心理学报 2009, Vol. 41, No. 2, 167 - 174
Acta Psychologica Sinica
DOI:10.3724/SP.J.1041.2009.00167

应征公民计算机自适应化拼图测验的编制*

田建全 苗丹民 杨业兵 何 宁 肖 玮

(第四军医大学航空航天医学系心理学教研室,西安 710032)

摘 要 在文献回顾和参考外军有关资料的基础上,根据项目反应理论和空间能力测验的有关理论编制试题库。首先采用纸笔测验的形式进行预实验,探讨采用 IRT 理论编制 CAT 拼图测验的可行性。然后,在预实验的基础上对试题进行修订并扩充试题数量,编制计算机辅助测验。选择三参数 Logistic 模型,采用铆题等值设计,分 7 份不同的试卷在全国征兵心理检测的过程中对 55777 名应征公民进行施测。根据测试结果,对题目进行分析,选择高质量的题目构成 CAT 试题库,采用 a 系数分层抽样的方法控制曝光率,并采用不同的测验终止策略编制 CAT 拼图测验。最后用 WAIS 智力测验积木分测验和三门功课的考试成绩为效标,通过 72 名被试对 CAT 拼图测验进行效度验证。结果显示该测验符合项目反应理论三参数 Logistic 模型的假设,各题目参数比较理想,所编制的测验具有较好的信度和效度,可用于应征公民心理选拔的实践。

关键词 IRT;CAT;空间智力;应征公民

分类号 B841.7

1 前言

随着高科技武器的不断出现和战争形态的不断改变,未来战争对士兵的心理素质提出了更高的要求。作为人类能力结构的重要组成部分,空间能力一直受到各国军队人力资源管理部门的重视。虽然对于空间能力的界定在相关文献中还不够统一,但几乎所有以选拔为目的的心理测验都从未忽视过空间因素的存在与价值。一方面,空间能力测验,尤其是那些通过积木或折纸来完成的操作测验,其实测量的是一般能力因素,而且空间能力通常被看作是创造性思维在社会和数学思维方面的高阶能力因素(Shepard, 1978);另一方面,空间能力又被认为是具体的、低水平的思维能力。因此,人们习惯于用空间能力来预测各种操作性和技术性职业,比如木工、汽车修理等。研究表明,很多职业的成功与否都与空间能力高度相关(Alley, 2004),空间认知能力对于飞行职业更为重要(Stumpf et al., 1999)。飞行员在短暂的时间内综合处理大量的信息、做出准确判断的决策过程已成为现代战斗机飞行员飞行认知加工的主要特征,这一过程在很大程度上依赖于良好的空间表象能力(McDonald, 1981)。Peterson 等(1987)研究指出,空间能力和心理运动能力一样,对于与意愿相关的工作(will do)绩效预测力很小,但对与能力相关的工作(can do)绩效的预测性却是明确的,相关系数分别达到 0.54 和 0.49。

正是由于空间能力在军事职业中的重要性,使其成为西方发达国家的军事人员选拔的重要内容。美军在 1995 年推行计算机自适应化武装部队职业能力倾向成套测验(Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery. CAT - ASVAB)时也保留了空间能力测试的内容。

计算机自适应性测验(Computerized Adaptive Testing. CAT)是基于项目反应理论(Item Response Theory. IRT)而发展起来的一种新的心理和教育测验方式。它通过应用计算机程序为每名被试建构一份与其能力水平最佳匹配的测验,从而减少测验误差和考试倦怠。另外,其优点还表现在可以显著减少测验时间,提高测验效率,有利于保密,能够提供及时判分和反馈、便于网络化施测等方面。尤其值得称道的是通过 CAT 可以构建新的能力倾向测验,

