计算机辅助口试评分稳定性历时研究
——以 PRETCO 口试为例

杨志强[1] 李志芳[2] 董曼霞[3]

(1. 重庆科技学院 外国语学院,重庆 401331/广东外语外贸大学 外国语言学及应用语言学研究中心,广东 广州 510420;
2. 陆军军医大学 基础医学院外语教研室,重庆 400038;3. 四川外国语大学 商务英语学院,重庆 400031)

摘 要:口语考试评分员的稳定性事关考试的效度、信度及公平性。本文对连续五次计算机辅助 PRETCO 口试评分进行历时分析,探讨 PRETCO 口试评分员在严厉度、评分准确度以及集中趋势三个方面的稳定性,并探究其背后的原因。

关键词:计算机辅助口试;PRETCO 口试;评分;稳定性

中图分类号:H319.3 **文献标志码**:A **文章编号**:1674-6414(2021)02-0126-10

0 引言

外语口语能力是外语水平的直接表现。随着英语口语越来越受到重视,参加口试的考生逐年增多,人工实考及评分已经无法满足操作需求。近 20 年来,计算机技术和测试理论的不断发展及融合,突显了计算机辅助考试的优势,如信度高、节约费用、考试时间灵活、便于组织等(曾用强, 2011)。因此,该技术已广泛应用于大规模考试实践中(金力,2011),包括 TOFEL 机考、CET 口试、TEM 口试以及高等学校英语应用能力口语考试(Practical English Test for College – Oral,简称 PRETCO 口试)等。PRETCO 口试是由国家高等学校英语应用能力考试委员会于 2005 年开始实施的口语考试,该考试是以人机对话方式进行的计算机辅助考试(刘鸿章 等, 2010)。由于计算机辅助口试自动评分技术还不成熟,目前仍采用人工评分。人工主观评分容易出现误差,所以有必要对评分的信度进行研究(Myford et al., 2004)。此外,评分员的评分可能随着时间的进展发生变化(Myford et al., 2004),而且评分的稳定性直接关系评分的质量、评分员的遴选,以及考试的信度、效度和公平性等(赵海燕 等, 2018),因此,对评分员的稳定性进行研究具有重要的实际意义。虽然近年来有关口语测试评分信度展开的研究日益增多(何莲珍等, 2008;刘建达, 2010;Attali, 2016;Kang et al., 2019)但这些研究都只对单次的评分作了分析,没有对评分的稳定性进行历时研究。为此,本文拟基于 PRETCO 口试连续五次的评分结果,调查评分员评分的稳定性,以期为 PRETCO 口试的评分提供一些启示,同时为其他高风险计算机辅助口试,如 CET 口试、TEM 口试的评分或评分培训提供一些参考。

1 文献回顾

国外有关口语测试的研究起步早,覆盖广,如口语测试的构念(Luoma, 2004)、口语测试的效度验证

收稿日期:2020-10-16

作者简介:杨志强,男,重庆科技学院讲师/广东外语外贸大学博士生,主要从事语言测试研究。
李志芳,女,陆军军医大学基础医学院外语教研室讲师,硕士,主要从事语言测试、语言教学研究。
董曼霞,女,四川外国语大学商务英语学院教授,博士,主要从事外国语言文字、教育理论与管理研究。

(Knoch et al. 2018)、口试的任务(Frost et al., 2020)、口试评分标准(Fulcher, 1996; Khabbazbashi et al., 2020)、受试的特征(Nakatsuhara, 2011)、评分培训及评分员对考生口试表现的影响(Kang et al, 2019)、口试的评分效度研究(Lumley et al., 1995;Elder et al., 2005;Attali, 2016),等等。其中,有关口试评分的研究占多数。虽然评分员的评分是动态变化的(Myford et al., 2004),但大部分研究只对单次的评分进行了分析。目前,仅有个别文献采用现代测试方法,比如基于项目反应理论的多层面 Rasch 模型,对口试评分进行了历时分析(Lumley et al., 1995;Bonk et al., 2003;Kim, 2015),然而这些研究的结果存在差异。Lumley 等(1995)分析了四名评分员三次职业英语口试(Speaking subtest of Occupational English Test)的评分结果,发现评分员评分的严厉度随着时间发生了变化,且宽严度变化的趋势不尽相同;Bonk 等(2003)基于对某校本英语口试两轮评分结果的分析,发现评分员的严厉度差异较大,而且不稳定,评分员的内部一致性随着其评分经验的积累不断加强;Kim(2015)通过采用定性的研究方法,对比了新、中、老口试评分员的三次评分行为,发现三组评分员历次的评分能力存在差异,新评分员改进较慢,中评分员通过不断培训得以不断改进,老评分员则相对较为稳定。

虽然国内有文献对口语测试的评分进行了研究(何莲珍 等, 2008;刘建达, 2010),但这些研究同样只对评分员某次的评分进行分析。截至目前,国内尚无文献从历时的角度探讨口试评分的稳定性。因此,本文将以此为出发点,基于多层面 Rasch 模型和 Myford 等(2009)写作评分漂移研究的框架,从评分员严厉度、准确度以及集中趋势三个方面对 PRETCO 口试的评分稳定性进行分析与研究。其中,评分员严厉度是指评分的宽严度,评分员准确度是指相对于其他评分员评分均衡性,集中趋势是指评分员高频率使用中间分数段(Myford et al., 2004)。

