北京大学
中国教育财政科学研究所

简　　报

怀仁怀朴　唯真唯实　　2011 年第 1 期（总第 50 期）

从当代教育测量学角度看我国高考研究

黄晓婷

（2011 年 1 月 19 日）

一、研究背景

高考是我国目前最重要的、对教育和社会具有最广泛和深远影响的考试。因此，关于高考的内容、形式和社会职能等方面的研究与评价，受到了教育研究者和社会各界的广泛关注。

从教育测量学的角度来说，对任何一个考试的研究，都不外乎两个方面：（1）该考试是否为其使用者提供了公平有效的受试个体之间的微观比较信息；（2）该考试是否能提供公平有效的跨时段或跨区域的教育质量宏观比较信息。评估微观比较信息的可靠性主要是通过研究单个考试的信度和效度来进行的，而宏观比较信息的科学性则取决于该考试与其他不同时间或区域内的考试间是否搭建了有效链接。

在过去十年中，我国学者运用传统测量理论（Classical Test Theory，简称

CTT）对高考的信度和效度进行了一些研究。但是，从上世纪六、七十年代起，以试题反应理论（Item Response Theory，简称 IRT)为核心内容的当代测量学已逐渐取代传统理论而成为国际测量学领域的主要研究方法。也是由于受到 CTT 的技术局限，我国在宏观信息方面还未能充分利用高考珍贵的数据资源。

随着高考改革的不断深入，如何全面提高高考各方面质量的研究必将成为一个重要的研究领域。本文以当代测量理论为视角，简述我国高考在微观和宏观比较信息质量方面的实证研究现状，并探讨未来发展的可能。

二、高考微观信息的质量研究

高考最重要的职能是为我国高等院校招生提供高质量的鉴别考生间知识和能力差异的信息。因此，高考成绩的信度和效度直接影响了我国高校招生的科学性和公平性。

1. 高考的信度

所谓信度（reliability），即指考试成绩的“可信”程度。我们常常会说，一次考试未必能反映一个人的真实能力。在考试中，难免会因为这样那样的原因，造成考分和真实能力间出现差异，这种差异被称为测量误差。误差越小，考试的信度也就越高。一般来说，考题的数量越多，考试的信度越高。这好比我们用不同的秤去称同一个物体，秤的次数越多，求平均之后得到的数值越接近该物体的实际重量。当考试受时间限制、题目数量一定时，试题的难度与受试者水平越匹配，则信度越高。这就好比物体的重量必须在我们所用的秤的承重范围内，才能得到精准的估值。假设某考试的信度不够高，100 分和 90 分之间的差异就可能仅由测量误差造成，而只有考 100 分以上的学生才能被录取，那就有悖高校招生的科学性和公平性原则了。

因此，越重要的考试，对信度的要求就越高。国外绝大多数考试在每次试测和实测之后，都会运用 IRT 中的统计模型计算考生成绩的信度指数。例如美国的 SAT 和 ACT，在每次考试结束后，这两个考试的组织机构都会向社会公布其信度指数，接受监督。

我国对高考信度的研究，主要通过 CTT 中的方法，计算某些地方部分考试

科目的难度、区分度和信度指数等[1]。但是 CTT 中的计算方法存在样本依赖的局限性，换言之，如果抽样另一批考生，指数就有可能发生偏移。因此，结论的可推广性有限。另外，由于信度研究的文献数量十分有限，我们对历年来全国各地高考的信度状况仍然缺乏全面、深入的了解。

信度研究中的另一重要方向是针对如何控制主观题判分的误差。我国目前已实施了后台动态监控的做法，包括测试跟踪法（比较阅卷人和专家对标准卷的打分误差）和统计分析法（计算每位阅卷人每天评分的统计分布是否与其整体分布类似）[2]。这两种方法前者并不直接分析实际判分的状况，而后者则依赖每位阅卷人每天评阅的试卷都是整体的随机抽样的假设，并且只提供粗略的整体估算。最后，合格阅卷人的标准由各省自定，但以省为单位的判分信度分析却很难检索到。

国外现在常用的主观题判分信度分析，包括计算两个独立阅卷人对同一试卷的判分相同率（每个主观题都要求有两位阅卷人独立判分），以及在 IRT 理论中通过统计模型估算具体到每个受试人的阅卷人影响的大小，甚至相应调整最后的考生能力评价结果[3]。这些做法都有值得我们学习借鉴的地方。

2. 高考的效度

如果说信度赋予一个考试生命，那么效度则赋予一个考试思想。效度的定义在测量学界曾有过激烈的争论，现在被广泛认可的是由 Samuel Messick 提出的广义概念，即通过理论和实证依据，对考试结果使用的适当性做出的评价。在这个定义下，效度研究不是单一的一个问题，而是围绕命题和考分使用后果这两个主题的“问题群”。

