DOI: 10. 16482/j. sdwy37 - 1026. 2018 - 03 - 003

基于多面 Rasch 模型的商务英语口语测试评分研究

揭薇
(上海交通大学 外国语学院，上海 200240 /
上海对外经贸大学 国际商务外语学院，上海 201600)

[摘要] 信度和效度研究较多考察通用英语测试，而对专门用途英语测试的效度微观研究则相对缺失。本文应用 FACETS 软件对某高校 VECTOR 商务英语主观题型测试进行分析，从微观上对评分者严厉程度、考生能力差异、试题难易度差异进行进一步分析。结果显示，与其他通用英语主观题型测试的经验性结论相比，商务英语话题任务难度差异大，评卷者严厉度存在显著差异。本研究的结论有助于解释商务英语口语测试效度，并且利用该模型对评分者进行培训，从而改善考试设计、控制评分质量，改进评分标准和提高考试效度，这对相关教学和测试意义重大。

[关键词] 商务英语口语测试; 评分员; 多面 Rasch 模型; FACETS
[中图分类号] H04 [文献标识码] A [文献编号] 1002-2643(2018)03-0022-12

A Study of the Scoring of Business English Oral Test Based on Many-facet Rasch Model

JIE Wei
(School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China /
School of Languages, Shanghai University of International Business and Economics, Shanghai 201600, China)

Abstract: Reliability and validity research mostly examined general English test, while the validity of the ESP test was relatively absent. In this research, the FACETS software is used to analyze the oral test of VECTOR Business English in a university. The severity of the score rater, the difference of examinee's ability and test item difficulty are further analyzed from a micro perspective. The results show that, compared with the empirical conclusions of other general English speaking test studies, there are great differences in the difficulty of the topic of business English oral test items and there are significant differences in the severity of raters. The results of this study are helpful to explain the validity of business English oral test and train raters, so as to improve the test design, control the quality of the scoring, expand scoring standard and test validity, which is of great significance to the business English teaching and testing practice.

Key words: Business English Oral Tests; Raters; Many-facet Rasch model; FACETS

收稿日期: 2018 - 03 - 24

作者简介: 揭薇，博士研究生，副教授，硕士生导师。研究方向: 语言测试。**电子邮箱:** neoweism@ aliyun. com。

1.0 引言

商务英语是专门用途英语的一个重要分支，商务英语有其自身的显著特点，体现在语言能力、专业知识、文体风格、外部语境等各个方面（对外经济贸易大学商务英语理论研究小组，2006）。针对专门用途英语的测试的内容和方法来源于分析特定的目标语言使用情况，测试任务和测试内容的设计强调真实性（authenticity），包括情景真实性（situational authenticity）和交互真实性（interactional authenticity）（Douglas，2000; Douglas，2001）。商务英语测试比其他任何类型的英语测试更强调展示情景的真实性以及测评考生的实际的运用和操作能力（O'Sullivan，2006）。因此，商务英语测试研究应该结合传统语言测试理论和方法，同时要兼顾商务英语自身特点，设计出符合学习者完成目标环境下的真实任务的测试内容和方法，并且根据完成任务的程度和效果来进行评分。

强调运用的语言测试主要测量语言使用者的实际应用能力而被广泛的使用于语言测试实践中。但是这种形式的测试过程因为引入了评分者、评分标准以及任务形式等因素从而使得到的分数往往更多的受到学生能力以外的因素的影响，从而对于确保测试的信度和效度有了更多的难度和要求。口语考试作为一种输出性的考试，对学生的语言能力进行直接的测量，如果设计合理，评分客观准确，能够达到高效度。随着经济全球化的步伐加快，商务英语口语能力是社会实践中非常重要而且很受欢迎的能力，在人才招聘和商务交流中是一项重要的能力指标，但是其口语考试的评分过程往往因为评分员专业知识的影响而带有主观性，要保证评分的准确性和一致性是一大难题。

2.0 文献回顾

多面 Rasch 模型是单参数 Rasch 模型的拓展（Linacre，1989; 1994），通过对测试过程中的每个层面参数化，同时假定这些层面会共同作用从而影响考生得到某个分值的概率。基于随机概率模型，多面 Rasch 模型将不同层面中的每个个体（学生、评分者、试题等）在共同的 logit 标尺上进行度量，并计算每个度量值的估算误差、对模型的拟合程度以及每个层面之间可能的交互作用。由此可见，利用多层面 Rasch 模型分析测试结果，特别是对口语、写作、翻译等考试中，考生成绩容易受到多个方面影响的试题类型，具有非常大的优势。多层面 Rasch 模型可以将这些影响因素的程度参数化，并用数值的形式体现，这样有助于我们最大程度的减少考试其他方面对于学生能力的影响，更真实的表现学生的能力水平。

