心理学探新 2023,Vol.43,No.1,84-89
PSYCHOLOGICAL EXPLORATION

双因子项目反应模型在研究生招生考试质量分析中的应用*

宋学玲[1],梁正妍[2]
(1.教育部教育考试院,北京 100084;2.华南师范大学,广州 510631)

摘 要:研究生招生考试中学科专业能力的考查,主要采用的是“大综合”形式的试卷,即将学科专业基础课程的知识集中在一张试卷上进行考查。“大综合”的试卷形式大大提高了试卷的命制难度,也对试题质量提出了极高的要求。针对2022年全国硕士研究生招生考试心理学专业基础科目的抽样作答数据,采用双因子项目反应模型对试题质量进行分析。研究显示,整套试卷命制基本符合“大综合”试卷的命制要求,其中心理学一般因子作为主要的考查维度,具有良好的区分度;而特殊因子(课程因子)的表现存在差异。从能力密度曲线来看,实验心理学、心理统计与测量两个因子的选拔性功能更强。

关键词:双因子项目反应模型;研究生招生考试;“大综合”试卷;质量分析
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2023)01-0084-06

1 引言

全国硕士研究生招生考试(简称“研究生招生考试”)的试卷质量事关高层次人才的选拔,其重要性不言而喻。与高考、公务员考试等不同,研究生招生考试中对于考生学科专业能力的考查,主要通过“大综合”形式的试卷进行,即多个学科专业基础课程的知识点集中在一张试卷上。考试时长限制了试卷的题量,而“大综合”试卷又需要涵盖多个专业基础课程的知识点,大大提高了试卷的命制难度。从题型的设计,到各基础课程知识所占试卷题量、总分的比例,以及如何在有限的题量限制下尽可能地区分出考生的能力,这些对于命题人员来说都是极大的挑战。

试题试卷的评价需要结合考试本身的目的、考试的具体形式而定。目前试题试卷的评价多采用经典测量理论(CTT)和项目反应理论(IRT;Baker & Kim,2004)。CTT的数学模型简单易懂、可操作性强、应用广泛,但是也存在着不少局限,比如测量结果拓广有限、测量分数依赖试题、统计量依赖样本、信度估计不精确、能力量表与难度量表不一致等(漆书青 等,1998,2002)。为了弥补CTT存在的缺陷,IRT应运而生。IRT主要考查被试的作答反应与被试能力之间的关系,通过项目特征曲线,将项目难度、项目区分度、被试能力值标记在同一个坐标系下,建立了被试能力与难度之间的直接联系。国内对于研究生招生考试的试题质量研究相对较少:冼利青等(1996)从经典测量理论的角度对医学硕士研究生入学考试的试题质量进行了分析;关丹丹等(2011)应用多元概化理论对全国硕士研究生入学考试心理学科目的试题质量进行了研究;赵守盈等(2012)采用Rasch模型对全国硕士研究生入学考试心理学科目的试题质量进行了分析;戴一飞等(2018)对法硕(非法学)专业学位联考的预测效度进行了分析。

总体而言,过往对于研究生招生考试试题质量的研究主要采用的是经典测量理论、项目反应理论以及概化理论三大理论,而其中基于项目反应理论的研究,主要采用的是单维Rasch模型。但是,采用Rasch模型的相关研究只对选择题部分做出了分析,同时也缺乏对“大综合”试卷中各基础课程试题间的比较分析,对“大综合”科目试卷质量的分析还不够全面。因此,探究双因子项目反应模型在“大综合”科目试卷质量分析中的应用路径,并采用该模型对研究生招生考试专业基础科目的试卷进行质

* **基金项目**:国家教育考试科研规划重点课题(GJK2021020),国家教育考试科研规划一般课题(GJK2021049)。
通讯作者:梁正妍,E-mail:2020010220@m.scnu.edu.cn。

