基于IRT的一种测验等值问题的参数估计方法

程黄金
(淮南联合大学 基础部,安徽 淮南 232038)

摘 要:基于项目反应理论,文章介绍了测验等值问题的意义和模型,然后分析了测验等值的原理,并采用最小二乘估计法对其中涉及到的转换系数进行了参数估计,真正实现了项目反应理论中的项目参数等值和真分数等值。

关键词:项目反应理论;测验等值;参数估计
中图分类号:TP311.131 **文献标识码**:A **文章编号**:1672-447X(2010)04-0004-03

1 测验等值的意义

项目反应理论(IRT)将被试潜在特质水平与项目参数置于同一量表,可比性大大增强,可是当项目参数与被试能力都未知时,它们之间就需要相互校准。如果两批被试群体 G_1 和 G_2 分别参加两个不同的测验 x 和 y,则 G_1、G_2 中的被试能力与 x、y 中的项目参数要分别相互校准。由于各自校准建立的坐标体系不同,所以,要对 G_1 与 G_2 中的被试特质进行比较,或对 x 与 y 中的项目参数进行比较,因此就必须对两种情况下的坐标系之间进行转换,这种转换就是项目反应理论中的测验等值。[1]简单的说,测验等值就是将测量同一心理品质的不同测验上的分数或项目参数实现单位系统转换,达到相互之间可以比较的过程。

测验等值无疑是测量中一个重要问题,项目反应理论与经典测量理论相比,或许最大的长处就是测验等值。[2]首先,测验等值是建设科学化题库的需要,要使题库能在长时间内为人们提供具有指定统计特性、且不同测验的结果等值可比、性能稳定一致的试卷,测验等值就是题库具有动态性的保证。其次,等值也是测验公平性的保证,尽管我们在命题过程中总是尽量保持考试难度的稳定性,但不同试卷之间在难度、信度、分数分布方面的差别很难完全避免,这种差别会影响到评价标准的客观性,造成使用不同试卷的被试受到不公平的对待,这就要解决测验总分等值问题。另外,测验等值是实现计算机化自适应性测验(CAT)的前提,实现CAT的意义不仅在于可以提高测验的效率,更重要的是可以提高测验的信度,CAT开发中的一个核心环节是实现各个被试所回答的不同项目之间的等值。

2 测验等值的模型

测验等值的常用方法是在两份待等值的测验x和y中,安排四分之一左右的相同项目,称为铆题,将两份测验分别施测于两批不同的被试群体G_1和G_2后,分别估计出项目参数和能力参数。[3,4]这样,作为铆题的项目就有了两套不同的项目参数,它们之间必然存在着某种关系,找到这种关系,就可以完成从一个系统向另一个系统的转换,从而实现测验的等值。下面基于双参数的*Logistic*模型,研究测验等值的具体方法。

设测验中的某项目j有k_j级记分 (K_j是整数,$j=1,2,\cdots,m$),且是等级反应项目,即所有相邻两级记

收稿日期:2010-02-28
基金项目:淮南联合大学自然科学研究项目(2009LYB0901)
作者简介:程黄金(1976-),安徽桐城人,淮南联合大学基础部讲师,硕士。

分的差相等。用 a_{jk}, b_{jk} 分别表示项目 j 第 k 级得分的区分度和难度参数,依据项目反应理论,能力为 θ_i 的被试在项目 j 上获得 k 级或 k 级以上得分的概率为:

$$P_{jk}^*(\theta_i)=\frac{1}{1+e^{-Da_{jk}(\theta_i-b_{jk})}} \quad (k=1,2,\cdots,k_j) \tag{1}$$

