等级反应多水平侧面模型及其在主观题评分中的应用*

康春花 孙小坚 曾平飞**
（浙江师范大学教师教育学院，金华，321004）

摘 要 探讨了康春花，孙小坚和曾平飞（2016）提出的等级反应多水平侧面模型（GR–MLFM）在包含被试及评分者层面预测变量（完整模型）下的返真性和适用性。结果表明：（1）GR–MLFM 完整模型具有逻辑上和数理上的合理性，可用于主观题的评分情境，能较好地检测出评分者效应、影响因素及其影响程度；（2）在数学问题解决的评分实践中，评分员存在两种类型的评分倾向（宽松和严格效应），但绝大多数评分员的宽严度不明显；评分者的责任心可正向预测其严格程度，自信心可正向预测其宽松程度，而情绪稳定性和评分经验的预测作用不显著。

关键词 等级反应多水平侧面模型 完整模型 评分者影响因素 评分者效应

1 引言

主观题在考查学生的分析、理解和问题解决等能力上具有客观题无法比拟的优势（康春花，辛涛，2010; Attali, 2014）。然而，主观题的评分相对比较复杂，往往需要多名评分者对同一份作答做出较为一致性的评定。评分者对评分标准的理解（Hamp-Lyons, 2011; He, Gou, Chien, Chen, & Chang, 2013），培训经历（Wolfe, Jurgens, Sanders, Vickers, & Yue, 2014），人格特质（Bernardin, Cooke, & Villanova, 2000; Breuer, Nieken, & Sliwka, 2013）及自我监控意识等不尽相同（Jawahar & Stone, 1997），这些均会对评分者一致性（评分者信度）产生一定的影响（孙晓敏，张厚粲，2006; Attali, 2014）。这种由于评分关注点、评分经验及人格特质等因素使得最终的评定分数与被试的实际表现不完全一致，导致评分者信度降低的现象称为评分者效应（Wolfe, 2004）。

随着教育与心理测验的发展，测验研究者和实施者对评分精度的要求也越来越严格。为此，研究者不仅需了解主观题评分中评分者层面的变异有多大，同时还想了解哪些因素会影响这些变异，且影响大小如何。在这方面，多水平模型突显其优势，它不仅解决了传统回归分析中‘两步走’所带来的不足（刘红云，骆方，2008）；还可将测验结果的变异进行多层面的详细分解（Fox, 2010），从而实现对被试能力的精确估计。Wang 和 Liu（2007）将多面 Rasch 模型和多水平模型相结合，发展出基于广义分部评分模型（Generalized Partial Credit Model, GPCM）的评分者模型——广义多水平侧面模型（Generalized Multilevel Facets Model, G-MLFM），使传统的多水平线性模型发展为多水平多侧面 IRT 模型。Tutz（1990）和 Andrich（1995）指出，PCM 对同时性加工（simultaneous process）任务情境（如 Likert 量表或论文写作）的适用性更好。以 Likert 量表为例，被试在作答项目时，可以同时看到所有的选项，只需从中选择与自己相符的选项，并不需要像数学运算题那样，等到前一个步骤完成才能进行后一步的操作（继时性加工，sequential process）。

康春花，孙小坚和曾平飞（2016）基于等级反应模型（Grade Response Model, GRM）将多面 Rasch 模型和多水平随机系数（Multilevel Random Coefficient Model, MRCM）结合，形成等级反应多水平侧面模型（Grade Response Multilevel Facets Model, GR-MLFM），通过两个模拟研究探讨了在不包含任何预测变量（无条件模型）下的固定效应和随机效应模型的拟合情况，发现相比较而言，在 GR-MLFM 无条件模型下，随机效应模型的返真性和稳健性均较好，说明随机效应模型可以检测出评分者的随机效应。由此，为进一步推进和完善 GR-MLFM，本研究拟在随机效应模型基础上加入评分

* 本研究得到浙江省自然科学基金 (LY15C090003) 和教育部人文社会科学研究规划基金 (16YJA190002) 的资助。
** 通讯作者：曾平飞。E–mail: zpf@zjnu.cn
DOI:10.16719/j.cnki.1671–6981.20170631

者效应的影响因素，使其发展为可同时检测评分者效应、影响因素、影响大小的完整模型，并通过模拟和实证研究同时考察 GR-MLFM 完整模型的合理性及其在评分实践中的应用效果。

2 等级反应多水平侧面模型的建构

2.1 模型定义

等级反应多水平侧面模型（GR-MLFM）基于多面 Rasch 模型、MRCM 和 GRM 构建而成，兼具以上模型的特征，无条件模型（unconditional model）和完整模型如（1a - 1b）和（2）所示，在无条件模型和完整模型之间可构建其它一系列模型。