量分析是非常必要的。

2　双因子模型与项目反应理论

2.1　双因子模型

双因子模型，又称一般－特殊因子模型（General－Specific Factor Model），其思想来源于能力结构的二因素理论（彭聃龄，2018）。双因子模型基于以下两点假设：（1）一般因子 G 的存在性，即存在一个可以解释所有项目共同变异的一般因子；（2）特殊因子 S_i 的存在性，即存在多个可以额外解释部分项目共同变易的特殊因子（Holzinger & Swineford，1937）。双因子模型的数学表达式如下所示：

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} G + \begin{bmatrix} b_{11}\cdots b_{1m} \\ b_{21}\cdots b_{2m} \\ \cdots\cdots \\ b_{n1}\cdots b_{nm} \end{bmatrix} \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_m \end{pmatrix} + \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} \quad (1)$$

其中，$\{x_1, x_2, \cdots, x_n\}$ 是一个测验的全部项目，G 为一般因子，$\{S_1, S_2, \cdots, S_m\}$ 是 m 个特殊因子，a_i 为项目 x_i 在 G 上的载荷，b_{ij} 是项目 x_i 在 S_j 上的载荷，δ_i 是项目 x_i 的测验误差。

双因子模型中，一般因子 G 与特殊因子 $S_1, S_2, \cdots, S_m$ 统称为公共因子（common factor），二者处于同一测量层次上，区别在于前者反映了所有项目的公共属性，而后者仅反映了部分项目的公共属性，因此每个变量仅在一般因子和一个特殊因子上的载荷非零，从而其因子载荷矩阵为分块矩阵。根据分析结果，双因子模型可以用来评估一般因子及特殊因子在整个测量中的重要性（顾红磊 等，2014）。

一般而言，测验的测量结构可以分为以下五种类型：单维模型、多个单维模型、相关特质多维模型、二阶因子模型、双因子模型。当各维度之间不相关或相关较弱时（相关系数在0.1以下），建议采用多个单维模型；当各维度之间存在中低等相关时（相关系数介于0.1到0.4），建议使用相关特质多维模型；当各维度之间存在中高等相关时（相关系数在0.4以上），建议采用双因子模型（顾红磊 等，2014；毛秀珍等，2018；Reise et al.，2007；Reise et al.，2010）。

2.2　项目反应理论

项目反应理论（IRT），又称潜在特质理论，是当前应用最为广泛的现代心理测量理论之一。IRT是在一定的假设下，用数学函数去刻画被试在项目上可观察的作答表现（得分）与其不可观察的特质水平（能力）之间的函数关系，即IRT模型。用概率密度函数来刻画被试的能力与其在项目上的正确反应情况之间的函数关系是自然的，相应的函数曲线称为项目特征曲线（闫成海 等，2014）。

IRT的理论假设主要包含以下三条：（1）单维性假设，即测验只测量被试的某一种能力（潜在特质），其他能力对测验结果的影响可以忽略不计。（2）局部独立性假设，即被试在各个项目上的作答反应相互独立。（3）项目特征曲线假设，即被试在项目上的正确作答概率遵循一定的函数关系。后来，多维项目反应理论打破了单维性假设，题组反应理论打破了局部独立性假设，所以第三条假设是IRT的核心假设。

依据评分规则的不同，IRT模型可以分为二级计分模型和多级计分模型。针对非对即错的选择题，选用二级计分模型进行试题质量分析；针对简答题、综合题等，一般采用多级计分模型进行试题质量分析。

二级计分模型中常用的有Rasch模型、Logistic模型等。Logistic模型可分为单参数、双参数、三参数Logistic模型，其对应的项目特征函数分别是：

$$p_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}; \quad (2)$$

$$p_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}; \quad (3)$$

$$p_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}. \quad (4)$$

其中，$p_i(\theta)$ 是能力水平为 θ 的被试在项目i上的正确作答概率；a_i, b_i, c_i 分别是项目i的区分度参数（又叫斜率参数）、难度参数、猜测度参数（又叫下渐近线参数）；$D = 1.7$（或1.701）是一个常量。