收稿日期:2008-03-12

* 中国博士后科学基金资助课题(基金号 20080431368)。

通讯作者:苗丹民,E-mail: psych@fmmu.edu.cn

甚至将一些需要动手的测试内容添加进来，开辟测量认知能力的新领域。目前国际上的一些著名考试系统，比如 TOEFL，GRE 等都有 CAT 化版本。

作为心理测量领域的重要力量之一，军队人力资源管理部门从 CAT 出现之初，就对其表现出了浓厚的兴趣（Martin 等，1989）。因为军队常常需要在短时间内对大量的候选者的能力水平进行评估从而实现人和岗位的良好匹配，而 CAT 化测验正好满足了军方的测验要求。目前，在西方国家征兵心理检测中广泛使用的 CAT-ASVAB 也是 CAT 化能力倾向测验的先驱，它由美军陆、海、空军以及海军陆战队等多军兵种联合研制，历时达 15 年之久，创造了心理测量历史上的多个第一。比如，是第一个完全自适应化的多重能力倾向测验；是第一个可以呈现图形类测验题目的计算机自适应测验系统；第一个网络化自适应测验；第一次验证了计算机自适应多重能力倾向测验和传统的纸笔测验具有相同的效力等等。可以说 CAT-ASVAB 标志着许多技术上的重大突破。研究表明，CAT-ASVAB 与传统的纸笔测验相比，具有更好的复本信度。其结构效度也与传统的纸笔测验相一致，在预测效度方面，甚至要优于传统的纸笔测验（Hetter & Sympson，1997）。

我国的征兵心理检测系统已经走过了 6 年历程，全国百分之九十以上的武装部已经配备了计算机检测系统，开展 CAT 测验的硬件设施已经基本具备。本研究的目的就是通过在我国征兵入伍心理检测系统中加入 CAT 化的空间能力测验，从而完善测量内容，并为实现应征青年心理检测系统的 CAT 化提供理论和经验支持。

2 研究对象和方法

2.1 研究对象

2006 年征兵心理检测五个省市的 55777 名应征青年，均为健康男性，平均年龄 18.61 ± 1.22 岁，城市人口占 41.32%，农村人口占 58.68%。

2.2 方法

在文献回顾的基础上，根据项目反应理论和空间能力测验的有关理论，参考明尼苏达纸板测验（Minnesota Paper Form Board，MPFB）和 2005 年版美军 ASVAB 中拼图测验的基本题型，编制双向细目表，由课题组人员和某大学三年级学生共同编制原始题目 1200 条，最后对照双向细目表，选择 300 题构建试题库。试题为横向排列的 5 个方框，第一个方框中为构图元素，后 4 个方框为被选答案，要求被试选出由构图元素正确连接形成的图形。如图 1 所示。

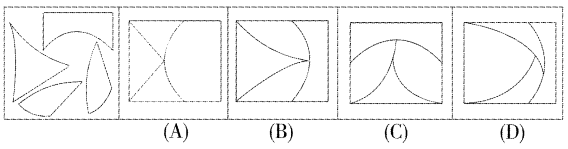

图 1 图形试题举例

首先选取 45 题采用纸笔测验的形式进行预实验，探讨采用 IRT 理论编制 CAT 图形智力测验的可行性。然后，在预实验的基础上对试题进行修订并扩充试题数量，编制计算机辅助测验。选择三参数 Logistic 模型，采用铆题等值设计，分 7 份不同的试卷，每卷 48 题，在全国征兵心理检测的过程中进行施测。根据测试结果，对题目进行分析，选择高质量的题目构成 CAT 试题库。最后，编制 CAT 图形智力测验，并以 WAIS 智力测验积木分测验和三门功课的考试成绩为效标进行效度验证。

所有图形利用 CoreDraw 12.0 和 Photoshop 7.0 绘制，动画呈现用 MACROMEDIA FLASH MX 6.0；计算机编程工具：Microsoft Visual Basic v 6.0 with SP5 Enterprise Version English；参数估计采用 BILOG MG3.0 软件；使用的数据处理工具包括 EXCEL XP 和 SPSS 13.0 软件包。

3 研究结果

3.1 预试验

预试验采用了纸笔测验的形式，共 45 题，被试为 2005 年新入伍的士兵 1450 名，平均年龄19.32 ± 1.30岁，均为健康男性，无精神及脑疾患史，受教育程度从小学到大专不等。

3.1.1 资料模型适合度检验 采用因子分析的方法，第一因子特征值为8.53，第二因子特征值为2.40，二者比值为3.55，说明该测验具有较好的单维型，可以应用 IRT 来进行指导编制和参数计算。图 2 是因子分析的碎石图。

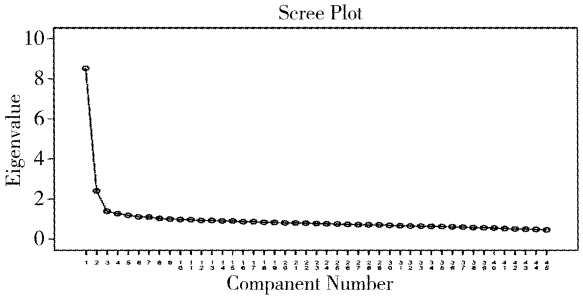

图 2 纸笔测验因子分析碎石图

3.1.2　项目参数和能力估计　选择三参数 Logistic 模型(3PL),利用 BilogMG 3.0软件包,顺利求出各项目参数和被试能力参数。统计结果见表 1。

