概化理论是独特的测量理论体系，对于分析测验结构合理性和探讨提升测验精度的方法是非常有帮助的，在内容和运用范围上是对经典测量理论的扩展和延伸。本文在介绍概化理论的基础上，结合教育部考试中心组织开发的《兴趣测验》，探讨了概化理论在测验设计中的作用。

概化理论及其在测验设计中的应用

□ 北京师范大学　朱小姝
教育部考试中心　关丹丹

人的心理特质具有主观性和内隐性的特点，在对其进行测量时，只能通过行为表现来获得相关信息，并依据一定的心理测量理论进行解释。长期以来，经典测量理论(Classical Test Theory，CTT)一直占据测量理论的统治地位，却存在误差分离过于笼统、"严格平行测验"很难在实际情境中实现等问题。针对CTT存在的问题，概化理论(Generalizability Theory，GT)将经典测量理论的内容和运用范围进行了扩展和延伸。本文在简要介绍概化理论的基础上，结合教育部考试中心组织开发的《兴趣测验》，探讨概化理论在测验设计中的作用，希望能够引起教育测量界对概化理论更深入关注和运用。

1　概化理论简介

在GT中，测量情境关系由测量目标(object of measurement)和测量侧面(facet of measurement)构成。测量目标，即测验中所要描述的特性，不仅仅是受测者的某种潜在特质，也可以是测验题目或评分者的某种特性(杨志明，张雷，2003)。测量侧面则是影响和制约测量目标的各种因素和条件，包括测量工具、测量环境、测量时间等。测量侧面又可分为随机侧面(random facet)和固定

侧面(fixed facet):随机侧面中,侧面各水平是从所有可能的水平中随机选取;固定侧面的各水平则是固定不变的,在GT模型中,至少需要包含一个随机侧面才能进行推广或概化(Brennan,2000b)。

一个测量侧面所有可能水平的全体称之为可接受的观察全域(universe of admissible observations)。测量对象在观察全域上的观察均分称为全域分数(universe-score),类似于经典测量理论中的真分数。实际测量活动中,测量面具有的特定条件样本对应的条件总体称为条件全域(universe)。因此,观察全域应该为所有测量侧面条件全域的集合,则全域分数应是所有条件全域上观察分数的均值。概括推论测验结果时所涉及的测量面条件全域的集合叫做概括全域或概化全域(Universe of Generalization)。

在GT中,将CTT中的"信度"转化为概化系数$E\rho^2$或可靠性指标φ系数。概化系数关注的是测量的相对误差,即测验设计中侧面和测量目标之间的交互作用。可靠性指数则关注的是绝对误差,即所有侧面的主效应和侧面及测量目标之间的交互效应的方差分量。

GT研究过程由两大部分组成,G研究和D研究。G研究是指在观测全域上,根据测量设计对测量目标、所有侧面以及它们之间的交互作用的方差协方差分量进行估计。在这个研究中,需要研究者明确测量对象和测量目标、测量侧面和观测全域以及它们的关系,还包括对测量设计和测量模式的确定。测量目标和测量侧面形成3种测量设计,分别是:交叉设计(cross design)、嵌套设计(nest design)和混合设计(mixed design)。每种测量设计都对应相应的测量模型及其假设。在测量模型中,将观察分数分解为总体均值和各种误差变异效应的累加。依据测量设计收集样本数据后,运用ANOVA的分析方法估计观察全域的方差分量,并确定测量目标、测量侧面及其交互作用的方差分量。这些方差估计值将为有效的测量方法提供一定信息。D研究则是在G研究基础上,通过改变测量侧面结构、测验模型等来考察概化系数和可靠性指数的变化,从而为有效控制误差、提高测验精度提供参考。其中,需要根据测量目的确定概化全域,也就是确定测验结果推广的侧面,及各侧面推广的范围。然后根据确定的概化全域,在各侧面条件样本水平上重新估计G研究中各因素的效应和交互作用的方差分量,获得特定概化全域上的整个测验的概化系数和可靠性指数。通过多次反复,获得不同概化全域上的系数指标,比较这些系数的估计精度,从而确定最佳的测量设计方案,将G研究中结果概化到新的全域上。

