心 理 学 报 2006,**38**(6):934 ~ 940
Acta Psychologica Sinica

应用项目反应理论创建图形推理测验题库*

肖 玮 苗丹民 朱宁宁 张青华

(第四军医大学心理学教研室,西安 710032) (北京师范大学心理学院,北京 100875)

摘 要 自编235个图形推理测验题目。采用铆测验等值设计,以72个联合型瑞文测验题目为铆题,对初中到大学各能力层次的1733名男性进行了测验。使用BILOG MG3.0(边际极大似然估计)对实测数据进行了分析,采用Logsitic 3参数模型。剔除数据与模型拟合不好的题目以及信息函数最大值小于0.3的题目,最终建立一个包含181道题目的题库。该题库可以用于淘汰智力较低的应征青年。

关键词 题库建立,项目反应理论,项目等值,图形推理测验。

分类号 B841

1 前言

军事人员心理选拔(psychological selection of military personnel)是根据军事职业的特殊需要,运用心理学方法,由军事专家和心理学专家共同对报名参军的候选者进行心理素质检测与评定,选拔那些心理素质适宜军事活动要求的候选者从事军事训练活动,淘汰心理素质不适宜的候选者。各发达国家几乎都有比较完善的军事人员选拔和分类系统,这些项目的实施对提高兵员质量,加速其军队质量建设,增强其部队战斗力起到了十分重要的作用。据1988年前联邦德国"人事管理和信息系统"数字统计[1],经心理选拔后的183631名候选者中有88.0%的人训练获得成功。假如不采用能力倾向测验测试,结果只有72%的人训练成功。通过对1988年实际情况统计,没有采用能力倾向测验,估计每年要在人员训练与安置上花费18762.3万德国马克;采用能力倾向测验后,每年仅需花费8091.2万德国马克,即每年可节省10671.1万德国马克。

从外军的征兵心理检测的发展规律看:(1)通常都要经历一个从功能简单到功能复杂的过程,先是发展一般能力测验,再逐步扩大能力测验的内容和范围,然后再增加职业兴趣、气质、人格等内容。这一发展过程和其所处的社会环境、国家兵役制度以及当时人们的认识水平相联系;(2)心理检测总时间有逐步增加的趋势;(3)为了保证在有限的时间内测量更多的内容,以及保证更高的测量精度,CAT测验成为发展趋势。以美军的军事人员选拔为例,美国军事心理学界在第一次世界大战期间对1726966名军事人员实施了历史上第一次大规模的人员甄选。多项选择的Alpha测验和非语言的Beta测验,提供了定量分析方法,将候选人员分配到不同部队或工作岗位。在Alpha测验和Beta测验的基础上,美军编制出《军队一般分类测验》(Army General Classification Test,AGCT),在第二次世界大战期间施测了上百万新兵。在Maxwell RT将军的积极参与和帮助下,军队人员选拔和分类项目(army selection and classification project, Project A)在1981年正式启动,历经20年的不断研究和完善,目前该系统已成为国际上军事人员心理选拔集大成者,该系统共包括5个部分[2]:

(1)武装部队职业能力倾向测验(armed services vocational aptitude battery, ASVAB)。2002年版的ASVAB包括4个版本的纸笔测验以及计算机自适应测验。内容包括9个分测验:常识(General Science, GS)、数学推理(Arithmetic Reasoning, AR)、词汇(Word Knowledge, WK)、段落理解(Paragraph Comprehension, PC)、车辆和购物知识(Automotive - Shop Information, AS)、数学知识(Mathematics Knowledge, MK)、机械理解(Mechanical Comprehension, MC)、电子信息(Electronics Information, EI)、装配(Assembling Objects, AO)。

收稿日期:2005 - 04 - 10

* 全军指令性课题(02L003).

