心 理 学 报 2008, 40 (9): 1030 ~1040
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2008.01030

多面 Rasch 模型在结构化面试中的应用*

孙晓敏[1] 薛 刚[2]

([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]Kennedy School of Government, Harvard University, MA 02138, USA)

摘 要 使用项目反应理论中的多面 Rasch 模型, 对 66 名考生在结构化面试中的成绩进行分析, 剔除了由于评委等具体测量情境因素引入的误差对原始分数的影响, 得到考生的能力估计值以及个体水平的评分者一致性信息。对基于考生能力估计值和考生面试分得到的决策结果进行比较, 发现测量误差的确对决策造成影响, 对个别考生的影响甚至相当巨大。进一步使用 Facets 偏差分析以及评委宽严程度的 Facets 分析追踪误差源。结果表明, 将来自不同面试组的被试进行面试原始成绩的直接比较, 评委的自身一致性和评委彼此之间在宽严程度上的差异均将导致误差。研究表明, 采用 Facets 的考生能力估计值作为决策的依据将提高选拔的有效性。同时, Facets 分析得到的考生个体层次的评分者一致性指标, 以及评委与考生的偏差分析等研究结果还可以为面试误差来源的定位提供详细的诊断信息。

关键词 结构化面试; 项目反应理论; 多面 Rasch 模型

分类号 B841.7

1 引言

1.1 面试及其误差

近年来, 人事测评在人员招聘中正发挥着越来越重要的作用, 测评的科学性与实用性也越来越得到人们的认可。在各种主要的测评技术中, 面试已经成为人员招聘中使用最为广泛的方法[1]。面试是一个或多个考官与一个求职者之间在有限时间内的人际互动, 旨在鉴别求职者的知识、技能、能力和行为等方面的特征, 这些特征将用于预测求职者在未来工作上的成功。对这种成功的操作性定义包括工作绩效、培训、晋升、任期等方面的指标[2]。面试按标准化程度可分为: 结构化面试、半结构化和非结构化面试三种。所谓结构化面试是指面试的内容、方式、评委构成、程序、评分标准及结果的分析评价等构成要素, 按统一制定的标准和要求进行的面试[3]。

随着面试技术的普遍使用, 出现了大量关于面试误差的研究。Wagner 曾总结道: 面试的信度和效度或许存在高度的情境特异性和考官特异性[4]。面试误差的来源有很多方面: 面试题目的有效性、面试实施的各个环节、面试评委的培训、面试记分维度的设定以及面试评分量表的设计等等。具体而言, 对面试研究的总结反复提到下述评分误差: 对比效应[5]、与我类似效应、第一印象偏差、晕轮效应、考官刻板印象、顺序效应[6]、考官对考生的个人感情、信息偏好[5]等等。关于面试误差的这些研究结果意味着评分往往在一定程度上反映的是面试过程的特征或者考官个人的评分技能, 而不完全是被试与工作相关的特征。因此, 这种来自情境的误差变异损害了面试的潜在效度和效用[7]。

1.2 面试误差的控制

为了降低面试过程中的情境误差, 提高实际面试的效度和信度, 研究者从三个方面做了大量工作。第一个方面是关注面试的内容、获取信息的维度的标准化, 从而使得考官尽可能在相同的工作相关信息的基础上对考生进行评价, 最终提高评分的一致性。例如, 不断提高面试过程的结构化, 确保提问的一致性, 将提问限制在与工作相关的问题上, 并且对考生反应的评价也结构化。第二个方面是重视考官

收稿日期: 2007 -10 -09

*2007 -2008 年度北京市教委重点实验室规划项目。

通讯作者: 孙晓敏, E-mail: sunxiaomin@bnu.edu.cn, 电话: 010 -58802101

培训。通过培训,使考官熟悉各种可能的评分误差,帮助他们在信息收集和解释的过程中认识并尽力消除这类误差。

除了上述两种方法,近年来,随着现代测量理论的发展,越来越多的研究者尝试使用现代测量学的方法,通过统计校正,改进面试评分技术,提高面试信度和效度。

在引入现代测量理论之前,面试分析的方法经历了经典测量理论(Classical Test Theory, CTT)和概化理论(Generalizability Theory, GT)两个阶段。

在 CTT 中,考生的观测值由真分数和误差项组成。误差项越大,观测值的信度就越低。但是,CTT 中所定义的误差很笼统,它不能说明测量误差究竟来自哪些误差源,以及各自产生的误差大小。对于面试中最容易出现的评分者误差,CTT 往往通过计算评委之间的一致性,如 Kendall 和谐系数等加以分析。由于这类评分者一致性系数属于事后检验,且 CTT 测量指标的样本依赖性又限制了其在类似面试情境中的推广,因此,Kendall 和谐系数等 CTT 提供的指标和方法对于面试决策科学水平的提高能够起到的作用非常有限。

虽然 GT 在 CTT 的基础上对各误差源的方差分量进行了估计,提供了概化系数作为评价测评结果信度的指标,并为进一步的实验设计提供了信息,但是 GT 并没有改良 CTT 的项目参数系统。它更多的是从整个测验的宏观结构及其与外部测验条件的关系上做了深入的计量分析,而 CTT 存在的一些问题,比如其测验结果的样本依赖性问题等,同是随机抽样理论的 GT 并没有从根本上克服。随着心理及教育测量中对测验的精度越来越高的要求,项目反应理论(Item Response Theory, IRT)应运而生,并为弥补 CTT 和 GT 的缺陷提供了新的思路。