由于在实际的测量中,常会涉及一个测量目标同时具有多个全域分数的问题。比如一个测验包括多个分测验,这些分测验的分数就可理解为同一测量目标所具有的多个全域分数。于是在单变量概化理论的基础上发展出了多元概化理论,多元概化理论在继承了单变量概化理论的思想基础上,还提供了测验目标、测量侧面等因素更为详细的方差协方差分量的信息,具有更为广泛的使用范围。

2 概化理论在兴趣测验设计中的运用

为了加强对中学生升学和就业过程的指导,提高升学与就业指导工作的科研水平,许多研究者尝试编制了职业兴趣测验,因为通过职业兴趣测量可以得到个体对于职业的兴趣类型,进而通过"匹配"原理为其找到适合的专业、职业提供更科学合理的指导。教育部考试中心组织开发的《升学指导测验》中的《兴趣测验》显示,我国高中生职业兴趣类型不同于霍兰德(Holland)职业兴趣理论中所描述的6种类型,而是7种,具体为:技术型、研究型、艺术型、社会型、经营型、事务型、自然型。其中自然型是具

有我国特色的一种兴趣类型，具有自然型兴趣的人喜欢户外活动，对大自然中的事物充满了浓厚的兴趣，喜欢探索生命现象，了解各种动植物的生活习性和生长发育规律，实干意识比较强。尽管以往对《兴趣测验》的信效度分析（依据CTT理论）均显示符合测量学的要求，但对于自然型分测验的测量精度并没有做过详细的分析，对于同一兴趣类型下（20个题目组成），是否所有题目反映的活动内容都相同；如果存在差异，是否会影响受测者最终的分数，多元概化理论为我们提供了一种新的分析思路。

这里，就以《兴趣测验》中的自然型分测验为例，在2005年网上测试数据中随机抽取1 005人，运用概化理论对自然型分测验的测验精度进行分析。基本思路是：1）分析同一类型下，题目之间形成的子兴趣类型对整个类型的得分的影响是否存在差异；2）为保证整体得分可靠性，应如何确定现有子类型题目数量。在这里，测量目标是受测者（p）在自然型兴趣类型上的喜好评定（6点量表）；观察全域为一个随机侧面——分测验中的题目形成的子兴趣类型（i），所有受测者需要完成整个测验的所有题目。因此本研究为单侧面交叉设计（$p \times i$）的多元随机效应概化模型。在同一职业兴趣类型内，对题目进行因素分析得到不同的子类型，即子兴趣类型，这些子类型构成概化研究中的变量，即“元”。这里得到的子类型是从同一个兴趣类型之下分析得到的，属于随机侧面。

2.1 G研究及结果分析

整个自然型分测验的20道题目经过因素分析形成3个子类型：因素N_1主要涉及动植物生长发育的探索活动；因素N_2主要涉及在自然环境下进行的考察活动；因素N_3主要涉及对生物生命现象的认识活动。

由于G研究为单侧面交叉设计（$p \times i$），通过mGENOVA分析可以得到受测者（p）效应、子类型（i）效应以及受测者和子类型之间的交互效应（或称残余效应）（pi）在3个子类型上的方差和协方差估计（见表1）。

由表1可知，对于受测者效应，自然型分测验的2个子类型中的子类型N_1的方差分量最大，子类型N_2的方差分量最小。说明在该兴趣类型中，子类型N_1对分测验总分的作用最大，子类型N_2的作用最小。另一方面，根据协方差估计值发现，3个子类型之间的协方差分量都大于0.81，说明受测者在3个子类型上的反应比较一致，能够反应受测者在该量表上较统一的反应模式。3个子类型间的相关系数中，两两相关系数都在0.65正相关水平之上，同样说明受测者在各个子类型上的反应基本一致，比如偏好动植物生长发育的探索活动的受测者同样也对自然环境下进行的考察活动有兴趣。在子类型效应上，子类型N_2具有最高的方差分量（0.229）；子类型N_3的方差分量最低（0.032）。子类型效应说明的是各个子类型所包括的题目对所在子类型总体得分的影响。子类型N_2的方差分量最大，则说明子类

