检验项目功能差异的两类方法—CFA 和 IRT 的比较

骆 方 张厚粲
(北京师范大学 心理学院, 100875)

摘 要:目前在验证性因素分析(CFA)和项目反应理论(IRT)两个领域,都有一些检验方法来识别项目功能差异(DIF)。该文主要针对单维的多级计分项目,分别介绍 CFA 和 IRT 检测 DIF 的方法,并进行二者的比较。

关键词 验证性因素分析;项目反应理论;项目功能差异
中图分类号:B841.2 文献标识码:A 文章编号:1003 - 5184(2006)01 - 0074 - 05

1 基本概念

1.1 测量偏差

如果来自不同组的特质(包括能力)水平相同的个体,对某个项目/测验有相同的答对率,那么该测量工具具有等同性(measurement equivalence),否则测验和项目有偏差(bias)。

1.2 项目功能差异

研究者对项目偏差的研究更多关注于性别、种族的分数差异,这就使偏差带有了社会评价意义。后来,研究者使用项目功能差异(differential item functioning,DIF)表示纯粹统计学意义上的偏差。

DIF 被定义为,在控制特质之后,一个项目在不同组中显示的不同统计特性。

DIF 的研究一般针对两个团体,如两种性别、种族以及其它特征的被试进行分析,指目标组和参照组在测验所测的特质相同的前提下,两组在某题上的答对率有差异,这种差异不是由于特质差异引起的,而是由与测验无关的因素引起的。

1.3 项目影响

如果,目标组与参照组在某题上的差异是由于在测验所测的特质上有差异,被称为项目影响(item impact)。传统上,检验两个群体在某个项目或者测验上是否存在差异,比如考察女生是否比男生在语文测验上得分更高,采用方差分析或者 t 检验进行均值检验。这种检验必须保证测量工具具有等同性,既不存在 DIF 才有意义。

2 DIF 的常用检测方法

DIF 方法可分为实际得分方法和潜在特质方法。二者根本区别在于,实际得分方法是直接把测量总分作为匹配变量,而潜变量方法是把潜在特质得分作为匹配变量,潜在特质得分是由实际得分估计出来的。

其中,实际得分方法,包括比较难度的 DIF 方法、卡方方法、标准化方法、SIBTEST 方法、逻辑斯蒂回归方法等[1]。

潜在特质方法,在验证性因素分析(CFA)和项目反应理论(IRT)两个领域,都有一些检验方法。目前多数的 IRT 研究都是针对单维数据的,最新的进展已经可以对多级计分的项目检测 DIF。本文的主要目的是,比较在对多级计分的单维项目的 DIF 检测上 CFA 和 IRT 两类方法的异同。

3 验证性因素分析进行 DIF 检测的方法

3.1 基本模型

验证性因素分析(CFA)的数学表达式为:$X = \Lambda_x\xi + \delta$ (1)

其中,X 为 p 阶的观测变量向量,ξ 是潜在特质变量(单维的),Λx 是 p 阶的因子载荷矩阵,δ 为 p 阶的测量误差向量。

假设误差 δ 的平均数为零,则 $E(x) = \Lambda_x\xi$ (2),也即项目 X 的真分数为 $\Lambda_x\xi$。

3.2 DIF 定义

CFA 定义 DIF 为具有相同特质的两个群体在某项目上真分数不同,由公式(2)可知,如果因素载荷 $\lambda_{iR} = \lambda_{iF}$,则如果具有相同的特质 ξ,真分数 $\Lambda_x\xi$ 相同,反之,如果 $\lambda_{iR} \neq \lambda_{iF}$,则该项目有功能差异(i 为某项目,R 为参照群体,F 为目标群体)[2]。

3.3 检测 DIF 的一般过程

CFA 通过定义嵌套模型,检验 Λ_x 是否存在差异,识别 DIF[3]。

(1) 模型 1(基准模型 baseline model):同时对两组进行分析,模型中的参数自由估计,验证不同组是否具有相同的模型结构;

(2) 模型 2(测量等同模型 measurement equivalence model) 在模型 1 的基础上限定不同组在所有项目上因素载荷相等；

(3) 模型 3(部分测量等同模型 partial measurement equivalence model) 在模型 1 的基础上限定不同组测量部分因素载荷相等；

Vandenberg 和 Lance(2000)[4]建议应在下述情况下部分测量等同模型才成立：(1)只允许极少数题目在两组的对应载荷不同；(2)有充分理论依据支持；(3)在不同样本重复得出相同结果。

