现代测量理论观点下的测验偏差评价

刘铁川[1]，戴海琦[1]，赵玉[2]
(1.江西师范大学心理学院，江西 南昌 330022；2.赣南医学院心理学系，江西 赣州 341000)

【摘要】 测验在当前社会广泛应用的同时，其公平性受到了社会各界的广泛关注。具备公平性的测验应是无偏差的。随着测量理论的快速发展，目前已经出现多类测验偏差评价技术用以维护测验公平，而国内测验研究与实践中所使用的方法却相对滞后。本研究从现代测量学的角度，介绍了评价测量偏差、预测偏差、等值偏差方法的最新进展，并给出了使用建议。这些方法关注测验偏差的不同角度，但紧密相联。我国各行业的测验工作者应充分利用这些理论技术来指导测验的编制、使用，以促进测验的公平性。

【关键词】 测验公平；测验偏差；项目功能差异；预测偏差；群体不变性

中图分类号：R395.1　　文献标识码：A　　文章编号：1005-3611(2012)03-0346-04

Assessment of Test Bias in the Perspective of Modern Measurement Theory

LIU Tie-chuan, DAI Hai-qi, ZHAO Yu
School of Psychology, Jiangxi Normal University, Nanchang 330022, China

【Abstract】 While tests are widely used in our society, its fairness has been concerned by every social class. Fair test should be unbiased. Currently, with the rapid development of the measurement theory, there have been a variety of methods for assessing bias for maintaining test fairness. Domestic research and practice of testing lag behind the current research of test bias. The present paper has reviewed new advances of methods for assessing measurement bias, predicting bias and equating bias through the perspective of modern measurement theory. Suggestions of using these methods in practice are also given. These methods are concerned about test bias in different perspectives, though closely related. Researchers in any area of using tests and examinations should take advantage of these theories and techniques to guide the development and use of tests for promoting the test fairness.

【Key words】 Test fairness; Test bias; Differential item functioning; Predictive bias; Population invariance

我国是一个测验使用大国，广泛应用于教育考试、资格认证、医学诊断等各个领域，测量结果对被测者的生活、教育与职业决策等影响很大。所以，人们非常关注这些测量工具的质量。测验质量的一个重要方面就是公平性(Fairness)。美国教育研究协会、美国心理学会以及美国教育测量学会所公布的《教育与心理测量标准》[1]以及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》中都有专章探讨测验公平问题[2]。后文将简称为《标准》和《原则》。在我国，《国家中长期教育改革和发展规划纲要》也强调考试招生制度改革“维护社会公平的原则”。

可以看出，测验公平性是个备受各界关注的测量学问题。近年来，此领域的研究发展较快，而国内相关研究与实践中所使用的方法却并未做出相应变革，事实上，只有极少数国内考试项目与测验编制研究采用项目功能差异报告了其公平性方面的质量，几乎没有测验报告其预测偏差和等值偏差(后文介绍)。本文将结合最新研究进展，从测验偏差检测的各个方面入手，探讨用以维护测验公平的各类偏差评价技术。

【基金项目】 本研究得到高等学校博士学科点专项科研基金联合资助课题(20103604110002)；国家自然科学基金(31100756)；江西省高校人文社会科学研究青年基金(XL1003)资助
通讯作者：戴海琦

测验公平(Fairness)有多种定义，一般来讲，公平的测验应在测验的设计、开发、施测、计分、报告、解释等方面对所有受测者一致，以保证受测者的测验分数差异体现了所测构念(Construct)水平的差异，而非其他无关因素(Irrelevant Factors)所导致。因此，测验公平性与效度关系密切，甚至有研究者直接将测验公平定义为，在测验开发与使用的各个阶段对所有子群体而言具有相同的效度[3]。测验是否公平涉及价值判断，通过实证研究较为困难，测量学中操作方案是通过分析测验是否存在偏差(Bias)来检验测验的公平性。为保障测验公平，测验专家通过审查整个测试过程努力保证能力水平相同的被试得到相同的测验分数。但实际上，测验仍可能对特定性别、文化背景、民族、经济背景的被试群体更有利，此时认为测验存在偏差，这是一种测量过程中的系统误差。正如《标准》所指出，公平测验的特征之一就是无偏差，在任何情况下都应尽量使用无偏差的测验[1]。

现代测量学认为，测验偏差表现为测验分析所使用的各种模型在不同人群中模型参数也存在差异。测量模型中，如项目反应理论(Item Response Theory，IRT)模型中题目的难度、区分度参数对不同子群体差异显著，则认为存在测量偏差；对预测模型，若回归方程的截矩、斜率对子群体差异显著，则认为存在预测偏差；在等值计算之中，若等值系数对子群体差异显著，则认为存在等值偏差。下面分别探讨此三类测验偏差评价技术。