口语考试是国内外大规模考试的重要组成部分，口语考试的评分是考试构念的体现，是考试信度和效度的重要保证（Fulcher，2003）。对口语考试的评分研究一直是各种口语考试的研究重点。多面 Rasch 模型在表现性评价（performance assessment）研究中具有诸

多优势，国外已有广泛的应用，如：分析同伴讨论口语测试任务的评卷人效应（Bonk & Ockey，2003）；写作评分的评卷人效应（Eckes，2008）；母语评分员在评阅二语学习者英文短文写作的评分员偏差分析（Schaefer，2008）；诊断性写作测试评分量表的开发与验证（Knoch，2009）；对现有特殊用途英语口语测试评分标准的扩展研究（Hagan et. al.，2015）；写作评分员评分行为的比较研究（Goodwin，2016）等。近年来，国内应用多面 Rasch 模型的口语考试研究也在逐年增加，研究侧重点主要集中在两个方面，一是应用多面 Rasch 模型对口语考试的效度进行验证以及对考试评分标准的质量检验（刘建达，2005；张洁，2008，2016；何莲珍、张洁，2008；赵南、董燕萍，2013；范劲松、季佩英，2015；高淼，2016）。另一方面是应用多面 Rasch 模型对口语考试的评分进行具体分析。刘建达（2010）分析了口语评卷人效应，戴朝晖、尤其达（2010）分析了大学生计算机口语考试评分者的评分偏差；李英、关丹丹（2016）对 PETS 口试教师评分的培训效果进行了分析，发现多面 Rasch 模型有助于发现评分异常情况，开展有针对性的评分培训。周燕、曾用强（2016）对比分析了听说考试中计算机自动评分和评分员评分的差异性。

这些研究都展现了多面 Rasch 模型在表现性评价中的各种应用，但是目前鲜有研究应用多面 Rasch 模型对商务英语口语考试进行评分研究。因此，本研究基于某高校的一次 VECTOR 商务英语会话考试的学生实际成绩，在多层面 Rasch 模型框架下对评分者、学生、试题等进行了探讨，旨在使用这种统计方法有效地研究商务英语测试中各个层面因素对于学生成绩的影响，从而使考试公平公正的反应学生的真实能力水平。

3.0 研究问题及研究设计

3.1 问题提出

本研究运用多面 Rasch 模型对某大学的一次商务英语口语测试的评分进行分析，具体回答以下问题：1）评分者的严厉程度是否一致，评分者的评分是否存在内在一致性？2）题项是否能够很好地区分考生能力？3）评分质量如何，是否存在显著偏差？

3.2 Rasch 模型设定

Rasch 模型作为 IRT 理论的主要模型之一，其基本想法是某个考生答对某道题的概率大小不仅取决于考生自身的能力，也取决于这道题目的难度。Rasch 模型的基本形式除了可以用来分析二分计分数据，其拓展形式（Andrich，1978；Wright & Masters，1982）还可以用来计算评分量表（Rating Scale）中的分步难度以及分析具有分部计分（Partial credit）的评分系统。如果将任务（或题项）由难度大小顺序从低往高排列，那么被试的能力大小应该与其通过的任务（或答对的题项）成正相关（朱正才等，2003；赵守盈、薛雯，2011），其模型的假设与语言测试理论评估试题质量的依据是一致的（刘建达，2010）。Rasch 模型估计方法有以下优势：估计得分能够有效反映潜在特征。受试个体得分可以通过计算其各个测试项目的总体反映得到。所有拥有相同项目得分的受试者具有相同的隐含特

征。考生在各题上的总分是个充分统计量，即考生能力参数的估计只与总分（即答对题目数）有关，而与具体的应答模式无关。考生与题目在模型中的地位的对称性，在 Rasch 模型下，可以同时求得考生能力与题目难度的估计。

Rasch 模型是离散选择模型中的重要分析模型。令y_{ij}为二元选择变量，其中，$i=1,\cdots,n$，$n=$对象，$m=$选项。有时“非平衡设计”也通过采用特定主题对应特定选项的方式进行估计。Rasch 模型中的潜在特征可以被认为由个体特征（如η能力）决定的固定效应或随机效应决定。