表 1　项目参数和被试能力参数统计

参数	*n*	Min	Max	*M*	*SD*
A 参数	45	0.32	1.54	0.78	0.25
B 参数	45	-2.87	0.98	-1.31	0.83
C 参数	45	0.08	0.33	0.20	0.05
能力值	1450	-3.46	1.96	0.00	1.00

3.1.3　测验信息函数　在能力参数为 -0.70处,可得到测验信息函数(Test Information Function, TIF)的最大值 12.30。结果见图 3。

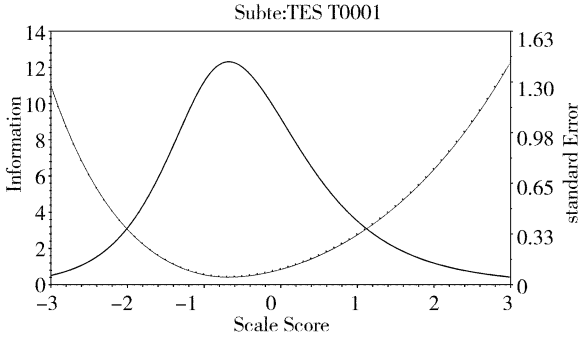

图 3　纸笔测验总信息函数分布图

从测验信息函数曲线来看,该测验对于能力分布在[-1.5,0]这一区间的被试有较好的测量效果。

3.1.4　测验对士兵绩效的预测性　所用效标为《中国士兵工作绩效评价问卷》,课题组前期研究表明该问卷有良好的信度(Luo, 2004)。根据班长对被试《中国士兵工作绩效评价问卷》中 41 个条目的评价结果得出被试的绩效成绩,计算该成绩和拼图测验成绩之间的相关系数。为了比较 CTT 和 IRT 两种测量理论所得结果之间的差异,对 CTT 和士兵绩效之间的相关也一并列出,结果见表 2。

表 2　两种能力值与工作绩效评价的相关分析 (n =91)

能力值	总绩效评价	任务绩效	关系绩效
CTT	0.28**	0.39**	0.14
IRT	0.28**	0.41**	0.19

注:**p<0.001

预试验结果表明,拼图测验满足 IRT 的理论假设,测验成绩和士兵工作绩效之间有显著相关,可以用于士兵的心理选拔。但还需要编制难度更大的项目。

3.2　计算机辅助测验

根据纸笔测验结果,在征求专家意见的基础上,对双向细目表进行修订,主要增加了试题的难度。由课题组人员根据命题要求对所编试题进行筛选,结果共保留 300 题。为了满足征兵心理检测时间的限制,共组合了 7 套试卷,每卷 48 道。采用中心放射式铆题设计,即在 7 套试卷中,有 15 道题是共同的。在 2006 年征兵心理检测中,全部采用计算机随机呈现试卷,记录反应时间并判断正误,共回收有效试卷 55777 份。

3.2.1　参数估计结果　所有题目均选用 3PL 模型,用贝叶斯估计法,由 Bilog MG 软件进行估计。在参数估计中采取了 EM 算法,将 7 份试卷同时估计,一次求出所有 246 个题目的参数。题目和能力参数的统计情况见表 3。

表 3　计算机辅助测验题目参数和被试能力参数统计结果

参数	*n*	Min	Max	*M*	*SD*
A 参数	246	0.61	3.02	1.44	0.44
B 参数	246	-1.65	2.79	-0.50	0.82
C 参数	246	0.05	0.40	0.19	0.07
能力值	55777	-3.18	2.66	-0.05	0.93

从表 3 可以看出,项目区分度参数平均为 1.44,猜测度参数平均为0.20,都比较理想,但项目难度参数平均数只有 -0.50,仍然偏简单。

3.2.2　参数不变性检验　为了验证能力参数的不变性,我们采取了将被试所作同一试卷上的题目进行分组,分别求取被试能力参数,然后将同一被试的成对估计值在直角坐标系上标点的方法。如果这些散点形成直线,就说明所选模型是恰当的(Qi 等, 2002)。同样,对于项目参数,我们也是将两个不同样组上获得的同一试卷项目的参数估计值描点,考察其是否呈线性关系。在此只选择试卷一作为样本说明。在检验能力参数的稳定性时,将试题随机分为 2 组,每组 24 题,分别估计能力参数。由于被试人数太多,很多散点直接重合,但基本可呈一条直线。对两组数据之间进行线性回归,标准化回归系数为 0.75,基本保持在同一条直线上。

