网络双评过程中作文评分误差以及评分者效应的分析

——以大规模英语考试作文评分为例

■ 李美娟　刘红云

摘要：目前大规模考试作文评分大都采用双评评分模式，本研究采用多侧面Rasch模型（MFRM）分析双评模式下大型英语作文评分中的评分者误差来源及主要影响因素。对57名评分者所评价的2 427篇作文分析发现：①评分者的宽严度存在显著的差异；②在作文评分中，约有22.8%的评分者之间的一致性较差，也存在约3.5%的评分者之间一致性过高；③约90%的评分者自身的一致性都较高，但仍有8.8%的评分者自身一致性很差，约2%的评分者出现评分自身一致性过高的情况；④从整体上讲，评分者在不同的评分标准（或维度）上、不同评分等级宽严程度的把握存在差异；评分者和被试，以及评分者、被试和评分标准三者的交互作用不显著；⑤评分者对男生和女生具有相同的宽严度。

关键词：主观题评分；多侧面Rasch模型；评分者误差分析

【中图分类号】G405　【文献标识码】A　【文章编号】1005-8427(2015)02-0039-10

1　问题提出

1.1　英语写作评分概述

主观题是大型考试必不可少的题型，其可以测量被试对某个概念的理解、掌握以及应用概念解决问题的过程，而不是简单的再认或者猜测[1]，但被试在主观题上的得分往往通过评分者依据既定的标准进行评定，由于每个评分者不同的个性特点可能会使得评分存在误差，从而影响考试结果的客观和公平。这种现象也就是所谓的评分者效应（Rater Effect）。评分者效应，如评分的准确性（Accuracy / Inaccuracy）、严厉度（Harshness / Leniency）和集中度（Centrality / Extremism）常常被认为是评分者评分的系统变异[2][3]。换句话讲，评分者效应与评定量尺的结构无关，与被试的能力无关，但却会影响评定的效度[4]。

以表现性为基础的语言评价已经越来越引起研究者们的注意，这种考试要求被试在现实生活中使用学到的知识和技能[5]，不论是GRE，TOEFL，还是国内四六级英语考试，或者初高中或者大学的升学考试，作文题是考查学生写作能力高低普遍采用的题型，但是评分过程必然会出现评分者效应。被试在特定写作任务的获得高分的概率不仅决定于被试的写作能力，而且可能受以下几个因素的影响①写作任务的特征；②评分者的个人特质；③评分时的特定情境；④学生的背景特征以及对写作任务

本文为北京市教育科学“十二五”规划2012年度青年专项课题“学业水平测验认知诊断功能的应用研究”（CHA12109）成果之一。

[作者简介]　李美娟，女，北京教育科学研究院，助理研究员（北京　100191）
刘红云，女，北京师范大学，教授（北京　100875）

或者问题的兴趣；⑤不同影响因素的交互作用。这些因素引起的评分变异就是评分者效应，所以有很多研究则关注评分者效应对评定结果的影响以及对于校正评分者效应的统计模型的发展。一些研究者强调原始分数和概化理论的应用，而更多的研究者现在则关注潜在特质模型的应用[6]。除了在分析方法和模型上的考虑外，在评分过程中目前大规模测试往往采用双评的评分模式以降低评分者带来的影响，然而双评的评分模式并没有从本质上消除评分者效应，同时又可能会带来评分趋中等一些负面效应。

1.2 评分效应模型

评分者效应的测量模型有三类，基于经典测量理论（CTT）基础上的评分效应模型，基于概化理论（GT）的评分效应模型和基于IRT的评分效应模型。

1.2.1 基于CTT和GT的评分效应模型

CTT将被试得分的总变异分解为真分数的变异和随机误差的变异。分析评分者效应时，通常只能通过计算评分者之间的一致性来进行判定。其中数据收集的设计方式不同，误差和信度系数的定义则不同。CTT最大的缺点就是题目和测验指标的样本依赖性，很难预测被试对特定的题目的作答情况，对不同形式评价的比较很困难，也没有办法计算测验结构不同水平的测量误差的变化[6]。另外，CTT下的很多的统计模型假设测量具有等距性。但是，许多里克特量表或者成就测验的原始分数都是顺序的，导致被试能力或者题目难度的有效比较就变得很困难。