（1）无条件模型。无条件模型是指不包含预测变量的模型，水平 2 的被试和评分者参数既可固定也可随机，当两者都固定时，则只需水平 1，即等级反应多侧面模型。

水平 1 公式：

$$\text{logit}\left(p^{*}_{nijr}\right)=\log\left[\frac{p\left(x\geq j\right)}{p\left(x<j\right)}\right]=a_i\left(\theta_n-B_{ij}-D_{rn}\right) \quad (1a)$$

水平 2 公式：

$$\theta_n=\gamma_{00}+\mu_{0n},\qquad \mu_{0n}\sim N\left(0,\ \sigma^2_{\vartheta}\right)$$

$$B_{ij}=b_{ij}$$

$$D_{rn}=\gamma_{r0}+\mu_{rn},\qquad \mu_{rn}\sim N\left(0,\ \sigma^2_{r}\right) \quad (1b)$$

水平 1 式中，$p^{*}_{nijr}=\dfrac{exp[a_i(\theta_n-b_{ij}-D_{rn})]}{1+exp[a_i(\theta_n-b_{ij}-D_{rn})]}$，$p^{*}_{nijr}$和 $p\left(x\geq j\right)$二者均表示被试得等级 j 及其以上等级的概率（康春花 等 , 2016）。$p\left(x<j\right)$为等级 j 以下的概率。θ 为被试能力参数，B 为项目难度参数，a为项目区分度参数，D 为评分者宽严度参数。水平 2 式中，θ_n 为被试 n 的能力值（刘红云，骆方 , 2008）；γ_{00} 为被试群体的平均能力，μ_{0n} 为被试随机效应，服从$N\left(0,\ \sigma^2_{\vartheta}\right)$的正态分布。同理，$\gamma_{r0}$ 为评分者 r 的平均宽严程度，μ_{r0} 为随机效应，表示评分者和被试间存在交互效应（Wang & Wilson, 2005），μ_{r0} 服从$N\left(0,\ \sigma^2_{r}\right)$的正态分布。

（2）完整模型。完整模型指被试和评分者层面均包含预测变量的模型，可用于探讨被试和评分者层面各自的影响因素。完整模型的水平 1 公式与无条件模型的水平 1 公式相同，故不再呈现，其水平 2 公式为：

$$\theta_n=\gamma_{00}+\sum_{k=1}^{K}\gamma_{0k}X_{kn}+\mu_{0n},\qquad \mu_{0n}\sim N\left(0,\sigma^2_{\vartheta}\right)$$

$$B_{ij}=b_{ij}$$

$$D_{rn}=\gamma_{r0}+\sum_{q=1}^{Q}\gamma_{1q}W_{rq}+\mu_{rn},\qquad \mu_{rn}\sim N\left(0,\sigma^2_{r}\right) \quad (2)$$

式中，X 和 W 分别为被试和评分者层面的预测变量，γ_{0k} 和 γ_{1q} 为回归系数，K 和 Q 则为各自预测变量的个数。

2.2 参数估计

康春花等（2016）用 SAS 软件，通过自编程序实现了两水平 GR-MLFM 无条件模型固定和随机效应的参数估计。当模型相对较为简单时，SAS 可以快速地进行参数估计，但是当模型较为复杂时，SAS 的估计效率会降低（Wang & Qiu, 2013）。最近，贝叶斯方法被广泛地应用于复杂模型的参数估计（Cowles, 2013; Wang, Su, & Qiu, 2014）。为此，本文将使用 OpenBUGS 软件，通过 MCMC 算法实现 GR-MLFM 全模型中各参数的同时估计。

3 模拟研究：GR-MLFM 完整模型的拟合验证

3.1 研究目的

在康春花等（2016）证明了 GR-MLFM 可行的基础上，将被试和评分者层面的预测变量加入到模型中，构建完整的 GR-MLFM，并通过模拟研究进一步考察其适用性和稳定性。

3.2 研究模型

研究将被试的性别（*gend*）和评分者的自信心（*conf*）、情绪稳定性（*emo*）、评分经验（*exp*）、以及责任心（*res*）加入到公式（2）中，得到水平 2 的公式为：

$$\theta_n=\gamma_{00}+\gamma_{01}X_{gend,n}+\mu_{0n},\ \mu_{0n}\sim N\left(0,\sigma^2_{\vartheta}\right) \quad (3a)$$

$$B_{ij}=b_{ij} \quad (3b)$$

$$D_{rn}=\gamma_{r0}+\gamma_{11}W_{r1}+\gamma_{12}W_{r2}+\gamma_{13}W_{r3}+\gamma_{14}W_{r4}+\mu_{rn},\ \mu_{rn}\sim N\left(0,\sigma^2_{r}\right) \quad (3c)$$