由于上面所定义的模型之间具有嵌套结构(一个模型中要估计的参数可以通过其它模型限定得到)，根据 Byrne 和 Shavelson(1987)[5]的方法，可以通过两个嵌套模型卡方值差异($\triangle\chi^2_{\triangle df}$)的显著性来检验模型中的限定条件是否成立，显著性水平可设定为 0.05 或 0.01。如果模型 3 成立，在两个群体中允许因素载荷自由估计的项目，即存在 DIF。

3.4　共同尺度

在 CFA 中为了避免由计数单位引起的不确定性，必须固定公共因素方差或者一个因子负荷为非零值(通常固定为 1)，这样公共因素才具有单位，参数才可以识别。在两个群体中，参数需要放在共同的尺度上才能比较，因而需要首先事先限定在某个项目上两个群体的因素载荷相同或者潜变量方差相同，其它估计参数才能够具有相同的尺度。Reise 等(1993)推荐使用这样的限定：固定目标组的方差为 1，限定目标组和参照组在某一项目上因素载荷相同[3]。

3.5　DIF 检测及示例

在模型 1 中，两组的因素载荷没有限制，自由估计，如果各拟合指数(绝对拟合指数，相对拟合指数)都比较好，则说明两个群体具有相同的测量模型。

在模型 1 的基础上，估计相对节省的模型 2，如果各拟合指数都比较好，而且 $\chi 2$ 差异显著，则整个测量不存在 DIF，是等同的。也有学者(Byrne，1998；Vadenberg & Lance，2000)[4,6]认为，如果模型有多个潜变量，则需要进一步检验潜变量的方差协方差是否相同，关注在两个群体中测量的理论结构是否相同，这个检验更为严厉。

如果模型 2 被拒绝，说明存在项目功能差异。需要进一步考察哪个项目上存在 DIF，模型 2 的修正指数(MI)可以提供这方面的信息。如果某些因素载荷自由估计的 MI 较大，达到显著水平，需要逐步放开这些载荷让其自由估计，直到模型拟合较好且与模型 1 的差异不显著，则模型 3 的成立。在两个群体中自由估计载荷的项目存在 DIF。

比如[3]，有 5 个项目测量负向情感，5 级计分。两个被试群体，540 名密尼苏达大学生和 598 名南京大学生，考察他们是否由于文化差异，使项目具有了功能差异。

表 1　三个模型的拟合指数表

参数	基准模型		测量等同模型		部分测量等同模型	
	密尼苏达	南京	密尼苏达	南京	密尼苏达	南京
λ_{11}	0.83		0.81		0.82	
λ_{21}	0.89	0.88	0.86		0.88	
λ_{31}	0.69	0.73	0.67		0.69	
λ_{41}	0.92	0.94	0.91		0.93	
λ_{51}	0.91	1.30	0.98		0.91	1.28
Φ	1.00	0.42	1.00	0.52	1.00	0.43
χ^2	74.84		90.03		75.15	
df	10		14		13	
$\Delta\chi^2$	–		15.19 **		0.31	
Δdf	–		4		3	
TLI	0.915		0.929		0.938	
NI	0.958		0.950		0.959	
RMSEA	0.076		0.069		0.065	

(参照组方差固定为 1；在模型中被限定在不同组是等同的参数，在两组的中间位置列出，比如基准模型中，λ_{11}在两组限定相同，估计值为 0.83)。

从表 1 可以看出，基准模型中的各项拟合指标都比较好，说明两个群体具有相同的模型结构。限定两组因素载荷矩阵相同，构建测量相同模型，模型估计的各项拟合指标也比较好，但是与基准模型的卡方差异显著($\Delta\chi^2 = 15.19$，$\Delta df = 4$)，该模型不能成立，有项目存在 DIF。该模型会估计项目因素载荷的修正指数，如果显著的修正指数数目小于总项目数的一半，可以构建部分测量等同模型。在该例中，修正指数显著的只有项目 5，修正指数为 13.97。在模型 3 中，让两组在项目 5 上的因素载荷自由估计，模型各项拟合指标都比较好，与基准模型的卡方差异不显著($\Delta\chi^2 = 0.31$，$\Delta df = 3$)，部分测量等同模型成立。其中，项目 5 在密尼苏达组载荷为 0.91，在南京组载荷为 1.28，说明该项目更能测量出南京组的负向情绪，相同潜在特质情况下，南京组在该项目上的真分数更大，存在 DIF。

