心理学报 2010, Vol. 42, No.3, 434−442
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2010.00434

含题组的测验等值*

吴 锐[1,2] 丁树良[1] 甘登文[1]

([1]江西师范大学计算机信息工程学院, 南昌 330022) ([2]华东交通大学图书馆, 南昌 330013)

摘 要 题组越来越多地出现在各类考试中, 采用标准的 IRT 模型对有题组的测验等值, 可能因忽略题组的局部相依性导致等值结果的失真。为解决此问题, 我们采用基于题组的 2PTM 模型及 IRT 特征曲线法等值, 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, 在几种不同情况下进行了大量的 Monte Carlo 模拟实验。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 等值的误差小且有显著性差异。另外, 用 6 种不同等值准则对 2PTM 等值并评价了不同条件下等值准则之间的优劣。

关键词 测验等值; 题组; 项目反应理论; Monte Carlo 模拟; 等值准则

分类号 B841.7

1 问题的提出

在实际考试中, 多种多样的试题形式应运而生, 一个测验其长度如果至少为 2, 其中某个“大题”由一个题干和附属于该题干的多个“小题”组成, 这个“大题”便称为“题组”, 如英语考试中阅读理解正是典型的题组题型。现有关于题组测验的等值问题, 国外的研究文献较少, 并且研究多是观察分数等值, 而不是项目参数等值。项目反应理论(简称 IRT)框架下的传统等值问题研究绝大部分是基于非题组题型。本文的研究目的是探索适合于题组的 IRT 等值的模型和方法, 希望本研究的结果对含题组测验的各类考试提供启发和借鉴。

IRT 建立在很强的统计假设之上, 其基本假设之一是局部独立性假设(local independence, LI), 即在给定能力的条件下, 同一被试在不同项目上的反应应该是相互独立的(Kolen & Brennan, 1995)。然而在含有题组的测验中, LI 假设就很难站得住脚(Wainer, Bradlow, & Wang, 2007), 因为同一题组内的各项目之间很可能存在相依性。此时, 在违背 LI 的情况下运用 IRT 方法等值, 其结果极可能产生偏差(Lee, Kolen, Frisbie, & Ankenmann, 2001; Li, Bolt, & Fu, 2005), 所以在包含有题组的测验中, 用 IRT 方法等值的结果失真程度是否严重是我们需要探察的问题。另外, 以前常用的等值方法是针对局部独立的情形开发的, 对于局部相依的题组, 这些等值方法还适用吗? 其他一些新方法是否有可能更好呢? 这也是需要我们进行探索的问题。

2 IRT 等值基本方法

2.1 IRT 基础模型、题组模型

IRT 模型是反映被试在测验中潜在的能力水平和可观察的反应之间的函数关系。若被试在题组题型的作答视为独立的, 则采用 IRT 基础模型等值。本文采用伯恩鲍姆在 1958 年提出的 0-1 评分的两参数逻辑斯蒂克模型(简记为 2PLM)(漆书青, 戴海崎, 丁树良, 2002), 其函数形式如下:

$$P_j(\theta) = \{1+\exp[-Da_j(\theta - b_j)]\}^{-1} \quad (1)$$

其中 D 为量表因子, 通常取为 1 或 1.7, a_j、b_j 分别为第 j 个项目的区分度和难度参数, $P_j(\theta)$则表示能力为θ 的那些被试答对第 j 个项目的概率。

随着测量实践的发展和测量研究的深入, 学者

收稿日期: 2008−09−26

* 国家自然科学基金(30860084), 教育部高校博士基金课题(编号 8020070414001), 江西省教育厅科技项目基金(GJJ08154), 卫生部项目(KY200704), 江西省教育厅科技项目(2675), 教育部人文社科项目(09YJCXLX012)资助。

通讯作者: 丁树良, E-mail: ding06026@163.com

们提出了几种针对题组的新模型，有人称之为题组反应理论模型(Testlet Response Theory Model，简记为 TRTM)。本文采用扩展 2PLM 的题组模型(简记为 2PTM)(Wang & Wilson, 2005)。该模型定义被试α 正确作答第 d 个题组中第 j 个项目的概率为

$$P_{\alpha dj}=\{1+\exp[-Da_{dj}(\theta_{\alpha}-b_{dj}-\gamma_{\alpha d})]\}^{-1} \quad (2)$$