通讯作者:苗丹民,E-mail:psych@fmmu.edu.cn

(2)武装部队职业资格测验(Armed Forces Qualification Test, AFQT)。由 ASVAB 的 4 个分测验(数学推理、段落理解、词汇、数学知识)构成,主要用于士兵基本素质的筛选。该测验的成绩分 5 个等级,I 级为最好,V 级最差。军队一般仅允许 AFQT 为 I 级至 IIIA 级的青年入伍。

(3)空间能力测验(Spatial Tests)。通过纸笔测验的形式完成的,包括拼图(Assembling Objects)、旋转(Object Rotation)、迷津(Maze)、地图(Map)和推理(Reasoning),主要测试被试的空间视觉旋转能力、空间扫视能力(Spatial Visualization - Scanning)、空间定向能力和归纳能力。

(4)认知和心理运动能力测验(Perceptual/Psychomotor Test)。采用计算机化测验形式,包括简单反应时、选择反应时、短时记忆、认知速度/准确率、数字记忆、目标鉴别、单(双)手轨迹追踪能力等。主要检测认知加工速度、短时记忆、心理运动准确性以及肢体协调性等能力。

(5)气质、兴趣和生平资料。测验工具为:《生活背景和经历评估量表》(Assessment of Background and Life Experiences, ABLE)和《军队职业兴趣测验》(Army Vocational Interest Career Examination, AVOICE)。

我军军事人员心理选拔工作起步较晚,但发展较快,先后开展了飞行学员、汽车驾驶员、领航员、航天员、陆军初级军官、通讯兵、潜艇艇员等的心理选拔研究工作。但除了飞行学员选拔已用于实际招飞工作,军校学员心理选拔正在试点外,其余研究成果由于种种原因均未被大面积使用或根本就未在实际工作中得以应用。就我国、我军当前人们的认识水平以及选拔体制来看,我们无法一开始就建立一个象 Project A 那样全面的征兵心理检测系统,而是必须针对我国国情、军情,重点解决征兵中迫切需要解决的问题,即防止一般能力较弱和有性格偏差的人进入部队。然后,随着人们认识水平的提高和心理检测系统开发、推广经验的积累再逐步完善这一系统。由于可以借鉴各国军队的心理检测发展经验,具有后发优势,因此可以预见,这一过程不会象美军那样需要几十年,而是完全可以在 10 年内完成。

在我国现有的兵役制度下,把好士兵入伍关是提高整体士兵素质的重要措施。从目前我国征兵的实际看,仅靠具有初中和高中文凭是不能将一般能力差的人挡在军队大门之外的。因此在征兵体检过程中采用智力测验具有重要的现实意义[3]。世界各军事强国士兵心理选拔的经验也证明了这点[4]。如美军正在使用的武装部队职业资格测验就是一般能力测验。其主要考察被试言语和数学能力[5]。从静态多因素智力理论的角度来考虑,无论是 Cattell 的流体晶体智力理论,还是 Gardner 的多元智力理论,言语、数学、图形空间三个维度是智力测验主要考察的内容[6]。经典智力测验,如:斯坦福 - 比纳测验第四版、韦克斯勒智力测验等均有言语、数学、图形分析测验题目[7],而瑞文测验则完全是图形分析的题目[8]。我国征兵心理检测系统第一版的智力测验部分主要从言语和数学两个方面对应征青年进行检测。通过三年的全国试用(2002 年 ~ 2004 年),得出以下几点经验[3]:(1)征兵心理检测为基本资格测验,加之参军入伍热情不高的社会现状,因此整个心理检测的策略为"淘劣"而非"选优",即对一个较低的划界分数附近的被试实现准确测量即可;(2)由于整个心理检测系统包括智力测验、人格及心理卫生测验、心理访谈等三个环节,根据报名及录取比例,智力测验部分的淘汰率总体控制在 5% 左右比较合适。由于采用单项淘汰策略,因此单项智力测验的淘汰率不应超过 5%;(3)由于征兵体检各地区不是同步进行,即便在一个体检中心,心理检测也是分批进行的,因此控制题目的曝光是一个应被优先考虑的因素。