其中 X 为被试的性别，W_1-W_4 分别为评分者的自信心、情绪稳定性、评分经验和责任心。

3.3 研究方法

3.3.1 模拟设计

评分设计：采用完全交叉设计，由 3 个评分者对 100 个被试在 4 道主观题上的作答情况逐一进行 1 到 4 级评分，1 表示被试作答最差，4 表示作答最好。

被试参数：设定被试群体的平均能力 γ_{00}= 0，随机效应从 N（0, 1）中随机生成。被试男女人数各占一半，并从 0~1 的二项贝努利分布中随机生成，0

为女生，1为男生。回归系数 γ_{00} 从标准正态分布 $N(0, 1)$ 中生成（Hung & Wang, 2012）。

评分者参数：3 个评分者的平均宽严度分别从均值为0，标准差为 .25 的正态分布中随机抽取（Hung & Wang, 2012）。三个评分者随机效应将从均值向量为（0, 0, 0），方差—协方差矩阵为 $\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ 的多元正态分布中随机生成。三个评分者层面的 4 个预测变量分别从 1~5 的五个等级中随机抽取三个表示；另外以李斌（2011）和宋慧鹏（2014）中关于 4 个评分者预测变量的回归系数为蓝本，假定其系数服从标准正态分布 $N(0, 1)$，并设定 4 个回归系数值在 -1 到 1 之间。项目参数设置见表 1（康春花 等，2016）。

表 1　各项目参数值的设定

项目（i）	等级（K）	区分度（ai）	B_{i1}	B_{i2}	B_{i3}
项目 1	4	1	-2	0	2
项目 2	4	1.5	-1	0	1
项目 3	4	2	-3	1	2
项目 4	4	2.5	-1.5	.5	1

注：每个项目的所有难度等级之和设定为 0，即 $\sum_{j=1}^{K-1} B_{ij} = 0$，另外，为保证模型在参数估计时可识别，研究将项目 1 的区分度固定为 1，在参数估计时不对其进行估计。

3.3.2　数据生成

使用 R 软件自编程序生成被试作答反应，步骤为：（1）生成被试能力；根据 3.3.1 中的能力参数设置生成各能力参数值，代入公式（3a）得到被试能力真值 θ_n。（2）生成评分者参数；将 3.3.1 中设定的评分者参数值代入式（3c）生成评分者的宽严度 D_m。（3）将生成的能力、项目及评分者参数真值代入式（1a）中，计算各项目的累积概率和相应的类别概率。（4）将步骤（3）中得到的累积概率和从均匀分布 U（0，1）生成的随机数 Ran 进行比较：若该随机数小于第一个等级的累积概率 $p(x<1)$，令评分等级为 1；若随机数大于等于第一个累积概率而小于第二个累积概率，即 $p(x<1) \le Ran < p(x<2)$，则令评分等级为 2，以此类推，最终得到所有评分者对所有被试在所有项目上的作答数据，重复 30 次以减少随机误差。

使用 OpenBUGS 软件进行参数估计，为使模型结果可收敛以及使蒙特卡罗误差（Monte Carlo error）小于参数标准差的 5%（Ntzoufras, 2011），模型将舍弃前 30000 次迭代，之后再迭代 60000 次，最终估计值为 30 次估计的均值。

3.3.3　评价指标

使用偏差（bias）、绝对百分比偏差（percentage bias, PB）和误差均方根（RMSE）3 个统计量作为评价指标（Flora & Curran, 2004; Wang, Su, & Qiu, 2014）。各指标的计算方法为：

$$Bias(\hat{\varsigma}) = \sum_{r=1}^{R}(\hat{\varsigma}_r - \varsigma)\Big/R$$

$$PB(\hat{\varsigma}) = 100 \times \sum_{r=1}^{R}\left|\frac{\hat{\varsigma}_r - \varsigma}{\varsigma}\right|\Big/R,\ \varsigma \neq 0$$

$$RMSE(\hat{\varsigma}) = \sqrt{\sum_{r=1}^{R}(\hat{\varsigma}_r - \varsigma)^2\Big/R}$$

