大学英语四级考试质量评估：基于经典测量理论和 Rasch 模型的数据分析

张 琳 陈琳丽
（上海交通大学，上海，200040）

摘要：大学英语四、六级考试近期在试卷结构和测试题型上再次进行了调整。本文运用经典测量理论和现代测量理论相结合的方法，基于考试数据对调整后的大学英语四级考试的试题质量进行了初步的评估。数据表明，试题总体符合考试质量要求，能够较准确地反映考生的水平，新题型设计得比较合理，考生对新题型总体比较适应。本文的数据分析结果初步论证了题型调整后四级考试的效度。

关键词：大学英语四级考试，考试质量评估，经典测量理论，Rasch 模型
［中图分类号］H319 ［文献标识码］A ［文章编号］1674-8921-(2015)10-0041-08
［doi 编码］10.3969/j.issn.1674-8921.2015.10.008

1. 引言

外语测试是外语教学的重要组成部分，是检查教学大纲执行情况、评定外语教学水平以及考核学生外语能力的一个重要手段。同时，外语测试还能对外语教学的内容和方法产生反拨作用(Bailey 1996；Hughes 1989)。大学英语四、六级考试作为评价在校大学生英语能力的主要手段之一，其科学性和公正性得到了社会的普遍认可。就考试的后效来看，四、六级考试对我国大学英语教学起了积极作用(王守仁 2011)，为我国大学英语教学质量的提高做出了巨大的贡献(吴启迪 2005)。为了满足社会发展的需求，更好地服务于大学英语教学，四、六级考试自 1987 年开始实施以来，在考试内容和形式上作过多次调整(Jin 2008，2011；Jin & Yang 2006)。自 2013 年 12 月起，四、六级考试在试卷结构和测试题型上再次进行了调整，旨在进一步提高考试的效度以及考试对教学的后效，更好地促进大学生英语综合应用能力的培养和提高。

对于一项涉及上千万考生的大规模高风险考试而言，确保考试的质量至为关键，因此有必要对题型调整后的考试质量进行评估，以检验新题型是否符合考试质量的要求，调整后的考试是否达到了设计者预期的效果。本文将采用经典测量理论和现代测量理论相结合的方法，基于考后数据对调整后四级考试的试题质量进行初步评估，从而对四级考试的效度作初步验证。

作者简介：张琳，上海交通大学外国语学院讲师。主要研究方向为语言测试。电子邮箱：zhang_lin@sjtu.edu.cn
陈琳丽，上海交通大学外国语学院讲师。主要研究方向为语言测试。电子邮箱：lynnchen@sjtu.edu.cn
* 衷心感谢金艳教授对本文的悉心指导。

2. 调整后的大学英语四级考试题型

Messick(1996)指出，应当通过改进考试设计来提高考试效度，为考试产生良好的后效打下基础。因此，四、六级考试委员会经过严格的科学论证，自 2013 年 12 月考次起对考试的内容和题型作了进一步调整，以期考试对大学英语的教学和学习产生更好的促进作用。

调整后的四级试卷由听力理解、阅读理解、翻译和写作三大部分组成，所占比例分别为：听力 35%，阅读 35%，翻译和写作 30%。调整后的四级试卷结构、测试内容、测试题型、分值比例和考试时间如表 1 所示：

表 1 现行的 CET-4 各部分测试内容、题型、所占比例和考试时间

试卷结构	测试内容		测试题型	比例	时间
听力理解	听力对话	短对话	多项选择	8%	30 分钟
		长对话	多项选择	7%	
	听力短文	短文理解	多项选择	10%	
		短文听写	单词及词组听写	10%	
阅读理解	词汇理解		选词填空	5%	40 分钟
	长篇阅读		匹配	10%	
	仔细阅读		多项选择	20%	
翻译和写作	汉译英		段落翻译	15%	30 分钟
	写作		短文写作	15%	

调整后的四级考试取消了多项选择题型的完形填空，另有三个题型作了局部调整：

(1) 单词及词组听写

原复合式听写调整为单词及词组听写，短文的长度、难度以及播放次数不变，所占分值比例不变。原复合式听写要求考生根据听到的短文内容填写空

缺的单词和句子，单词要求用听到的原文填写，句子可以在理解原文内容的基础上用自己的语言表述。此部分调整后，要求考生在听懂短文的基础上填写空缺的单词或词组，所有单词和词组均要求用听到的原文准确填写。

(2) 长篇阅读

原快速阅读理解调整为长篇阅读理解，篇章长度和难度不变，所占分值比例不变。原快速阅读理解要求考生阅读一篇较长篇幅的文章后作答 7 道多项选择题及 3 道句子填空题。现调整为篇章后附有 10 个句子，每句一题，每句所含的信息出自篇章的某一段落，要考生找出与每句所含信息相匹配的段落。

(3) 段落翻译

原单句汉译英调整为段落汉译英。原单句翻译共 5 句，每句一题，要求考生根据中文提示将每句的部分内容翻译成英语，使句子意思完整。调整后，要求考生将所给的中文段落全部译为英语。整个部分所占比重由原来的 5％上升到 15％，答题时间也由原来的 5 分钟增加至 30 分钟。翻译内容融入了中国元素，涉及中国的历史、文化、经济、社会发展等各个方面。为了保证翻译评分的信度，考试委员会制定了统一的翻译评分标准。评分采取整体印象法(holistic marking)，主要考虑意思表达的准确程度和语言的质量两个方面。满分 15 分，分六个档次，每个档次应达到的水平都有详细的文字描述。