以上分析表明,征兵心理检测第二版除了应对已有内容进行修订外,还应该增加图形测验。以几何图形为基本元素构建的题目由于难以记忆和描述,因此"天生"具有控制题目曝光的功能,若能够建立一个大型题库,在使用时根据预先设定的规则生成若干平行试卷,或是建成计算机自适应测验系统(CAT),则可以更好地控制题目曝光[9]。由于项目反应具有项目参数不变性,测验信息函数的概念代替了测验信度等诸多优点,比经典测验理论更能准确反应测验项目的优劣。因此本研究拟采用项目反应理论来进行项目分析及题库构建。

2　方法

2.1　题目编写

以几何图形如:方框、点、园、线等为基本元素,首先构建一个基本图形,然后运用填充、组合、类比、系列关系、套合、互换、交错、相加、相减、位移、旋转等多种图形操作方式生成新的图形,要求被试找出其中变化规律,并预测下一个图形是什么(从被选答案中选择正确答案)。通过运用规则的数量来控

制难度。如图 1 所示,该图包括三种图形,每种图形有三种角度,被试只有发现这两个规则才能得出正确答案。由于该题仅采用了两个规则,因此是一个较容易的题目。较难的题目采用 4 个和 5 个规则联合应用的策略,共自编 235 个题目。

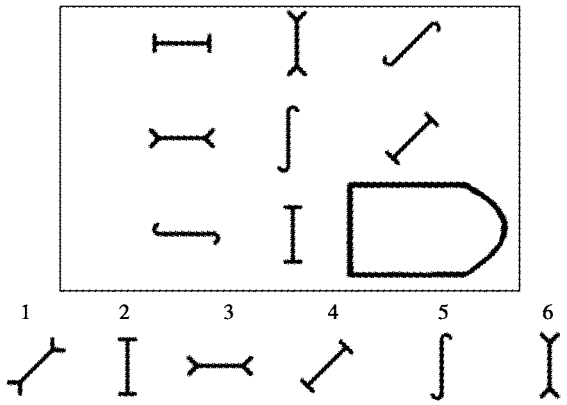

图 1 图形推理测验示例

2.2 铆测验等值设计

由于题目较多,无法在一次测验中测完所有题目,因此采用铆测验等值设计[10,11]。由于本测验的出题策略与瑞文测验类似,因此以华东师范大学李丹等修订的瑞文测验联合型版本[12]作为铆题。将题目分成 5 个测验版本,版本 1 ~ 版本 4 由自编题和联合型瑞文测验的部分题目组成,版本 5 为联合型瑞文测验,最后都等值到版本 5。为了保证各测验版本题目难度大体一致,首先将 235 个自编题目和 72 个瑞文测验题目分别粗分成低、中、高三个难度水平,由三名心理学工作者完成。每个难度水平的题数及分配到各版本中的题数如下:

自编题:低难度 113 题(29,29,29,26),中难度 58 题(14,14,14,16),高难度 64 题(16,16,16,16)。

瑞文测验联合型:低难度 32 题(8,8,8,8),中难度 28 题(7,7,7,7),高难度 12 题(3,3,3,3)。

然后将 235 道自编题及 72 道瑞文测验题按照难度水平,均分成 5 个测验版本,各版本题数及构成为:

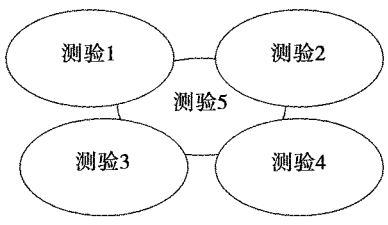

图 2 铆测验等值设计示意图

版本 1:77(29 +14 +16 +8 +7 +3);版本 2:77(29 +14 +16 +8 +7 +3);版本 3:77(29 +14 +16 +8 +7 +3);版本 4:76(26 +16 +16 +8 +7 +3);版本 5:瑞文测验联合型(72 个题)。铆测验等值设计示意图如图 2。