在检验项目参数时,将被试分为两组,分别估计 a,b,c 三个参数,然后检验两组参数是否呈线性分布,图 4 是难度参数 b 的散点图。

从总体结果看,项目区分度和难度参数的线性趋势比较明显,说明这两项指标的稳定性比较好,但

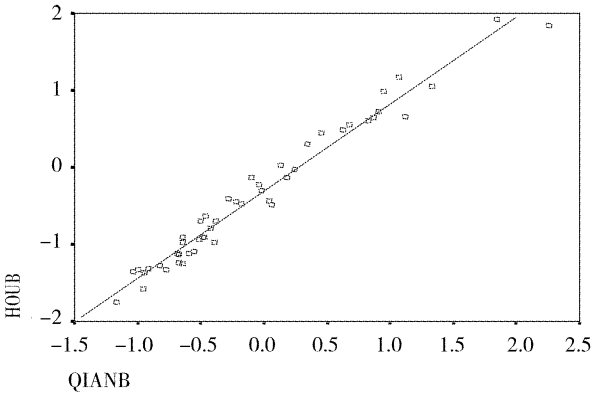

图 4 被试分组后同一套题目配对 b 参数散点图

猜测指数的线性分布不够理想。

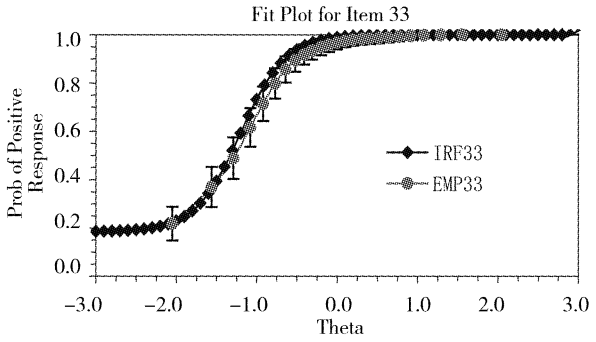

图 5 预测性良好题目举例

(Item 33: a = 2.09; b = -1.20; c = 0.18)

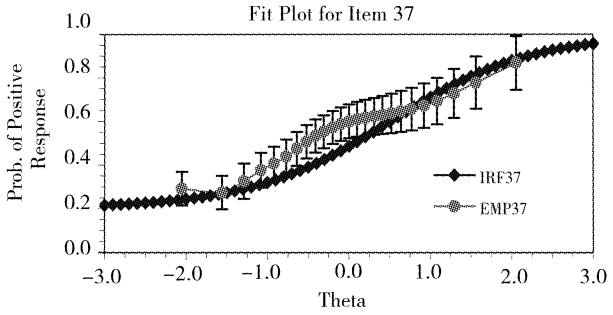

图 6 预测性较差题目举例

(Item 37: a = 0.68; b = 0.51; c = 0.20)

3.2.3 模型资料拟合度检验 在模型资料拟合度检验时我们应用了项目残差分析。这种方法就是,选定一个项目反应理论模型,估出能力与项目参数,并在假定所选模型有效的情况下对各能力水平组的成绩作出预测,然后对预测成绩和实测成绩进行比较。我们采用美国 Illinois 大学编制的 MODFIT 软件来绘制估计的项目反应曲线(Item Response Curve, IRC),并与实测资料的 IRC 进行比较,这也是建立 CAT 题库时选择题目的主要依据之一。由于题目太多,这里分别列举一个预测性好和预测性差的题目作为示例,分别见图 5 和图 6。

图中 IRF 曲线为实测资料的 IRC,EMP 曲线为估计 IRC,所有试题均来自试卷一。

最后根据检验结果,共有 20 道题目的预测性比较差,其余 226 道题的预测性良好。

3.2.4 时间对测验结果的影响 IRT 的一个重要假设就是测验不能为速度测验(No Speeding Test)。为了探索答对率和反应时之间的关系,我们将答题所用时间以 5 秒钟为单位分为 14 个区间,分别统计每个题目每个区间的答对频数,然后对数据进行拟合。从数据录入情况看,反应时和答对频数之间不可能是线性关系,我们选择了四种模型对数据进行拟合,分别是 Quadratic 模型、Inverse 模型、Compound 模型和 Cubic 模型。所有数据回归和曲线拟合以及模型评估均通过 SPSS 13.0 软件包处理完成。

