心理学报 2013, Vol. 45, No.9, 1039–1049
*Acta Psychologica Sinica* DOI: 10.3724/SP.J.1041.2013.01039

# 无领导小组讨论的多侧面 Rasch 模型应用*

姚若松[1] 赵葆楠[1] 刘 泽[1] 苗群鹰[2]

([1]广州大学教育学院，广州 510006) ([2]广州大学外国语学院，广州 510006)

摘 要 采用项目反应理论(IRT)的多侧面 Rasch 模型(MFRM)，分析评价中心技术中无领导小组讨论(LGD)的测评结果，探讨被试能力水平、评委评分宽严度、评分内部一致性、维度难度和评定等级等问题，进而讨论各种偏差。通过 MFRM 分析人事测评结果，可深入了解被试能力的真实差异、甄别维度难度、探查测评误差源，从而完善测评试题编制、评估或诊断评委合格性、提高测评维度与测评目的匹配性，为拓展项目反应理论在人事测评中的应用提供独特视角。

关键词 无领导小组讨论；多侧面 Rasch 模型；项目反应理论；人事测评

分类号 B849: C91; B841

## 1 引言

无领导小组讨论(Leaderless Group Discussion, LGD)是组织招聘使用频率较高的测评方法。据统计，无领导小组讨论在国外的评价中心技术中使用率为 59%，而在国内的使用率则达到 85% (何广陵，袁翎，1998)。作为常用的主观评价中心技术，LGD 评分偏差易受维度、测题、评分标准、评委因素(何琪，2003；彭平根，丁彪，苏永华，2005；铁鑫鑫，2010)等设计与实施的影响。如何准确衡量测评组织者的测评的有效性，对于拓宽无领导小组讨论甚至评价中心技术在企业和政府部门招聘和晋升实践应用具有重要的意义。

研究者一直致力于以测评结果的信效度衡量测评实施的科学性。众多研究采用 Kendall 和谐系数或 Cronbach's $\alpha$ 系数检验评委一致性(田效勋，车宏生，2009；姚若松，梁乐瑶，苗群鹰，2011)，然而 Kendall 和谐系数或 Cronbach's $\alpha$ 系数只反映所有评委在各维度的评分差异，不能体现单一评委的评分一致性。因而有研究者结合 Kendall 和谐系数和 Spearman 相关系数分析，以评价评委间评分一致性和所有评委评定信度(郭朝晖，2011)但是这些信度指标严重依赖样本，对解释评委在人事测评中差异和重要性分析的力度有限，难以为测评者选拔评委和衡量评委评分准确性、客观性提供有意义的反馈作用。除此之外，研究者还重视评价中心的效度问题。王忠军和龙立荣(2006)采用维度相关性和验证性因素分析验证评价中心结构效度，为测评维度建构提供了良好的信息。由于跟踪被试工作绩效效标工作量大且耗时较长，有关评价中心效标关联效度的研究不多(张笑菡，2011)，同时关于评价中心的区分效度和聚合效度证据不足，因而评价中心的结构效度一直不太理想(王忠军，龙立荣，2006；卞冉，高钦，车宏生, 2013)。

LGD 的有效实施及应用，不仅受到统计信效度的影响，还有赖于整体设计和实施过程的严格控制，其中评委是决定 LGD 测评有效性的关键因素之一。评委的认知负荷较高，需在短时间内过滤无关信息、评定应聘者胜任特征水平，容易产生评分系统偏差。同时受评委经验、应聘者特点、测评题目、测评维度等因素影响，评估分数存在标准不一、跨情境不稳定的特点。因此，如何降低各种偏差(特别是评委偏差)不仅是 LGD 实证研究的重点，也是人事测评实践函需解决的现实问题。

评委培训是降低评分偏差、保证评分质量的重要方法之一。早期研究指出，针对评分误差、绩效

收稿日期: 2012-12-27

* 广东省哲学社会科学"十一五"规划项目(GD10CGL08)和广州市哲学社会科学发展"十二五"规划项目(13G59)资助。

通讯作者: 姚若松, E-mail: yaoruosongmmm@163.com

维度及绩效标准的培训都有利于促进评分的准确性(Smith, 1986)。三种培训方法中，针对评分误差培训效果最明显(Gatewood, Lahiff, Deter, & Hargrove, 1989)。评分误差培训为评委提供各种评分误差现象(包括宽松效应、严紧效应和晕轮效应等)。经评分误差培训的评委能降低宽松效应和严紧效应的发生(Bernardin & Pence, 1980; Noonan & Sulsky, 2001; Roch & O'Sullivan, 2003; Uggerslev & Sulsky, 2008)。Woehr 和 Huffcutt 于 1994 年提出了使用行为观察培训取代早期的培训方法。随着对评委培训的进一步研究，有研究者指出参照系培训法更能增加行为和能力特质评分的可靠性(Jackson, Atkins, Fletcher, & Stillman, 2005)。研究者针对评价中心的不同测量技术使用的培训方法多种多样，毋庸置疑的是，加强评委培训确实能改善人事测评中的偏差问题(Highhouse, 2008; Segrest, Perrewe, Gillespie, Mayes, & Ferris, 2006)。

Rasch 于 1960 年为改善早期项目反应理论的不足，开发了单参数 Rasch 模型，该模型可得到独立于项目难度的被试能力值。其后, Linacre 提出多侧面 Rasch 模型(Many-Facet Rasch Model, MFRM)。MFRM 增加了被试能力、项目难度、评委评分、评分标准和任务性质等变量(或侧面)。MFRM 根据评委宽严度数据，评估评委一致性程度，纠正评分差异明显的成绩，检查评定标准的功能，并检测不同侧面交互作用。研究者将该模型用于统计主观评定测验。目前关于 MFRM 的实证研究集中评定高等教育学生学业成绩、生活质量的调查和病患部位损伤康复程度的鉴定。MFRM 能清晰地察觉评委评定的变化并分析评委评分是否存在偏差(Allen & Schumacker, 1998; Farrokhi & Esfandiari, 2011; Myford & Wolfe, 2004)。