表 2　参数估计值的描述统计结果

参数	真值	估计值	*SD*	参数	真值	估计值	*SD*
a_2	1.5	1.499	.118	B_{43}	1	1.015	.078
a_3	2	2.033	.107	γ_{00}	0	.008	.052
a_4	2.5	2.501	.079	γ_{10}	.136	.124	.047
B_{11}	-2	-2.014	.107	γ_{20}	-.023	-.064	.04
B_{12}	0	.025	.09	γ_{30}	.245	.208	.05
B_{13}	2	1.989	.104	γ_{01}	.14	.179	.081
B_{21}	-1	-1.022	.096	γ_{11}	-.09	-.085	.041
B_{22}	0	.027	.057	γ_{12}	.3	.287	.042
B_{23}	1	.995	.083	γ_{13}	-.53	-.498	.052
B_{31}	-3	-2.99	.134	γ_{14}	.04	.055	.038
B_{32}	1	.994	.078	σ_θ	1	.982	.117
B_{33}	2	1.98	.131	σ_1	1	.952	.072
B_{41}	-1.5	-1.525	.115	σ_2	1	.955	.083
B_{42}	.5	.517	.071	σ_3	1	.98	.085

式中，R 表示重复次数，实验中 R=30，$\hat{\varsigma}$为估计值，ς为参数真值。Flora 和 Curran（2004）指出，PB 小于 5 为偏差无关紧要，5~10 之间为中等偏差，10 以上为显著偏差。

3.4 研究结果

3.4.1 参数真值和估计值

表 2 为项目区分度（a_2 - a_4）、项目难度参数（B_{11} - B_{43}）、被试和评分者固定效应（γ_{00} - γ_{30}）、被试及评分者预测变量回归系数（γ_{01} - γ_{14}）、随机效应参数（σ_ϑ - σ_3）的真值及在 30 次实验下的估计值均值和标准差。从表 1 可以看出，所有参数的估计值均值都比较接近真值，且标准差较小，表明 GR-MLFM 全模型具有较好返真性和稳定性，适宜用于主观题评分的评分者效应及其影响因素的评定。

3.4.2 参数评价指标

模型各参数的三个评价指标结果见表 3。由表可知，在所有 28 个参数中，除了评分者 1 的随机效应（σ_1）以及评分者 2 的固定和随机效应（γ_{20} 和 σ_2）三个参数的 Bias 绝对值大于 .04 外，其它 25 个参数估计值的 Bias 绝对值均在 .04 以内，且所有参数的 Bias 绝对值均在 .05 以下。另外，RMSE 的范围在 .04 到 .132 之间，误差相对较小，其平均误差为 .083，标准差为 .027，可见模型各参数的 RMSE 相对比较小且较集中，模型的拟合情况较好。需要注意的是，在计算 PB 指标时，分母（即真值）对 PB 的影响比较大，PB 值可能并不适合评价分母趋近于 0 的参数。对于本研究中的 PB 值，25 个存在 PB 值的参数中，有 7 个参数的 PB 值大于 5（粗体标记部分），而纵观这 7 个参数，除了评分经验的回归系数值（γ_{13}）离 0 相对较远外，其它 6 个参数的真值均在 0 附近，可见此 6 者的 PB 值比较大的原因是其分母接近于 0，没有参考意义。而其余 18 个参数的 PB 值均在 5 以下，根据 PB 的界定标准可知（Flora & Curran, 2004），这些参数估计值与其真值间的差异可以忽略不计。

总而言之，即使在样本容量较小（100 人）的情况下，模拟研究的结果表明 GR-MLFM 的完整模型可获得相对准确的参数估计值，参数估计值的稳定性也相对比较好，且模型在各评价指标上的表现也比较理想,说明模型具有较好的适用性和稳定性。

表 3 完整模型下各参数的拟合结果

参数	*Bias*	*RMSE*	*PB*	参数	*Bias*	*RMSE*	*PB*
a_2	-.001	.116	.086	B_{43}	.015	.079	1.47
a_3	.033	.111	1.673	γ_{00}	.008	.051	—
a_4	.001	.077	.029	γ_{10}	.039	.089	**28.14**
B_{11}	-.014	.106	.7	γ_{20}	-.012	.048	**8.463**
B_{12}	.025	.092	—	γ_{30}	-.041	.057	**178.536**
B_{13}	-.011	.103	.562	γ_{01}	-.037	.061	**15.065**
B_{21}	-.022	.097	2.216	γ_{11}	.005	.041	**5.878**
B_{22}	.027	.063	—	γ_{12}	-.013	.044	4.494
B_{23}	-.005	.082	.531	γ_{13}	.032	.06	**6.099**
B_{31}	.01	.132	.336	γ_{14}	.015	.04	**38.083**
B_{32}	-.006	.077	.634	σ_ϑ	-.018	.116	1.838
B_{33}	-.02	.131	1.021	σ_1	-.048	.085	4.811
B_{41}	-.025	.116	1.667	σ_2	-.045	.093	4.477
B_{42}	.017	.072	3.397	σ_3	-.02	.086	1.963

