文章编号:1000-5862(2017)04-0383-11

统计测量视角下考试公平推动教育公平的对策

汪文义[1],张华华[2,3*]

(1. 江西师范大学计算机信息工程学院,江西 南昌 330022;2. 伊利诺伊大学香槟分校心理系,伊利诺伊州 香槟 61820;
3. 华东师范大学教育学部,上海 200062)

摘要: 考试不公平将影响被试受教育机会的公平性和社会公平性. 针对我国考试公平性中统计分析长期被忽视问题,该文主要从统计测量视角,在介绍测验公平性评价在国外盛行情况之后,深入剖析测验公平性统计分析的项目功能差异方法. 最后,针对高厉害考试的公平性问题,提出促进考试公平的详细并可行的举措,以供读者借鉴.

关键词: 考试公平; 教育公平; 项目功能差异; 统计测量; 高考

中图分类号: B 841.7 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2017.04.10

0 引言

2012年2月,国际经济合作与发展组织(OECD)出版了《教育的平等和质量:支持弱势学生和学校》一书[1]. 该书开篇中给出了教育公平的定义. 教育公平有2个方面:一是公平(fairness),即个人与社会背景,如性别、种族、家庭背景或社会经济地位等,不应该成为潜能发展的障碍;二是全纳(inclusion),即要确保全民教育至少达到基本的最低技能水平(例如,保证每个人都能够阅读、书写和进行简单的算数等). 鉴于联合国千年发展目标未能如期完成,继"仁川宣言"之后,2015年11月,联合国教科文组织(UNESCO)在巴黎总部通过并发布了"教育2030:仁川宣言和行动框架",旨在"为所有人确保包容、公平的优质教育并促进终身学习机会"[2]. 公平和全纳的缺失造成学业上的失败,而最明显的是表现为退学. 根据国际经济合作与发展组织报告显示,平均有20%的青年人在完成高中教育之前就退学了. 而学业失败和退学的社会和经济代价很高.

2010年中国政府颁布的《国家中长期教育改革和发展规划纲要》(2010—2020年)提出:把促进公平作为国家基本教育政策、把提高质量作为教育改革发展的核心任务. 近10年来我国采取一系列有力措施促进教育公平,推进教育信息化,让农村、边远、贫困和民族地区的孩子们共享优质教育资源;健全完善公平公正入学制度,为困难群体就学创造更多便利;落实《国家贫困地区儿童发展规划(2014—2020年)》,实施营养改善计划;完善义务教育质量评价体系和督导制度;实施《特殊教育提升计划(2014—2016年)》,促进残疾人全面发展;促进考试公平公正,考试作弊入法;已连续实施8年的"支援中西部地区招生协作计划",推进大学入学机会公平等,教育公平迈出了重大步伐.

教育公平是社会公平的基石,也是一个人感受社会的起点. 尽管走向现代化的中国,已使国人的成长成才之路变宽变多,但不能不承认,通过接受高等教育深造以改变命运、成就未来,依然是其中最重要的一条途径. 在所有向上流动的制度通道中,高考是最有效的一种公平制度,是寒门学子改变自身命运的最大希望,是守护教育公平及社会公平的重要底线. 但是,考试公平性的统计分析长期被忽视,缺少用"数据"说话.

高考作文题所占分数比重大,它对于不同群体考生的公平性,格外牵动人心. 倍受争议的高考作文题:2017年全国卷Ⅰ"中国关键词"、2016年浙江卷《虚拟现实》、北京卷二选一大作文题《"老腔"何以

收稿日期:2017-02-20

基金项目:国家自然科学基金(31500909,31360237,31160203,30860084),中国国家汉办HSK研究、美国国家科学基金(NSF-DRL1252389),国家留学基金(201509470001)和江西省自然科学基金(20161BAB212044)资助项目.

通信作者:张华华(1953-),男,上海出生,教授,博士生导师,长江学者,主要从事计算机化自适应测验和项目功能差异等研究. E-mail:hhchang@illinois.edu

让人震撼》、2015 年全国卷 I《给违反交规父亲一封信》、1991 年高考"三南"卷《妈妈爱吃鱼头,我从小就知道》[3] 等. 来自不同地域、城乡或民俗等能力水平相同的考生,在这些考题甚至其他考题上得分可能存在较为明显的差异. 例如,大多数农村考生可能并没有掌握虚拟现实相关知识,这不仅会引起考生紧张,还会影响考生正常发挥等,作文题分数差异可能会影响考试公平. 考试结果的不公平,将影响受教育机会的公平性.

2014 年,国务院发布《关于深化考试招生制度改革的实施意见》,明确提出要深化高考考试内容改革. 2015 年起增加使用全国统一命题试卷的省份. 2016 年,新一轮考试招生制度改革已进入具体实施阶段,高考考生 900 多万人,是美国高考 ACT 和 SAT 人数的 3 倍. 使用全国统一命题试卷的省份从 2015 年的 18 个增至 2016 年的 26 个,全国统一命题试卷增至 3 卷,全国 I、II 和 III 卷. 使用全国统一命题试卷,在提升效率的同时,有利于各省考生公平竞争和录取. 同时,给考试公平性也带来了前所未有的挑战. 高厉害考试的公平性,同一份全国卷对如此大范围内不同群体的考生的公平问题[3-4],亟需借助科学方法来保证.

1 测验公平性评价在国外盛行

在 1960 年之前,科研人员和公众对测验公平性关注较少. 而在 1960 年之后,测验公平在测验设计、开发和使用各个环节中成为关注重点[5]. 在 1964 年 7 月 2 日美国总统约翰逊签署民权法案(Civil Rights Act of 1964)之后,教育测验和就业测验的公平性问题成为公众关注的焦点[6-8],测验公平性曾径一度成为法律问题[8]. 最典型的例子是发生在美国的教育测评中心(Educational Testing Service,ETS)与伊利诺斯金簌保险公司(Golden Rule Insurance Company)长达 8 年(1976—1984)的法律纠纷. 后者认为,ETS 为伊利诺伊保险代理机构(Illinois Insurance Agents)开发的测验存在种族歧视,白人申请者比黑人测试者在项目正确作答概率上高出 15% 或更多[5,6,8].

此件事件后,ETS 在测验公平性评价方法之项目功能差异(Differential Item Functioning,DIF)方法方面做出了十分重要的贡献[6]. 项目(测验)功能差异指匹配能力水平下不同群体被试对试题(测验)上表现的差异. 美国教育研究协会(AERA)、美国心理学会(APA)以及美国教育测量学会(NCME)出版的《教育与心理测量标准》[9] 包含 4 章(第 2 部分的第 7,8,9,10 章)涉及公平和测验的标准,如第 9 和 10 章分别对特定群体(不同语言和文化背景、残疾考生)被试讨论测验公平标准. 为了强调公平在整个测验过程中重要性和对参加测验的所有被试及子群体的同等公平性,2014 版《教育与心理测量标准》[10] 在第 1 部分第 3 章集中介绍测验公平的背景和 4 大标准. 美国及工业与组织心理学会发布的《人事选拔与效度验证原则》中均有专章探讨测验公平问题[11-12].

2014 版《教育与心理测量标准》指出"测验公平一词并没有单一的技术含义,而在社会中往往以不同形式广泛使用",因此在《标准》中只给出公平的 4 个核心内容:测试平等性,测量无偏性,测验无障碍性,测验分数有效性. 测验公平是测验效度的基本问题,这 4 个方面均与效度相关. 测量偏差是指与测验所测构念无关因素导致相同构念能力被试组在测验分数上的系统差异(偏高或偏低),更强调定性方面. 测验的测量偏差检查方法通常是采用项目功能差异分析(更强调定量方面),并结合专家评价法对项目敏感性进行评价. 项目功能差异分析是测验公平性和测验效度证据的重要组成部分.