2　研究方法

2.1 评分员及阅卷量

由于本研究中 PRETCO 口试阅卷点每次评阅的数量不统一,评分员的数量不定,一般在 10--20 人之间,评分员分别来自 15 所不同的高校。本研究所选取的五次 PRETCO 口试阅卷结果共涉及到 6525 份,其中第一次为 1493 份,第二次为 1356 份,第三次为 1351 份,第四次为 870 份,第五次为 1455 份,参加阅卷任务的评分员共 45 名。每位考生的口语由两名评分员进行评分,因此总阅卷数为 13050 份。本研究评分员 R0、R1 和 R2 连续参加了五次阅卷任务,评分员 R4、R5 和 R6 连续参加了前四次阅卷任务,其具体信息见表 1:

表 1　评员基本情况

项目 / 评分员	评分员基本信息				阅卷员阅卷情况				
	阅卷经验	性别	职称	年龄	第一次	第二次	第三次	第四次	第五次
R0	7 年	女	副教授	51	√	√	√	√	√
R1	8 年	女	教授	54	√	√	√	√	√
R2	8 年	女	教授	53	√	√	√	√	√
R3	9 年	男	副教授	44	√	√	√	√	×
R4	8 年	男	讲师	43	√	√	√	√	×
R5	新评分员	女	讲师	38	√	√	√	√	×

2.2 PRETCO 口试及其评分标准

RRETCO 口语考试形式为人机对话,主要由朗读、问答、翻译(汉译英)以及口头陈述四部分任务组成,整个考试过程约为 20 分钟(《高等学校英语应用能力考试大纲》修订组, 2016)。每次 PRETCO 口试会采用2--4 套平行试题,每项任务总分为4 分,采用七级记分制(0,1,2,2.5,3,3.5,4),为方便计算,本研究将其转换为 1,2,3,4,5,6,7 七个等级。“朗读”主要从语音、语调以及流利程度三方面进行评分(见表 2);“问答”“翻译”和“陈述”主要从内容、表达、语言三方面进行评分(见表 3)。两位评分员分别独立对考生四项任务的表现进行评分,然后再根据每个任务的得分算出口试总分。如果两者评分出现等级差异,由第三位高级评分员(评分组长)进行仲裁,重新进行整体评分。

表 2 朗读任务评分标准

分数	等级	语音语调	流利程度
4	7	语音语调正确,重音、停顿、意群恰当	朗读流利清晰
3.5	6	语音语调基本正确,重音、停顿、意群基本恰当	朗读比较流利清晰
3	5	有少量语音语调错误,但不影响理解	朗读基本顺利
2.5	4	有较多语音语调错误,但还能理解	朗读有时有停顿、重复
2	3	有较多语音语调错误,影响理解	朗读不很顺利,常有停顿、重复
1	2	朗读基本听不懂	朗读很困难
0	1	不能朗读或没有朗读	

表 3 陈述评分标准

分数	等级	内容与表达	语言
4	7	能清楚连贯地介绍画面所包含的重要信息,有评述	语句符合规范
3.5	6	能基本清楚连贯地介绍画面所包含的重要信息,有评述	语句比较符合规范
3	5	能基本连贯地传达画面所包含的重要信息	语句基本正确,用词基本恰当
2.5	4	能表述画面重要信息,有少量停顿或重复	语句用词有一些错误,但不影响理解
2	3	能勉强表述画面主要信息,但遗漏较多,经常停顿或重复	语句有较多用法错误,个别地方费解
1	2	基本不能表述画面信息,叙述混乱,层次不清,仅个别的地方可以理解	语句错误很多,很难理解
0	1	没有答题或答题无法理解	

注:(1)由于 FACETS 要求使用整数数据,所以本文将所有原始分数换算成相应的七个等级(1,2,3,4,5,6,7);(2)囿于篇幅,而且考虑到问答、翻译和陈述都是从内容、表达、语言三方面进行评分,故只列出其中一种评分标准

2.3 数据分析依据

本研究基于多层面 Rasch 模型,采用 FACETS 软件(版本 3.71.3)(Linacre, 2013)对历次 PRETCO 口试评分结果进行分析。模型包括四个层面,考生能力、评分员、口试的四项任务以及评分次序。鉴于 PRETCO 四项任务具体的评分标准不一致,所以本研究采用多层面 Rasch 模型中分部记分模型(Partial Credit Model)(Bonk et al. ,2003)。此外,以往研究忽略了数据链接(connectivity)的重要性(Wind et al. , 2018)而探究评分员历时评分的稳定性需要链接(link)历次评分的数据。本研究中评分员 R0 五次评分的各项指标,比如严厉度和加权均方拟合度都在合理的范围,所以选用该评分员的总体评分作为链接数据,以观察另外五位评分员(R1、R2、R3、R4 和 R5)评分的稳定性。同时,本研究借鉴 Myford 和 Wolfe (2009)对于评分员写作评分漂移研究的框架,从评分员严厉度、准确度以及集中趋势三个方面对

