SCHOLARS' FORUM

ACADEMIC TREND

PRACTICE & EXPLORING

RETROSPECTIONS

HOTSPOTS

FOREIGN TESTING

# 考试研究

KaoShi YANJIU

《考试研究》编辑部

●天津市教育招生考试院

2002 第二辑

天津人民出版社
TIANJINRENMINCHUBANSHE

# 考试研究

天津市教育招生考试院《考试研究》编辑部

天津人民出版社

2002年第二辑

**图书在版编目（CIP）数据**

考试研究．2002年．第2辑／天津教育招生考试院《考试研究》编辑部编．—天津：天津人民出版社，2002
ISBN 7-201-04310-2

Ⅰ．考… Ⅱ．天… Ⅲ．考试学—文集
Ⅳ．G424.74-53

中国版本图书馆CIP数据核字(2002)第087605号

天津人民出版社出版、发行
出版人：赵明东
（天津市张自忠路189号 邮政编码：300020）
邮购部电话：(022)27307107
网址：http://www.tjrm.com.cn
电子邮箱：tjrmchbs@public.tpt.tj.cn
天津美术印刷厂印刷
*
2002年12月第1版 2002年12月第1次印刷
787×1092毫米 16开本 10.75印张 2插页
字数：200千字 印数：1-2,000
定价：28.00元

《考试研究》 2002年 第二辑

## 理论研究



## 实践探索



## 热点争鸣

## 博士论坛



## 旧文新读



## 异域考试





## 学术动态

# Testing Research

2002 年 第二辑

## *Theory Research*



## *Practice & Exploration*



## *Hotspots*

## *Doctors' Forum*



## *Review*



## *Foreign Tests*



## *Academic Trend*

# 关于考试公平性的一些思考

谢小庆　王　洋

【摘　要】 考试公平性是评价考试质量的重要方面，也是一个受到广泛关注的问题。本文探讨了一些与考试公平性有关的问题。

【关键词】 测验　公平　偏见　题目功能差异

考试公平性是评价考试质量的重要方面，也是一个受到广泛关注的问题。但是，什么样的考试才是公平的考试？回答这个问题，并不容易。

**考试并不是天然公平的**

今天，在许多中国人的观念中，还认为考试是天然公平的。“考试面前人人平等”，“是英雄，是好汉，考场上，比比看。”如果因考试成绩不好而被拒绝，被考人心服口服，并无怨言。在一些西方人眼中却并非如此。虽然西方人运用考试的历史比我们晚整整一千年，却形成了一些更科学的考试观念，就好像他们虽然从我们这里学会了造纸术和印刷术，今天却在向我们出口造纸和印刷的成套设备。例如，如果一个美国人因某项考试成绩不佳而被拒之于校门、厂门之外，他可能会追究：这项考试是否可靠？因为这项考试而拒绝我的根据是什么？何以见得我在此项考试中成绩不高就说明我一定完不成学业、做不好工作？

今天，许多考试受到“高分低能”的批评。以一个“高分低能”的考试将一些人挡在校门、厂门之外，显然没有什么公平可言。

一些经过培训班强化辅导的学生在“托福”上考分很高，但实际英语

作者简介：谢小庆，博士，教授，北京语言文化大学汉语水平考试中心主任；王洋，北京语言文化大学人文学院，北京，100083。

交际能力并不高。这个问题已经引起美国教育测验服务中心（ETS）的注意，下决心对“托福”进行大的改革。在国内许多考试中也存在类似现象。在这种情况下，考试对于那些没有机会得到特殊辅导的考生，是没有太多公平可言的。