另外，此次题型调整后，构建型作答试题(constructed response items)所占比重进一步增加，达到了整卷的 40％，从而更好地测试学生的语言综合应用能力。

3. 2013 年 12 月四级考试数据分析

调整后的四级考试于 2013 年 12 月首次实施，笔者从此次四级考试所采用的试卷中随机抽取一份试卷，并在作答所选四级试卷的考生总体中抽取了部分考生的数据进行分析。根据分层随机抽样原则抽取了 3427 名四级考生，所选的样本覆盖了全国不同地区不同层次的本科院校，因而是一个容量较大、代表性也比较好的样本数据。在样本数据的基础上，笔者既采用了经典试题分析方法对评价试题质量的主要指标(试题的难易度和区分度、试卷的内部相关等)进行考察，另外还运用了项目反应理论中的 Rasch 模型对试题的质量作进一步分析。

3.1 试卷总体难度与各部分难度

四、六级考试报道成绩时对原始分要进行等值处理，所以理论上试卷平均难度的高低对考生能力的测量不会产生影响，但难度过高或过低对教学的后效都较差，因而需要将试卷的难度控制在合理的范围内。就大规模标准化考试而言，整卷的平均难度在 0.6 左右是合适的(杨惠中、Weir 1998)。表 2 是对 3427 名四级考生所得原始分数的描述统计。其中，翻译和写作作为一个大的部分计算平均难度。

表 2 CET-4 各部分原始分数统计

	满分	均值	标准差	平均难度
Ⅰ 听力理解部分	35	20.22	7.19	0.58
1 短对话	8	5.22	2.00	0.65
2 长对话	7	4.42	1.72	0.63
3 听力篇章	10	6.18	2.35	0.62
4 短文听写	10	4.40	2.60	0.44
Ⅱ 阅读理解部分	35	24.94	7.05	0.71
5 选词填空	5	2.54	1.35	0.51
6 长篇阅读	10	7.99	2.22	0.80
7 仔细阅读	20	14.40	4.45	0.72
Ⅲ 翻译和写作	30	17.68	4.75	0.59
8 翻译	15	8.79	2.79	0.59
9 写作	15	8.89	2.61	0.59
整卷	100	62.84	16.98	0.63

从表 2 各个题型的数据可以看出，四级试卷中传统题型的难度大多在 0.6 左右，难度比较适中，仅选词填空题略难(0.51)，仔细阅读题略易(0.72)。从三个新题型的难度来看，段落翻译题的平均难度为 0.59，难度适中；单词和词组听写题偏难(0.44)，同时也是整个试卷最难的部分，但从历年考试数据来看，学生作答听写题的表现一直相对较弱，平均得分率不到 50％，因此调整后的听写题就其整体难度而言与以往大致相当，属正常水平；长篇阅读题难度较低(0.80)，但由于其他两个阅读题的难度相对较高，因而整个阅读理解部分的平均难度在合理范围内。

从试卷各个大的部分来看，听力理解部分的总体难度(0.58)与翻译和写作部分的总体难度(0.59)基本相当，难度适中。而阅读理解部分的整体难度为 0.71，相对略易。因此，整份试卷的平均难度为 0.63，表明四级试卷的难度是适中的。另外，数据还显示总分标准差较大，说明考生能力分布的离散程度较大，试卷能够很好地将考生能力区分开来。

3.2 试题项目分析

除了试卷的总体难度以外，试题项目分析也是考试质量评估的一个重要方面，即根据样本计算每道试题的难易度和区分度。试题的难易度和区分度是衡量试题质量的两个重要指标。对于大规模语言考试而言，一般要求试题的难易度在 0.3～0.7 之间，区分度在 0.2 以上(杨惠中、Weir 1998)。表 3 为四级试卷中全部客观题的难易度和区分度的双向汇总表：

表3 CET-4难易度和区分度双向汇总表

R\P	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	总计
<0.1											
0.1-0.2					14						1
0.2-0.3			25	39	61		49		1		5
0.3-0.4			43		6		2,13 17,24	12,22 40,45	46,47 54,56		14
0.4-0.5					36,38 41	7,8,11 21,44	4,9,10 19,53 57,62	3,20 50,63 65	16,48 58,64		24
0.5-0.6						18,23 42	59	5,55 60	15,51 52		10
0.6-0.7							37				1
0.7-0.8											
0.8-0.9											
0.9-1.0											
总计			2	1	6	8	14	12	12		55

注:R=区分度;P=难易度。

所有55道客观题中,1~25题为听力题,36~65题为阅读题。从各题数据可以看出,难易度在0.3~0.7之间的题目共29题,高于0.7的有24题,低于0.3的有2题,因而整卷全部客观题的平均难度为0.68。另外,就试题的区分度来看,在所有客观题中,仅有1题的区分度在0.2以下,说明四级题目有着很好的区分度,能将不同水平的考生区分开来。

