# 高风险语言测试的公平性检验框架研究*
## ——以高考英语为例

罗 娟 肖云南
(湖南大学,长沙 410082/中南林业科技大学,长沙 410004;湖南大学,长沙 410082)

**提 要:** 大规模高风险测试对社会及利益相关者的影响极大,测试公平性检验成为教育测量领域的研究重点。本文梳理语言测试界对公平性的定义及理论框架,从计量学与社会学两个层面界定公平性的定义,并从测量公平与社会公平两个维度构建测试公平性的检验框架。结合我国高考英语,从实证角度明确从两个维度进行公平性检验的具体内容及步骤,并论证两者间的关系,探讨该检验框架对我国大规模测试走向公平化的指导意义。

**关键词:** 高风险测试; 公平性; 测量公平; 社会公平
**中图分类号:** H319.5 **文献标识码:** A **文章编号:** 1000-0100(2018)01-0086-6
DOI **编码:** 10.16263/j.cnki.23-1071/h.2018.01.013

## Evaluating Fairness of High-stakes Language Test: An Empirical Approach Based on Chinese Matriculation English Test

Luo Juan Xiao Yun-nan
(Hunan University, Changsha 410082, China/Central South University of Forestry and Technology, Changsha 410004, China; Hunan University, Changsha 410082, China)

Due to the increasingly important role that large scale and high stakes language testing plays in today's society and different stakeholder groups, test fairness has become one of the key research topics in educational measurement field. Based on a review of the definitions and theoretical frameworks of test fairness, this paper proposes a definition from the aspects of metrology and sociology, and constructs a theoretical framework which contains two inter-related dimensions: measurement fairness and social fairness. And then it proceeds to provide detailed steps on how to integrate the two dimensions in test fairness evaluation practices based on Chinese Matriculation Test (English Test); moreover, the relation of the two dimensions is clarified and suggestions on how to promote better test fairness in China is discussed. The study helps develop professional standards for language testing fairness in large scale and high stakes language tests theoretically and practically.

**Key words:** high-stakes test; test fairness; measurement fairness; social fairness

## 1 引言

近年来,语言测试工作者的研究重点逐步从提高语言测试信度与进行效度验证转向对语言测试公平性问题的关注(何莲珍 吕洲洋 2013:164),目前探讨的热点主要围绕测试公平性的定义及检验框架。虽然语言测试界认识到测试公平对大规模测试的重要意义,但在很多方面未达成共识。本研究从测量公平与社会公平两个维度构建测试公平性检验框架,明确两者的关系,并以高考英语为例探讨测试公平性检验实践。

## 2 语言测试公平性定义及检验框架

对测试公平的定义随不同的社会与政治环境而变化,近年来研究重点逐步转向语言测试对社

* 本文系湖南省教育科学"十二五"规划项目"网络多媒体课堂外语教学绩效评估研究"(XJK015BGD091)和湖南省社科基金项目"基于认知诊断理论的 ESL 分级测试体系研究"(16YBA392)的阶段性成果。

会的影响，诸多学者开始从社会视角判断测试公平性，探讨其概念并尝试构建检验框架。Kunnan（2000: 1，2004: 27）基于考试心理测量学属性，对测试公平性定义从考试效度、机会均等、无偏差、施考条件与社会后果 5 个部分进行拓展；并强调考试应促进社会公平，减少测试带来的负面影响。该框架提出迄今最为全面的公平性检验框架，但操作性不强，无法给研究人员的公平性检验提供切实、有效的指导（Xi 2010: 147）。将效度验证与公平性验证相互统一，并将公平性检验的各个部分形成连贯的论证链，有助于深入理解测试分数的使用情况及产生的社会后果。但该操作框架中，公平性检验与效度验证存在明显交叉（李清华 2016: 549），让研究者在实践操作中无所适从。Walters（2012: 469）提出从微观分析与宏观分析两个方面检验公平性。前者基于计量分析，依靠技术质量检测；后者使用质性方法，从社会视角来判断。该模式提出的微观和宏观之分看似较为全面又具体，但实际上两方面之间交叉较多，对实践的指导意义有限。参照“语言测评使用论证”，李清华（2016: 549）构建的公平性检验理论框架将公平性划分为“测量公平性”与“社会公平性”两部分，认为公平性检验既有技术属性，又有社会属性，并明确公平性检验的具体步骤及研究问题，具有理论突破意义。