**组间差异与公平性**

既然考试并不是天然公平，我们就面临这样的问题：什么样的考试才是公平的？上世纪初，在比奈智力测验的研究过程中，人们就发现了不同人口群组之间的差异，发现母语不是英语的儿童会因语言上的障碍而影响其智商分数，发现经济地位较低阶层的儿童由于受到家庭的影响而在一些题目上得分较低。在西方，大规模的考试始于第一次世界大战。其间，大约 200 万人参加了军队甲种和乙种测验（The Army Alpha and Beta Tests）。那时，黑人与白人之间的组间差异是明显的，但并未引起太多关注，组间差异被认为是考生真实水平的反映。在那个时候，测验的公平性并不是关注的焦点。20 世纪 40 年代末，一些心理学家才开始考虑到一些考试表现中的差异并不一定反映实际能力，并开始通过控制那些低收入阶层的孩子们不熟悉的测试内容来提高考试的公平性。直到 60 年代中期，一些心理学家才感到考试偏见（bias）问题的严重性，才开始关注黑人学生与白人学生之间的考试分数差异，开始考虑这种组间差异可能对较低分数组造成的伤害。这一时期，许多测量专家开始关注测验和题目的公平性问题。Berk 曾经写道：“60 年代晚期和 70 年代早期心理测量学家们都一窝蜂地用具有客观标准的术语给偏见下定义，寻找研究偏见的细致精确的方法，对测验偏见进行实证调查。”[1](P370)美国 60 年代的民权运动和稍后的女权运动更强化了人们对公平性问题的兴趣。这一时期，在美国教育研究协会（American Educational Research Association, AERA）和美国国家教育测量协会（National Council on Measurement in Education, NCME）的年会中，公平性是一个重要的主题。

1968 年，Cleary 提出了一个关于考试公平性的定义：如果一项考试不会系统地高估或低估某一组人，这个考试就是公平的。[2](P371)1985 年版的《教育与心理测量标准》（Standards for Educational and Psychological Testing）认可了 Cleary 的这个定义。[3](P12)很长时间，人们根据 Cleary 的定义，认为像《学术评价测验（SAT）》这样的大学入学考试低估了黑人的能力，并因此认为这些考试存在偏见。

但是，更深入细致的研究发现，如果严格按照 Cleary 的定义来考察测

验分数与效度标准行为之间的关系，在许多情况下，SAT不仅没有低估黑人考生的实际能力，反而高估了黑人考生的实际能力。今天，教育测量学术界已经普遍改变了最初的看法，许多测量学家已经认识到，低估或高估不能成为考试偏见的一个定义。测量专家们几乎一致地认为有效的组间差异也可能是公平的。

**有效的考试就是公平的考试**

既然低估或高估不能成为测验偏见的定义，那么，怎么来定义测验偏见呢?1989年，Cole和Moss将测验偏见定义为"对不同可界定的被试子团体的效度差异"。[4](P205)这里，公平性问题就与考试效度问题联系在了一起。大学入学考试对白人黑人是否公平的问题，同时也是这项考试在预测白人和黑人考生的大学表现方面是否同样有效的问题。汉语水平考试对日韩考生和对欧美考生是否公平的问题，同时也是这项考试在测量日韩考生和欧美考生的汉语水平方面是否同样有效的问题。从此以后，公平和效度的关系成为公平性研究的核心问题。

关于"公平"，我们至少可以有两种选择：常模参照的公平和效率参照的公平。例如，4个白人和4个黑人通过考试申请4个职位。根据"常模公平"，可以按考试成绩取白人中的前两名和黑人中的前两名；根据"效率公平"，不管白人黑人，按成绩取前4名。有可能，被录取的4个人都是白人。

美国人早就选择了"效率公平"。在以往有关就业公平的文件的基础上，1978年，美国国会公平就业机会委员会、人事管理总署、劳动部、司法部、财政部等机构共同颁布了《雇员选拔的统一准则》，并于1980年颁布了对这一文件的补充说明。[5]在这两个文件中，明确选择了"效率公平"。在整个文件中，表达了"有效的考试就是公平的考试"的精神。如果在前述的招聘中经过考试录取的全部是白人，黑人全部落选，只要有证据说明考试确实是有效的，就不能认为存在种族歧视。但是，如果不能提供关于考试效度的证据，被拒绝的黑人就可以指控招聘过程中存在种族歧视。

这样，考试的使用者就需要提供有关考试的"效度证据"。否则，他就可能受到那些因考试成绩较低而遭到拒绝的求职者的指控。根据美国1964年确立的《人权法案》的精神，凭一纸靠不住的考试就剥夺一个人的学习或就业机会，是对人权的一种侵犯。如果被拒绝的是妇女或黑人，人们就可以控告考试的使用者歧视妇女和黑人。以这种眼光来看，由于缺乏足够的效度研究和效度证据，我们今天的许多考试(如高考、研究生考试、

招工考试、职业资格证书考试、职业技能鉴定考试等）的公平性都可能受到质疑。

**考试的主要功能不是维护社会公平**

将科学化考试用于人员选拔的逻辑依据主要是“效度资料”，即可以用事实说明：在这项考试中考高分的人的平均能力高于考低分的人，以考试选拔的学生比随机选拔的学生的平均能力更强，以考试选拔的职工可以比随机选拔的职工带来更高产量，以考试选拔的飞行员比不用考试选拔的飞行员具有更高的训练成功率。