表1 G研究中各效应在3个子类型上的方差协方差分量估计

效应	N_1	N_2	N_3
p	0.844	0.841	0.818
	0.651	0.710	0.870
	0.676	0.660	0.809
i	0.103		
		0.229	
			0.032
pi	0.789		
		0.933	
			1.011

注：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
主对角线以上元素为各效应在不同子类型间协方差分量的估计
主对角线以下元素为子类型间的相关系数的估计。

型 N_2 中题目可能在内容上与相同子类型下其他题目的差异要相对大一些。在残余效应上，子类型 N_3 具有最大的方差分量(1.011)，子类型 N_1 的方差分量最小，说明子类型 N_3 中受到的其他测量因素影响要比其他子类型大。这里需要说明的是，方差分量的数值大小直接反映了子类型在某种效应上的作用大小，方差数值越大，则对效应的作用也越大。通过分析可以发现，受测者在子类型 N_1 中题目的作答情况对其在自然型分测验的总分高低有最大的影响，并且子类型 N_1 所受到的无关因素的干扰相对小于其他两个子类型。

以上分析将为进一步的D研究提供最初步的数据基础，对整个分测验结构合理性和改进可能，以及概化到其他测量情景的分析将在D研究中进行。

2.2 D研究及结果分析

D研究在G研究估计的方差协方差矩阵基础上，进一步估计受测者在3个子类型上的全域分数以及相应的误差估计的方差分量，从而获得概化系数。从表2可见，子类型 N_3 的相对误差和绝对误差在3个子类型中最大，说明子类型 N_3 受到受测者和题目之间的交互效应以及子类型主效应对测验分数的影响相对其他两个子类型来说要大。3个子类型的概化系数都在0.8以上，说明该分测验各个子类型的测量精度都达到较好水平。子类型 N_1 的概化系数最高(即信度最高)，说明这一测验中关于动植物生长发育的探索活动的测量精度相对其他子类型要好些；另一方面，在子类型 N_3 上的概化系数最低，说明对生物生命现象的认识方面的测量精度相对于其他子类型要稍差一些，但由于概化系数同样达到0.8，说明该子类型题目还是具有较高的信度。

另外，根据G研究中得到的各子类型的方差和协方差分量矩阵，在D研究中将改变各个子类型的题量从而获得不同子类型对分测验整体测量精度的影响。D研究中改变题量的方法，一是在不改变整体题量数的条件下，将3个子类型的题量均分。因素分析提取的3个子类型题量存在差异，由于在编制测验之前并没有明确提出某一子类型重要性要强于其他子类型，因此将题量均分，既是赋予各个子类型相同权重，考察子类型权重一致情况下分测验的测量精度，并与原有设计进行比较。二是同样不改变整体题目量数情况下，单独增加某一子类型题量，相应减少其他子类型题量，考察是否可以由一个子类型来代替其他子类型，以便提高测验的精度。D研究中各子类型题量设计方案见表3。

同样，在改变了原有分量表的

表2 D研究中受测者在3个子类型上估计的方差分量值

	N_1	N_2	N_3
全域分数	0.844	0.710	0.809
相对误差	0.099	0.133	0.202
绝对误差	0.111	0.166	0.209
均值误差	0.014	0.034	0.007
概化系数	0.895	0.842	0.800
可靠指数 Φ	0.883	0.811	0.795

表3 G研究和D研究的题量设计情况

		N_1	N_2	N_3
G研究				
原有设计题量		8	7	5
D研究				
D0均分模式		7	7	7
D1增加子类型 N_1	D1.1	10	6	4
	D1.2	18	1	1
D2增加子类型 N_2	D2.1	7	9	4
	D2.2	1	18	1
D3增加子类型 N_3	D3.1	7	6	7
	D3.2	1	1	18