3.3 试卷的内部相关性

根据Alderson等人(1995)的观点,对考试的各个组成部分进行相关分析可以提供考试构念效度方面的证据。一般来说,如果某两部分之间的相关系数太高,说明两者考核的是相同的能力,在考试设计上存在不必要的重复;如果两部分之间的相关系数太低,则说明两者考的是完全不同的能力。就语言测试而言,如果各个部分考核的是语言能力的不同侧面,那么它们之间的相关系数应在适中的水平,一般认为在0.3~0.7之间是合适的(杨惠中、Weir 1998)。但Alderson等人同时指出,由于总分是整体语言能力的体现,因此各组成部分与总分的相关应相对较高,以0.7左右或更高水平为佳。以下是对四级试卷所作的内部相关分析的结果。

3.3.1 各题型之间的相关

笔者首先分析了四级试卷各题型之间的相关以及各题型与总分的相关。表4是四级试卷各题型的相关系数矩阵。

表4 CET-4各题型的相关系数

	LC1	LC2	LC3	LC4	RD1	RD2	RD3	TR	WT	Total
LC1(8%)	1	.55	.61	.57	.56	.47	.53	.50	.47	.74
LC2(7%)		1	.59	.53	.49	.47	.48	.44	.42	.69
LC3(10%)			1	.63	.56	.50	.54	.49	.47	.77
LC4(10%)				1	.63	.52	.56	.57	.55	.80
RD1(5%)					1	.57	.62	.57	.55	.78
RD2(10%)						1	.63	.50	.48	.75
RD3(20%)							1	.53	.53	.83
TR(15%)								1	.55	.75
WT(15%)									1	.73
Total										1

注:LC=听力部分;LC1=听力短对话;LC2=听力长对话;LC3=听力篇章;LC4=短文听写;RD=阅读部分;RD1=选词填空;RD2=长篇阅读;RD3=仔细阅读;TR=翻译;WT=写作;括号内百分比=各题型所占分值比例。

就各题型之间的相关来看，表 4 中的数据显示听力篇章(LC3)与短文听写(LC4)、长篇阅读(RD2)与仔细阅读(RD3)的相关最高，相关系数达到了 0.63。听力篇章与短文听写同属听力理解部分，考核的均为听力方面的技能，仔细阅读与长篇阅读同属阅读理解部分，考核与阅读相关的能力，所以相关程度较高是合理的。另外，短文听写(LC4)与选词填空(RD1)的相关达到了 0.63，虽然两者分别属于听力题型和阅读题型，但实际上两者都在很大程度上考核了词汇的理解和运用能力，因此两者之间有较高的相关也是可以理解的。听力长对话(LC2)与写作(WT)的相关最低，但也达到了 0.42。其他相关系数大多在 0.4～0.6 之间，呈中等程度相关，说明各题型既具有一定的独立性，又存在相互联系，试卷设计得比较合理。

从表 4 还可以看出，各题型与总分的相关系数大多都在 0.7 以上，值得一提的是，仅占整卷 5%的选词填空题(RD1)与总分的相关也达到了 0.78。在各个题型中，仔细阅读(RD3)与总分的相关最高，相关系数达到了 0.83。由于仔细阅读部分所占分值比例最高，占到整卷的 20%，所以此部分与总分的相关最高是可以理解的。另外，短文听写(LC4)与总分的相关也达到了 0.8。短文听写要求学生既能听懂内容，还能正确书写，是对领会能力和表达能力的综合考核，能较好地反映出学生的整体语言水平，因此与总分相关较高也是可以理解的。与总分相关最低的是听力长对话(LC2)，但相关系数也达到了 0.69，且此部分仅占整卷的 7%，因此，与总分达到这样的相关程度也是比较理想的。

以上是对各题型相关数据的总体描述，下面将着重分析三个新题型的相关数据：单词及词组听写(LC4)与其他各题型之间的相关在 0.52～0.63 之间，长篇阅读(RD2)与其他题型的相关在 0.47～0.63 之间，段落翻译(TR)与其他题型的相关在 0.44～0.57 之间，说明这三个新题型与其他各题型之间呈中度相关；另外，三个新题型与总分的相关都很高，长篇阅读和翻译与总分的相关均达到了 0.75，听写与总分的相关则更高，达到了 0.8。各项数据表明，三个新题型的设计均比较合理，符合测试意图。

3.3.2 各部分之间的相关

笔者进一步分析了四级试卷各个部分之间的相关以及各部分与总分的相关。表 5 是四级试卷各部分的相关系数矩阵，其中翻译和写作仍作为一个整体进行分析。

从表 5 的数据来看，在试卷各个部分中，听力(LC)和阅读(RD)的相关最高，相关系数为 0.72，其他各部分之间的相关系数均为 0.67。虽然各部分的相关系数在 0.7 左右，处于较高水平，但仍在合理范围之内，表明试卷各部分既考核了语言能力的不同方面，同时又是相互关联的，各个部分的综合能够准确而有效地反映学生的总体语言水平。另外，各部分与总分之间的相关都很高，其中听力部分和阅读部分与总分的相关均高达 0.91，而翻译和写作部分与总分的相关也达到了 0.8 以上。

表 5 CET-4 各部分的相关系数

	LC	RD	TR & WT	Total
LC	1	0.72	0.67	0.91
RD		1	0.67	0.91
TR & WT			1	0.84
Total				1