评价中心测评体系包括一系列测评形式，如案例分析、情景判断、结构化面试、公文筐、无领导小组讨论、演讲等。国内应用 MFRM 研究测评中心技术相关测评形式多集中于结构化面试(孙晓敏, 张厚粲, 2006; 孙晓敏，薛刚, 2008)。研究者根据结构化面试不同影响因素(考生、评委、性别及时间)，检测评委的偏差行为。同时，研究者也通过 MFRM 分析各种主观评价技术(俞宗火，唐小娟，王登峰, 2009; 田清源, 2007)。但是不同形式的评价技术在人事测评的适用性和有效性各不相同。此外，由于各种评价中心技术的操作过程和判断指标各异，关于 MFRM 在人事测评的应用还需进一步深入探究。更加重要的是，基于 MFRM 对无领导小组讨论的分析讨论，能让测评者深入探索 LGD 操作影响因素，进一步理解评分的各种差异。无领导小组讨论的 MFRM 研究帮助测评者更好的控制和预防评分偏差，从而全面准确的理解和提高人事测评中的评价技术。

回顾以往研究，大量关于 LGD 测评的实证研究侧重使用 CTT 方法，到目前为止还较少有研究应用 MFRM 分析 LGD 实证的偏差问题，多方面综合考虑评委评分结果、被试能力、评价维度、评定量表等级影响的实证研究不足。本文将 MFRM 引入 LGD 测评分析，一方面拓宽 LGD 测评统计研究的局限性，进一步检验 MFRM 的合理性，另一方面为深化评价中心技术的多种测评研究提供更广泛的参考价值。

## 2 方法

### 2.1 测评被试

研究数据来自某大学 77 名被试参加 LGD 测评结果，编号为 1~77。所有被试随机分为 11 组，每组 7 人，评委由 6 人组成，其中评委 A、B、C 经评分培训(培训包括了各种评分误差现象的授课说明及试评)，而评委 D、E、F 未培训。所有评委与被试互不相识，消除因熟悉性导致的评分误差。

### 2.2 测评施测程序

评委组织被试参加测评，并根据被试表现对语言表达能力、分析归纳能力、组织协调能力、应变与压力承受能力、行为表现及风度五维度实行 10 点量表评分。

### 2.3 测评结果处理

采用 MFRM 的计算机统计程序 FACETS 软件，该软件由 Linacre 编制，版本为 3.70.1。研究建构被试、评委和评分维度三个侧面。每个输出侧面(Facets)结果包括观测平均值(Obsvd Average)、符合度统计量(Fit Statistics)等统计指标，其中符合度统计量分为加权拟合统计量(*Infit*)和未加权拟合统计量(*Outfit*)。通过这些指标考察被试能力值、评委评分宽严度、评分内部一致性、维度难度和评定等级，并具体探讨各种偏差。

## 3 结果

### 3.1 被试能力的结果分析

以被试能力值为侧面一，分析各被试在 LGD 的表现。表 1 是部分被试的能力估计值结果。第二

列观测平均值是各评委对被试五维度评分的平均值。第三列是 MFRM 模型计算的被试能力估计值。第五列是评委对被试评分的一致性程度。*Infit* 数值大小反映评委使用评分量表评定被试成绩的一致性。测评者可通过观测值与模型预测变异值关系判断侧面是否拟合模型(Eckes, 2009)。*Infit* 在一定范围内浮动, 其取值范围没有严格规定, 评判标准由研究目的和数据量而定。测评精度随可接受范围值域的缩小而提高, 有研究认为 *Infit* 在 0.5~1.5 (Linacre, 2012)为可接受范围, 但更精确的 *Infit* 值域为 0.8~1.2 (孙晓敏, 薛刚, 2008)。可接受范围越窄越能体现测评的规则性和权威度, 并增强测评者对各侧面的监控力度。当 *Infit* 大于 1.2 为非拟合(misfitting), 评委评分明显大于模型预测值, 评分一致性差; 相反, *Infit* 小于 0.8 为过度拟合(overfitting), 即评委评分明显小于模型预测值, 评分过于一致, 并未区分不同的评分等级。第六列 *Outfit MnSq* 是未加权均方拟合统计量, 其值容易受极端数据的影响, 因而一般以 *Infit* 值作为检验侧面的主要指标。

经 MFRM 分析, 77 名被试的能力值范围是 –2.72 Logits 至 4.58 Logits, 平均能力为 0.46 Logits, 其中 49 号被试能力水平最高, 75 号被试能力水平最低。被试能力的 *Infit* 值高于 1.2 的有 21 名, 占总被试的 27%, 表明不同评委对这 21 名被试评分各有差异, 评分有偏差; 51 号被试的 *Infit* 值高达 4.51, 偏离可接受范围最大。*Infit* 值低于 0.5 的有 34 名, 占总被试的 44%, 表明这些评委对其评分过于一致, 需要校正某些异常分数; 其中 61 号被试 *Infit* 值最低, 即评委对其评分差异太小, 存在一定的集中趋势。

MFRM 分析的分隔信度(*Separation Reliability*)值越大, 则差异越显著。MFRM 的分隔信度在 0~1 之间浮动, 数值越接近 0 越无法区分被试差异, 数值越接近 1 越能辨别被试能力。分隔指数(*Separation*)是调整测量误差后的标准偏差估计值(*Adj S.D.*)除以标准误差均方根(*RMSE*, Root Mean Square Standard Error)后得到的数值, 表示测量的有效性。表 1 可知分隔指数为 7.46, 分隔信度为 0.98, 表明被试能力水平存在显著差异。对被试能力估计值进行卡方检验($\chi^2(76)= 4071.3$, $p<0.01$), 表明被试能力估计值差异显著。

### 3.2　评委宽严度和评分内部一致性结果分析

被试得分受评委宽严度和评分内部一致性的影响, 以 FACETS 分析 6 位评委的评分特点, 结果见表 2。宽严度数值越高, 代表评委评分严格; 反之, 评委评分宽松。由表 2 可知, 6 位评委的宽严度跨度为 1.57 Logits, 其中评委 C 最严格, 宽严度为 0.65 Logits; 评委 E 最宽松, 宽严度为-0.92 Logits; 评委 D 的宽严度低于 0, 说明其评分也偏宽松。其中培训组的评委 A、B、C 之间的宽严度跨度为 0.22 Logits, 而未培训组的评委 D、E、F 之间的宽严度跨度为 1.08 Logits。评分差距较未培训组评委评分差距小, 这在一定程度上反映培训组评委评分较为一致, 表明培训作用显著。第五列 *Infit* 值, 反映评委评分内部一致性信息。MFRM 容许评分内部一致性在一定范围内波动, 但超出可接受范围说明评分稳定性较差。从表 2 得知, 评委 C 的 *Infit* 值小于 0.8, 该评委对各被试的评分小于模型预期数值, 说明该评委采取保守策略, 不轻易评定高分, 呈现过度一致性; 而评委 F 的 *Infit* 值为 1.60, 该评委评分超过模型预期的变化幅度, 其内部一致性较低。而其他的评委 *Infit* 值为 0.84~0.99, 表明内部一致性较好。对评委宽严度进行卡方检验, 结果显示, $\chi^2(5)= 680.5$, $p<0.01$, 表明 6 位评委宽严度存在显著差异。除了评委 C 和 F, 其他评委呈现较好的内部一致性, 即他们对自身宽严度标准掌握较好。