3.6　组间均值差异的检验

两个群体在某个测量上是否存在均值差异，也是研究者通常关注的问题，传统上采用方差分析来比较，但必须保证测量工具具有等同性。对于部分测量等同模型，不能进行方差分析，但是 CFA 中的

均值结构比较，可以通过直接比较两个群体的潜变量的差异，来考察群体差异。比如，语文测验，如果存在 DIF，可以考虑采用部分测量等同模型，直接考察男女生潜在的语文能力上的差异。在本例的模型 3 中，估计两组潜在特质（负向情绪）的均值差异，固定密尼苏达均值为 0，则南京组均值为 - 0.378，标准误为 0.060，差异显著。

4 项目反应理论进行 DIF 检测的方法

4.1 基本模型

对于多级计分数据，依据 Samejima(1969) 的等级反应模型讨论。有 m 个类别的项目，有 m - 1 个分类阈限函数 $P_{ik}^{*}(\theta)=\frac{\exp[Da_i(\theta_s-b_{ik})]}{1+\exp[Da_i(\theta_s-b_{ik})]}$ (3)

则 m 个类型反应函数为 $P_{ik}(\theta)=P_{i(i+k)}^{*}(\theta)-P_{ik}^{*}(\theta)$ (4)

则某个项目的真分数为

$E(x_i)=t_i(\theta)$，$t_i(\theta)=(1)P_{i1}+(2)P_{i2}+\cdots+(m_i-1)P_{i(m_i-1)}+(m_i)P_{i(m_i)}$，也即是 $t_i(\theta)=1+P_{i1}^{*}+P_{i2}^{*}+\cdots+P_{i(m_i-1)}^{*}$ (5)，其对应着项目反应函数（*IRF*）[2]。

整个量表的真分数为 $T(\theta)=\sum_{i=1}^{n}t_i(\theta)$，对应着测验反应函数。

4.2 定义 DIF

IRT 领域中，已有不少学者探讨过 DIF，定义检测指标，开发出相应的估计程序，比如 Lord(1980) 的卡方统计量，Raju(1988，1990) 的面积测度等。这里采用 DFIT 框架，它是唯一的可以分析多级计分、同时进行项目水平和测验水平 DIF 检验，而且可以进行总体参数估计的 IRT 程序[2]。

在 DFIT 框架中，定义 DIF 为具有相同能力的两个群体在某项目上真分数不同，即项目反应函数 IRF 存在差异。在图 1 中，参照组和目标组的两条 IRF 线不重合，可能存在 DIF。

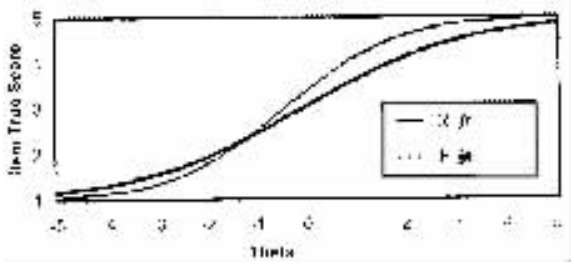

图 1 参照组和目标组的项目反应曲线

下面介绍 DIF 的检验指标和标准[7]。

在项目 i 上，两个群体在特质 θ 的某个水平（s）上的真分数的差异，由(6)表示，

$$d_{is}=t_{isF}-t_{isR} \quad (6)$$

在整个测验上，两个群体在特质 θ 的某个水平（s）上的真分数的差异，由(7)表示，

$$D_s=(T_{sF}-T_{sR})=d_{1s}+\cdots+d_{ns}=\sum_{i=1}^{n}d_{is} \quad (7)$$

项目在特质 θ 的所有水平，真分数差异的平方的平均值，由(8)表达，

$$NCDIF=E(d^2)=\mu_{d^2} \quad (8)$$

测验在特质 θ 的所有水平，真分数差异的平方的平均值，由(9)表示，

$$DTF=E(D^2)=\mu_{D^2} \quad (9)$$

NCDIF 和 DTF 分别是项目水平和测验水平上 DIF 的指标，$\sqrt{NCDIF}$ 和 $\sqrt{DTF}$ 分别代表两个群体真分数的平均差异。

由于 $\chi^2_{N-1}=\frac{N\times DTF}{\hat{\sigma}^2_D}$（N 为目标组的人数，$\hat{\sigma}^2_D$ 为 D 的方差的无偏估计），如果 χ^2 显著，说明该测验存在功能差异。