与随机抽样理论不同,IRT 认为测量的目标不是考生在特定测验上得到的真分数,而是由这个分数体现出来的考生能力。IRT 中称这种能力为潜在特质。虽然潜在特质不能被直接测量,但受测者在测验项目上的表现与该项目所要测量的潜在特质之间存在着一种单调递增的函数关系。IRT 数学模型的核心是项目特征函数,它是受测者在项目上的答对概率对其能力值(或潜在特质)的回归曲线。与 CTT 不同,IRT 通过项目特征函数将项目难度和考生特质水平定义在同一度量系统上,使得项目参数和考生特质参数都不依赖于样本。IRT 的优势吸引了众多研究者对此展开了大量的研究,IRT 也因此得以迅速发展,各种符合实际需要的新模型不断涌现。IRT 常用的有单参数、双参数和三参数模型。由丹麦数学家 Rasch 独立开发的单参数 Rasch 模型以其统计上的优点和参数估计的便利性而著称。下面为单参数 Rasch 模型的函数表达式。

$$\log\left(\frac{P_{ni}}{1-P_{ni}}\right)=B_n-D_i$$

P_{ni}是考生 n 正确作答项目 i 的概率

$1-P_{ni}$是考生 n 答错项目 i 的概率

B_n 是考生 n 的能力 ($n=1, 2, \cdots, N$)

D_i 是项目 i 的难度 ($i=1, 2, \cdots, L$)

单参数 Rasch 模型是一个两面的模型,即包含考生能力和项目难度两个侧面。通过应用该模型,可以得到项目难度和考生能力的估计值,且该能力估计值独立于考生遇到的特定项目的难度。

与单参数 Rasch 模型所处理的测验情境相比,面试的测量情境由于引入了评委而变得更加复杂。因为考生得到某一特定分值的概率不仅取决于考生的能力、项目的难度,而且受到评委的宽严程度、评定量表上特定分数等级的难度等因素的影响。面对这样的情境,如果不仅能确定项目的难度、评委的宽严程度,而且还能估计出考生独立于特定项目难度和特定评委宽严程度的能力值,这对于提高测验的信度将具有十分重要的作用。Linacre[28] 提出的多面 Rasch 模型(Many Facets Rasch Model, MFRM)就成功实现了这一目标。MFRM 在两面 Rasch 模型的考生侧面和项目侧面的基础上增加了评委侧面,因而被称作多面 Rasch 模型。以下是 MFRM 的函数表达式:

$$\log\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right)=B_n-D_i-C_j-F_k$$

P_{nijk}是考生 n 在项目 i 上被评委 j 评定为 k 等的概率

$P_{nij(k-1)}$是考生 n 在项目 i 上被评委 j 评定为 $k-1$ 等的概率

B_n 是考生 n 的能力参数 ($n=1, 2, \cdots, N$)

D_i 是项目 i 的难度参数 ($i=1, 2, \cdots, L$)

C_j 是评委 j 的宽严程度 ($j=1, 2, \cdots, J$)

F_k 是分部记分模型(Partial Credit Model)中考生得分从 $k-1$ 等到 k 等的等级难度(step difficulty),每个项目均为 K 级评分 ($k=1, 2, \cdots, K$)

测验过程中的每个因素或侧面都被设定为一个独立的参数。这些参数的估计值统一用 logits 作为单位表示。Logits 具有可加性,相当于一种成功的

概率。由于各种可能偏差的存在，面试中考生的得分是考生能力、评委宽严程度、项目难度以及等级难度等因素共同作用的结果。MFRM 在统计分析过程中剔除评委和特定项目特征的影响，得到的考生能力值是独立于考生遇到的特定评委特点以及特定项目难度的[10]。因此，基于 MFRM 估计得到的考生能力值做出的用人决策将更为客观公平。

MFRM 的这一优势使得它在面试这种评委有可能引入大量误差的测量情境中展示了巨大的应用前景。因此，在国外 MFRM 自提出即受到多个领域研究者的关注。这种分析方法被广泛地用于外语口语面试评分[11]、作文评分[12~16]、医师资格认证考试评分[17~19]、体育比赛评分[20]、档案袋[21]形式的表现性评价[22~24]，以及其它多种测评情境中[25]。事实上，从更广泛的意义上来讲，凡是存在多个评委主观判断的测评情境，MFRM 几乎总能找到它的用武之地。

但是，在我国，对于面试中各种误差的研究还停留在比较初级的阶段。大多数面试研究仍局限于 CTT 的方法。鉴于此，本文拟将 MFRM 引入结构化面试分析。这一方法在面试中的应用不但有利于有效区分不同能力水平的考生，而且为进一步完善评分规则、识别问题评委、解决面试等值等问题都提供了全新的解决思路[26]。

项目反应理论由于其参数估计的复杂性，必须通过计算机程序才能实现。John M. Linacre 编制了用于 MFRM 的计算机程序 Facets。Facets 采用非条件极大似然法(Unconditional Maximum Likelihood)对 MFRM 中的各个参数进行估计。目前该软件的最新版本是 Facets for Windows 3.63.0[27]。