多级计分模型中常用的有称名反应模型、评定量表模型、等级反应模型等（漆书青 等，1998，2002）。等级反应模型适用于按步骤作答的题目，且步骤间的难度是递增的。能力为 θ_i 的被试在项目 j 上的得分不低于 k 分的概率为：$P^*_{ijk} = \frac{e^{a_j(\theta - b_{jk})}}{1 + e^{a_j(\theta - b_{jk})}}$，其中 a_j 是项目 j 的区分度参数，b_{jk} 是在项目 j 上得 k 分的等级难度参数：$b_{j1} < b_{j2} < \cdots < b_{jk}$。于是，能力为 θ_i 的被试在项目 j 上得 k 分的概率为：$p_{ijk} = P^*_{ijk} - P^*_{ij,k+1}$。

多维项目反应理论（MIRT）建立在单维项目反应理论和因子分析的基础之上，克服了单维项目反应理论的单维性缺陷，可在多个维度上分析被试的

作答表现。下面所述的双因子项目反应模型就是多维项目反应模型在双因子模型假设下的特殊形式(毛秀珍 等,2018)。

2.3 双因子项目反应模型

1992 年,Gibbons 和 Hedeker 将双因子模型引入项目反应理论。之后,Cai,Yang 和 Hansen 等(2011)详细描述了双因子 Logistic 模型、双因子多级计分模型及其参数估计方法。以三参数 Logistic 模型为例,其对应的双因子 Logistic 模型的概率密度函数为

$$p(u_{ij}=1 \mid \theta_{0i},\theta_{sj}) = c_j + \frac{1-c_j}{1+e^{-D(d_j+a_{0j}\theta_{0i}+a_{sj}\theta_{si})}} \quad (5)$$

其中,$p(u_{ij}=1 \mid \theta_{0i},\theta_{sj})$ 表示被试 i 在项目 j 上的正确作答概率;$\theta_i=(\theta_{0i},\theta_{si})$ 是被试 i 的能力向量参数;a_{0j},a_{sj} 分别是项目 j 在一般因子和特殊因子上的斜率参数,代表了项目 j 在相应维度上的区分度;c_j 是下渐近线参数,反映了项目 j 内容的模糊程度;$d_j=-(a_{0j}b_j+a_sb_j)$ 是项目 j 的截距参数,与项目的难度参数 b_j 负相关。多级计分的双因子项目反应模型的密度函数也可以由双因子 Logistic 模型的密度函数推导得到。

3 研究生招生考试“大综合”试卷质量分析

以 2022 年全国硕士研究生招生考试《心理学专业基础(312)》为例,采用双因子项目反应模型对试卷质量进行分析。在被试作答数据中,随机抽取 22953 份样本,剔除小题数据缺失的 827 份样本,实际研究可用作答样本为 22126 份。数据分析均采用 SPSS 21.0 以及 R 软件中的 mirt 包(沈励,万雅琦,2022)。

3.1 试卷结构

全国硕士研究生招生考试《心理学专业基础(312)》科目主要涉及心理学导论(简称“普心”)、发展与教育心理学(简称“发教”)、实验心理学(简称“实验”)、心理统计与测量(简称“统测”)四个学科基础课程的内容。试卷结构见表 1。

表 1 试卷结构

	单项选择题 每题 2 分	多项选择题 每题 3 分	简答题 每题 10 分	综合题 每题 30 分	总计
普心	1 - 19 题	66 - 69 题	76 - 77 题	81 题	26 题,100 分
发教	20 - 31 题	70 - 71 题	78 题	82 题	16 题,70 分
实验	32 - 48 题	72 - 73 题	79 - 80 题	- -	21 题,60 分
统测	49 - 65 题	74 - 75 题	- -	83 题	20 题,70 分
总计	65 题,130 分	10 题,30 分	5 题,50 分	3 题,90 分	83 题,300 分