DIF 在国外已经相当成熟. 仅近年来,许多研究对国际学生评估项目(PISA)[13]、美国教育进步评价(NAEP)[14]、美国学术成就测验(SAT)[15]、佛罗里达综合成就测验(FCAT)[16]、南亚中学毕业考试(SSC)[17]、澳大利亚全国心理健康与幸福调查(NSMHWB)[18]、大学生一般自我效能感量表(GSE)[19] 和其他测验[20] 进行了分析,发现许多高质量测验仍有不少项目存在 DIF.

幸运的是,在测验设计、开发和使用中,测验公平性评价已经成为国外众多专业测验公司或测评项目考虑的中心问题[5,7]. 为了保证测验质量和公平,ETS[21-23]、ACT[24-25]、SBAC(Smarter Balanced Assessment Consortium)[26]、PARCC(Partnership for Assessment of Readiness for College and Careers)[27] 已经编制了大量技术文档,内容涉及在测验设计、开发、实施、分析和分数解释等过程中测验公平保证框架和方法,同时形成并定期发布测验公平性报告.

国内虽然开展了一些测验的 DIF 应用研究,如汉语词汇测验[28]、人格问卷[29]、高考数学[30] 和高考英语[31]、汉语水平考试(HSK)阅读测验[32-33]、英语四级[34]、2000 年临床医师内科学测验[35] 和其他测验[36-39]. 但是国内研究主要局限于 0-1 评分项目分析,DIF 成因分析薄弱和测验范围有待扩展等[40]. 而鲜有高厉害考试的公平性技术和结果报

告. 究竟如何从具体数据分析入手，借助什么统计测量方法，以保证考试公平，这是本文关注的重点.

2　项目功能差异

2.1　DIF 中重要概念

在介绍项目功能差异之前，先介绍 DIF 中一些重要概念[41-42].

1) 参照组(reference group)和目标组(focal group)，是根据人口变量划分，例如黑人/白人、男/女、城/乡、不同文化背景、不同民族、不同经济背景等. 记参照组为 R 和对照组为 F.

2) “研究”(studied)项目是指特定分析中 DIF 待检验的项目. 特别注意的是，测验中所有项目均需要进行 DIF 检验.

3) 匹配标准或匹配变量，即含或不含“研究”项目的测验观察或潜在分数. 如果不使用匹配标准，而单纯比较参照组和目标组在一个项目的正确作答比率，2 种原始比率差异，因为混杂了被试能力分布差异等变异，称之为“影响”(impact)，并非项目功能差异.

4) 非一致性 DIF 和一致性 DIF，根据 2 组被试在某项目(集)上的表现是否与匹配变量(如能力或总分)有交互作用[43]，DIF 类型分为非一致性 DIF 和一致性 DIF. 例如，有研究者在分析 ACT 的实验性数学表现型测验(含 21 个 0-1 记分题和 6 个多级评分题)时，对不同性别被试分析发现，在 0-1 记分题中，7 个题存在一致性 DIF 和 1 个题存在非一致性 DIF($p<0.005$)，发现多级评分题中存在一致性 DIF 和非一致性 DIF($p<0.005$)各 2 个[44].

2.2　DIF 方法分类

DIF 方法按匹配变量的类型可分成 2 类[45-47]：基于观察分数匹配的方法和基于潜在变量匹配的方法. 根据匹配变量与项目反应是否有数学模型假设，DIF 方法可分为参数方法和非参数方法[45-46,48]，参数方法依赖于模型指定和参数估计共线性问题，而非参数方法没有此问题，但是在样本较小时抽样误差易导致检验结果不稳定[45]. 本文下面介绍的方法分类情况见表 1 所示. 由于篇幅限制，基于项目反应理论的方法在此不做介绍.

2.3　DIF 的数学定义

为了对 DIF 进行量化处理，必须从数学上定义 DIF. 根据采用的匹配变量类型，下面列出 3 种 DIF 形式化定义[46].

定义 1(潜在变量 DIF 虚无假设)　令 Y 表示待探查项目(集)上总分，$E_R[Y\mid\theta]$ 和 $E_F[Y\mid\theta]$ 表示参照组和对照组下潜变量 θ 对 Y 的回归. $\forall\theta$ 没有 DIF 的项目须满足 $E_R[Y\mid\theta]=E_F[Y\mid\theta]$，其中 $\forall\theta$ 表示任意 θ 值，即对任意 θ 均成立. 建立在潜在变量 DIF 虚无假设下的 SIBTEST，直接通过统计检验方法判断此虚无假设是否成立[46].

表 1　项目功能差异侦测方法分类

	参数方法	非参数方法
匹配变量为观察分数	LR/LDFA	MH/GMH、STD/SMD
匹配变量为潜在变量	基于项目反应理论的方法	SIBTEST/Poly-SIBTEST

注：LR = logistic regression; LDFA = Logistic discriminant function analysis; MH = Mantel-Haenszel; GMH = generalized Mantel-Haenszel; STD = thestandardization approach; SMD = standardized mean difference; SIBTEST = simultaneous item bias test; Poly-SIBTEST = simultaneous item bias test for polytomous item.

定义 2(观察分数 DIF 虚无假设)　令 Y 表示待探查项目(集)上总分、X 表示匹配测验分数、$E_R[Y\mid X]$ 和 $E_F[Y\mid X]$ 表示参照组和对照组下匹配分数 X 对 Y 的回归. $\forall X$ 没有 DIF 的项目须满足 $E_R[Y\mid X]=E_F[Y\mid X]$，其中 $\forall X$ 表示任意 X 值，即上式要对任意 X 均成立. 尽管 MH 和 SMD 方法均采用了定义 2，但是 MH 方法是通过统计检验方法判断此定义 2 中虚无假设是否成立[46].

定义 3(真分数 DIF 虚无假设)　令 Y 表示待探查项目(集)上总分、V 表示匹配真分数、$E_R[Y\mid V]$ 和 $E_F[Y\mid V]$ 表示参照组和对照组下匹配真分数 V 对 Y 的回归. $\forall V$ 没有 DIF 的项目须满足 $E_R[Y\mid V]=E_F[Y\mid V]$，其中 $\forall V$ 表示任意 V 值，即上式要对任意 V 均成立.

在 Rasch 模型条件下，有研究证明了上面定义 1 和定义 2 等价[42]. 但有研究显示在非 Rasch 模型下某些条件下，2 类方法表现明显不同[49]. 对于多级评分，令 $P_{t,g}(\theta)$ 表示在组 g 下能力为 θ 的被试在项目类别反应函数(item category response function，ICRF)，其中 $t=0,1,\cdots,T$，知项目反应函数(itemresponse function，IRF)为 $E_g[Y\mid\theta]=\sum_{t=1}^{T}tP_{t,g}(\theta)$. 如果能判断 2 组被试在某项目(集)上具有完全相同的项目类别反应函数或项目参数(测量不变性)，则项目无 DIF.

张华华等[50]证明了如果 2 个项目服从同一项目反应理论模型，如分部评分模型、拓广分部评分模型或等级反应模型，并且具有相同的项目反应函数，

则这 2 个项目的项目类别数相等,且对应的项目类别反应函数相同. 因此,侦测项目有或无 DIF,只须判断 2 组被试下项目反应函数是否相等. 因此,在常用多级评分模型下,DIF 定义仍与定义 1 或定义 2 相同[46],只是项目反应函数随模型不同而不同.

