型标准误	加权均方拟合值	未加权均方拟合值
1	.12	.03	1.20	1.12
2	－.12	.03	.80	.83
均值	.00	.03	1.00	.97
标准差	.17	.00	.28	.20

分隔系数：5.85；　分隔信度：.97　卡方：35.3；　df＝1；　p＝.00

4.3　读后续写评分信度

完成上述总体情况分析之后，我们用 Rasch 模型专门对读后续写分数进行了分析。模型与前面提到的相似，即 $\log (P_{nijk}/P_{nijk-1}) = B_n-T_i-R_j-F_k$，唯一区别是，$T_i$ 为续写形式 i 的难度而非任务 i 的难度，因为此次分析仅限于读后续写的分数，不包括书面表达分数，重点看评分员信度（表 3）、评分量表（表 4）和续写难度（表 6）的分析结果。

表 3. 读后续写评分员层面

评分员	严厉度	模型标准误	加权均方拟合值	未加权均方拟合值
1	.10	.06	1.05	.95
2	－.10	.06	1.01	.91
均值	.00	.06	1.00	.93
标准值	.14	.00	.28	.03

分隔系数：2.16；　分隔信度：.82；　卡方：5.7；　df＝1；　p＝.02

表 3 显示，两位评分员对读后续写的评分总体可信，与模型的拟合度较好，加权均方拟合值分别为 1.05 和 1.01，未加权均方拟合值分别为 0.95 和 0.91（表 3 第 4 和 5 列），均接近 1，在 0.70-1.30 的可接受范围之内，说明内部一致性较好。但是，她们的严厉程度相差 0.2 logits，1 号评分员较为严厉，2 号评分员较为宽松，差异具有显著意义（分隔系数 2.16，分隔信度 0.82，卡方值 5.

7，df＝1，p＝.02）。评分员因素对写作评分的影响已有不少研究（如 Lumley & McNamara 1995；刘建达 2010），有研究者采用 Rasch 模型分析评分的内部一致性、宽严程度等，为评分员提供即时反馈，使其了解自己的评分情况，以此改善评分信度与效度。此类研究成果有望用于实际评分操作，提高评分质量（Elder *et al*. 2005；Knoch 2011）。如果在大规模考试中采用读后续写题型，对评分员的培训尤为重要，需要根据实证研究结果采取有效措施，提高培训效率和评分信度。除评分员因素外，对信度产生影响的另一重要因素是评分标准和评分量表，下面讨论读后续写评分量表的使用情况。

4.4　读后续写的评分量表

本研究共用了两个评分量表，书面表达和读后续写量表，前者采用全国高考英语卷多年沿用的评分量表，后者参照前者制定。两个量表均采用综合评分法（holistic marking），总分 15 分[3]，分为五个档次，每档三个分值。因为本研究的重点是读后续写，在此仅分析读后续写评分量表的使用情况，量表各档次对续写要求有具体描述，涵盖四方面：内容、结构、语言准确性和语言丰富性。续写内容要求与所读文章的内容高度相关，情节发展连贯合理；续写结构要求续写部分有效使用了语句间的连接词语，使作文结构紧凑；语言准确性和丰富性要求续写中使用多样化的词语和句型，并且表达准确。Rasch 模型分析结果列于下页表 4。

表 4 显示，读后续写评分量表基本达到要求，但需改进。第 1 列为分值，第 2 和第 3 列是各分值使用的频率和百分比，从中可以看出：评分员使用了全部分值，两头使用较少，0 和 1 分及 14 和 15 分共使用 18 次，占总数的 5%。

检验评分量表质量的一个重要指标是平均能力度量值（average measure，见第 4 列）。能力强的考生得高分，能力弱的得低分，因此，能力度量值应从低到高单向递增（Bond & Fox 2001）。第 4 列的能力度量值总体趋势从低到高递增，但对应第 1 列 11 分和 15 分的能力度量值低于前边的度量值，意味着得到这两个分数的考生能力低于得到低一个分数的考生，评分出现偏差。