各维度得分的相关系数如表 2 所示。可以看出,试卷所包含的四个维度的考核内容相关系数均在 0.8 左右,属于高相关,可以采用双因子项目反应模型来分析被试的作答反应。

表 2 各维度原始得分相关矩阵

	普心	发教	实验	统测
普心	1			
发教	0.83	1		
实验	0.81	0.80	1	
统测	0.76	0.77	0.80	1

3.2 模型拟合

针对样本数据,采用单维项目反应模型、多维项目反应模型以及双因子项目反应模型对数据进行了拟合检验,拟合结果如表 3 所示。

其中,模型拟合评价指标 AIC 是 Akaike 信息准则,BIC 是贝叶斯信息准则,SABIC 是样本校正的 BIC,HQ 为 Hannan - Quinn 准则,这四个指数的值越小,表示模型对数据的拟合越好;对数似然函数 logLik 的绝对值越小,模型对数据的拟合也越好(潜变量建模与 Mplus 应用 · 进阶篇,王孟成,毕向阳,2018)。

表 3 三种模型的拟合指标比较

	AIC	SABIC	HQ	BIC	logLik
单维	2153245	2154273	2153800	2154950	-1076410
多维	2242154	2243182	2242709	2243859	-1120864
双因子	2147630	2149059	2148402	2149999	-1073519

从模型拟合结果来看,多维项目反应模型的拟合结果是最差的,其次是单维项目反应模型,拟合表现最好的是双因子项目反应模型。采用 R 软件 mirt 包中的 anova 函数对单维项目反应模型和双因子项目反应模型进行比较后发现,双因子项目反应模型的拟合显著优于单维项目反应模型,详见表 4。

表4　单维与双因子项目反应模型比较

	AIC	SABIC	HQ	BIC	logLik	X^2	*df*	*p*
单维	2153245	2154273	2153800	2154950	-1076410	NA	NA	NA
双因子	2147630	2149059	2148402	2149999	-1073519	5781.047	83	0

综上，选用双因子项目反应模型来分析作答数据是合适的。

3.3　项目参数估计

本套试卷共有83道试题：选择题75道，单项选择题每题2分、多项选择题每题3分；简答题5道，每题10分；综合题3道，每题30分。二级计分题（选择题）采用双因子双参数Logistic模型；多级计分题（简答题和综合题）采用双因子等级反应模型，其中简答题每2分合并为一个等级，共5个等级难度，（分数(0,2]合并为一个等级，此等级所估难度为难度1；分数(2,4]合并为一个等级，此等级所估难度为难度2；以此类推）；综合题每3分合并为一个等级，共10个等级难度（分数(0,3]合并为一个等级，此等级所估难度为难度1；分数(4,6]合并为一个等级，此等级所估难度为难度2；以此类推）。

二级计分题和多级计分题的部分项目参数估计结果如表5、表6所示。多级计分题与难度相关的参数（截距参数）设定为（$d_1,d_2,d_3,d_4,d_5,d_6,d_7,d_8,d_9,d_{10}$），具体测量的参数数量因等级数量不同而不同；每个项目都有2个区分度参数，即一般因子区分度和特殊因子区分度。表格中*MDISC*是多维项目的总区分度$MDISC=\sqrt{\sum a_i^2}$；*MDIFF*是多维项目的难度，其中$MDIFF_k=\frac{-d_k}{MDISC}$。

表5　部分二级计分题区分度及难度参数

题号	a_1	a_2	a_3	a_4	a_5	*MDISC*	*MDIFF*	题号	a_1	a_2	a_3	a_4	a_5	*MDISC*	*MDIFF*
1	0.29	0.21				0.36	3.73	42	1.59			0.22		1.60	0.47
5	1.30	-0.18				1.31	0.49	45	-0.66			-0.14		0.67	0.97
6	2.97	0.49				3.01	-0.81	46	2.34			0.43		2.38	-1.54
15	3.26	0.72				3.34	-0.96	55	2.29				0.90	2.46	-0.89
20	1.79		-0.17			1.80	-0.70	61	2.53				0.68	2.61	-0.95
21	1.73		-0.34			1.77	-1.33	66	0.93	0.40				1.01	-3.09
26	2.44		0.22			2.45	-0.72	67	-0.25	0.16				0.30	4.48
28	2.13		-0.22			2.14	-1.09	74	0.25				0.27	0.37	3.65
37	-0.56			0.21		0.60	-0.36	75	0.72				0.18	0.74	-1.80