3.4 Rasch 模型分析

Rasch 模型是一种单参数项目反应理论模型，因其克服了传统测量理论的局限之处，实现了测量的客观等距目标，为社会科学领域内的测量建立了一套客观标准，现已广泛应用于教育、心理学、医学等诸多领域。近年来，Rasch 模型也越来越多地应用于语言测试领域，国内外有不少学者运用 Rasch 模型对测试的信效度进行了研究(如 Eckes 2005；Bonk & Ockey 2003；刘建达 2005；江进林、文秋芳 2010)。本文尝试使用 Rasch 模型从另一个视角对试题的质量进行了分析。笔者采用 Rasch 分析软件 FACETS 3.58 对 3427 名四级考生除翻译和作文之外的全部客观题(即听力部分和阅读部分)的作答结果进行了分析。听写题尽管采用的是主观评分，但采用的计分方式与客观题相同，即只有正确和错误两种作答结果，满足 Rasch 模型分析的基本要求，因此对听写题的数据也进行了 Rasch 分析。以下是对考试整体情况的分析结果。

3.4.1 试题难度与考生能力的对应关系

Rasch 模型将试题难度和考生能力都转化为以 logit 为单位的统一度量值，并将两者在共同的标尺上进行度量，因而可以直接比较考生与考生、考生与试题、试题与试题的差异。图 1 直观地展现了试题难度与考生能力之间的对应关系。

图 1 左边一列是 logit 量尺，是后面两列参照的共同标准。中间一列呈现的是考生能力的分布情况，每个 * 代表 35 名考生，每个圆点表示不足 35 名考生，考生能力从下往上依次递增。右边一列呈现的是 65 道试题难度水平的分布，题目难度自下而上依次增加。听力题的编号为 1～35，阅读题的编号为 36～65，其中 26～35 题为采用 0/1 计分的听写题，

其余全部为客观题。考生间的距离代表考生能力水平的差异,试题间的距离代表试题难度水平的差异。考生水平分布越分散,说明考生水平差距越大,题目对考生的区分能力强;反之,分布较集中,说明考生水平差距不明显,题目的区分能力弱。另外,理想的试题分布状况是题目能够覆盖所有水平的考生,而且在考生水平分布相对密集处,题目数量相应较多;试题难度水平与考生能力水平越接近,对考生能力水平的估计越精确。

Measr	+student	–items
5		
4	.	
	.	34
	*.	
3	*.	
	*.	
	*.	
	****.	34
2	*****.	25 35
	*****.	30 39
	******.	33
	********.	14 26
1	*****.	6 36 38 41 61
	*******.	23 44
	*********.	7 8 18 29 32
	******.	10 11 13 21 24 42
0	*****.	27 31 49 57
	*****.	4 9 17 19 28 37 53 59
	******.	2 3 12 22 45 60 62 65
	****.	5 20 40 50 63
–1	**.	48 55 58
	**.	1 15 51 56
	*.	16 46 47 54 64
	.	52
–2	.	
	.	
	.	
–3	.	
	.	
–4		
Measr	*=35	–items

图1 试题难度与考生能力对应图

从图1可以看出,考生能力基本呈正态分布,且分布较分散。试题的难度覆盖了绝大多数考生的语言能力水平,分布比较均匀,考生的水平与试题分布基本匹配,说明试卷可以对考生的能力水平做出比较精确的估计。同时,图1也清晰地呈现了试题难度的顺序,其中52题最简单,34题最难。从图1可以看出,仅有1题与其他试题相距较远,此题为34题,是一道单词/词组听写题,在65道试题中难度最高,与其他题目的难度水平差异较大。绝大多数试题集中分布在±2个logit范围内,总体上试题的难度分布是合理的。

3.4.2 试题分析结果

Rasch模型对试题的难度和考生的能力进行估计后,对每个考生在每道试题上答对的理论概率进行估算,并与实际的观测分数进行比较,用两者之间的差异来评估数据与模型的拟合情况。图2是65道试题的拟合分析结果,按照试题难度的度量值由高到低排列。

Rasch模型通常报告Infit MnSq和Outfit MnSq两个拟合统计量,前者是加权均方拟合统计量,后者是未加权均方拟合统计量。由于后者更容易受到个体差异大的数据的影响,因此一般以前者作为判断个体是否拟合模型的依据。拟合统计量的值为1,表示数据与模型预测完全符合。对于Infit MnSq的取值范围没有严格规定,鉴于此处分析的大多为选择题,因而采用较严格的拟合控制,Infit MnSq值在0.7～1.3之间认为数据与模型拟合较好(Wright & Linacre 1994)。若试题的Infit MnSq值大于1.3,视为非拟合题目,表明考生的作答方式与模型设定的不一致;小于0.7,则视为过度拟合题目,表明考生的作答结果差异较小或题目不能区分考生之间水平的差异。Rasch标准误(Model S.E.)表示试题测量考生能力的误差大小,误差越小表示对考生能力的估计越精确,题目的信度越高,一般认为0.03～0.05是可接受的范围(Green 2013)。相关系数(Corr. PtBis)表示试题与其测量目标的拟合程度,相关系数越高,说明题目与其测量目标越接近。