真分数(true score,$X = V + E$),又称之为条件期望分数或能力对观察分数的回归. 因为项目反应理论模型下单调性假设,知真分数是能力 θ 的严格单调函数[46,51-52],所以真分数(或期望分数)与能力 θ 可以建立一一映射关系. 如果项目满足定义 1,由张华华等所证明的结论[50],相同的项目类别反应函数确定的 $E_R[Y|V]$ 和 $E_F[Y|V]$ 相等,以及真分数与能力 θ 之间的一一映射关系,可知,定义 1 与定义 3 等价. 基于该结论,SIBTEST 就是基于定义 3 而进行假设检验,并且多级评分版的 SIBTEST[46],只需要将 0-1 评分下真分数估计过程中使用的库德-理查德森信度系数,变化成克隆巴赫 α 信度系数即可,极大地简化了方法的复杂性.

2.4 常见的项目功能差异检验方法

2.4.1 MH 和 GMH 方法 1) MH 方法[41-42]是基于观察分数匹配的非参数方法. 该方法源于卡方检验[43,53-56]和基于美国著名生物统计学家 Nathan Mantel 等[57]提出用于匹配研究(the study of matched groups)的卡方检验过程,故此命名. 下面介绍 MH 方法的基本过程.

(i) 基于匹配变量($k = 0,1,\cdots,K$)构建样本或总体列联表(见表 2).

表 2 0-1 评分下基于匹配变量构建样本(总体)列联表

		"研究"项目上得分		
		1	0	总计
组	参照组	$A_k(p_{Rk})$	$B_k(q_{Rk})$	$J_{Rk}(1)$
	目标组	$C_k(p_{Fk})$	$D_k(q_{kj})$	$J_{Fk}(1)$
总计		J_{1k}	J_{0k}	J_k

(ii) 建立 MH 方法检验中的虚无假设,H_0: $p_{Rk}/q_{Rk} = p_{Fk}/q_{Fk}$ 或 H_0: $\alpha_k = p_{Rk}q_{Fk}/(q_{Rk}p_{Fk}) = 1$, $k = 1,2,\cdots,K$,其中 α 表示机率比.

(iii) 构建 MH 方法的检验统计量,$\chi^2_{MH} \equiv \Delta_{\mathrm{MH}} = \left(\left|\sum_{k=0}^{K} A_k - \sum_{k=0}^{K} E[A_k]\right| - 0.5\right)^2 \Big/ \sum_{k=0}^{K} Var[A_k]$,该检验统计量服从自由度为 1 的卡方分布[57],其中 $E[A_K] = J_{Rk}J_{1k}/J_k$, $Var[A_K] = J_{Rk}J_{Fk}J_{1k}J_{0k}/(J_k^2(J_k - 1))$ · $\hat{\Delta}_{\mathrm{MH}}$ 的方差公式、项目反应理论模型下的 MH 检验统计量也有相关研究[41-42].

另外,也可以通过机率比的估计 $\hat{\alpha}_{MH}$ 来得到 $\hat{\Delta}_{\mathrm{MH}}$,公式为 $\hat{\Delta}_{\mathrm{MH}} = -4\ln(\hat{\alpha}_{\mathrm{MH}})/1.7 = -2.35 \cdot \ln(\hat{\alpha}_{\mathrm{MH}})$,其中 $\hat{\alpha}_{\mathrm{MH}} = \left(\sum_{k=1}^{K} A_kB_k/J_k\right) \Big/ \left(\sum_{k=1}^{K} C_kD_k/J_k\right)$. $\hat{\alpha}_{\mathrm{MH}} > 1$ 或 $\hat{\Delta}_{\mathrm{MH}} < 0$ 意味着在研究项目上参照组表现好于目标组.

2) GMH 方法[58-59]是 MH 方法的推广,用于多级评分($y_1,y_2,\cdots,y_T$)下项目 DIF 检验. GMH 方法源于有序反应的卡方检验过程[60]. 下面介绍 GMH 方法的基本过程.

(i) 基于匹配变量($k = 0,1,\cdots,K$)构建样本或总体列联表(见表 3,"+"表示对相应下标求和).

表 3 多级评分下基于匹配变量构建样本(总体)列联表

组别	"研究"项目上得分				总计
	y_1	y_2	…	y_T	
参照组(R)	A_{R1k}	A_{R2k}	…	A_{RTk}	A_{R+k}
目标组(F)	A_{F1k}	A_{F2k}	…	A_{FTk}	B_{F+k}
总计	A_{+1k}	A_{+2k}	…	A_{+Tk}	A_{++k}

(ii) 建立 GMH 方法检验中的虚无假设,H_0: $p_{Rk}/q_{Rk} = p_{Fk}/q_{Fk}$ 或 H_0: $\alpha_k = p_{Rk}q_{Fk}/(q_{Rk}p_{Fk}) = 1$, $k = 1,2,\cdots,K$,其中 α 表示机率比.

(iii) 构建 GMH 方法的检验统计量,$\chi^2_{GMH} = \left(\sum_{k=0}^{K} F_k - \sum_{k=0}^{K} E[F_k]\right)^2 \Big/ \sum_{k=0}^{K} Var[F_k]$,该检验统计量服从自由度为 $T - 1$ 的卡方分布,其中 $F_k = \sum_{t=1}^{T} y_tA_{Ftk}$, $E[F_k] = \sum_{t=1}^{T} y_t \frac{A_{F+k}A_{+tk}}{A_{++k}} = \frac{A_{F+k}}{A_{++k}} \sum_{t=1}^{T} y_tA_{+tk}$, $Var[F_k] = \frac{A_{R+k}A_{F+k}}{A^2_{++k}(A_{++k} - 1)} \left[A_{++k} \sum_{t=1}^{T} y_t^2A_{+tk} - \left(\sum_{t=1}^{T} y_tA_{+tk}\right)^2\right]$. 若 $T = 2$,即 $y_1 = 1, y_2 = 0$,则 $\chi^2_{GMH} = \chi^2_{MH}$.

2.4.2 STD 和 SMD 方法 1) STD 方法是一种用于 SAT 项目公平性侦测的一种早期方法[61-63].

$$STD = \sum_{k=0}^{K} J_{Fk}(P_{Fk} - P_{Rk}) \Big/ \sum_{k=0}^{K} J_{Fk}$$

,其中 $P_{Fk} = A_{Fk}/J_{Fk} = \sum_{j \in J_{Fk}} Y_{Fj}/J_{Fk}$, $P_{Rk} = C_{Rk}/J_{Rk} = \sum_{j \in J_{Rk}} Y_{Rj}/J_{Rk}$, J_{gk} 表示参照组($g = R$)或目标组($g = F$)中在有效项目集上总分 k 的被试集合,$J_{gk} = |J_{gk}|$ 为参照组或目标组中在有效项目集上总分 k 的被试数. STD 处于 -0.05 至 0.05 可忽略; STD 小于 -0.1 或大于 0.1 则需要仔细检查项目; 其他值没有太大的影响[64].

2) STD 方法的多级评分版本[62],又称为 SMD 方法[46,58-59]: $STD \equiv SMD = \sum_{k=0}^{K} \frac{A_{F+k}}{A_{F++}} \left(\frac{1}{A_{F+k}} \sum_{t=1}^{T} y_tA_{Ftk} -$

$\frac{1}{A_{R+k}}\sum_{t=1}^{T} y_t A_{Rtk}$).

2.4.3　LR 和 LDFA 方法　1) LR 方法[65] 即逻辑斯蒂克回归方法. LR 方法主要克服了以下问题：基于对数线性模型方法[43] 忽视能力为有序变量本质的不足；而基于项目反应理论模型方法虽视能力为有序变量，但是又对样本数和模型拟合敏感；MH 方法并不适合侦则非一致性 DIF. 该方法基于逻辑斯蒂克回归模型，既考虑了能力为有序变量的事实，且能用于侦测一致性 DIF 和非一致性 DIF. MH 方法对应随机区组设计模型，而 LR 方法对应协方差模型，MH 方法只看成是 LR 方法的特例，其中 MH 采用离散能力水平，且不考虑能力与组别的交互作用. LR 方法的步骤如下.