1 测量偏差

测量偏差(Measurement Bias)关注的是测验的心理计量特征(如题目难度、区分度、正确作答概率)是否存在群体间差异[2],此时只涉及测验分数(或由 IRT 模型估计出的能力参数)与人口统计变量,而不涉及效标变量,所以有研究者称为测验偏差检验的内部方法(Internal Methods)[4]。若测验无测量偏差,则称测验具备测量不变性(Measurement Invariance)。

可以通过项目功能差异 (Differential Item Functioning, DIF)与多组验证性因素分析(Multi-Group Confirmative Factor Analysis)检验测验题目是否存在测量偏差。DIF 分析通常根据人口统计变量将被试分成两组, 一组为焦点组(Focus Group,F),题目对其不利;另一组为参照组(Reference Group, R),题目对其有利。可通过条件概率来理解 DIF:

$$P(Y_j=1 \| \theta, G=F)=P(Y_j=1 \| \theta, G=R) \quad (1)$$

上式中,θ 为匹配变量,G 为分组变量,Y_j 为题目 j 作答反应,1 为正确,0 为错误。条件概率可以根据实际数据选用不同的测量模型进行计算。若上式成立,表示被试能力水平为 θ 的被试,无论是来自组 F 还是组 R,其答对题目 j 的概率都相同,题目 j 无测量偏差,否则存在测量偏差。若两组被试在题目上的条件概率大小关系与 θ 取值大小有关,称为非一致性 DIF (non-uniform DIF), 否则称为一致性 DIF(uniform DIF)。

DIF 检验一般经过以下几个基本步骤。

第一步,选择匹配变量。根据测验开发、质量分析所使用的测量学理论,匹配变量可以选择使用测验总分、或 IRT 模型估计出的潜在特质水平。

第二步,检查匹配变量的信度。为保证检验结果的可靠性,匹配变量的信度应较高。对于测验总分,可使用 alpha 系数等指标检查测验分数的信度;对于潜在特质水平值,可检查其估计标准误。

第三步,根据匹配变量,选择不同的检验方法,逐个题目进行检验。

若以观察分数为匹配变量, 可选择 MH 检验、Logistic 回归、SIBTEST 等方法。研究显示,对于一致性 DIF,MH 检验比 Logistic 回归要更有效,而对于非一致性 DIF,Logistic 回归的检验力比 MH 检验要更大[5]。SIBTEST 方法的优势在于通过回归方法校正了观察分数的测量误差,而且同时可以检验题组的功能差异(Item Bundle)[6]。

若以潜在特质水平为匹配变量,在 DIF 检验前还要通过等值将焦点组的题目与能力参数转换到参照组的量尺之上,然后可选择 Lord 卡方统计量、Raju 面积测度法、似然比检验等方法进行 DIF 检验。研究显示,三种方法的检验结果比较一致[7]。但 Lord 卡方统计量与 Raju 面积测度的扩展性不如似然比检验方法,尤其是 Raju 面积测度法不仅在三参数模型下难以求积,其取值也会受到数量分布较少的被试条件概率差异的较大影响[8]。

第四步,对 DIF 检验显著的题目进行内容分析。通过学科专家、命题专家对存在项目功能差异的题目的题干材料、提问方式、选项等的细致分析,确定是否存在与测量目标无关的其它因素导致不同组作答结果的差异。

第五步,删除存在测量偏差的题目,重复进行第三步和第四步,直至所有题目无测量偏差。

测量偏差检验通常逐个题目进行,有些题目可能对参照组有利,有些题目可能对焦点组有利,但对整个测验而言,各个题目的偏差效应可能相互抵消, 导致整个测验无测量偏差。所以也需要检验整个测验是否存在功能差异(Differential Test Functioning,DTF),可以通过 DFIT 方法进行检验[6]。另外, 大多数 DIF 分析都是比较两组被试在题目上的作答表现,但实际上人群可能分成 2 个以上的组别,此时可以使用多层次 IRT 模型(Multilevel IRT)将组别作为第三层,设置题目参数的随机效应,用来进行 DIF 分析[9]。

根据单一方法的检验结果难以做出题目是否存在测量偏差的准确决策,实践中应综合考虑多种 DIF 方法的检验结果与专家判断做出最终决策。目前能同时完成多种 DIF 分析方法的程序不多,这也是测验偏差研究较少的重要原因。建议使用 R 语言的 difR 4.1 程序包[10],其功能相对而言比较全面。