图2的数据显示,Infit MnSq值基本都在可接受范围内,而且大多数非常接近于期望值1,仅1题(即听力部分的14题)的Infit MnSq值为1.31,略大于1.3,处于非拟合的边缘。因此,试题数据总体而言与Rasch模型拟合较好。绝大多数试题的Rasch标准误都在可接受的水平,仅最难的34题和最简单的52题两道试题相应的Rasch标准误略大于其他题目,因而整体来看误差较小,说明试题对考生能力的估计比较准确,试题的信度较高。相关系数都处于可接受水平,表明所有题目与测量目标之间有较好的一致性。

除此之外,图2最下方的分隔系数(Separation)和分隔信度(Reliability)用以衡量个体之间存在差异的程度,数值越大说明越有把握认为个体之间存在明显差异。信度的取值范围为0～1,越接近于1表明差异越大。卡方检验(Fixed chi-square)旨在检验个体之间是否具有统计学意义上的显著差异。图2的数据显示,分隔系数为24.33,信度达到了1.00,卡方值为33023.2(d.f.=64),显著性为0.00,这些都表明各题目之间的难度有显著意义的区别,符合试卷设计的要求。

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S. E	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. discrm	Corr. ptBis	Nu	items
301	3427	0.09	0.05	3.57	0.06	0.93	−1.5	0.76	−2.4	1.06	0.29	34	34
782	3427	0.23	0.18	2.25	0.04	0.99	−0.6	0.93	−1.4	1.06	0.37	43	43
914	3427	0.27	0.22	2.00	0.04	1.11	5.1	1.44	9.0	0.76	0.24	25	25
923	3427	0.27	0.22	1.98	0.04	0.95	−2.6	0.98	−0.4	1.07	0.40	35	35
1114	3427	0.33	0.28	1.65	0.04	1.19	9.0	1.29	8.0	0.62	0.26	39	39
1123	3427	0.33	0.29	1.63	0.04	0.94	−3.3	0.86	−4.4	1.14	0.46	30	30
1199	3427	0.35	0.31	1.51	0.04	0.94	−3.6	0.87	−4.4	1.15	0.46	33	33
1365	3427	0.40	0.37	1.25	0.04	1.31	9.0	1.53	9.0	0.22	0.15	14	14
1377	3427	0.40	0.38	1.23	0.04	0.91	−5.5	0.87	−5.2	1.21	0.49	26	26
1463	3427	0.43	0.41	1.10	0.04	0.92	−5.1	0.89	−4.7	1.20	0.49	36	36
1524	3427	0.44	0.43	1.00	0.04	0.89	−7.4	0.86	−6.0	1.28	0.52	38	38
1548	3427	0.45	0.44	0.97	0.04	0.92	−5.3	0.92	−3.5	1.20	0.49	41	41
1548	3427	0.45	0.44	0.97	0.04	1.23	9.0	1.39	9.0	0.38	0.23	61	61
1595	3427	0.47	0.46	0.90	0.04	1.17	9.0	1.22	8.8	0.57	0.29	6	6
1766	3427	0.52	0.52	0.64	0.04	0.93	−4.5	0.92	−3.3	1.17	0.49	23	23
1769	3427	0.52	0.52	0.64	0.04	0.98	−1.4	0.95	−2.52	1.07	0.45	44	44
1801	3427	0.53	0.53	0.59	0.04	0.95	−3.4	0.93	−2.8	1.14	0.48	18	18
1835	3427	0.54	0.55	0.54	0.04	1.00	−0.1	0.99	−0.2	1.01	0.43	8	8
1857	3427	0.54	0.55	0.51	0.04	1.04	2.5	1.04	1.8	0.90	0.40	7	7
1909	3427	0.56	0.57	0.43	0.04	1.03	2.0	1.05	2.2	0.91	0.41	29	29
1912	3427	0.56	0.57	0.42	0.04	0.92	−5.4	0.87	−5.5	1.22	0.51	32	32
1974	3427	0.58	0.60	0.33	0.04	0.98	−1.4	0.96	−1.6	1.06	0.45	21	21
2013	3427	0.59	0.61	0.27	0.04	0.87	−8.3	0.81	−7.7	1.32	0.54	42	42
2033	3427	0.59	0.62	0.24	0.04	1.07	4.2	1.07	2.7	0.84	0.37	13	13
2037	3427	0.59	0.62	0.23	0.04	1.08	4.6	1.11	4.2	0.81	0.36	11	11
2084	3427	0.61	0.64	0.16	0.04	1.02	1.2	1.07	2.5	0.94	0.41	10	10
2100	3427	0.61	0.64	0.13	0.04	1.09	5.3	1.16	5.6	0.77	0.35	24	24
2128	3427	0.62	0.65	0.09	0.04	0.84	−9.0	0.79	−7.9	1.35	0.57	27	27
2150	3427	0.63	0.66	0.06	0.04	1.03	1.4	1.00	0.1	0.96	0.41	31	31
2203	3427	0.64	0.66	−0.03	0.04	1.00	0.2	1.03	0.9	0.98	0.42	57	57
2255	3427	0.66	0.70	−0.11	0.04	1.23	9.0	1.35	9.0	0.51	0.22	49	49
2266	3427	0.66	0.70	−0.13	0.04	1.04	2.2	1.01	0.3	0.94	0.39	53	53
