项目功能差异在临床问卷分析中的应用

刘曦[1,2]，张建新[1]
（1.中国科学院心理研究所，北京 100101；2.中国科学院研究生院，北京 100101）

【摘要】 在跨文化问卷分析中，对项目功能差异的分析是保证测量公平性和有效性的重要依据。与经典测量理论相比，在测查项目功能差异时，运用项目反应理论有较强的前提假设和优越性。本研究对项目功能差异的相关概念进行解释；介绍项目功能差异的常用分析方法；简介国内外有关项目功能差异在跨文化临床问卷分析中的应用。

【关键词】 临床问卷；项目功能差异；跨文化

中图分类号：R395.1 文献标识码：A 文章编号：1005-3611(2006)04-0349-03

Application of Differential Item Functioning in Clinical Questionnaire

LIU Xi, ZHANG Jian-xin
Chinese Acedemy of Sciences, Institute of Psychology, Beijing 100101, China

【Abstract】 In the cross-cultural psychology, the Differential item functioning is an important evidence to assure the equitableness and validity of the measurement. Item response theory is considered as one of the best hypothesis to detect the DIF compared with the classical measurement theory. This study explains the related conceptions of DIF. The common method of DIF has also be introduced. At last, the application of Differential item functioning in Clinical questionnaire at home and abroad is discussed.

【Key words】 Clinical questionnaire；Differential item functioning；Cross-culture

项目功能差异（Differential item functioning，简称DIF）是指，当来自两个不同团体、但特质相等的个体选择某一测验项目的同一答案时，他们的应答概率之间存在差异。既是说，两组受测者选择某一项目答案的概率差异，除了受到他们自身心理特质水平的影响之外，还会受到项目本身因素的影响，从而在同一项目上表现出不同的特征曲线[1]。

对项目功能差异的研究最早始于20世纪60年代。人们越来越关注测验的性别、种族或民族公平性问题。心理测验对于不同受测者团体是否公平、有效，这不仅仅是心理计量学的研究范畴，而且还涉及到社会其他领域。

1 项目功能差异的主要概念

1.1 项目偏差、项目功能差异

项目偏差定义为：如果具有相同能力或熟练程度的不同团体受测者在回答某个项目上的正确率存在差异，且这种差异不是由测验所测能力、而是由其它与测验对象无关的因素引起，则该项目就存在偏差[2]。项目偏差和测验偏差的研究最早始于1910年，Binet将智力测验中这些差异反应归结为测题的偏差，而不是智力本身的问题[3]。但项目偏差不仅显示了测验分数与真分数之间的差异，而且含有受测者的主观判断和评价，故不适于用作统计学指标。所以，Angoff提出用项目功能差异（DIF）的概念来代替项目偏差[4]。因此，项目功能差异只是项目偏差的必要条件，而不是充分条件，仅表示该项目的答案正确率在不同团体受测者间存在着差异或者相对难度[5]。

1.2 参照组和目标组

进行DIF分析时，通常只对两组受测者进行分析，把研究目标群体称为目标组，而把另一群体称为参照组。

1.3 一致性DIF和非一致性DIF

在80年代以前，人们只发现一种一致性DIF，当受测者的能力水平与其组别之间不存在交互作用时，项目的DIF即为一致性DIF。也就是说，在所有能力水平上，一组受测者回答某一项目的正确率总大于另一组。但后来研究者们又发现，受测者能力与项目功能差异并不总是平行的。当受测者能力水平与分组（如性别组）间有交互作用时，项目的DIF就表现为非一致性DIF[6]。

2 项目功能差异常用分析法

目前探索DIF的统计方法主要分为两大类：一类是参数方法（项目反应理论和逻辑斯蒂回归方法）；另一类是非参数方法（MH方法，标准化方法和SIBTEST方法）。其中MH、SIBTEST、LRDIF和标准化方法主要适用于“0”、“1”记分的项目，而IRT和多重LRDIF法则适用于多重记分的项目。