表 **1**　部分被试的能力估计值统计结果

| 被试 | 观测平均值 | 能力值 | *S.E.* | *Infit MnSq* | *Outfit MnSq* |
|---|---|---|---|---|---|
| 49 | 8.27 | 4.60 | 0.31 | 1.39 | 1.45 |
| 48 | 7.70 | 3.23 | 0.26 | 1.02 | 1.00 |
| 51 | 7.70 | 3.23 | 0.26 | 4.51 | 4.43 |
| 61 | 4.57 | –0.96 | 0.18 | 0.34 | 0.36 |
| 75 | 2.50 | –2.72 | 0.17 | 1.20 | 1.27 |
| 平均值 | 5.71 | 0.46 | 0.20 | 1.00 | 1.00 |
| 标准差 | 1.26 | 1.55 | 0.03 | 0.59 | 0.59 |

注：*RMSE*：0.21　*Adj S.D.*：1.54　*Separation*：7.46　*Separation Reliability*：0.98

### 3.3　各维度评分结果分析

维度评分结果表明被试维度难度。表 3 难度估计值纵列显示, 组织协调能力难度最大, 行为表现及风度难度最小。*Infit* 值显示所有维度大致吻合模型期望, 只有语言表达能力略低于 0.8, 仍在可接受范围。对维度难度进行卡方检验, 结果表明, $\chi^2(4)= 324.0$, $p<0.01$, 即不同维度难度存在显著差异。

表 2 6 位评委的宽严度及评分一致性统计结果

| 评委 | 观测平均值 | 校正平均值 | 宽严度 | *Infit MnSq* | *Outfit MnSq* |
|---|---|---|---|---|---|
| C | 5.16 | 5.34 | 0.65 | 0.56 | 0.58 |
| B | 5.35 | 5.52 | 0.44 | 0.84 | 0.88 |
| A | 5.36 | 5.53 | 0.43 | 0.87 | 0.88 |
| F | 5.60 | 5.76 | 0.16 | 1.60 | 1.78 |
| D | 6.33 | 6.48 | –0.75 | 0.99 | 0.97 |
| E | 6.46 | 6.61 | –0.92 | 0.86 | 0.94 |
| 平均值 | 5.71 | 5.87 | 0.00 | 0.95 | 1.00 |
| 标准差 | 0.50 | 0.49 | 0.61 | 0.32 | 0.37 |

注： *RMSE*：0.06 *Adj S.D.*：0.61 *Separation*：10.90 *Separation Reliability*：0.99

表 3 五维度评分统计结果

| 维度 | 难度估计值 | *S.E.* | *Infit MnSq* | *Outfit MnSq* |
|---|---|---|---|---|
| 组织协调能力 | 0.60 | 0.05 | 1.10 | 1.15 |
| 应变与压力承受能力 | 0.28 | 0.05 | 0.97 | 1.02 |
| 分析归纳能力 | 0.01 | 0.05 | 1.00 | 1.08 |
| 语言表达能力 | –0.39 | 0.05 | 0.78 | 0.83 |
| 行为表现及风度 | –0.49 | 0.05 | 0.86 | 0.94 |
| 平均值 | 0.00 | 0.05 | 0.94 | 1.00 |
| 标准差 | 0.41 | 0.00 | 0.11 | 0.11 |

注: *RMSE*：0.05 *Adj S.D.*：0.40 *Separation*：7.95 *Separation Reliability*：0.98

### 3.4 评定等级分析

表 4 为 6 位评委评定等级统计结果。由表 4 可知评委们并未使用第 10 等级，同时大多数评委过分使用第 4~8 等级。其中评分为第 6 等级的次数最多，第 5 等级次之，可见评委评分呈现一定的集中趋势。平均能力值代表等级与被试能力关系。一般而言，能力值随评定等级增加而提高。表 4 结果呈现此趋势，说明 6 位评委评分大致给予与被试能力相符的分数，评分准确度较高。从表 4 可知，不同等级的平均能力值并非以相同的指数递增，等级 7 到等级 8 能力平均增值为 1.15 Logits，是九个等级中升幅最高的，表明被试要获得等级 8 必须付出最多努力。

表 4 评定等级统计表

| 等级 | 次数 | 频率% | 平均能力值 | 预测能力值 | *Oufit MnSq* | 估计能力阈限 | *S.E.* |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 2 | –2.87 | –2.65 | 0.7 | | |
| 2 | 54 | 2 | –2.17 | –2.20 | 1.1 | –2.61 | 0.18 |
| 3 | 117 | 5 | –1.66 | –1.65 | 0.9 | –2.71 | 0.13 |
| 4 | 295 | 13 | –0.93 | –1.01 | 1.2 | –2.26 | 0.10 |
| 5 | 458 | 20 | –0.41 | –0.30 | 0.9 | –1.10 | 0.07 |
| 6 | 561 | 24 | 0.55 | 0.51 | 0.9 | –0.11 | 0.06 |
| 7 | 441 | 19 | 1.58 | 1.49 | 1.0 | 1.22 | 0.06 |
| 8 | 288 | 12 | 2.73 | 2.71 | 1.0 | 2.51 | 0.08 |
| 9 | 51 | 2 | 3.29 | 3.91 | 1.4 | 5.06 | 0.16 |

图 1 为评定等级变化概率曲线图。其中，纵轴为被试获得某一评分等级的概率，横轴为能力全域，范围为–6~8 Logits。等级 2 和 3 没有明显划分的独立峰值，且等级 1~3 与等级 4~7 峰值接近。观察等级高峰可预测一定能力区域的被试获得相应等级评价的可能性(孙晓敏，张厚粲, 2007)。如能力值为 4 Logits 的被试最有可能获得评委们等级 8 的评定。