综上所述，近年来语言测试界以更广阔的视角从计量学与社会学两个层面来界定测试公平性，逐渐将测试公平性研究从测试命题、施测、评分扩展到分数解释、测试决策及产生的社会后果，着眼于整个测试始终。借鉴以上学者的观点，本文将测试公平定义为在测试命题、施测、分数评定及进行分数解释、作出测试决策、使用测试结果的一系列过程中，所有受试群体及个人得到相同的待遇，不存在有利/不利某受试个体/群体的现象。基于以上定义，本文尝试从测量公平性与社会公平性两个维度提出语言测试公平性检验框架：

其中，测量公平性维度侧重从测量学范畴检验测试公平性的计量指标，体现为测试命题、施测、评分阶段所有受试个体/群体接受无偏颇的评估内容及形式、同等的评估条件及评分方式，不存在有利/不利某受试个体/群体的现象；社会公正性维度注重从社会、政治视角对公平性进行质性检验，体现为测试的分数解释及测试决策使所有受试者得到同等待遇，不存在有利/不利某受试个体/群体的现象，并且测试结果的使用对教育体系、社会环境产生系统、显著的积极影响。

表 1　语言测试公平性检验框架

| | | |
|---|---|---|
| 测量公平性 | 测试命题 | 测试内容避免偏颇、与构念无关的内容；<br>测试形式为所有考生熟悉；<br>为考生提供充分发挥能力的平等机会。 |
| | 测试施测 | 施测环境使所有考生（残疾生、少数民族等）受到公平对待。 |
| | 分数评定 | 避免评分误差、对所有考生一律平等。 |
| 社会公平性 | 分数解释 | 对所有考生一视同仁。 |
| | 测试决策 | 为所有考生提供平等受教育的机会。 |
| | 测试使用 | 对教育体系的影响：正面反拨效应；<br>对社会环境的影响：促进社会公平。 |

## 3　研究设计

### 3.1 研究问题

测试公平性是一个较为主观、相对的概念，必须置于特定社会、文化环境中进行研究（McNamara，Roever 2006: 197）。我国人口众多，教育发展不平衡，考生群体复杂，其他社会环境下建立的测试公平性理论并不一定完全适用于我国国情。基于我们已经构建的测试公平检验框架，下文将以中国高风险测试——高考英语为例，结合我国国情从测量公平与社会公平两个维度检验分省命题下的测试公平，探讨以下问题：（1）如何从测量公平与社会公平两个维度检验语言测试公平性；（2）如何看待两者间的关系；（3）以上结论对改革我国测试现状有何指导意义。

### 3.2 实验设计

自 2000 年，在分省命题政策下，各省根据教育部《全日制普通高级中学教学大纲》（以下简称《教学大纲》）制定出十几套高考试卷，试题内容、题型各不相同，各省录取分数线也不相同。鉴于各年与各省的高考试卷与考生相互独立，且高考实测数据的保密性，本文利用等值研究中的共同组设计（common-group design），通过高考模拟测试收集实验数据进行计量分析回答研究问题（1），并基于分析结论对研究问题（2）及（3）展开探讨[①]。

### 3.3 试卷结构

经过比较各省试卷，笔者发现上海卷与江西卷在试卷结构与测试微技能等方面很相似，因而选取 2008 年上海卷（简称卷 A）、2009 年上海卷（简称卷 B）、2009 年江西卷（简称卷 C）用于实验。选择 2009 年上海卷与 2009 年江西卷旨在检验同年各省间高考英语的测试公平性，选择 2008 年与 2009 年上海卷旨在探究同省历年高考英语的测试公平性。

3.4 测试对象

依据高中统考成绩，本研究以高、中、低 3 个水平抽取湖南省 3 所高中 1157 名高三学生参加测试，3 套试卷相隔 1 周施测 1 卷，以保证考生能力的同质性。该批考生处于高考备考阶段，且模拟成绩计入月考成绩，因此与高考测试群体在能力分布与测试动机上有很高的同质性。