题量设计基础上，也可以得到改变后分测验每个子类型的概化系数和贡献率，结果见表4。

在这里，最重要指标为合成的全域总分的概化系数。本研究关注的是相对决策，因此使用概化系数作为信度指标。在原有设计的基础上合成的G系数为0.9404，属于较高水平。整个分量表的相对误差方差分量仅为0.045，说明自然型这一分量表的总体测量信度是可以接受的。

D0研究中，将各子类型的题量进行均分，保证所有子类型的题量一致，使得题量相对原有题量增加一题，达到21题。结果发现，合成后的G系数(0.9420)要高于原有设计模式(0.9404)，即将该类型的3个主要子类型赋予相同的权重能够提高测验的精度。但这里还需要考虑由于增加了总体题量所带来的测量精度的提高。另一方面即使赋予相同权重情况下，子类型N_1对分测验总分的贡献率还是要高于其他子类型。

另外，子类型N_1题量变化对合成G系数的影响较其他两个子类型要更显著。在子类型N_1题量数增加3个的情况下，合成的G系数要大于子类型N_2和N_3题量数增加3个的情况下合成的G系数，达到0.9424；同时在固定题量数为17题的情况下，子类型N_1可使合成的G系数达到最大为0.9525。改变子类型N_1的两种情况下所获得的合成G系数都要高于原有设计题量下合成的G系数。相反，子类型N_2和N_3在增加题量的情况下，却使合成的G系数低于或近似等于原有设计方式下的G系数。

表4 D研究中各子类型G系数、贡献率及合成G系数

			N_1	N_2	N_3	合成
原有设计模式		G系数	0.8953	0.8420	0.7999	0.9404
		贡献率	41.63%	33.44%	24.92%	
均分模式	D0研究	G系数	0.8821	0.8420	0.8484	0.9420
		贡献率	34.26%	31.90%	33.84%	
增加子类型N_1	D1.1研究	G系数	0.9145	0.8204	0.7618	0.9424
		贡献率	52.47%	28.07%	19.46%	
	D1.2研究	G系数	0.9506	0.4323	0.4443	0.9525
		贡献率	91.75%	4.04%	4.21%	
增加子类型N_2	D2.1研究	G系数	0.8821	0.8727	0.7618	0.9396
		贡献率	36.28%	43.81%	19.92%	
	D2.2研究	G系数	0.5167	0.9320	0.4443	0.9378
		贡献率	4.72%	90.51%	4.76%	
增加子类型N_3	D3.1研究	G系数	0.8821	0.8204	0.8484	0.9395
		贡献率	35.99%	28.50%	35.51%	
D3.2研究		G系数	0.5167	0.4323	0.9350	0.9401
		贡献率	4.36%	4.23%	91.41%	

总体上来看，由于测验本身的测量精度较好，即使增加子类型N_1题量，概化系数的增加也只有0.01左右，所以，无论是增加子类型N_1(涉及动植物生长发育的探索活动)中的题量、减少其他子类型题量，还是对各子类型的题量进行均分，都不能在较大程度上提高分测验整体的测量精度。

3 总结

本文简单的介绍了概化理论的原理，并运用多元概化理论对职业兴趣测验中自然型分测验的设计结构进行分析，结果证明具有我国特色的"自然型"分测验具有较好的测量精度，组成分测验的3个子类型对总体测量精度的影响是存在差异的，偏好动植物生长发育的探索活动的子类型对分测验整体得分影响要大于其他兴趣子类型，但整个分测验结构还是合理的。

总之，概化理论具有独特的理论体系，将原有的测量条件的总体和特定的测量条件进行联系，G研究针对观察全域、D研究针对条件全域，对测验结果做出不同推论。概化理论对于分析测验结构合理性和探讨提升测验精度的方法是非常有帮助的，在内容和运用范围上是对经典测量理论的扩展和延伸。