其中θ_{α}是被试α 的能力参数，a_{dj}、b_{dj}分别是第 d 个题组中第 j 个项目的区分度和难度参数。$\gamma_{\alpha d}$是 Li, Bolt & Fu(2005)提出的题组因子，独立于θ_{α} 的随机影响参数，代表被试α 与第 d 个题组之间的交互效应。在某一个题组中，被试α 的$\gamma_{\alpha d}$值是一个常量，但是因人而异。对于不同的题组，$\gamma_{\alpha d}$的方差可以是不同的，其方差的大小反映每个题组内局部依赖的程度。另外，在题组模型等值的过程中，为了识别，他们除了限定被试的能力的分布为平均数为 0，标准差为 1 以外，还限定每个题组因子的均值$\mu_{\gamma 1}$, $\mu_{\gamma 2}$,…, $\mu_{\gamma D}$为 0，或者满足$\sum_{d=1}^{D}\mu_{\gamma_d}=0$，方差自由估计。在等值的过程中，就像我们假设两个组被试的能力之间有线性关系从而推出项目参数之间的等值关系一样，两个组的被试在共同题上的题组因子的平均数可能也是不一样的，但参数估计的时候两组被试在题组因子上的平均分都是限定为 0，所以这时项目的难度参数可能就存在一个波动，这也就是 Li, Bolt 和 Fu (2005)提出的除了传统的 A 和 B 以外还有一个$\gamma_{\alpha d}$的原因。

2.2　2PTM 重新参数化

由于题组内的各项目之间很可能存在局部相依，如果把 IRT 基础模型直接运用到含有题组的测验等值中，则违背 LI，从而导致等值结果偏差。我们试图采用题组模型，考虑题组内部各小题之间的相互依赖，使模型与资料更加拟合，期望能得到较好的等值效果。2PTM 公式(2)与 2PLM 公式(1)相比，增加了随机影响参数$\gamma_{\alpha d}$，等值也就不像用传统 IRT 模型那么直接了当。尽管如此，为使等值变得更简单，Glas, Wainer 和 Bradlow 在 2000 年提出(Li, Bolt & Fu, 2005)把公式(2)中的$a_{dj}(\theta_{\alpha}-b_{dj}-\gamma_{\alpha d})$表示为$a_{dj}(\xi_{\alpha d}-b_{dj})$，其中$\xi_{\alpha d}=\theta_{\alpha}-\gamma_{\alpha d}$。假设被试的能力参数$\theta_{\alpha}$和随机影响因子$\gamma_{\alpha d}$分别服从正态分布且相互独立，$\theta_{\alpha}$的均值为 0，方差等于 1，服从标准正态分布；$\gamma_{\alpha d}$的均值为 0，方差为$\sigma^2_{\gamma_d}$。所以在给定$\theta$ 的情况下，ξ_d服从$N(\theta,\sigma^2_{\gamma_d})$。在重新参数化的模型当中，对每一个题组而言，某个被试的能力参数$\xi_{\alpha d}$可以看成是来自于以θ为均值，且以$\sigma^2_{\gamma_d}$为方差的正态分布的一个随机的抽样。采用这样的参数化后，能力值为$\xi_{\alpha d}$的那类被试正确作答第 d 个题组中第 j 个项目的概率即为

$$P_{\alpha dj}=\left\{1+\exp[-Da_{dj}(\xi_{\alpha d}-b_{dj})]\right\}^{-1} \quad (3)$$