(i) 估计参数列向量 $\boldsymbol{\tau} = [\tau_0, \tau_1, \tau_2, \tau_3]^{\mathrm{T}}$ 和方差-协方差矩阵 $\boldsymbol{\Sigma}$. 给定数据下似然函数为

$$L(D_{ata} \mid \boldsymbol{\tau}) = \prod_{j=1}^{J} P(U_j)^{U_j} [1 - P(U_j)]^{1-U_j},$$

其中 $J = J_R + J_F$，$P(U_j = 1) = \frac{e^{Z_j}}{1 + e^{Z_j}}$，$Z_j = \tau_0 + \tau_1 X_j + \tau_2 G_j + \tau_3 (X_j G_j)$，已知数据 $D_{ata} = (\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{G})$ 含所有被试某个项目的得分向量、匹配的能力或总分向量和被试所在组别向量. 参数向量可通过极大似然估计方法[66] 估计 $\hat{\tau}$，估计量的方差-协方差矩阵 $\boldsymbol{\Sigma}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^{\mathrm{T}}}\right]$. 由极大似然估计的渐近正态性，知 $\hat{\tau} \sim N(\boldsymbol{\tau}, \boldsymbol{\Sigma})$.

(ii) 建立虚无假设. 在 Z_j 的表达式中，τ_2 表示 2 组被试在项目上表现的平均差异，τ_3 表示组别与能力的交互作用. 根据 DIF 定义，若 $\tau_3 \neq 0$ 时，知项目为非一致性 DIF；若 $\tau_2 \neq 0$，$\tau_3 = 0$ 时，知项目为一致性 DIF. 因此虚无假设为 $H_0: \boldsymbol{C\tau} = 0$ $v.s$ $H_1: \boldsymbol{C\tau} \neq 0$，其中 $\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

(iii) 构建检验统计量为：$\chi^2 = \hat{\boldsymbol{\tau}}' \boldsymbol{C}' \cdot (\boldsymbol{C\Sigma C}')^{-1} \boldsymbol{C}\hat{\boldsymbol{\tau}}$. 该检验统计量服从自由度为 2 的卡方分布，当检验统计量值超过自由度为 2 的卡方分布的上 α 分位数时（$\chi^2_{2,\alpha=0.05} = 5.99$），则拒绝原假设，认为项目存在 DIF. 如果项目仅存在一致性 DIF，由于失去一个自由度，上面检验过程的检验力会受到影响. 如果根据上面方法识别出了项目存在 DIF，可借助卡方统计量[67] 或效应量[68] 进一步确定项目是存在一致性 DIF 还是非一致性 DIF. 下面仅给出前者.

建立 3 个嵌套模型：$Z_j^{(1)} = -\tau_0 - \tau_1 X_j$（模型 1、零模型或虚无模型）、$Z_j^{(2)} = -\tau_0 - \tau_1 X_j - \tau_2 G_j$（模型 2）、$Z_j^{(3)} = -\tau_0 - \tau_1 X_j - \tau_2 G_j - \tau_3 (X_j G_j)$（模型 3）. 然后分别估计并得到各个回归模型参数向量（$\hat{\boldsymbol{\tau}}^{(1)}$，$\hat{\boldsymbol{\tau}}^{(2)}$，$\hat{\boldsymbol{\tau}}^{(3)}$）及其对数似然函数值 $L^{(1)} = L(D_{ata} \mid \hat{\boldsymbol{\tau}}^{(1)})$，$L^{(2)} = L(D_{ata} \mid \hat{\boldsymbol{\tau}}^{(2)})$ 和 $L^{(3)} = L(D_{ata} \mid \hat{\boldsymbol{\tau}}^{(3)})$，然后计算自由度均为 1 的对数似然比卡方检验统计量值 $G^2_{Uniform} = \chi^2_{Uniform} = -2\ln(L^{(1)}/L^{(2)}) = -2(\ln L^{(1)} - \ln L^{(2)})$ 和 $G^2_{Non\text{-}Uniform} = \chi^2_{Non\text{-}Uniform} = -2(\ln L^{(3)} - \ln L^{(2)})$，其中 $\chi^2_{Uniform}$ 用于判断项目是否存在一致性 DIF，$\chi^2_{Non\text{-}Uniform}$ 用于判断项目是否存在非一致性 DIF.

2) LDFA 方法[44] 即逻辑斯蒂克判别函数分析方法. LDFA 方法是一种以观察分数为匹配变量以模型为基础的参数 DIF 检测方法. 该方法是在 LR 方法的基础上，并在克服多级评分下逻辑斯蒂克回归法（P-LR）缺点的过程中发展起来的，用于多级评分情形. 多级评分下逻辑斯蒂克回归法，尽管可以采用不同编码方式，如连续比率 logits（Continuation-ratio logits）、累积比率 logits（cumulative logits）、或毗邻比率 logits（adjacent-ratio logits），对多级评分（$y_1, y_2, \cdots, y_T$）建模和 DIF 检验，但是一个项目，通常会产生 $T-1$ 个回归方法，并且需要进行多次假设检验并汇总结果. 显然，无论此 3 种中的何种编码方式均将大大地增加模型数量、计算工作量等，并且分开估计得到的假设检验结果解释较为困难，这在实际运用极为不便. 正因为如此，尽管 P-LR 方法在鉴别不同种类 DIF（尤其是非一致性 DIF）上存在优势，但在实际工作中却很少被采用[31,44].

在 LR/P-LR 方法中，将被试作答反应变量视 $\boldsymbol{U}$ 为随机变量，而将被试匹配变量（能力或总分）$\boldsymbol{X}$ 和群组变量 $\boldsymbol{G}$ 视为固定变量，即估计 $Prob(\boldsymbol{U} \mid \boldsymbol{X}, \boldsymbol{G})$. 但变量 $\boldsymbol{U}$ 的多个取值，使得方程需重新编码而十分麻烦，这一点成为了其在实际工作中运用的最大障碍. 尽管 $\boldsymbol{G}$ 是固定变量，$\boldsymbol{U}$ 是随机变量，仍然可以建立逻辑斯蒂克回归方法来估计 $Prob(\boldsymbol{G} \mid \boldsymbol{X}, \boldsymbol{U})$. $Prob(\boldsymbol{G} \mid \boldsymbol{X}, \boldsymbol{U})$ 其实是判别分析中的后验概率的逻辑斯蒂克形式. Miller 等提出 LDFA 方法建立的模型如下：$L(D_{ata} \mid \boldsymbol{\tau}) = \prod_{j=1}^{J} Prob(G_j \mid X_j, U_j)^{G_j} \cdot [1 - Prob(G_j \mid X_j, U_j)]^{1-G_j}$，其中 $J = J_R + J_F$，$Prob(G_j \mid X_j, U_j) = \frac{e^{(1-G_j)Z_j}}{1 + e^{Z_j}}$，$Z_j = -\tau_0 - \tau_1 X_j - \tau_2 U_j - \tau_3 (X_j U_j)$，已知数据 $D_{ata} = (\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{G})$ 含所有被试某个项目的得分向量、匹配的能力向量（或匹配的总分向量）和所在组别.

在这一回归方程中,因变量为群组变量 G,其取值为 1 表示参照组,取值为 0 表示目标组. 参数列向量 $\boldsymbol{\tau} = [\tau_0, \tau_1, \tau_2, \tau_3]^{\mathrm{T}}$ 为回归系数向量. 系数 τ_2 用来检测一致性 DIF,τ_3 用来检测非一致性 DIF. 值得注意的是,在 LR 方法中,被试作答反应变量 U 是作为因变量的,所以其取值受到限制,必须是二分的. 而在 LDFA 中,作答反应变量 U 为自变量,因此其取值可以是多个等级. 这样就解决了 P-LR 中多等级题须重新编码麻烦. 对该方程的求解可以采用极大似然法获得.