表6　部分多级计分题区分度及截距参数

题号	a_1	a_2	a_3	a_4	a_5	*MDISC*	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
76	1.56	-0.46				1.63	-0.77	-1.30	-1.57	-1.56	-1.22					
78	0.92		0.68			1.14	-1.24	-1.99	-2.38	-1.57	0.88					
79	0.66			-0.21		0.70	-1.61	-3.46	-4.46	-4.44	-2.37					
83	0.63				0.99	1.17	-0.23	-0.41	-0.19	0.38	0.71	0.90	1.19	1.31	1.89	2.16

经转换计算，表6中所涉及试题的难度参数如下：第76题的难度参数$MDIFF_{76}=(0.47,0.80,0.96,0.96,0.75)$，第78题的难度参数$MDIFF_{78}=(1.09,1.75,2.09,1.38,-0.77)$，第79题的难度参数$MDIFF_{79}=(2.30,4.94,6.37,6.34,3.39)$，第83题的难度参数$MDIFF_{83}=(0.20,0.35,0.16,-0.32,-0.61,-0.77,-1.02,-1.12,-1.62,-1.85)$。

项目反应理论认为，项目的难度参数应在[-3,3]之间，项目的区分度参数应在[0,3]之间（罗照盛，2012）。难度参数的数值越高代表试题难度越大。从难度参数来看，整套试卷中绝大多数试题难度合理，难度参数在[-3,3]范围内，但极少数试题难度偏高，如第67题。结合区分度来看，第67题在主测维度“发教”上区分度过低，可能是由于其难度过高(4.48)导致的，即便在维度“发教”上能力高的被试在该题上正确作答的概率也很小，而其他被试却依然有一定概率通过猜测答对这道选择题。*MDIFF*值也可以用来分析多级计分题等级划分的合理性。比如第79题的难度$MDIFF_{79}=(2.30,4.94,6.37,6.34,3.39)$，前三个等级的设置有一定的递增梯度，比较合理，但是后面两个等级的难度相

关参数递减，等级设置不够合理，还需改进。

MDISC 是一个总的概念，可以通过每一个 a_i 值来细致分析每个项目在各维度上的区分度。数据显示，二级计分题在一般因子上具有较好的区分度（表中 a_1），但是具体到特殊因子上，不同试题的区分度表现存在差别。其中，在“发教”维度共有 4 道试题的特殊因子区分度（表中 a_3）为负数，说明这些试题测试该维度的能力时，能力高的被试反而正确作答率低，但是这几道题在一般因子上的区分度表现却很好。多级计分题在一般因子上的整体表现也优于特殊因子。其中，多级计分题在“普心”和“实验”两个维度上的区分度（表中 a_2、a_4）表现一般；在“发教”和“统测”两个维度上的区分度（表中 a_3、a_5）表现良好。