PRETCO 口试的评分稳定性进行历时分析与研究。

首先,对于严厉度的稳定性,传统方法是采用分离模型和交互模型计算各个时间段的严厉度 logit 值,然后进行显著性检验(Myford et al. , 2009)。然而,由于交互模型存在混合测量误差(Dobria, 2011),所以本研究未采用该方法计算评分员严厉度稳定性的偏差,而是将评分员在每次评分中视作不同的评分员,可以根据评分员的 logit 值直接观察评分员严厉度的变化。其次,关于评分员评分准确度的历时变化,可以基于评分员的点二列相关系数($r_{SR\text{-}ROR}$,即 Point-biserial Correlation 或 Point Measure)进行判断(Myford et al. , 2004)。检验评分员准确度的变化趋势需要根据公式(一)将相关系数转化为 Fisher's Z 值,然后再通过 Z 检验(公式二)来判断评分员评分准确度的稳定性是否具有统计意义上的显著性(Myford et al. , 2009)。

$$Z_{SR\text{-}ROR} = \frac{LN(1+r_{SR\text{-}ROR})\text{-}LN(1-r_{SR\text{-}ROR})}{2}$$ 公式(一)

其中 SR-ROR 是指评分员的点二列相关系数$r_{SR\text{-}ROR}$。

$$Z_{SR\text{-}RORc,\ SR\text{-}RORb} = \frac{Zr_{SR\text{-}RORc}\text{-}Zr_{SR\text{-}RORb}}{\sqrt{\frac{1}{Nc\text{-}3}+\frac{1}{Nb\text{-}3}}}$$ 公式(二)

其中$Zr_{SR\text{-}RORb}$表示评分员$r_{SR\text{-}ROR}$系数在第 b 次评分转化后的 Fisher's Z 值;$Zr_{SR\text{-}RORc}$表示评分员$r_{SR\text{-}ROR}$系数在第 c 次评分转化后的 Fisher's Z 值;Nb 和 Nc 分别表示评分员在第 b 次和第 c 次的阅卷量。如果$Z_{SR\text{-}RORc,\ SR\text{-}RORb}$的值大于 1.96,则表明评分员在第 c 次与其他评分员的一致性显著高于在第 b 次与其他评分员的一致性(显著性水平 0.05,下同);如果$Z_{SR\text{-}RORc,\ SR\text{-}RORb}$小于-1.96,则表明评分员在第 c 次与其他评分员的一致性显著低于在第 b 次与其他评分员的一致性。

最后,关于集中趋势的稳定性,历次评分阈值(Threshold)的标准差可以用作判断集中趋势稳定性的参数。所谓阈值是指相邻分数段概率曲线的交叉值(Bond et al. , 2015)。本研究基于 Rasch 的混合模型(Hybrid Model 2)(Myford et al. , 2004),通过计算单个评分员每次评分中对各项任务评分标准的使用情况,然后根据评分员每次评分阈值的标准差来判断其评分集中趋势的波动情况。分数段之间阈值离散程度越大,表明评分越集中。本研究在评分结束后对评分员进行了半结构式访谈,主要问题为"你是如何阅朗读/问答/翻译/陈述任务的?""你认为你历次的评分是否稳定?""哪些因素可能会影响评分的稳定性?"等。研究者对访谈录了音并转写为文字,最后根据 Given(2008)的归纳法对访谈内容进行分析和归纳。

3　研究结果

文章从评分员的严厉度、准确度和集中趋势三个方面报告评分员历次评分的稳定性。

3.1 评分员严厉度的稳定性

为了探讨评分员严厉度的稳定性,本研究分别将评分员 R0 的评分作为链接数据,以观察另外五位评分员的评分表现。通过 FACETS 的运算,五次评分总体评分严厉度 logit 的均值为 0.41,标准差为0.43 logits。评分员 R2 在第四次出现了明显的偏差,logit 值为 0.53,而第五次的 logit 值为 -0.41(见图 1),相差 0.94logits,大于两个标准差。其次,评分员 R3 第一次和第二次评分的偏差较大,分别为 0.55logits 和 1.06logits,相差 0.51logits,大于一个标准差。

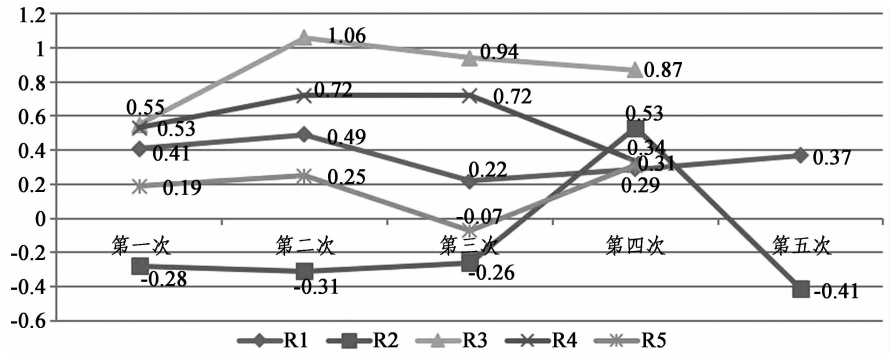

图 1 评分员评分严厉度的稳定性对比(logit)