2 预测偏差

用于选拔的测验通常假定测验能够相对准确地预测被试在今后学习工作中的实际表现。测验的预测能力也是考查测验效度的重要手段。然而,在不同的人群中,测验对效标的预测功能可能存在群体间差异(Differential Prediction),称之为预测偏差(Predictive Bias)[2]。这也是一种测验偏差,此时除测验分数与人口统计变量外,还要使用效标变量,所以有研究者称之为测验偏差检验的外部方法(External Methods)[4]。

传统的预测偏差分析做法是通过允许回归方程的斜率、截矩存在差异来检验预测偏差。具体而言,通过层次回归的做法依次纳入测验分数、分组变量、以及二者的交互作用为预测变量,以效标分数为因变量,检验分组变量及交互作用项的显著性。若分组变量显著,则截矩参数存在组间差异;若交互作用项显著,则斜率参数存在组间差异[11]。对美国医学院入学考试的研究显示,以学分绩点为效标的回归方程的截矩和斜率系数在性别、种族间均存在差异,这意味着使用唯一的回归方程会产生预测偏差[12]。

附表 截矩参数差异的各种原因的可能影响

原因	导致测验均值差异?	导致效标均值差异?	测验存在问题?
测验测量偏差	是	否	是
效标测量偏差	否	是	否
遗漏重要预测变量	否	是	否
测验信度	是	是	否

传统的预测偏差分析认为,无论是分组变量显著还是交互作用项显著,测验均存在预测偏差。斜率参数存在组间差异表示的是测验预测效度不同,必然导致预测偏差,但近期研究显示,截矩参数的差异可能是由测验的测量偏差、效标的测量偏差、测验信度、遗漏重要的预测变量等多种原因造成,并不一定是由测验本身存在问题[13]。若测验不存在问题,

则不影响测验使用，只需要使用不同的回归方程进行预测即可，所以若截矩参数存在显著差异，必须仔细探查其原因。附表呈现了各种情况的可能结果。

因此，为排除以上混淆因素，应使用如下步骤来检验截矩参数差异的具体原因[14]。

第一步，作测验分数对分组变量的回归，或对测验的组间差异进行 t 检验，并计算测验 X 分数差异的效应量 d_x。

第二步，作效标分数对分组变量的回归，或对效标的组间差异进行 t 检验，并计算 Y 测验 Y 效标的效应量 d_y。

第三步，作效标分数对分组变量、测验分数、以及二者交互作用的层次回归。交互作用项显著，表示各组回归方程斜率不同，测验存在预测偏差；交互作用项与分组变量均不显著，测验无预测偏差；交互作用项不显著，但分组变量显著，需要进一步分析：①若测验分数无组间差异，而效标分数存在组间差异，表示测验本身无问题，可能是效标测验存在测量偏差或遗漏重要预测变量导致的预测功能差异，可以使用 DIF 检验方法或加入其它预测变量进一步探查原因；②若测验分数存在组间差异，而效标分数无组间差异，表示测验存在测量偏差，从而导致预测功能差异；③若测验分数与效标分数都存在组间差异，此时需要比较效标效应量观察值 d_y 与其期望值 $\hat{d}_Y$ 的大小。$\hat{d}_Y$ 的计算公式如下[14]：

$$\hat{d}_Y = \frac{2r_{YX} \cdot d_X}{\sqrt{r_{XX} \cdot (d_X^2 + 4) - (r_{YX} d_X)^2}} \quad (2)$$

上式中，为 r_{XX} 测验信度，r_{YX} 为测验与效标的相关系数。当 $d_Y < \hat{d}_Y$，可能是测验的测量偏差导致效标的期望效应量更大，而观察值较小；当 $d_Y > \hat{d}_Y$，可能是效标测量偏差或遗漏了重要预测变量造成了二者的差异；当 $d_Y = \hat{d}_Y$，此时可能是由于测验信度导致截矩的组间差异。

因此，只有在斜率参数存在差异或由于测验的测量偏差造成截矩参数存在差异的情形下，需要停止使用测验，而其他原因造成的预测功能差异并不一定是由于测验本身存在质量问题，在预测各组的效标分数时使用不同的回归方程即可避免预测偏差。需要指出的是，在各种情形下，都应使用 DIF 检验方法确认测验与效标是否存在测量偏差以得到更准确的预测偏差分析信息。