以第一卷第一题为例,分别采取以上四种模型进行数据拟合并绘制拟合曲线,各模型的拟合指标见表 4。

表 4 模型拟合指标及 ANOVA 检验

模型	拟合指标				ANOVA	
	R	R^2	修正 R^2	$S.E.$	F	p
Compound	0.36	0.13	0.06	0.82	1.81	0.20
Inverse	0.61	0.37	0.31	67.92	6.96	0.02
Quadratic	0.86	0.74	0.69	45.80	15.35	0.00
Cubic	0.96	0.93	0.91	25.27	42.31	0.00

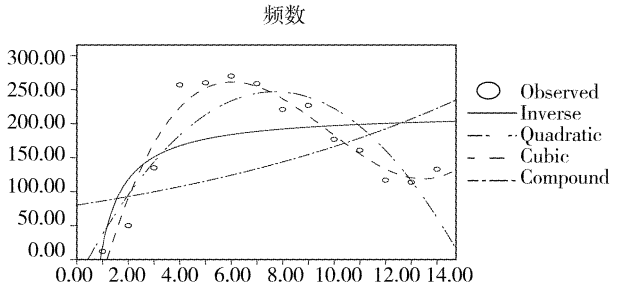

图 7 四种模型对题目 1-1 数据进行拟合的曲线图

从四种模型的拟合情况来看,Cubic 模型拟合最好,复相关系数值达到 0.96,判定系数的修订值 R^2 达到 0.91。方差分析的 F 值和概率值都很理想。对于答对频数为何会在第 65 至 70s 之间出现反向增加,可能与猜测和答题策略有关。根据这一结果,最后将每一道题的答题时间限制在 70s。

3.2.5 基于概化理论对结果的分析 为了进一步验证本测验的科学性,我们用一元概化理论(Univariate Generalizability Theory, UGT)对计算机辅助测

验的数据进行了分析,所用软件为 Brennan 研制开发的 GENOVA,分别研究了 $p \times i$ 设计随机测量模式的 G 研究变异分量和 D 研究变异分量,以第三卷为例,结果如下。

表 5　第三套试卷 $P \times I$ 设计随机测量模式的 G 研究变异分量

效应或变异来源	*df*	*SS*	*MS*	变异分量估计值
被试(p)	7881	13809.77699	1.75229	0.0332216
试题(i)	47	6824.01547	145.19182	0.0184007
交互效应(pi)	370407	58395.15120	0.15765	0.1576513

表 6　第三套试卷 $P \times I$ 设计随机测量模式的 D 研究变异分量变异来源变异分量估计 D 研究

变异来源	变异分量估计	D 研究			
		48	45	40	35
被试(p)	$\sigma^2(P)$	0.03322	0.03322	0.03322	0.03322
试题(i)	$\sigma^2(I)$	0.00038	0.00041	0.00046	0.00053
交互效应(pi)	$\sigma^2(PI)$	0.00328	0.00350	0.00394	0.00450
相对误差方差	$\sigma^2(\delta)$	0.00328	0.00350	0.00394	0.00450
绝对误差方差	$\sigma^2(\Delta)$	0.00367	0.00391	0.00440	0.00503
样本均值方差	$\sigma^2(XPI)$	0.03651	0.03672	0.03716	0.03773
概化系数	$E\rho^2$	0.91003	0.90461	0.89395	0.88060
可靠性指标	Φ	0.90057	0.89464	0.88302	0.86850

一元概化理论 G 研究各方差分量估计值如表 5 所示,被试效应、试题效应、交互效应与残差效应三者相比,由被试提供的方差分量最大,这说明,该测验能够有效的引起不同被试不同的反应,造成分数呈宽分布。一元概化理论 D 研究研究结果见表 6,该结果进一步表明,无论是作为常模参照测验,或是标准参照测验,该测验的相对误差和绝对误差都较小,具有较高的测验精度;其中,概化系数和可靠性指标均大于 0.90。此外,借助于概化研究,我们还可以考察当题目数量发生变化时,信度指标体系所发生的变动。我们设计了三种不同的题目数量,分别为 45、40 和 35。显然,随着题目数量的逐步减少,概化系数和可靠性指标都略有下降,尤其是当题目数量降至 35 时,出现了较为明显的低落,这也为编制 CAT 测验时测验的终止策略提供了依据。

