·教学评价·

2010至 2012 年全国硕士研究生入学西医综合考试多元概化理论研究

白娟

【摘要】 **目的**　该研究全面评估 2010 至 2012 年西医综合考试的总体信度、试卷结构、二级学科比例分配的合理性等指标，为今后修订大纲和改进测验质量提供参考依据。**方法**　抽取 2010至 2012 年 15 765 名考生考后数据，应用多元概化理论，分别从学科因子和题型因子角度进行概化研究和决策研究的统计分析。**结果**　①关于学科因子分析，基础学科的方差分量相对较高，全域总分的相对误差很小，概化系数均在 0.930 以上。②关于题型因子分析，总体上 X 型题和 A1 型题的方差分量相对较高，全域总分的相对误差很小，概化系数均在 0.930 以上。**结论**　①从考查的学科内容看，各学科测量精度均较高，基础学科的贡献相对高于临床学科，进一步提高试题区分度有利于进一步提高测量精度。②从题型设计看，各题型测量误差小，测量精度高，各题型的分量分布较合适，A3 型题的试题分量可进一步提高。

【关键词】 西医综合考试；　概化理论；　试卷结构；　测量精度

【中图分类号】 R-05

Study of 2010–2012 western medicine comprehensive test in postgraduate entrance examination based on multivariate generalizability theory

Bai Juan. National Education Examination Authority, Beijing 100084, China

【**Abstract**】 **Objective**　To systematically evaluate the overall reliability, test paper structure and rationality proportional distribution of the secondary discipline of western medicine comprehensive test papers from 2010 to 2012 to provide references for future improvement to the syllabus and test quality. **Methods**　Data of 15 765 candidates from 2010 to 2012 were extracted and statistical analysis was performed to do generalized research and decision-making research from discipline factors and question type factors. **Results**　①Based on the discipline factor analysis, variance components of basic disciplines were relatively high , global score relative error was very small and the generalizability coefficient was above 0.930. ②Based on the question type factor analysis, variance components in X-type questions and A1-type questions were relatively high, global score relative error was very small and the generalizability coefficient was above 0.930. **Conclusions**　①From the discipline content, the accuracy of various disciplines is high; contribution of basic disciplines is relatively higher than that of clinical disciplines; further improving the discrimination of questions can help to improve measure-ment accuracy. ②From the question design, question measurement error is small, accuracy is high, and the question distribution is appropriate; scale of A3 type questions can be further improved.

【**Key words**】 Western medicine comprehensive test;　Generalizability theory;　Construct of the test paper;　Measurement accuracy

DOI:10.3760/cma.j.issn.2095-1485.2014.01.001

基金项目:全国教育科学规划 2011 年度教育部重点课题、教育考试研究专项立项课题(GFA111012)

作者单位:100084　北京,教育部考试中心

Email: baijuan2001@163.com

全国硕士研究生入学西医综合考试是为高等院校和科研院所招收医学专业的硕士研究生而设置的具有选拔性质的全国统一入学考试科目，目的是科学、公平、有效地测试考生是否具备继续攻读硕士学位所需要的基础医学和临床医学有关学科的基础知识与基本技能[1]。根据《西医综合考试大纲》(以下简称《考试大纲》)的规定，西医综合考试涵盖生理学、生物化学、病理学、内科学和外科学 5 门课程，设有 A1 型题、A3 型题、B 型题、X 型题 4 种。对照《考试大纲》，西医综合考试的试卷结构是否合理、各学科比例是否分配得当、考试结果的信度是否达到统计要求等，多元概化理论为这些问题的研究与论证提供了一个框架。