2281	3427	0.67	0.71	−0.16	0.04	0.79	−9.0	0.69	−9.0	1.43	0.61	37	37
2324	3427	0.68	0.72	−0.23	0.04	1.15	7.6	1.15	4.3	0.73	0.29	17	17
2329	3427	0.68	0.72	−0.24	0.04	1.03	1.7	1.09	2.7	0.92	0.39	19	19
2348	3427	0.69	0.73	−0.27	0.04	0.98	−0.8	0.96	−1.0	1.03	0.43	4	4
2369	3427	0.69	0.74	−0.30	0.04	0.94	−3.0	0.86	−4.4	1.13	0.47	59	59
2396	3427	0.70	0.74	−0.35	0.04	1.04	1.9	1.02	0.7	0.94	0.38	9	9
2407	3427	0.70	0.75	−0.37	0.04	1.00	−0.1	1.12	3.3	0.97	0.41	28	28
2415	3427	0.70	0.75	−0.38	0.04	1.03	1.5	0.99	−0.2	0.96	0.39	22	22
2432	3427	0.71	0.76	−0.41	0.04	1.10	4.6	1.14	3.6	0.83	0.33	62	62
2452	3427	0.72	0.76	−0.45	0.04	1.04	2.1	0.99	−0.1	0.94	0.38	2	2
2464	3427	0.72	0.77	−0.47	0.04	1.12	5.6	1.24	5.8	0.78	0.30	12	12
2469	3427	0.72	0.77	−0.48	0.04	0.93	−3.3	0.83	−4.6	1.13	0.47	60	60
2494	3427	0.73	0.78	−0.52	0.04	0.96	−1.8	0.90	−2.5	1.07	0.44	3	3
2495	3427	0.73	0.78	−0.53	0.04	1.06	2.7	1.15	3.7	0.89	0.35	65	65
2522	3427	0.74	0.79	−0.58	0.04	1.03	1.3	0.99	−0.2	0.97	0.38	45	45
2612	3427	0.76	0.81	−0.75	0.04	0.95	−2.2	0.90	−2.3	1.08	0.44	40	40
2619	3427	0.76	0.81	−0.76	0.04	0.86	−6.4	0.78	−5.1	1.20	0.52	5	5
2655	3427	0.77	0.83	−0.83	0.05	0.93	−3.1	0.88	−2.6	1.10	0.45	50	50
2664	3427	0.78	0.83	−0.85	0.05	1.02	0.7	1.04	0.8	0.98	0.37	20	20
2673	3427	0.78	0.83	−0.87	0.05	0.99	−0.2	0.92	−1.7	1.03	0.40	63	63
2696	3427	0.79	0.84	−0.92	0.05	0.87	−5.4	0.75	−5.4	1.17	0.50	55	55
2781	3427	0.81	0.86	−1.10	0.05	0.95	−2.0	0.88	−2.1	1.07	0.41	48	48
2783	3427	0.81	0.86	−1.11	0.05	0.97	−0.9	1.01	0.1	1.02	0.38	58	58
2809	3427	0.82	0.87	−1.17	0.05	1.09	3.3	1.19	3.1	0.89	0.27	1	1
2823	3427	0.82	0.87	−1.20	0.05	1.06	1.9	1.30	4.8	0.90	0.28	56	56
2847	3427	0.83	0.88	−1.26	0.05	0.86	−4.9	0.75	−4.6	1.16	0.48	15	15
2871	3427	0.84	0.89	−1.32	0.05	0.81	−6.7	0.63	−6.8	1.21	0.52	51	51
2899	3427	0.85	0.89	−1.39	0.05	0.98	−0.5	1.07	1.0	1.00	0.35	16	16
2929	3427	0.85	0.90	−1.47	0.05	0.95	−1.4	0.88	−1.8	1.05	0.38	46	46
2929	3427	0.85	0.90	−1.47	0.05	0.91	−2.9	0.74	−4.2	1.10	0.43	64	64
2948	3427	0.86	0.90	−1.52	0.05	0.99	−0.3	0.97	−0.4	1.01	0.34	54	54
2961	3427	0.86	0.91	−1.56	0.05	0.98	−0.5	0.98	−0.2	1.01	0.34	47	47
3005	3427	0.88	0.92	−1.69	0.06	0.87	−3.6	0.69	−4.5	1.13	0.44	52	52
2147.2	3427.0	0.63	0.65	0.00	0.04	1.00	−0.3	0.99	−0.3		0.40	Mean (Count:65)	
612.9	0.0	0.18	0.21	1.06	0.01	0.10	4.6	0.18	4.5		0.09	S. D. (Population)	
617.7	0.0	0.18	0.21	1.07	0.01	0.10	4.6	0.18	4.5		0.09	S. D. (Sample)	

Model. Pouln: RMSE .04 Adj(True) S. D. 1.06 Separation 24.33 Strata 32.78 Reliability 1.00
Model. Sample: RMSE .04 Adj(True)S. D. 1.07 Separation 24.52 Strata 33.03 Reliability 1.00
Model. Fixed (all same) chi-square: 33023.2 d. f. : 64 significance (probability): .00
Model. Random (normal) chi-square: 63.9 d. f. : significance (probability): .45