今天的许多考试尚缺乏效度资料的支持，还受到“高分低能”的批评。即使积累了足够的效度证据，即使实现了“高分高能”，我们也仅仅可以说这个考试是“有效的”，也不足以说这个考试是“公平的”。在以一张试卷考查不同学生时，公平的前提是这些学生“以往所处的教育环境基本相同”。以同样的考试考查学习条件迥异的学生，是没有什么公平可言的。同样的一个高考分数，对于一个艰苦环境中的自学者和一个优越环境中接受特殊辅导的学生，具有很不同的意义。

从“卷子总比条子好，考官总比跑官好”的意义上讲，考试具有一定维护社会公平的功能。但是，这种功能是非常有限的。就主要功能讲，考试的意义在于提高效率，而不在于维护社会公平。



是否应该降低少数民族考生的分数线？是否应该对教育条件较差的落后地区考生予以照顾？这些关系到维护社会公正的问题，需要教育部门做出选择和决策。这些，属于考试以外的价值判断和政策选择，基本与考试本身无关。关于这个问题的深入探讨，可以参看参考文献[6](P16～P19)、[7](P120～P122)。

**建立考试分数之间的可比性是保证考试公平性重要任务**

我们以考试作为尺子对人的能力进行测量。公平的前提是不同尺子的测量结果之间的可比性。如果对同一个东西，这把尺子量得一个长度，那把尺子量得另一个长度，这样的测量就不能保证公平。

在常模参照考试中，分数的可比性不是很严重的问题。在标准参照考试中，分数的可比性就直接关系到考试的公平性。今天，社会上各种属于标准参照的考试很多，如自学考试、语言水平证书考试、会计师、经济师、律师的资格考试等等。每次考试的试卷不同，试卷难度也会有差异。水平较低的考生可能因碰巧遇到较容易的试卷而取得资格，水平较高的考生可能因碰巧遇到较难的试卷而得不到资格。这种情况，对考生是不公平的。

命题者总是希望不同的试卷具有相同的难度，但实际上不同试卷之间的难度差异很难完全避免，以数学方法对不同试卷之间的难度差异进行校正的过程即等值(equating)。“托福”、GRE 以及被称为“汉语托福”的中国汉语水平考试(HSK)等考试都经过了等值处理。但是，今天国内多数的资格考试和自学高考都没有经过等值处理，这种状况对于考生是不公平的。

**从“偏见(bias)”到 DIF**

在 2001 年中央机关的国家公务员考试中有这样一道题：

下列哪个城市不是经济特区？

A 深圳　B 厦门　C 广州　D 汕头（正确答案是广州）

在事后的试卷分析中发现，广州考区的多数考生都做出了正确回答，成都考区的许多考生却答错了。这道题目存在明显的组间差异，有利于广州考生，不利于成都考生。

研究发现，一些涉及武器、足球的题目存在男女之间的组间差异，有利于男性考生，不利于女性考生。在中国，一些涉及空调、微波炉、地下铁道等现代设施的题目存在城乡考生之间的组间差异，有利于城市考生而不利于农村考生。

人们曾经将这一类组间差异视为考试偏见，认为这些题目会造成对一部分考生的偏见。在关于克服偏见的研究中，人们提出许多有助于提高题目和测验公平性的统计方法，如 TID 方法、MH 方法、标准化方法、Sibtest 方法、IRT 方法等等。在这些统计法中，通常将测验总分作为“匹配变量”，即反映考生实际水平的指标，比较具有相同总分的不同群组的考生正确回答某一题目的比例。这些统计方法得到广泛应用，对 DIF 的检验已经成为许多考试研制中的常规步骤。上面那道关于经济特区的题目中所存在的问题，就是在公务员考试之后的常规 DIF 分析中发现的。

但是，根据这些统计方法并不能确定一道题目是否公平。如果我们认为具备有关经济特区的知识是胜任一个国家公务员必备的条件，就不能认为这道题目对成都考生存在偏见。又如，那些涉及计算机和互联网的题目，在一般的能力考试中，可能对来自落后农村的考生存在“偏见”。但是，在某些职位的选拔考试中，这些知识是胜任工作所必需的，根据“有效即公平”的原则，这些题目并不能被认为造成“偏见”，也不应被排除在这类选拔考试之外。再如，那些涉及武器的题目，在一般的学习能力考试中，可能会对女生造成“偏见”。但是，在有关安全的职位的选拔考试中，关于武