评分量表的另一个重要指标是未加权均方拟合值（见第 5 列）。由每个分数段考生的平均能力度量值与模型对考生能力的预测值进行对比而得出该指数，接近 1 为理想值，大于 2.0 则说明得到该分值考生的预测分数与实际分数有较大差异，即分数不能反映其实际水平（Linacre 1999）。表 4 中未加权均方拟合值绝大部分低于 2.0，唯一例外是对应 11 分的指数。这可理解为读后续写评分量

[3]　高考英语采用的评分量表总分为 25 分，分为五个档次。本研究评分培训时发现各档次内部分级太多（5 分），不易操作，为了提高评分效率和信度，我们将每个档次的分数减至 3 分，总分为 15 分。

表可以比较准确地区分各能力段的考生，只有 11 这个分值不能准确反映取得该分数考生的能力。

表 4.　续写评分量表层面

分值	频率	百分比（%）	平均能力度量值	未加权均方拟合值	阈值	标准误
0	4	1	－.12.37	1.0		
1	9	2	－.11.76	.4	－12.83	.62
2	21	5	－5.58	.8	－10.43	.65
3	38	10	－2.07	1.1	－3.88	.35
4	54	14	－.78	1.1	－1.80	.23
5	50	13	.11	.9	－.17	.19
6	45	12	1.06	.9	.73	.19
7	40	10	1.63	.8	1.42	.18
8	32	8	2.05	.8	2.06	.19
9	34	9	2.43	1.1	2.20	.19
10	15	4	2.81	.7	3.40	.22
11	7	2	2.55*	2.1	3.59	.24
12	21	5	3.26	.6	1.94	.25
13	9	2	3.29	1.2	4.11	.34
14	3	1	4.47	.6	4.80	.59
15	2	1	4.09*	1.8	4.85	.86

还有一个重要指标是阈值（threshold or step calibration）。阈值反映相邻分值概率曲线的相交点，对该指标的要求与能力度量值相同，须单向递增，否则被视为无序阈值（disordered）（Linacre 1999）。第 6 列的阈值总体呈单向递增，但对应 12 分值的阈值（1.94）低于前边四个分值的阈值，出现逆反（reversed），说明该分值不能准确反映得到该分数考生的能力。此外，相邻阈值的差距也是检验评分量表质量的重要指标。对于划分三个分数等级的量表，相邻阈值的差距需达到 1.4 logtis，才能说明每个等级相对独立，等级之间的差异能够反映能力差别。但是，差距并非越大越好，而是随着等级的增加而减少，对于五级量表，差距只需达到 1.0 logits。而且，差距达到或大于 5 logits 时，则说明量表等级之间的差距过大，影响测量的准确性（同上）。本研究采用的评分量表含 15 个等级，相邻阈值差距略小于 1 logits 应可接受，不过其中有些差距过小，如 6 与 7、7 与 8、8 与 9、10 与 11、14 与 15 之间（见表 4 第 6 列），均小于 1.0 logits，未达要求。该评分量表分为五个档次，每档 3 分，1～3 分为第一档，13～15 分为最高档。差距过小的分值大都在同一档次之内，说明档次之内的三个分值容易混淆。

五个档次之间的阈值三个大于 1.0 logits(3 与 4、9 与 10、12 与 13),说明评分量表对这几个档次标准的描述比较清晰,易于操作。但是,6 分与 7 分之间的差距小于 1.0 logits,说明第二档和第三档的区分不明显,难以把握。

从分析结果来看,读后续写评分量表有 15 个等级,划分过细偏多,多个等级之间的区分不明显,影响评分质量。参看国际相关评分量表,等级均较少,TOEFL 网考写作评分量表为 6 个等级,IELTS 为 9 个等级。而我国的大规模考试作文评分量表等级过多,如大学英语四级作文评分量表为 15 分,英语高考全国统一卷为 25 分。在实际评分操作中,评分员能否准确区分相邻等级?减少等级能否提高评分质量?这些问题需要进一步研究。