## 4　测量公平性检验

测试的公平性首先体现在测量公平上，贯穿测试命题、施测与评分 3 个阶段，本节侧重从测试命题方面进行试卷的计量分析。测量公平主要体现为测量有效、测量误差小、分数具有可比性和可解释性等（杨惠中 2015: 2），这样测试才能为考生提供充分发挥能力的平等机会。测量有效是指测试不涉及与构念效度无关的因素，误差小要求测量信度高，可比性是指不同考次的测试分数可直接比较，可解释性是分数表示的意义可以解释，为用户决策者提供依据。下文将从试卷效度、信度、分数可比性方面对高考试卷进行测量公平维度的计量分析。

4.1 构念效度验证

在参详《教学大纲》后，实验组 3 位语言测试专家以经验判断，卷 A，B，C 基本以此为准编制试题，总体覆盖考纲技能，测试内容及结构符合标准。经 Bartlett 球度检测，3 套试卷适合进行因子分析（$P < .01$）；然后采用主成分分析法显示，卷 A，B，C 因子分析抽取的因子 1 的值较高，均能解释该卷绝大部分方差（卷 A: 66%；卷 B: 56%；卷 C: 75%），按照《教学大纲》要求，高考英语应强调英语综合应用能力，因此可认定因子 1 即综合英语应用能力（分析表略）。

4.2 试卷信度

本文采用项目反应理论（Item Response Theory，简称 IRT）首先对试题进行参数估计，同时估计试卷信息函数（test information function，简称 TIF），参数估计软件为 IRTPRO（Cai，Thissen，du Toit 2011）。在 IRT 理论中，采用 TIF，也就是测验对受试能力估计所提供的信息量多少来表示测量的精度，并能估计不同能力受试的测量精度，代替传统的信度概念。

在高风险测试中，划界分数处的考生能力估计精度对测试决策的误差大小产生关键影响，在此处测试应具有较高的测量精度，将划界分数附近的受试准确区分，决定是否录取，将误判率降到最低。笔者参考当年全国高考录取率（2008 年 57%，2009 年 62%），假设高考分数呈正态分布，对照正态分布表可见划界分数点的能力估计值在 [−0.4，0] 之间。在此区间，虽然 3 套试卷的 TIF 值均达到最高值，测量标准误差为最低（见表$_2$），但显然存在差异：卷 C 的 TIF 值在该区间最高，在划界分数处的测量精度最高，而卷 B 则为最低。

表$_2$　卷 A、卷 B、卷 C 测验信息值分布

| | 受试能力值 $\theta$ | −2.8 | −2.4 | −2 | −1.6 | −1.2 | −0.8 | −0.4 | 0 | 0.4 | 0.8 | 0 | 1.2 | 1.6 | 2 | 2.4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 卷 A | 试卷信息函数 | 10.57 | 14.28 | 18.64 | 22.93 | 26.07 | 27.03 | 25.7 | 22.95 | 19.62 | 16.1 | 12.71 | 9.77 | 7.44 | 5.69 | 4.4 |
| | 标准误差 | 0.31 | 0.26 | 0.23 | 0.21 | 0.2 | 0.19 | 0.2 | 0.21 | 0.23 | 0.25 | 0.28 | 0.32 | 0.37 | 0.42 | 0.48 |
| 卷 B | 试卷信息函数 | 10.1 | 13.24 | 16.65 | 19.71 | 21.62 | 21.96 | 20.94 | 19.1 | 16.97 | 14.74 | 12.48 | 10.34 | 8.3 | 6.41 | 4.9 |
| | 标准误差 | 0.31 | 0.27 | 0.25 | 0.23 | 0.22 | 0.21 | 0.22 | 0.23 | 0.24 | 0.26 | 0.28 | 0.31 | 0.35 | 0.39 | 0.45 |
| 卷 C | 试卷信息函数 | 10.36 | 14.98 | 21.34 | 29.38 | 37.74 | 41.19 | 36.04 | 27.29 | 19.39 | 13.57 | 9.61 | 6.99 | 5.23 | 4.02 | 3.15 |
| | 标准误差 | 0.31 | 0.26 | 0.22 | 0.18 | 0.16 | 0.16 | 0.17 | 0.19 | 0.23 | 0.27 | 0.32 | 0.38 | 0.44 | 0.5 | 0.56 |

