

谢小庆 张晋军 / 主编

考试研究文集

(第3辑)



经济科学出版社

考试研究文集

(第3辑)

谢小庆 张晋军 主编

经济科学出版社

责任编辑：卢元孝 柳 敏

责任校对：徐领弟

版式设计：代小卫

技术编辑：潘泽新

考试研究文集（第3辑）

谢小庆 张晋军 主编

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲28号 邮编：100036

总编室电话：88191217 发行部电话：88191540

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

富达印刷厂印刷

华丰装订厂装订

880×1230 32开 10.25印张 270000字

2006年6月第一版 2006年6月第一次印刷

ISBN 7-5058-5662-6/F·4921 定价：22.00元

（图书出现印装问题，本社负责调换）

（版权所有 翻印必究）

前 言

就人均占有量而言，中国不仅算不上自然资源大国，甚至可以算是一个自然资源贫国。但是，中国毫无疑问是一个人力资源大国。中国未来的持续发展，很大程度上要依靠开发自己的人力资源，而不是开发自己的自然资源。怎样将“人口负担”转变为“人力资源”，这将是关系到中国命运的大问题。

在人力资源的开发中，考试占据着非常重要的地位。今天，中国对人类在科技和文化艺术方面的贡献，与得到世界公认的中国人的“聪明与勤劳”并不相称。虽然导致这种局面的原因很多，但不能说与“学以应考、教以应考”的“应试教育”没有关系。这种“应试教育”，从童年起就挫伤了中国儿童的好奇心和创造力，影响到从基础教育到高等教育的整个国民教育体系，造成学生为应试而学、教师为应试而教的局面，造成普遍存在的“学一本书、教一本书、背一本书、考一本书”的局面。这种局面，严重地影响到我国的人力资源开发，严重地摧残着中华民族的自主创新能力。

2005年5月在巴黎进行的中欧贸易谈判中，商务部部长薄熙来曾说，“为了买回一架空中客车A380，中国需要

出口 8 亿件衬衣”。这句话，突出地体现了中国经济发展中缺乏自主创新的问题。今天，“自主创新”已经成为朝野共同关注的核心话题之一。

为了促进我国人力资源的开发，需要对考试改革问题给予更多的关注。好的考试，促使学生去读厚厚的一大摞书，不好的考试，逼迫学生去背薄薄的一小本复习资料。好的考试，使学生“爱学”，不好的考试，使学生“厌学”。好的考试，可以有助于将“人口负担”转化为“人力资源”，不好的考试，可以将“人力资源”转化成“人口负担”。本来，能够进入大学的人是我们民族最优秀的一批青年人。但是，据教育部公布的大学毕业生就业情况，高达 30% 的大学生未能在毕业当年找到工作，最优质的“人力资源”经过“应试教育”被转化成“人口负担”，实在是一件让人沮丧和伤心的事情。

考试研究工作者们承担着推动中国考试改革的使命。在这本文集中，汇集了部分考试研究工作者的研究成果，内容涉及到一些考试工作者共同关注的问题，包括关于“考试公平”的深入思考，关于“考试应该体现谁的意志”的思考，关于测验分数解释问题的探讨，标准参照考试中及格线的确定问题，难度与区分度的关系问题，测验长度的确定问题，作文的电子评分问题，测验分数等值问题，档案式评价问题，等等。还有一些文章探讨了语言测试方面的问题，如语言测试中的语法点控制问题，语言测试中的语段测试问题，写作能力测验的方式问题，口语的测试方式问题，词汇和汉字等级大纲编写中的问题，等等。此外，还有一些文章涉及到 HSK、MHK 和 ZHC 等语

言测验的改进问题。

我们希望，我们的这些初步的探索能够得到考试开发研究同行们的指教，也希望这些思考能对同行们有所启发。我们希望，这些思考和探索能有助于我国考试的科学化改革，能通过考试的科学化改革来开发我国丰富巨大的人力资源，能通过考试改革使我国儿童的好奇心得到保护，能通过考试改革来促进我国经济、文化方面自主创新能力的增长。我们希望，经过我们共同的努力，中国出口的产品将不仅是衬衣、鞋袜和玩具，而是包含更多高知识含量的产品。我们希望，经过我们的共同努力，中国能成长起一批像英特尔、微软、松下、索尼、夏普等一样不是靠“勤劳和汗水”、而是靠“聪明和智慧”赚钱的企业。