注：γ_{00}、B_{12} 和 B_{22} 的真值为 0，故不存在 PB 值，最终有 25 个参数存在 PB 值。

4 实证研究：GR-MLFM 全模型在实践中的应用

4.1 研究模型

为考察 GR-MLFM 在复杂的实际评分情境中的应用效果。将 GR-MLFM 的完整模型应用于小学四年级数学问题解决的评分情境。模型包括一个学生层面预测变量和四个评分者层面预测变量。其水平 1 公式为式（1a），水平 2 公式为：

$$\theta_n = \gamma_{00} + \gamma_{01} X_{gend,n} + \mu_{0n}$$

$$B_{ij} = \gamma_{ij}$$

$$D_{rn} = \gamma_{r0} + \gamma_{res} W_{r1} + \gamma_{emo} W_{r2} + \gamma_{conf} W_{r3} + \gamma_{exp} W_{r4} + \mu_{rn}$$

$$\mu_{0n} \sim N\left(0, \sigma_\vartheta^2\right)$$

$$\mu_{rn} \sim N\left(0, \sigma_r^2\right) \quad (4)$$

其中，W_1 —W_4 分别为责任心、情绪稳定性、自信心、以及评分经验。

4.2 研究方法

4.2.1 研究材料

测试项目：邀请有关学科领域的命题专家根据本研究对题目的要求编制 4 道试题，再用这 4 道题对某市 260 名小学四年级学生进行预测试以检验试题的合理性。试卷收集回收后请一位专家对所有学生的作答情况进行 1~4 级评分。然后使用 Amos 软件对数据进行验证性因素分析(CFA)所得结果表明：模型与数据间的匹配度很好（ χ^2 = .492，df = 2，p > .05），且模型的各项拟合指标均较好（GFI=.999，NFI=.993），说明这 4 个主观题可以很好地体现学生的数学问题解决能力。

评分者层面预测变量：采用宋慧鹏（2014）修订的《评分者影响因素》调查问卷（李斌，2010）获取评分者层面的预测变量。问卷由 20 个项目组成，各项目的评分等级为 1~4，表示完全不同意到完全同意。问卷有四个维度，为式（4）中的 W_{r1}—W_{r4}，其中情绪稳定性分数越高，则说明评分者的情绪越不稳定。四个维度的信度分别为 .776、.884、.742、.726。

4.2.2 被试及评分者

被试：考虑到评分设计为交叉设计，只施测了 90 名学生，剔除无效试卷 10 份，最终得到 80 名学生的作答数据，女生 32 人，男生 48 人。

评分者：招募 20 名数学专业的研究生作为此次测验的评分员。

4.2.3 评分过程

先请学科专家制定参考答案，再请 20 位招募的评分员对所有学生的作答进行评分。正式评分前先进行系统培训以及预评，之后才进行正式评定。为保证局部独立性假设能够成立，评分时每一位评分者均独自对项目进行评分，避免评分者之间相互商量，以影响最终的评分结果。采用 Epidata 软件录入数据，然后使用 OpenBUGS 软件进行参数估计。

4.3 研究结果

4.3.1 被试参数

学生的固定和随机效应参数如表 4 所示。学生的整体平均能力为 -.058，t 值和 95% 置信区间（95%CI）均表明学生能力无性别差异。此外，能力随机效应（σ_ϑ）的显著性检验结果表明，能力随机变异不显著，说明各被试间的能力差异相对较小，无需加入其它因素建立三水平模型进行后续分析。

表 4 学生参数的固定效应和随机效应

效应	估计值	se	t 值	χ^2	95% CI
固定效应					
γ_{00}	-.058	.880	-.066	—	[-1.534, 1.875]
γ_{01}	-.331	.357	-.927	—	[-1.035, .395]
随机效应					
σ_ϑ	1.690	.237	—	74.83	—

4.3.2 项目参数

四个项目的难度和区分度参数见表 5。罗照盛（2012）指出，IRT 框架下的项目区分度应该在 .3 以上，由此，项目 2 的区分度较差，项目 1 和项目 3 的区分度尚可，项目 4 的区分度最好。四个项目的第一个难度阈值（B_{i1}）均很低，学生在这些项目上比较容易得 2 分，其中项目 1 和 4 的第二个难度阈值也比较小，学生在这两个项目上比较容易得 3 分，而学生在项目 2 和 3 上获得 3 分和 4 分的难度相对较大。