研究中被广泛使用的是参数方法，项目反应理论（Item response theory，简称IRT）又被译作潜在特质理论（Latent Trait Theory）、或者项目特征曲线理论（Item Characteristic Curve Theory）。IRT是对受测者能力的一种估计，并将受测者对单个测验项目的反应概率与三参数logistic曲线图的一定特征联系起来。这一理论是由潜在特质、项目特征曲线等基本概念组成。潜在特质是指受测者某种相对比较稳定的、支配其做出测验反应、并使其反应表现一致的内在特征。由于特质不能直接被观察到，所以被称为潜在特质（多用θ表示），是测验所要测量的目标。项目特性曲线描述受测者测验得分与受测者内在特质（能力）水平之间的关系，它能统一地将答对概率与受测者特质水平和项目功能（质量）有机地结合起来。

项目反应理论（IRT）为DIF分析提供了一个良好的理论

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

框架。项目特征曲线（ICC）不应受样本的影响，即估计出的项目参数值应该是一致的。而如果参数值不一致，就要考虑该项目是否存在 DIF。DIF 的定义是：如果参照组和目标组之间的项目参数或项目特征曲线存在差异，则该项目具有 DIF；如果不存在差异，则该项目没有 DIF。用 IRT 法计算 DIF，通常都是先分别计算参照组和目标组的项目参数，然后把参数都放在同一尺度上，比较两者之间的差异。在 IRT 方法中，最简单的 DIF 形式表现为只有项目难度参数（b）在存在差异；当然，项目功能差异也可能表现为 a（区分度）、b（难度）和 c 参数（猜测度）中任何一个存在着差异，或者它们当中的两个或三个参数同时存在差异。

因为项目反应理论建立在严谨的数学统计模式基础上，它借助于电脑科技在近一、二十年取得了突飞猛进的发展，因此当代很多研究者都根据 IRT 理论来测查项目的 DIF。测查项目功能差异的 IRT 法主要是通过软件 MULTILOG 程序来实现。我国的学者也在 DIF 研究方面进行了初步的探讨。漆书清等人[7]经过多年的研究，编制出《现代教育与心理测量通用分析系统》（1.6 版本），可用于进行有关项目功能差异 DIF 的分析。

3　项目功能差异在跨文化临床量表分析中的应用

国外关于项目功能差异在跨文化问卷分析中的研究，主要集中在三个方面，临床量表、智力量表和人格问卷。参照组和对照组多以性别、国籍、种族来划分。

目前国外已有很多关于测查临床量表的项目 DIF 的研究报告。例如，使用 Thalbourne 躁狂抑郁量表（MDS）分别对澳大利亚和美国受测者群体进行测量，然后对该量表进行项目功能差异测查。结果发现：在同等抑郁水平上，女性比男性更为担心贫穷问题。这一结果排除了女性的抑郁反应仅仅是反复出现的强烈消极情绪的副产物的预假设。美国受测者在同等的 MDS 抑郁水平上，男性比女性更少报告自己的抑郁症状。针对 MDS 在性别和文化方面存在 DIF 的项目，研究者制定出一套只有 12 个项目的修订版 R-MDS。比较 R-MDS 和原版 MDS 的结果发现，R-MDS 能使测验在性别上的项目功能差异大大降低[8]。

为了检查不同种族间抑郁症状在表现上的变化，有学者分析了在美国出生的黑人、西班牙人、不在美国出生的西班牙移民、以及非西班牙白人群体在流调中心抑郁量表（CES-D）上产生的项目功能差异。DIF 分析发现，非西班牙白人和其他三个群体之间在 CES-D 的一半项目上都存在着 DIF[9]。

Stricker、Reynolds 和 Willson 用偏相关法对考夫曼的儿童攻击性评估量表（K-ABC）问卷进行 DIF 分析，探察该问卷与整体能力无关的项目平均数在种族（黑人和白人）和性别上的差异。在 10 种 K-ABC 的测验上，基于 2932 个受测者的比较，确定出在各种测验当中有 23 个项目在黑人上有偏差，有 13 个对白人有偏差。在考虑性别变量的时候，12 个项目在男性上有偏差，16 个项目在女性上有偏差。无论是在性别或者是在种族变量上出现的偏差项目，其正确应答概率都超过了偶然机率水平。总的来说，这些有偏差项目对组间总分的差异存在着不恰当的影响，使 K-ABC 表现出了明显的种族和性别系统偏差[10]。