Rasch 模型可以采用以下一般化的描述：

$$\operatorname{logit}\Pr(y_{ij}=1|\eta_i)=\eta_i-\theta_j$$

其中η_i可以被理解为受试者的能力参数，θ_j为某一选项的难度参数。继而，给定η_i能力参数条件，y_i^*为自变量（局部独立），即其主要特征为具有充分统计条件$\{\eta,\theta\}$的$\{y_{+j},y_{i+}\}$集合。当处理各η和θ参数（固定效应）时，常用的最大似然估计量在渐进条件下不再为一致的估计量($n\to\infty,m$ fixed)。这在一定程度上由以下特征反映：如果我们观测到每增加的对象存在m个增加的观测值，但这样也增加了相应的一个待估能力参数η。那么，观测值的数目，即每个参数的信息随样本数n的增加而保持特征不变。

举例而言，具有相同结构的估计其最大似然估计存在非一致性。令$y_{ij}\backsim\ \text{Normal}(\mu_i,\nu^2),\ i=1,\cdots,n;j=1,2$。当参数$\mu_i$和$\nu^2$被估计时，其方差$\nu^2$的最大似然量为

$$\sum_{ij}(y_{ij}-y_{i.})^2/2n$$

其数学期望的概率极限为$(1/2)\nu^2$，而非ν^2。

3.3 数据来源

本次研究的数据来自某校商务英语学习平台的一次人机练习，一共有 120 名非英语专业的大二学生参加，这些学生来自某校金融专业、经贸专业。评分员共有 5 名，其中三名高校教师，教授商务英语课程，并且有过数次的评分经验，另外二名教师也教授商务英语课程，但是第一次参与这个考试的评分，5 名评分员分别对 120 名被试评分。所有被试参加的口语练习测试题每道大题包括两个部分，第一部分是模仿跟读，第二部分是情景模拟，给出设定商务场景，然后要求学生根据场景给出回应，采用的是人机对话的形式录制学生的回应。本次测试一共有三道大题，每道大题 10 分，被试要求在 18 分钟内做完这三道大题。由于教学班级规模较大，实施直接面试型口语考试难度大，因此采用计算机化的口语考试，之后教师通过回听学生考试录音的方式进行评分，每位教师是独立评分。评分方式为总体评分，具体的评分依据有：1）发音（包括声音大小、重音、语调、语气）；2）准确性（包括语法、专业用词及说话方式）；3）流利程度（包括语速、长短句搭配）。

3.4 分析模型

本研究的基本分析模型如下：

$$\log(P_{njk}/P_{nij(k-1)})=B_n-C_j-F_{jk}$$

P_{njk}表示评分员j给学生n打k分数段的概率; $P_{nj(k-1)}$表示评分员j给学生n打$k-1$分数段的概率; B_n为学生n的能力; C_j为评分员j的严厉程度; 而F_{jk}代表评分员j在k分数段的打分情况。

4.0 结果分析

4.1 各层面总体分析

此次研究使用的软件是 FACETS(Linacre, 2008a) ,我们对评分结果数据进行了多面 Rasch 模型分析。本研究将学生、评分者、测试题设为三个“面”(见图 1) 。