与CTT相比，Cronbach等人的GT将方差分析的技术引入信度的研究，允许对不同来源的误差进行深入分析，如题目本身，评分者误差等。GT将这些被视为误差的无关变量引入测量模型，用统计方法分别估计出这些因素或者因素之间的交互作用对测验分数的影响。虽然GT中G研究中得到的方差变异可以用来优化未来的实验设计，而D研究可以将各种误差的变异最小化。但是GT对原始分数进行分析，并没有从根本上解决测验结果的样本依赖性，另外GT也假设测量具有等距性。CTT和GT在考虑题目难度、测量误差指标时，并没考虑到被试能力的差异，所以它们对于信度的估计都可被描述为在所有能力水平上的一个平均信度。

1.2.2 基于IRT的多面Rasch模型

基于IRT的评分者效应模型主要有多面Rasch模型、对多面Rasch校正的阶层评分者模型（Hierarchical Rater Model，HRM[7]）和评委束模型（Rater Bundle Model，RBM[8]）。其中应用最广泛和最简洁的是多侧面测量模型（MFRM），其他两种模型仅限于理论上的探讨，在实际中上没有广泛应用。

MFRM是在以IRT为基础的单参数Rasch模型的基础上发展来的，以下是用来分析写作任务的多侧面Rasch模型：

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right]=\theta_n-\beta_i-\alpha_j-\tau_k$$

其中，p_{nijk} 代表评分者j在评分标准i上对被试的作文评定为k等级的概率；p_{nijk-1} 代表评分者j在评分标准i上对被试的作文评定为$k-1$等级的概率；θ_n 代表被试的写作能力；β_i 表示评分标准的难度；α_j 表示评分者j的宽严度；τ_k 表示评定量表模型（Rating Scale Model）或者分步记分模型（Partial Credit Model）中被试得分从等级$k-1$到k的等级难度（step difficulty）[9]。MFRM是一个潜在特质模型，是在两侧面Rasch模型的被试侧面和项目侧面的基础上增加了评分者侧面，并将评定的观测值转化为logits值。评定量尺模型假设每个题目或者标准的评定量尺是相同的，而分步记分模型（PCM）则假设每个题目或者标准均可以有自己独立的评分结构。

Weigle（1998）研究发现，没有经验的评分者可能会更严格或者缺乏评分的一致性[10]。但是，基于评分表现的Rasch分析和经验反馈又可能使评分更加一致。总之，基于个体特征的评分者差异是很难

通过一般背景变量预测的，精细的挑选和高强度的培训不足以使评分者的评分等价。对评分者个体特征的统计上的校正可以确保分数的可靠性[11]。MFRM从统计上对被试的能力进行校正，整个连续体测量单位相同，模型中的参数估计独立，即被试的能力估计与题目的特征、评分者的评价没有关系，题目的难度和评分者的宽严度与数据收集设计的其他面的分布特征没有关系。由于logits具有可加性，不同侧面的logits可以进行比较。该模型一方面可以提高主观评分的信度，另一方面还可以提供给评分者更多的信息。

1.3 研究目的

目前大规模考试中的作文评分大多采用网络双评的技术，在评分过程中通过一定的技术手段控制评分者之间的一致性以及评分者内部的一致性[11]。本研究通过对某大型英语考试作文评价中的评分者效应分析，考查目前网络双评过程中影响评分者效度的因素，并为更好地估计被试的写作能力提供一定的方法，进一步拓展国内使用MFRM在主观题分析应用的范围，就作文评分培训以及网络双评过程中应该关注的问题提供一些建议。

2 研究方法

2.1 样本描述性统计

随机抽取参加某大型英语写作考试的2 427名学生，其中男生936名（38.6%），女生1 491名（61.4%）。这些学生的作文由57个评分员随机阅评，每篇作文随机由2名评分者评分，评分者评价作文的篇数为8~601篇。

2.2 作文评分过程及评定量表

某大型英语水平测试的写作考试中，要求所有学生根据同一题目写一篇文章。作文完成后，不作任何改动扫描到电脑中，将每篇作文随机分给2名评分员，该作文的总分为11分。

对于该英语作文的评分，评分者将从内容（information），语言（gracture），结构（mechanics），长度（length），和连贯性（coherence）五个维度分别对每篇作文进行评价，其中，对内容和语言的评价采用4点计分量表，对结构采用3点计分量表进行评价，对长度的评价采用二级计分量表，文章连贯性则采用3点计分量表进行评价。