3.3　CAT 测验

在计算机辅助测验的基础上选择题目参数和拟合曲线都比较理想的题目构建试题库,采取由张华华等人提出的 a 系数分层抽样的策略呈现试题。采取预定测验信息量(TIF≥10)和固定试题数目(n = 35)两种方式终止测验,分别标为 test1 和 test2。被试为某中学初中二年级学生 52 名,年龄 15.00 ± 1.10岁,均为健康男性,无家族性脑疾患史。将被试随机分为三组,分别标记为组 1、组 2、组 3,各组以不同顺序参加 test1、test2 和 WAIS 测验。test1 和 test2 由计算机集体施测,WAIS 测验由一名主试统一进行个别测试。根据学校档案记录提取被试初二年级第一学期学校统一考试语文、数学、物理三门功课考试成绩。

3.3.1　两种测验测试结果分析　两种不同终止规则之间在测验题目数量上出现了明显差异,尤其是通过测验信息量来终止测验(test1)时,施测题目数最少的只有 9 个,而最多的达到 82 个,个体之间差异显著(SD = 16.86)。但在平均测验成绩上并没有出现明显差异。两测验的描述性分析见表 7。

表 7　两种不同测验终止策略测试结果以及效标分析(n = 52)

		n	Mini	Max	*M*	*SD*
限定测验信息函数(test1)	成绩	52	35.00	101.43	64.72	12.66
	答题数	52	9	82	40.04	16.86
	正确回答数	52	1	68	23.02	13.47
	错误回答数	52	1	42	16.60	6.66
	未回答数	52	0	5	0.46	1.13
	测验用时	52	67.42	2857.30	1091.72	694.52
	平均用时	52	2.70	114.29	43.67	27.78
限定测验题目数量(test2)	成绩	52	37.70	103.90	65.57	12.06
	正确回答数	52	4	34	24.58	4.08
	错误回答数	52	1	30	10.25	3.98
	未回答数	52	0	2	0.17	0.47
	测验用时	52	163.92	1150.11	833.15	242.40
	平均用时	52	6.56	46.00	25.33	19.70
	测验信息函数	52	4.97	11.58	8.61	1.78
效标考核成绩	WAIS	52	16	45	30.62	6.33
	数学	52	17	89	49.37	18.54
	语文	52	25	85	58.44	14.07
	物理	52	25	92	66.60	18.80

从表 7 的时间结果来看,两种测验差异非常明显,test1 的平均测验用时为 1091.72s,而 test2 的平均测验用时为 833.15s,都小于 20min,但 test2 的最长用时为 2857.30s,超过 40min,这在征兵心理检测中是不可能满足的。Test2 的测验信息函数介于 4.97 – 11.58 之间,平均数为 8.61。而 test1 的终止规则要求测验信息函数大于等于 10 方可结束测验,因此测验结果更为精确。

如果不考虑效度因素,单纯从测验的可行性来看,test2 更适合在征兵心理检测的实际中应用。

表 8 两种测验成绩与效标之间相关矩阵($n=52$)

测验	test2	test1	WAIS	数学	语文
test1	0.69**				
WAIS	0.55**	0.60**			
数学	0.31*	0.33*	0.37**		
语文	0.04	0.17	0.16	0.34*	
物理	0.19	0.28*	0.26	0.63**	0.58**

注: $*p<0.05$, $**p<0.01$

从表 8 可以看出,test1 与 WAIS 智力测验积木分测验的相关系数最高,达 0.60,并与数学和物理成绩之间的相关性都达到显著水平($p<0.05$);test2 也与 WAIS 智力测验和数学考试成绩有显著相关。但三个测验都和语文考试成绩没有表现出显著性相关。

4 讨论

CAT 相对于纸笔测验有许多优越性,主要包括更高的测验效率、更高的安全性能、测验时间的个体化、被试的倦怠和受挫感减少、可以及时给被试以反馈以及可以应用新的或自己创造的测题形式等等(Howard, 1990)。这些优点使得 CAT 日益成为现代大规模测试的主流。

Reckase 列出了 CAT 测验的四大组成部分,包括题库建设、项目选择方法,能力估计方法和终止规则,除了这些基本部分,还有两项 CAT 测验经常涉及到的新课题——内容的均衡性和曝光率控制方法(Reckase, 1989)。