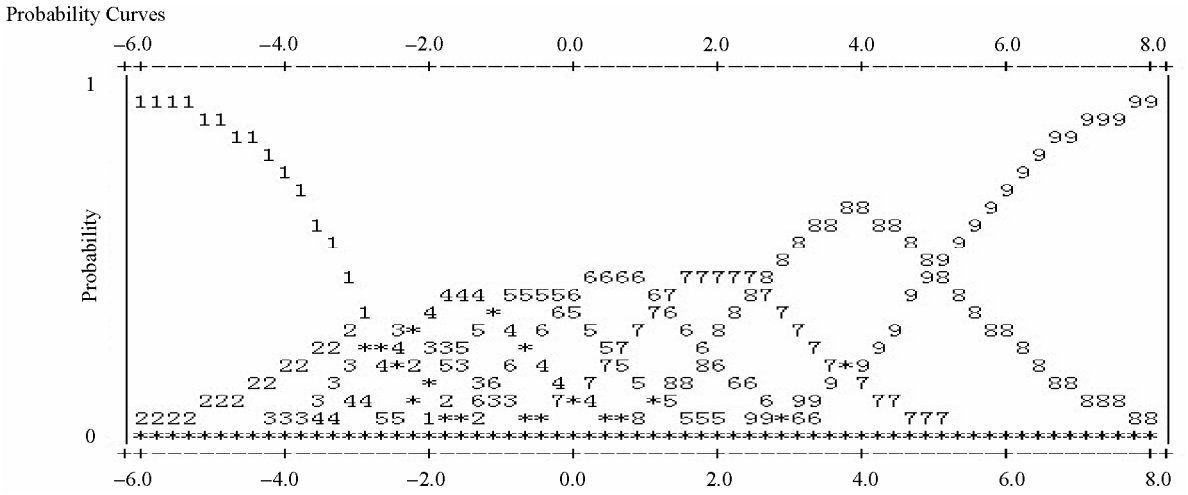


图 1 评定等级变化概率曲线图

### 3.5　偏差分析(Bias Analysis)

偏差分析可判断不同研究侧面间的交互作用(如评委与被试、评委与维度、评委、被试与维度三者间)是否存在显著偏离模型估计值的评分，以此追踪评委对被试或维度的评分差异问题、识别评委能否保持一贯的严宽度，判断评委对测评维度理解的透彻性和评分准确性。

#### 3.5.1　评委与被试的偏差分析

MFRM 估算的偏差 $t$ 值若大于 2 或者小于–2，则被视为差异显著 (Farrokhi, Esfandiari, & Schaefer, 2012)、评分存在较大分歧。经汇总，评委与被试偏差显著次数为 79 次，约占 462 个交互作用组合的 17%，其中正负向的偏差各占 50%，正的 $t$ 值表示评委对被试的评分比 FACETS 模型所预测分值更宽松，反之亦然。统计结果表明各评委都发生评分偏差，其中评委 F 发生评分显著偏差次数最多，共发生了 21 次显著偏差，占所有偏差总数的 26.6%。图 2 列出评委 F 所有的评分偏差，其中对 51 号被试评分偏差结果最显著，$t$ 值达-7.84。卡方检验结果($\chi^2(462)= 1068.7$, $p<0.01$)表明在整体上评委与被试评分的交互作用显著。

#### 3.5.2　评委与维度的偏差分析

评委与维度的显著次数为 9 次，约占 30 个交互作用组合的 30%。表 5 可看出评委对维度评分宽松还是严格，其中评委 C 对维度评分没有发生显著偏差，评委 E 评价组织协调能力和行为表现及风度维度显著偏宽，与上述分析评委 E 评分宽松的结果一致。培训组 A、B、C 评委发生偏差的次数仅占 2/9。偏差次数最多的维度是组织协调能力，说明评委在此维度容易出现评分偏差。

#### 3.5.3　评委、被试与维度的偏差分析

评委、被试与维度的偏差结果以 FACETS 软件的异常分数表示。异常分数可作为评委、被试与维度的整体吻合性指标，帮助测评者快速定位某个测量差异值。MFRM 先估算评委对被试某维度评分的预测值，然后转换计算观测分数和模型预测分数差异的标准残差。当标准残差(StRes)在正负 3 个标准差之外时，则为异常分数。分析数据后得出 21 个异常分数，如表 6 所示。

从表 6 可知，异常分数只出现于对个别被试个别维度的评定，可见评分总体上较合理。其中异常分数较多出现在对被试组织协调能力、应变与压力承受能力、行为表现及风度的判断。所有被试中，评委对 51 号被试评分异常分数最多。另外，统计所有评委评分异常分数发现评委 F 评分异常次数最多，而培训组评委 A、B、C 异常分数少甚至没有。