4.5　读后续写的难度

读后续写的难度属于任务层面,这里从两个角度进行分析,一是与书面表达难度的比较,一是对有段首语和无段首语的续写进行比较。Rasch 模型分析结果显示(参看表 5),读后续写题(任务 2)难度大于书面表达题(任务 1),前者的难度值是 0.27,后者为 0.06,两者相差 0.21 logits,差异虽然不大,但具有统计上的显著意义(分隔系数 5.02,分隔信度 0.96,卡方值 26.2,df=1, p=.00)。导致该结果的主因可能是考生对题型的熟悉程度。书面表达是高考英语的传统题目,从 1985 年沿用至今,许多省市的中考也采用该题型,考生大量操练,熟悉程度高,高中生做起来相对容易。参与本研究的考生从未接触过读后续写,自然有一定难度。如果该题型被大规模考试所采用,教学中定会针对它开展训练,难度会因此有所下降。此外,两项任务的拟合值均在 0.7～1.3 的可接受范围之内,说明与模型的拟合较好(见表 5 第 4 和第 5 列)。

表 5. 任务层面

任务	难度	模型标准误	加权均方拟合值	未加权均方拟合值
2	.27	.03	1.06	1.01
1	.06	.03	.93	.93
均值	.16	.03	.99	.97
标准差	.15	.00	.10	.05

分隔系数:5.02;　分隔信度:.96;　卡方:26.2;　df=1;　p=.00

如前所述,读后续写分为有段首语和无段首语两种形式,针对这两种形式的 Rasch 模型分析结果显示,提供段首语对续写有一定影响,增加了任务的难度。表 6 第 2 列显示续写形式 1(无段首语)的难度值是－0.12,形式 2(有段首语)难度值为 0.12,两者的差异是 0.24 logits,差异不大,但具有统计上的显著意义(分隔系数 2.73,分隔信度 0.88,卡方值 8.5,df=1, p=.00)。这样的结果说明,读

后续写若提供段首语，将提高任务难度，很可能是因为段首语在一定程度上限制了考生的思维，使其无法根据自己的英语表达能力自由创造内容。不过两种续写形式所得数据均与模型拟合，拟合值在 0.7～1.3 的可接受范围之内（见表 6 第 4 和第 5 列）.

表 6. 续写形式层面

续写形式	难度	模型标准误	加权均方拟合值	未加权均方拟合值
2	.12	.06	1.00	.96
1	−.12	.06	1.07	.90
均值	.00	.06	1.03	.93
标准差	.17	.00	.04	.04

分隔系数：2.73；　分隔信度：.88　卡方：8.5；　df＝1；　p＝.00

4.6　读后续写的共时效度

本研究对读后续写效度的探讨仅限于共时效度，从两个角度达到目的，一是了解它与阅读理解和书面表达题的关系，二是与熟悉考生英语水平的教师所提供的学生排名进行比较。

测量阅读理解能力采用了三篇短文与 12 道多项选择题。如前所述，在本研究之前，这组题目在等级考试中得到验证，其信度、难度、区分度均达到可接受的标准。本研究中写作的评分总体可信（见 4.2 节），因此，我们采用两位评分员所给分数的平均值作为考生的书面表达和读后续写分数，对这两组分数和阅读理解分数进行相关和回归分析，探索三者之间的关系。分析结果表明，续写与书面表达和阅读理解存在一定相关，具有统计上的显著意义（$p<0.01$），与书面表达的相关高于与阅读理解的相关，（$r=0.680$ vs. $r=0.458$），阅读理解与书面表达也呈显著相关（$r=0.535$）。

多元线性回归分析显示，以学生的读后续写分数作为因变量，阅读理解和书面表达得分为自变量，回归模型检验差异有统计学意义（$F=90.27$，$P=.001$），阅读理解和书面表达得分与续写分数相关（$R=0.689$），可以建立续写得分与阅读理解和书面表达得分的回归方程（见表 7）。此外，阅读理解和书面表达进入回归方程后，可以解释 47％的续写得分变异。该结果说明，读后续写在一定程度上考查到学生的阅读理解与书面表达能力。