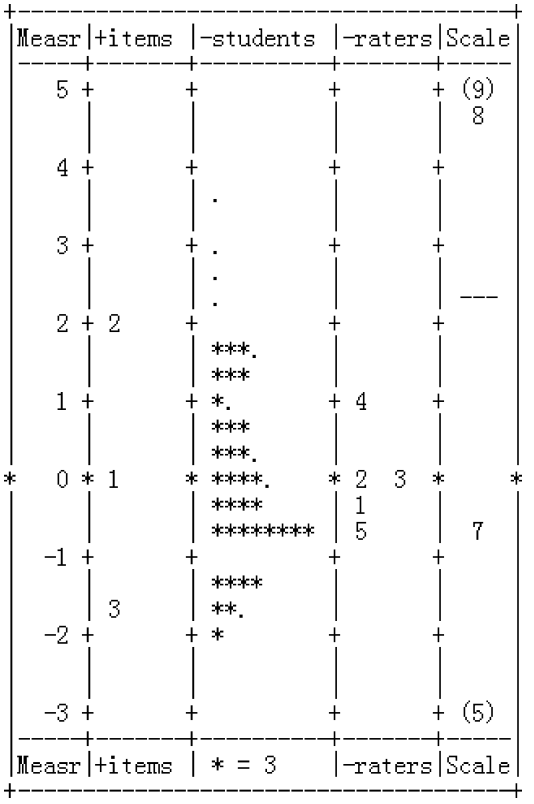

图 1　分层图

图 1 是所有层面的分布情况。最左列是 logit 尺度,是用来衡量各层面的真实测量值的尺度。第二列为测试题分布,这里体现的是试题的难度分布,有图可知,评分项的难度分布较为均匀,第二大题最难,第三大题最容易。第三列是学生能力,该图按照学生能力的高低自上而下排列,每个星号代表 3 个学生,每个圆点则代表少于 3 个学生。图 1 的结果显示,所有学生的能力介于 -2 到 +4 之间。第四列为评分员严厉度,严厉度高的评分员排在上面,严厉度低的评分员排在下面。由图可见 5 名评分员(编号为 1 -5) 的严厉度介于 -1 和 +1 之间,分布比较集中,同时最严厉的评分

员和最宽松的评分员之间的差异大约在 2 个 logits，不到考生能力度量跨度（大约 8 logits）的四分之一。这表示评分员之间严厉度的差异在总体上不会对考生的成绩产生决定性的影响（Myford & Wolfe，2004）。最后一列是模型给出的各能力段学生应该获得的分数（expected score）。如：logit 值为 0 的学生应该得到的分数大约为 7 分，括号内表示的是最高分和最低分。

4.2 分层面分析

多层面 Rasch 分析有个重要的优点就是它可以给出模型中各个层面的估量和度量并且提供每个层面甚至与每一个个体的单位统计量。

评分员层面

Rasch 模型分析显示，5 位评卷人的严厉度有差异（图 1），其中 4 号评分员最为严厉（1.12 logits），而 5 号评分员最宽松（－0.83 logits），他们之间相差 1.95 logits。5 位评分员的平均严厉程度为.00 logits，标准差为.63，其中 4 位评分员的严厉程度小于.00 logits，这显示出评分总体偏宽松。评分员的 Infit Mnsq 反应了评分员评分的内部一致性（internal self-consistency），对于这个取值，有不同的取值范围，比较严格的拟合取值是在 0.7－1.3 的范围内（Bonk & Ockey，2003），如果这个值在这个范围则认为评分员有较好的内部一致性。表 1 是评分员层面的统计数据，其中，分隔指数（Separation）为 5.15，分隔指数信度（reliability）为 0.96，一般认为分隔指数大于 2 即表示该层面的每个个体之间存在显著差异（Myford & Wolfe，2004）。这里的分隔指数信度是指该层面的个体之间有显著差异的程度。数据表明，评分员严厉度差异达到显著水平，不容忽视。同时从表 1 中可以看出这五位评分员中，有经验的评分员（1、2、3 号）虽然总体偏宽松，但是他们比首次评分的评分员（4、5 号）的评分严厉度差异更小。

表 1　评分员层面

Raters	Obsvd Average	Fair（M）Average	Measure	Model S. E.	Infit MnSq	Outfit MnSq
4	6.95	6.99	1.12	.11	1.14	1.18
3	7.18	7.12	－.01	.12	.82	.81
2	7.20	7.13	－.09	.12	.65	.61
1	7.22	7.15	－.19	.12	1.57	1.68
5	7.33	7.26	－.83	.12	.77	.62

Separation：5.15　　Reliability：.96

Chi-square：146.8　　Significance：.00

学生层面

表 2 为学生层面分析结果的一部分，因学生人数比较多，这里我们只看其中的一部分。这个层面是按照学生能力的高低排序的，能力高的学生排在上面。这里 Observed Average 是考生的实际平均得分，而 Fair Average 是结合题目难度而得到的期望分值，这个值更能体现学生的实际能力。Measure 值是表示学生能力的度量值，值越大表示学生的能力越高。本次考试学生的能力范围从－1.97 到 3.63 logits，跨越 5 个 logits，说明学生的

能力分布差异不是很大。Model S. E. 是指该模型估算的精确度。Infit Mnsq 是指学生的拟合统计值，我们可以根据这个值来判断哪些学生拟合模型以及哪些学生非拟合，并且可以计算出非拟合学生所占的比例。Linacre(2008b) 提出 0. 5 – 1. 5 可以作为拟合取值范围，那么本研究中大于等于 1. 5 拟合值的属于非拟合，共有 10 名学生非拟合，大约占总学生的 8%，这表示有 8% 的学生内部答题行为不太一致，可以进一步进行偏差分析，检查学生和试题项之间的交互作用。如这学生在哪些试题项上的成绩与其他试题项不同，是否是学生的答题方式的问题。(Linacre，2008b) 。