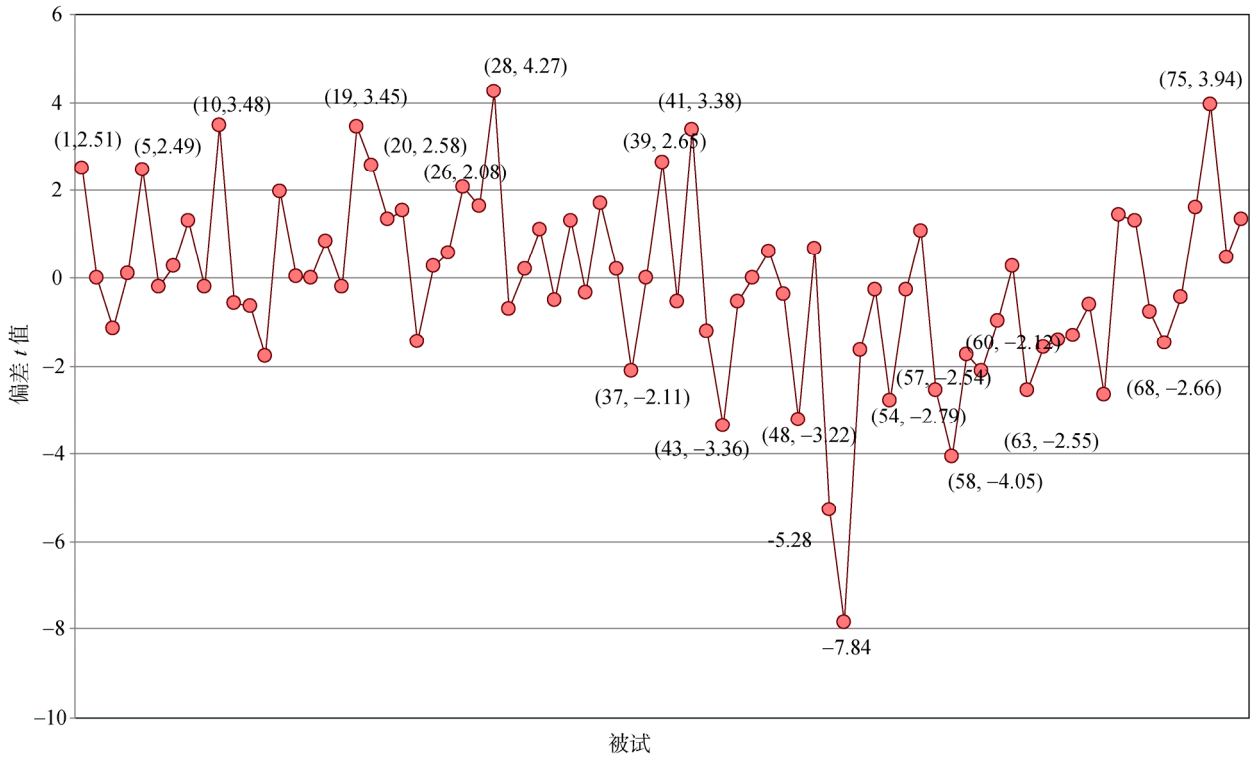


图 2　评委 F 对被试的偏差评定图

表 5 评委与维度的偏差数(评分严格/评分宽松)

| 评委 | 组织协调能力 | 分析归纳能力 | 语言表达能力 | 行为表现及风度 | 总计 |
|---|---|---|---|---|---|
| A | 1/0 | | | | 1/0 |
| B | 1/0 | | | | 1/0 |
| D | 0/1 | | | 0/1 | 0/2 |
| E | 0/1 | | 1/0 | 1/0 | 2/1 |
| F | | 1/0 | 0/1 | | 1/1 |
| 总计 | 2/2 | 1/0 | 1/1 | 1/1 | 5/4 |

表 6 LGD 评定异常分数表

| 被试 | 评委 | 维度 | 观测分 | 预测值 | *StRes* |
|---|---|---|---|---|---|
| 51 | F | 分析归纳能力 | 4 | 7.7 | –5.2 |
| 51 | F | 应变与压力承受能力 | 4 | 7.5 | –4.8 |
| 51 | F | 行为表现及风度 | 5 | 7.9 | –4.4 |
| 50 | F | 组织协调能力 | 3 | 6.5 | –4.1 |
| 71 | D | 分析归纳能力 | 2 | 5.6 | –3.9 |
| 54 | F | 应变与压力承受能力 | 4 | 7.0 | –3.7 |
| 69 | E | 行为表现及风度 | 5 | 7.6 | –3.7 |
| 10 | F | 语言表达能力 | 9 | 5.7 | 3.6 |
| 62 | D | 分析归纳能力 | 2 | 5.3 | –3.6 |
| 40 | E | 行为表现及风度 | 4 | 6.8 | –3.4 |
| 59 | F | 分析归纳能力 | 4 | 6.9 | –3.4 |
| 28 | F | 语言表达能力 | 9 | 6.1 | 3.3 |
| 62 | D | 应变与压力承受能力 | 2 | 5.1 | –3.3 |
| 28 | F | 行为表现及风度 | 9 | 6.2 | 3.2 |
| 43 | E | 分析归纳能力 | 5 | 7.4 | –3.2 |
| 71 | D | 组织协调能力 | 2 | 5.1 | –3.2 |
| 30 | A | 组织协调能力 | 2 | 5.0 | –3.1 |
| 54 | E | 行为表现及风度 | 6 | 8.0 | –3.1 |
| 68 | F | 应变与压力承受能力 | 3 | 5.8 | –3.1 |
| 70 | D | 语言表达能力 | 3 | 5.8 | –3.1 |
| 75 | F | 应变与压力承受能力 | 5 | 1.9 | 3.1 |

# 4 讨论

本研究运用 MFRM 模型分析无领导小组讨论数据，利用 FACETS 软件将被试、评委、维度三个侧面置于共同的 Logits 量尺上，在分离各侧面之间的相互作用情况下单独获得被试的能力、评委宽严度、评分内部一致性和维度难度的估算值，使被试能力值独立于评委各种偏差和维度难度，更接近真实能力。此外，利用 *Infit* 值检测评分出现的误差分数，提供更多个体得分反馈信息，提高评分的准确性和公正性。

## 4.1 被试能力水平的估计

MFRM 最大特点在于能独立于各种偏差估计被试能力值，为测评提供全新的视角—— 根据能力排序筛选被试。同一能力水平的被试，会因评委界定评分标准的特异性而获得差异悬殊的分数。如 48 号和 51 号的被试的能力值相同，但 *Infit* 值差异悬殊，体现在评委对 48 号被试表现评价趋于一致，对 51 号被试表现判断异议，意见难达一致。基于传统测评，两位被试测评排名相同，应有相似的人事决策结果；但基于 MFRM 模型探测的异常 *Infit* 值，将提醒测评者判断两被试的差异，复查测评评分并继而甄别被试其他综合能力。根据传统测评的决策精简易行，但容易忽略评委评分特异性，导致决策的片面性。孙晓敏和薛刚(2008)运用 MFRM 估计被试面试成绩，发现相同能力值与其面试分差距较大，导致被试排名变化。以能力水平作为用人决策标准，避免因评委偏差而发生评定偏高或偏低的情况。当被试能力值相同时，测评者可根据规定范围的 *Infit* 值，排除超出值域的被试评分，提高评判合适人选标准，使人事决策过程具科学性。

## 4.2 评委因素对评分的影响

评委因素影响评分，评委在不同的环境会表现各种偏差。这种差异有多方面的原因。例如，在无领导小组讨论施测中，评委需全程保持高度注意力集中，并避免无关干扰影响；不同组被试产生对比效应；评委对被试期望度、评分等级的把握受个体经验影响等。以评委的宽严度和内部一致性作为考察评委主观评分差异的指标能准确地区分各评委评分误差，为及时校正评委的错误评分提供可靠依据。

评委宽严度在一定程度影响被试得分，评委越严厉，被试获得高分的可能性则越低。MFRM 将各评委宽严度数字化，为鉴别评委提供客观指标。以上结果显示 6 位评委评分宽严度差异显著。一般来说，评委宽严度的跨度越小，则表明评分更为客观。其中，最严格的评委 C 与最宽松的评委 F 相差 1.56 Logits，被试能力水平高低的跨度值为 7.28 Logits，评委的宽严度跨度低于被试能力水平跨度值的四分之一，因此 6 位评委宽严度的差异在总体上不会对被试得分产生决定性的影响。