2.3 分析方法

由于本研究中5个评定量表具有各自的评定标准，所以本研究采用的是以分步计分模型（PCM）为基础的三侧面多面Rasch模型，其中三个侧面分别是被试的英语写作能力，评分者的宽严度和评分标准的难度。

本研究采用Facets3.62.0[12]软件，该程序采用无条件极大似然估计法（Unconditional Maximum Likelihood）对MFRM中的各个参数进行估计，其中每个侧面都进行了校正，并且每个侧面的分析都是与其他侧面独立的。为了建立logit量尺的原点，以及使模型得到识别，本研究将评分者侧面，评分标准侧面的均值固定为0，另外一个模型识别的限制条件就是评分等级的系数之和限定为0，并按照惯例，被试侧面非中心化。

3 分析结果

3.1 整体拟合指标

根据Linacre & Wright（2008）的标准，较好模型拟合的条件是大于等于2的（绝对）标准化残差不超过5%，大于等于3的（绝对）标准化残差不超过1%。本研究数据结果显示，大于等于2的（绝对）标准化残差占5%，大于等于3的（绝对）标准化残差占1.2%[12]。

模型假设残差为正态分布，当模型的参数得到精确的估计时，残差的均值为0。如果数据与模型越拟合，标准化残差的均值越接近0，样本标准差的期望值也越接近1.0，数据结果显示，标准化残差的均值为0，样本标准差的期望值为1，并且模型的卡

方值不显著(固定卡方=24844.3, p>0.05)。总之,从以上几种指标来看,模型的整体拟合较好。

另外,估计得到的原始分数的误差是(0.4/1.0)2×100%=16%,说明模型所能解释变异比例是(1-0.16)×100%=84% 。

3.2 对被试参数,评分者参数,和评分标准的校正

图1将评分标准,评分标准的等级难度和被试能力以及评分者的宽严度放在同一个量尺上进行比较,从而为评分标准的调整和改进提供依据。

图1中,第一栏为被试的写作能力分布,从上到下表示被试能力从高到低;第二栏为评分者的宽严程度,从上向下表示评分者的宽严程度从严到宽;从图中可以看出,评分者宽严度的测量全距为5.33logits,占了被试写作能力17.1logits测量全距的1/3,评分者之间的一致性影响了评分者的评价信度。第三栏为评分标准的难度分布,从上向下表示评分标准的难度从难到易,被试在连贯性上获得高分较难,在内容上获得高分较容易;第四栏到第八列分别为四个评定标准(内容,语言,结构,连贯性)的等级难度,从上到下表示等级难度从难到易,图中的横线代表(等级+0.5)的logits值,在最左侧纵向的一列是这几个统计量共用的logits量尺,可以将不同统计量进行比较。

3.2.1 评分者测量

图1表明评分者之间的一致性较差。评分者的logits测量值是-1.99~3.34,均值为0.00,标准误为0.16,其中分离指标为6.27,表明评分者的变异是测量误差的6.27倍,分离信度为0.98,固定的卡方值3989.4(df=56,p<0.01),表明评分者之间的一致性很差,其中433号评分者严厉度水平最高,416号和441号评分者的严厉度水平最低。

Linacre(2002)建议,把0.5和1.5作为infit和outfit的下限和上限[13],其他研究者则建议使用更加严格的范围,以0.7(或者0.75)和1.3作为其上限和下限[14]。本研究以0.5和1.5作为infit和outfit的下限和上限,从整体上来看,7%的评分者的infit小于0.5,这些评分者自身的一致性过好,分别是评分者431,465,459,468;8.8%的评分者的infit大于1.5,这些评分者自身的一致性过差,5个评分者分别是438,473,460,421,415;8.8%的评分者的outfit小于

