学业水平测试中作文评分误差的多面 Rasch 分析*

刘红云[1] 陈 阅[2] 骆 方**[1] 王云峰[3]

([1]北京师范大学心理学院,应用实验心理北京市重点实验室,北京,100875)

([2]中科院心理研究所,北京,100192)([3]北京教育科学研究院,北京,100031)

摘 要 使用多面 Rasch 模型,从评分量表、评分员等层面对参与 2007 年八年级语文学业水平测试作文评分的 17 名评分员的评分情况进行了研究。结果发现:(1)评分员的评分等级所对应的能力值呈正常的变化趋势,大部分评分员有较好的内部一致性;(2)不同评分员的宽严程度有显著差异,评分员之间的一致性整体较好;(3)此外,本文还就评分内部一致性较差的几个评分员的评分做了进一步研究。

关键词 作文 评分误差 项目反应理论 多面 Rasch 模型

1 问题提出

教育测量中,按试题评分方式可将试题分为客观题和主观题两类。客观题有评分客观、公平等优势,但是难以全面考查考生的能力。近年来主观题在测验中的地位越来越受到重视。虽然主观题对考生能力的测评比较全面、真实,但其评分的客观性和有效性却是一直困扰着人们的问题。写作就是一种受到众多关注又常引起很大争议的主观题。

写作能力在语言测试中非常重要。许多大规模的语言测试都把写作作为一个重要部分。在写作测试中,考生的表现由经验丰富的评分员根据所设定的标准来评定。评分员的经验为评分的有效性奠定了基础,但每个评分员都有其独特的特点,在对考生的评定中难免掺杂个人因素的影响[1]。研究发现:考生的得分是考生能力和评分员特点的函数,并不仅仅取决于其真实的能力水平[2]。主观题评分中评分员偏差的存在对客观、准确地评定考生水平产生了较大的影响[3]。为了提高主观题评分的信度,研究者在评分过程中引入了许多措施。如构建结构化的测验项目,使用标准化的评定准则和施测程序,对评分员进行充分的培训等[4]。虽然这些措施在一定程度上可减小评分员偏差,但其效果十分有限,比如无论如何进行事前培训,评分员也无法在评分宽严程度上保持一致,即评分员偏差的影响依然存在[5]。因而,近年来研究者们开始尝试通过现代测量理论所提供的方法来分析和控制主观评分中的误差。

多面 Rasch 模型(Many－Facets Rasch Model, MFRM)是项目反应理论(Item Response Theory, IRT)模型中 Rasch 模型的延伸。该模型可对考生能力、评分员宽严程度、项目难度等进行估计,常用在主观题的分析中,以识别和消除评分员偏差,提高主观评分的信度[6][7]。本研究拟通过对语文学业水平测试中作文评分进行多面 Rasch 分析,来考察造成作文评分误差的原因,探讨评分标准对评分准确性的影响等。研究结果可以帮助我们设想出更有效的方法来减小误差,使作文评改的流程与方式更加科学化,使作文评分更加合理、公平。

2 研究方法

2.1 评分数据

2007 年八年级语文学业水平测试中,要求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,评分员在内容选择(12 分)、语言表达(12 分)、结构安排(8 分)、书写和标点(6 分)四个维度上对作文进行评分。随机抽取考生作文 3679 份,由 17 个评分员完成阅评。按评分标准中对分数等级划分的说明将考生在各评分维度上的原始得分转化为 4 个等级。

3 结果分析

3.1 作文分项评分的多面 Rasch 模型分析

使用 Conquest 2.0 对数据进行分析。分别考察了各维度评分之间独立和相关两种模型。结果显示维度评分相关模型中大部分未加权的拟合指数介于 0.8 到 1.2 之间,表明数据与模型拟合理想,可以据此模型结果进行以下研究[8]。