4.3 试卷难度

基于 IRT 理论，我们对试卷的两级计分选择题用双参数模型进行项目参数估计，除写作题外的主观题用等级评分模型分析，然后对全卷项目参数进行描述性统计，以比较 3 套试卷难度。

表$_3$　卷 A、卷 B、卷 C 试题难度参数 $b$ 描述性统计

| | 题量 | 全距 | 最小值 | 最大值 | 均值 | 标准差 |
|---|---|---|---|---|---|---|
| 卷 A | 84 | 3.48 | −1.24 | 2.24 | .54 | .86 |
| 卷 B | 84 | 4.14 | −1.63 | 2.51 | .37 | .99 |
| 卷 C | 85 | 6.06 | −2.39 | 3.67 | .98 | 1.09 |

表$_4$　卷 A、卷 B、卷 C 试题区分度 $a$ 描述性统计

| | 题量 | 全距 | 最小值 | 最大值 | 均值 | 标准差 |
|---|---|---|---|---|---|---|
| 卷 A | 84 | .83 | .06 | .89 | .49 | .19 |
| 卷 B | 84 | .83 | −.17 | .66 | .24 | .18 |
| 卷 C | 85 | 2.72 | .01 | 2.73 | 1.17 | .53 |

由表$_3$和表$_4$可见，卷 C 的试题难度 $b$ 及试题区分度 $a$ 的均值在 3 卷中均为最高（$b_{mean} = .98$；$a_{mean} = 1.17$），在 3 套试卷中难度最大，区分度最好；卷 B 试题难度 $b$ 及区分度 $a$ 的均值（$b_{mean} =$

. 37; $\alpha_{mean}$ = . 24) 均为最低，难度最小，区分度欠佳；而卷 A 的两个指标均值处于两卷之间。由此可见，无论是同省跨年试卷，还是同年跨省试卷，均出现试题难度、区分度不稳定的现象。

4.4 测试分数可比性

试卷间因难度差异对分数可比性产生的影响，一般通过等值将分数转换到统一量表后验证，本文采用共同组设计的分数等值，向参加实验的所有考生先后施测 3 套试卷后将卷面分进行等值。高考为常模参照考试，依据考生成绩在各省考生群体中的相对排名而非绝对的考试分数择优录取，因而采用等百分位法( equipcentile method) 将分数进行等值。其等值原理为: 两个不同测验形式的分数，如它们的百分等级相同，即被认为是等值的，实质是基于在考生群体中的相对排名的等值方法。在 3 套试卷中，卷 A 的难度、区分度及信度都居中等，现将卷 A 定为基准卷，采用经过平滑处理的等百分位法进行等值，将卷 B、卷 C 分数转化到卷 A 上来。