LDFA 方法关于 DIF 的显著性检验类似于 LR 方法,在此不再赘述. 如果多级评分项目存在 DIF,说明该项目至少在某个分数等级上存在 DIF,可以进一步通过绘制零模型的 $Prob(\boldsymbol{G} = 0 \mid \boldsymbol{X})$ 变化趋势图、全模型的 $Prob(\boldsymbol{G} = 0 \mid \boldsymbol{X}, \boldsymbol{U} = \boldsymbol{u}_t)$ 变化趋势图及 95% 置信区间图,如果某分数段上全模型图偏上(偏下),则说明该项目对目标组(参照组)中此分数段上的被试更为有利.

2.4.4 SIBTEST 和 P-SIBTEST 方法 1) SIBTEST[68] 是一种计算简单的非参数方法,并不使用项目反应函数、潜在能力估计,并提供一种显著性检验的方法. 若测验中项目数 N,得分随机向量为 $\boldsymbol{U} = (U_1, U_2, \cdots, U_N)$. 不仿假设前 K 个项目组成的集合为有效项目集,而剩下的项目组合的集合为探查项目集;U_{gij} 表示组 g 中被试 j 在有效项目 i 上的得分;Y_{gj} 表示组 g 中被试 j 在待探查项目集上的总分为 $Y_{gj} = \sum_{i=K+1}^{N} U_{gij}$;$X_{gj}$ 表示组 g 中被试 j 在有效项目集上的得分和 $X_{gj} = \sum_{i=1}^{K} U_{gij}$;$\overline{U}_{gi} = \sum_{j \in J_{gk}} U_{gij}/J_g$ 表示组 g 中参加测验的所有被试在项目 i 上的得分率;$\overline{U}^{*}_{gi} = \max\left((\overline{U}_{gi} - c)/(1-c)\right), 0)$,其中 c 为猜测参数;$\overline{Y}_{gk} = \sum_{j \in J_{gk}} Y_{gj}/J_{gk}$ 表示给定匹配变量条件下($k = 0, 1, \cdots, K$),参照组或目标组上待探查项目集上总分的条件样本均值,即参照组或目标组分别对应 $\overline{Y}_{Rk}$ 和 $\overline{Y}_{Fk}$,$\hat{p}_{gk} = J_{gk}/J_g$ 表示组 g 中在有效项目集上总分 k 的被试频率.

(i) 建立 SIBTEST 方法检验中的虚无假设(一般采用单边假设检验,以推断目标组是否存在有偏差,即不利于对照组的偏差):$H_0: \beta_U = 0$ $vs.$ $H_1: \beta_U > 0$.

因为真值未知,首先需要对真实全局偏差量 β_U 进行估计. 通常认为在有效项目集上总分相同的被试具有十分接近的目标能力,在偏差侦查中,这些被试的总分具有直接可比性. 将 $\overline{Y}_{Rk} - \overline{Y}_{Fk}(k = 0, 1, \cdots, K)$ 作为目标组上偏差的衡量指标. 按频率 $\hat{p}_k = \hat{p}_{Fk}$ 将观察偏差进行加权求和,即得到偏差估计量 $\hat{\beta}_U$:$\hat{\beta}_U = \sum_{k=0}^{K} \hat{p}_k(\overline{Y}_{Rk} - \overline{Y}_{Fk})$.

由目标能力分布差异引起偏差的问题,基于有效项目集上观察分数匹配的偏差估计量 $\hat{\beta}_U$,并不能十分有效地将目标能力分布差异从测验偏差中分离出来. 目标能力分布差异只会引起 $\hat{\beta}_U$ 增大,而形成测验偏差的假象. 即使测验没有偏差,能力分布差异将导致偏差估计量 $\hat{\beta}_U$,最终增加一类错误率. 由此可以看出,基于有效项目集上观察分数匹配划分参照组和目标组的被试,对于在有效项目集上分数相同的 2 个被试组,在目标能力分布而不尽相同. 因此,有必要对目标能力分布差异进行修正.

(ii) 估计真分数. 可使用回归方法[68] 得到真分数的估计 $\overline{Y}^{*}_{gk}$,替换原来观察分数 $\overline{Y}_{gk}$. 该方法主要认为有效项目集上分数 X 是真分数 V 的组成成分,并假设 X 对 V 的回归是线性关系 $X = V + E$,直接采用 $E[V \mid X = k]$ 作为回归估计,并记 $\hat{V}_g(k)$ 为 $E[V \mid X = k]$ 的估计,采用 $\hat{V}_g(k)$ 处一阶泰勒展开式(或中值定理)近似估计 $\overline{Y}^{*}_{gk}$:$\overline{Y}^{*}_{gk} = \overline{Y}_{gk} + \hat{M}_{gk}(\hat{V}(k) - \hat{V}_g(k))$,其中 $\hat{M}_{gk} = (\overline{Y}_{g,k+1} - \overline{Y}_{g,k-1}) \Big/ (\hat{V}_g(k+1) - \hat{V}_g(k-1))$,$\hat{V}(k) = (\hat{V}_R(k) + \hat{V}_F(k))/2$,$\hat{V}_g(k) = (\overline{X}_g + n(1 - (\hat{\sigma}^2(e \mid g))/(\hat{\sigma}^2(X \mid g)))(k - \overline{X}_g)/(n-1))/n$,$\hat{\sigma}^2(e \mid g) = \sum_{i=1}^{K} \overline{U}^{*}_{gk}(1 - \overline{U}^{*}_{gk})$,$\hat{\sigma}^2(X \mid g) = \sum_{j=1}^{J_g} (X_{gj} - \overline{X}_g)^2/(J_g - 1)$,$k = 0, 1, \cdots, K$;$g$ 为 R 或 F.

(iii) 构建检验统计量. $\overline{Y}^{*}_{Rk}$ 和 $\overline{Y}^{*}_{Fk}$ 可以看成对应有效项目集上同一真分数估计 $\hat{V}(k)$ 处的 2 组被试在探查项目集上的真分数估计. 也是原来是采用有效项目集上相同观察分数 k 匹配,而现在通过有效项目集上同一真分数估计 $\hat{V}(k)$ 进行匹配. 匹配有效项目集上真分数估计 $\hat{V}(k)$,得到下面的真实偏差 β_U 的无偏估计量 $\hat{\beta}^{*}_{U}$ 及其检验统计量:

$$\hat{\beta}^{*}_{U} = \sum_{k=0}^{K} \hat{p}_k(\overline{Y}^{*}_{Rk} - \overline{Y}^{*}_{Fk}),$$

$$B = \sum\nolimits_{k=0}^{K} \hat{p}_k(\overline{Y}^{*}_{Rk} - \overline{Y}^{*}_{Fk}) \Big/ \left(\sum_{k=0}^{K} \hat{p}_k^2\left(\frac{1}{J_{Rk}}\hat{\sigma}^2(Y \mid k, R) + \frac{1}{J_{Fk}}\hat{\sigma}^2(Y \mid k, F)\right)\right)^{1/2},$$

其中组 g 的样本条件方差分别为 $\hat{\sigma}^2(Y \mid k,g) = S_{gk}^2 = \frac{1}{J_{gk}-1}\sum_{j\in J_{gk}}(Y_{gj}-\bar{Y}_{gk})^2$.

最后只需要比较 B 与标准正态分布上 α 分位数 z_α，如果 $B > z_\alpha$（如 $z_{0.05} = 1.96$），则拒绝原假设；否则，没有充分证据拒绝原假设. 对于 SIBTEST 具体计算过程，在此不再作更详细的叙述. 需要注意的是，由于考虑在两端分数的目标能力估计误差（较大）、各划分组被试数量下限（如 30）、线性插值、p_k 的不同估计 $\hat{p}_k$ 等问题，在 SIBTEST 具体计算过程中，与上面列出的公式不完全相同，详情可参见文献[68] 附录.