评分员 R1、R4 和 R5 评分的严厉度较为稳定,波动较小,严厉度最高值和最低值之差分别为0.27 logits、0.38logits 和 0.36logits,均小于 1 个标准差(S. D. =0.43logits)。

3.2 评分员准确度的稳定性

表 4 中 $r_{SR\text{-}ROR}$为评分员每次评分的点二列相关系数值,$Z_{SR\text{-}RORc,SR\text{-}RORb}$为检验评分员准确度稳定性的 Z 值。$r_{SR\text{-}ROR}$可以判断评分员的评分与其他评分员评分的一致性,如果评分员 $r_{SR\text{-}ROR}$的值越大,则表示该评分员的一致性越好,不存在随机性(Myford et al., 2004; 刘建达 2010)。

表 4 评分员准确度稳定性相关数据统计

		评分次序				
项目	评分员	第一次	vs 第二次	vs 第三次	vs 第四次	vs 第五次
$r_{SR\text{-}ROR}$	R1	0.71	0.73	0.80	0.75	0.77
	R2	0.64	0.64	0.67	0.80	0.59
	R3	0.75	0.76	0.79	0.78	/
	R4	0.75	0.73	0.71	0.75	/
	R5	0.71	0.79	0.67	0.73	/
$Z_{SR\text{-}RORc,SR\text{-}RORb}$	R1	/	0.73	4.04	1.68	0.94
	R2	/	0.00	0.97	6.52	-8.18
	R3	/	0.37	1.62	1.23	/
	R4	/	11.68	-1.60	0.00	/
	R5	/	2.93	-1.24	0.73	/

由表4 可知,评分员五次评分的 $r_{SR\text{-}ROR}$值位于0.59 -0.81 之间,评分员历次评分中和其他评分员一致性较好,但所有 $Z_{SR\text{-}RORc,SR\text{-}RORb}$值中,大于 1.96 或小于 -1.96 的次数为五次,其中评分员 R3 评分的准确性波动不明显,Z 值均小于 1.96 或大于-1.96($p<0.05$);评分员 R1、R4 和 R5 各出现一次显著性的波动(Z 值分别为 4.04、11.68 和 2.93,$p<0.05$);评分员 R2 出现两次波动,第三次评分准确性明显高于第一次,而第四次评分又明显低于第一次(Z 值分别为 6.52 和-8.18, $p<0.05$。由此可以看出,所有评分员历次评分的准确性均不稳定,其中四位评分员出现了五次明显的波动,仅占评分员阅卷总次数的 22.7%。

3.3 评分员集中趋势的稳定性

相邻阈值的差一般要求大于 1.0 logit,但不超过 5.0 logits (Linacre, 2002)。由表 5 可见,大多数评分员评分阈值的差位于 1.0 -5.0 logits 之间,阈值的标准差位于 2.1 -4.0logits 之间。历次 PRETCO 口试的总体评分较为稳定,多数评分员总体不存在明显的集中趋势。然而,评分员 R5 第一次评分没有使用分数段 1 和分数段 7,而且分数段 3、4、5 的使用率达到 90%,因此该评分员第一次的评分较为集中;评分员 R3 历次评分中分数段 3、4、5 之间的阈值差较小,均小于 1.0 logit;评分员 R2 第二次、第三次和第五次

评分中分数段 5、6、7 之间的阈值差较小,同样小于 1.0 logit。这两位评分员可能对上述几个分数段难以把握或理解出现偏差。

表 5　评分员四项任务总体阅卷量(%)及阈值(logits)