1.3 研究目的

使用 Facets 3.63.0 对考生的能力值进行估计，得到剔除特定情境误差的考生能力估计值；在此基础上，将基于能力估计值进行的决策与基于考生面试原始分进行的决策进行比较，定位问题考生。进一步使用 MFRM 偏差分析，对问题考生的误差源进行追踪，分析造成偏差的深层原因。

2 方法

2.1 被试

本研究的数据来自某机构的结构化面试。每个面试组由经过培训的 7 名评委组成。具体面试工作在两天内完成，为防止评委作弊，21 名评委在每天面试前随机抽签分成 3 组同时进行面试。

本研究随机抽取面试第一天和第二天各一个面试组的评分信息作为分析所用数据。所抽取的第一天评委组 1 由评委 A、B、C、D、E、F、G 七人组成。该组当天共面试考生 34 名，编号 1 –34。研究所抽取的第二天评委组 2 由评委 A、E、H、I、J、K、L 七人组成。该组当天共面试考生 32 人，编号 35 –66。

2.2 实验程序

该结构化面试由一名主考官按既定问题顺序对应试者提问。每个评委对每名被试在仪表风度举止、口头表达能力、应变能力、综合分析能力和逻辑思维能力 5 个维度上使用 10 点量表进行独立评定。决策部门采用专家法为上述各维度设定的权重依次为：2.5、2.0、1.0、2.5、2.0。

2.3 数据分析

采用 IRT 的 MFRM 模型，使用该模型的配套软件 Facets for Windows 3.63.0 对面试数据进行处理。

3 结果

3.1 考生能力值的 MFRM 分析结果

采用 Facets 3.63.0 对面试数据进行分析，得到参加面试的 66 名考生能力估计值见表 1。

在表 1 中，对考生按照能力估计值从高到低，从左到右的顺序进行了排列。表中第 1 列是考生编号。第 2 列是考生能力估计值。第 3 列是考生能力估计值的标准误，表明了该估计值的精确程度。在其它条件不变的情况下，一个估计值所基于的观测次数越多，其标准误越小[29]。

66 名考生的能力值范围为 –1.46 ~3.15 logits，全距为 4.61 logits，平均数是 0.94 (SE = 1.2)。其中，53 号考生是能力水平最高的，能力值是 3.15 logits (SE =0.22)；而 37 号考生是能力水平最低的，能力值是 –1.46 logits (SE =0.13)。

第 4 列的考生 infit 值(Infit MnSq)表明了评委们对于该考生的评分一致性程度，即该考生在多大程度上得到了来自评委们的比较一致的评价。在 MFRM 中，fit 值是对模型预期值和观测值之间的差异进行描述的统计量。通常情况下，不同评委对特定考生的评价不可能完全一致，总会出现一定程度的差异。MFRM 允许其在正常范围内的波动。但是如果实际观测分数中，多个评委对特定考生的评定的一致性超出了一定的范围，即评定差异过于悬殊或评定过于一致，fit 统计量将探测出这种异常。对于 fit 统计量的可接受范围，MFRM 并没有给出严格的规定，fit 值的范围常常根据具体研究的需要而

定。通常测评要求的精度越高, fit 值的可接受范围越窄。有的研究将合适的 fit 值范围定在 0.5 ~1.5 之间[13]。但大多数研究设定为 0.8 ~1.2[14]。本研究也选择此标准。即如果某考生的 *infit* 值大于 1.2, 表明评委们对该考生的评分变异大于模型期望的变异程度, 对该考生的评分一致性比较差。相反, 如果某考生的 *infit* 值小于 0.8, 则表明评委们对该考生的评定与模型期望相比过于一致。从表 1 可以看到, 考生 3 的 infit 值高达 2.28, 远远大于 1.20 的警戒线, 是所有考生中最有争议的一个。

表 1　66 名考生的能力估计值

考生	能力估计值	*S.E.*	Infit MnSq	考生	能力估计值	*S.E.*	Infit MnSq	考生	能力估计值	*S.E.*	Infit MnSq
53	3.15	0.22	0.61	19	1.57	0.18	1.05	64	0.14	0.16	0.89
23	3.05	0.21	1.44	32	1.53	0.18	0.93	40	0.13	0.16	0.99
11	3.02	0.21	1.19	62	1.47	0.19	0.96	27	0.1	0.15	1.02
25	2.84	0.21	0.63	13	1.42	0.18	1	63	0.06	0.16	0.84
16	2.76	0.21	0.83	1	1.4	0.18	1.05	54	0	0.16	1.18
61	2.69	0.21	1.05	47	1.38	0.19	0.86	7	-0.02	0.15	0.77
3	2.67	0.21	2.28	49	1.38	0.19	0.77	44	-0.02	0.16	0.81
8	2.48	0.21	2.02	45	1.35	0.19	0.97	58	-0.07	0.16	0.72
48	2.31	0.21	0.71	56	1.33	0.19	0.78	60	-0.17	0.15	1.05
41	2.18	0.21	0.94	2	1.33	0.18	0.63	26	-0.17	0.15	0.67
29	2.17	0.2	1.32	30	1.11	0.17	0.64	46	-0.18	0.15	1.75
52	2.12	0.2	1.07	33	0.87	0.17	0.96	5	-0.24	0.15	0.86
10	2.05	0.2	0.74	4	0.8	0.16	1.15	6	-0.49	0.14	1.18
9	2.03	0.2	0.55	36	0.69	0.17	0.83	42	-0.57	0.15	0.62
14	1.98	0.19	1.72	20	0.61	0.16	0.88	66	-0.65	0.15	1.23
17	1.9	0.19	1.07	22	0.59	0.16	1.53	51	-0.67	0.15	1.87
24	1.9	0.19	0.94	50	0.54	0.17	0.75	28	-0.84	0.14	0.69
15	1.87	0.19	0.5	39	0.52	0.17	1.27	35	-0.97	0.14	1.25
31	1.85	0.19	0.61	43	0.36	0.16	1.08	38	-0.99	0.14	0.52
12	1.83	0.19	0.5	18	0.36	0.16	1.08	57	-1.11	0.14	0.98
65	1.78	0.2	1.02	55	0.32	0.16	0.85	59	-1.14	0.14	0.54
21	1.71	0.19	1.4	34	0.23	0.15	0.8	37	-1.46	0.13	0.78
								平均数	0.94	0.17	0.99
								标准差	1.2	0.02	0.36