表 2　学生层面(部分)

Students	Obsvd Average	Fair-M Average	Measure	Model S. E.	Infit Mnsq	Outfit Mnsq
4	6. 33	6. 53	3. 63	. 52	2. 06	2. 25
2	6. 53	6. 73	2. 86	. 50	1. 38	1. 32
…	…	…	…	…	…	…
101	7. 47	7. 44	– 1. 60	. 62	1. 05	1. 04
76	7. 53	7. 54	– 1. 97	. 60	. 79	. 68

分隔系数(separation index) 为 1. 63(表 3) ，分隔指数信度的值在 0 到 1 之间，其值表明区分学生的能力的信度。这里的分隔指数信度是 0. 73，同时也通过卡方检验验证这种差异具有显著意义。说明此次考试较好地区分了学生的能力水平。

表 3　学生整体能力情况

Students (count: 120)	Mean	Obsvd Average	7. 18
		Fair M Average	7. 15
	Separation		1. 63
	Reliability		. 73
	Chi-square test		. 00

试题层面

从表 4 的结果来看，分隔系数为 29. 35，信度 1. 0，卡方值 2780. 5，显著性 = . 00，这些都说明了本次口语试题的难度在统计上存在显著差异，考试结果体现出来的试题难易差异较大(measure 最高和最低之间的差异为 3. 75 个 logits) 。口语考试中，由于试题的特点和话题的因素会造成难度上的差异，这也是研究者们关注的焦点和难点。不仅因为影响试题难度的因素很难确定，而且这些因素和学生之间的交互作用也是非常重要的(Bachman，2002) 。商务英语口语考试的话题任务设计涉及范围广，专业跨度大，话题可以涵盖经济、贸易、财政、金融等各个方面。本次考试三道大题的区别主要在模拟商务场景的主题上，三道大题的主题分别是 “解决营销问题”、“洽谈贸易折扣”、“国际清算业务会话”。根据 measure(表 4) 一栏可以看出每个题目的难度，这次考试难度最高的是第二大题(2. 02 logits) ①，但是题目的拟合分析没有发现有非拟合或过度拟合，说明题目的难度差异还是符合考试的要求。

表 4　试题层面

Item Number	Obsvd Average	Fair(M) Average	Measure	Model S. E.	Infit MnSq	Outfit MnSq
2	7.86	7.89	2.02	.10	1.13	1.10
1	7.08	7.06	-.05	.10	1.11	1.09
3	6.60	6.66	-1.73	.08	.81	.76

Separation: 29.35　　Reliability: 1.0

Chi-square: 2780.5　　Significance: .00

从表 4 我们可以推测，造成难度差异的主要原因是学生对于商务口语话题的熟悉程度、兴趣、以及对于话题所涉及信息的商务专业词汇的掌握差异比较大，专业知识的掌握影响到了学生口语能力的发挥。如果要进一步了解话题的难度，我们需要再分析成绩之间的统计差异，以及结合学生的个人特点，专业知识结构和学生考试时的话语进行具体分析。本次研究结果表明虽然试题在难度上体现出较大差异，这个难度差异总体上还对学生的考试成绩有一定影响，但是这几个任务的 Infit MnSq 都在 0.7 - 1.3 的范围内，说明评分员对各个试题的评分还是较为一致的，符合模型的期望。

4.3 评分量表使用分析

表 5 是评分量表各个分数段的使用统计，可以评估评分量表是否能够达到预期的使用目的。其中频数统计(Counts，Cum%) ，拟合均分指数(Outfit MnSq) 和 Rasch - Andrich 阈值(Rasch-Andrich Thresholds) 是分析量表使用情况的主要指标。通常拟合值小于 2.0 且阈值随分值递增且没有出现逆序，不同分数值之间的阈值差距相对均匀，说明评分量表的使用情况良好，评分员能够比较准确地区分各个分数段(Park，2004; 刘建达，2005; 张洁，2016) 。分析表 5 可以看出评分员总体上能够较好地使用评分量表，但是表 5 第二栏的频数统计了分数的中间段(7 分) 使用频次(54%) 远高于其他分数段，且 0 - 4 分数段评分员没有使用，这说明评分员在评分时有可能存在趋中性，为了明确这一点，我们还可以再看一下表 2 学生层面的能力分析，如果处于中间能力的学生确实要多于能力两端的学生，或是学生的能力差异比较小，那么评分员的趋中表现恰恰是非常合适的(Myford & Wolfe，2004) 。本研究中学生的能力跨度 5 个 logits，处于中间段的学生比较多，学生能力分布比较均匀。