		第一次		第二次		第三次		第四次		第五次	
	分数段	阅卷量/%	阈值	阅卷量/%	阈值	阅卷量/%	阈值	阅卷量/%	阈值	阅卷量/%	阈值
R1	1	18/2%		9/1%		22/2%		33/5%		24/3%	
	2	70/9%	-4.9	33/6%	-5.4	60/9%	-4.7	73/10%	-3.6	69/9%	-3.1
	3	219/27%	-3.0	129/25%	-3.5	136/21%	-2.9	161/22%	-2.6	178/23%	-2.3
	4	262/32%	-1.0	169/33%	-1.1	191/29%	-1.3	223/31%	-1.3	241/31%	-0.9
	5	194/24%	0.4	133/26%	0.6	209/32%	0.2	184/25%	0.3	183/24%	0.4
	6	52/6%	2.4	38/7%	2.9	45/7%	3.3	46/6%	2.7	71/9%	1.8
	7	1/0%	6.2	1/0%	6.6	5/1%	5.4	8/1%	4.5	6/1%	4.1
R1 阈值标准差			3.6		4.0		3.5		2.9		2.4
R2	1	3/0%		19/3%		24/4%		25/4%		13/2%	
	2	22/3%	-3.8	23/4%	-3.0	34/6%	-3.4	100/15%	-4.6	30/4%	-2.9
	3	107/13%	-2.7	71/11%	-3.1	93/16%	-3.1	180/27%	-2.7	79/11%	-2.3
	4	253/31%	-1.1	153/24%	-1.8	188/32%	-2.1	163/24%	-1.1	175/24%	-1.4
	5	278/34%	0.5	280/44%	-0.4	205/35%	-0.5	143/21%	0.0	266/37%	-0.3
	6	126/16%	2.4	83/13%	2.8	35/6%	2.7	62/9%	2.0	146/20%	1.7
	7	19/2%	4.6	7/1%	5.5	1/0%	6.4	3/0%	6.2	7/1%	5.0
R2 阈值标准差			2.9		3.2		3.5		3.5		2.7
R3	1	10/2%		44/8%		25/4%		45/7%			
	2	84/17%	-4.4	73/14%	-3.0	69/12%	-2.9	85/13%	-2.9	/	/
	3	159/32%	-1.5	209/39%	-2.9	226/38%	-2.5	229/34%	-2.4	/	/
	4	63/12%	0.6	98/18%	-0.3	98/16%	0.2	102/15%	0.2	/	/
	5	99/20%	-0.2	85/16%	0.0	92/15%	0.2	120/18%	0.1	/	/
	6	82/16%	1.2	25/5%	2.0	81/14%	1.0	76/11%	1.6	/	/
	7	7/1%	4.3	2/0%	4.1	9/2%	3.9	23/3%	3.3	/	/
R3 阈值标准差			2.6		2.5		2.3		2.1		
R4	1	20/9%		15/3%		27/4%		23/3%		/	/
	2	32/15%	-3.0	80/16%	-4.8	78/11%	-3.4	100/15%	-4.6	/	/
	3	65/31%	-2.4	162/31%	-2.6	165/23%	-2.4	201/30%	-2.6	/	/
	4	50/24%	-0.6	139/27%	-0.8	213/30%	-1.2	149/22%	-0.6	/	/
	5	35/17%	0.4	93/18%	0.6	172/24%	0.2	132/20%	0.4	/	/
	6	8/4%	2.4	25/5%	2.6	60/8%	2.1	50/8%	2.5	/	/
	7	2/1%	3.1	2/0%	4.9	5/1%	4.8	9/1%	4.8	/	/
R4 阈值标准差			2.2		3.2		2.8		3.1		
R5	1	/		5/1%		18/4%		60/8%		/	/
	2	16/3%		36/8%	-6.1	25/5%	-2.8	79/12%	-3.5	/	/
	3	145/25%	-3.5	174/38%	-2.6	102/20%	-3.2	77/11%	-2.5	/	/
	4	172/30%	-0.4	123/27%	1.0	166/33%	-1.7	158/23%	-2.6	/	/
	5	198/35%	0.6	102/22%	2.3	148/29%	-0.2	278/41%	-1.5	/	/
	6	41/7%	3.3	16/4%	5.5	43/9%	2.2	27/4%	3.1	/	/
	7	/		/		2/0%	5.7	1/0%	6.9	/	/
R5 阈值标准差			2.4		4.0		3.1		3.8		

通过对单项任务的分析可知,评分员阅读任务历次评分中阈值的标准差位于 3.9 -8.6logits 之间(见表 6),明显高于其四项任务总体评分阈值的标准差。以评分员 R1 为例,其朗读任务历次评分的阈值标

准差分别为 6.8、5.3、8.6、7.4 和 5.4（logits）。评分员 R1、R2、R4 和 R5 的历次评分都过多地使用了分数段 4 和分数段 5，评分员 R3 则过多地使用了分数段 5 和分数段 6，比例多数超过 70%，评分员 R3 第三次的使用频率甚至达到 90%。由此可见，评分员在阅读任务的历次评分中都存在明显的集中趋势。