对于重新参数化后的题组模型，与 2PLM 有相同的表达形式，更加有利于等值的实施。

2.3　两种模型参数对比

如果是不存在题组效应的情况，题组模型 2PTM 也就化简成为纯粹的 2PLM；但如果存在题组效应，选用 2PLM 则必定要影响参数估计。此时，b_{ydj}便应该包含这方面的信息，不论你是否注意到，它都客观存在。所以标准的 IRT 模型事实上在等值的过程中是受到题组因素影响的。以极大似然估计(MLE)而论，求项目参数，要对所有被试在各项目上的反应进行分析。然而，IRT 项目参数的估计是独立于被试样本的，一个项目只有一个难度参数，并不因人而异，故$\gamma_{\alpha d}$要对α求和再求平均得$\mu_{\gamma_d}=\sum_{\alpha=1}^{N}\gamma_{\alpha d}/N$，所以等值转换过程中，对项目参数应加上$\mu_{\gamma d}$, $b_{ydj}=Ab_{xdj}+B+\mu_{\gamma d}$，用 2PLM 和 2PTM 对题组测验进行等值，对具体的模型参数做了一个比较，如表 1。

3　Monte Carlo 模拟实验

本文重点是在存在题组效应时考虑如何使得测验等值更加准确的问题。我们采用 Monte Carlo 方法模拟出铆题项目参数等进行测验等值的数据，而且预先给定两测验的等值系数真值，模拟整个测验等值的过程，计算估计的等值系数与真值的差异，从而评价等值效果的优劣。鉴于其他文献(丁树良，熊建华，2003；丁树良，熊建华，罗芬，吴锐，甘小方，涂白，2005)已经给出具体的等值实施步骤，本文中不作赘述。

假设(1)被试的能力值θ~$N(0,1)$；(2) 各铆题的项目参数a_{xdj}, b_{xdj}(1na_{xdj}~$N(0,1)$, b_{xdj}~$N(0,1)$)；(3) 参数估计的随机误差ε_{adj} ~ $N(0,\sigma_1^2)$，ε_{bdj}~ $N(0,\sigma_2^2)$根据 MULTILOG 使用说明书(Thissen, 1991)可知，区分度的估计标准误差(SE)大部分记在 0.05 ~ 0.15 之间，而难度的估计标准误差大部分在 0.1 ~ 0.25 之间，故取σ_1在 1/20~1/6 之间，σ_2在 1/10~1/4 之间；(4) 题组因子$\gamma_{\gamma d}$~ $N(\mu_{\gamma_d},\sigma^2_{\gamma_d})$其中$\mu_{\gamma d}$和$\sigma_{\gamma d}$分别是题组 d 的影响因子$\gamma_d$的均值和标准差；由(1)到(4)模拟作答矩阵 U。

3.1　等值的评价指标

假设对于某一对 A 和 B 的真值，不同误差条件下重复求等值系数 k 次，则可以得到 k 组不同的等

表 1 2PLM 和 2PTM 等值过程中的模型参数对比

模型	2PLM	2PTM(Testlet)
数据准备	模拟 X 测验铆题的项目参数 a_{xdj}, b_{xdj} 模拟 N 个被试的能力值$\theta_{x\alpha}$ 模拟铆题组的题组因子$\gamma_{\alpha d}$ 给定等值系数的真值 True_A 和 True_B 说明: 因实验比较的需要, 2PLM 采用与 2PTM 对应的的项目参数来表示。	
X 测验 ICC	$P_{Lx\alpha dj}=\left\{1+\exp\left[-Da_{xdj}\left(\theta_{x\alpha}-b_{xdj}\right)\right]\right\}^{-1}$	$P_{Tx\alpha dj}=\left\{1+\exp\left[-Da_{xdj}\left(\xi_{x\alpha d}-b_{xdj}\right)\right]\right\}^{-1}$ $\xi_{x\alpha d}=\theta_{x\alpha}-\gamma_{\alpha d}$
参数转换	$a_{ydj}=a_{xdj}/A+\varepsilon_{adj}$ $b_{ydj}=Ab_{xdj}+B+\mu_{\gamma_d}+\varepsilon_{bdj}$ $\theta_{y\alpha}=A\theta_{x\alpha}+B$ 其中 $\mu_{\gamma_d}=\sum_{\alpha=1}^{N}\gamma_{\alpha d}/N$	$a_{ydj}=a_{xdj}/A+\varepsilon_{adj}$ $b_{ydj}=Ab_{xdj}+B+\varepsilon_{bdj}$ $\xi_{y\alpha d}=A\xi_{x\alpha d}+B=A\left(\theta_{x\alpha}-\gamma_{\alpha d}\right)+B$
Y 测验 ICC	$P_{Ly\alpha dj}=\left\{1+\exp\left[-Da_{ydj}\left(\theta_{y\alpha}-b_{ydj}\right)\right]\right\}^{-1}$	$P_{Ty\alpha dj}=\left\{1+\exp\left[-Da_{ydj}\left(\xi_{y\alpha d}-b_{ydj}\right)\right]\right\}^{-1}$
等值准则	(以 Hcrit 为例) $Hcrit=\sum_{\alpha=1}^{N}\sum_{j=1}^{m}(P_{Lx\alpha dj}-P_{Ly\alpha dj})^2$	(以 Hcrit 为例) $Hcrit=\sum_{\alpha=1}^{N}\sum_{j=1}^{m}(P_{Tx\alpha dj}-P_{Ty\alpha dj})^2$
计算结果	求出等值系数 A、B 用 A、B 与 True_A、True_B 计算 评价指标 RMSD_L	求出等值系数 A、B 用 A、B 与 True_A、True_B 计算 评价指标 RMSD_T
比较结果	比较 RMSD_L 和 RMSD_T 的值, 较小者说明等值得更准确	