3.4 被试能力参数估计

采用双因子项目反应模型对被试能力参数进行估计，基于不同因子能力绘制密度曲线图如图 1 所示。

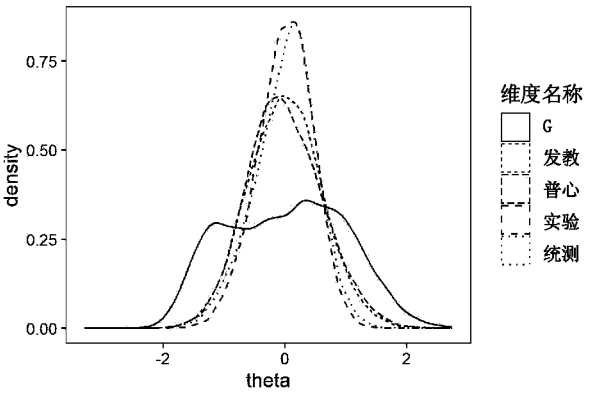

图 1 能力密度曲线

在双因子模型中，G 因子即一般因子，代表了心理学一般素养，它蕴含在考核的各部分知识内容中。被试在心理学一般素养的能力分布广，测验信度的大部分方差均由心理学一般能力所解释，从数据分析结果可以看出，测验项目一般因子的区分度（表中 a_1）比特殊因子的区分度（表中 $a_2 \sim a_5$）更好。从图 1 中四个特殊因子的能力分布图来看，“实验”和“统测”维度上的能力分布比“普心”和“发教”维度上的能力分布更高狭，且能力均值更高，说明“实验”和“统测”更能考查出被试的高阶思维能力。

总的来说，此次试卷的命制达到了“大综合”考试形式的目的，即对学科综合素质的考查。

4 结论与思考

4.1 结论

针对 2022 年全国硕士研究生招生考试《心理学专业基础（312）》科目，采用双因子项目反应模型对试卷进行了质量分析，在多个维度上分析了被试的作答表现，并绘制了被试在各维度的能力密度曲线图，充分解读和分析这些测评信息，可以为提高试题质量提供有针对性的启发。主要结论如下。

（1）整套试卷命制符合“大综合”科目试卷的命制要求，基本达到了考试大纲中所设定的考核要求，实现了考查学科综合素质的目的。

（2）从项目特征参数来看，心理学一般因子作为主要的考查内容，具有较好的区分度；而特殊因子（课程因子）的表现存在差异。二级计分题的特殊因子“发教”、多级计分题的特殊因子“普心”和“实验”，在其主测维度上的测量精度有待提高。

（3）从能力密度曲线来看，相较“普心”和“发教”两个因子，“实验”和“统测”两个因子对被试高阶思维能力的考核更加有效，选拔性功能更强。

4.2 思考

双因子项目反应模型符合研究生招生考试中“大综合”试卷的结构特征。用双因子项目反应模型来处理被试在项目上的原始反应数据，比起传统的线性双因子模型的间接处理，保留了更多的被试作答信息。相较单维项目反应理论而言，双因子项目反应模型对“大综合”试卷的分析更加精细，它对每个项目都做了的细致的分析，对试卷总体和涉及的基础课程维度也进行了分析，能够看到被试能力在各个维度上的具体表现，从而能够全方位了解试题的质量情况，其最突出的优点是能够更加精确区分出专业基础“大综合”试卷中的鉴别性维度，有利于后续对考核内容和试卷结构进行针对性调整。

基于以上分析，对研究生招生考试专业基础“大综合”试卷的命制提出以下建议：

（1）明确“门槛性”考核内容和“鉴别性”考核内容。根据被试在特殊因子上的反应，区分出考试的“门槛性”因子和“鉴别性”因子。对于专业必需的“门槛性”知识，不必过分追求项目的难度和区分度，应该更加强调考核知识点的重要性和覆盖度，但是对于“鉴别性”知识则要求尽量提高项目质量，以实现考试的选拔目的。

（2）适当调整“门槛性”项目与“鉴别性”项目的题量和分值。根据双因子项目反应模型的分析结果，适当调整各个维度考核内容的比重，在适度考核专业“门槛性”知识的基础上，尽量提高“鉴别性”项目的比重和质量，以提高人才选拔的有效性。

针对研究生招生考试《心理学专业基础（312）》科目，建议在后续修订考试大纲时，对试卷结构进行

如下调整:第一,在心理学导论、发展和教育心理学的维度上,以适度、必需为原则,认真斟酌项目的取舍,在此基础上尽量使项目的特征参数在合理区间范围内。第二,在实验心理学、心理测量与统计两个维度上,适当提高考核内容占比,提高命题质量,加强对被试高阶思维能力的考核。

双因子项目反应理论的引入,拓宽了研究生招生考试质量评价的路径,为研究生招生考试的内容改革提供了更加丰富的分析资料,在提高研究生招生考试的科学性方面具有较广的应用前景。

参考文献

戴一飞,邢博特.(2018).法律硕士(非法学)专业学位联考的预测效度分析.*中国考试*,(3),19-26.