图 2 试题分析结果

3.4.3 考生能力分析结果

鉴于考生人数较多,这里仅报告整体的考生能力情况,不再一一罗列个体的数据。表6显示了考生整体情况的分析结果。

表6 考生能力整体情况

Fit	Infit MnSq	0.7～1.3	98.8%
		0.60～0.69	0.1%
		1.3-1.4	1.1%
Separation	Separation	3.42	
	Reliability	0.92	
Chi-square test	Fixed (all same) chi-square	37149.0 (d.f.=3426)	
	Significance	0.00	

从表6中总结的考生拟合数据的取值范围及其所占考生的百分比来看,仅1.2%的考生的Infit MnSq值略超出可接受范围,一般来说非拟合考生的比例应控制在2%左右(Pollitt & Hutchinson 1987),因而考生的答题行为整体上符合Rasch模型的预期。这里的分隔系数为3.42,分隔信度为0.92,表明考生能力具有很大差异。这种差异是否显著可以通过卡方检验进行验证。卡方值为37149.0(d.f.=3426),显著性为0.00,结果显示考生能力的差异具有统计上的显著意义,表明试题具有较好的区分度,能够区分出不同考生的能力。

除了了解考试的整体情况以外,笔者也对听力部分和阅读部分分别作了Rasch分析,受篇幅所限不再细述,总体而言每个部分的试题质量都比较理想,试题难度与考生能力匹配得较好,试题能够准确地反映考生的水平,符合考试的质量要求。

4. 结语

为了验证调整后四级考试的效度,本文以3427名抽样考生的答题数据为基础,对四级试题的质量进行了初步分析。本文首先采用传统试题分析方法考察了试题难易度、区分度以及试卷内部相关等衡量试题质量的主要指标。从初步的数据分析结果来看,题型调整后的四级试卷总体难度适中,除了各个传统题型的难度总体保持稳定以外,单词及词组听写、长篇阅读和段落翻译这三个新题型的难度也处在比较合理的水平。尤其值得一提的是,段落翻译题属主观性试题,且占到了整卷的15%,而数据显示其平均得分率达到了60%左右,表明考试设计者在命题过程中对此部分难度进行了较好的控制。从四级学生的答题情况来看,学生对各个新题型总体比较适应,在新题型上的表现整体比较理想。但是,在测试一定程度表达能力的听写题上,学生的表现仍差强人意,平均得分率仍然不到50%。另外,根据对试卷中全部客观题所做的试题项目分析结果,四级试题的难易度和区分度分布总体符合考试质量要求。对试卷所作的内部相关分析的结果显示,各题型之间呈中等程度相关,说明各题型既互相独立又存在关联,整份试卷设计得比较合理。同时,各题型与总分之间大多呈现高相关,达到了比较理想的相关水平。从新题型的相关数据来看,三个新题型与其他题型之间的相关总体上比较适中,三个新题型与总分的相关也比较理想,表明新题型设计合理,基本符合考试设计者的意图。

此外,本文还对试卷中的全部客观题及采用0/1计分的听写题进行了Rasch分析。结果显示,试题的难度水平总体上与考生的能力水平相匹配,试题覆盖了绝大多数考生的能力水平,能够对考生的能力做出比较准确的估计。同时,绝大多数试题集中分布在±2个logit范围内,试题难度的分布是比较合理的。就试题数据和考生能力数据与Rasch模型的拟合分析结果来看,加权均方拟合统计量的取值绝大多数都在可接受的范围内,表明数据与模型的拟合比较理想。各个题目的Rasch标准误和相关系数也都在可接受的水平,表明试题对考生的能力水平进行估计时误差较小,试题能够较好地测量出所要测量的目标。此外,数据还显示试题具有良好的区分度,能够将不同考生的能力区分开来。Rasch分析结果进一步表明四级试题的难易度和区分度分布比较理想,试题质量符合考试的要求。

此外,教师、学生及媒体对此次考试题型调整也普遍反映良好。从考试委员会对部分教师进行的考后访谈结果来看,教师对题型调整给予了充分的肯定,认为调整后的考试更综合地测试学生的英语应用能力。教师们还一致认为翻译题的调整是此次题型调整的最大亮点:首先,调整后采用的段落翻译题型可以更有效地测试学生的翻译技能,能够对翻译教学产生良好的后效;其次,翻译题融入的中国元素有助于增加学生对中国的历史、文化、经济和社会发展等各方面的了解,从而提高学生的跨文化交际能力。对部分考生进行的考后访谈结果显示,考生总体上也持肯定态度。不少考生指出调整后的翻译题更具真实性,对学生的能力提出了更高的要求,但同时也能更好地反映出学生的语言综合运用能力。国内有不少主流媒体也关注了此次的题型调整,并给予了正面报道。如有报道指出,多项选择题的减少和主观性试题的进一步增加使考试能够更好地测试大学生的英语实际应用能力,从而引导师生更加重视语言实际运用能力的培养。

对考试数据进行科学地分析和评价是考试质量评估的重要组成部分,数据分析和评价的结果可以为

(下转66页)