评分内部一致性是衡量同一评委对所有被试评分稳定性的指标。FACETS 通过 *Infit* 值判断评委的内部一致性，*Infit* 值的增大表明评委评分过度一致到过度随机的变化。评分是一致还是随机，影响测评结果的严密性。经典测量理论在分析总体评委评分在多大程度上达到一致，无法检测独立评委评分的稳定性。MFRM 提供评委个体内部一致性得分，

从被试、维度、评分等级等角度深入分析评委于不同情境的评分一致性，因而测评者可跟踪、监控或培训内部一致性明显不一致的评委。

此外，结果发现培训组的评委宽严度比未培训组评委宽严度集中，表明了培训的有效性。专业化培训，有助于消除评委偏见，保持评分独立，增强对测量维度的敏感性。评委宽严度和内部一致性都将影响被试得分，因而需要增加评委测评前的评分练习。一方面，加强评委测评信息收集能力，更大程度地记录与测评维度相关的被试行为；另一方面，讲授详细的评分标准及典型行为，帮助评委理解测量量表，准确区分和鉴别被试各项表现。此外研究指出评委类型是影响评委水平的重要因素(Lievens, 1998)。相比于人力资源经理，心理学家的知识更为丰富，更公平和客观地评价被试的各项行为表现(Gaugler, Rosenthal, Thornton, & Bentson, 1987; Sagie & Magnezy, 1997)。根据经验及专业知识，并利用 MFRM 探测的宽严度不一的数据结果，作为评委筛选标准，从而降低评委偏差，将是人事测评的重大突破。

### 4.3　维度难度的 MFRM 分析

无领导小组讨论评分主要通过评委对被试测评行为表现的观察评价进行，不同维度评定会影响被试的总体得分。结果表明组织协调能力估计的难度最大，行为表现及风度维度估计难度最小。相比于其他维度，评委容易根据被试的外在特征判断其行为表现及风度维度。而被试展现个人优势时间有限，且组织协调能力难以通过被试的简短言语或其他外在行为展现，使得评委对组织协调能力维度的判定标准更严格，倾向给予较低的评价。同时，组织协调能力维度的异常分数较多的结果也可表明评委对此维度的评价尺度把握困难。从此可见，测评者不仅要加强此维度评分标准定义，还要明确规定此维度不同评分标准。另外，不同维度难度差异显著说明维度难度设置能较好地区分被试的能力，展示不同被试的行为和能力差异。MFRM 对维度难度的鉴定，协助测评者针对岗位胜任特征，设置不同的能力维度难度，以更好地辨别与岗位匹配的被试。

### 4.4　评定等级的 MFRM 分析

评定等级的使用次数、每个等级与被试能力符合度是衡量评分等级设置合理性的重要依据。理想状态下，被试能力跟得分呈正比例，即高能力被试对应高分数段的分数。观察评定等级变化概率图可直接推断将各等级估计的能力范围。若概率曲线有独立峰值，那么此能力段的被试最有可能得到该分数段的评定等级。如评定等级分布图所示，除等级 3 没有独立的峰值，其他等级均有独立的峰值。此外，统计各等级的使用频率还可判断评委是否存在集中趋势，以检验评委是否正确鉴别被试能力水平。等级 5、6、7 的高使用频率，一方面可推测评委评分采取安全策略，经常使用中间段分数以避免造成与其他评委评分冲突，表明评委存在一定的评分集中趋势，另一方面也表明参加测量的部分被试能力处于平均水平。在无领导小组讨论测评中，要求评委能客观依据评分标准，体现不同等级或标准之间的差异性，而不是过于集中评定趋中等级或分数。如果评委集中使用某一阶段的分数值，将无法判断被试、维度特殊性，其他分数值也不存在设置意义。通过观察每评委评定等级的使用情况能判断评委的评分倾向，快速地为建立综合的评委评分反馈报告，以此作为考核和选拔评委的标准之一。

### 4.5　偏差分析

偏差分析结果帮助测评者快速追踪评委具体的评分差异，健全评委评价方案，修正被试得分，保持测评公正性与准确率。尽管某些评委有良好的评分内部一致性，但受测评过程诸多因素影响，无法在整个测评过程一直保持准确评分，因而引入偏差分析，协助测评者鉴定评委具体的评分问题，快速定位详细的测量结果，为筛查评委及测评分数提供便捷方式。

根据各种偏差分析，汇总每位评委在不同偏差上的表现(见表 7)。

由表 7 可以发现，非培训组的各种偏差次数多于培训组，其中评委 F 评分偏差最多，说明培训可使评委更好理解测评维度的内涵，增强一致性，减少不必要的偏差，这与过往研究结果相符(Lievens, 1998; 彭平根等, 2005)。由此可见，降低人事测评误差简单直接的方法就是加强评委培训。

在评委与被试偏差的分析中发现，各评委都表现出一种双向变化：有时对被试的评分比 FACETS 模型预测的更宽容，有时则更严厉。卡方检验结果表明，尽管大部分评委的内部一致性较好，仍不可避免地对特定被试评分时而宽松时而严格。这些都与无领导小组讨论实施的复杂性有关。例如，所有被试无法在一次无领导小组讨论中完成，需被分为不同组别。评委无法跨组别对比所有被试，极易造成评价上的动态变化。此外，无领导小组讨论施行

表 7 偏差分析汇总

| 组别 | 评委 | 评委与被试偏差次数(百分比) | 评委与维度偏差次数(百分比) | 评委、被试、维度偏差次数(百分比) | 评委偏差次数汇总 |
|---|---|---|---|---|---|
| 培训组 | A | 14 (17.95%) | 1 (11.11%) | 1 (4.76%) | 16 (14.81%) |
| | B | 15 (19.23%) | 1 (11.11%) | 0 (0.00%) | 16 (14.81%) |
| | C | 3 (3.85%) | 0 (0.00%) | 0 (0.00%) | 3 (2.78%) |
| 总计 | | 32 (41.03%) | 2 (22.22%) | 1 (4.76%) | 35 (32.41%) |
| 非培训组 | D | 11 (14.10%) | 2 (22.22%) | 5 (23.81%) | 18 (16.67%) |
| | E | 15 (19.23%) | 3 (33.33%) | 4 (19.05%) | 22 (20.37%) |
| | F | 20 (25.64%) | 2 (22.22%) | 11 (52.38%) | 33 (30.56%) |
| 总计 | | 46 (58.97%) | 7 (77.78%) | 20 (95.24%) | 73 (67.59%) |