◎ 概 论 ◎

- “考试公平”的三种不同含义 谢小庆 (1)
- 考试应该体现谁的意志 谢小庆 (7)
- HSK 的分数解释问题 谢小庆 (15)
- 基于档案评价的汉语口语水平考试 彭恒利 (27)
- 语法在第二语言教学中的作用 胡友珍 衣 莉 (40)
- 难度与区分度关系之讨论 马新芳 (58)
- 小议母语、第一语言和第二语言 马新芳 伯 冰 (65)
- 关于测验长度的研究 马新芳 (71)
- 谈汉语水平考试中的语段测试问题 张宝林 (84)
- 关于《汉语水平词汇和汉字等级大纲》的思考 姜德梧 (99)

◎ 实 证 研 究 ◎

关于作为第二语言的汉语写作能力

- 测验方式的实验研究 赵 亮 谢小庆 (117)
- 对 HSK 部分等级的验证性研究 武晓宇 徐 静 (134)

▲ 考试研究文集 (第3辑) ▲

HSK 作为密歇根中文系水平考试的适用性、

可行性分析报告 刘 威 (149)

Tucker 线性等值法中内铷与外铷两种等值

参数估计结果的比较 任 杰 (160)

汉语测试电子评分员实验研究报告 张晋军 任 杰 (171)

◎ MHK ◎

MHK 一、二级命题中的语法点控制研究 李桂梅 (187)

MHK (四级) “听后写” 题型设计理念及实测验证 李桂梅 (199)

MHK (三级) 口试个案分析 鲁新民 (209)

吉林省 2004 年 MHK (三级) 作文评分简述 鲁新民 (221)

◎ ZHC ◎

细说 ZHC 谢小庆 (233)

ZHC 口试实施研究方案 谢小庆 张晋军 (253)

ZHC 改进设想 张晋军 (272)

ZHC 拼卷工作介绍 张晋军 (292)

对一次作文考卷的调查 姜德梧 (302)

◎ 文献综述 ◎

第六届海峡两岸心理与教育测验学术研讨会述评 谢小庆 (307)

“考试公平”的三种不同含义

谢小庆

摘要：本文讨论了对“考试公平”的不同理解，提出了程序公平、条件公平和事实公平三种不同的“考试公平”概念，讨论了考试公平与效率之间的关系，讨论了考试工作者的责任。

关键词：考试 公平性 效率

在谈论“公平”的时候，人们赋予这一概念的含义是不同的。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。”（American Educational Research Association，第80页）显然，对于什么是“公平”，还需要

做出进一步的分析。

第一考试公平：程序公平

“公平”的第一种用法是指“程序公平”，即对所有参加考试的应考者都一视同仁，使所有考生得到相同的对待。这种用法体现为1968年Cleary提出的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（参见Cole & Zieky，第371页）。这种用法也体现为1989年Cole和Moss将考试偏见定义为“对不同可界定的被试子团体的效度差异”（Cole & Moss，第205页）。我们可以将此称为“第一考试公平”。

为了实现“程序公平”，今天还有许多事情要做。在北京单独命题之前，高考总分550分的北京考生可以上清华、北大，在同样试卷上得到同样分数的山东、湖北考生可能还不能进重点院校。考生受到如此不同的区别对待，显然损害了“程序公平”。

据2002年7月22日《南方日报》一篇题为“高考作文两次改判背后的幸与不幸”的文章披露，在2002年四川省高考语文阅卷工作中，一篇作文因“与传统观点格格不入”，评卷老师只给20分。“将此文提出来与整个阅卷室的数十位评卷教师讨论，结果争议很大，从20多分改成了40多分，最后请示评卷指导委员会，最终被认为是一篇不可多得的佳作，给出了相当高的分数。”（满分60）控制主观题的评分误差，是保证“程序公平”的一个方面。

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，尽管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。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。

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,今天国内的许多资格考试和自学高考尚没有经过等值处理,这种状况也违背了“程序公平”。

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。通过“题目功能差异(differential item functioning, DIF)”分析,可以克服这种由试题DIF导致的不公平。

第二考试公平:条件公平

“公平”的第二种用法是指“条件公平”,即不仅追求所有参加考试的应考者在考试中受到相同的对待,而且要求参加考试的考生在教育条件方面也能受到相同的对待。就这种意义上的“公平”而言,以同样的考试考查学习条件迥异的学生,并不能算公平。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。我们可以将此称为“第二考试公平”。

今天,不论中外,都在考试中采用了降低少数民族考生的分数线、照顾教育条件落后地区的政策。这些政策,体现了对“第二考试公平”的追求。今天高考中存在的分数线地区差异,有些可能在一定程度上体现了对这种“条件公平”的追求,如青海、新疆等教育条件较差地区的录取分数线较低。有些却不是,如北京等教育条件较好的地区的录取分数线反而较低。