注: *RMSE* (Model): 0.18,　Adj *SD*: 1.19　Separation: 6.75,　Separation Reliability: 0.98

表 1 下方的指标 *RMSE* 是 Root Mean - Square Standard Error 的缩写, 即考生能力估计值标准误(表 1 第 3 列)的均方的平方根。*RMSE* 的平方即 *MSE* (Mean-Square Standard Error), *MSE* 代表了误差变异。Adj. *SD* 是校正了测量误差之后的估计值的标准差。Adj. *SD* 的平方即为真实变异。观测变异(表 1 第 10 列最后 1 行标准差的平方即为观测变异)减去误差变异 *MSE* 就可以得到真实变异。

Separation 是 Adj. *SD* 除以 *RMSE* 得到的数值, 它标志着测量分数整体的有效性。如果来自考生的真实变异与来自测量误差的变异相等, 则 Separation 为 1.0。要达到传统的 0.90 水平的置信度, 则 Separation 需要等于 3.0[30], 而本研究得到的 Separation 指标已高达 6.75。

Separation Reliability 是真实变异除以观测变异得到的数值, 即表明在总观测变异中真实变异所占的比例[26]。该信度值越大, 表明考生之间的能力差异也就越大。考生之间能力差别不大, 这个值接近于 0; 反之, 则该值接近 1。就表 1 而言, Separation Reliability 为 0.98, 表明考生之间在能力水平上存在较大差异。对考生之间能力的差异大小进行 χ^2 检验, 结果表明, $\chi^2(65)=3236.4$, $p<0.01$, 即考生之间在能力上差异显著。

3.2 考生原始分与考生能力估计值结果比较

在面试的实际操作中，考生面试成绩的计算方法是：以每个评委在各维度上的加权平均分作为原始评定，去掉原始评定中的最高分和最低分，再求算术平均值为考生的最后面试成绩。按照这个方式计算出来的面试成绩就是决策部门人员选录的依据。在表 2 中，称其为考生的面试分。

为了对比基于面试分和基于考生能力估计值得到的考生排序之间的差异，及面试误差对每个考生产生的影响的性质和大小，现将考生的面试分和 Facets 能力估计值及相应的排序列入表 2。

表 2 考生面试分、Facets 能力估计值及其排序对比表

考号	面试分	面试分排序	MFRM 成绩(logits)	MFRM 排序	排序差别	考号	面试分	面试分排序	MFRM 成绩(logits)	MFRM 排序	排序差别
1	83.5	31	1.4	27	4	34	78.5	49	0.23	44	5
2	83.7	30	1.33	32	−2	35	71.5	64	−0.97	62	2
3	88.8	10	2.67	7	3	36	82.4	34	0.69	36	−2
4	82.1	35	0.8	35	0	37	69.8	66	−1.46	66	0
5	75.1	57	−0.24	56	1	38	73.4	60	−0.99	63	−3
6	72.3	61	−0.49	57	4	39	81.5	40	0.52	40	0
7	76.8	53	−0.02	51	2	40	79.2	45	0.13	46	−1
8	88.8	11	2.48	8	3	41	88.9	8	2.18	10	−2
9	86.7	13	2.03	14	−1	42	75.7	55	−0.57	58	−3
10	86.6	14	2.05	13	1	43	81.8	39	0.36	41	−2
11	90.8	3	3.02	3	0	44	79.1	46	−0.02	50	−4
12	85.8	25	1.83	20	5	45	86	22	1.35	30	−8
13	83.1	32	1.42	26	6	46	77.1	52	−0.18	55	−3
14	86.4	15	1.98	15	0	47	85.7	26	1.38	29	−3
15	85.9	23	1.87	18	5	48	88.9	9	2.31	9	0
16	89.2	6	2.76	5	1	49	85.9	24	1.38	28	−4
17	86.2	20	1.9	17	3	50	82	36	0.54	39	−3
18	78.1	51	0.36	42	9	51	75.2	56	−0.67	60	−4
19	84.8	27	1.57	23	4	52	88.5	12	2.12	12	0
20	79.8	43	0.61	37	6	53	91.4	1	3.15	1	0
21	84.7	29	1.71	22	7	54	78.9	47	0	49	−2
22	79.6	44	0.59	38	6	55	81.9	37	0.32	43	−6
23	89.4	4	3.05	2	2	56	86.4	16	1.33	31	−15
24	86.3	17	1.9	16	1	57	72	63	−1.11	64	−1
25	89.4	5	2.84	4	1	58	78.8	48	−0.07	52	−4
26	74.1	59	−0.17	54	5	59	72.2	62	−1.14	65	−3
27	76.3	54	0.1	47	7	60	78.3	50	−0.17	53	−3
28	70	65	−0.84	61	4	61	90.9	2	2.69	6	−4
29	86.3	18	2.17	11	7	62	86.3	19	1.47	25	−6
30	82.8	33	1.11	33	0	63	80.3	41	0.06	48	−7
31	86	21	1.85	19	2	64	80.2	42	0.14	45	−3
32	84.8	28	1.53	24	4	65	89	7	1.78	21	−14
33	81.8	38	0.87	34	4	66	74.9	58	−0.65	59	−1