表 7. 考生读后续写得分与阅读理解和书面表达得分的回归分析（n＝203）

	回归系数（B）	决定系数（R^2）	t	p
（常数项）	0.519		0.911	0.363
阅读	0.197	0.474	2.183	0.030
写作	0.617		10.033	0.000

把英语教师的学生排名分别与读后续写、书面表达和阅读理解分数进行相关分析，三者均有显著意义（p＜.01）。相比之下，排名与读后续写的相关系数最高，说明读后续写测试结果与教师所观察的学生英语水平有较高的相关，能够较为准确地反映学生的英语水平，意味着共时效度良好（见表 8）。

表 8. 教师排名与读后续写、书面表达及阅读理解的相关分析结果（n＝51）

	读后续写	书面表达	阅读理解
相关系数（r）	0.677	0.620	0.551
显著性（p）	0.000	0.000	0.000

5. 结语

本文首次报道读后续写题型的一项可行性研究，对此题型的效度和信度进行了初步检验，结果是肯定的。从效度方面看，读后续写与阅读理解和书面表达显著相关，说明该题型能够有效测量学生的阅读与写作水平；同时，续写还与教师排名显著相关，说明总体上能够反映学生的英语水平。从信度方面看，续写题型的可靠性在很大程度上取决于评分工具的质量和评分员素质及评分的操作，而非题型本身。本研究聘请的两位评分员经过训练，每人的内部一致性较好，但评分的严厉程度存在一些差异，需在未来的培训中注意。评分量表基本上能够达到预期目的，续写分数也能较好地将各能力段的学生区分开来。研究结果还显示，对写作这样的主观题型，评分量表的等级不宜太多。为了进一步验证读后续写的效度和信度，将来的研究应该拓宽范围，不仅在高中，还可以在大学甚至初中收集数据和证据，测定难度，完善题型，使之更好地为外语考试服务。

需要注意的是，读后续写尽管面临主观题型如何提高评分信度的问题，但还是具有明显优于其他写作题型的地方[4]。它不仅能够有效考查学生的读写综合能力，还能够借助考试的反拨效应，促使外语教学和学习将理解和产出紧密结合起来，将内容的创造与语言的模仿有机结合起来，藉此提高外语教学和学习效率。此外，读后续写题型中的读物会影响考生的续写表现。读物选得好，可直接产生良性学习效应（参阅王初明 2012），这正是外语教学所期盼的，也是读后续写的促学魅力所在，恰好说明此题型极具考试用途的研发价值。

开发读后续写考试应用价值的研究仍有许多工作要做，也值得做，本文仅为抛砖引玉。未来调查可涉及续写题型的多个方面，其中包括：（1）哪种体裁的阅

[4] 由于考试对作文长度和时间有限制，读物和续写作文不可能很长，会因此掣肘学习能力的充分发挥。需要注意的是，任何好的学习任务，一旦用于考试，其促学效果都会打折扣。

读材料更适合考试续写任务，记叙文抑或议论文？(2)读后续写的指令对续写有何影响？(3)篇幅多长的读物有助于续写和评分？(4)有无必要在阅读材料之中划出有用的词语，要求考生在续写中采用，并针对这些词语给分？(5)评分量表如何制定才能更科学地反映考生的语言能力？采用几个等级才算合适？(6)有无必要考察对阅读材料的理解？(7)续写质量与阅读理解之间存在怎样的联系？(8)针对读后续写的评分员培训如何开展？(9)如何开发机考或网考形式？(10)如何应用于对外汉语测试？随着研究的推进，值得探讨的问题还会更多，一旦大规模考试采用该题型，针对其实际反拨效应的研究就应立即跟进。请切记：考试终归要服务于教学和学习。

参考文献

Bachman, L. & A. Palmer. 2010. *Language Assessment in the Real World: Developing Language Tests and Justifying Their Use* [M]. Oxford: OUP.

Bloomfield. L. 1933. *Language* [M]. New York: Holt.

Bond, T. & C. Fox. 2001. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* [M]. Mahwah, N. J.: Lawrence Erlbaum Associates.