概化理论(GT)是一种把测量误差作为模型参数来处理的测量理论，是经典测量理论与实验设计、方差分析相结合的产物。根据测量目的的不同，研究者可以设计不同的概括全域，通过多个决策研究计算出不同概括全域上的多个信度指标，尤其是多元概化理论，在处理含有多个潜在能力因子、且因子之间存在相关的测量问题方面有独到之处。多元概化理论作为新一代的测量理论，目前已被广泛应用于国内外各种选拔性考试、标准参照性考试及教学评价中。其主要用途有：根据不同的需要，对各种标准化测验提供多个更为恰当的测量信度估计，给测验一个正确和公正的评价；对非标准化测验，准确地分析测量误差的来源并估计各种误差对分数总变异的影响程度，为改进测验提供具体的有益信息；同时估计出整套测验及其各分测验的测量信度，使得测验结果的解释更加准确和合理[2]。

在此，应用多元概化理论评估 2010 至 2012 年全国硕士研究生入学西医综合考试的试卷结构和测量精度等。在对 3 次考试分别进行多元概化理论分析的基础上，获取“平均”或“趋势”的结果，从整体上评价西医综合考试，并为今后修订《考试大纲》、改进命题方案、提高考试质量提供参考依据。

1　对象与方法

1.1　研究对象

本研究采取分层等距抽样法，抽取 2010 至 2012 年度考生 15 765 人及相关考试数据进行统计分析(表 1)。

1.2　研究方法

根据多元概化理论的研究设计，从测量学科能力的角度，可以将西医综合能力看成是包含生理学、生物化学、病理学、内科学和外科学 5 个学科因子的单面 $p^0 \times i^0$ 设计。另外，可以将西医综合考试看成是包含 A1 型题、A3 型题、B 型题、X 型题 4 种题型因子的 $p^0 \times i^0$ 设计(表 2)。

表 1　2010 至 2012 年西医综合考试成绩的描述统计($\bar{x} \pm s$)

年份	人数	成绩(分)	难度	α 信度
2010	5 225	150.58 ± 39.80	0.502	0.940
2011	2 540	156.70 ± 41.58	0.522	0.944
2012	8 000	162.01 ± 46.38	0.540	0.958

表 2　2010 至 2012 年西医综合考试的试卷结构

学科	A1型题数量	A3型题数量	B型题数量	X型题数量	合计
生理学	24	0	6	6	36
生物化学	16	0	6	6	28
病理学	16	0	6	6	28
内科学	20	20	6	6	52
外科学	14	10	6	6	36
合计	90	30	30	30	180

1.3　统计学分析方法

应用 mGENOVA 软件，分别进行考生(p)、试题(i)以及考生与试题之间交互效应(p × i)在 5 个学科因子和 4 种题型因子上的方差和协方差分量的估计矩阵。

2　结果

2.1　关于 5 个学科因子的分析

2.1.1　G 研究：各效应在 5 种学科因子上的方差协方差分量

在 5 个学科因子中，2010 至 2012 年方差分量最大的均是生物化学，其次为病理学、生理学和内科学，外科学的方差分量相对较低。2010 和 2011 年各学科方差分量高低相差幅度较大，2012 年相对趋于平均。这说明在研究生入学西医综合考试中，各学科作用逐渐平均，生物化学发挥作用最大，其次为病理学和生理学，外科学的作用还应进一步加强。同时，根据协方差分量的估计值，研究发现连续 3 年各个学科与其他 4 个因子的协方差分量值都小于 0.100，说明考生在各学科所得分数高低的一致性程度相对较低，各学科因子在预测考生能力水平方面发挥各自不同的作用。另外，与考生方差相比，内科学和外科学学科因子的试题方差较大，这表明对于考生来说，这两门学科的试题在难度上具有较大的

变异性,而生物化学等其他学科的试题在难度上变异性较小。考生与试题之间的交互方差在 5 个学科因子上都远大于考生方差，这表明各学科因子的相对难度是因人而异的[3](表 3)。