值系数 $A^{(h)}$,$B^{(h)}$, h=1,2,…,k, 本文中 k=50 固定不变。计算估计的等值系数与真值的偏移均方根 (root mean square deviation, 简记为 RMSD)

$$RMSD=\sqrt{\sum_{h=1}^{k}\left[(A^{(h)}-A)^2+(B^{(h)}-B)^2\right]/k} \quad (5)$$

其中 A 和 B 表示等值系数的真值, 故 RMSD 的比较原则是基于估计值对真值的修复(recovery)程度。若 RMSD 的值越小, 说明估计值与真值越接近, 即修复程度越好, 并认为这种等值方法越好。相反, RMSD 的值越大, 说明估计值与真值偏离越远, 该等值方法不能精确地估计出等值系数。

3.2 题组相依性检验

对模拟数据进行题组相依性检验, 是保证实验有效的前提。我们对题组进行相依性检验的方法是, 根据模拟产生的项目参数模拟出被试的作答 U 矩阵, 然后对 U 矩阵计算 Q3 统计量, 若两个题目的 Q3 绝对值大于 0.2, 则认为这两题存在相依性 (Chen & Thissen 1997)。

将项目参数用 2PTM 模拟 U 阵, 依题组因子方差 $\sigma_{\gamma d}$ 的大小检验题组内各小题的相依性, 以 |Q3|>0.2 的成对小题数量和 Q3 绝对值的大小评判题组的相依程度。

(1) $\sigma_{\gamma d}$=0 的题组内, 各小题之间计算的 Q3 绝对值无大于 0.2 者, 即没有表现出相依;

(2) $\sigma_{\gamma d}$<1 的题组, |Q3|>0.2 的成对小题只占个别, 而且|Q3|大多小于 0.3, 可以认为是轻度相依;

(3)1≤$\sigma_{\gamma d}$<2 的题组, |Q3|>0.2 的成对小题约为题数的一半, |Q3|的值适中, 平均约在 0.3 至 0.4, 可认为是中度相依;

(4) $\sigma_{\gamma d}$≥2 的题组, |Q3|>0.2 的成对小题较多, |Q3|的值较大, 有些甚至在 0.5 以上, 则认为是重度相依。

经检验, 模拟产生的实验数据在题组相依性上基本符合实验要求。

3.3 实验设计

在这个模拟研究中采用非等组铆测验等值设计。出于比较的目的, 同一批生成的模拟数据, 分别在 2PTM 和 2PLM 这两种模型下进行等值。模拟的每个测验版本的铆题组题量包含两种情况: 6 个题组或 4 个题组, 并且每个题组含有 5 个小题, 共计 30 个或 20 个题。在 6 个铆题组的情况下, 其中有 4 个题组内部含有依赖关系, 有 2 个题组内的各小题之间是相互独立的, 且这两题组因子的方差为 0。这样铆题就可以看成 4 个题组和 10 个独立的小题, 它们来自于一个含题组的测验。在 4 个铆题组的情况下, 这 4 个题组内部都存在相依性, 可以看成是完全的题组测验。在以前的研究中, Li, Bolt 和 Fu 于 2004 年发现当每题组含有 5 个小题以上, IRT