时间较长，而多组别的实施更增加了评委负担，评委无形受个人因素主导，忽略评分标准的划分方式，从而带来一定的偏差结果。分析所有偏差结果发现，51 号被试的评分偏差发生次数最多，测评者可着重检验此被试测评分数，避免因评委偏私而影响测评秩序的现象发生。

观察评委与维度偏差次数发现，评委在组织协调能力维度发生的偏差评价更多。一方面是被试自身对题目理解不同，导致回答不一致而造成评分的困难；另一方面是评委对这维度把握不到位。对于偏差较多的维度，测评者既要考虑试题设置的难度问题，又要增强维度评分标准的细致性。

从评委、被试与维度的分析可直接获取测评过程中不合理的分数，如对于 51 号被试的分析归纳能力维度，模型预测的分数值为 7 分，而评委 F 的 4 分评定结果显著低于其他评委评分。跟踪评委 F 偏差评定图可清晰判断其偏差分数。测评者可根据偏差分数调整相关信息，进一步分析原因，更换或弃用该分值。而组织协调能力、应变与压力承受能力判断偏差多，这主要是这三个维度是评委对被试内在素质的评价，容易出现判断不一情况。

在关于考试的研究中，为提高评分的准确性，组织者会在评分前举行动员会和加强评分监督(李中权，孙晓敏，张厚粲，张立松, 2008)。而在人事测评实践中，测评者一般以评分培训来评委培训达到控制评委偏差、提高评委信度的目标。培训的内容包括制定统一的评分标准、介绍各种评分错误现象、加强以行为观察为基础的练习、增加特殊行为的记录及试评等。通过评委培训进行人事测评的最大优点是能直接降低评委层面的系统偏差，这是影响人事测评主观评分的关键因素。在实践中，既能根据评委某次评分的偏差对其可能产生的问题进行普遍性培训与特殊性培训，又能根据评委的评分偏差次数作为选择与筛选评委合格性的标准。另外，维度评分标准也是影响评分稳定性的因素之一，维度难度越大，评委越不能保持评分的一致性(Graham, Milanowski, & Miller, 2012)。测评者需创建具有明晰评价标准的评分量表，根据偏差显著的维度来评识别存在问题或会导致风险决策的维度，开展评分指导或完善维度评分标准。

## 5 结论

目前大多数评价中心测评的理论和实践研究多基于经典测量理论，本研究应用项目反应理论的多侧面 Rasch 模型，对无领导小组讨论测评中被试能力值的真实水平、评委宽严度和评分内部一致性、测评维度的难度与评定等级等特征进入更深入、细致的分析，探讨被试能力水平值差异、评委有效性、测评维度差异及各种偏差情况，并提出相应的应对措施。研究结果表明项目反应理论在研究评价中心的主观性测评具有优势，本研究可推广到评价中心技术其他测评形式研究和应用。

基于项目反应理论建立全面反映被试作答反应或表现行为与测评试题及能力水平间关系的非线性模型，估计出的能力水平不依赖于特定的试题样本，可将被试能力水平置于同一标尺上，有利于直接对比被试能力，使人员测评与选拔更具科学性；通过跟踪、监控测评分数，筛选内部一致性差异显著、评分过宽或过严的评委，达到降低评委偏差、提高评分准确性的效果；对测评维度的分析，能提升测评维度与测评目的匹配性；评委各种偏差情况作为选拔与培训评委的标准，促进了评估的客观化。项目反应理论在人事测评领域的应用与拓展，可突破经典测量理论无法定位到个体差异的不足，提供人事测评情景中各关键侧面的详细分析信息，因而，将项目反应理论应用于评价中心的研究，将

促进评价中心测评体系的科学化、规范化、精确化发展。

在评价中心测评的体系中，测评设计应同时包含 2~3 种甚至更多种测评。不同的测评形式，其对应测评的素质能力有差异，如公文筐主要是测评行政处理能力、严谨细致性、书面沟通与思维逻辑能力等，无领导小组讨论主要测评团队合作性、领导能力、人际关系处理能力、语言沟通能力，结构化面试主要考察被试的应急能力、应对压力能力、分析和解决问题能力等，由于测评形式特点的差异，其在评价中心测评体系中亦有差异。目前项目反应理论在评价中心领域的研究多为针对单种测评形式的验证与分析，未来可将测评形式作为影响测评结果的侧面进行设计，深化项目反应理论对同一测评体系的多种测评形式、以及各测评形式中的相关问题的综合研究，进一步探讨各种测评形式的特征(如测评稳定性)、差异及适用范围，为人事测评的发展提供科学的理论依据。

## 参 考 文 献


Allen, J. M., & Schumacker, R. E. (1998). Team assessment utilizing a many-facet Rasch model. *Journal of outcome Measurement, 2*(2), 142–158.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*(1), 60–66.

Bian, R., Gao, Q., & Che, H. S. (2013). The construct validity puzzle of assessment centers: Are we measuring dimensions or exercises? *Advances in Psychological Science, 21*(2), 358–371.

[卞冉，高钦，车宏生．(2013). 评价中心的构想效度谜题: 测量维度还是活动? *心理科学进展, 21*(2), 358–371.]

Eckes, T. (2009). Section H: 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*, 1–48.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531–1540.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *Japan Association for Language Teaching, 34*(1), 79–102.

Gatewood, R., Lahiff, J., Deter, R., & Hargrove, L. (1989). Effects of training on behaviors of the selection interview. *Journal of Business Communication, 26*(1), 17–31.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington, DC: Center for Educator Compensation Reform. Retrieved March, 23, 2012.

Guo, Z. H. (2011). How to test the reliability of the judges in Leaderless Group Discussion. *Chinese Talents,* (5), 55–57.

[郭朝晖．(2011). 无领导小组讨论，如何检验评委可信度．*中国人才,* (5), 55–57.]

He, G. L., & Yuan, L. (1998). “Leaderless group discussion” is a effective method to select personnels. *Chinese Pubillic Servants,* (2), 33.

[何广陵，袁翎．(1998). 选拔人才的一种有效方法—— “无领导小组讨论”．*中国公务员,* (2), 33.]

He, Q. (2003). Leaderless group discussion: The effective methods of evaluating modern Leadership’s qualities. *Administrative Tribune,* (5), 4.

[何琪．(2003). 无领导小组讨论：现代领导人才素质测评的有效方法．*行政论坛,* (5), 4.]

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1,* 333–342.

Jackson, D. J. R., Atkins, S. G., Fletcher, R. B., & Stillman, J. A. (2005). Frame of reference training for assessment centers: Effects on interrater reliability when rating behaviors and ability traits. *Public Personnel Management, 34*(1), 17–30.