表 2 中，第 1 列为考生编号，共 66 名考生；第 2 列为考生面试分；第 3 列为以面试分为依据对考生进行排序的结果；第 4 列为考生 Facets 能力估计值；第 5 列为以能力估计值为依据对考生进行排序的结果；第 6 列为第 3 列减去第 5 列，即考生在不同条件下排序结果的差值。

由表2 可以看出, 当决策基于的分数不同时, 有些考生的名次发生了一定程度的变化。名次变化最大的是 56 号考生, 当采用原始分进行排序时, 该考生排名第16, 但是当采用剔除了考官宽严程度等测量情境因素之后的 Facets 能力估计值进行排序时, 该考生排名第 31, 前后相差 15 名之多。假设决策机关划定前 25 名考生录取的话, 那么基于不同分数得到的不同排序结果将直接影响到 56 号考生的录用与否。

为了进一步了解 56 号考生面试分数的误差来源, 有必要使用 Facets 程序所提供的偏差分析功能, 对其做进一步分析。

3.3　评委与考生的 Facets 偏差分析结果

评委与考生的偏差分析通过比较特定的评委对特定的考生进行评价时的观测值与期望值之间的差异, 提供了每个评委在面对不同考生时发生偏差的程度。对参与评价56 号考生的 7 位评委的评分数据作进一步偏差分析, 提取与 56 号考生有关的分析结果, 见表 3。

表3　七位评委与 56 号考生的偏差分析结果

评委	宽严程度	观测分	期望分	偏差分	偏差 logits 值	Model*S. E.*
A	-0.35	89	86.3	-2.7	-0.77	0.56
E	0.24	86	83.8	-2.2	-0.51	0.51
H	-0.89	90	88.3	-1.7	-0.54	0.57
I	0.23	84	83.9	-0.1	-0.09	0.46
J	0.67	78	81.8	3.8	0.75	0.4
K	-0.98	89	88.5	-0.5	-0.14	0.56
L	-0.5	85	86.9	1.9	0.6	0.48

表 3 中第 1 列为评委代号; 第 2 列为评委的宽严程度估计值; 第 3 列为特定评委对该考生的原始评定; 第 4 列为模型依据该考生的能力水平以及特定评委在评价其他考生时的宽严程度所预期的该评委应该给出的分数; 第 5 列为评委实际给出的分数与模型期望分数的差值。

从表3 可以初步看出, 对56 号考生进行评价的 7 名评委中, 有 5 位的偏差值均为负数, 即 5 位评委对该考生都给出了比模型期望要高的分数, 在一定程度上造成了该考生原始分数的夸大。

此外, 为了深入挖掘 56 号考生分数的偏差原因, 考虑到66 名考生来自不完全相同的两个面试评分组, 我们对两组的 12 名评委的宽严程度进行了 Facets 分析。

3.4　评委宽严程度的 Facets 分析结果

对两组评委的宽严程度进行 Facets 估计, 得到结果见表4。

表 4 中, 第 1 列为评委代号; 第 2 列为该评委所属的评分组, 前已提及, 由于面试两天随机分组, 因此评委 A 和 E 参加了两个组的评分; 第 4 列是该评委所评价的考生人数; 第 5 列为评委宽严程度估计值。12 位评委按照从宽到严的顺序自上而下排列。

对第 1 组评委与第 2 组评委的宽严程度进行比较, 第 1 组的平均宽严程度值为 0.2629, 第 2 组的平均宽严程度为 -0.2771, t =1.982, p =0.071。这表明第 1 组评委比第 2 组更加严格, 两者的差异接近但尚未达到 0.05 水平上的统计显著。

表 4　评委宽严程度 Facets 分析结果

评委	组别		考生个数	宽严程度值	Model *S. E.*
K	2	宽	32	-0.98	0.08
H	2	松	32	-0.89	0.08
L	2	↕	32	-0.50	0.08
A	1&2	↕	66	-0.35	0.06
D	1	↕	34	-0.22	0.08
C	1	↕	34	0.00	0.08
I	2	↕	32	0.23	0.07
E	1&2	↕	66	0.24	0.05
B	1	↕	34	0.31	0.08
J	2	↕	32	0.67	0.07
F	1	严	34	0.70	0.07
G	1	格	34	0.80	0.07
平均数:				0.00	0.07
标准差:				0.58	0.01