特征曲线等值法用于题组模型的精确度并不随题组内题量的增加而相应地受影响(Li, Bolt, & Fu, 2005)。所以, 在本文中每个题组包含的题目数量设定为 5, 这在现实的阅读理解测验当中大多也是如此。

我们实验的主要目标是分别考察 2PLM 和 2PTM 应用于题组等值的表现, 从参数随机误差的大小, 被试人数, 题组相依性程度等方面比较题组等值的效果。表 2 给出这几方面的实验条件变化值。

分别采用 Haebara 和 Stocking-Lord 两种等值准则(漆书青等, 2002)处理 2PLM 和 2PTM 的等值, 目的是考察在题组测验中 2PLM 和 2PTM 是否因等值准则的不同而表现不一致, 也即模型的稳定性。在采用 Haebara 准则等值的实验中, 还对不同题组因子的数据集(采用 Li 等人(2005)的研究中 Table2 的 Condition 1, 共 10 个数据集 Data Set 1~10, 每个数据集中都分别列出 4 个题组的题组因子均值)作为实验条件, 以考察不同的题组因子对等值效果的影响。所以, 对不同的条件进行交叉组合, 2PLM 和 2PTM 的模型比较共有 9 项实验, 某种条件改变的情况下, 各实验的具体安排如表 3。

表 2　实验条件的变化值

条件		变化值
参数随机误差(ε_{adj}, ε_{bdj})		(1/6, 1/4), (1/10, 1/6), (1/20, 1/10)
被试人数 N		300, 1000, 5000
题组相依性程度 $(\sigma^2_{\gamma_1},\sigma^2_{\gamma_2},\cdots,\sigma^2_{\gamma_D})$	6(30)*	(0,0,0.5,0.5,1,1), (0,0,1,1,2,2), (0,0,2,2,4,4)
	4(20)	(0.25,0.25,0.25,0.25), (0.5,0.5,0.5,0.5), (0.75,0.75,0.75,0.75), (1,1,1,1), (2,2,2,2), (3,3,3,3), (4,4,4,4)

注: * 6(30)代表 6 个题组共 30 小题

表 3　模型比较的实验安排

等值准则	参数随机误差	被试人数	题组相依性程度	题组因子数据集
Hcrit	实验 1	实验 2	实验 3(6 题组) 实验 4(4 题组)	实验 5
SLcrit	实验 6	实验 7	实验 8(6 题组) 实验 9(4 题组)	

其中实验 1 的内容是分别用 2PLM 和 2PTM 进行等值, 采用 Haebara 准则, 考察改变参数随机误差的大小对等值效果的影响。实验 2 至 9 的内容与实验 1 类似, 改变不同的条件下, 采用 Hcrit 或 SLcrit 准则等值的情况。

3.4　实验结果与分析

本文的每个实验, 等值系数的真值 A 从 0.5 到 2.0, B 从–0.4 到 0.4, 以步长为 0.1 变化, 每一对真值(A,B)的不同组合, 我们都进行 50 次等值模拟, 计算偏移均方根 RMSD。所以对给定真值(A,B)的 144 个组合, 每个实验均进行 7200 次模拟。

在实验中, 确定实验条件后, 每一组(A,B)的取值分别用 2PLM 和 2PTM 计算 RMSD, 以及对两种不同模型的等值结果进行 Wilcoxon 符号秩检验。我们对 A、B 的每一种组合, 从两模型的结果中选出一个最优胜者, 也即 RMSD 最小者。并且经统计检验后确认有显著差异的, 给它添上括号。其中 2PLM 用 L 表示, 2PTM 用 TM 表示。