CAT 题库要求有足够数量高质量的题目,而且难度范围也要够大。另外,还要考虑题库在各个领域内都有足够数量的题目以满足特殊测验的要求(Howard, 1990)。题库的大小由测验的长度、被试量、项目曝光率和测验重复率的要求来决定(Bergstrom et al., 1999)。一般认为合适的题库大小应为纸笔测验题数的 6 到 12 倍(Stocking et al., 1993; Patsula et al., 1997; Yu ,1992),但是由于项目的曝光率,项目的淘汰和题库的轮替等因素,题库的数目要远比这个大。由于许多 CAT 都是连续施测的,所以项目或题库的有效期就受到了限制。Luecht (1998)提出 CAT 题库的大小应达到 3800 到 21000 之间。本测验的题库只有 200 多个题目,远远不能满足大型题库建设标准。这也提示我们可能采取确定测验信息函数的终止规则在目前的情况下可能还不可行。

CAT 中两个最常用的项目选择方法是最大信息法和欧文的贝叶斯估计法(Luecht, 1998),被选中的是对当前被试能力估计值来说能提供最大信息的题目。我们在本研究中,首先是出于对曝光率的控制,采取了 a 参数分层抽样的方法,在这一原则下,采用了信息最大法,对于这一策略的效果,尚需后续研究来验证。

关于被试能力水平的估计,我们采用了贝叶斯后验期望估计(Expected a posteriori, EAP),它将先验分布分成许多积分点而不是将其作为连续分布来进行评估(Bock et al., 1981)。由于知道先验信息,所以在测题数量相同时,贝叶斯法的标准误要小于 MLE。

对于 CAT,人们提出了许多方法来决定何时终止测验并计算最后的能力估计值。在本研究中我们尝试了固定测验题目数量和确定 TIF 值两种方法。Bergstrom 和 Lunz 等(1999)研究指出,确定 TIF 值的方法对题库的应用效果最好,但本研究却没有表现出这种效果,原因可能是题库高质量的题目还不充足。另外,采用确定 TIF 值的方法,对测验时间要求高,可能在目前的征兵心理检测工作实际中不便于推广。

由于本研究只涉及图形的拼凑和排列,只是测量空间表象加工能力,内容单一,所以不存在内容均衡方面的问题。

为了控制试题曝光率,人们提出了 Sympson-Hetter 法、随机法、0.10 对数法、限制性最大信息法等,都各有利弊,不能一概而论。本研究也应用了目前比较流行的 a 参数分层抽样法(Chang 等, 1999)。这种方法要求测验编制者按题目的区分度将题库分为 k 个层。按照区分度将各层升序排列。依据层数将测验分成相应的阶段,每个阶段从对应的层中选取一定量的题目,开始用区分度最低的,最后用区分度最高的。在一层中,随机从难度与被试能力估计值最接近的两个项目中选择施测。层的数量取决于一些因素,比如题库中项目区分度的变异越大,需要的层数越多。但如果区分度非常相近,层数就应减少。而且距区分度等级间难度范围越大,所需要的层数也越多,最后,测验长度和题库大小也必须考虑。在大型题库中,分的层数应与测验和长度相近,这样就可以从每层中抽取一个题目来组成测验。这种方法在曝光率控制上效果比较突出,因为区分度低的或高的项目被选择的可能性相同(Hau 等, 2001)。本研究根据 a 参数的大小和测验长度,将 a 分为 10 层,每层大约 20 题,并要求同一

被试同一测验过程中试题不能重复,既提高了题库中试题利用率,也有效地控制了曝光率。

我们所编制的 CAT 拼图测验测量了被试的视觉组织能力、空间想象能力以及知觉整体与部分的能力,概括起来就是空间表象加工能力。结果表明,该图形智力测验与士兵任务绩效评价表现出显著相关性($r = 0.41, p < 0.01$), CAT 拼图测验与 WAIS 智力测验积木分测验和文化课成绩也表现出显著性相关,说明该测验有较好的效度。研究结果也表明,该测验试题库还不够大,还需要补充大量高质量的题目,对于曝光率控制的效果等问题,还需要进行深入的研究.

参 考 文 献

Alley, W. E. (2004). *Experimental Test Development for the AFOQT Form "S"* (pp. 23 -27). Draft Report. Operational Technologies Corporation (OpTech), San Antonio, TX.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67 -91). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometric*, *46*, 433 -459.