2.3 被试选择

由于本研究用于对应征青年的智力评价,因此被试的选择应参照应征青年的构成,同时期望被试的能力有一广泛分布,只有这样才有可能得到比较准确的参数估计结果。选择初中到大学各能力层次的男性被试 1733 人,平均年龄 18.05 岁。年龄最大 23 岁,最小 15 岁。被试具体构成如下:偏远农村中学初三年级学生 153 人、高一 141 人、高二 145 人、高三 139 人,中专一年级 126 人、中专二年级 137 人、中专三年级 144 人、大专一年级 164 人、大专二年级 153、重点高校本科一年级 207 人、二年级 224 人。将上述 11 个年级的被试各分成 5 份,分别接受 5 套测验。最终各测验所接受的测验人数分别为:测验 1:320,测验 2:369,测验 3:340,测验 4:350,测验 5:354 人。

2.4 施测及数据处理

施测采用纸笔测验形式,分组进行,每组不超过 40 人。测验时间 100min,确保无时间压力[13,14]。采用“0”、“1”记分。同时搜集初三年级组的学业成绩进行相关分析,以考察测验的效度。数据处理采用 SPSS 12.0 软件及 BILOG MG 3.0 软件进行[15]。将 5 个测验版本视为一份“大”测验同时进行估计,这种等值方法称为同时参数标定的等值方法[16]。这份“大”测验包括 5 个部分:仅包含于测验 1、2、3、4 版本的题目和共同题(测验 5)。这样可以将 5 组被试视为同一组被试,只不过回答的题目不同而已,这样估计出来的参数自然具有统一性和可比性。BILOG 软件采用的就是这个策略。实际操作时采用 3 参数 Logistic 模型,以测验 5 为基准测验,都等值到测验 5 的参数量表上。估计的内容包括:数据与模型拟合检验、各题目的参数(难度、区分度、猜测度)、信息函数曲线、被试能力估计等 4 个部分。对极容易的题目(难度小于 -3)和极难的题目(难度大于 +3)的题目分别强性收敛为 -3 和 +3。因为这些极端题目的参数估计已经不够准确,而且这些极端值会夸大等值的差异。

3　结果与分析

3.1　题目单维性检验及模型－数据拟合检验

首先根据点二列相关的结果,剔除相关系数为负值的题目,共剔除 2 个题目。然后采用探索性因素分析的方法对 5 个测验的被试应答情况进行分析,结果显示 5 个版本测验数据的第 1 因子特征根与第 2 因子特征根之比均大于 5,并且 5 个测验的 Cronbach's alpha 系数均大于 0.9,基本满足数据单维性的要求,详见表 1。需要说明的是:上述两种常用的方法在检验数据单维性上都存在着局限,特别是 Cronbach's alpha 依赖于题目的数量,当题目数量较大时,不是单维的题目也可以得出很高的 Cronbach's alpha 值[17]。因此题目单维性的检验除了采用上述两种方法外,还应该从内容效度的角度对题目进行审查,本研究从出题思路上及题目形式上尽量地控制了题目的单维性。本研究单维性检验的目的是为模型选择的恰当性提供依据,而最根本的检验还是模型－数据拟合检验。BILOG MG 3.0 提供数据与模型的似然比 χ^2 检验,结果表明:除 19 个题目外,大部分题目都符合 3 参数 Logistic 模型,遂剔除这 19 个题目,保留 216 个题目使用 3 参数 Logistic 模型进行等值及参数估计。

表 1　5 个版本测验因子分析结果

版本	第 1 因子特征根	第 1 因子可解释总方差比例(%)	第 2 因子特征根	第 1 和 2 因子特征根之比	Cronbach's alpha
1	17.22	22.37	3.25	5.30	0.902
2	15.21	19.76	2.71	5.61	0.920
3	14.29	18.57	2.49	5.74	0.922
4	16.44	21.63	3.11	5.29	0.914
5	16.54	20.43	3.15	5.25	0.907