以实验 1 为例, (1)采用 Haebara 准则, (2)固定 1000 个被试, (3)6 个铆题组共 30 小题(其中各题组因子的方差为(0,0,0.5,0.5,1,1), 2 个题组不含相依性(方差为 0), 2 个题组轻度相依(方差为 0.5), 还有 2 个稍重度的相依(方差为 1)), (4)题组因子均值用 Date Set 1。在以上四个条件不变的情况下, 改变项目参数随机误差的标准差为(1/6, 1/4)、(1/10, 1/6)、(1/20, 1/10)分别进行等值模拟实验, 得到的实验结果显示在表 4 和附录的附表 1、附表 2 当中。

我们可以看出, 2PTM 在实验结果的 3 个表中都是绝大部分占优。其中当σ_1=1/6, σ_2=1/4 时, 在 A、B 的 144 个组合中, 2PTM 有 125 个胜出且 96 个具有显著性优势, 2PLM 只有 19 个占优, 2 个具有显著差异, 主要集中在 A>1.7 的范围内; 当σ_1=1/10, σ_2=1/6 时, 2PTM 有 141 个胜出且 117 个具有显著性优势; 当σ_1=1/20, σ_2=1/10 时, 144 个组合全部都是 2PTM 等值更准确, 只有 1 个无显著差异。随着误差的减小, 2PTM 也表现得更加出色, 这一结果提示我们, 对题组测验实际等值过程中要得到更精确的结果, 应该尽量提高题组题型项目参数估计的精度。

表 4 实验 1 当随机误差取值(σ_1=1/6, σ_2=1/4)的结果

B	A															
	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
–0.4	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	(TM)	TM	TM	L
–0.3	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	TM	L	(L)	L
–0.2	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	TM	TM	L	TM	(L)
–0.1	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	(TM)	(TM)	TM	L	TM	L	L
0	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	TM	TM	L
0.1	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	(TM)	TM	L	L	L
0.2	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	L	L	TM	TM	TM
0.3	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	L	TM	TM	L
0.4	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	TM	TM	L	L

实验 2~9 采用的实验方法与实验 1 相同, 为节省篇幅, 略去实验 2~9 的结果表格。现就所有的实验结果作一小结:

(1) 无论项目参数的随机估计误差是较大、适中或较小, 考虑了局部相依性的题组模型 2PTM 绝大部分情况下的等值误差比不考虑相依性的局部独立模型 2PLM 的要小。

(2) 改变被试的人数, 2PTM 始终保持优势, 说明不论抽样误差大或小, 2PTM 都更加适合含题组测验的等值。

(3) 改变题组相依性程度的情况下, 相比 2PLM, 2PTM 的等值效果表现得更好些。特别是测验若全为题组, 2PTM 表现出显著的优势, 这更加说明, 对于题组题型的等值, 题组模型更有优势。

(4) 不同的题组因子均值对等值是有影响的, 但是影响并不大。在绝大多数情况下, 2PTM 都表现得比 2PLM 要好。

3.5 对 2PTM 用不同准则的比较实验

我们从前面对 2PLM 和 2PTM 两种模型的比较知道, 2PTM 要更适用于题组等值, 那对 2PTM 采用不同等值准则的等值效果又将如何呢? 为了得到答案, 我们对此也进行了准则的比较实验。分别采用 Haebara 准则 (Hcrit)、Stocking-Lord 准则 (SLcrit)、对称相对熵准则(SREcrit) (丁树良, 熊建华, 2003)、对数对比准则(LCcrit) (丁树良, 熊建华, 毛萌萌, 2003)、Haebara 加权准则(Wcrit) (熊建华, 丁树良, 2005)、平方根等值准则(SQRcrit) (丁树良等, 2005)这 6 种准则进行题组模型下的等值, 探查用 2PTM 解决含题组测验的等值问题时, 不同的准则对等值精度的影响。与模型比较的实验设计类似, 同样是从参数随机误差的大小, 被试人数, 题组相依性程度等方面的条件下模拟实验, 因此准则的比较共有 4 项实验。

表 5 列出了实验安排情况, 其中实验 10 的内容是用 2PTM 进行题组等值, 同时采用 6 种不同的等值准则, 考察改变参数随机误差的大小各准则的适用范围。实验 11 至 13 的内容与实验 10 类似。