Christopher 等对酒精依赖量表（ADS）进行了深入分析。研究抽取了高危饮酒人群中的 101 名男性和 93 名女性。对 ADS 使用最大似然率因素分析表现出一维的因素结构。然后为 ADS 的项目建立了项目特征曲线（ICC）。确定出 25 个项目中的 9 个项目可以把有酒精过量或酒精滥用现象的群体从无酒精问题或极少酒精问题的群体中明确的区分出来。在性别变量上，项目功能差异并没有被检查出来，证明这些项目在评估男性或女性时功能是一致的。研究证明从 ADS 中缩减出来的 9 个源于经验的项目在样本中是有效的。它们与原量表的相关很高，并有很高的因子负载，证明这 9 个项目有良好的心理测量特性[11]。

贝克抑郁量表在临床诊断上受到广泛的运用，有研究测查晚期病人和中期病人在贝克抑郁量表（BDI）上的作答是否存在与年龄相关的测量差异。在研究中使用项目反应理论模型来平衡这两个测量样本之间因为量表带来的差异和年龄的组间差异。研究的 BDI 的数据由 218 个晚期病人和 613 个中期病人组成。结果表明低抑郁水平的晚期病人更倾向于报告较少的感知觉症状。相反地，高抑郁水平的病人更倾向于报告较多的躯体症状[12]。

美国的学者对最新翻译的西班牙语的创伤后压力障碍量表-普通译本（PCL-C）和已经出版的英文版本这两种测量版本进行了等值性评估。运用项目反应方法作为检验等值性的工具，研究中还具体讨论了运用项目反应理论所面临的基本原则和实际问题。研究中的被试由西班牙母语被试（N=120 人）和英语母语被试（N=284 人）组成，调查他们在社会暴力事件后的心理健康结果。项目功能差异的分析结果表明这两个版本并不是完全等值的，第 6 至第 17 个项目表现出了一致性的 DIF，所以这些学者推测两个不同语言的版本并不完全等值[13]。

许多研究者认为无论是同质量表或异质量表之间都存在着测量偏差，他们关心如何使用项目反应理论的方法把真实的组间差异从由量表带来的偏差中分离出来。明尼苏达多项个性检查表（MMPI）包括 12 个一维的因子量表（Waller, 1999）和 13 个多维的效度量表和临床量表。研究者回顾了有关 MMPI 的偏差研究，使用逻辑斯蒂双参数模型对加里福利亚大学 511 名黑人和 1277 名白人学生在 MMPI 的各因子、效度量表和临床量表的得分来进行了项目功能差异和测验功能的分析。结果证明了由具有偏差的项目组成的同质量表或异质量表不一定在测验总分上也表现出偏差[14]，认为 MMPI 可以被推广使用。

精神变态检查表修订版（PCL-R）在应用和研究领域都是非常重要的一个测量。它的效度指标多来源于对男性白种人的研究。使用验证性因素分析和项目反应理论分析两种方法比较了 359 名美国白种人和 356 名在美国的非洲移民。使用 IRT 方法在 PCL-R 的 5 个项目上发现了显著的项目功能差异，但测验整体的功能差异没有被发现，这表明项目间的功

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

能差异彼此抵消了。所以得出结论 PCL-R 被认为对美国的非洲移民没有偏差,可以被广泛使用[15]。

除了临床方面的 DIF 研究以外,研究者还关注 DIF 在智力量表和人格问卷上的应用。此外,Thissen 等人[16]在人格研究中进行性别的 DIF 分析。Huang,Church 和 Katigbak[17]利用了三种 DIF 技术,跨文化比较分析了著名的大五人格问卷(NEO 英语版),发现了 DIF 的存在。

4 小 结

国外学者已经开发了大量的 DIF 检测方法,在跨文化临床问卷分析中的应用也受到学者们的广泛讨论,并比较了这些方法的优缺点和效率。当前国外对 DIF 的研究主要致力于:①分析方法的开发;②分析方法的比较;③用 DIF 方法对各种测验的项目进行分析,确定有功能差异的项目,分析其产生的原因,以形成可能原因的假设,并加以逻辑论证;④在跨文化问卷的分析中,将常见的组间总分平均值比较深入到项目分析水平上,使得测量工具更具合理性和公平性。但由于 DIF 研究还相对比较新,产生 DIF 的原因很复杂,因此如何利用 DIF 技术,探讨国外引进量表和问卷对中国文化背景的契合程度,并尝试编制适合中国文化的心理测验,是需要我国心理学工作者们继续给予关注的课题。