Brinkley, W. 1988. *The Last Ship* [M]. New York: Viking.
Burns, J. J. 2009. *Alternate Endings: The Post-apocalypse in American and British Film and Literature* [D]. University Of Arkansas.
Curtis, C. 2010. *Postapocalyptic Fiction and the Social Contract: "We'll Not Go Home Again"* [M]. Lanham: Lexington Books.
DuPrau, J. 2003. *The City of Ember* [M]. New York: Random House.
Forstchen, W. R. 2009. *One Second After* [M]. New York: Forge.
Frank, P. 1959. *Alas, Babylon* [M]. New York: Harper Perennial.
Grossman, J. R. 2011. *Keeping the Lights on: Post-apocalyptic Narrative, Social Critique, and the Cultural Politics of Emotion* [D]. Colorado State University.
Hegland, J. 1996. *Into the Forest* [M]. New York: Bantam Books.
King, S. 1990. *The Stand* [M]. New York: Doubleday.
LaFaille, G. 1992. Review of *On the Beach* [J]. *Wilson Library Bulletin* (10): 131.
McCarthy, C. 2006. *The Road* [M]. New York: Vintage.
Matheson, R. 1971. *I Am Legend* [M]. New York: Berkley.
Miller, W. M. 1960. *A Canticle for Leibowitz* [M]. New York: J. B. Lippincott.
Niven, L. & J. Pournelle. 1977. *Lucifer's Hammer* [M]. Chicago: Playboy Press.
Rawles, J. W. 2009. *Patriots: A Novel of Survival in the Coming Collapse* [M]. Berkeley: Ulysses.
Shute, N. 1957. *On the Beach* [M]. New York: William Morrow.
Stewart, G. R. 2006. *Earth Abides* [M]. New York: Ballantine Books.
Swartz, Z. C. 2009. *Ever Is No Time at All: Theological Issues in Post-apocalyptic Fiction and Cormac McCarthy's* The Road [D]. Georgetown University.
Wagar, W. W. 1982. *Terminal Visions: The Literature of Last Things* [M]. Bloomington: Indiana University Press.
Wolfe, G. 1983. The remaking of zero: Beginning at the end [A]. In E. S. Rabkin, M. H. Greenberg & J. D. Olander (eds.). *The End of the World* [C]. Carbondale & Edwardsville: Southern Illinois University Press. 1-19.
葛红兵、肖青峰. 2008. 小说类型理论与批评实践——小说类型学研究论纲[J]. 上海大学学报(社会科学版)(5):63-74.
威廉·福岑. 2012. 一秒之后(符瑶译)[M]. 北京:新星出版社.

(责任编辑 玄 琰)

(上接 47 页)

设计者提供考试质量方面的重要信息,为进一步改进考试提供重要依据。本文通过对考试数据的分析对此次题型调整后的四级考试作了初步的质量评估,从而初步论证了调整后四级考试的效度。然而,考试分数只是开展效度研究过程中所需收集证据的其中一个方面,今后还需要不断收集其他各方面证据,以更全面地论证考试的效度,从而及时地发现考试可能在某方面存在的局限,不断地改进和完善考试,更好地为教学服务。

参考文献

Alderson, J. C., C. Clapham & D. Wall. 1995. *Language Test Construction and Evaluation* [M]. Cambridge: Cambridge University Press.
Bailey, K. M. 1996. Working for washback: A review of the washback concept in language testing [J]. *Language Testing* 13(3): 257-79.
Bonk, W. J. & G. J. Ockey. 2003. A many-facet Rasch analysis of the second language group oral discussion task [J]. *Language Testing* 20(1): 89-110.
Eckes, T. 2005. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis [J]. *Language Assessment Quarterly* 2 (3): 197-221.
Green, R. 2013. *Statistical Analyses for Language Testers* [M]. London: Palgrave Macmillan.
Hughes, A. 1989. *Testing for Language Teachers* [M]. Cambridge: Cambridge University Press.
Jin, Y. 2008. Powerful tests, powerless test designers? —Challenges facing the College English Test [J]. *English Language Teaching in China* 31(5): 3-11.
Jin, Y. 2011. Fundamental concerns in high-stakes language testing: The case of the College English Test [J]. *Journal of Pan-Pacific Association of Applied Linguistics* 15(2): 71-83.
Jin, Y. & H. Yang. 2006. The English proficiency of college and university students in China: As reflected in the CET [J]. *Language, Culture and Curriculum* 19 (1): 21-36.
Messick, S. 1996. Validity and washback in language testing [J]. *Language Testing* 13(3): 241-56.
Pollitt, A. & C. Hutchinson. 1987. Calibrated graded assessments: Rasch partial credit analysis of performance in writing [J]. *Language Testing* 4(1): 72-92.
Wright, B. D. & J. M. Linacre. 1994. Reasonable mean-square fit values [J]. *Rasch Measurement Transactions* 8 (3): 370.
江进林、文秋芳. 2010. 基于 Rasch 模型的翻译测试效度研究[J]. 外语电化教学(1):14-18.
刘建达. 2005. 话语填充测试方法的多层面 Rasch 模型分析[J]. 现代外语(2):157-69.
王守仁. 2011. 关于高校大学英语教学的几点思考[J]. 外语教学理论与实践(1):1-5.
吴启迪. 2005. 教育部 2005 年第 2 次新闻发布会:介绍大学英语四、六级考试改革有关情况[EB/OL]. [2005-2-25]. http://www.moe.edu.cn/edoas/website18/info8745.htm.
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.

(责任编辑 杨 丽)