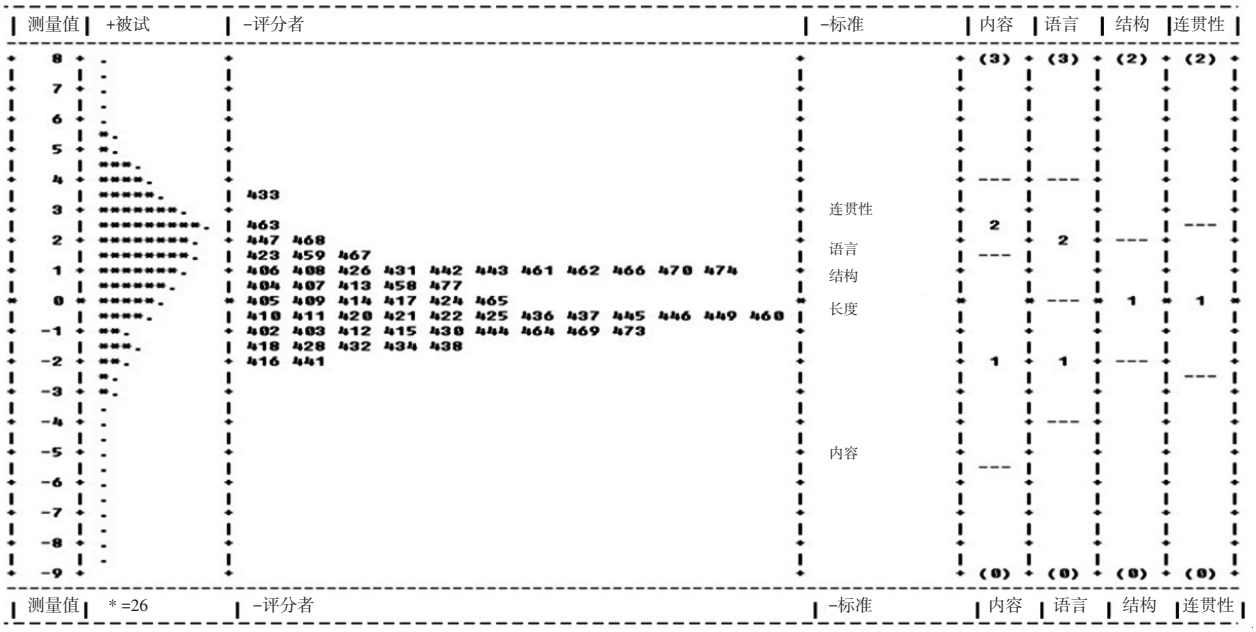

图1 被试参数、评分者参数和评分标准参数的校正图

0.5，说明这些评分者出现了评分者趋同现象，分别是436，445，459，431，466，467，444；22.8%的outfit大于1.5，表明评分者之间的一致性很差，这些评分者分别是442，438，411，421，460，473，402，417，410，441，415，428，416。

3.2.2 评分标准的测量结果

表2中列出了每个标准的logits测量值，5个标准之间的差异均显著（t检验），说明5个标准的难度是不同的，具体来讲，被试在连贯性量表获得高分的难度要大，在内容量表获得高分的难度要小。文章信息的outfit大于2.0，说明5个评分标准并不是单维，推翻了Rasch模型的假设。从固定的卡方值显著（固定卡方=22901.5，p<0.01）、分离信度为1.00，可以看出五个标准的难度是不一致的，分离指标79.79说明评分标准的变异是测量误差的79.79倍。

表3列出了每个评分标准的测量情况，其中包括模型分析的等级分数频数，频数所占百分比，实际的平均测量值，期望测量值，outfit值，以及PCM模型为基础的等级难度参数估计值、标准误。检测评定量表的效度一个重要的指标就是评定等级的平均测量值。这个值是特定评定等级条件下的被试能力测量值的均值。随着评定等级越高，平均测量值越大。另一个重要的指标是每个评定等级下的outfit指数，其将被试能力的平均测量值与Rasch模型估计的期望值进行比较，两者差异越大，outfit的值越大，往往这个值不能大于2.0。评定量尺效度还可以通过等级难度的排序来判断，等级难度应该随着评定等级的增加而增加。并且等级难度的差异值应该大于1.4小于5个logits。当k+1到k的等级难度大于k到k−1等级难度1.4以上logits时，从理论上讲，可以将评定量表视为独立的二级计分题目，也就是说，评定等级具有较好的区分度[15]。如果相邻的两个等级难度的差值大于5个logits，这种情况可能导致信息丢失，这时需要更多的等级[15]。

从表3可以看出，用于分析内容评定量表每个等级分数的频数分布差异较大，并且等级分数0，1，2的outfit值均大于2，并且0到1的等级难度值与1到2的等级难度的差值远远大于5，因此，该量表的质量较差，测量效度也很低。将标准1的0，1，2三