4　讨论

使用 Facets for Windows 3.62.0 对考生能力值进行独立于特定测量情境的估计, 一方面得到了考生能力估计值, 另一方面得到的 infit 值可以作为判定评委们在对特定被试进行评价时的评分者一致性

指标。这个指标是评分者一致性程度在每个特定考生身上的体现。Facets 分析亦可提供关于评分者差异信度的报告[26]，但评分者差异信度是从整体水平上对评分者之间在宽严程度上存在的差异进行的分析，而这里所报告的考生 infit 值作为评分者差异信度的补充，从个体水平上给出了评委对特定考生进行评价时的一致性程度。Infit 值可以使研究者精确追踪在对哪些被试进行评价时，评分者之间存在较大分歧。MFRM 的这种个体定位的分析深度，体现了 IRT 相比于 CTT 整体定位的诊断优势[31]。

对基于考生的面试分数和能力估计值进行的排序结果进行比较，发现得到的决策结果会出现较大差异，这种差异对于有些考生来说将意味着能否被录取。以56 号考生考生为例。该考生是在两种排序条件下名次变化最大的考生。为了进一步考察造成 56 号考生评分偏差的深层次原因，我们对面试数据做了进一步的偏差分析，并对 12 位评委宽严程度值进行了 Facets 估计。结果表明：

第一，两个评委组的宽严程度存在差异。

通过对评委宽严程度的 Facets 分析结果进行比较可知：与第 1 组评委相比，第 2 组评委整体而言更宽松，尽管这种差异接近但尚未达到统计显著。基于此，当把所有的考生放在一起进行比较时，参加了第 2 组面试的 56 号考生与参加第 1 组面试的同样能力水平的考生相比就更有可能得到更高的面试分。

其次，评委评分的自身一致性波动。

从表 3 可知：对 56 号考生进行面试的 7 位评委中，有 5 位评委的偏差值都为负数。即，有 5 位评委在对这个考生进行评价时都给出了比自己惯常的模式要宽松的评定。

综合上述两方面因素可知：首先，56 号考生被分到了比较宽松的评委组；其次，即使是比较宽松的这个评委组中还有 5 位评委在给该考生打分时表现出比自己一般的水平更为宽松的尺度，从而导致该考生幸运地得到了超出其自身真实能力水平的高分。Facets 对这些特定的测验情境引起的偏差进行校正，剔除掉原始分中被夸大的部分，因此对 56 号考生基于能力值进行的排序便明显落后于基于面试分进行的排序了。

为了从总体上认识考生面试分和能力估计值的关系，分别以面试分与考生能力估计值为横坐标和纵坐标绘制散点图 (见图 1)。

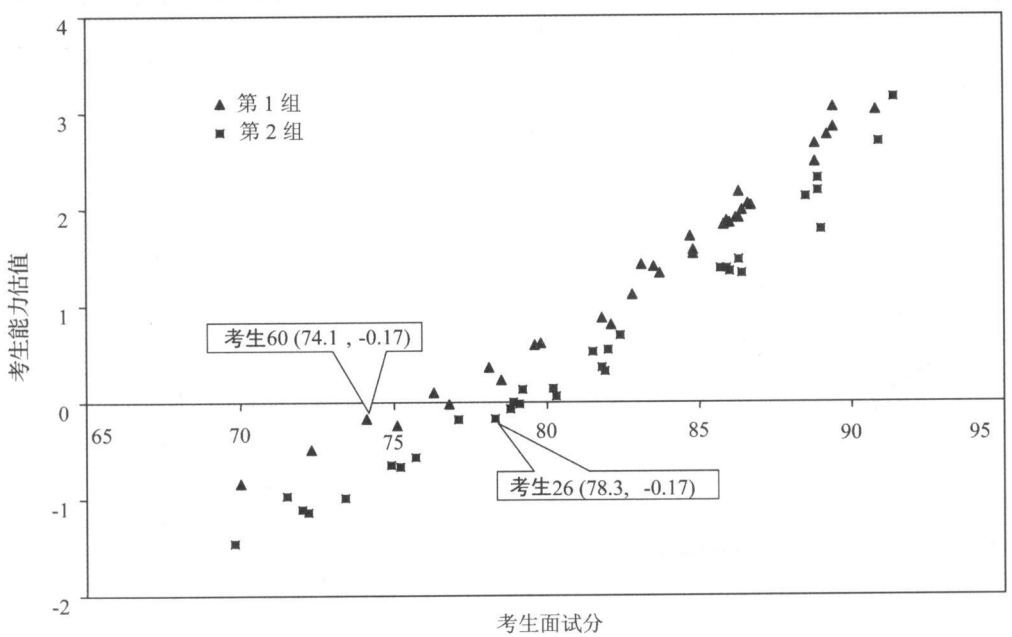

图 1 考生面试分与能力估计值的散点图

图 1 中，三角代表面试 1 组，方块代表面试 2 组。结合图 1 与表 2 可以看到：

第一，能力值相同的考生得到的面试分差别可能较大。例如 26 号考生与 60 号考生的能力值均为 -0.17logits，但是他们的面试分却分别是 78.3 和 74.1，相差约 4 分之多。由表 2 可知，面试分的极差为 21.6，66 名考生分布在这样窄的一个值域区间中，平均而言，每 1 分相差约 3 人，那么 4 分的面试