Li, Z. Q., Sun, X. M., Zhang, H. C., & Zhang, L. S. (2008). Application of many-facet Rasch model in rater training for subjective items. *China Examinations,* (1), 26–31.

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

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection & Assessment, 6*(3), 141–152.

Linacre, J. M. (2012). *A user's guide to facets Rasch-model computer programs*. Chicago: MESA Press.

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

Noonan, L. E., & Sulsky, L. M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*(1), 3–26.

Peng, P. G., Ding, B., & Su, Y. H. (2002). A study of the Leaderless group discussion applied to choosing senior managers in enterprises. *Psychological Science, 25*(5), 576–579.

[彭平根，丁彪，苏永华．(2002). LGD 在选拔企业中高级管理人才方面的实证研究．*心理科学, 25*(5), 576–579.]

Roch, S. G., & O'Sullivan, B. J. (2003). Frame of reference rater training issues: Recall, time and behavior observation training. *International Journal of Training & Development, 7*(2), 93–107.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*(1), 103–108.

Segrest-Purkiss, S. L., Perrewe, P. L., Gillespie, T. L., Mayes, B. T., & Ferris, G. R. (2006). Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processing, 101,* 152–167.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*(1), 22–40.

Sun, X. M., & Gang, X. (2008). A many-faceted Rasch model analysis of structured interview. A*cta Psychologica Scienca, 40*(9), 1030–1039.

[孙晓敏, 薛刚. (2008). 多面 Rasch 模型在结构化面试中的应用. *心理学报, 40*(9), 1030–1039.]

Sun, X. M., & Zhang, H. C. (2006). An IRT analysis of rater bias in structured interview of national civilian candidates. *Acta Psychologica Sinica, 38*(4), 614–625.

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

Sun, X. M., & Zhang, H. C. (2007). A rating scale analysis by using the modern test theory. *Chinese Journal of Applied Psychology, 13*(3), 250–256.

[孙晓敏, 张厚粲. (2007). 结构化面试评定量表的现代测量学分析. *应用心理学, 13*(3), 250–256.]

Tian, Q. Y. (2007). Rasch experimental analysis of HSK performance test rating. *Psychological Exploration, 27*(1), 65–69.

[田清源. (2007). HSK 主观考试评分的 Rasch 实验分析. *心理学探新, 27*(1), 65–69.]

Tian, X. X., & Che, H. S. (2009). A study of the predictive validity of behavioral interview and Leaderless group discussion. *Psychological Science, 23*(1), 187–189.

[田效勋, 车宏生. (2009). 行为面试与无领导小组讨论的预测效度研究. *心理科学, 23*(1), 187–189.]

Tie, X. X. (2010). The issues need to be attentioned in Leaderless group discussion. *Operation and Managemen*t, (11), 93.

[铁鑫鑫. (2010). 浅谈无领导小组讨论中需要注意的问题. *经营管理者,* (11), 93.]

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711–719.

Wang, Z. J., Long, L. R. (2006). The construct validity of assessment center. *Advances in Psychological Science, 14*(3), 426–432.

[王忠军, 龙立荣. (2006). 评价中心的结构效度研究. *心理科学进展, 14*(3), 426–432.]

Woehr, D. J., & Huffcutt, A. L. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology, 4*, 189–216.

Yao, R. S., Liang, L. Y., & Miao, Q. Y. (2011). The validity of Leaderless group discussion in accessment center. *Leadership in Science,* (29), 39–41.

[姚若松, 梁乐瑶, 苗群鹰. (2011). 评价中心无领导小组讨论测评效度的实证研究. *领导科学,* (29), 39–41.]

Yu, Z. H., Tang, X. J., & Wang, D. F. (2009). A comparison of GT and IRT: An analysis of performance rating of men's 10 meters platform diving in Beijing Olympic Games. *Psychologica Sinica, 41*(8), 773–784.

[俞宗火, 唐小娟, 王登峰. (2009). GT 与 IRT 的比较: 北京奥运会男子 10 米跳台跳水分析. *心理学报, 41*(8), 773–784.]

Zhang, X. H. (2011). The reliability and validity of Leaderless group discussion. *Business Cuture,* (3), 163–164.

[张笑菡. (2011). 关于影响无领导小组讨论信度和效度因素的研究. *商业文化,* (3), 163–164.]

## The Application of Many-Facet Rasch Model in Leaderless Group Discussion

YAO Ruosong[1]; ZHAO Baonan[1]; LIU Ze[1]; MIAO Qunying[2]

([1]*Department of Education, Guangzhou University, Guangzhou* 510006, *China*)

([2] *School of Foreign Studies, Guangzhou University, Guangzhou* 510006, *China*)

### Abstract

Many-Facet Rasch model (MFRM) of Item Response Theory (IRT) is applied to performance assessment. Domestic and foreign researches applied MFRM in many fields such as analysis of various examinations, medical diagnosis, judgments of life quality and so on. In these assessment tests, ratings were influenced by a variety of factors among which judges played the most important part. This thesis mainly probed into issues covering subjects, judges, rating scales and rating deviation in Leaderless Group Discussion (LGD) of personnel assessment center in personnel assessment to improve the effectiveness and stability of assessment.

This study adopted the FACETS software, a MFRM computer statistics program, to establish 3 facets of subjects, judges and rating dimensions to analyze subjects' abilities, rater severity, inter-rater reliability, dimension difficulty and rating scales. Meanwhile, this study got results of deviation analysis of subjects and judges, judges and dimensions, deviation among judges, subjects and dimensions.

The results illustrated significant differences existed among levels of subjects' ability, rater severity,

dimension difficulty and the rating scale. Differences of rater severity generally did not affect the test scores of subjects. Except some judges, other judges' ratings had good internal consistency. Dimension difficulty could better distinguish subjects' ability but judges tended to concentrate on using an intermediate rating scale; The results of deviation analysis of judges and subjects, judges and dimension showed that untrained judges E, F had more rating deviations, so it was necessary to monitor their scores and strengthen the training of the two judges.

The application of MFRM, IRT's expansion, to assessment center evaluation could enable evaluators to make the employment decision by estimated ability level of subjects, design tests according to dimension difficulty, set the standards for training and selection referring to examine judges' ratings rater severity and inter-rater reliability, improve the assessment process based on a variety of deviation analysis, and finally promote scientific, standardized and precise development of evaluation system of assessment center.

**Key words**　leaderless group discussion; many-facet Rasch model; item response theory; personnel assessment