3.2　题目整理

剔除信息函数峰值小于 0.3 的题目 33 个,累计剔除 52 个题目,最终保留 181 个题目。然后将题目按信息函数曲线峰值所处的位置将题目分成 4 个级别:< -2.5 为极低难度,$-2.5 \leqslant$ 且 $\leqslant -1.5$ 为低难度,$-1.5 <$ 且 $\leqslant 1.5$ 为中等难度,> 1.5 为高难度。从分类情况看,极低难度:32 个题目,低难度:79 题,中等难度:69 题,高难度:1 题。从题库构成上看,本研究建立的题库适合于测量低能力被试和中等能力的被试,尚不具备测量高能力被试的功能。但考虑到本研究的目的是对应征青年进行筛查,淘汰那些智力较弱的应征青年,因此对"决断点"水平附近的被试有良好的鉴别力即可。结合人群智力的正态分布以及我国征兵上站与录取比例等情况,单项智力测验的淘汰率应低于 5%,大致相当于 -1.64(正态分布推算)水平。-1.64 划界分数附近($-2 \sim -1$ 区间)有 74 个题目,信息函数峰值从 0.30 ~ 1.96,算术均值为 0.57 ± 0.14,可以基本胜任出题需要。

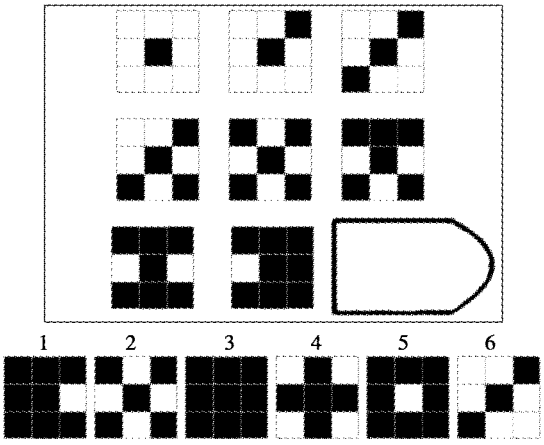

图 3　低难度题目示例

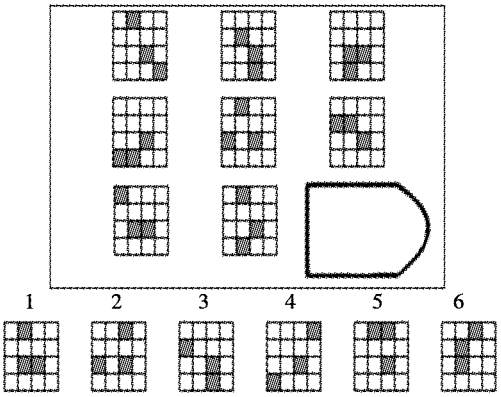

图 4　高难度题目示例

3.3　题目难度分析

本研究的题目编制运用了多种操作规则,包括:填充、组合、类比、系列关系、套合、互换、交错、相加、相减、位移、旋转等。将 1 个规则以及两个规则组合运用作为低难度题目,将 3 个规则同时应用作为中

等难度题目,将 4 个规则和 5 个规则同时应用作为高难度题目。如:图 3 所示为一个低难度题目,该题仅考察数量关系,只要发现该规则即可正确作答;图 4 所示为一个高难度题目,该题包括 5 个规则:(1)图形是每行为一组,从左到右依次变化;(2)一个色块是水平移动的;(3)一个色块是垂直运动的。(4)一个色块是静止不动的;(5)若色块移出边缘将从对面的边缘移入。只有将这 5 个规则都掌握了才能得出正确答案。将高、中、低 3 个难度的题目的难度值进行单因素方差分析,组间有显著差异,详见表 2。接着进行两两比较,采用 Student – Newman – Keuls,SNK 法,3 组题目间有显著差异($p < 0.05$)。3 组的难度均值分别为:-2.598 ± 0.780、-0.638 ± 0.274、0.569 ± 0.584。这一结果证明了当初的构想:题目的难度主要由编题时所使用的规则数所决定。

表 2 各组题目难度单因素方差分析结果

方差来源	离差平方和	自由度	均方值	F	p
组间	417.882	2	208.941	257.356	0.000
组内	268.731	331	0.812		
合计	686.613	333			