4.3.3 评分者参数

与模拟研究类似，20 位评分员的平均宽严程度（系统的评分者偏差）和预测变量的回归系数，越接近 1 表示越严格，越接近 0 表示越不存在评分者效应，越接近 -1 表明越存在宽松效应。所有评分员的平均宽严度在 -.809 到 .808 之间。评分员 9（γ_{90}）的平均宽严度达到显著水平（γ_{90} = -.809，t = -2.18*），说明该评分员表现出明显的宽松效应，评分员 17 则较为严格（γ_{170} = .808），而评分员 6 和 15 则较为客观和稳定（γ_{60} = .067，γ_{150} = -.033）。图 1 中的固定效应趋势线进一步刻画了 20 位评分员从最宽松到最严格的表现情况。

表 5 四个项目的参数估计结果

	a_1	B_{i1}	B_{i2}	B_{i3}
项目 1	.408 (.054)	-5.518 (.871)	-1.007 (.598)	.834 (.618)
项目 2	.245 (.035)	-3.846 (.785)	.984 (.629)	4.624 (.906)
项目 3	.646 (.075)	-4.198 (.742)	3.473 (.708)	7.451 (1.035)
项目 4	3.334 (.451)	-2.587 (.645)	-1.817 (.614)	1.094 (.607)

注：括号内为参数的标准误 (*se*)

对于回归系数，责任心（*res*）和自信心（*conf*）两个预测变量对评分员的评分存在显著的预测作用（γ_{res} = 0.276，*t* = 2.19*；γ_{conf} = -.176，*t* = -4.889*）。其中，责任心（*res*）对评分严厉度有正向的预测作用，评分员责任心越强，则他们的评分越严格；而自信心（*conf*）则对评分宽松度起正向的预测作用，评分员的自信心越强，则评分标准变得宽松，越倾向于给高分。情绪稳定性（*emo*）和评分经验(*exp*)对评分员宽严度并无显著影响(γ_{emo}= 0.009；γ_{exp}= -0.015）。

各评分员的随机效应均相对较小，分布范围在 .088 到 .473 之间，均值为 .130，所有评分员的随机效应均未达到显著性水平，说明这些评分员的内部一致性程度相对较高。其中，评分员 7 的随机效应最大（.473），且其 *se* 也相对较大（*se* = .337）；而其他评分员的 *se* 在 .057 ~.150 之间，说明这些评分员的评分准确性较评分员 7 更好。另外，由图 1 可知，所有评分员在评定被试时，要么始终保持宽松倾向，要么始终保持严厉倾向。

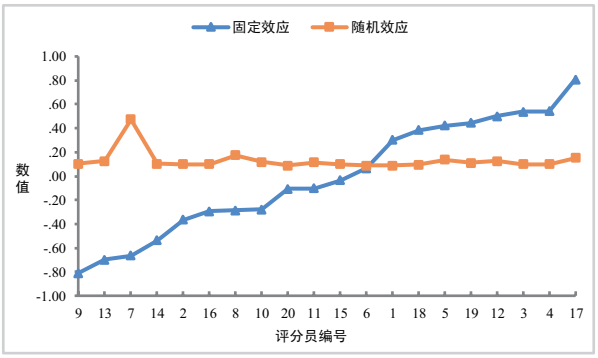


图 1 评分员固定和随机效应分布图

5 讨论

5.1 模型具有较高的可行性

模型是否可行可由两方面来验证：逻辑上的可行性和模型对数据的拟合情况，只有当这两个方面都符合时，模型才是可行的（Wang & Liu, 2007; Wang, Su, & Qiu, 2014）。康春花等（2016）构建了 GR-MLFM，证明了其逻辑上的可行性，并通过两个模拟研究探讨了在不包含任何预测变量（无条件模型）下的固定效应和随机效应模型的拟合情况，发现随机效应的结果整体要好于固定效应，说明进一步发展为完整模型是必要的。为此，本研究在随机效应模型的基础上加入评分者效应的影响因素，使其发展为可同时检测评分者效应、影响因素、影响大小的完整模型。模拟研究结果表明：即使在样本容量极小的情况下，在所有的项目参数、能力参数、预测变量的回归系数和评分者固定效应及随机效应参数中（共 28 个），只有极少数参数存在稍微的偏差外，其它参数的拟合效果均很好。说明在完整模型的情况下，GR-MLFM 也能较好地实现模型的参数估计，GR-MLFM 的完整模型是可行的。在大尺度测评的实践中，样本容量远远超过 100，而 IRT 模型和 MCMC 算法往往也需要较大容量的样本，因此可以预测 GR-MLFM 在大尺度测评实践中会有更好的精确性和稳定性。