表 6　评分员朗读任务阅卷量及阈值

		第一次		第二次		第三次		第四次		第五次	
	分数段	阅卷量/%	阈值	阅卷量/%	阈值	阅卷量/%	阈值	阅卷量/%	阈值	阅卷量/%	阈值
R1	1	1/0%		1/0%		2/0%		1/0%		/	
	2	8/1%		/		2/0%		3/1%		2/1%	
	3	21/11%	-7.6	14/9%		13/8%		17/10%	-8.2	11/6%	-6.2
	4	87/45%	-3.0	42/37%	-5.0	48/30%	-10.0	70/39%	-2.8	44/23%	-3.6
	5	75/39%	2.7	54/47%	-0.6	77/48%	-3.3	84/47%	1.6	76/40%	-0.5
	6	12/5%	8.0	17/8%	5.6	21/13%	3.3	7/3%	9.4	54/28%	2.9
	7	/		/		4/1%	10.1	/		6/3%	7.5
R1 阈值标准差			6.8		5.3		8.6		7.4		5.4
R2	1	/		1/0%		1/0%		1/0%		/	
	2	/		2/1%	-3.5	/		2/1%	-10.1	/	
	3	5/2%		7/5%	-3.3	18/9%		10/6%	-8.9	12/6%	
	4	58/30%	-6.7	35/23%	-2.3	53/39%	-4.5	58/35%	-6.5	51/30%	-4.1
	5	91/47%	0.8	93/60%	1.0	63/46%	-0.5	61/37%	-0.3	85/50%	-0.2
	6	48/21%	5.9	21/11%	8.1	10/6%	5.0	35/21%	3.5	31/15%	4.4
	7	/		/		/		2/1%	12.2	/	
R2 阈值标准差			6.3		4.9		4.7		8.4		4.3
R3	1	1/1%		1/0%				/			
	2	/		2/2%				2/1%			
	3	/		12/9%	-6.9			/			
	4	7/6%		30/23%	-4.5	6/4%		7/4%			
	5	55/44%	-7.6	64/48%	-2.0	67/45%	-6.5	74/44%	-7.0		
	6	57/46%	-0.3	23/17%	3.9	68/45%	0.5	71/42%	0.8		
	7	4/3%	7.8	2/2%	9.6	9/6%	6.1	16/9%	6.2		
R3 阈值标准差			6.3		4.7		5.1		5.4		
R4	1	/		/		/		2/1%			
	2	2/1%		/		/		3/2%			
	3	4/2%		25/18%		9/5%		24/14%			
	4	43/26%	-7.6	53/43%	-8.1	58/32%	-6.9	50/30%	-5.7		
	5	84/50%	-2.3	41/33%	-2.5	81/45%	-2.1	65/39%	-2.1		
	6	27/16%	3.3	7/6%	4.0	29/16%	2.1	17/10%	2.4		
	7	8/5%	6.6	1/1%	6.8	3/2%	6.9	5/3%	5.4		
R4 阈值标准差			5.4		5.8		5.1		4.3		
R5	1	/		1/0%		/		2/1%			
	2	1/0%		0/0%		/		4/2%	-5.2		
	3	12/8%		20/18%		13/10%		10/6%	-4.3		
	4	54/39%	-4.8	34/30%	-5.3	41/33%	-4.7	26/16%	-3.0		
	5	63/45%	-0.1	50/44%	-1.6	54/43%	-0.3	125/75%	-0.6		
	6	13/8%	4.9	9/8%	6.9	17/13%	4.9	3/1%	13.7		
	7	/		/		/		/			
R5 阈值标准差			4.0		5.1		3.9		7.0		

对于评分员其他任务的历次评分,问答任务都不存在集中现象。翻译和陈述任务历次评分中,个别评分员偶尔会出现集中趋势现象,比如评分员 R1 在第一次的陈述评分中出现了集中趋势。需要指出的是,评分员在五次翻译和陈述评分中,分数段 7 的使用率非常低,平均每次的使用率为 0.13 次和0.33 次。

4　讨论

4.1 评分员严厉度稳定性

数据显示多数评分员评分严厉度的总体趋于稳定,评分员历次评分中宽严度变化的趋势却不尽相同,这与 Lumley 等(1995)的研究发现相似。评分员评分严厉度总体波动不大,原因可能是:(1)评分员不断熟悉评分标准,比如评分员在每次评分前都接受培训并认真学习评分标准;(2)评分员评分时结合了教学和评分经验,比如评分员 R1 根据考生的语音、语调、断句和流利度推断考生的口语水平。但数据同样显示,评分员 R2 和 R3 分别在第四次和第二次评分中出现了明显的波动,这与 Kim(2015)的研究结果不一致,即使是有经验的评分员,其评分也可能会出现波动。虽然评分员 R2 阅卷经验丰富,而且每次都认真接受评分培训,但依然在第四次出现了明显的偏差。通过对评分员 R2 的访谈得知,该评分员的历次评分都严格按照评分标准进行阅卷,不应该存在明显的波动。为了究其原因,研究者同时对比了相邻两次考试的评分结果(第四次和第五次)。第四次评分的总量较少,当时考试只使用了两套试题,评分员 R2 只评阅了第一套试题的考生,其余评分员所阅考生均使用了两套试题。通过对两套试题的分析得知,其难度存在显著差异,比如第一套试题朗读任务的易读度为 76.5,明显比第二套(易读度为 65.6)①*简单,所以试题难度不同可能会影响评分员评分的稳定性。对于评分员 R3,其评分的严厉度也出现了较为明显的波动。通过访谈得知,该评分员第二次阅卷时除了正常教学和承担一定的行政工作外,还要准备博士研究生的考试,当时阅卷出现波动可能和压力大、身心疲惫有关。由此可见,“平行试题”中某些题型可能存在难度差异,影响评分员评分的严厉度。评分员评分时的身心状态也会影响评分结果。