3.4 效度检验

由于客观原因,只搜集到初三年级的学业成绩。将该学业成绩与被试的能力估计进行相关分析,结果见表 4。被试的能力与学业有显著相关,特别是对数、理、化等抽象推理能力为主的学科的相关更高一些,说明本题库具有较好的同时效度。此外,这一结果也表明本研究所创建的题库对低能力组被试的测量精度及区分度达到可接受水平。对中高能力被试的效度研究需要进一步研究来证实,但从题目的难度构成上可以推测对这部分被试的测量效度较差。

表 3 初中被试能力与学业成绩的相关($n = 130$)

变量	化学	语文	代数	英语	几何	物理
能力	0.357**	0.308**	0.353**	0.303**	0.386**	0.347**

注:** $p < 0.01$

4 讨论

为了适应新军事变革的发展,我军的建军思想发生了巨大的转变,对士兵的要求由过去的“高身体素质型”向“心理、身体素质并重型”转变。但在现有的征兵工作中仅有身体检查和政审两项,没有关于心理检测的内容,因此在征兵体检中增加心理检测的内容是历史的必然要求。由于征兵心理检测被统一安排在征兵体检中进行,加之征兵体检工作流程已经形成多年,因此这些外部条件决定了:(1)心理检测的时间非常有限,单项的智力测验时间不应超过 10min,否则将会成为整个体检流程的瓶颈;(2)应将其界定为一个基本的资格测验,应以淘劣为目的而不是选优。人员的选优和分类测验应在新兵营中进行,这样更有利于人员的分类和安置;(3)控制曝光是一个应优先考虑的因素。本研究所开发的征兵专用图形推理测验题库,基本上可以满足上述要求,可对 5% 划界点附近的被试提供足够精度的测量,从而淘汰智力较差的应征青年,而中高能力被试则可以轻易、迅速地通过测验,节省了测验时间。图形测验难以记忆和描述的“天性”是目前可以找到的最不怕曝光的题型。当然,若要建立一个适合不同能力水平被试的题库,还需要再编写相当数量的高难度题目,这样才可能在将来建立一个适合不同能力水平被试的 CAT 测验系统。具体做法可以在每年的征兵心理检测过程中,将新编的题目加入正式测验中,但只记录被试的反应不算成绩。然后进行参数估计和等值,最后将合适的题目放入题库,成为正式题目。这样可以不用专门搜集数据,节省人力物力。

本研究在题目的编制和难度控制上主要采用了通过规则数来控制难度的策略。但对具体题目难度的考察时发现有难度倒置的情况,如:个别两个规则的题目难度高于三个规则的题目难度。可见除了规则数外,其它一些因素可能对题目的难度也会产生影响,这主要包括图形的复杂程度和规则本身固有的难度,如:“相加减”规则可能比“互换”规则难一些。当然,这些推想需要从认知心理学的角度来加以证实。若可以实现对规则的固有难度以及基本图形的构成难度进行界定,那么就可以建立图元库(如点、园、线、三角、方框等)和规则库,然后计算机根据预先设定的难度首先从图元库中随机提取图元任意组合生成基本图形,然后再根据预先设定的难度区间从规则库中提取若干规则对基本图形进行加工操作形成新的图形,这样可以实现计算机智能生成题目的目的。在理论上,这个系统可以生成接近无限多的题目,这样的解决方案可以从根本上控制曝光问题。

本研究在被试选择上参照我国应征青年的人员

构成,并尽量保证了能力分布范围尽可能广泛,包括从农村中学初三学生到重点大学的大二学生,从而较好地保证了参数估计的准确性。本研究将难度在 -3 以下的极容易题目强行指定为 -3,因为这些极端值会夸大等值时两次估计的差异,比如一个分别被估计为 -3 和 -5 的题目在实际难度上差异是很小的,但在数值上却相差 2,因此本研究采用了上述的策略。从表 2 可知,瑞文测验大部分题目都过于简单,不适合用于正常成人的智力评估。这可能和瑞文测验过多的曝光,以及人们智力水平的普遍提高有关。