分差距将导致该被试的名次发生相当大的变化。总之,同样能力的考生得到了不同的面试分,表明来自评委宽严程度等方面的测量情境的误差的确对考生的面试分数造成了影响。而且,这种影响必然会影响到录用决策的有效性。

第二,整体而言,同一能力值的考生在面试 2 组可能得到更高的面试分。表现在图 1 中即第 2 组考生比同一能力水平的第 1 组考生整体向横轴的右侧偏移。造成这一点主要原因前面已经提到,由于第 2 组评委平均而言比第 1 组评委更宽松,因而同样能力水平的考生在第 2 组更容易得到高分。

第三,就每个面试组内部而言,同一能力的考生面试分之间的差异小于组间的这种差异。造成这个现象的主要原因在于:当考生由同一组考官进行评价时,对考生排序造成误差的是评委的自身一致性程度。而评委之间在宽严程度上的系统性差异不会影响考生排序。但是,当把来自不同评分组的考生放在一起进行比较时,由于不同评委组之间存在的差异,基于考生面试原始分进行的决策不但受到评委自身一致性的影响,还会受到评委之间宽严程度差异等因素的影响,因此就会出现更为严重的问题。同一能力水平的考生由于被不同的评委组评定而将得到不同、甚至差别很大的面试分数。在面试实践中,考虑到人力、物力、保密性和时间等各方面的要求,往往是由多个评委组同时对考生进行面试。决策时直接将来自不同评委组考生的面试分进行排序和比较。这种做法使得决策的客观和公平性受到很大影响。MFRM 通过纠正评分者之间宽严程度差异、评分者自身一致性等具体测量情境对考生原始分数造成的偏差,使得决策更加客观和公平。

5　结论

本研究综合使用 MFRM,对 66 名考生的结构化面试数据进行了分析,得到一下主要结论:

(1)通过 MFRM 对考生的原始分数进行校正,剔除了由于评委等具体测量情境因素引入的误差对原始分数的影响,得到考生的能力估计值以及个体水平的评分者一致性信息(infit 值)。

(2)对基于考生能力估计值与考生面试分得到的决策结果进行的比较发现,测量误差的确对决策有效性造成影响,对个别考生的影响甚至相当巨大。进一步使用 Facets 偏差分析以及评委宽严程度的 Facets 分析追踪误差源。结果表明,将来自不同面试组的被试进行面试成绩的直接比较,评委的自身一致性和评委彼此之间在宽严程度上的差异均将导致误差。

采用 Facets 的考生能力估计值作为决策的依据将在很大程度上提高选拔的有效性。同时,MFRM 分析得到的考生个体层次的评分者一致性指标,以及评委与考生的偏差分析等研究结果还可以为面试误差来源的定位提供详细的诊断信息。总之,面试评价力求客观公正的目标受到诸多方面的影响。在改进面试各个环节的同时,探询有效的统计处理方法以控制面试过程中各方面误差的影响是一个思路。本研究所展示的 MFRM 的分析方法为解决上述问题提供了一个途径。

参　考　文　献

1　Maurer T J, Solamon J M. The science and practice of a structured employment interview coaching program. Personnel Psychology, 2006, 59(2): 433 ~456

2　Wiesner W H, Cronshaw S F. A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. Journal of Occupational Psychology, 1988, 61(4): 275 ~290

3　Schmidt F L, Zimmerman R D. A counterintuitive hypothesis about employment interview validity and some supporting evidence. Journal of Applied Psychology, 2004, 89(3): 553 ~561

4　Walters L C, Miller M R, Ree M J. Structured interviews for pilot selection: No incremental validity. International Journal of Aviation Psychology, 1993, 3(1): 25

5　Arvey R D. The employment interview: A summary and review of recent research. Personnel Psychology, 1982, 35(2): 281 ~322

6　Schmitt N. Social and situational determinants of interview decisions: Implications for the employment interview. Personnel Psychology, 1976, 29(1): 79 ~101

7　Maurer S D, Fay C. Effect of situational interviews, conventional structured interviews, and training on interview rating agreement: an experimental analysis. Personnel Psychology, 1988, 41(2): 329 ~344

8　Sun X M, Zhang H C. A comparative study on methods used in estimating reliability of performance assessment: Methods based on correlation, percent of inter-rater reliability and Generalizability theory. Psychological Science, 2005, 28(3), 646 ~649 (in Chinese)
(孙晓敏, 张厚粲. 表现性评价中评分者信度估计方法的比较研究——从相关法、百分比法到概化理论. 心理科学, 2005, 28(03): 646 ~649)

9　Shavelson R J, Webb E A. Generability theory: A primer. Newbury Park, CA: SAGE Publications, Inc., 1991

10　Myford C M, Wolfe E W. Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part II. Journal of Applied Measurement, 2004, 5(2): 189 ~227