由于项目的 NCDIF 受样本量影响很大，Raju，Van der Linden，& Fleer(1995)[8] 建议采用经验临界值，即模拟 2000 个没有 DIF 项目，99% 的项目不显著所在的点（I 类错误）。比如三级计分项目的临界点是 0.016，五级计分的是 0.096。在 Raju(1995) 提供的程序里，有临界值的模拟研究。

DTF 是代偿性的指标，即有的项目存在正向的 DIF，有的项目存在负向的 DIF，在特质 θ 的某个水平（s）上，测验 n 个项目的真分数差异可能会互相抵消（$\sum_{i=1}^{n}d_{is}=0$）。因此，DTF 是一个很有弹性的指标，很多时候编制测验无法避免使用存在 DIF 的项目，研究者更关心测验是否等价。而其他更多检测 DIF 的方法，包括 CFA 的 χ^2 指标都是非代偿的。

4.3 检测的一般过程

(1) 在每一个子群体中分别估计所有项目的参数。

(2) 求取所有项目两套参数之间的关系转换函数，即链函数。利用链函数实现项目参数的转换。

(3) 在所有项目参数转换完成后，进行 DIF 分析，如果项目的 NCDIF 超过临界值，被排除。

(4) 对剩下的所有项目重新求取链函数，利用新的链函数对所有项目进行参数转换。

(5)再次进行 DIF 分析,排除有偏的项目。

(6)重复上两步的过程,直至 DTF 的 $\chi 2$ 不显著,部分测量等同模型成立。

项目参数和 θ 值在 PARSCALE 程序(Muraki&Bock,1997)[9]中估计,采用的是边际最大似然估计和 EM 算法估计项目参数,采用 Bayesian 程序估计 θ 值。在 EQUATE 程序中估计链接系数(Backer, 1995)[10]。DFIT 指标由 Raju(1995)提出的一个 FORTRAN 程序估计[11]。

4.4　共同尺度

在 IRT 中,要限定目标组的 θ 符合 N(0,1)分布,同时利用链函数转换参照组的参数的尺度,使得参照组与目标组的潜在结构匹配,具有共同尺度。

存在两个常数 A 和 B($A \neq 0$),使得两套项目参数满足如下关系式(链函数):

$$\theta_{Fj} = A\theta_{Rj} + B \; ; a_{Fi} = a_{Rii}/A \; ; b_{Fi} = Ab_{Ri} + B \; ; c_{Fi} = c_{Ri}$$

其中 $\theta_{Fj}\, a_{Fi}\, b_{Fi}\, c_{Fi}$ 在目标群体 F 上估计的参数,$\theta_{Rj}\, a_{Ri}\, b_{Ri}\, c_{Ri}$ 在参照群体 R 上估计的参数。上述式子中的 A,B 称为链接系数[12]。

4.5　*DIF* 检测及示例

用十个项目测量工作满意度,对两个群体黑人和白人施测,5 级计分[2]。由 *PARSCALE* 程序估计的项目参数,和 *DFIT* 框架估计的 *NCDIF* 指标列在表 2 中。

表 2　参数估计表

item	A	b_1	b_2	b_3	b_4	NCDIF
Item1						0.000
Blank	0.537	-3.739	-2.814	-1.907	0.806	
White	0.557	-3.816	-2.603		-1.769	0.726
Item2						0.144
Blank	0.671	-1.566	-0.715	0.363	1.825	
White	0.913		-1.744	-0.978	-0.237	1.159
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Item10						0.017
Blank	2.321	-1.520	-0.964	-0.351	0.986	
White	1.806	-1.832	-1.033	-0.446	1.089	

表 2 显示出,在项目 2 上 NCDFI 显著,把它删除掉后,测验的 DTF 不再显著。

4.5　均值差异检验

在 PARSCALE 程序可以估计出每名被试的 θ 值,经过链函数转换,在同一尺度上,可以进行两组之间的各种差异检验。

5　两种方法的比较

本文就上述两种方法,首先总结两者的相似性,进而分析其差异。

5.1　两种方法的相似性:

(1)都是建立在潜在结构的基础上,使用潜在特质作为匹配变量。

(2)DIF 的定义都是指两个群体在项目和量表水平上的相同特质水平对应的真分数不一致。

这与平行的概念类似,平行测验是指两个测验的真分数相同,这里是指如果潜变量相同,两个群体的真分数一致。

(3)潜在特质的分布可以不一致,比如,分布的平均数差异就体现为项目影响(item impact),指目标组和参照组在测验所测的特质相同的前提下,项目或者测验的真分数是否一致。