许祖慰报告联合型瑞文测验所有题目的信息函数峰值没有一个超过 1.0,另外有 10 个题目的峰值小于 0.1[18,19],结合本研究可以推测图形类测验信息函数峰值不高可能是一个"通病",这类测验的测验效能低于一般的成就测验。若要从本研究建立的题库的 -2 ~ -1 区间中随机选择题目构成试卷的话,每份试卷平均约需 35 道题目,才能使测验信息函数的峰值达到 20 的水平,此时测量标准误约为 0.22。当然若将一些优质题目(测验信息函数峰值大于 1)做为每份试卷必考题目,则可将平均测验题数减少到 30 以下。在具体操作时可以此方案编制多个测验版本储存在计算机软件中,随机呈现给被试,这样更有利于控制测验的曝光。

参 考 文 献

1 Jones A. A case study in utility analysis. Guidance and Assessment Review, 1988,4:3 ~6

2 Tian Jianquan. The Reference of Project A for the Soldier's Selection and Placement System of the PLA. Advances in Psychological Science, 2006,14(2):164 ~168
(田建全. Project A 对我军士兵心理选拔研究的启示. 心理科学进展, 2006, 14(2):164 ~168)

3 Xiao Wei. The construction of the Intelligence Test System for Nationwide Conscription Based on Classic Test Theory and Item Response Theory(in Chinese). Xi'an: The Fourth Military Medical University, 2005. 6 ~7
(肖玮. 基于经典测验理论及项目反应理论的征兵用智力测验系统的研制. 博士学位论文. 西安:第四军医大学, 2005. 6 ~7)

4 Driskell J E, Olmstead B. Psychology and the military: Research applications and trends. American Psychologist, 1989, 44:43 ~54

5 Campbell J P. An Overview of the Army Selection and Classification Project(Project A). Personnel Psychology,1990, 43(2): 232 ~239

6 Ackerman P L, Heggestad E D. Intelligence, personality and interests: evidence for overlapping traits. Psychological Bulletin, 1997, 121(2): 219 ~245

7 Sternberg R J, Kaufman J C. Human Abilities. Annual Review of Psychology, 1998, 49: 479 ~502

8 Zhang Houcan, Wang Xiaoping. Standardization research on Raven's standard progressive matrices in China. Acta Psychological Sinica, 1989, 21(2): 113 ~121
(张厚粲,王晓平. 瑞文标准推理测验在我国的修订. 心理学报, 1989, 21(2): 113 ~12)

9 Sands W A, Waters B K, McBride J R. Adaptive testing: Inquiry to operation. Washington, DC: American Psychological Association, 1997. 121 ~143

10 Qi Shuqing, Dai Haiqi, Ding Shuliang. Principles of Modern Educational and Psychological Measurement (in Chinese). Beijing: Higher Educational Press, 2002. 142 ~149
(漆书青,戴海崎,丁树良. 现代教育与心理测量学原理. 北京:高等教育出版社,2002. 142 ~149)

11 Embretson S E, Reise S P. Item response Theory for Psychologists. Mahwah: Lawrence Erlbaum Associates, Inc, 2000. 249 ~273

12 Li Dan. The handbook of Combined Raven's Test in Chinese version(in Chinese). Shanghai: East China Normal University, 1989
(李丹. 瑞文测验联合型 (CRT) 中国修订版手册. 上海:华东师范大学,1989)

13 Liu Zhengkui, Shi Jiannong. A review of researches on inspection time and intelligence(in Chinese). Advances in Psychological Science, 2003, 11(5): 511 ~515
(刘正奎,施建农. 检测时与智力关系的研究述评. 心理科学进展,2003,11(5): 511 ~515)

14 Hambleton R K, Swaminathan H, Rogers H J. Fundamentals of Item response theory. London: SAGE Publications, 1991. 7 ~21

15 Mislevy R J, Stocking M L. A consumer's guide to LOGIST and BILOG. Applied psychological measurement, 1989, 13: 57 ~75