表 5　分数类别使用统计

Data			Quality control			Rasch-Andrich Thresholds	
Score category	Counts	Cum%	Avge Meas	Exp. Meas	Outfit MnSq	Measure	S. E.
5	27 (2%)	2%	-5.05	-4.78	1.0		
6	232(13%)	14%	-3.11	-3.17	1.0	-6.15	.22
7	971(54%)	68%	-.58	-.55	1.0	-3.46	.09
8	536(30%)	98%	3.76	3.67	.9	2.25	.08
9	34(2%)	100%	4.61	5.21	1.1	7.36	.18

图 2 展示了 5 个分数段的概率曲线图，可以更直观地看各个分数段的使用质量，通过这 5 个分数段的峰值曲线，可以看出各分数段的峰值较为独立，间隔度相当，这说明评分员对于各个分数段可以较好地区分，也就是说能力处于某个分数段的学生能够得到这个分数段的分数。由图 2 可知本次考试所使用的各个分数段表现尚可，基本达到预期。

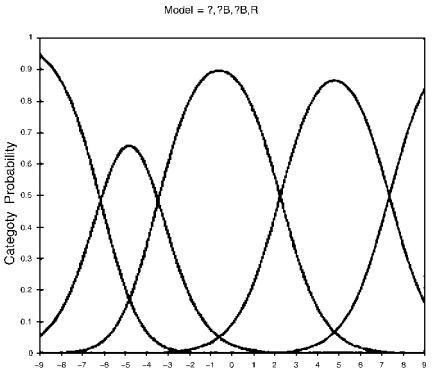

图 2　分类别概率曲线图

4.4 偏差分析

即使在阅卷前仔细地挑选评分员，进行阅卷前培训，评分员效应仍然可能存在（Bonk & Ockey，2003）。FACETS 的多面 Rasch 分析可以对评分数据进行偏差交互（Bias interaction）分析，所谓的偏差是指评分员在给分时出现了异常高分和异常低分的情况，通过偏差分析可以调查评分的哪个方面引起了评分员偏差，对哪些学生产生了评分偏差，尤其是在评分培训时分析偏差并反馈给评分员可以帮助他们修正评分偏差。

表 6　评分员 – 学生显著偏差交互统计

Raters	Severity Measure	Students	Ability Measure	Observd Score	Expected Score	Bias Size	Model S. E.	t
1	–0.19	48	–0.41	24	21.91	3.74	1.31	2.84
4	1.12	77	–0.81	23	21.44	3.01	1.29	2.34
4	1.12	68	–1.21	23	21.65	2.61	1.29	2.03
4	1.12	69	–1.21	23	21.65	2.61	1.29	2.03
4	1.12	70	–1.21	23	21.65	2.61	1.29	2.03
4	1.12	71	–1.21	23	21.65	2.61	1.29	2.03
1	–0.19	113	–0.03	20	21.72	–2.57	1.12	–2.3
3	–0.01	48	–0.41	20	21.82	–2.78	1.12	–2.48

本研究分析了评分员和学生的偏差交互，t（偏差统计量）的绝对值大于 2 即视为显著偏差。表 6 列出了出现显著偏差评分的评分情况，可见出现评分偏差数量较多的是 4 号评分员（首次评分），他对第 77 号等 5 位学生的评分过于宽松，这 5 位学生的能力均为中等偏下，这在首次评分的“新评分员”中普遍存在，对于差距较小的中等能力水平学生，他们往往难以区分，对评分标准把握不够准确，从而导致评分偏差。1 号评分员对 48 号学生评分过于宽松而对 113 号学生评分过于严厉，3 号评分员对 48 号学生的评分则过于严

厉,这一行为和 1 号评分员存在较大的差异,1 号和 3 号评分员都是有经验的评分员,因此在评分培训中,我们可以进一步询问这三位评分员,从而推断为什么出现这样的偏差。