第三考试公平：事实公平

“公平”的第三种用法是指“事实公平”。即不仅追求所有参加考试的应考者在考试中受到相同的对待，不仅追求参加考试的考生在以往教育条件方面也受到相同的对待，而且追求所有具有“天赋人权”的儿童都具有平等分享优质教育资源的权利。优质教育资源有限，通过考试把有限的优质教育资源提供给强者而不是弱者，使强者更强，弱者更弱，自然会扩大社会的不平等，会加速社会的两极分化。我们可以将“事实公平”称为“第三考试公平”。

今天，在小学、初中入学中所采用的“就近入学、电脑派位”方式，就是对这种“事实公平”的追求。

考试公平与效率

通过考试，学校、机关、企业可以选拔到优秀的人才，可以拒绝不称职的庸才。显然，考试可以带来效率。就“靠分儿总比靠钱好，卷子总比条子好，考官总比跑官好”而言，考试也可以抑制权力和金钱的影响，可以带来社会的公正。但是，三种不同的“考试公平”概念，与效率之间具有并不相同的关系。

“第一考试公平”与效率是基本统一的。只有坚持“程序公平”，考试才能发挥“优胜劣汰”的作用，才能成为提高工作效率的一个有效工具。在主要关注考生今天已有水平的标准参照性考试中，“程序公平”与效率具有更高的一致性。在关注考生未来发展潜力的预测性考试中，严格的“程序公平”可能会影响到考试的效度，从而影响到人力资源的使用效率。在招生、招工中，如果简单地坚持“一视同仁”，不考虑不同考生的成长环境和教育条件，未必能够达

到最好的人才选拔效果。有时候，为了发现那些真正具有潜力的人才，我们不仅要看他今天已经走到了哪里，还需要看他是从哪里出发，看他走过了什么样的路程，看他已经走过了多远的路程。

“第二考试公平”与效率之间具有很高的统一性，尤其是在预测性考试中。不仅看一个考生今天已经掌握了哪些知识和技能，而且看他是在什么样的条件下、在什么样的老师指导下、通过什么方式达到了今天的水平。这种选拔人才的方式，可能会比单纯看考试分数的方式具有更高的效率。在美国的大学招生中，学校常常会录取 490 分的学生而拒绝 500 分的学生，很少听说 500 分的学生会为此去诉讼校方。

“第三考试公平”的实现通常是以牺牲效率为代价的。通常，将优质教育资源提供给强者而不是弱者，会产生更大的效率。对“事实公平”的追求，势必会影响到社会的经济发展速度。

通过以上分析可以看出，虽然在一定条件下考试具有维护公平的意义，但考试的主要功能并不是维护社会公平，而是提高工作效率。

考试工作者的责任

作为一个考试工作者，首先需要追求“第一考试公平”。今天，不同省份之间高考分数线还存在很大差异，考试中有时还存在较大的主观评分误差，许多考试尚未实现等值，试卷中还时常会有一些存在 DIF 的题目，等等。因此，实现“第一考试公平”尚任重道远。

考试效度是考试工作者的职业追求目标。为了使考试具有更高的效度，考试工作者也需要对“条件公平”予以适当的关注，需要注意对考试分数做出更合理的解释，促进社会更合理地使用考试分数。

考试工作者也是人，也是知识分子，也是学者。作为一个人、

一个知识分子、一个学者,他还需要关注社会上的弱势群体,还需要致力于避免过度的社会两极分化。我个人认为,一个考试工作者也需要考虑到“条件公平”和“事实公平”的问题,他应该支持那些向少数民族、落后地区倾斜的考试政策,他也应该支持那些旨在追求“事实公平”的“就近入学、电脑派位”入学政策。

参考文献

- [1] 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
- [2] Cole, N. S. & Moss (1989), Bias in test use, in Linn ed. Educational measurement, Macmillan Publishing, 201 - 219
- [3] Cole, N. S. & Zieky, M. J. (2001), The new faces of fairness, Journal of Educational Measurement, 38 - 4, 369 - 382
- [4] 谢小庆、王洋 (2002): 关于考试公平性的一些思考, 考试研究 (2002 - 2), 天津人民出版社

Three Different Implication of Test Fairness

Abstract: This paper discusses different understandings of test fairness, and proposes the concepts of procedure fairness, conditional fairness and entity fairness. It also discusses the relationship between fairness and efficiency, as well as the responsibility of a professional testing practitioner.