4.2 评分员准确度稳定性

评分员单次评分和其他评分员的一致性较好,但历时来看,五位评分员的准确度都不太稳定,其中四位评分员共出现五次明显的波动。评分员 R2 出现两次显著的波动,评分员 R1、R4 和 R5 分别出现一次显著的波动。原因可能来自两方面,首先评分员阅卷队伍不稳定性。虽然每次阅卷员的数量为 10—20 名左右,但参加五次评分的评分员只有三名,即评分员 R0、R1 和 R2,连续参加四次评分的评分员也只有三名,即评分员 R3、R4 和 R5。出于公平性和实际情况的考量,阅卷员来自不同的高校,而且每次可能会有个别新评分员加入评分队伍。由于评分员评分的准确度涉及和其他评分员评分的一致性,故评分员队伍不稳定可能会导致评分员准确度出现波动;其次,评分的准确度的稳定性可能和考生的水平相关。由于每次报考 PRETCO 口试的学校和学生存在变化,不同批次考生的口语水平会存在一定的差异,从一定程度上可能会影响评分员评分的稳定性。

4.3 评分员集中趋势稳定性

评分员历次的总体评分不存在明显的集中趋势,但评分员 R5 第一次评分的集中趋势较为明显,分数段 3、4、5 的使用次数占其评分总数的 90%。该评分员可能第一次参加 PRETCO 口试评分,对评分标准的把握不准确,四项任务均没有使用分数段 1 和分数段 7。由此可见,新评分员随着评分经验的积累,其评分会逐渐改进(Kim, 2015)。虽然历次总体评分的集中趋势不明显,但所有评分员朗读任务的历次评分却均呈现明显的集中趋势,主要集中在分数段 4、5、6。一方面,原因可能是朗读任务的评分标准存在问题。Linacre(2002)指出,如果某分数段的使用频率低于 10 次,那么该分数段需要修改或者与相邻分数段合并。另一方面,评分员评分时可能结合了评分标准以外的参数,比如教学或阅卷经验。以评分员 R1

① *根据 Flesh 易读度参考量表,易读度值越高,篇章难度越低。

为例,该评分员在评阅朗读任务时会根据考生能否读准较难词汇(比如单词 circumstances)来判断其朗读水平是否属于高分数段。问答任务历次的评分都不存在集中趋势,这可能和该题型的计分方式有关,问答任务的答案相对"封闭"(《高等学校英语应用能力考试大纲》修订组, 2016),只需计算考生答对的数量即可,该题型没有翻译或陈述任务"开放"。翻译和陈述任务对分数段 7 的使用频率非常低,这可能和评分员对该分数段描述语的理解偏差有关(杨志强 等, 2016)。通过访谈得知,由于分数段 7 为最高分数段,象征各项任务的最高水平,评分员认为考生的回答需要接近完美才能获得该分数,因此评分员在翻译和陈述任务的评分中对该分数的使用较少。

5 结语

本文采用定量为主,访谈为辅的方法对 PRETCO 口试连续五次的评分进行分析,探讨了评分员的严厉度、评分准确度以及集中趋势三个方面的稳定性及其背后的原因。结果发现:多数评分员历次总体评分的严厉度比较稳定,其中一位评分员某次评分的严厉度波动明显;所有评分员历次评分的准确度均不稳定,但显著波动的次数占比不高;评分员历次总体评分不存在明显的集中趋势,虽然新评分员第一次的总体评分较为集中,但随着该评分员评分经验的不断积累,其评分质量逐渐改进;评分员个别口试任务,比如"朗读"任务的历次评分均呈现集中趋势,且朗读、翻译和陈述三项任务个别分数段使用次数过少,比如陈述任务分数段 7,这些评分标准本身可能存在问题,需要改进。基于此,本研究对计算机辅助口试以及 PRETCO 口试的评分及其改进提出一些参考性的建议。

(1)使用有经验的评分员并保持评分员队伍的稳定性。无论是计算机辅助口试还是 PRETCO 口试,其评分都应尽量使用有教学经验和评分经验的评分员,他们能够结合多方面因素进行综合评分,以保证评分的内部一致性。此外,应保持评分员队伍相对稳定,以增强评分的外部一致性,提高历次评分的信度和稳定性。(2)加强对评分员的培训。如果是新评分员,应充分利用评分培训加强其对评分标准和所评考生总体水平的把握,同时增强新老评分员之间的交流,帮助新评分员改进评分质量。即使有经验的评分员,也有可能出现评分偏差。每次阅卷前,无论是经验丰富的评分员还是新评分员,都需要认真接受培训。另外,在阅卷过程中可以组织阅卷员结合考生的答题情况和评分标准进行讨论,从而加强阅卷员对评分标准的理解。(3)提高口试试题的效度。通过对 PRETCO 口试题目的分析可以看出,试题的难度可能不一致。为了确保历次考试的公平性,需要对平行试题进行质量分析,比如计算朗读任务的易读度,或者通过专家判断以及试测,降低其他口试任务难度的差异。(4)改进评分标准中描述语的质量。评分标准是考试构念的体现,评分标准描述语须简单、明了, 没有歧义(曾用强, 2011)。本研究发现,历次评分中朗读任务第一个分数段、翻译和陈述任务第七个分数段的使用频次极低。鉴于现实评分的需要,不能简单将这些分数段和相邻的分数段合并。因此,有必要对这些分数段的描述语进行改写,以确保评分员理解的准确性和一致性,防止出现理解偏差(杨志强 等, 2016)。

参考文献:

Attali, Y. 2016. A Comparison of Newly-trainedand Experienced Raters on Astandardized Witing Asessment [J]. *Language Testing*(1): 99-115.