5.2 GR-MLFM 在实践情境中的适宜性

实证研究表明，所有 20 位评分员中，虽然绝大多数评分员不存在显著的系统偏差(评分员 9 除外)，但依然存在两种评分偏差，一种是评分相对较为宽松，另一种是评分相对严厉。根据评分员随机效应的大小可知，尽管有些评分员倾向宽松，有些评分员倾向严厉，但这种宽严度并不会因被试的不同而产生较大的波动性，即其宽严度在被试个体间保持一定的稳定性，根据前人的观点（Wang & Liu, 2007; Wang & Wilson, 2005），说明这些评分员具有较高的评分者内信度。

评分者的责任心和自信心会显著地影响评分结果，而情绪稳定性和评分经验则对评分结果的预测作用不显著。责任心可以正向预测评分者的严格效应，自信心则可以正向预测评分者的宽松效应。这一结果与 Bernardin 等（2000）的研究结果相同：责任心越强，则越不会出现宽大效应。同样，评分经验与评分者效应间的关系与 Leckie 和 Baird（2011）的结果相同：不管是经验丰富者，还是新手评分员，均无法预测其评分者效应。但是，本研究得到的关于情绪稳定性的研究结果与前人的研究结果（Knez, 1995; Knez & Enmarker, 1998）存在差异，前人的研究表明在积极情绪下，评分员容易产生宽松效应，而在消极情绪条件下，则评分相对严格。这可能与研究者对情绪的操作分类存在差异。在 Knez 和 Enmarker（1998）的研究中，他们将情绪分为积极和消极情绪，然后探讨在这两种情绪状态下的评分者宽严程度；而本研究是依据稳定程度对情绪进行分类，将情绪分为稳定和不稳定两种情况。

5.3 研究展望

未来研究可以考虑：第一，评分情境方面的拓展。本研究采用的完全交叉设计，该设计的不足

之处在于评分者的评分任务相对较大，未来可探讨GR-MLFM在嵌套或混合设计中的适用情况；第二，模型的拓展。本研究中的模型是在单维性假设的基础上建立的，即假设项目只测量了学生的一种能力特质，未来可将单维的GR-MLFM拓展成多维GR-MLFM；第三，多种评分者效应的检测研究。未来可将GR-MLFM用于探讨其它的评分者效应如极端化、中心化等，以进一步检验模型的适宜性。

6 结论

通过模拟研究和实证研究对GR-MLFM完整模型的适宜性进行了探讨，得到以下结论：

（1）GR-MLFM完整模型具有逻辑上和数理上的合理性，可用于主观题的评分情境，较好地估计项目、被试能力、评分者及影响因素各参数，较好地检测出评分者效应、影响因素及其影响程度；

（2）在数学问题解决的评分实践中，评分员存在两种类型的评分倾向（宽松效应和严格效应），但绝大多数评分员的宽严度不明显，只有极个别评分员具有明显的宽松效应；评分者的责任心可正向预测其严格程度，自信心可正向预测其宽松程度。

参考文献

康春花，孙小坚，曾平飞. (2016). 基于等级反应模型的多水平多侧面评分者模型. *心理科学，39(1)*，214–223.

康春花，辛涛. (2010). 基于IRT的评分者效应模型及其应用展望. *中国考试，8*，3–8.

李斌. (2010). 评分者一致性自身特征结构影响因素的验证性研究. *山东师范大学学报（人文社会科学版），55(5)*，101–105.

李斌. (2011). *评分者信度及其影响因素结构研究*. 北京师范大学博士学位论文.

刘红云，骆方. (2008). 多水平项目反应理论模型在测验发展中的应用. *心理学报，40(1)*，92–100.

罗照盛. (2012). *项目反应理论基础*. 北京：北京师范大学出版社.

宋慧鹏. (2014). *嵌套设计下的主观题评分者效应研究——基于多水平随机系数模型*. 浙江师范大学硕士学位论文.

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.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19(1)*, 101–119.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement, 74(5)*, 795–808.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology, 85(2)*, 232–236.

Breuer, K., Nieken, P., & Sliwka, D. (2013). Social ties and subjective performance evaluations: An empirical investigation. *Review of Managerial Science, 7(2)*, 141–157.

Cowles, M. K. (2013). *Applied bayesian statistics: With R and OpenBUGS examples*. New York, USA: Springer.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9(4)*, 466–491.

Fox, J. P. (2010). Multilevel item response theory models. In J. P. Fox (Ed.). *Bayesian item response modeling* (pp. 141–191). New York, USA: Springer.

Hamp–Lyons, L. (2011). Writing assessment: Shifting issues, new tools, enduring questions. *Assessing Writing, 16(1)*, 3–5.