参 考 文 献

1 Thissen D, Steinberg L, Wainer H. Detection of Differential Item Functioning Using the Parameters of Item Response Models. Differential Item Functioning,1993. 67-117

2 Thelander H, et al. Learning disabilities associated with lesser brain damage. Journal of Pediatrics, 1958, 53:23-29

3 李伟明. 重视对项目差异功能(DIF)的分析和研究. 中国考试,1998, 4:13-16

4 Angoff WH. Use of difficulty and discrimination indices for detecting item bias. In RA Berk(Ed). Handbook of Methods for Detecting Test Bias.Baltimove: John Hopkins Univ. Press, 1982. 96-116

5 曹亦薇,张厚粲. 汉语词汇测验中的项目功能差异初探. 心理学报,1999,4(1):460-467

6 曾秀芹,孟庆茂. 项目功能差异及其检测方法. 心理学动态,1999,2:41-57

7 漆书清,等. 现代教育与心理测量学原理. 高等教育出版社,1998

8 Rense Lange, Michael A, James Houran. Depressive response sets due to gender and culture-based differential item functioning. Personality and Individual Differences, 2002, 33:937-954

9 Noboru Iwata, Stephen Buka. Race/ethnicity and depressive symptoms: a cross-cultural ethnic comparison among university students in East Asia, North and South America. Social Science and Medicine, 2002, 12(55):2243-2252

10 Victor L, Rebecca FN. Kamphaus. Race and gender effects on item functioning on the Kaufman assessment battery for children. Journal of School Psychology,1989,3(27):289-296

11 Christopher W, David R, Gregory L. Item functioning of the alcohol dependence scale in a high-risk sample. Drug and Alcohol Dependence, 2003,72(2):83-192

12 Yookyung, Paul A, Ellen. Differential Functioning of the Beck Depression Inventory in late-life patients: use of item response theory. Psychology and Aging, 2002,17(3):379-391

13 Maria Orlando, Grant N. Differential item functioning in a spanish translation of the PTSD checklist detection and evaluation of impact. Psychological Assessment, 2002,14(1):50-59

14 Niels G, Jane S, Ernst. Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous acales: An illustration with the MMPI. Psychological Methods, 2000, 5(1):125-146

15 David J, David S, Christine. Psychopathy and ethnicity: structural, item, and test generalizability of the psychopathy checklist-revised (PCL-R) in Caucasian and African American participants. Psychological Assessment, 2001,13(4):531-542

16 Colom R, García-López. Sex differences in fluid intelligence among high-school graduates. Personality and Individual Differences, 2002, 32(3):445-451

17 Thissen D, Steinberg L,Gerard M. Beyond group-mean difference: The concept of item bias. Psychological Bulletin, 1986,99:118-128

(收稿日期:2005-12-19)

(上接第348页)

基于人性的哲学修订量表已有了明确的双因子的理论模型,本研究选择验证性因素分析,用拟合指数衡量测量数据与量表理论结构的拟合优度。结果显示,该量表模型的拟合指数 RMSEA 和 SRMR 均小于 0.08,这与 Hu 和 Bentler[4]所提出的当 SRMR 小于 0.08 时,模型可以接受的观点相一致。其它拟合指数 GFI、AGFI、NNFI 以及 CFI 和 IFI 均达到 0.90 以上,也证明数据与量表因子模型有较高的拟合优度,支持了量表的双因子模型。通过对效标效度的研究证明,人性的哲学修订量表亦具有较好的效标关联效度,可预测人际信任度的高低。

参 考 文 献

1 汪向东主编. 心理卫生评定量表手册. 中国心理卫生杂志社.1999(增刊):168-180

2 Wrightsman LS.Measurement of human nature.Psychological Reports,1964,14:743-751

3 张雅明,曾盼盼,俞国良. 中小学生心理健康量表的信效度检验. 中国临床心理学杂志,2004,12(1):7-8

4 Hu L, Bentler PM. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods,1998,3:424-453

(收稿日期:2005-12-27)

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net