Elder, C., U. Knoch, G. Barkhuizen & J. von Randow. 2005. Individual feedback to enhance rater training: Does it work? [J]. *Language Assessment Quarterly* 2: 175-196.

Engelhard, G. 1992. The measurement of writing ability with a Many-Faceted Rasch Model [J]. *Applied Measurement in Education* 5: 171-191.

Knoch, U. 2011. Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study [J]. *Language Testing* 28: 179-200.

Lado, R. 1961. *Language Testing: The Construction and Use of Foreign Language Tests* [M]. London: Longman.

Linacre, J. 1999. Investigating rating scale category utility [J]. *Journal of Outcome Measurement* 3: 103-122.

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

Lumley, T. & T. McNamara. 1995. Rater characteristics and rater bias: Implications for training [J]. *Language Testing* 1: 54-71.

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

Oller, J. 1979. *Language Tests at Schools* [M]. London: Longman.

刘建达，2010，评卷人效应的多层面 Rasch 模型研究[J]，《现代外语》(2)：185-193。

王初明，2012，读后续写——提高外语学习效率的一种有效方法[J]，《外语界》(5)：1-7。

收稿日期：2013—06—03；修改稿，2013—07—25；本刊修订，2013—08—04

通讯地址：510420 广东省广州市 广东外语外贸大学外国语言学及应用语言学研究中心（王）
510420 广东省广州市 广东外语外贸大学英语语言文化学院（亓）

acquisition tempo and ultimate achievement of Chinese as a second language than the CPH and MCPH.

A study of the continuation task as a proficiency test component (p. 707)

WANG Chuming (Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou 510420, China)

QI Luxia (Faculty of English Language and Culture, Guangdong University of Foreign Studies, Guangzhou 510420, China)

This article reports on a study that sets out to examine if the continuation task which has been shown to facilitate L2 learning is also suitable for use in a L2 proficiency test. Data were collected from a sample of senior high school L2 learners of English and statistically analyzed following a Rasch model and other methods. Results show that the learners' performance of the continuation task correlated significantly with their performance of the reading and writing tasks used in a standardized English proficiency test, suggesting that the continuation task is a valid measure of the learners' reading and writing abilities. Furthermore, reliability of the task hinges on the design rigor of its scoring scheme, the training of raters and the proper use of the scheme. The scores yielded in accordance with the scheme could reliably discriminate the learners' L2 proficiency. The findings pave the way for more in-depth explorations into the continuation task for the testing purpose.

A study on the Chinese lexical attrition of advanced English learners in China (p. 719)

LIU Xueli & LIN Lihong (Faculty of Foreign Languages, Ningbo University, Ningbo 315211, China)

This study explores whether the Chinese lexicons of advanced English learners are affected by attrition or not. Thirty English major postgraduates participate in the study and thirty non-English major college freshmen serve as a control group. Chinese data are elicited through a written story-retelling task. Lexical diversity, lexical sophistication, lexical density and lexical errors are the variables investigated, using ICTCLAS, AntConc and SPSS. The results show that there is not a decrease in the first three aspects, but a significant increase in the lexical errors of the experiment group. The findings do not accord with those of previous research nor the prediction of this study completely. Finally, the findings are discussed from four aspects.

A theoretical framework for effective college English classroom environment construction and evaluation (p. 732)

REN Qingmei (School of Foreign Language, Qufu Normal University, Qufu 273165, China)

The paper proposes a theoretical framework for effective college English classroom environment construction and evaluation. The construction of the framework consists of the following parts. Firstly, based on the theory of effective teaching, research into classroom environment and characters of EFL classroom instruction and aligned with objectives of college English instruction in China, the concept of effective college English classroom environment construction is delineated. Secondly, enlightened by educational quality standards and instruments used in previous research, evaluation standards for effective college English classroom environment construction are set and explicated. Thirdly, initial revisions of the measurements are made according to feedback from experts and findings from in-service teacher interviews and lesson studies. Fourthly, the theoretical model is formulated and interactive relationships between the theoretical framework and