11　Lunz M E. Variation among examiners and protocols on oral

examinations. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA: 1989
12 Linacre J M. The calibration of essay graders. Paper presented at the Midwest Objective Measurement Seminar, Chicago, IL: 1987
13 Du Y. Raters and single prompt – to – prompt equating using the FACETS model in a writing performance assessment. Paper presented at the Ninth International Objective Measurement Conference, Chicago, IL: 1997.
14 Kondo – Brown K A. FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing, 2002: 3 ~31
15 Yamauchi K. Comparing Many – facet Rasch Model and ANOVA model: Analysis of ratings of essays. Japanese Journal of Educational Psychology, 1999, 47 (3): 383 ~392
16 Myford C M. Looking for patterns in disagreements: A Facets analysis of human raters ánd e – raters´scores on essays written for the graduate management admission test (GMAT). New Orleans, LA.: American Educational Research Association, 2002
17 Linacre J M. An extension of the Rasch Model to multi – faceted situation. Chicago: University of Chicago, 1987
18 Lunz M E, Wright B D. Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 1990, 3 (4): 331
19 Allen J M, Schumacker R E. Team assessment utilizing a Many – Facet Rasch Model. Journal of Outcome Measurement, 1998, 2 (2): 142
20 Looney M. A many – facet Rasch analysis of 1994 Olympic figure skating. Research Quarterly for Exercise &Sport, 1997, 68 (1): 1 ~53
21 Myford C M, Mislevy R J. Monitoring and improving a portfolio assessment system, Center for Performance Assessment Research Report. Princeton, NJ: Educational Testing Service, 1995
22 Chi E. Comparing holistic and analytic scoring for performance assessment with many – facet Rasch model. Journal of Applied Measurement, 2001, 2 (4): 379 ~388
23 Heller J I, Sheingold K, Myford C M. Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. Educational Assessment, 1998, 5 (1): 5
24 Lunz M E. Performance examinations: Technology for analysis and standard setting. Paper presented at the Annual Meeting of the National Council of Measurement in Education. Chicago, IL: 1997
25 Kline T L, Schmidt K M, Bowles R. Using LinLog and FACETS to model item components in the LLTM. Journal of Applied Measurement, 2006, 7 (1): 74 ~91
26 Sun X M, Zhang H C. An IRT analysis of rater bias in structured interview of national civilian candidates. Acta Psychologica Sinica, 2006, 38 (4), 614 ~625 (in Chinese)
(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 (04): 614 ~625)
27 Linacre J M. Facets for Windows. 3.63.0 ed. Chicago, IL: MESA Press, 2007
28 Linacre J M, Wright B D. Understand Rasch measurement: Construction of measures from Many – facet Data. Journal of Applied Measurement, 2002, 3 (4): 486
29 Linacre J M. Understanding Rasch measurement: Optimizing rating scale category effectiveness. Journal of Applied Measurement, 2002, 3 (1): 85
30 Lunz M E, Wright B D, Linacre J M. Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 1990, 3 (4): 331
31 Linacre J M, Wright B D. Construction of measures from many – facet data. Journal of Applied Measurement, 2002, 3 (4): 486 ~512

A Many – faceted Rasch Model Analysis of Structured Interview

SUN Xiao – Min [1], XUE Gang [2]
([1] *School of Psychology, Beijing Key Lab of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875 *China*)
([2] *Kennedy School of Government, Harvard University, MA* 02138, *USA*)

Abstract

Being one of the most important techniques in personnel selection, structured interview has attracted more and more research interest in improving its reliability and validity. Some researches focus on the standardization of its content and dimensions, others try to decrease rater bias by intensive rater training. The third one, to handle the possible bias in statistic way, has attracted more and more attention. Many – faceted Rasch Model (MFRM), an extension to Rasch model, served as such kind of techniques. By parameterizing not only interviewee s´ability and item difficulty but also judge severity, MFRM offers an effective way to estimate interviewee s´latent trait, that is, the ability, and provides detailed information of inter – rater reliability as far as a specific interviewee is concerned. This study used MFRM to analyze the result of a structured interview and demonstrated a creative way to locate the source of bias for a specific

interviewee.

Data came from a structured interview. There were 7 raters in each interview panel. Since the interview last two days, these 21 raters were randomized into 3 panels in the morning of each day in order to prevent cheating. Rating scores of two panels were used in this study. A、B、C、D、E、F、G were raters of one panel and interviewed interviewees numbered 1 ~34. A、E、H、I、J、K、L were raters in the other panel that interviewees numbered 35 ~66 were interviewed. Each rater rated each interviewee independently on five dimensions using a 10 points rating scale.

Using Facets 3. 62. 0, a computer program based on MFRM, the abilities of 66 interviewees was estimated, accompanied with a Infit MnSq, which demonstrated the degree to which raters in the panel agreed with each other on the evaluation of a specific interviewee. The ranking order based on interview raw scores and Facets estimated logits values were compared. Difference were found between those them. To track the source of error for interviewee numbered 56, bias analysis of Facets was also made.

The ability of 66 interviewees were reported with infit MnSq, showing inter – rater reliability at individual level;

The ranking order based on interview raw score and estimated ability score were quite different, especially for some interviews. Taking interviewee numbered 56 for example, the ranking difference was as large as 15.

Bias analysis aimed at locating the source of error for interviewee number 56 show that not only rater consistency, but also rater severity contribute to the ranking difference.

The results confirmed the utility of MFRM analysis. The application of MFRM in the analysis of structured interview was proved to be not only an effective way in personnel selection, but also provided diagnostic information for sources of error locating at individual level.

Key words　structured interview; IRT; many-faceted Rasch Model