表 3　G 研究中各效应在五个学科因子上的方差和协方差分量估计

年份	效应	生理	生化	病理	内科	外科
2010	p	0.035	0.929	0.964	0.902	0.902
		0.050	0.082	0.964	0.822	0.822
		0.047	0.072	0.067	0.910	0.897
		0.035	0.048	0.048	0.042	0.945
		0.031	0.044	0.043	0.036	0.035
	i	0.149	0.085	0.066	0.132	0.128
	pi	0.438	0.481	0.488	0.606	0.599
2011	p	0.067	0.974	0.980	0.905	0.901
		0.078	0.095	0.947	0.849	0.826
		0.068	0.078	0.071	0.936	0.940
		0.046	0.051	0.049	0.038	0.972
		0.037	0.040	0.040	0.030	0.025
	i	0.083	0.048	0.095	0.143	0.193
	pi	0.480	0.482	0.461	0.618	0.528
2012	p	0.055	0.961	0.897	0.861	0.723
		0.058	0.066	0.952	0.869	0.780
		0.049	0.058	0.055	0.960	0.917
		0.041	0.046	0.046	0.042	0.919
		0.030	0.035	0.038	0.033	0.031
	i	0.107	0.054	0.058	0.080	0.068
	pi	0.374	0.518	0.518	0.650	0.617

注:p：主对角线上的元素为各效应在相应子类型上的方差分量估计;主对角线以下元素为各效应在不同因子间协方差分量的估计;主对角线以上元素为因子间的相关系数的估计。i 和 pi:由于假设题目是从各个侧面下随即抽取的,各个维度之间的题目没有相关,是独立的,因此仅报告 i 和 pi 的方差分量,并以横向形式列出

2.1.2　D 研究：各学科因子全域分估计及全域总分估计的精度问题

按照 5 个学科因子试题量所占比重来决定权系数,对 5 个学科因子全域分进行合成,可以得到全域总分的方差以及相应误差的方差分量估计，进而估计全域总分的概化系数[4]。2010 和 2011 年生物化学的相对误差小,概化系数最高,2012 年概化系数略有小幅降低,整体上比较稳定,这表明生物化学的测量精度相对较高。生理学在 2011 和 2012 年概化系数相比 2010 年有所提高,这表明生理学的测量精度得到进一步提高。病理学和内科学的概化系数一直较为稳定，表明这两门学科的测量精度比较稳定,也较为理想。外科学的概化系数在 2010 至 2012 年间比较稳定,介于 0.600~0.700 之间,其测量精度处于可接受水平,但与其他学科相比相对较低,有待进一步提高。整体上,分析 3 年的考后数据,按照 5 个学科因子合成的全域总分的相对误差很小(小于 0.005),概化系数均在 0.930 以上,表明西医综合考试的整体测量精度比较理想(表 4)。

2.2　关于 4 种题型因子的分析

2.2.1　G 研究：各效应在四种题型因子上的方差和协方差分量

2010 年 4 种题型因子的方差分量最大的是 X 型题,其次为 A3 型题,最小的为 B 型题;2011 年方差分量最大的为 A1 型题,其次为 X 型题,最小的为 A3 型题;2012 年方差分量最大的为 X 型题,其次为 A3 型题,最小的为 B 型题。3 年的结果各有不同,这说明在西医综合考试中,各题型都发挥了重要作用,总体上,X 型题和 A1 型题作用相对较大,B 型题作用相对较小,试题内容本身会直接影响到题型的测试效果。同时,根据协方差分量的估计值,研究发现各个题型与其他 3 个因子的协方差分量值都小于 0.100,说明考生在各题型上所得分数高低的一致性程度相对较低，各题型因子在预测考生能力水平方面发挥各自不同的作用[5]。另外,与考生方差相比,各题型因子的试题方差均较大,这表明对于考生来

表 4　考生全域分数等在 5 个学科因子以及全域总分上估计的方差分量值

年份	指标	生理	生化	病理	内科	外科	全域总分
2010	全域分数方差	0.035	0.082	0.067	0.042	0.035	0.044
	相对误差	0.012	0.017	0.017	0.012	0.017	0.003
	概化系数	0.740	0.827	0.794	0.784	0.675	0.937
2011	全域分数方差	0.067	0.095	0.071	0.038	0.025	0.049
	相对误差	0.013	0.017	0.016	0.012	0.015	0.003
	概化系数	0.835	0.846	0.812	0.762	0.633	0.944
2012	全域分数方差	0.055	0.066	0.055	0.042	0.031	0.043
	相对误差	0.01	0.018	0.019	0.013	0.017	0.003
	概化系数	0.840	0.782	0750	0.769	0.642	0.934