2) Poly-SIBTEST 方法是 SIBTEST 方法的多级评分版本[46]. 由于证明了定义 1 和定义 3 的等价，因此 Poly-SIBTEST 方法与 SIBTEST 方法的计算过程基本类似，需将 SIBTEST 方法中库德-理查德森信度系数改为克隆巴赫 α 信度系数. 其研究结果显示：(i) 对于无 DIF 的待探查项目，在待探查项目的平均区分度与匹配测验项目的平均区分度有较大差异时，GMH 和 SMD 拒真率较高，同时 GMH、SMD 的拒真率随样本量增加而增大. 而 Poly-SIBTEST 的拒真率对区分度和样本量的依赖程度十分小，接近设定的显著性水平值 0.05，即 Poly-SIBTEST 对违反 Rasch 模型条件更稳健；(ii) 对于有 DIF 的待探查项目，在区分度较大且固定难度偏移量时，项目的偏差较大，各方法的拒绝率均较高，GMH、SMD 方法稍高于 Poly-SIBTEST；而对于区分度较小时，GMH、SMD 方法犯错的概率高于 Poly-SIBTEST，即 GMH、SMD 方法的拒绝率低于 Poly-SIBTEST 的拒绝率.

2.5　各方法的优势与不足

MH/GMH 方法优势与不足：MH 方法并不能用于侦测非一致性 DIF；MH 方法以总分作为匹配变量，如果作答反应数据拟合复杂的项目反应理论模型，MH 方法倾向于将没有 DIF 的项目误认为有 DIF[46-47]；但易受目标组和参照组的能力分布差异的影响.

STD/SMD 方法优势与不足：STD 与 MH 方法表现相当，并且存在与 MH 相同的不足. STD/SMD 方法用于组别上选项选择差异和未完成项目速度差异[64]，有助于确定产生 DIF 的原因并进一步提纯匹配变量[47]；但易受目标组和参照组的能力分布差异的影响.

LR/LDFA 方法优势与不足：LR 方法与 MH 探查一致性 DIF 基本相当；LR 方法可较好侦测非一致性 DIF[65]；LR 方法以总分作为匹配变量，存在与 MH 和 STD 类似问题；LR 方法可以在回归方法中添加其他自变量或协变量，以探查引起 DIF 的原因[44,47,69]；LDFA 对项目数没限制，可用于探查项目集 DIF，但是有待进一步研究[44]；除非进行测验等值，否则易受目标组和参照组的能力分布差异的影响.

SIBTEST/Poly-SIBTEST 方法优势与不足：基于真分数估计的 SIBTEST 方法的表现，并不受目标组和参照组的能力分布差异的影响；SIBTEST 能对由于多个项目引起的测验偏差进行评价，即使这些项目中的单个项目含相同方向的较小偏差，SIBTEST 仍可较好地识别项目集的偏差；SIBTEST 方法对违反 Rasch 模型条件更稳健；SIBTEST 是一种计算简单的非参数方法，并不使用项目反应函数、潜在能力估计，能较好地侦测单向性偏差（比一致性偏差条件弱）；但并不能侦测非一致性 DIF.

3　推动考试公平的具体举措

在提出高厉害考试面临的公平性问题和详细介绍了测验公平性评价的统计方法之后，提出促进考试公平的详细而可行举措，以供读者借鉴.

1) 科学命题是考试公平的根本. 命题和审题是一项长期而艰难的工作，美国 ACT 的技术报告称一份新试卷开发需耗时 2.5 年[25]. 在严格命题和审题规范指导下，要进行科学命题以保证试题对各群体考生的公平性，需要保证命题和审题专家具有广泛的代表性和独立性. 由于诸多高厉害考试过度保密，极易造成命题和审题专家小组所代表的群体过小，从而编制出对一部分考生有利的试题. 在注重公平性的同时，可采取命题专家开发少量试题，以减少泄题的风险，保证高厉害考试的安全性.

2) 科学评分和科学分析是考试公平的保障. 考试公平性对于考试开发者和使用者无疑是十分必要的. 考试公平性作为信条历史悠久. 宋代创设了糊名法和誊录法[70]. 试卷交上来后，先由弥封官将卷面折叠，封藏应试者的姓名，编上红号；然后由誊录人员将试卷用朱笔滤誊写，称为“朱卷”，将它送考官评阅. 放榜的时候，按取中的“朱卷”红号调取“黑卷”拆封，最后唱名写榜. 其目的在于避免阅卷者通过考生信息和考生字迹徇私舞弊，这样可以较好地控制评分效应.

许多考试开发者、老师或家长认为，不同群体的学生参加同一份考试，考试分数就可以公平地比较.

从测量学来看,事实并非如此. 需要“数据”和“专家”来“说话”,分析试卷和试题质量和公平性. 必须使用项目功能差异方法,以探查试卷和试题是否对某些群体比较有利,如探查《给违反交规父亲一封信》作文题是否对乡村或欠发达地区考生不公平. 美国 ACT 的公平性报告称[24],2011—1012 年 ACT 考试进行了 10 320 次项目功能差异检查,标记了 132 个考题并反馈给专家再审. 不同卷考生分数比较,需要借助于测量学中连接设计和等值技术等.

虽然我国教育测量领域的学者层出不穷,也有众多享誉世界的研究成果,但理论与实践之间却始终存在着巨大空白. 大型公共考试的数据大多封锁,也导致专家们无法用科学方法客观测量考试的公平性,用数据分析结果提高下一次考试公正性,使得考试公平问题成为亘古难题;考题的产生和筛选缺乏科学理论支撑,缺少严格的流程监控,导致考后争议频发.

3) 科学决策是考试公平的重点. 科学决策离不开各种测量误差控制,如主观题评分松严、合格分数线处测量误差、分数转换误差等控制. 通常要求各分数段考生的测量误差基本相同. 各种测量误差的控制与最终决策密切相关,直接影响考试结果的公平性. 高考主观题评分的网上阅卷日趋成熟. 高考一本、二本、三本等各分数线的测量误差如何,是否基本一致,是否控制在一定的范围之内? 通俗地讲,高考这把尺子能否公平地对待不同分数段的考生? 有报告显示,国际学生评估项目 PISA,2009 年考试对低水平考生的测量精度不高.

4) 科学研究是考试公平的关键. 培养并建立专业的测量学研究团队,并与国内外测量学团队合作,基于考试大数据,针对性开展高厉害考试理论、方法和技术等研究,加速新方法和技术的诞生和应用,从而促进考试公平. 高厉害考试对各个考生群体的公平性、高厉害考试的计算机化和纵向比较、高考分数对大学一年级成绩的预测效果、高考分数和其他背景信息辅助新生专业和课程选择等,都需要进行长期研究. 国际许多考试机构,具有庞大的研究团队,发表了大量高质量研究成果并迅速应用到考试中.

除 SIBTEST 方法[68]、LR[65]、LDFA[44] 由伊利诺伊大学厄巴纳-香槟分校、马萨诸塞大学和 ACT 研究人员提出外(其实 WilliamStout 和 Hariharan Swaminathan 同时也是 ETS 的十分重要的研究员),本文介绍的其他项目功能差异方法均是 ETS 研究人员提出,如 MH 方法由 ETS 的 Paul W. Holland 于 1985 年提出[41],GMH 由[58-59,71] 由 ETS 的 Rebecca Zwick 于 1993 提出,STD 和 SMD 由 ETS 的 Neil J. Dorans 等于 1983 和 1991 年提出[62-63]. MH 方法的前身卡方方法的提出者 Janice D. Scheuneman[53,55]、多级评分 SIBTEST 方法的提出者 Hua-hua Chang[46] 均曾经在 ETS 工作过.