表1 评分者拟合指数概况

拟合指数范围	Infit	Outfit	拟合指数范围	Infit	Outfit
窄			宽		
fit <0.70（overfit）	5(8.8%)	9(15.8%)	fit< 0.50（overfit）	4(7.0%)	7(12.3%)
0.70≤fit≤1.30	45(78.9%)	32(56.1%)	0.50 ≤ fit ≤ 1.50	48(84.2%)	37(64.9%)
fit>1.30（misfit）	7(12.3%)	16(28.1%)	fit > 1.50（misfit）	5(8.8%)	13(22.8%)

表2 评定量表的整体测量结果

统计指标	内容	语言	结构	长度	连贯性
测量值	−4.83	2.06	0.31	0.15	2.31
标准误	0.04	0.03	0.03	0.03	0.04
Infit	1.41	1.05	0.93	0.86	0.91
Outfit	3.08	1.03	0.87	0.85	0.84
固定χ^2	22901.5**				
自由度	4				
分离指标	79.79				
分离信度	1.00				

个评分等级进行合并，从表3可以看出，等级分数0的outfit大于2，说明这个等级的划分还是存在问题。另外，图2~图6表示每个评分标准的概率曲线，从表3和图2~图6可以看出，语言、结构、长度和连贯性量表质量较好，评定效度较高。

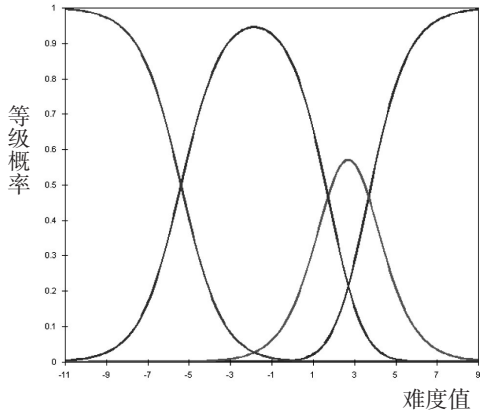

图2 内容量表的概率曲线

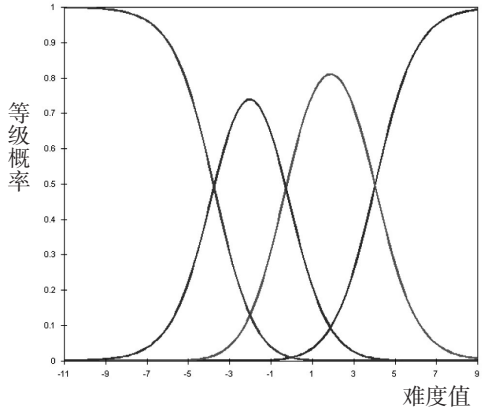

图3 语言量表的概率曲线

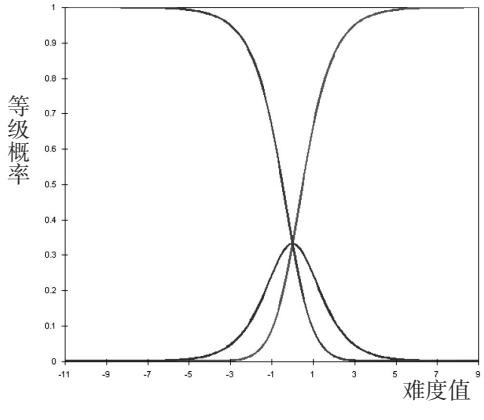

图4 结构量表的概率曲线

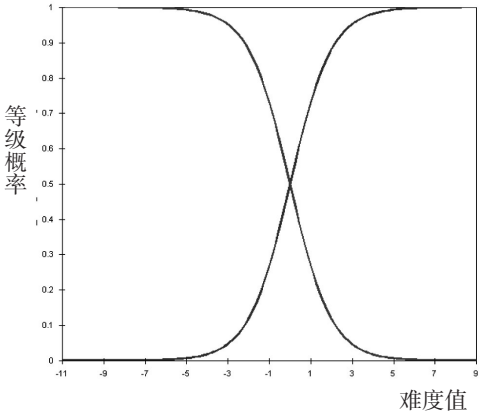

图5 长度量表的概率曲线

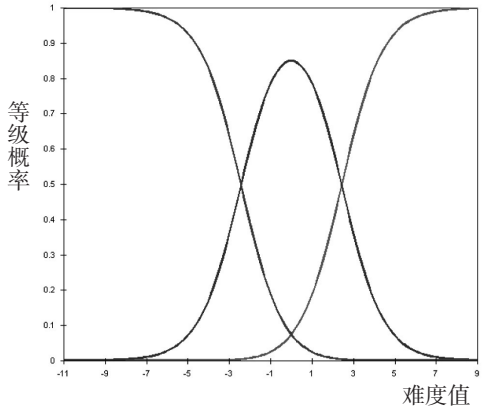

图6 连贯性量表的概率曲线

3.3 交互作用分析

3.3.1 评分者和被试的交互作用

从整体上来讲，被试和评分者的交互作用不显著（固定卡方=3424，$p>0.05$）。从个体水平上来讲，仍有30个评分者和被试存在交互作用，从图7中可以看出，评分者对被试能力的高估或者低估与被试能力的分布是没有关系的。