He, T. H., Gou, W. J., Chien, Y. C., Chen, I. S. J., & Chang, S. M. (2013). Multi–faceted rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports, 112(2)*, 469–485.

Hung, L. F., & Wang, W. C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics, 37(2)*, 231–255.

Jawahar, I. M., & Stone, T. H. (1997). Appraisal purpose versus perceived consequences: The effects of appraisal purpose, perceived consequences, and rater self–monitoring on leniency of ratings and decisions. *Research and Practice in Human Resource Management, 5(1)*, 33–54.

Knez, I. (1995). Effects of indoor lighting on mood and cognition. *Journal of Environmental Psychology, 15(1)*, 39–51.

Knez, I., & Enmarker, I. (1998). Effects of office lighting on mood and cognitive performance and a gender effect in work–xrelated judgment. *Environment and Behavior, 30(4)*, 553–567.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48(4)*, 399–418.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. New Jersey, USA: John Wiley & Sons.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43(1)*, 39–55.

Wang, W. C., & Liu, C. Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67(4)*, 583–605.

Wang, W. C., & Qiu, X. L. (2013). A multidimensional and multilevel extension of a random–effect approach to subjective judgment in rating scales. *Multivariate Behavioral Research, 48(3)*, 398–427.

Wang, W. C., Su, C. M., & Qiu, X. L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement, 51(3)*, 260–280.

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random–effects facet model. *Applied Psychological Measurement, 29(4)*, 296–318.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46(1)*, 35–51.

Wolfe, E. W., Jurgens, M., Sanders, B., Vickers, D., & Yue, J. (2014). *Evaluation of pseudo–scoring as an extension of rater training*. Research Report,. Retrieved from http://researchnetwork.pearson.com/wp–content/uploads/Wolfe PseudoScoring 04142014.pdf

Formulating and Applying Grade Response Multilevel Facets Model in Scoring Open-Ended Item

Kang Chunhua, Sun Xiaojian, Zeng Pingfei
（College of Teacher Education, Zhejiang Normal University, Jinhua, 321004）

Abstract Open-ended items played an important role in evaluating students' skills, such as analysis, synthesis, and problem solving skills. During scoring open-ended items, the rater effects would occur due to lack of standard answers, as well as lack of consistent cognition for the rating rules between different raters. As a consequence, the scoring results would be affected by the rater effects. How to estimate the person, item and rater parameters precisely is an important issue. Some researchers formulated the GRM-based multifaceted model, which was called grade response multifaceted model (GR-MFM), to estimate one's ability and handle the rater effect when the tasks were successively processing. They used two simulation studies to examine parameter recovery for the unconditional model (no predictors were added to the model) of GR-MFM. Results showed that the model could recover all the parameters very well, and GR-MFM was useful and reasonable; also, results showed that the random effect model were more suitable than the fixed model. The purpose of current study was to examine the reasonableness of GR-MFM when both person and rater predictors were added to the model, which was called full model of GR-MFM. One simulation study and an empirical study were conducted to evaluate the feasibility of GR-MFM.

For the simulation study, 2 levels were formulated. Level 1 was an IRT model, and in level 2, the gender of students and 4 predictors of raters were considered. The R software was applied to generate one's response matrix. After that, the OpenBUGS software, based on MCMC algorithm, was used to estimate the parameters of the model. Bias, root mean square error (RMSE), and percentage bias (PB) were used to evaluate the recovery. The results indicated that all of the estimates of parameters were closed to the true values, and the absolute differences between the estimates and true values were less than .05 across all the parameters. Meanwhile, the RMSEs of these estimates were small enough, and the values ranged from .04 to 0.132. Furthermore, although the 7 PB values were larger than 5, most of them could be attributed to smaller denominators, which meant almost all of these parameters showed an acceptable result. Based upon these results, it could be considered that the model could fit the data precisely and stably, and was promising to apply the model to detect the rater effect.

For the empirical study, 4 open-ended items were used to detect students' problem solving skills in mathematics, then 20 raters were recruited to rate the responses of 80 persons who answered these items. Also, the gender of students and 4 predictors of raters, responsibility and stability of emotion and confidence and rating experience, were added to the level 2 model to investigate the rater effects. Results showed that among these 20 raters, almost all raters showed non substantial rater effects (severity/leniency), and only rater 9 displayed significant severity. Furthermore, 2 predictors of raters had significant effects on the rater effects, among these, responsibility had positive effects on severity, and confidence had positive effects on leniency; while the rating experience and stability of emotion of raters produced non-significant effects on rating results.

Key words grade response multilevel facets model, full model, rater characteristics, rater effect