需要指出的是,如果是 0-1 记分,k 取 0 表示的是被试答错,k 取 1 表示的是被试答对。依据公式(1)的定义,能力为 θ_i 的被试在项目 j 上恰好获得 k 级得分的概率 $P_{jk}\theta_i$ 应由下列模型:

$$\begin{cases} P_{j0}(\theta_i)=1-1-P_{j1}^*(\theta_i) \\ \cdots\cdots\cdots\cdots \\ P_{jk}(\theta_i)=P_{jk}^*(\theta_i)-P_{j,k+1}^*(\theta_i) \quad (k=1,2,\cdots,k_j-1) \\ \cdots\cdots\cdots\cdots \\ P_{j,k_j}(\theta_i)=P_{j,k_j}^*(\theta_i) \end{cases} \tag{2}$$

来确定,在公式(2)中,$P_{jk}(\theta_i)$为模型的运算特征函数。

因此,一个能力为 θ_i 的被试在项目 j 上得分的期望值(即真分数)为:

$$X_{ij}=\sum_{k=1}^{kj} k\cdot P_{jk}(\theta_i) \quad (j=1,2,\cdots,m) \tag{3}$$

该被试在整个测验中的总得分的期望值(即全卷真分数)为:

$$X_i=\sum_{j=1}^{m} X_{ij}=\sum_{j=1}^{m}\sum_{k=1}^{kj} k\cdot P_{jk}(\theta_i) \tag{4}$$

如果在两份待等值的测验x和y中,共有n道铆题,能力为 θ_i 的被试分别参加这x和y两个系统的测验,由公式(4),可以得到在这n道铆题项目上,该被试的真分数分别是:

$$X_{xi}=\sum_{j=1}^{n}\sum_{k=1}^{kj} k\cdot P_{xjk}(\theta_i) \quad (5) \text{和} \quad X_{yi}=\sum_{j=1}^{n}\sum_{k=1}^{kj} k\cdot P_{yjk}(\theta_i) \quad (6)$$

其中,公式(5)和(6)中,下标 x 和 y 表示它们对应了上述两个测验系统。

3　测验等值的原理和参数估计

在两份不同的测验 x 和 y 中,同一项目(铆题)的两套参数值间存在的线性关系设为:[5,6]

$$a_{yj}=\frac{a_{xj}}{\alpha} \quad (7) \qquad \text{和}\ b_{yj}=\alpha\cdot b_{xj}+\beta \quad (8)$$

其中,a_{xj},b$_{xj}$ 分别是项目 j 在测验 x 中估计出的区分度和难度值,a_{yj}, b_{yj} 分别是项目 j 在测验 y 中估计出的区分度和难度值,α 与 β 是两套参数的转换系数。

同时,如果用 θ_{xi}, θ_{yi} 分别表示被试 i 在测验 x 和测验 y 中所估计出的能力值,那么对于同一被试而言,他在两个测验系统上被估计出的两套能力参数之间具有的关系为:

$$\theta_{yi}=\alpha\cdot\theta_{xi}+\beta \tag{9}$$

因此,测验等值的实质就是估出这一对转换系数 α 与 β,下面结合项目特征曲线等值原理,应用最小二乘法来实现参数 α 与 β 的估计。

由于能力为 θ_i 不的被试在两个测验系统 x 和 y 中应答同一个项目 j,其得分的期望值(真分数)在理论上是相等的,即:

$$X_{xij}=X_{yij}(X_{xij}=\sum_{k=1}^{kj} k\cdot P_{xjk}(\theta_i), X_{yij}=\sum_{k=1}^{kj} k\cdot P_{yjk}(\theta_i) \tag{10}$$

因此,当测验中有 n 道铆题,被试的数目是 N 的时候,就有:

$$\sum_{i=1}^{N}(X_{xi}-X_{yi})^2=0(X_{xi}=\sum_{i=1}^{n} X_{xij}, X_{yi}=\sum_{i=1}^{n} X_{yij}) \tag{11}$$