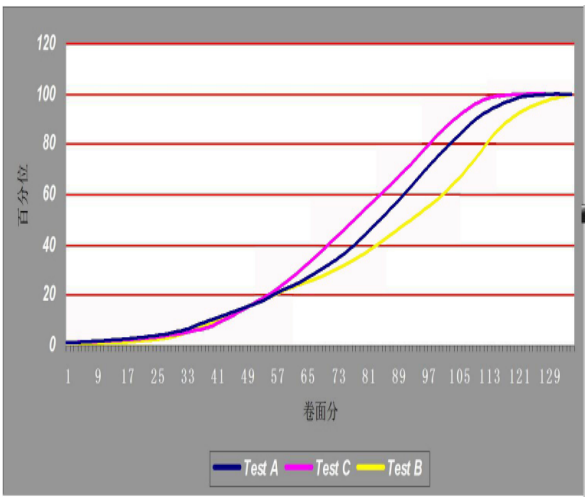


图 1　3 套试卷卷面分—百分位曲线对照表

图 1 显示，经等值处理后，3 套试卷的相同卷面分在考生群体中对应的百分位差异显著，卷 C 的卷面分对应的百分位最高，卷 B 则最低。换言之，因试卷难度差异较大，3 套试卷的相同卷面分表面上看似分值相等，但实质反映考生的不同能力，因此，在考生中的相对排名截然不同。例如，依据等值结果，卷 A 的 100 分处于考生群体的百分位为 58，而卷 B 与卷 C 的 100 分对应的百分位分别为 47 与 70。由此可见，各省、各年的高考分数本身不具有直接可比性，并且各省考生的常模团体不同，如不经等值依据各省考生排名制定录取决策，显然对试卷偏易的考生群体有利，而对试卷偏难的考生群体不利。

由此可见，在大规模测试的分数解释阶段应实现对不同测试群组间分数的可比性，基于这一前提，对各受试群组作出的测试决策才具有合理性( Kane 2010: 177) 。未经过等值，测试成绩间不具备可比性，评价标准也会因试卷难度差异的影响而对测试公平造成威胁( He，Qi 2010: 359，Kobayashi，Negishi 2008: 244) 。

## 5　社会公平性检验

测试公平的另一维度是社会公平，检验在特定社会环境下，测试分数的解释、决策是否存在有利/不利某受试个体/群体的情况，测试结果的使用是否对教育系统产生系统、显著的正面反拨( washback) ，是否发挥积极的社会性功能，对社会环境是否有正面后效。该维度涉及社会层面较广，主要为测试用户及利益相关群体，如政府机关、教育机构、公司、考生、教师等，检验方法以质性研究为主。目前国内外对于测试的社会公平性研究不多，相关研究以教学反拨为主探讨其对教育体系的影响，对社会环境的影响关注不足。

5.1 分数解释及决策

虽然高考各省、各年试卷在计量指标上存在明显差异，且各省考生团体常模存在差异，在分省命题政策下，高考采用常模参照性评价，根据考生原始分在各省常模中的相对位置转化成标准分进行分数解释。考生的相对等级随着用来比较的常模团体的不同而变化，对高考分数的解释也会产生显著、系统性的影响，所以处于教育欠发达地区的考生群体因此会受益，而对教育相对发达地区的考生群体不利。

在录取政策上，高考实际未经各省试卷分数等值，采取地区配额制度实行全国招生，即高校拥有招生自主权，独立分配各省招生人数，按照考生分数在各省相对排名的先后择优录取。地区配额招生制度表面上照顾到各省教育资源差异及教育发展不平衡的国情，但导致一系列负面社会影响: 各大高校招生指标分配明显偏向于本地考生，严重歧视外地考生接受高等教育的平等权力。各省试题不一，分数没有可比性，高考就丧失统一衡量、平等选拔的功能，因而掩盖了全国高校录取指标分配不公的现实，恶化了招生地域歧视，限制了广大考生接受高等教育的平等权利。

5.2 教学反拨

纵观近年来高考英语反拨效应研究( 董连忠 2014; 朱明瑛 2012; 陈丽珍 2009; 洪小祥 2008; 亓鲁霞 2004，2007) ，高考英语对高中课程设置、教

学内容、教学方法、教学评估、师生教学态度等产生不同程度的影响，总体上呈现出对高中英语教学正面反拨作用增大、负面反拨效应相对缩小的趋势。尽管国内高中的总体教学目标向新课标中“培养学生的综合语言应用能力”靠拢，但“应试教育”现状依然严重，尤其是毕业班。高考分数被误用作评估学校、师生的唯一量化指标，师生压力较大。总而言之，高考的反拨效应在大体上有利于我国高中英语教学，但负面反拨在毕业班的教学中较为明显。

5.3 社会后效

高考是我国最有影响的高风险大规模考试，是教育教学和高等人才选拔的基本制度，对于促进教育发展与稳定社会发挥着重要作用，但我们应全面、客观、公正地看待高考的社会性作用。