器的知识是必需的，这些题目并不能被认为造成“偏见”，也不应被排除在这类选拔考试之外。

20世纪80年代中期以后，人们开始回避具有价值判断性质的“偏见”概念，代之以价值中性的“题目功能差异（differential item functioning, DIF）”概念。DIF仅仅是题目的一种统计特征。具有DIF的题目是否造成偏见？还需要专家根据测验目的和测验内容做出经验的判断。经过统计分析，我们可以认定上面那道关于“经济特区”的题目存在DIF。存在DIF的题目是否造成偏见？还需要专家做出经验的判断。

**考试公平性的问题尚未解决**

虽然人们越来越多地将公平性和效度联系在一起，并不意味着人们对公平性的看法已经取得一致。在美国，许多考试机构和考试的应用机构，在考试分数的解释方面，都强调对少数族裔和低收入阶层的照顾。在我国，多年来在考试中一直坚持着照顾少数民族的政策。这些，都是对“效度公平”观念的校正和调整。

在1999年新版的《教育与心理测验标准》中写道：“不论是就整个社会而言，还是就测量专业的学术界而言，近期都还看不到人们在测验公平问题上取得一致意见的前景。……公平性概念可以从多种角度来定义，公平并不完全是一个技术概念，关于公平的定义和解释随不同的社会和政治环境而变化。……需要再一次强调，本标准仅仅从技术角度提供了一些专门的指导，对测验负责任的使用，还需要有关价值和社会政策方面的考虑。”[8](P80)

2000年卸任的ETS前总裁Cole和ETS命题负责人Zicky于2001年发表在《教育测量杂志》的一篇题为“公平性的新面孔”的文章中写道：“现阶段关于公平性的研究尚不能对任何测验公平问题做出简单的回答。60年代涌现出的对公平问题的研究最终是令人失望的。没有一种普遍接受的方法可以决定一项测验是否公平，没有一种统计方法可以清晰明确地证明一个题目是否公平，也没有一种技术上的解决之道。简而言之，过去的三十多年没有研究出任何分析方法可以表明公平或不公平，也没有一种清楚的程序可以避免不公平。”[9](P375)

## 参考文献


[1] Cole, Nancy S. & Zieky, Michael J. (2001) *The new faces of fairness*, Journal of Educational Measurement, 38 - 4, 369 - 382.

[2]同[1]。

[3] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985) *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

[4]Cole, N. S. & Moss(1989) *Bias in test use, in Linn ed. Educational measurement*, Macmillan Publishing. 201 - 219.

[5]EEOC(1978)UNIFORM EMPLOYEE SELECTION GUIDELINES, http://www.uniformguidelines.com/uniguideprint.html#1

[6]谢小庆等,《洞察人生》,山东:山东教育出版社,1992 年。

[7]谢小庆,《考试观念的变革》,《开放时代》,1999 年第 126 期。

[8] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999) *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

[9]同[1]。




# Insight Into Test Fairness


Xie Xiaoqing, HSK Testing Center, Beijing Language University

Wang Yang, Humanity College, Beijing Language University 100083



**【Abstract】** Fairness is one of the most important features of a test. Widely concern is concentrating on the problem of test fairness. This paper discussed some topics related to test fairness.



**【Key word】** test, fairness, bias, differential item functioning (DIF)

# 全息项目因素分析简介

## ——一种新的多维、多级评分的 IRT

王　权

【摘　要】 本文介绍 Bock 和 Muraki 等学者创立的“全息项目因素分析”。这种新颖的项目分析方法是因素分析与项目反应理论的有机结合，也是一种多级评分的多维项目反应理论。全文分三个部分：(1) 多维、多级评分项目的全息项目因素分析的数学模型，以及模型的几何解释；(2) 多维 IRT 模型的参数解释；(3)项目参数的极大似然估计和能力参数的期望后验估计。