Key word: test fairness, efficiency

考试应该体现谁的意志

谢小庆

摘要：许多人谈到：考试要体现国家意志。本文认为，考试所要体现的不应是国家意志，而是用户意志。本文认为，行政权力的介入阻碍了考试的发展，行政权力应该退出考试研究领域，行政权力要让位于用户权力。

关键词：考试 行政权力

一、考试应该体现谁的意志

考试应该体现谁的意志？原教育部考试中心副主任马金科同志认为：“高考是国家考试，必须进行改革，其改革走势应体现国家意志，这也是不争的事实。”（马金科，2001）全国高教自学考试办公室综合处刘粤平处长认为：“高等教育自学考试……作为国家考试制度，其代表着国家意志，行使学历认定和检验职能，具有高度的权威性和严肃性。这一点已为社会所广泛认同。”（刘粤平，2000）湖北省教育厅厅长路钢同志认为：“教育统一考试是国家意志的体现，是政府行为。”（路钢，2002）华中师范大学廖平胜教授认为：“考试制度属政治制度范畴，凡成文法规性考试的创建与实施，都是国家意志的体现。”他认为考试的行政功能突出表现为“强化国家意志，凝聚国民心向，促进政局稳定”（廖平胜，2002）。

这些同志，都认为考试是国家意志的体现。但是，原教育部高校学生司司长瞿振元同志提出了不同的看法。他在2001年全国招

生工作总结研讨会上讲：“在传统计划经济下……高校招生权是国家专属权力；不承认高校的独立利益，高校没有自主招生权。但是，随着社会主义市场经济的逐步建立和民主法制建设的进步，随着高等教育规模的日益扩大。招生考试权的性质已经并将继续发生变化……现在的招生考试制度是国家教育考试制度的最主要的种类，但它不再是国家行政性质的考试，而是国家统一组织的社会性质的考试。所以，高等学校招生考试权已经不是国家专属权力，而是一个社会权利体系。”（瞿振元，2001）

我非常赞成瞿振元同志的看法。考试是测量人的生理、心理属性的工具。作为一个工具，它只能是工具的使用者意志的体现。当这个工具被国家机关使用时（如国家公务员考试），它可以是国家机关意志的体现。当这个工具被其他人使用时，它就应该其他人意志的体现。

考试应该体现谁的意志呢？我认为，考试只能体现使用者或用户的意志。用户可以是国家机关，也可以是学校和企业。例如，“托福”、学术评价测验（SAT）、研究生水平考试（GRE）等考试的用户主要是美国的数百所大学，这些考试所体现的只能是这些用户（大学）的意志。对于编制这些考试的美国教育测验服务中心（ETS）来说，这些大学就是它的“上帝”。ETS 必须小心翼翼地讨这些大学用户的欢心。这些考试的生命力维系在这些用户的认可之上。如果这些考试不能体现这些用户的意志，ETS 就会垮台。因此，ETS 不断地根据这些用户的需要而改进着自己的考试。1982年，SAT 用基于现代项目反应理论的新的等值方法取代了原有的线性等值方法。1993 年推出了“无纸化”的计算机 GRE 考试。1994 年推出自适应性 GRE 考试。1998 年在美国本土取消了纸笔的 GRE 考试。1998 年，在全世界大部分地区开始计算机化的“托福”考试。2002 年，在 GRE 中增加了写作考试。2005 年，将推出全新的“新托福”。

需要指出，考生并不是考试的用户，考试也不必体现考生的意

志。多数情况下考生参加考试都是被动的或被迫的。

二、国家意志的介入阻碍了考试的改革

早在 20 世纪 80 年代初期,“片面追求升学率”问题就已经引起方方面面的高度重视,《教育研究》曾经长期开辟专栏进行讨论。20 年过去了,为什么我国教育界“学以应试、教以应试”的局面基本没有改变呢?

一种意见认为,不能把“应试教育”的问题都归咎于考试。我不同意这种说法。正是由于考试所考查的不是“素质”,才会有“素质”与“应试”的矛盾。如果一旦考试可以考查“素质”,素质教育与应试教育就得到了统一。高考改革已经进行了 20 年,“应试教育”问题仍然严重困扰着我国的教育,这种状况与高考改革一直坚持稳妥、保守的思路有很大关系。^①

“国家意志”和行政权力介入过多是导致这种保守思路的重要原因。正是由于政府权力的过多介入,才使考试改革缺乏动力,才使考试质量缺乏有效的控制,才使考试缺乏优化机制。

由于存在竞争关系,“乐凯”、“柯达”和“富士”面对消费者就必须兢兢业业;由于“联通”的出现,“中国电信”的服务质量大大提高。竞争导致优化,竞争带来质量,竞争形成制约。今天我国的考试机构缺乏必要的改革动力和优化机制。考试的科学化水平高低,考试质量的好坏,考试改革进度的快慢,与考试主管机构和主管人员的切身利益没有多大关系。

ETS 主持着相当于我国高考的 SAT。在大学入学考试这一领域中,ETS 面临着另一家考试机构美国大学考试中心 (ACT) 的强烈

^① 参见《湖北招生考试》2002 年 8 月号 (下) 15 页拙文中关于这种保守思路的分析。