3.3.2 评分者和评分标准的交互作用

从表4中可以看出，固定的卡方值显著，从整体上讲，评分者和评分标准的交互作用显著，从个体水平上讲，有124（43.5%）的观测值和期望值存在显著的差异显著（$t\geq2$ 和 $t\leq2$），其中58个差异值是正的，即评分者评分较严厉，66个差异值是负的，即评分者评分较宽松。对于评分者来讲，评分者具有的

表3 评定量表的测量情况

评定量表	分数	频数	百分比	平均测量值	期望测量值	Outfit	等级难度参数	标准误
内容	0	3	0	−4.58	−2.20	3.6		
	1	281	6	1.56*	1.10	4.6	−5.39	0.59
	2	710	15	4.16	3.90	3.2	1.72	0.09
	3	3814	79	7.00	7.09	1.2	3.68	0.05
语言	0	695	14	−4.53	−4.36	0.8		
	1	1885	39	−1.46	−1.43	0.8	−3.75	0.06
	2	1950	41	1.02	0.92	0.9	−0.28	0.04
	3	278	6	3.20	3.20	1.0	4.03	0.07
结构	0	752	16	−2.64	−2.64	0.8		
	1	2077	43	0.7	0.70	1.0	−1.85	0.06
	2	1979	41	3.00	3.00	1.2	1.85	0.04
长度	0	1571	33	−1.21	−1.12	0.9		
	1	3237	67	2.5	2.50	0.7		
连贯性	0	1408	29	−3.57	−3.48	0.9		
	1	2737	57	−0.26	−0.29	0.8	−2.43	0.04
	2	663	14	2.31	2.27	1.0	2.43	0.05
内容(修订)	0	820	18	0.45	−0.04	3.3		
	1	3814	82	3.27	3.37	1.5		

表4 交互作用的整体测量情况

	交互作用		
	被试*评分者	评分标准*评分者	被试*评分标准*评分者
N	4808	285	24040
显著*t*值%	107(2.22%)	124(43.5%)	279(1.16%)
*t*最大值	3.25	12.60	3.06
*t*最小值	−3.09	−7.70	−6.05
M	−0.02	−0.02	−0.07
SD	1.09	1.06	1.29
固定卡方	3424.0	2327.7**	8895.3
Df	4808	285	24040

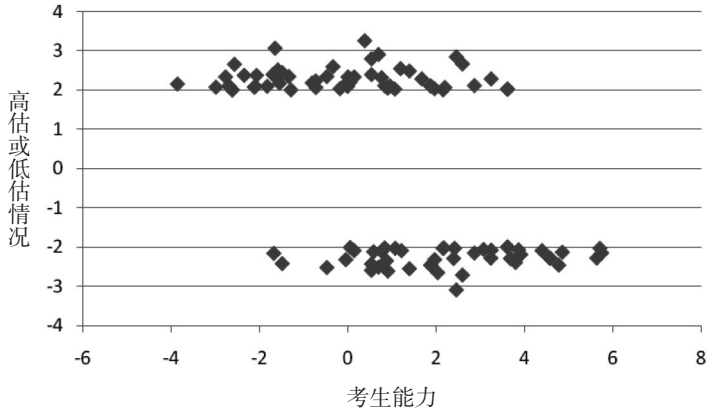

图7 评分者对被试能力的高估或者低估的散点图

差异值个数的范围是1~5,其中评分者412,433,415具有5个显著的差异值,评分者406,407,415,416,424,434,443具有4个显著的差异值。评分者对四个评分标准严厉或者宽松的程度相当,信息(14 vs. 12),gracture(8 vs. 9),文章结构(15 vs. 17),文章连贯性(12 vs. 13),而对长度(9 vs. 15),则表现为评分宽松的评分者较多。