顾红磊,温忠麟,方杰.(2014).双因子模型:多维构念测量的新视角.*心理科学*,*37*(4),973-979.

关丹丹,王博,车宏生.(2011).2007-2010年心理学专业基础综合考试的多元概化理论研究.*心理科学*,*34*(4),950-956.

罗照盛.(2012).*项目反应理论基础*.北京:北京师范大学出版社.

毛秀珍,夏梦连,辛涛.(2018).全信息项目双因子分析:模型、参数估计及其应用.*心理科学进展*,*26*(2),358-367.

彭聃龄.(2018).*普通心理学*(第5版).北京:北京师范大学出版社.

漆书青,戴海崎,丁树良.(1998).*现代教育与心理测量学原理*.南昌:江西教育出版社.

漆书青,戴海崎,丁树良.(2002).*现代教育与心理测量学原理*.北京:高等教育出版社.

沈励,万雅奇.(2022).高中学业水平等级性考试数据分析拓展研究.*中国考试*,(5),54-63.

王孟成,毕向阳.(2018).*潜变量建模与Mplus 应用·进阶篇*.重庆:重庆大学出版社.

洗利青,程铭光,盛翠华.(1996).硕士研究生入学考试的初试质量分析.*中国高等医学教育*,(5),40-41.

闫成海,杜文久,宋乃庆,张健.(2014).高考数学中考试评价的研究——基于CTT与IRT的实证比较.*华东师范大学学报*(教育科学版),*32*(3),10-18.

赵守盈,何妃霞,陈维,罗杰,关丹丹.(2012).Rasch模型在研究生入学考试质量分析中的应用.*教育研究*,*33*(6),61-65.

Baker,F. B.,& Kim,S.(Eds).(2004). *Item response theory parameter estimation techniques*(2nd ed.). New York:Marcel Dekker,Inc.

Cai,L.,Yang,S.,& Hansen,M.(2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3),221-248.

Holzinger,K. J.,& Swineford,S.(1937). The bi-factor method. *Psychometrika*,*7*,41-54.

Reise,S. P.,Moore,T. M.,& Haviland,M. G.(2010). Bifactor models and rotations:Exploring the extent to which multidimensional data yield univocal scores. *Journal of Personality Assessment*,*92*,544-559.

Reise,S. P.,Morizot,J.,& Hays,R. D.(2007). The role of the bifactor model in resolving dimensionality issues in health outcome measures. *Quality of Life Research*,*16*,19-31.

The Application of Bi-factor Item Response Model in the Quality Analysis of Postgraduate Entrance Examination

Song Xueling[1],Liang Zhengyan[2]

(1. National Education Examinations Authority,Beijing 100084;2. South China Normal University,Guangzhou 510631)

Abstract:The examination of professional ability in the postgraduate entrance examination mainly adopts the "big comprehensive" test paper,that is,the knowledge of basic professional courses is concentrated on one test paper. The form of "big comprehensive" test paper greatly improves the difficulty of making test paper,and also puts forward high requirements for the quality of items. In this study,the answering data of subjects in the psychology major of the postgraduate entrance examination in 2022 were taken as the research object,and the bi-factor item response model was used to analyze the quality of the items. The research showed that the test paper basically met the requirements of "big comprehensive" test paper,in which the psychological general ability factor,as the main examination content,had a good differentiation;However,there were differences in the performance of special factors(course factors). From the ability density curve,experimental psychology,statistics and measurement were more selective.

Key words:bi-factor item response model;postgraduate entrance examination;"big comprehensive" test paper;quality analysis