实验也采用简化的表格, 每一组(A,B)的取值分别用 6 种不同的准则计算出 RMSD, 选出 RMSD 最小者填入表中对应单元格, 以便从中找出等值准则的一些规律。此时由于比较的准则较多, 故未进行显著性检验。为简洁起见, 表格中 RMSD 的最优者均用准则名称的大写字母部分代替, 即用 H、SL、SRE、LC、W、SQR 表示。

目标二以实验 10 的结果为例说明, 得到的实验结果见表 6 及附录的附表 3、附表 4。

从实验 10 中 3 个表的结果可以知道, 对于题组测验的等值 SLcrit、SQRcrit 和 Hcrit 胜出的情况占据了较大的范围。其中 SLcrit 在 A 偏小的区域表现更好, Hcrit 在 A 较大的范围内表现更好, SQRcrit 则是在 A 适中的范围内表现突出。而且随着误差的减小, 它们的适用范围会沿着 A 增大的方向转移,

表 5 等值准则比较的实验安排

等值准则	参数随机误差	被试人数	题组相依性程度
Hcrit、SLcrit、SREcrit、LCcrit、Wcrit、SQRcrit	实验 10	实验 11	实验 12(6 题组) 实验 13(4 题组)

表 6　实验 10 当随机误差取值(σ_1=1/6, σ_2=1/4)的结果

B	A															
	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
–0.4	SL	SL	SL	SL	SL	SQR	SQR	SQR	W	SQR	SQR	SQR	H	H	H	H
–0.3	SL	SL	SL	LC	SL	W	SQR	SQR	SQR	W	H	H	SRE	H	H	H
–0.2	LC	SL	SL	SL	SQR	SQR	SQR	W	SQR	SQR	SQR	H	H	H	H	H
–0.1	SL	SL	LC	SL	LC	W	SQR	SRE	SQR	SQR	SQR	H	H	H	H	H
0	SL	SL	SL	SL	SL	SL	W	SQR	W	SQR	H	H	SQR	H	H	H
0.1	LC	LC	SL	SL	SL	W	SQR	SQR	SQR	H	H	H	H	H	H	H
0.2	LC	SL	SL	SL	SL	SQR	W	W	W	SQR	H	H	H	SRE	H	H
0.3	SL	SL	SL	SL	W	W	W	SQR	SQR	SQR	SQR	H	SQR	H	H	H
0.4	LC	SL	SL	SL	SL	W	W	SQR	SQR	SQR	H	SQR	H	H	H	H

即 SLcrit 表现较好的范围由 A<1 扩大到 A<1.2, SQRcrit 表现突出的范围由 1.0<A<1.5 转移到 1.2<A<1.7, Hcrit 胜出的区域 A 的取值范围由 1.4~2.0 减少到 1.8~2.0。LCcrit、Wcrit、SREcrit 胜出的情况不太多, 都只是零星的表现出来。除了 Wcrit 相对集中于 0.9<A<1.6 的区域内, LCcrit 和 SREcrit 都无固定的占优区域。

实验 11~13 采用的方法与实验 10 相同, 限于篇幅, 不列出实验结果。对准则比较的实验结果的表现小结如下:

(1) SLcrit、SQRcrit 和 Hcrit 胜出的情况占据了较大的范围。LCcrit、Wcrit、SREcrit 占优的情况不多, 胜出的范围也没有规律。

(2) 当项目参数估计精度较低时, 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。当参数估计精度较高时, SLcrit、SQRcrit 和 Hcrit 表现较好的范围依次是, A 取值 0.5 ~ 1.1, 1.2 ~ 1.6, 1.7 ~ 2.0。

(3) 题组相依程度逐渐加重, SQRcrit 和 Hcrit 胜出的情况也增多, SQRcrit 在 0.7 ~ 1.3 的范围内, Hcrit 在 1.4 ~ 2.0 的范围内表现得比其他准则要好。