评分量表的使用偏差

首先对评分量表进行评估,即检查是否所有的评分员都使用了量表的所有分数段,是否每个分数段都体现了考生相应的能力[9]。对评分维度的参数而言,加权拟合指数参数(MNSQ)大于 1.30 或小于 0.77 表明可能存在等级限制误差,即某些评分等级被过多或者过少使用。表 1 显示,“书写和标点”评分维度的加权拟合指数为 1.71,大于 1.3,表明在“书写和标点”的评分中,评分员不能有效区分考生能力。分离信度＝1.000,$\chi^2=16555.25$,$df=3$,$p<0.05$,即不同评分维度的难度差异显著,尤其是“书写和标点”难度较低。

3.1.2 评分员宽严程度

评分员的宽严程度,是指评分员在评分时稳定地过高或过低评分的趋势[10]。表 2 列出了评分员宽严程度的估计值及其误差。结果表明编号为 17 的评分员评分最松(－1.079),9 号评分员最严(0.889)。14、7、5、6 号评分稳定,而 17 号评分最不稳定(0.236)。$\chi^2(16)=1118.68$,$p<0.001$,分离信度为 0.979,表明评分员宽严程度有显著差异。此外,结果表明评分员的评分等级所对应的能力值总体呈现正常的变化趋势。

3.1.3 评分员内部一致性

评分员内部一致性考察的是评分员在对不同考生的表

* 本研究是“教育部课程与教材研究中心”主持的“建立中小学生学业质量分析、反馈与指导系统”项目成果之一,本文所用数据由项目组提供。本研究得到国家自然科学基金项目(30870784)资助

** 通讯作者:骆方。Email:luof@bnu.edu.cn

表 1 评分量表参数

变量			未加权的拟合值			加权后的拟合值		
评分维度	估计值	标准误	*MNSQ*	*CI*	*t*	*MNSQ*	*CI*	*t*
内容选择	2.095	0.019	0.96	(0.95, 1.05)	−1.8	0.99	(0.95, 1.05)	−0.2
语言表达	1.220	0.020	0.71	(0.95, 1.05)	−13.9	0.78	(0.95, 1.05)	−8.8
结构安排	0.288	0.020	1.00	(0.95, 1.05)	0.2	1.04	(0.95, 1.05)	1.5
书写和标点	−3.603*	0.034	1.93	(0.95, 1.05)	31.5	1.71	(0.94, 1.06)	20.7

注:"*"表示该数据是根据约定的条件计算获得

表 2 评分教师评分宽严程度

评分员	宽严程度	误差	评分员	宽严程度	误差	评分员	宽严程度	误差
17	−1.079*	0.236	6	−0.070	0.044	11	0.358	0.048
15	−0.615	0.057	1	−0.033	0.050	4	0.403	0.058
10	−0.361	0.051	16	0.005	0.117	5	0.481	0.042
12	−0.350	0.056	7	0.013	0.042	14	0.672	0.041
3	−0.277	0.053	8	0.046	0.089	9	0.889	0.048
13	−0.259	0.048	2	0.177	0.047			

注:"*"表示该数据是根据约定的条件计算获得

现进行评定时是否能够根据统一标准使用各个评分等级,这里由加权的拟合指数来描述。当其值小于或等于 0.8 时,说明该评分员将不同等级的表现归属于少数几个甚至某一个评分中,整个评分过程中表现出过度的一致性,未能对考生的优劣做出有效的区分;当其大于 0.8 且小于 1.2 时,这些评分员有较好的内部一致性;而当其值大于或等于 1.2 时,该评分员内部一致性欠佳。