具体说来，围绕命题的研究，主要包括四个方面：（1）考题的内容覆盖与权重是否符合考核目的的需要；（2）考题对不同背景的学生是否会造成不公（如性别、地域等）；（3）考生答题的思维过程是否符合考核设计（如能否通过猜测答对等）；（4）考题与同科目其他考试的相似性与区别。

围绕考分使用后果的研究，主要有两个方面：（1）基于该考试所作分流决定是否成功（如：高考考分是否准确预测了学生在大学阶段的学习情况等）；（2）考试结果反馈到课程设计、课堂教学、教育政策制定等，造成的一系列社会效应是否符合考试设计的初衷。

美国 SAT 和 ACT 的组织机构每年都会发布大量的效度研究报告，显示其科学性、有效性和公平性。在命题内容方面，这两个考试都有很详细的、基于认知理论和教学要求的知识点比重和难度的规划，最后实测卷的内容严格对应规划。考题的公平性常常通过 IRT 理论中的试题反应差异计算（Differential Item Functioning，简称 DIF）来检验。对在试测中发现有公平性问题的题目会进行修改或删除。此外，有很多研究通过计算 SAT 或 ACT 成绩与该组学生在大学阶段成绩（GPA）的相关性，来检验这两个考试的预测效度[4]。但是，尽管有大量的研究都为 SAT 和 ACT 的效度提供了正面的佐证，仍有很多学者提出质疑，认为这两个考试未能很好地预测学生在大学阶段的成功与否。还有学者认为，这些考试对少数民族和低收入家庭的学生进入大学造成了更大阻力，造成了负面的社会影响[5][6]。

我国关于高考命题方面的研究主要有两类：一类是以考试大纲为中心的，对内容、权重、难度等方面的提纲挈领式的理论规划；另一类是对实测的知识点、难度、形式等情况的汇总分析[7]。对考题公平性、命题逻辑严谨性以及考试与教学的关系等方面的研究则相对稀少。

关于高考成绩使用方面的实证研究，我国的研究者们基本着眼于分析高考成绩与大学学业成绩的相关性[8]。但是不同的研究常常得出不一致的结果。例如，有研究发现高考数学成绩与大学学习成绩成正相关，而也有研究报告了截然相反的结论。这可能是由于研究中使用的样本不是随机的、样本间差异较大，也有可能是因为各省份、不同年份间考试的内容和形式都存在较大差异而导致的。

最后，虽然我们针对高考社会效应的讨论非常热烈，但基于实证的学术研究却为数不多。因此，在深入加强高考效度研究时，我们可以参考国外对效度的多角度定义，运用更科学严谨的当代测量学工具，争取从更多方面完善高考制度。

三、高考宏观信息的研究

在保障单次高考质量的基础上，我们还要进一步利用高考的数据资源，为教育研究者和政策制定者提供地区间教育质量比较和逐年教育质量变化等参考信息。这需要运用当代教育测量学 IRT 中的等值（equating）和链接（linking）技术。所谓等值，顾名思义，就是找到两个独立考试间（同科目）难度值相等的点。

所谓链接，简单说就是找到两个考试之间试题难度的关系。打个比方，等值就好像对齐了两把尺的零刻度，链接好比找到了两把尺单位刻度间的换算关系。通过这两个方法，我们就可以把两个考试中的题目用同一难度单位来表述，进而科学地比较两次考试的结果。

1. 跨时间比较

目前世界上很多大型国际化的考试，如 PISA、TIMMS 等,都向使用者提供不同年度纵向比较的依据。美国的 SAT 和 ACT 考试，每年分别有七次和六次考试机会，每次考题各不相同，因此更是不断进行着严谨的等值链接工作，以保证各次考试间的成绩完全可比。

在英国，大学招生的重要依据 A-level 考试,过去的评分方法是给最好的 10%的学生 A，其次 15%的学生 B，以此类推。这种做法受到研究者们的强烈反对。自上世纪 80 年代中后期开始，A-level 的评分标准改为各科设定固定的内容和难度标准[9]。但是由于没有进行严格的等值和链接，对近年来发现的 A-level“分数膨胀”（Grade inflation）现象，就无法断定究竟是学生整体水平提升了，还是考试在革新中越来越简单了[10]。

我国的高考研究主要使用 CTT 理论，对难度的估算会随着样本特性的变化而改变，因此无法十分严谨地比较两个不同考试的结果[11]，造成了高考在提供纵向比较信息方面的作用十分有限。逐步建立与完善一个宏观的、纵向多角度的科学评估体系，可以进一步充实高考的内涵。