显然，高考改革历程中的分省命题及地区配额招生制度引起一系列负面社会影响。首先，它造成大学生源的地方化和录取标准的严重不公；然后，经济、文化发达地区形成高度集中的教育资源优势，以低标准录取当地考生，增强发达地区对人才与资源的吸引力，催生“高考移民”现象，导致该地区人才、物质、财富更加集中，进一步加剧资源配置失衡；其次，资源相对集中不利于全国范围内的人才流动，教育发达地区的毕业人才就业压力过大，而欠发达地区则人才日益匮乏。如此恶性循环，高校招生地方化只能进一步扩大城乡差别，人才与资源不断从农村流向城市的形式日益严重。最后，高考招生制度饱受社会各阶层诟病，成为社会不和谐的重要因素。据中国青年报调查显示，89.3% 的民众认为全国重点大学招生指标分配不公平。高考招生歧视侵犯全国大多数地区考试的利益，引起公众普遍不满，容易激化地区矛盾，影响共建和谐社会。

## 6　讨论

基于本文构建的测试公平性检验框架，笔者对分省命题的 3 套高考英语试卷从测量公平与社会公平两个维度进行检验。

首先，对高考命题的计量分析显示，3 套试卷在难度、区分度及信度方面存在较大差异，试卷难度的起伏无疑对考生的测试表现会造成系统性的影响，并直接导致试卷分数的不可比，试卷信度的差异也意味着测试对考生能力评估的准确性存在差异。显而易见，计量分析揭示出的命题缺陷，致使高考试题无法为考生提供发挥能力的平等机会，也直接影响测试决策的公平性。其次，高考的分数解释及地区配额招生决策违背所有受试享有接受高等教育平等权利的原则；高考结果的使用对教育反拨的负面影响虽然呈减少趋势，但引发一系列负面社会影响，妨碍社会公平的实现。

总而言之，分省命题下的高考英语在测量公平性与社会公平性两个维度上有所欠缺，真正实现测试公平有待进一步改革。

6.1 两个维度的关系

本文围绕大规模测试的公平性定义展开探讨，从测量公平与社会公平两个维度构建测试公平性检验框架。基于该框架对高考英语试卷的实证分析可见，两个维度的公平性检验贯穿测试的全过程，两者既有独立要求，又紧密联系，缺一不可。

首先，测量公平仅是测试公平性研究的一部分，是决定测试公平的前提与基础。该维度主要由测试机构及测试工作者负责，涉及心理测量、教育学、心理学等多学科的交叉应用，以技术性手段保证学术行为决定。测量公平先于社会公平，只有实现测量公平才能谈社会公平，才能保障社会公平（杨惠中 2015：2）。

然后，社会公平维度是测试公平性研究的重要方面，是体现测试社会功能的关键因素。该维度超出测试工作者能控制的范围，主要由我国某些政府职能部门负责，涉及政治、经济、道德及价值观等多种复杂因素，公平性检验多以质性研究方法为主。有悖社会公平，将削弱测量公平的作用，最终阻碍测试公平的实现。

只有清晰地界定测试公平性研究的维度、明确各方在维护测试公平性中应承担的责任，才能最后形成连贯的、系统的测试公平性框架。要实现测试的公平性，不仅要确保测试开发机构在考试过程中的专业行为，也要确保相关行政机构对测试结果的解释合理、决策得当，确保将促进教学、促进社会公平作为测试改革的基本价值取向。

6.2 对测试实践的指导意义

测量有效、测量可信、分数具有可比性与可解释性是测量公平的基础。我国诸多考试为超大规模考试，参考人数众多，考生群体复杂，出于试题保密性和可操作性等原因，采用平行卷是常见做法。但众多大规模测试未实现等值，如高考、高中会考、公务员考试等。为使考生间的分数具有可比性，必须对平行卷进行等值处理，并逐渐建立试题库系统，以克服命题的片面性、随意性，从而实现命题标准化、施测标准化、评分标准化、分数解释标准化，为实现测试的测量公平性提供前提。