5.0 结语

通过以上讨论,我们可以回答本研究提出的问题: 1) 评分者的严厉程度是否一致,评分者的评分是否存在内在一致性? 2) 题项是否能够很好地区分考生能力? 3) 评分质量如何,是否存在显著偏差? 根据前文评分者层面的分析,我们可以看到评分者的分隔指数和信度指数都很高,表明评分者之间总体严厉度差异显著,但是能保持评分员内部评分的一致性。本研究中的试题项难易区分度较大,这个结果表明了学生对商务话题的熟悉程度、兴趣以及话题中涉及的商务专业知识可能会影响学生口语能力的发挥。本研究中的评分质量较好,评分员可以有效地使用提供给他们的评分量表,但是评分员也存在评分偏差,应有针对性的给予评分培训。

多面 Rasch 模型分析结果使我们看到商务英语口语测试评分作为一种主观性评价可能会产生各种问题和偏差,影响对学生真实商务口语能力的评价。此外,对商务英语口语能力的评分除了要考虑影响通用英语口语考试的因素,也需要考虑到专业知识和能力对口语能力的影响。可以做如下几方面的改进: 1) 设计明确详细,具有可操作性的评分标准。商务英语口语考试的评分标准不能参照或是照搬通用英语口语考试的评分标准。描述清晰的、标准明确的、有针对性的评分标准是对商务英语口语能力准确评估的必要前提,对于评分者把握评分标准,保证一致性和评分信度至关重要。2) 加强对评分员的训练,特别是评分前培训,使其对评分标准充分理解以便更好地使用,最大限度地达到评分者总体宽严度的一致以及评分者内部的一致,避免评分偏差。

本研究是将多层面 Rasch 模型应用于商务英语口语测试评分研究中的一次尝试,尚存在一些局限需要在进一步的研究中改进,主要有以下三个方面:

1) 本研究通过 MFRM 定量分析商务英语口语测试的评分效应,并未收集和分析评分员以及学生的定性数据,因此在对定量统计结果的解释上缺少定性数据的佐证和补充。

2) 由于本次商务英语口语测试采用的是总体评分标准,因此研究者在微观层面上无法探索商务英语口语各分项评分维度的评分效应,在进一步的研究中,可以应用多层面 Rasch 模型对比分析总体评分和分项评分模式下评分员的评分依据,有助于改进和拓展评分标准。

3) 本次口语考试采用的是间接性的测评方法(人机对话,对学生的答题进行录音) ,进一步的研究可以收集学生对于这种考试形式的评价以及他们所希望的考试形式(录音或是面对面) ,从而探索适合校本商务英语口语测试的最佳形式。

注释:

① 具体题目请联系本文作者索取。

参考文献

[1] Andrich, D. A general form of Rasch's extended logistic model for partial credit scoring [J]. *Applied Measurement in Education*, 1978,4: 363 – 378.

[2] Bachman, L. F. Some reflections on task-based language performance assessment [J]. *Language Testing*, 2002,19: 453 – 476.

[3] Bonk, W. J. & G. J. Ockey. A many-facet Rasch analysis of the second language group oral discussion task [J]. *Language Testing*, 2003,20(1): 89 – 110.

[4] Douglas, D. *Assessing Languages for Specific Purposes* [M]. Cambridge: Cambridge University Press, 2000.

[5] Douglas, D. Language for specific purposes assessment criteria: Where do they come from [J]. *Language Testing*, 2001,18(2): 171 – 185.

[6] Eckes, T. Rater types in writing performance assessments: A classification approach to rater variability [J]. *Language Testing*, 2008,25: 155 – 185.

[7] Fulcher, G. *Testing Second Language Speaking* [M]. London: Pearson ESL, 2003.

[8] Goodwin, S. A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes [J]. *Assessing Writing*, 2016,30: 21 – 31.

[9] Hagan, S., J. Pill & Y. Zhang. Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective [J]. *Language Testing*, 2015,33: 195 – 216.

[10] Knoch, U. The development and validation of a rating scale for diagnostic writing assessment [J]. *Language Testing*, 2009,26(2): 275 – 304.

[11] Linacre, J. M. *Many-facet Rasch Measurement* [M]. Chicago: MESA Press, 1989.

[12] Linacre, J. M. Constructing measurement with a many-facet Rasch model [A]. In M. Wilson (ed.). *Objective Measurement: Theory in Practice Vol. II* [C]. Newark: Ablex, 1994.

[13] Linacre, J. M. FACETS: version 3.63.0 [CP/DK]. Chicago: Winsteps.com, 2008a.

[14] Linacre, J. M. *A User's Guide to FACETS: Rasch-model Computer Program* [M]. Chicago: MESA Press, 2008b.