2. 跨区域比较

近年来，我国高考的另一新格局就是越来越多的省、直辖市开展自主高考。如果我们在全国统一高考和各自主考试间建立链接,就可以形成以全国统考为参照的“考试群”，既为地区间比较提供科学依据，也对统考和各自主考试的质量起到监督和促进的作用。

实际上，在自主考试越来越多的情况下，缺少必要的参照，可能还会造成严重的问题。比如在日本，很多私立高校不采用“全国共同学历第一次考试”（或称“大学入学中心考试”），而单凭自主考试决定招生。有研究显示，极少数精英学校的考试内容和难度远远超出了课程范围，大大增加了学生的应试负担[12]。而

更多学校则因为面临无法满额招生的困境，大幅降低考试难度，使得这部分学校的学生素质和教学质量都无法得到保证，被日本媒体称为“学力危机”、“学力崩溃”[13]。 这一教训值得我们警醒和深思。

最后，如果我们能在高考和其他针对高中毕业生的国际考试或其他国家和地区的高考间建立联系，那么我们还可以获得国际比较方面的信息。比如由 OECD 组织的 PISA 考试已成为颇具影响力的大型国际考试，因而目前有不少国家都在努力构建本国对应年龄段、对应科目考试与其的链接关系。

四、总结和讨论

我国的高考招生制度正处在革新的转折点。革新的需求和当代教育测量学理论的发展为我国建立一套多角度、多层次的高考质量科学评价体系提供了契机。

采用IRT中的分析方法，以大量实测数据为基础，我们可以对高考的信度和效度进行更全面、更深入的探索[14]。在提高高考科学性的同时，研究者发现的问题又可以被快速反馈到命题乃至教学中，从而形成一个循环的、不断完善的过程。大量的高考质量研究还有利于让教师、考生、家长、高校和社会各界关心高考的人士有渠道正确认识高考、监督高考，了解其科学性和局限性，从而更好地运用高考所提供的信息。

在保证单次高考质量的基础上，建立跨年度、跨区域的考试间链接关系，能够进一步发挥高考的作用，在提供多层次的教育质量评估信息之余，既监督促进了自主考试的质量，还提升了自主考试的公信力。

高考作为我国现阶段最有影响的大规模教育考试，肩负着为高等教育选拔人才、公平分配高等教育资源等重要社会职能，对高中课程改革、高校招生制度改革等也有着信息反馈的作用。因此，高考质量研究具有重大的政策意义。当代测量学理论在我国的研究和运用尚处于起步阶段，如果善加利用，定能促进高考更好地服务于社会。

参考文献

[1] 赵海燕,臧铁军.CTT框架下基于数据分析的高考试题质量评价标准[J].中国考试,2009,3-15.

[2] 扈涛.高考主观题评分误差控制的研究与实践[J].河南大学学报,1996,36(4):31-35.

[3] Wilson, M. Constructing Measures: An Item Response Modeling Approach. Mahwah, New Jersey: Lawrence Erlbaum Associates. 2005.

[4] Gose, B., & Selingo, J. The SAT's greatest test: Social, legal and demographic forces threatening to dethrone the most widely used college entrance exam. The Chronicle of Higher Education, 2001. A10-15.

[5] Stricker, L. J., Rock, D. A., & Burton, N. W. Using the SAT and High School Record in Academic Guidance. Educational and Psychological Measurement, 1996. 56(4), 626−641.

[6] Miller, D. W. Scholars say High-Stakes test deserve a failing grade: Studies suggest students and educators are judged by faulty yardsticks. The Chronicle of Higher Education. 2001. A14.

[7] 郭崇义,王青. 1993年以来高考英语出题方式统计分析[J].教学与管理,2000,66-68.

[8] 吴根洲.高考效度研究文献述评[J].考试研究,2009,49-51.

[9] Choppin, B. Is Education Getting Better? British Educational Research Journal, 1981.7, 1-15.

[10] AQA. GCSE Design and Technology. Manchester The Assessment and Qualifications Alliance. 2010.

[11] 宋宝和,高黎明,张振鸿.影响高考选做题等值性的因素及其控制策略[J].中国考试,2009,26-29.

[12] Underwood, P. A Comparative Analysis of MEXT English Reading Textbooks and Japan's National Center Test. RELC Journal, 2010. 41(2), 165–182.

[13] 王丽燕.日本大学招生制度改革及其发展趋势述评[J].教育科学,2008,24(4):86−89.

[14] 王蕾.客观等距量尺在高考评价指标体系中的应用与构想[J].中国考试,2008,17-22.