(4)都可以检测出有 DIF 的项目,允许部分测量等同模型存在。通过潜在特质的差异性检验,考察特质的组间差异。

(5)项目反应信息图,可以帮助识别 DIF 的类型、程度和存在的特质位置。

5.2　两种方法的区别:

(1)对于二级计分的数据,用 IRT 更好一些;多级计分的数据二者都是可以的。但是 CFA 可以比较多个组和多个潜变量,而 IRT 目前的一些多级计分和多维的模型正处在迅速发展阶段。

(2)IRT 提供每个项目的类型反应函数,通过分析类型阈限值,帮助寻找造成 DIF 的原因。而 CFA 没有这个功能。

(3)IRT 的 DFIT 框架提出的识别指标 DTF 是代偿性的,而 CFA 是依靠协方差矩阵 E 与 S 的差异值(χ^2)识别 DIF 的,由于 χ^2 是由协方差矩阵中各元素差异的平方和计算得到,所以项目间的不同方向的 DIF 是不能抵消的,因而是不可代偿的。

(4)IRT 优于经典反应理论的一个主要原因是,IRT 更加关注个体,考察 Person - fit 的程度。在 DFIT 框架中,DIF 识别指标 NCDIF 和 DTF,都是个体在项目上的反应偏差的总和,对个体真分数的反应差异是非常敏感的。CFA 关注的是协方差矩阵的差异,不关注个体的差异。

参考文献

1　曾秀琴. 项目功能差异及其检测方法. 心理学动态, 1999, 7(2):41 - 57.

2　Raju N S, Laffitte L J, Byrne B M. Measurement Equivalence; A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory. Applied Psychological Measurement, 2002,87(3),517 - 529.

3 Reise S P, Widaman K F, Pugh R H. Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance. Psychological Bulletin, 1993, 114(3), 552 – 566.

4 Vadenberg R J, Lance C E. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. organizational research. Organizational Research Methods, 2000, 3, 4 – 69.

5 Byrne B M, Shavelson R J. Adolescent self – concept: Testing the assumption of equivalence structural across gender. American Educational Research Journal, 1987, 24, 365 – 385.

6 Byrne B M. Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum. 1998.

7 Flowers C P, Oshima T C, Raju N S. A description and demonstration of the polytomous – DFIT framework. Applied Psychological Measurement, 1999, 23(4), 309 – 326.

8 Raju N S, van der Linden W, Fleer P. An IRT – based internal measurement of test bias with applications for differential item funtioning. Applied Psychological Measurement, 1995, 19, 353 – 368.

9 Muraki E, Bock R D. PARSCALE: IRT based test scoring and item analysis for graded open – ended exercises and performance tasks. Chicago, IL: Scientific Software. 1997.

10 International. Raju, Backer F B. EQUATE 2.1: Computer program for equating two metrics in item response theory. [Compter program]. Madison: University of Wisconsin, Laboratory of Experimental Design. 1995.

11 N. S. DFIT5P: A Fortran program for calculating DIF/DTF [Compter program]. Chicago: Illinois Institute of Technology. 1999.

12 漆书青,等. 现代教育与心理测量学. 南昌: 江西教育出版社. 1998. 278.

Differential Item Functioning: A Comparison of Methods Based on CFA and IRT

Luo Fang, Zhang Houcan

(Psychology College, Beijing Normal University, Beijing 100875)

Abstract: Currently, there are several methods for identify Differential Item Functioning, which are in confirmatory factor analysis and item response theory fields. The major purpose of this article is to offer a comparison of these two methods with a special emphasis on their methodological similarities and differences, for polytomous unidimensional case.

key words: confirmatory factor analysis; item response theory; Differential Item Functioning

(上接第 69 页)

Unfolding IRT Model and the Non – cumulative Response Mechanism in Personality Tests

Guo Qingke[1], Miao Jinfeng[2], Wang Weili[1]

(1. Department of Psychology, Laoning Normal University, Dalian 116029;
2. HaiHua College, Liaoning Normal University, Dalian 116029)

Abstract: People with high trait level will not necessarily get high scores when responding to non – cognitive items, this is called non – cumulative response mechanism. Generalized Graded Unfolding Model (GGUM) is an IRT model developed to solve this problem in personality. In this study EPQ and NEO – FFI were administered to 991 and 1017 students, the results show that GGUM can fit the data better and provide more information than cumulative IRT models (2PLM and SGRM). The study also suggests that the issues of model fit and response mechanism in personality should be further studied.

Key words: Non – cumulative IRT Model; Generalized Graded Unfolding Model; Model – Data Fit