然而实际中得到的项目参数和能力参数都是样本估计值,误差的存在不可避免,所以公式(11)的右端不会正好等于零,而应当接近于零。令

$$Q=\sum_{i=1}^{N}(X_{xi}-X_{yi})^2 \tag{12}$$

则上面的函数 Q 中包含了已知参数 $a_{xjk}, b_{xjk}, \theta_{xj}, a_{yjk}, b_{yjk}$,和 θ_{yj},其中符号 a_{xjk} 表示在测验 x 中项目 j 第 k 级得分的区分度,换用字母 b 和 θ 分别表示项目难度和被试能力,其余符号类似解释。

在实际测试中,难以安排同一批被试参加 x 和 y 两个不同的测验,因此就不存在两个对应的能力参数值 θ_{xi} 和 θ_{yi},但是它们之间存在着关系式(9),将该关系式代入公式(12),函数 Q 就变成了包含已知参数 $a_{xjk}, ba_{xjk}, \theta_{xi}, a_{yjk}, b_{yjk}$ 和未知参数 α 与 β 的函数,表示的是所有被试在整个测验中的真分数误差的平方和,下面用最小二乘法来求在 Q 极小条件下的 α

与β的估计值。[6]

结合公式(9)将公式(12)两端同时关于α与β求偏导,得:

$$\begin{cases} f_1=\frac{\partial Q}{\partial \alpha}=-2\sum_{i=1}^{N}(X_{xi}-X_{yi})\cdot\frac{\partial X}{\partial \alpha}=0 \\ f_2=\frac{\partial Q}{\partial \beta}=-2\sum_{i=1}^{N}(X_{xi}-X_{yi})\cdot\frac{\partial X_{yi}}{\partial \beta}=0 \end{cases} \tag{13}$$

其中,$X_{xi}=\sum_{j=1}^{n}(X_{xij}=\sum_{j=1}^{n}\sum_{k=1}^{kj}k\cdot(P_{xjk}^{*}-P_{xjk+1}^{*}$ (14)

$$X_{yi}=\sum_{j=1}^{n}X_{yij}=\sum_{j=1}^{n}\sum_{k=1}^{kj}k\cdot(P_{yjk}^{*}-P_{yjk+1}^{*} \tag{15}$$

而 $P_{xjk}^{*}=\frac{1}{1+e^{-Da_{jk}(\theta_j-b_{jk})}},P_{xjk+1}^{*}=\frac{1}{1+e^{-Da_{xjk}(\theta_{xi}-b_{xjk+1})}}$ (16)

$$P_{yjk}^{*}=\frac{1}{1+e^{-Da_{yjk}(\theta_{yi}-b_{yjk})}},P_{yjk+1}^{*}=\frac{1}{1+e^{-Da_{yjk}(\theta_{yi}-b_{yjk+1})}} \tag{17}$$

因此,$\frac{\partial X_{yi}}{\partial \alpha}=D\theta_{xi}\sum_{j=1}^{n}\sum_{k=1}^{kj}k\cdot[a_{yjk}p_{xjk}^{*}(p_{xjk}^{*}-1)$

$$-a_{yjk+1}P_{xjk+1}(P_{xjk+1}^{*}-1)] \tag{18}$$

$$\frac{\partial X_{yi}}{\partial \beta}=D\sum_{j=1}^{n}\sum_{k=1}^{kj}k\cdot[a_{yjk}p_{xjk}^{*}(p_{xjk}^{*}-1)$$

$$-a_{yjk+1}^{*}P_{xjk+1}^{*}(P_{xjk+1}^{*}-1)] \tag{19}$$

基于以上公式(14)-(19),用 Newton 迭代法就可以解出上述非线性方程组(13),迭代公式为:

$$\begin{pmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{pmatrix}=\begin{pmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{pmatrix}-\begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix}^{-1}\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \tag{20}$$

其中$f_{11}=\frac{\partial^2 Q}{\partial \alpha^2}$,$f_{12}=\frac{\partial^2 Q}{\partial \alpha\partial \beta}$,$f_{21}=\frac{\partial^2 Q}{\partial \beta\partial \alpha}$,$f_{22}=\frac{\partial^2 Q}{\partial \beta^2}$,

迭代终止条件为:$\sqrt{(\alpha^{(t+1)}-\alpha^{(t)})^2+(\beta^{(t+1)}-\beta^{(t)})^2}<\varepsilon$(ε是给定的正实数)。

有了转换系数α与β,就可以根据公式(7),(8)和(9),来实现项目反应理论下的测验等值。

参考文献:

[1]戴海琦.基于项目反应理论的测验编制方法研究[J].考试研究,2006,2(4):31-44.