说,各种题型在难度上具有较大的变异性;考生与试题之间的交互方差在 4 个题型因子上都大于考生方差,这表明各题型的相对难度是因人而异的(表 5)。

表 5　G 研究中各效应在四个题型因子上的方差和协方差分量估计

年度	效应	A1型题	A3型题	B型题	X型题
2010	p	0.045	0.857	0.961	0.956
		0.040	0.048	0.810	0.840
		0.039	0.034	0.037	0.918
		0.052	0.048	0.045	0.067
	i	0.093	0.153	0.068	0.115
	pi	0.419	0.801	0.455	0.697
2011	p	0.055	0.866	0.992	0.968
		0.041	0.040	0.9890	0.846
		0.053	0.040	0.051	0.972
		0.052	0.039	0.050	0.053
	i	0.058	0.216	0.086	0.125
	pi	0.440	0.726	0.414	0.724
2012	p	0.048	0.848	0.950	0.950
		0.045	0.059	0.800	0.786
		0.040	0.038	0.038	0.869
		0.061	0.056	0.050	0.087
	i	0.079	0.129	0.108	0.122
	pi	0.411	0.810	0.371	0.631

2.2.2　D 研究：各题型因子全域分估计及全域总分估计的精度问题

按照 4 个题型因子试题量所占比重来决定权系数,对 4 种题型因子全域分进行合成,可以得到全域总分的方差以及相应误差的方差分量估计,进而估计全域总分的概化系数[6]。综合 3 年来数据,A1 型题和 B 型题的相对误差小,A3 型题和 X 型题的相对误差较大;考虑到误差因素,测量精度(概化系数)最高的是 A1 型题,3 年均在 0.900 以上,其次为 X 型题和 B 型题,A3 型题的测量精度相对较低。3 年西医综合考试全域总分的相对误差都非常小(小于 0.005),概化系数均在 0.930 以上,这表明,从题型合成角度来看,整体测量精度非常理想(表 6)。

表 6　考生全域分数等在 4 个题型因子以及全域总分上估计的方差分量值

年份	指标	A1型题	A3型题	B型题	X型题	全域总分
2010	全域分数方差	0.045	0.048	0.037	0.067	0.044
	相对误差	0.005	0.027	0.015	0.023	0.003
	概化系数	0.905	0.644	0.707	0.742	0.937
2011	全域分数方差	0.055	0.040	0.051	0.053	0.049
	相对误差	0.005	0.024	0.014	0.024	0.003
	概化系数	0.919	0.626	0.786	0.686	0.944
2012	全域分数方差	0.048	0.059	0.038	0.087	0.049
	相对误差	0.005	0.027	0.012	0.021	0.003
	概化系数	0.913	0.685	0.753	0.806	0.946

3　讨论

3.1　从考试内容看西医综合考试

从考试内容来看,各学科因子的方差和协方差成分的估计是比较稳定的,这说明各个学科试题的命制是严格遵循《考试大纲》的,各个学科的分量分配是基本合适的。相对而言,基础学科的贡献总体较高,外科学的分量和贡献还需要进一步平衡。从 D研究的结果看,5 个学科因子的测量误差小，测量精度均较高[7]。

从概化理论 D 研究的角度来讲，提高样本容量,信度会有一定程度的提高。但是考虑到西医综合考试的试卷共有 180 道题,已经是比较长的试卷,因此如果简单增加题目数量无疑会加大考生负担。对各学科题目数量进行分析发现,基础学科的测量精度高,是否可以适当减少基础学科的题目数量,增加外科学的题目数量,以提高外科学的测量信度,还有待进一步研究探讨。另外,进一步提高外科学的试题区分度也是改进外科学试题质量的一个方向。