教育财政专业委员会会员第二期共享资料：

《公共教育支出统计的可比性研究》[1]

（会员可登陆北京大学中国教育财政科学研究所官方网站（www.ciefr.pku.edu.cn），在右下角“会员登录”处注册、待注册成功并审核通过后，在“学会信息”栏目中免费下载。）

目　录

第一篇　公共教育支出统计的可比性研究

一、导言

二、公共教育支出统计的国际标准

三、外国案例考察

四、我国公共教育支出统计

五、政策建议

附件一　全国教育经费统计报表填表说明

附件二　UOE 手册教育财政表

第二篇　UOE 教育系统数据采集手册（卷一）[2]

一、简介

二、UOE 数据采集的定义、解释和说明

三、关于 ISCED–97 在 UOE 数据采集中的使用说明

附件一 弗拉斯卡蒂手册摘录

附件二 1993 年国民账户体系摘要

附件三 社会缴款推算实际操作指南

第三篇　OECD 国际比较教育统计手册第四章（节选）[3]

[1] 王蓉，魏建国. 公共教育支出统计的可比性研究[M]. 北京：中国财政经济出版社，2008.

[2] UNESCO/OECD/Eurostat,2006.UOE Data Collection on Education Systems Volume1 Manual Concepts, Definitions and Classifications Montreal, Paris, Luxembourg.

[3] OECD. 2004. Chapter 4 Definitions and Classifications , OECD International Education Statistics, Paris: OECD.

一、教育机构

二、教育支出

第四篇　教育支出统计指南[4]

前言

一、导言

二、什么是 UOE 数据采集系统

三、概念、定义和分类

四、财政数据采集的范围

五、棘手案例

六、资金来源和转移支付：财政表 1 中的支出分类

七、财政表 2：按性质、资源种类及教育层级划分的教育支出

八、关于债务附表的说明

九、对研究和开发支出附表的说明

十、结语

第五篇　欧盟 2005 年国家概览调查报告[5]

一、模块 1：按政府级别和教育层级划分的支出类型

二、模块 2：政府间教育专项转移支付

三、模块 3：支出类型：教育机构人员的养老金计划

四、模块 4：支出类型：资本性支出

五、模块 5：家庭支出

六、模块 6：辅助服务

七、模块 7：学生资助

八、模块 8：私有企业为特定教育活动所作的支出

九、模块 9：教育机构财政

第六篇　UOE 数据采集覆盖范围快速调查报告[6]

[4] European Communities, 2005. A Guide to Educational Expenditure Statistics, Luxembourg: European Communities.

[5] European Communities, 2005. Survey on Country Profiles, Luxembourg: European Communities.

[6] European Communities, 2001. Results of the Quick Survey on the Coverage of the UOE Data Collection, Luxembourg: European Communities.

一、导言

二、按照大类对反馈进行的分析

三、最终评论

四、结论

五、快速调查中适用的大/小类别列表

第七篇　教育支出统计信息系统[7]

前言

导论

一、领域的界定

二、支出的概念

三、支出中心

四、支出的经济分类

五、支出的功能分类

六、信息来源

七、调查程序

八、估计过程

九、支出分类过程

十、指标

附录 关于一些支出中心的财力和支出状况的问卷样本

第八篇　美国公共教育财政调查指导手册[8]

一、概述

二、如何使用财政数据

三、美国人口普查局

四、数据采集的步骤

五、如何提交财政数据

六、数据的准备、采集、审查和提交

[7] UNESCO,1998. Statistical Information System on Expenditure in Education, Gérard Lassibille and Bernard Rasera, Institute for Education Economics Research (IREDU), the Division of Statistics of UNESCO, Paris: UNESCO.

[8] U.S. Department of Education, Institute of Education Sciences 2007, The National Public Education Financial Survey Instruction Booklet, The National Center for Education Statistics (NCES).

七、准备和提交可比性数据

八、其他事项

九、公立初等和中等教育所有来源收入

十、公立初等和中等教育支出

上期回顾

2010 年第 10 期（总第 49 期）

关于代课教师若干问题的初步思考

摘要：改革和完善教师制度包括教师收入制度成为当前最为重大的制度建设挑战。在目前我国的代课教师群体中，存在相当比例的年轻新型代课教师。对于解决与他们相关问题的政策设计，需要新的思维和更加纵深的视角，以及更加广泛地借鉴国际经验，也必须与建设和完善我国教师制度这一根本问题联系起来。

本《简报》由北京大学中国教育财政科学研究所主办；旨在反映本所最新的学术科研活动；相关内容仅体现作者本人观点，并不必然代表本所的立场。

简报内容仅供参考，如需转载须事先征得本研究所同意。

本期印发：1600 份

下载网址：http://ciefr.pku.edu.cn

主办单位：北京大学中国教育财政科学研究所　　**电子信箱：**ciefr@pku.edu.cn

责任编辑：毕建宏　　**电　　话：**010-6275-9700

地　　址：北京市海淀区颐和园路 5 号
北京大学教育学院大楼 413 室（100871）　　**传　　真：**010-6275-6183