3　等值偏差

等值(Equating)是实现平行测验分数相互比较的重要技术。锚测验不等组设计(NEAT)是等值最常用的一种数据搜集方法，这种设计通过锚题来链接两个平行测验。近年来，研究者对锚测验代表性、不同等值方法的稳健性进行了大量研究，花费了大量时间和精力收集测验数据以进行等值，但很少有实践者关注等值过程是否真正实现了测验分数可比。下面介绍关于锚题参数漂移和群体不变性假设违反导致等值偏差研究的新进展。

虽然锚题是经过精挑细选的优良试题，但研究显示，若锚题受到了与测量目标无关的外部因素的影响，如题目曝光、作弊、两次的评分标准不同、印刷错误等，就可能导致量尺的不稳定[15]。具体表现为同一锚题，但其参数不同，测量学中称之为项目参数漂移(Item Parameter Drift，IPD)[16]。有研究者通过模拟作答数据考察了平均数/标准差法、特征曲线法和同时校准三种等值方法的表现，发现随着发生难度参数下降的锚题数量的增加，三种方法都有明显的等值偏差[17]。所以，在正式开展等值工作之前，应移除参数漂移锚题。具体来说，可以通过比较平均数/平均数法和平均数/标准差法两种等值结果的差异或 DIF 方法来进行，在 Rasch 模型下也可以使用“0.3 难度单位”进行筛选。

等值是刻画两个平行测验对应关系的函数，应与被试选取无关。但 Flanagan 指出，测验的本质是抽样，在内容上的差异(包括平行测验)可能会使等值函数对不同群体的敏感性存在差异[18]，从而破坏等值的重要假设——群体不变性(Population Invariance，PI)。等值的群体不变性在评估考试分数的公平性时起着重要的作用，在其不成立时使用统一的等值函数对被试而言是不公平的。在期望均方根差异 (Root Expected Mean Square Difference，REMSD)[19]这一分析等值群体不变性的量化指标提出后，近年来此领域的研究才开始逐渐增多，有研究者对美国大学预修课程考试[20]、学术性向测验[21]、法学院入学考试[22]等大型考试等值或链接的群体不变性进行了分析。REMSD 计算方法如下：

$$REMSD = \frac{\sqrt{\sum_j w_j \sum_{x=\min(x)}^{\max(x)} v_{xj} [e_{P_j}(x) - e_P(x)]^2}}{\sigma_{YP}} \quad (3)$$

上式中，w_j 为期望权重，表示对子群体与全部被试等值关系差异的权重。期望权重可采用子群体被试比例，此时应保证样本数据中的被试比例结构能够代表测验的真实使用情境，也可以进行等量加权；v_{xj} 为题目分数权重，有三种选择，其一是分数在子群体中的比例，其二是分数在总人群中的比例，其三是等量加权。公式 3 只适用于单组或等设计下的观察分数等值，Von Davier，Holland 和 Thayer 将 REMSD 推广至 NEAT 设计下的等值群体不变性评价，但关注的仍是观察分数等值[23]。目前，von Davier 和 Wilson 已将其成功用于不等组设计下 IRT 真分数等值的群体不变性研究[20]。目前生活质量评估等大型题库基本都通过基于 IRT 的等值方法构建，REMSD 指标的推广对于检验等值假设是否成立和提高题库质量具有极大的实际意义。

以往 REMSD 指标分数权重的选择取决于研究者的经验。笔者的模拟研究发现，期望权重与分数权重都采用等权时的 REMSD 会更大，子群体分数分布无差异时分数权重采用子群体比例还是总体比例时的 REMSD 计算结果一致，所以一般情况下分数权重应使用子群体比例加权。需要指出的是，群体不变性反映的是子群体间等值函数的差异大小程度，不是成立或者不成立的概念。一般而言，严格按照相同测验蓝图编制的平行测验，在信度相似 (等值的另一重要假设)、子群体平均水平差异不大的情况下等值函数在子群体间不会有太大差异。但对信度相似、测量目标不同的链接，群体不变性通常不成立。

前文对锚题参数与群体不变性的分析介绍不涉及具体的等值方法，也就是说，不论使用何种等值方法，都应检验锚题是否发生参数漂移以及等值函数的群体间差异。理论上，等值函数群体间差异较大时，应使用多个等值函数，但社会难以接受，实践中如何在二者之间取得平衡是需要进一步研究的课题。