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Measurement in Education*, *23*, 211 -222.

Hau, K.-T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first?. *Journal of Educational Measurement*, *38*, 249 -266

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In William Sands, Brian K. Waters, and James R. McBride (Eds.), *Computerized adaptive testing-from inquiry to operation* (pp. 141 -144). Washington, D. C.

Howard W. (1990). *Computerized adaptive testing: A primer* (pp. 17 -102) Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (1998). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Luo Zhengxue. (2004). *A Constructive Analysis and Predictive Study of Soldiers´Job Performance* (pp. 42 -63). Unpublished doctoral dissertation. Fourth Military Medical University.

Wiskoff, M. F., & Rampton G. M. (1989). *Military personnel measurement testing, assignment, evaluation* (pp. 3 -7). New York.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100 -117

Patsula, L. N., & Steffan, M. (1997). *Maintaining item and test security in a CAT environment: A simulation study*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Peterson, N., Hough, L., Ashworth, S., & Toquam, J. (1987). *New predictors of soldier performance*. Paper presented at the mid-year conference of the Society of Industrial and Organizational Psychology, Atlanta. (pp. 23 -56).

Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). *Principles of modern education and psychological measurement* (in Chinese) (pp. 15 -115). Beijing higher education Publishing Press.

[漆书青, 戴海崎, 丁树良. (2002).*现代教育与心理测量学原理* (pp. 15 -115). 北京:高等教育出版社.]

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice*, *8*, 11 -15.

Shepard, R. N. (1978). Externalization of mental images and the act of creation. In B. S. Randhawa & W. E. Coffman (Eds.), *Visual learning, thinking, and communication* (pp. 133 -190) New York: Academic Press.

Stumpf, H., & Eliot, J. (1999). A structural analysis of visual spatial ability in academically talented students. *Learning and Individual Difference*, *11* (2): 137 -15

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277 -292.

Yu, J. Y. (1992). *Item response theory and application* (pp. 24 -56). Jiang Su publishing press.

[余嘉元. (1992). *项目反应理论及其应用* (pp. 24 -56). 江苏出版社.]

The Development of Computerized Adaptive Picture Assembling Test for Recruits in China

TIAN Jian-Quan, MIAO Dan-Min, YANG Ye-Bing, HE Ning, XIAO Wei

(*Department of Psychology, School of Aerospace Medicine, Fourth Military Medical University, Xi'an* 710032, *China*)

Abstract

According to the literature review and references of psychological testing about foreign armies, a spatial ability item bank was developed based on the principles of item response theory (IRT) and theories about spatial abilities.

At first, we conducted a Paper and Pencil (P&P) test as a pilot study to explore the possibility of developing a picture assembling test using computerized adaptive testing (CAT) form. And correlation of this P&P test and task performance subscale of Soldier Performance Assessment Scale was analyzed. Then at the basis of the pilot study, a computer-administrated test (as a part of Nationwide Psychological Testing for Recruitment) was completed to modify and enlarge the item bank further. In this test, 7 different test forms, which were linked using anchor items, were developed and tested. Data was analyzed under 3 Parameter Logistic Model (3PL) using Bilog-MG software. Qualified items were selected to form the item bank, and a CAT form picture assembling test was developed. At last, we conducted validation test using subtest of block design in WAIS and scores of three curriculums as criterion.

From the results, we could see that the CAT form picture assembling test could satisfy the three assumptions of 3PL model, including unidimensionality, local independence and no speeding. And discrimination, location and guessing parameters, which were estimated using marginal maximum likelihood estimation (MMLE), were satisfying. So were ability parameters, estimated using Bayes expected a posterior estimation (EAPE). Furthermore, both test's and items' information functions were good. On the other side, item exposure can be controlled properly by strategy of using maximum information procedure and a-stratified method in this test. And on the basis of current item bank, using fixed-length stopping rule is appropriate for Nationwide Psychological Test for Recruitment. As for the validation study, we found that results of P&P form of picture assembling test had significant positive correlations with task performance subscale of Soldier Performance Assessment Scale. Scores of CAT form had significant positive correlations with subtest of block design in WAIS and examination scores of math and physics, but had no significant correlation with scores of Chinese.

However, a good CAT test need a good and large item bank to be based on, so the item bank in this study should be enlarged continuously.

Key words IRT; CAT; spatial ability; enlistees