【关键词】 全息项目因素分析　多维、多级评分项目的全息项目反应理论极大似然估计　期望后验估计



分析鉴定测验的测量维度是研制各种标准化教育测验和心理量表的不可缺少的一个重要环节，过去多用因素分析方法验证。但不论是探索性因素分析还是实证性因素分析，方法的实质是属于“协方差分析”，分析的基础是观测样本的协方差阵或相关矩阵，而不是从原始观测分开始。所以当观测变量是二分变量时，我们面对的是 Φ 系数矩阵或是四项相关系数矩阵。对于 Φ 系数矩阵，当项目难度不一致时，一般的多因素分析就会混入虚假因素。而对于四项相关系数矩阵，虽然可以抑制虚假因素的混入，但由于四项相关系数矩阵几乎永远不正定，所以严格来讲不能应用公共因素模型。而且，若测验中的某对项目的四格表中出现零观测次数，四项相关系数矩阵中的相应元素的绝对值就会出现 1，因而由此就会发生“海伍德”(Heywood) 现象，即分析结果出现负方差或绝对值大于 1 的相关系

作者简介：王权，教授，浙江大学教育系，浙江杭州，310002。

数的不合理现象。基于以上原因，Bock 和 Aitkin(1981 年)提出了一种直接与"项目反应理论"(IRT)相结合的新的项目因素分析法。这种方法直接使用每个被试对全部项目的反应向量，最大限度地利用反应数据载荷的可用信息，所以称作"全息项目因素分析"(Full - Information item factor analysis)。

Bock 和 Aitkin 巧妙地借用 Thurston 的多因素公共因素模型来描述被试对项目刺激的反应过程，即公共因素模型：

$$Y_{ij} = \alpha_{i1}\theta_{1j} + \alpha_{i2}\theta_{2j} + \cdots + \alpha_{im}\theta_{mj} + \varepsilon_{ij} = a'_i\theta_j + \varepsilon_{ij}$$

$$i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, N, \qquad (1)$$

式中的因变量 $Y_{ij}$ 并非表示连续型的观测变量，而是表示被试 $j$ 对项目 $i$ 的一个不可观测的反映过程变量，当 $Y_{ij} \geq \gamma_i$ 时，被试 $j$ 对项目 $i$ 产生正确反应，记作 $Y_{ij} = 1$；当 $Y_{ij} < \gamma_i$ 时，则反应错误，记作 $Y_{ij} = 0$。$\gamma_i$ 称作项目 $i$ 的阈限(Threshold)。按因素分析惯例，(1)式假设 $Y_i$ 的数学期望为 0，方差为 1；$\theta \sim N(0,1)$；$\varepsilon \sim N(0,\sigma^2)$。所以被试 $j$ 对项目 $i$ 产生正确反应的条件概率是

$$P(Y_{ij} = 1 \mid \theta_j) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{Z_i(\theta_j)}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt = \Phi_i(\theta_j) \qquad (2)$$

其中 $Z_i(\theta_j) = \frac{\gamma_i - a_i\theta_j}{\sigma_i}$。于是反应错误的概率 $P(Y_{ij} = 0 \mid \theta_j) = 1 - \Phi(\theta_j)$。

对于有猜测因素的多重选择题，反应函数(2)式可取

$$\Phi_i^*(\theta_j) = g_i + (1 - g_i)\Phi_i(\theta_j)$$

式中的 $g_i$ 是项目 $i$ 的猜测参数。在此基础上，Bock 和 Aitkin 还建立和导出了参数估计方法等一系列重要结果。1995 年 Muraki 和 Carlson 在 Samejma 等人的多级反应的逻辑斯蒂模型的基础上，将 Bock 和 Aitkin 的二分变量的全息项目因素分析推广到类别变量的多级评分项目，形成了一种新型的"项目反应理论"(IRT)，称作类别变量的全息项目因素分析。

**一、数学模型**

测验项目为二分变量时，Bock 和 Aitkin 由公共因素模型(1)式导出了二级反应项目的项目反应函数(2)式。当测验项目是类别变量时，被试的项目反应须用多级评分表示。设被试 $j$ 对 $n$ 个项目的反应模式是

$$W_j = (w_{1j}, w_{2j}, \cdots w_{nj})$$

其中当被试 $j$ 对项目 $i$ 的反应变量 $Y_{ij}$ 超过阈限 $\gamma_{i,k-1}$、但没有达到 $\gamma_{i,k}$ 时，即 $\gamma_{i,k-1} \leq Y_{ij} < \gamma_{i,k}$ 则记 $w_{ij} = k \quad k = 1, 2, \cdots, K_i$。于是根据公共因素模型(1)式，被试 $j$ 对项目 $i$ 产生第 $k$ 级反应的条件概率是