[2]漆书青,戴海琦.项目反应理论及其应用研究[M].南昌:江西高校出版社,1992:10-136.

[3]戴海琦.等级反应模型项目特征曲线法等值研究[J].心理学探新,2000,20(3):49-53.

[4]陈希镇.关于测验等值几个问题的研究[J].应用概率统计,2000,16(2):213-219.

[5]Lord F M.Applications of item response theory to practical testing problems[M]. Hillsdale, NJ: Erlbaum,1980:21-85.

[6]周骏,欧东明,等.等级反应模型下项目特征曲线等值法在大型考试中的应用[J].心理学报,2005,37(6):832-838.

责任编辑:胡德明

A Parameter Estimation Method for Test Equivalence Based on IRT

Cheng Huangjin

(Department of Basic Courses, Huainan Union University, Huainan 232038, China)

Abstract: Based on item response theory, the article firstly introduces the importance and model of test equivalence, then analyses the principles of test equivalence, adopts least square method to estimate the relevant transform-parameters, and finally realizes item parameters and true score equivalence based on IRT.

Key words: item response theory; test equivalence; parameter estimation

基于IRT的一种测验等值问题的参数估计方法

作者： 程黄金， Cheng Huangjin
作者单位： 淮南联合大学,基础部,安徽,淮南,232038
刊名： 黄山学院学报
英文刊名： JOURNAL OF HUANGSHAN UNIVERSITY
年，卷(期)： 2010，12(5)

参考文献(6条)

1. 戴海琦 基于项目反应理论的测验编制方法研究 2006(4)

2. 漆书青.戴海琦 项目反应理论及其应用研究 1992

3. 戴海崎 等级反应模型项目特征曲线法等值研究 2000(3)

4. 陈希镇 关于测验等值几个问题的研究 2000(2)

5. Lord F M Applications of item response theory to practical testing problems 1980

6. 周骏.欧东明.徐淑媛.戴海琦.漆书青 等级反应模型下项目特征曲线等值法在大型考试中的应用 2005(6)

相似文献(10条)

用一种基于题组的模掣——两参数题组模型(2Parameters Testlets Model，2PTM)，它由IRT两参数逻辑斯蒂克模型(2 ParametersLogistic Model，2PLM) 加入与每个题组相关的随机影响参数扩展而来的。这一模型考虑了铆题组中题目的局部依赖。本文给出了利用IRT特征曲线法求解等值系数的方法和具体步骤。以等值系数估计值的误差大小作为衡量标准，以Wilcoxon符号秩检验为依据，进行了大量的Monte Carlo模拟实验。实验分别从项目参数随机误差的大小，被试人数，题组相依性程度等方面考察对含题组的测验等值的效果，将2PTM与标准的IRT的2PLM进行比较，其中2PLM并没有考虑题组内部的依赖关系。实验结果表明，考虑了局部相依性的题组模型2PTM绝大部分情况下都比2PLM等值的误差小而且有显著性差异，更加适用于题组测验的等值。另外，对6种不同等值准则用2PTM等值的情况也做了相应的比较。结果表明，一般来讲，等值系数A取值在0.5～0.9之间SLcrit表现更好，1.0～1.4之间SQRcrit表现突出，1.5～2.0之间Hcrit表现较好。随参数估计精度的提高，SLcrit和SQRcrit的表现更加突出，胜出的范围也更大。题组相依程度逐渐加强，SQRcrit和Hcrit胜出的情况也增多。LCcrit、Wcrit、SREcrit占优的情况不多，胜出的范围也没有规律。

6.期刊论文 范晓玲.廖利国 测验等值综述 -科技信息2009(34)

对于教育评价计量单位的"分"来说,只有分数等值,测验同一心理特质的不同版本的测验分数之间才具有可比性,从而保证测验的公平性.要使这些分数可以直接比较就必须进行测验等值,测验等值已成为教育测量研究和应用的一个重要问题,也成为教育管理部门重视的问题.与经典测验理论(CTT)相比,基于项目反应理论(IRT)的等值方法,被认为是比较理想的.