4　结　　语

本文从测验偏差检验角度探讨了维护测验公平的各种

技术手段，尽管这些方法关注的角度不同，但相互之间存在紧密联系。如探讨测验的预测不变性时需要对预测测验与效标测验的测量偏差进行检验，因为预测功能差异可能是由于效标测验的测量偏差所导致。对于平行测验的等值，也需要使用测量偏差的检验方法检测锚题是否发生参数漂移，这是因为参数漂移在本质上是纵向的、跨时间的测量偏差，而传统的测量偏差关注的是横向的跨群体的测量偏差。

在实践中研究者使用这些技术时应注意以下几点：①用于等值的锚题不少于全长测验的四分之一，甚至更多，否则删除参数漂移锚题后无法实现准确等值；②要搜集尽可能详细的人口统计信息，测量偏差和等值函数的群体间差异可能体现在各个方面；③应对测验偏差的结果进行细致分析，以指导测验的编制、使用、解释。对测量偏差，应分析题目内容的文化差异、教学重点等方面的可能原因。对预测偏差，应分析是否遗漏重要的预测变量、以及预测测验和效标测验的质量。对锚题参数漂移，应从文化变迁、课程改革等方面分析原因。对等值的群体间差异，应从两测验的内容差异、信度水平差异、子群体间差异等方面进行分析。

参 考 文 献

1 American Educational Research Association/American Psychological Association/National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC, American Educational Research Association, 1999

2 Society for Industrial and Organizational Psychology. Principles for the validation and use of personnel selection procedures. Bowling Green, OH, Society for Industrial and Organizational Psychology, 2003

3 Xiaoming Xi. How do we go about investigating test fairness. Language Testing, 2010, 27(2): 147–170

4 Camilli G, Shepard LA. Methods for identifying biased test items. CA: Sage. Thousand Oaks, 1994

5 Hidalgo MD, López-Pina JP. Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. Educational and Psychological Measurement, 2004, 64(6): 903–915

6 Penfield RD, Camilli G. Differential item functioning and item bias. In Rao CR, Sinharay S. Handbook of statistics: Vol. 26. Psychometrics. Amsterdam: Elsevier, 2007. 125–168

7 Seock-Ho K, Cohen AS. A comparison of Lord's Chi-Square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. Applied Measurement in Education, 1995, 8(4): 291–312

8 Camilli G. Test fairness. In Brennan RL. Educational measurement (4th edition). CT: ACE/Praeger. Westport, 2006

9 Pastor DA. The use of multilevel item response theory modeling in applied research: An illustration. Applied Measurement in Education, 2003, 16(3): 223–243

10 Magis D, Beland S, Tuerlinckx F, De Boeck P. A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 2010, 42(3): 847–862

11 Cleary TA. Test bias: Prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement , 1968, 5(2): 115–124

12 Kyei-Blankson, Lydia S. Predictive validity, differential validity, and differential prediction of the subtests of the medical college admission test. PhD dissertation, Ohio University, United States, 2005

13 Meade AW, Tonidandel S. Not seeing clearly with cleary: What test bias analyses do and do not tell us. Industrial and Organizational Psychology, 2010, 3(2): 192–205

14 Meade AW, Fetzer M. Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. Organizational Research Methods, 2009, 12(4): 738–761

15 Kim W, Nering M. Evaluation of equating items using DFIT. Annual meeting of the national council on measurement in education. Chicago, IL, 2007

16 Bock R, Muraki E, Pfeiffenberger W. Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 1988, 25(4): 275–285

17 Huiqin H, Rogers WT, Vukmirovic Z. Investigation of IRT-based equating methods in the presence of outlier common items. Applied Psychological Measurement, 2008, 32 (4): 311–333

18 Kolen MJ. Population invariance in equating and linking: Concept and history. Journal of Educational Measurement, 2004, 41(1): 3–14

19 Dorans NJ, Holland PW. Population invariance and the equatability of tests: Basic theory and the linear case. Journal of Educational Measurement, 2000, 37(4): 281–306

20 von Davier AA, Wilson C. Investigating the population sensitivity assumption of item response theory true-score equating across two subgroups of examinees and two test formats. Applied Psychological Measurement, 2008, 32(1):11–26

21 Yi Q, Harris DJ, Gao X. Invariance of equating functions across different subgroups of examinees taking a science achievement test. Applied Psychological Measurement, 2008, 32(1): 62–80

22 Liu M, Holland PW. Exploring population sensitivity of linking functions across three law school admission test administrations. Applied Psychological Measurement, 2008, 32 (1): 27–44

23 von Davier AA, Holland PW, Thayer DT. The chain and post-stratification methods for observed-score equating and their relationship to population invariance. Journal of Educational Measurement, 2004, 41(1): 15–32

(**收稿日期**:2011-11-28)