16 Xie Xiaoqing. Comparison of 15 equating methods. Acta Psychologica Sinica, 2000, 32(2): 217 ~223
(谢小庆. 对 15 种测验等值方法的比较研究. 心理学报,2000,32(2): 217 ~223)

17 Hau Kit - tai. Reliability and dimensionality: Scales with high alpha coefficients are not necessarily unidimensional (in Chinese). Education Journal, 1995,23(1):135 ~146
(侯杰泰. 信度与度向性:高 alpha 量表不一定是单度向. 教育学报,1995,23(1):135 ~146)

18 Xu Zuwei. The application of item response theory in test(in Chinese). Shanghai: East China Normal University Press,1992. 208 ~216
(许祖慰. 项目反应理论及其在测验中的应用. 上海:华东师范大学出版社,1992. 208 ~216)

19 Xiao Wei, Miao Danmin, Zhu Ningning. An Item Analysis of Combined Raven's Test on the Item Response Theory. Psychological Science, 2006 ,29 (2) : 389 ~391
(肖玮,苗丹民,朱宁宁. 应用项目反应理论对瑞文测验联合型的分析. 心理科学, 2006 ,29 (2) :389 ~391)

The Development of the Item Bank of Graphic Deductive Test Based on Item Response Theory

Xiao Wei[1], Miao Danmin[1], Zhu Ningning[2], Zhang Qinghua[2]
([1] *Department of Psychology, The Fourth Military Medical University, Xi'an* 710032, *China*)
([2] *college of Psychology, Beijing Normal University, Beijing* 100875, *China*)

Abstract

Introduction

With the application of high-tech weapons in the military arena and the changes in the pattern of warfare, the future high-tech local wars require much more of soldier's psychological qualifications. In order to improve the quality of the Chinese soldiers, it is vitally important and also necessary to add psychological measurement system to the physical examination of the enlisted men. The history of military personnel psychological selection shows constructing such system is historical development trend. The Psychological Selection System (Version 1.0) is soldier's qualification test. The objective of the test is to eliminate recruited young men with low intelligence. The content includes: Chinese Vocabulary Reasoning Test (CVRT), Number Operation Test (NOT) and Digital Search Test (DST). Version 2 needs to add nonverbal test to measure pattern-recognition and spatial reasoning.

Method

235 graphic deductive items imitating Combined Raven's Test (CRT) were developed and administered to 1,733 males with different education levels— junior high school, senior high school, technical secondary school, freshman and sophomore. Using Anchor-Test design, the participants were divided into 5 groups. The 235 items were divided into 4 tests. The CRT was used as test 5 while the 72 CRT items were distributed to the 4 tests as anchor items. The items were calibrated using BILOG-MG 3.0 (Marginal maximum likelihood estimation and three-parameter logistic model). The scale of test 5 (CRT) serves as the reference in the calibration. The items were then deleted if their data-model fitness were not good or the maximum information were less than 0.3. The Cronbach's alpha and information of each item were calculated for testing reliability. Correlation coefficients between ability of subjects and their scholastic performance were used as criterion-related validity.

Results

The item bank with 181 items were established with a maximum information between 0.30 and 1.13. Based on the location of maximum information, 181 items were divided into 4 groups: < -2.5 (32 items), $\leqslant -2.5$ and $\leqslant -1.5$ (79 items), $-1.5 <$ and $\leqslant 1.5$ (69 items), and > 1.5 (1 items). This means that the item bank can only be used to estimate the subjects whose ability are low, but it is not suitable for measuring high ability subjects. The cutoff scores for candidate of conscription was confirmed at -1.64 based on 5% elimination rate. The reliability and validity for low ability person are satisfactory.

Conclusions

Subjects' performance on graphic deductive test was affected by education level and by the characteristics of the item. Item contents were found to be difficult to remember and to describe, which suggests that they are good for personnel selection. The degree of difficulty of the items was affected mainly by the number of the principles being used in reasoning process. The item bank have satisfactory reliability and validity for individuals with low ability, rendering it fit for being used for elimination of low ability individuals in recruitment.

Key words development of item bank, item response theory, item equating, graphic deductive test.