3.3.3 评分者,被试和评分标准的交互作用

表4中,固定的卡方值不显著,从整体上来讲,被试,评分者和评分标准的交互作用不显著,从个体水平上来讲,43个评分者和被试,评分标准存在交互作用。279个显著差异值中,内容最多(162),连贯性(20),语言(49),长度(7),结构(46)。由于个体水平上被试,评分者,和评分标准的详细结果过于烦琐,本研究将不再详述。

3.4 性别偏差分析

从表5中可以看出,性别分离指数为1,分离信度为0,说明评分者对于男生和女生的严厉度并不存在偏差。

表5 性别偏差的测量结果

	统计指标
组水平	
女	
M	0.00
SE	0.02
男	
M	0.00
SE	0.02
卡方(df)	0(1)
性别分离指数	1.00
分离信度	0
个体水平	
残差分析	
显著t值%	0
卡方(df)	44.4(114)
配对分析	
显著t值%	0

4 讨论

表现性评价中关于评分者行为的许多研究都很关注评分者变异,这些变异与评分者自身的特征有关,与被试的表现无关[3]。本研究中评分者之间的宽严度存在显著差异,约22.8%的评分者评分之间的一致性较差,约3.5%的评分者评分一致性过高,出现评分趋同现象;影响评分者宽严度的因素有很多,例如,是否受过专业的培训,个人的特质,态度,背景变量特征,工作压力以及评价的目的等等。评分者自身的不一致也是影响作文评分信度和效度的重要因素,评分者内部一致性的分析结果表明约90%的评分者自身的一致性均较高,但仍有8.8%的评分者自身一致性很差,约2%的评分者出现评分自身一致性过高的情况;评分者可能认为她/他必须为其他评分者提供标准,对被试写作能力的评价波动较小,导致"评分者自身一致性过高",而有些评分者未受过培训,喜欢质疑被试,尤其在两个被试能力相近的时候,导致"评分者自身一致性过差"。另外,评分者对评分标准理解的不同导致被试在不同评分标准上获得高分的难易程度不同,被试在连贯性量表获得高分相对较难,在内容量表获得相对容易。

MFRM允许进行偏差的分析,即模型中多个侧面交互作用的分析,例如,评分者和被试、评分者和评分标准、以及评分者、被试、评分标准三者的交互作用。本研究中,评分者对被试能力的高估或者低估与被试能力的分布是没有关系的,也就是说,评分者并没有低估高能力被试,高估低能力被试;Thomas Eckes(2008)使用two-mode聚类分析技术对评分者和评分标准进行联合分类,结果发现,6种不同类型的评分者具有不同的评分准则[16],具体来讲,评分者并不会将注意平均放在每个评分标准上,并且评分者的背景变量可以部分解释评分准则

之间的差异，本研究中，评分者对信息、语言、文章结构、文章连贯性四个评分标准严格或者宽松的程度相当，而对于长度，较多的评分者评分宽松。

许多研究发现，偏差分析对于评分者的培训很有意义，评分者不同的宽严度，独特的反应模式，评分者之间的一致性，评定的效度是通过培训可以提高的[10]。Myford 和 Wolfe 提出，MRFM 可以提供每个侧面，每个评分者的评分模式，这样就可以帮助我们提供给评分者更加精确的个人反馈，并且帮助他们了解如何使用评定量表，进而提高评分的效度[17]。但是也有研究发现评分者培训并不能按照预期有效减少评分者的变异，即使经过大量的培训[18]或者提供个体评分的反馈，评分者的变异并不会明显减少[19]，因此不仅要通过对评分者的培训提高评分信度，还要通过统计模型对评分者效应进行校正，进而实现评分的客观性。

本研究对如何改善和提高英语写作评分的效度有着十分重要的意义。首先，可以通过对评分量表的修改进一步提高测量工具的效度，其次，考察单个评分者的表现情况有利于评分者的选拔，以及对评分不准确的评分者进行进一步的培训或者替换，另外，具有高风险的大规模测试，可以通过模型对评分者，评分标准进行校正，确保考试的客观性和公平性。

参考文献

[1] 李中权，孙晓敏，张厚粲，张立松. 多面 Rasch 模型在主观题评分培训中的应用[J]. 中国考试(研究版)，2008(1)：26-31.

[2] Myford. C. M., Wolfe. E. W. Detecting and measuring rater effects using many-facet Rasch measurement: part I[J]. Journal of Applied Measurement, 2003, 4(4)：386.