在国内,对于考试的公平性这一领域的探索,早在几十年前就有大批优秀专家学者投入其中,并且做出了杰出贡献. 项目功能差异的理论和应用,由于直接关乎考生成绩,对于考试实施的成功与否影响甚大,自 20 世纪 80 年代初至今,都是学者们经久不衰的研究话题,每年的美国教育研究协会年会,作为教育测量学领域影响力和参与度最大的会议,甚至单独分出一个版块供学者们讨论交流项目功能差异. 可见如何用科学的方法促进考试公平,早已成为并一直是测量学中研究热点之一.

另外,在实际 DIF 分析应用中仍存在诸多需要深入研究的问题[7,71-72],如匹配标准“纯化”问题、假设检验、统计模型(HLM)[69]、项目和测验功能差异(differential functioning of items and tests,DFIT)、专家评判一致性等问题. 还有,项目功能差异对测验效度的影响[73],如项目功能差异主要关注 DIF 项目集对组的公平性的影响,而 DIF 项目集对个体的公平性的影响有待考虑.

5) 科学施测是考试公平的途径. “互联网 +”测评已经成为未来的趋势[74],“互联网 +”测评将在 DIF 原因探讨、测验安全性、测验效率等方面更好地保证测验公平性. “美国士兵职业倾向成套测验”、“美国研究生入学考试”、中国军队入伍考试和全国计算机等级考试等均已采用计算机化考试. 据“美国新闻和世界报导”2015 年 5 月报道,近 2 年10 000 名学生将试行美国 ACT 高考. 2015 年 11 月,国务院副总理刘延东指出,教育信息化成绩显著,“宽带网络校校通”、“优质资源班班通”、“网络学习空间人人通”取得突破性进展. 2015 年 12 月“计算机化考试”写入美国的“每一个学生成功法案”.

2015 年 2 月,教育部办公厅印发《2016 年教育信息化工作要点》,指出 2016 年教育信息化工作目标之一是实现全国中小学互联网接入率达到 95%. 习近平总书记强调,“没有信息化就没有现代化”. 伴随着教育信息化、云计算、大数据技术、移动互联网和人工智能的发展,考试信息化迫在眉睫,“互联网 +”测评即将走入大众视野,将给个性化学习提供新的契机. 特别是,依托国家题库的建立,高厉害考试计算机化将给考试公平性、考试安全性注入新的活力.

简言之，在让农村、边远、贫困和少数民族地区的孩子们共享优质教育资源、健全完善公平公正入学制度、实施“支援中西部地区招生协作计划”等，迈出教育公平重大步伐的同时，借助专家、科学方法和技术，保证考试内容和结果的公平性，做好考试科学性、公正性这道时代命题，永远不能停下脚步. 因为，考试（特别是高考）的背后，是万千学子改变命运的可能，是民众对于社会公平正义的信心. 希望考试更加公平，切实让“寒门学子”成就梦想，促进教育公平和社会公平.

4　参考文献

[1] OECD. Equity and quality in education: Supporting disadvantaged students and schools [EB/OL]. [2016-07-26]. https://www.oecd.org/education/school/50293148.pdf.

[2] UNESCO. Education 2030: Incheon declaration and framework for action-towards inclusive and equitable quality education and lifelong learning for all. [EB/OL]. [2016-07-26]. http://www.uis.unesco.org/Education/Documents/incheon-framework-for-action-en.pdf.

[3] 王旖旎. 教育测评中的不公正问题：项目功能差异[J]. 中国远程教育，1999(8)：39-41，63.

[4] 中国教育学会教育测量与统计分会. 项目功能差异[J]. 中国考试，2003(Z4)：51.

[5] Zieky M. Fairness reviews in assessment [C] // Downing S M，Haladyna T M. Handbook of test development，lawrence erlbaum associates. Inc: Mahwah，NJ，2006.

[6] Dorans N J，Sinharay S. Looking back: proceedings of a conference in honor of Paul W. Holland [M]. New York，NY: Springer，2011.

[7] Osterlind S J，Everson H T. Differential item functioning [M]. 2nd. Thousand Oaks，CA: SAGE Publications，Inc，2009.

[8] Holland P W，Wainer H. Differential item functioning [M]. New York: Routledge，Taylor & Francis Group，1993.

[9] Association A E R，Association A P，Education N C o M i. Standards for educational and psychological testing [M]. ishington，DC: AERA，1999.

[10] Association A E R，Association A P，Education N C o M i. Standards for educational and psychological testing [M]. ishington，DC: American Educational Research Association，2014.

[11] Psychology S f I a O. Principles for the validation and use of personnel selection procedures [M]. Bowling Green，OH: Society for Industrial and Organizational Psychology，Inc，2003.

[12] 刘铁川，戴海琦，赵玉. 现代测量理论观点下的测验偏差评价[J]. 中国临床心理学杂志，2012，20(3)：346-349.

[13] Jehangir K，van den Berg S M，Glas C A W. Correcting for differential item functioning in multi-level regression models in cross-national surveys [J]. Measurement，2015，66：263-271.

[14] Lee H，Geisinger K F. The matching criterion purification for differential item functioning analyses in a large-scale assessment [J]. Educational and Psychological Measurement，2015，76(1)：141-163.

[15] Tay L，Huang Q，Vermunt J K. Item response theory with covariates (IRT-C)：assessing item recovery and differential Item functioning for the three-parameter logistic model [J]. Educational and Psychological Measurement，2015，76(1)：22-42.

[16] Koo J，Becker B J，Kim Y S. Examining differential item functioning trends for English language learners in a reading test: a meta-analytical approach [J]. Language Testing，2013，31(1)：89-109.

[17] Latifi S，Bulut O，Gierl M，et al. Differential performance on national exams: evaluating item and bundle functioning methods using english，mathematics，and science assessments [J]. SAGE Open，2016，6(2)：1-14

[18] Cavanagh A，Wilson C J，Caputi Petal. Symptom endorsement in men versus women with a diagnosis of depression: a differential item functioning approach [J]. International Journal of Social Psychiatry，2016，62(6)：549-559.

[19] Chalmers R P，Counsell A，Flora D B. It might not make a big DIF: improved differential test functioning statistics that account for sampling variability [J]. Educational and Psychological Measurement，2015，76(1)：114-140.

[20] Berger M，Tutz G. Detection of uniform and nonuniform differential item functioning by item-focused trees [J]. Journal of Educational and Behavioral Statistics，2016，41(6)：559-592.

[21] ETS. ETS standards for quality and fairness [M]. Princeton，NJ: Educational Testing Service，2014.

[22] ETS. ETS guidelines for fair tests and communications [M]. Princeton，NJ: Educational Testing Service，2015.

[23] ETS. ETS international principles for fairness review of assessments [M]. Princeton，NJ: Educational Testing Service，2009.

[24] ACT. Fairness report for the ACT ® tests [M]. Iowa City，IA: ACT，Inc，2012.

[25] Testing A C. Technical manual the ACT ® [M]. Iowa City，IA: ACT，Inc，2014.

[26] SBAC. Smarter balanced assessment consortium: 2014-2015 technical report [M]. Los Angeles，CA: Smarter Balanced Assessment Consortium，2016.

[27] PARCC. PARCC accessibility features and accommodaions

manual 2016-2017 [M]. Parcc Inc. ishington,DC: PARCC Assessment Consoria,2016.

[28] 曹亦薇,张厚粲. 汉语词汇测验中的项目功能差异初探[J]. 心理学报,1999,31(4):460-467.

[29] 曹亦薇. 项目功能差异在跨文化人格问卷分析中的应用[J]. 心理学报,2003,35(1):120-126.

[30] 相阳. 利用数理统计方法进行测验偏差分析[J]. 数学的实践与认识,1993(3):26-33,98.

[31] 宋丽红. LDFA 方法及其在项目功能差异分析中的应用研究:以高考英语试卷分析为例[D]. 南昌:江西师范大学,2008.