Bond, T. & C. Fox. 2015. *Applying the Rasch Model*: *Fundamental Measurement in the Human Sciences* (3*nd Edition*)[M]. New York;LondonRoutledge.

Bonk, J. & G. Ockey. 2003. A Many-facet Rasch Analysis of the Second Language Group Oral Discussion Task[J]. *Language Testing*(1):89-110.

Dobria, L. 2011. *Longitudinal Rater Modeling with Splines*. Unpublished Doctoral Dissertation, Chicago:University of Illinois at Chicago.

Elder, C., Knoch, U., Barkhuizen, G. & J. Von Randow. 2005. Individual feedback to enhance rater training: Does it work? [J]. *Language Assessment Quarterly*(3):175-196.

Frost, K., Clothier, J., Huisman, A. & G. Wigglesworth. 2020. Responding to a TOFEL iBT Integrated Speaking Task: Mapping Task Demands and Test Takers' Use of Stimulus Content[J]. *Language Testing*(1), 133-155.

Fulcher, G. 1996. Does Thick Description Lead to Smart Tests? A Data-based Approach to Rating Scale Construction[J]. *Language Testing*(2): 208-238.

Given, L. M. 2008. *The SAGE Encyclopedia of Qualitative Research Methods* (Volume 1 & 2)[M]. London: Sage Publications Ltd.

Kang, O., Rubin, D. & A. Kermad. 2019. The Effect of Training and Rater Differences on Oral Proficiency Assessment[J]. *Language Testing*(4): 481-504.

Khabbazbashi, N. & E. D. Galaczi. 2020. A Comparison of Holistic, Analytic, and Part Marking Models in Speaking Assessment [J]. *Language Testing*(3): 333-360.

Kim, H. J. 2015. A qualitative Analysis of Rater Behavior on an L2 Speaking Assessment[J]. *Language Assessment Quarterly* (12): 239-261.

Knoch, U. & C. A. Chapelle. 2018. Validation of Rating Processes Within an Argument-based Framework[J]. *Language Testing* (4): 477-499.

Linacre, M. 2002. Optimizing Rating Scale Category Effectiveness[J]. *Journal of Applied Measurement*(1): 85-106.

Linacre, M. 2013. *A User's Guide to FACETS: Rasch-Model Computer Program* (Computer Program Manual)[M]. Chicago: MESA Press.

Lumley, T. & T. F. McNamara. 1995. Rater Characteristics and Rater Bias: Implications for Training[J]. *Language Testing* (1): 54-71.

Luoma, S. 2004. *Assessing Speaking*[M]. Cambridge: CUP.

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.

Myford, C. M. & E. W. Wolfe. 2009. Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use[J]. *Journal of Educational Measurement*(4):371-389.

Nakatsuhara, F. 2011. Effects of Test-taker Characteristics and the Number of Participants in Group Oral Tests[J]. *Language Testing*(4): 483-508.

Nitta, R. & F. A. Nakatsuhara. 2014. Multifaceted Approach to Investigating Pre-task Planning Effects on Paired Oral Performance[J]. *Language Testing*(2): 147-75.

Wind, S. A. & M. E. Peterson. 2018. A Systematic Review of Methods for Evaluating Rating Quality in Language Assessment [J]. *Language Testing*(2): 161-192.

《高等学校英语应用能力考试大纲》修订组. 2016. 高等学校英语应用能力考试(口试)大纲和样题[Z](2 版). 北京:高等教育出版社.

何莲珍, 张洁. 2008. 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(4): 388-398.

金力. 2011. 计算机辅助大学英语口语测试研究[J]. 外国语文(4): 126-130.

刘鸿章, 孔庆炎,陈永捷. 2010. 高等学校英语应用能力考试十年回顾与展望[J]. 中国外语(4):12-15.

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.

杨志强, 全冬. 2016. PRETCO 口试评分标准效度验证[J]. 外语测试与教学(1): 13-21,31.

曾用强. 2011. 计算机辅助英语口语考试研究[M]. 北京:科学出版社.

赵海燕, 辛涛,田伟. 2018. 主观题评分中的评分者漂移及其传统检测方法[J]. 中国考试(8): 20-27.

A Longitudinal Study of the Rating Stability of Computer Assisted PRETCO – Oral

YANG Zhiqiang　LI Zhifang　DONG Manxia

Abstract: The stability of raters' rating of an oral test might pose a threat to its reliability, validity and fairness. This paper investigates raters' stability of scoring PRETCO – Oral longitudinally on the basis of the recent five times' rating results from perspectives of severity, accuracy and centrality. Interview is also employed to shed some light on the reasons of raters' ratings.

Key words: computer assisted oral test; PRETCO – Oral; rating; stability

责任编校:路小明