国内大规模测试均由各级教育或考试主管部门实施，基于分数进行决策是行政行为多于学术行为，与测试开发者的预想存在一定脱节，由此产生社会公平性问题是国内语言测试公平性最突出的问题（李清华 2016：549）。由于其权威性，测试决策的公平性很少受到公开质疑，相关行为无法得到有效监督与约束。因此，一方面研究者关于测试使用的后效，如对教育体制、社会各层面影响的研究亟待加强；另一方面，单靠测试机构无法确保测试的社会公平性，应委托独立研究机构进行社会公平性检验，其研究报告应向公众公开。权威机构也应自觉将相关工作置于社会监督之下，积极促进考后分数解释的科学化、录取政策的公开化、测试使用的科学化。

## 7　结束语

大规模高风险测试对考生、教育及社会的影响极大，其公平性检验不容忽视。本文构建的公平性检验框架将促使语言测试界的研究重点从心理计量学范畴向社会学范畴延伸，对两者间关系的探讨具有重要理论价值及现实意义：帮助测试机构及测试工作者进一步理解公平性的内涵，同时促使相关行政部门提高测试公平性意识，从政策上保障测试公平性，减少测试结果的误用及滥用。双方的共同协作对于推动我国语言测试的公平性及专业化进程极为重要。

分省命题已成为高考改革历程中的一个背影，但其弊端对促进我国大规模测试的公平性提供诸多借鉴。2016 年我国高考逐渐实现全国统考，是我国高风险测试走向公平化的一项重要举措，标志着新一轮考试招生改革的全面推进。

**注释**

①本文实验数据来自国家社科规划项目“全国高考分省命题的英语分数等值模型研究”。

**参考文献**

陈丽珍. 高考英语测试对高三英语教学反拨效应的研究［D］. 福建师范大学硕士学位论文，2009.

董连忠. 全国高考北京市英语考试对高中英语教学的反拨效应研究［D］. 上海外国语大学博士学位论文，2014.

洪小祥. 高考英语考试反拨效应的调查研究［D］. 湖南师范大学硕士学位论文，2008.

何莲珍　吕洲洋. 语言测试研究的新视角：批判语言测试［J］. 浙江大学学报，2013(6).

李清华. 语言测试的公平性检验框架［J］. 现代外语，2016(4).

亓鲁霞. 意愿与现实：中国高等院校统一招生英语考试的反拨作用研究［M］. 北京：外语教学与研究出版社，2004a.

亓鲁霞. NMET 的反拨作用［J］. 外语教学与研究，2004b(5).

亓鲁霞. 高考英语的期望后效与实际后效——基于短文改错题的调查［J］. 课程· 教材· 教法，2007(10).

杨惠中. 有效测试、有效教学、有效使用［J］. 外国语，2015(1).

朱明瑛. 新高考英语反拨效应研究［D］. 福建师范大学硕士学位论文，2012.

Cai，L.，Thissen，D.，du Toit，S. H. C. IRTPRO for Windows［CP］. Lincolnwood：Scientific Software International，2011.

He，L.，Qi，L. Gui Shichun：Founding Father of Language Testing in China［J］. *Language Assessment Quarterly*，2010(4).

Kane，M. Validity and Fairness［J］. *Language Testing*，2010(2).

Kobayashi，M.，Negishi，M. An Interview with Professor Kenji Ohtomo：The Founding Father of Language Testing in Japan［J］. *Language Assessment Quarterly*，2008(5).

Kunnan，A. J. Fairness and Justice for All［A］. In：Kunnan，A. J.(Ed.)，*Fairness and Validation in Language Assessment*［C］. Cambridge：Cambridge University Press，2000.

Kunnan，A. J. Test Fairness［A］. In：Milanovic，M.，Weir，C.(Eds.)，*European Language Testing in a Global Context*［C］. Cambridge：Cambridge University Press，2004.

McNamara，T. F.，Roever，C. *Language Testing*：*The Social Dimension*［M］. Oxford：Blackwell，2006.

Walters，F. S. Fairness［A］. In：Fulcher，G.，Davidson，F.(Eds.)，*The Routledge Handbook of Language Testing*［C］. New York：Routledge，2012.

**定稿日期：**2017－11－20　　【责任编辑　孙 颖　滕 琳】