表 3 评分员一致性分析

变量	未加权的拟合值			加权后的拟合值		
评分员	*MNSQ*	*CI*	*T*	*MNSQ*	*CI*	*T*
1	1.03	(0.88, 1.12)	0.6	1.00	(0.87, 1.13)	0.1
2	0.97	(0.88, 1.12)	−0.5	0.96	(0.87, 1.13)	−0.5
3	1.08	(0.87, 1.13)	1.2	1.09	(0.86, 1.14)	1.2
4	1.28	(0.85, 1.15)	3.4	1.24	(0.84, 1.16)	2.7
5	1.05	(0.89, 1.11)	0.8	1.02	(0.88, 1.12)	0.3
6	0.97	(0.89, 1.11)	−0.6	0.99	(0.88, 1.12)	−0.2
7	1.20	(0.89, 1.11)	3.4	1.17	(0.89, 1.11)	2.8
8	1.07	(0.75, 1.25)	0.6	1.03	(0.73, 1.27)	0.2
9	1.01	(0.87, 1.13)	0.1	1.04	(0.86, 1.14)	0.5
10	1.22	(0.87, 1.13)	3.1	1.19	(0.86, 1.14)	2.5
11	1.32	(0.88, 1.12)	4.6	1.25	(0.87, 1.13)	3.4
12	1.16	(0.86, 1.14)	2.2	1.14	(0.85, 1.15)	1.7
13	1.06	(0.88, 1.12)	1.1	1.06	(0.87, 1.13)	1.0
14	1.06	(0.89, 1.11)	1.0	0.96	(0.88, 1.12)	−0.7
15	1.62	(0.85, 1.15)	7.0	1.44	(0.84, 1.16)	4.8
16	1.26	(0.64, 1.36)	1.4	1.29	(0.62, 1.38)	1.4
17	1.41	(0.49, 1.51)	1.5	1.25	(0.46, 1.54)	0.8

由表 3 可见,编号为 4、11、15、16、17 的评分员在评分过程中自身一致性较差,其余 12 名评分员都具有较好的内部一致性。

3.1.4 评分员之间的评分一致性

评分员之间评分的一致性,是指在彼此独立评分的情况下,两个或多个评分员对同一篇文章所给分数的一致程度。数据与模型拟合统计量是评价评分教师间评分一致性的重要依据[11]。在评分员参数中,当未加权的拟合指数等于 1 且 *T* 等于 0 时,表明评分员之间具有良好的一致性,当未加权的拟合指数大于 1.5 时,则表明评分员之间存在较大差异。表 3 中,未加权的拟合指数都小于 1.5(除了 15 号),表明大多数评分员的评分比较一致。2,6 两名评分员的未加权的拟合指数小于 1,又说明评分员之间的一致性还不够理想。

3.2 一致性较差评分员的评分误差

3.1.3 的结果显示评分员 4、11、15、16、17 评分一致性较差,下面对其评分情况做进一步检验。从这 5 位评分员给出的所有评分中每人随机选取 10 份评分,使用 FACETS 软件中的三面(写作能力、评分员宽严度、评分维度难度)测量模型进行分析。对评分维度分析结果显示,1 分到 4 分四个分数等级的难度及标准差分别是 −4.90 (0.70)、−2.89 (0.30)、2.89(0.70)和 10.11(1.00),呈单调递增状态,表明评分等级与考生能力的对应是合理的。

表 4 评分员宽严程度分析

评分员	宽严程度	标准差	*INFIT*	*OUTFIT*
4	−2.48	0.41	1.05	0.85
15	−3.26	0.49	0.80	0.94
16	−3.40	0.41	0.92	0.73
11	−3.59	0.47	1.06	0.78
17	−5.75	0.46	0.50	0.28

表 4 给出了评分员的宽严程度、标准差、*INFIT* 和

OUTFIT 统计值。*OUTFIT* 统计量是传统卡方值除以自由度,是未加权的均方,对偏离期望的极端值比较敏感;*INFIT* 统计量是用该项目的方差作权重求得的联合均方,受极端值影响较小。

整体来看,这 5 名评分员的评分都偏宽松,17 号评分员最宽松,其评分也没有区分度(*OUTFIT* = 0.28 < 0.6),同时,该评分员可能没有用全部的分数段来评分(*INFIT* 值低于均值 2 个标准差)。检查由该评分员评分的全部 30 名学生,发现没有学生得到合格以下的分数。因此,对于该评分员应该重新进行培训甚至进行撤换。偏差分析的结果显示,没有发现明显的交互作用,即没有出现 Z 值 > 2.0 或 < -2.0 的偏差。因此,个别评分员的评分虽不够准确,但没有出现对特定考生或在某些特定评分维度上表现出特殊偏向的情况。

对拟合不良的评分(misfitting ratings),其中包括那些拟合值在合理范围内的评分员偶然给出的缺乏一致性的评定)的进一步分析发现,这些个别的意外评分几乎都出现在"书写和标点"评分维度上。