[32] 柴省三. 汉语水平考试(HSK)阅读理解测验公平性研究[J]. 语言文字应用,2013(4):107-116.

[33] 黄春霞. 第二语言学习者专业背景对 HSK 阅读成绩影响的项目功能差异检验[J]. 考试研究,2011(5):59-66.

[34] 肖园园. 大学英语四级考试对不同学术背景和不同性别学生的项目功能差异研究[D]. 广州:广东外语外贸大学,2013.

[35] 张颖,赵世明. 医师资格考试中的项目功能差异研究[J]. 中国考试,2004(10):23-26.

[36] 李现文,刘海宁,安静. 老年抑郁量表城乡项目功能差异分析[J]. 中国全科医学,2016,19(9):1002-1005.

[37] 耿亮,竺培梁. 情绪智力量表(EIS)中文版的项目功能差异分析[J]. 外国中小学教育,2008(9):42-46.

[38] 肖影影,毕重增,狄轩康. 一般自我效能感量表的性别与跨文化项目功能差异分析[J]. 心理研究,2013,6(5):38-41.

[39] 王蕾,黄晓婷. 国际教育成效评价协会儿童认知发展状况测验项目功能差异分析[J]. 考试研究,2006,2(4):94-107.

[40] 朱乙艺,韦小满. 我国成就测验的项目功能差异研究述评[J]. 教育与考试,2012(1):78-81.

[41] Holland P W. On the study of differential item performance without IRT [C]//Proceedings of the 27th Annual Conference of the Military Testing Association, San Diego, CA: 1985,282-287.

[42] Holland P W, Thayer D T. Differential item functioning and the Mantel-Haenszel procedure [C]// Wainer H, Braun H I. Test validity, L. Erlbaum Associates: Hillsdale, NJ,1988: 129-145.

[43] Mellenbergh G J. Contingency table models for assessing item bias [J]. Journal of Educational Statistics,1982,7(2):105-118.

[44] Miller T R, Spray J A. Logistic discriminant function analysis for DIF identification of polytomously scored items [J]. Journal of Educational Measurement,1993,30(2):107-122.

[45] Dorans N J, Potenza T M. Equity assessment for polytomously scored items: a taxonomy of procedures for assessing differential item functioning (Research Rep. RR-94-49) [M]. Princeton, NJ: Educational Testing Service, 1994.

[46] Chang Hua-hua, Mazzeo J, Roussos L. Detecting DIF for polytomously scored items: an adaptation of the SIBTEST procedure [J]. Journal of Educational Measurement, 1996,33(3):333-353.

[47] Millsap R E, Everson H T. Methodology review: Statistical approaches for assessing measurement bias [J]. Applied Psychological Measurement,1993,17(4):297-334.

[48] 张龙,涂冬波. 多级计分题项目功能差异常用检测方法及比较[J]. 江西师范大学学报:自然科学版,2015,39(5):441-448.

[49] Roussos L A, Stout W F. Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance [J]. Journal of Educational Measurement,1996,33(2):215-230.

[50] Chang Hua-hua, Mazzeo J. The unique correspondence of the item response function and item category response functions in polytomously scored item response models [J]. Psychometrika,1994,59(3):391-404.

[51] Chang Hua-hua. A note on the monotonicity of the IRFs for polytomous IRT models [M]. Princeton, NJ: Educational Testing Service,1994.

[52] Lord F M. The relative efficiency of two tests as a function of ability level [J]. Psychometrika,1974,39(3):351-358.

[53] Scheuneman J. A method of sssessing nias in test items [J]. Journal of Educational Measurement,1979,16(3):143-152.

[54] Baker F B. A criticism of Scheuneman´s item bias technique [J]. Journal of Educational Measurement,1981,18(1):59-62.

[55] Scheuneman J D. A response to Baker´s criticism [J]. Journal of Educational Measurement,1981,18(1):63-66.

[56] Marascuilo L A, Slaughter R E. Statistical procedures for identifying possible sources of item bias based on χ^2 statistics [J]. Journal of Educational Measurement,1981,18(4):229-248.

[57] Mantel N, Haenszel W. Statistical aspects of the analysis of D_{ata} from retrospective studies of disease [J]. Journal of the National Cancer Institute,1959,22(4):719-748.

[58] Zwick R, Donoghue J R, Grima A. Assessing differential item functioning in performance tests (Research Rep. RR-93-14) [M]. Princeton, NJ: Educational Testing Service,1993.

[59] Zwick R, Donoghue J R, Grima A. Assessment of differential item functioning for performance tasks [J]. Journal of Educational Measurement,1993,30(3):233-251.

[60] Mantel N. Chi-square tests with one degree of freedom ex-

tensions of the mantel-haenszel procedure [J]. Journal of the American Statistical Association, 1963, 58(303): 690-700.

[61] Dorans N J, Kulick E. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test [J]. Journal of Educational Measurement, 1986, 23(4): 355-368.

[62] Dorans N J, Schmitt A P. Constructed response and differential item functioning: a pragmatic approach (Research Rep. RR-91-47) [M]. Hillsdale, NJ: Educational Testing Service, 1991.

[63] Dorans N J, Kulick E. Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in december 1977: An application of the standardization approach (Research Rep. RR-83·9) [M]. Princeton, NJ: Educational Testing Service, 1983.

[64] Dorans N J, Schmitt A P, Bleistein C A. The standardization approach to assessing comprehensive differentiali item functioning [J]. Journal of Educational Measurement, 1992, 29(4): 309-319.

[65] Swaminathan H, Rogers H J. Detecting differential item functioning using logistic regression procedures [J]. Journal of Educational Measurement, 1990, 27(4): 361-370.

[66] Jodoin M G, Gierl M J. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection [J]. Applied Measurement in Education, 2001, 14(4): 329-349.

[67] Zumbo B D. A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores [M]. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999.

[68] Shealy R, Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias DTF as well as item bias/DIF [J]. Psychometrika, 1993, 58(2): 159-194.

[69] Noortgate W V d, Boeck P D. Assessing and explaining differential item functioning using logistic mixed models [J]. Journal of Educational and Behavioral Statistics, 2005, 30(4): 443-464.

[70] 吉丽. 科举考试公平公正研究 [J]. 扬州大学学报: 高教研究版, 2011, 15(1): 28-32.

[71] Zwick R. A review of ETS differential item functioning assessment procedures: flagging rules, minimum sample size requirements, and criterion refinement (Research Rep. RR-12-08) [M]. Princeton, NJ: Educational Testing Service, 2012.

[72] Karami H, Nodoushan M A S. Differential item functioning (DIF): current problems and future [J]. International Journal of Language Studies, 2011, 5(3): 133-142.

[73] Salehi M, Tayebi A. Differential item functioning: implications for test validation [J]. Journal of Language Teaching and Research, 2012, 3(1): 84-92.

[74] 张华华，汪文义. “互联网 +”测评自适应学习之路 [J]. 江西师范大学学报: 自然科学版, 2016, 40(5): 441-455.

A Practical View of Test Fairness to Improve Equity in Education from Statistical Measurement

WANG Wenyi[1], CHANG Hua-hua[2,3*]

(1. College of Computer Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China;
2. Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA;
3. Faculty of Education, East China Normal University, Shanghai 200062, China)

Abstract: If the result of a test is unfair, it will affect the fairness of the educational opportunity and social fairness. Statistical analyses for test fairness in our country has been neglected and even ignored for a long time. The purpose is to give a review of key aspects concerning differential item/test functioning from the perspective of statistical measurement. Finally, regarding the problem of the fairness of testing in the context of high-stakes test use, some detailed and practical suggestions for test fairness are presented for readers´ reference.

Key words: the fairness of testing; equity in education; differential item functioning; statistical measurement; the national college entrance examination

(责任编辑: 冉小晓)