7.学位论文 程德巧 绝对值等值准则及求解算法的应用 2005

计算机智能组卷和计算机自适应测验是都需要大型测验题库.题库建设离不开测验等值.本文在分析已有项目反应理论(Item Response Theory简称IRT)等值方法基础上,提出两种更为稳健的等值方法:普通绝对值等值方法和平方根绝对值等值方法.已有的IRT等值准则都具有较好的数学性质,通常采用于N-R迭代算法求解.而N-R算法无法处理新的等值准则,为此,我们引入两种现代启发式搜索算法——模拟退火(Simulated Annealing,简称SA)和进化策略(Evolutionary Strategy,简称ES),以求解新等值准则下的等值系数.经过统计检验,Monte Carlo模拟试验表明:新的等值方法比几个常用等值方法更为稳健.

8.期刊论文 谢小庆.Xie Xiaoqing 对15种测验等值方法的比较研究 -心理学报2000,32(2)

此项研究通过试验方法对4种基于经典测验理论的等值方法和11种基于项目反应理论的等值方法进行了比较研究。研究数据为HSK正式考试的数据，研究采用了较为可靠的检验标准。研究结果表明，在有些情况下，进行等值处理并非是最好的选择；在题库建设中，某些IRT方法是可行的；至少对于HSK数据，不论是单、双、三参数，不论是ms方法和mm方法，IRT参数转换等值方法的误差都较大，均不足取。

9.学位论文 邓湘云 CTT与IRT等值方法比较研究 1996

10.学位论文 廖利国 湖南省高考理科数学的等值研究 2010

目的：应用经典测量理论(CTT)和项目反应理论(IRT)对湖南09年高考理科数学与模拟考试进行等值分析，尝试使用新的等值模型研究模拟考试和高考等值，为避免高考“一考定终身”提供理论与实践的参考。
方法：采用锚测验等值设计。基于经典理论的等值方法选用Tucker等值方法、Levine观察分数等值方法、Levine真分数等值方法和等百分位等值方法。基于项目反应理论理论的等值分析采用的是“混合模型”，选择题选用三参数模型，解答题选用Samejima等级反应模型。
结果：所选锚题与两次考试的相关系数分别为0.778和0.725，锚测验与需等值的两次考试显著相关。线性等值方法中，总体权重等于0.4811时，Tucker方法，截距为-2.7136，斜率为0.9336；Levine观察分数方法，截距为-5.0561，斜率为0.9487。用ST和Polyst_wcv1.0程序计算出等值转换常数α和β，选用Stocking-Lord方法得出的等值常数，选择题的等值常数α=1.0216，β=0.1008，解答题的等值常数α=0.6943，β=-0.3151。线性等值方法中用Tucker方法等值标准误最小为1.6425，等百分位方法等值标准误最大为5.2639。IRT观察分数方法等值标准误选择题为1.1508，解答题为4.1408，IRT真分数方法等值标准误选择题为1.1924，解答题为3.3592。
结论：针对本研究所取等值数据与样本，线性等值方法优于等百分位等值方法，其中Tucker方法比Levine观察分数方法更好一些，频数估计中等百分位方法等值误差较大，不足取。在高考数学成绩等值工作中，线性等值方法选择不同总体权重对等值结果影响很小。用项目反应理论等值方法对高考数学的考试数据采用“混合模型”对其进行分析是可行的。