4 讨论

作文是语文能力测试的重要组成部分,本研究运用多面 Rasch 模型对不同评分方式下,作文评分中可能出现的误差进行了分析。该方法不但可以确定评分员之间在宽严程度上的差异,而且可以得到每个评分员宽严程度的度量值,从而识别出过于严厉和过于宽松的评分员。

不同评分维度的难度有显著差异,其中书写和标点与其余 3 个评分维度难度差距较大,且在评分过程中表现出了等级限制,没有能很好的区分不同能力水平的考生。一方面可能是因为该维度的评分标准设定不合理或描述不准确;另一方面,可能是因为这一评分维度难度过低,不能够对考生的写作能力进行良好区分。因此,可以考虑对这一评分维度进行调整。

评分员自身评分的不一致也是威胁作文评分信度和效度的重要因素。影响评分员内部一致性的因素有很多,比如除了写作能力之外的一些考生自身的特质,像性别、种族、民族、社会阶层等[12]都可能产生影响。对于这种自身缺乏一致的评分员,可以让他们结合评分细则多做练习,达到比较稳定的状态后再进行正式评分。同时,也可以考虑使用测量技术对考生的原始分数进行调整,得到对考生能力更准确的估计值。

从本研究可看出即使有经验的阅卷员在经过正式培训后仍可能出现评分偏差和不一致。通过统计方法及网上阅卷等方法的使用,可实现对评分员的评分进行实时监控,及时给以具有针对性的反馈,指出评分员在评分过程中出现的问题,有利于减小主观评分中的误差。

5 参考文献

1 孙晓敏,张厚粲. 国家公务员结构化面试中评委偏差的 IRT 分析. 心理学报,2006,38(4): 614 - 625

2 Cason G J, Cason C L. A deterministic theory of clinical performance rating, Evaluation and the Health Professions, 1984, 7: 221 - 247

3 Lunz M E, Stahl J A, et al. Variation among examiners and protocols on oral examinations, The annual meeting of the American Educational Research Association, San Francisco, 1989

4 Lumley T, McNamare T F. Rater characteristics and rater bias: Implications for training. Language Testing, 1995, 12(1): 54 - 71

5 Lunz M E, Wright B D, et al. Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 1990, 3(4):331 - 345

6 Lumley T, Mcnamara T F. Rater characteristics and rater bias: Implications for training. Language Testing, 1995, 12:54 - 71

7 Linacre J M, Wright B D. Understand Rasch measurement: Construction of measures from Many-facet Data. Journal of Applied Measurement, 2002, 3(4):486

8 Barrett S. Question choice: Does marker variability make examinations a lottery? HERDSA Annual International Conference, Melbourne, 1999, 12 - 15

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

10 Saal F E, Downey R G, Lahey M A. Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 1980, 88(2):413 - 428

11 Engelhard G J, Stone G E. Evaluating the quality of ratings obtained from standard-setting judges. Educational and Psychological Measurement, 1998, 58(2):79 - 196

12 Brown R C. Testing black student writers. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman, 1986 - 108

The Measurement of Writing Scoring Quality with the Many-Facet Rasch Model

Liu Hongyun[1], *Chen Yue*[2], *Luo Fang*[1], *Wang Yunfeng*[3]

([1] School of Psychology, Beijing Normal University, Beijing, 100875)

([2] Beijing Academy of Educational Sciences, Beijing, 100031)([3] Institute of Psychology, Beijing, 100192)

Abstract Based on the scoring data of writing tests of the Chinese Students' Academic Achievement Assessment, the writing scoring quality of 17 raters were analyzed with the Many-Facet Rasch Model. The results revealed that (1) Corresponding to the students' ability, the trend of the raters' grading was normal, and most raters exhibited an intra-rater stability. (2) The raters were significantly different in their strictness, but the consistence of the inter-rater scoring was generally good. (3)Besides, the paper did further studies on the raters with low rating reliability.

Key words writing test, rater effort, Item Response Theory, Many-Facets Rasch model