[15] Myford, C. M. & E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement – Part II [J]. *Journal of Applied Measurement*, 2004,5(2): 189 – 227.

[16] O'Sullivan, B. (ed.). *Issues in Testing Business English: Studies in Language Testing, Volume 17* [M]. Cambridge: Cambridge University Press, 2006.

[17] Park, T. An investigation of an ESL placement test of writing using Many-Facet Rasch Measurement [J]. *Teachers College, Columbia University, Working Paper in TESOL & Applied Linguistics*, 2004,4(1): 1 – 21.

[18] Schaefer, E. Rater bias patterns in an EFL writing assessment [J]. *Language Testing*, 2008,25(4): 465 – 493.

[19] Wright, B. D. & G. N. Masters. *Rating Scale Analysis* [M]. Chicago: MESA Press, 1982.

[20] 戴朝晖,尤其达. 大学英语计算机口语考试评分者偏差分析 [J]. 外语界,2010,(5): 87 – 95.

[21] 对外经济贸易大学商务英语理论研究小组. 论商务英语的学科定位、研究对象和发展方向 [J]. 中国外语,2006,(9): 4 – 8.

(下转第 49 页)

[18] 习近平. 决胜全面建成小康社会　夺取新时代中国特色社会主义伟大胜利 [N]. 人民日报，2017－10－28.
[19] 虞建华. 谈我国高校英语专业“两个走向”问题——兼及英美文学教学 [J]. 中国外语，2010，(3)：14－18.
[20] 赵世举. 语言是保障国家经济安全的要素 [N]. 中国教育报，2013－12－13.
[21] 仲伟合. 拔尖创新型国际化人才培养模式的探索与实践——以广东外语外贸大学为例 [J]. 广东外语外贸大学学报，2013，(1)：98－101.
[22] 仲伟合. 《英语类专业本科教学质量国家标准》指导下的英语类专业创新发展 [J]. 外语界，2015：(3)：2－8.
[23] 仲伟合，王巍巍. “国家标准”背景下我国英语类专业教师能力构成与发展体系建设 [J]. 外语界，2016，(6)：2－8.
[24] 仲伟合，王巍巍，黄恩谋. 国家外语能力建设视角下的外语教育规划 [J]. 语言战略研究，2016，(5)：45－51.
[25] 庄智象. 我国外语专业建设与发展的若干问题思考 [J]. 外语界，2010，(1)：2－10.

（责任编辑：刘焱）

（上接第32页）

[22] 范劲松，季佩英. 口语测试中分析性评分量表的构念效度研究 [J]. 中国外语教育，2015，(3)：85－94.
[23] 高淼. 基于多面 Rasch 模型的初中英语口语测试 EBB 评分标准研究与效度验证 [J]. 中国考试，2016，(12)：29－47.
[24] 何莲珍，张洁. 多层面 Rasch 模型下大学英语四六级考试口语考试(CET-SET)信度研究 [J]. 现代外语，2008，(4)：388－398.
[25] 李英，关丹丹. PETS 口试评分培训效果的多面 Rasch 分析 [J]. 外语教学理论与实践，2016，(3)：43－47.
[26] 刘建达. 话语填充测试方法的多层面 Rasch 模型分析 [J]. 现代外语，2005，(2)：157－169.
[27] 刘建达. 评卷人效应的多层面 Rasch 模型研究 [J]. 现代外语，2010，(2)：185－193.
[28] 张洁. PETS 三级口语考试评分质量控制研究——基于多侧面 Rasch 模型(MFRM)的方法 [J]. 考试研究，2008，(4)：65－78.
[29] 张洁. 基于多层面 Rasch 模型的评分员评分质量诊断 [J]. 外语测试与教学，2016，(2)：47－54.
[30] 赵南，董燕萍. 基于多面 Rasch 模型的交替传译测试效度验证 [J]. 解放军外国语学院学报，2013，(1)：86－90.
[31] 赵守盈，薛雯. Rasch 模型和 IRT 在学生成就测验统计分析中的对比研究 [J]. 中国考试，2011，(6)：8－12.
[32] 周燕，曾用强. 机助英语听说考试计算机自动评分的多层面 Rasch 模型分析 [J]. 外语测试与教学，2016，(1)：22－31.
[33] 朱正才，杨惠中，杨浩然. Rasch 模型在 CET 考试分数等值中的应用 [J]. 现代外语，2003，(1)：70－75.

（责任编辑：刘琛）