[3] Eckes T. Examining rater effects in TestDaF writing and speaking performance assessments: A many- facet Rasch analysis[J]. Language Assessment Quarterly, 2005, 2(3)：197-221.

[4] Bachman L. F. Statistical analyses for language assessment[M]. 2004, Cambridge Univ Pr.

[5] Kondo-Brown, K. A FACETS analysis of rater bias in measuring Japanese second language writing performance[J]. Language Testing, 2002, 19(1)：3.

[6] Sudweeks. R. R, Reeve. S., Bradshaw. W. S. A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing[J]. Assessing Writing, 2004, 9(3)：239-261.

[7] Patz. R. J., Junker. B. W., Johnson, M. S., Mariano, L. T. The hierarchical rater model for rated test items and its application to large-scale educational assessment data[J]. Journal of Educational and Behavioral Statistics, 2002, 27(4)：341.

[8] Wilson. M., Hoskens. M. The rater bundle model[J]. Journal of Educational and Behavioral Statistics, 2001, 26(3)：283.

[9] Eckes, T. Many-facet Rasch measurement. Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment[C]. Strasbourg: Council of Europe, Language Policy Division,2009.

[10] Weigle. S. C. Using FACETS to model rater training effects[J]. Language Testing, 1998, 15(2)：263.

[11] Lunz, M. E., Wright, B. D., Linacre, J. M. Measuring the impact of judge severity on examination scores[J]. Applied Measurement in Education, 1990, 3(4)：331-345.

[12] Linacre, J. M., Wright, B. D. A user's guide to FACETS: Rasch-model computer programs[C].Computer software manual. Retrieved May 28, 2008, from http://www. winsteps. com/aftp/facets. pdf.

[13] Linacre, J. M. What do infit and outfit, mean-square and standardized mean[J]. Rasch Measurement Transactions, 2002, 16(2)：878.

[14] Bond, T. G., Fox, C. M. Applying the Rasch model: Fundamental measurement in the human sciences[J]. Lawrence Erlbaum，2007.

[15] Linacre. J. M. Investigating judge local independence[J]. Rasch Measurement Transactions, 1997, 11(1)：546-547.

[16] Eckes, T. Rater types in writing performance assessments: a classification approach to rater variability[J]. Language Testing, 2008, 25(2)：155.

[17] Myford, C. M, Wolfe. E. W. Detecting and measuring rater effects using many-facet 17 Rasch measurement: Part II [J]. Journal of Applied Measurement, 2004(5)：189-227.

[18] Barrett, S. The impact of training on rater variability[J]. International Education Journal, 2001, 2(1)：49-58.

[19] Elder, C., Knoch.U., Barkhuizen. G., Von Randow. J. Individual Feedback to Enhance Rater Training: Does It Work?[J]. Language Assessment Quarterly, 2005, 2(3)：175-196.

Error Analysis of Inter-rater Reliability of the Second Language Writing Performance Assessment

LI Meijuan & LIU Hongyun

Abstract: This research would investigate the extent to which that second language writing performance scores were influenced by rater effect in large scale assessment in China. Writing samples were obtained from 2427 (1491 females, 936 males) first grade students in Junior high school. The 54 raters in this study were all experienced specialists in the field of Teaching English as the second language. Each examinee was randomly scored by two raters. Each writing sample was scored according to five criterion:①Information, a 4-point scale was use to measure content; ②Gracture, which is a 4-point scale used to evaluate the sentence; ③ Mechanics, a 3-point scale is for the overall structure; ④Length, a 2-point scale used to measure the number of words; and ⑤ Coherence, a 3-point scale u the expression. The MFRM analysis was completed using Facets software. Three facets were analyzed including persons, raters, and rating criteria based on Partial credit Model. The findings in this study indicated that ①Raters differed in severity or leniency. ②Some raters could not follow the rating scale consistently, while others could not stay close to their own scoring standard. ③Raters could be able to maintain an constant level of severity across all the examinees, but not to all five criteria. ④There was no differential rater functioning related to the gender of examinees, which also means that the raters maintained a consistent severity or leniency across male and female examinees. MFRM study had a number of implications for rating issues in L2 writing assessment. Individual feedbacks can improve the efficiency of rater training to ensure objectivity and fairness of the writing performance assessment.

Keywords: Writing Performance Assessment; MFRM; Rater Effect