3.2　从题型设计看西医综合考试

从题型来看,各题型的方差和协方差成分的估计是比较稳定的,不同年度的试卷在结构上可看成是“平行的”。从 D 研究的结果来看,按照题型合成分数,由于都是客观的选择题型,总体上测量误差小,测量精度非常理想。

A3 型题是临床病例分析题,是在一个同样的病例资料题干背景下,设置关于病因、诊断、治疗等不同方面的 2~3 个问题,能够较为灵活地考查学生的临床思维和分析问题、解决问题的能力。从 3 年数据来看，相对而言,A3 型题的测量精度略低于其他题型,一方面是适当增加 A3 型题的题目数量,另一方面，进一步提高 A3 型试题本身的题目质量和区分度,是提高 A3 型试题测量信度的方法。

西医综合考试连续多年采用的都是客观题的形式,题型相对统一,答案精确,评分误差影响很小,因此采用多元概化理论进行试卷分析的适用度很高。但是,由于国家大规模考试的局限性,本研究在使用多元概化理论分析试卷时也存在一定不足;例如,由

·教学评价·

以培养应用型医学人才为目标的机能实验考核体系改革

朱丽艳　杨秀红

【摘要】 机能实验学在基础理论与临床实践之间起着重要的衔接作用,为培养高素质医学人才,各医学院校改进了机能实验教学方法,取得一定效果,但仍缺乏全面综合的考核体系。我校根据机能实验学教学基本要求,以 2010 级临床医学学生为考核对象,设计考核方法,制订考核内容及标准,采用平时成绩、技能操作考核、笔试并结合实验竞赛对学生进行综合评价。实践表明,该考核体系对提高实验教学质量、培养学生实践能力和综合素质具有一定意义。

【关键词】 机能实验学; 考核标准; 应用型医学人才

【中图分类号】 R33-33

Exploration of functional experiment evaluation system to cultivate practical medical talents

Zhu Liyan, Yang Xiuhong. Function Laboratory of Basic Medicine College, Hebei United University, Tangshan 063000, China

Corresponding author: Yang Xiuhong, Email: yangxiuhong38@yahoo.com.cn

【**Abstract**】 Functional experiment plays an important role in basic theory and clinical practice. For training high-quality medical talents in medical colleges, many improvements have been made in the functional experiment teaching and certain effects have been obtained. But a comprehensive evaluation system was still needed. Clinical medical students of 2010 grade were taken as the appraisal objects. Assessment content, methods and standards were designed according to the basic requirements of functional experiment. Students were evaluated by ordinary performance, test papers and experiment competition. Practice showed that the system was of significant importance in improving the experimental teaching quality and developing students' practical ability and comprehensive quality.

【**Key words**】 Functional experiment; Assessment criteria; Practical medical talents

DOI:10.3760/cma.j.issn.2095-1485.2014.01.002

作者单位:063000 唐山, 河北联合大学基础医学院机能实验室

通信作者:杨秀红, Email: yangxiuhong38@yahoo.com.cn

于是国家统一考试,每年都使用不同的试卷,没有共同题和共同人,无法对不同年度的分数进行等值处理,可能在一定程度上会影响到分数的可比性。

参考文献

[1] 教育部考试中心. 全国硕士研究生入学统一考试: 西医综合考试大纲[M]. 北京: 高等教育出版社, 2010.

[2] 杨志明, 张雷. 测评的概化理论及其应用[M]. 北京: 教育科学出版社, 2003.

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

[4] 陆一萍. HSK 高等考试信度的多元概化理论研究[J]. 中国考试, 2011, 3(2): 56-58.

[5] 李灿, 曾有良, 王晓红. 2010 年湖南高考理科数学试卷分析[J]. 当代教育理论与实践, 2011, 3(2): 56-58.

[6] 田霖, 韦小满, 王桥影. 基于概化理论构建自学考试信度分析框架[J]. 考试研究, 2013, 40(5): 27-33.

[7] 王晓华, 文剑冰. 多元概化理论在高等教育达标性考试中的应用[J]. 心理科学, 2010, 33(5): 1223-1226.

(收稿日期:2013-10-12)

(本文编辑:张学颖)