4　展望

对于含题组的测验等值问题, 本文的研究结论是考虑题组效应影响的等值效果比忽略题组影响的更加准确。但实验条件还存在一定的局限性, 目前只考察了两参数的情形, 三参数的情形是否依然如此, 还需要我们进一步探讨。另外, 本文讨论的是 0-1 记分的模型, 对于现在研究较热的多级评分模型, 它对应的题组模型应该如何表示, 用于题组的测验等值该如何进行? 这些都是值得我们深入研究的问题。

另外, 铆题设计如何进行, 这是一个难题。特别是, 在铆测验等值设计中, 铆题应是整个测验的一个"微缩版", 其铆题目比例是一个重要的考虑因素, 但一份试卷中题组题不一定太多, 有时甚至只有一、二题, 如果在铆题中安排了题组题, 这种类型题所占的份额可能超出了其他题所占比重。比如铆题量占测验长度 1/3, 则各类题目也应占相同的份额。由于题组题的题量在测验中一般不会太多, 纵使在铆题中安插了一个题组题, 也可能超过了 1/3 这个比例。故这对于等值设计来讲也是一个值得讨论的问题。

还有一点要特别说明的是, 本文专门讨论等值误差, 仿照文末的参考文献 2 和 3(丁树良等 2003; 2005)所提供的方法, 并没有动用程序估计 a、b 和题组影响因子的方差, 以免讨论等值误差时混杂了参数估计误差。

参　考　文　献

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Ding, S. L., & Xiong, J. H. (2003). Study of several equating issues on IRT (in Chinese). *China Examinations, 12*, 14–15.

[丁树良, 熊建华. (2003). 项目反应理论框架下几个等值问题的探讨. *中国考试, 12*, 14–15.]

Ding, S. L., Xiong, J. H., Luo, F., Wu, R., Gan, X. F., & Tu. B. (2005). A new equating criterion and its behaviorsl (in Chinese). *Acta Psychologica Sinica, 37*(5), 674–680.

[丁树良, 熊建华, 罗芬, 吴锐, 甘小方, 涂白. (2005). 一种新的等值准则及其实用范围的探讨. *心理学报, 37*(5), 674–680.]

Ding, S. L., Xiong, J. H., & Mao, M. M. (2003). Logcontrst

method for equating test based on IRT (in Chinese). *Acta Psychologica Sinica*, *35*(6), 835–841.

[丁树良, 熊建华, 毛萌萌. (2003). 项目反应理论框架下的新等值方法—— 对数对比等值法. *心理学报*, *35*(6), 835–841.]

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and Practices* (pp. 169–173). New York: Springer-Verlag.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scaores from tests composed of testlets. *Applied Psychological Measurement*, *25*(4), 257–372.

Li, Y. M., Bolt, D. M., & Fu, J. B. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement*, *29*(5), 340–356.

Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). Principles of modern educational and psychological measurement (pp. 201–224) (in Chinese). Beijing: Higher Education Press.

[漆书青, 戴海崎, 丁树良. (2002). *现代教育与心理测量学原理*(pp. 201–224). 北京: 高等教育出版社.]

Thissen, D. (1991). MULTILOG User's Guide [Computer software]. Scientific Software, Inc. Examples3-7–3-9.

Wainer, H., Bradlow, E., T., & Wang, X. (2007). *Testlet response theory and its application* (pp. 11–12). New York: Cambridge University Press.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Xiong, J. H., & Ding, S. L. (2005). Haebara equating method and weighted-Haebara equating method (in Chinese). *Journal of Jiangxi Normal University (Natural Sciences Edition)*, *29*(5), 434–437.

[熊建华, 丁树良. (2005). Haebara 等值方法及其加权准则. *江西师范大学学报(自然科学版)*, *29*(5), 434–437.]

附录

附表 1 实验 1 当随机误差取值(σ_1=1/10, σ_2=1/6)的结果

B	A															
	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
–0.4	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	(TM)	TM	(TM)	L
–0.3	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	TM	TM	L	TM
–0.2	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	(TM)	L
–0.1	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	(TM)	TM	TM
0	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	TM
0.1	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM
0.2	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	(TM)	TM	(TM)	(TM)
0.3	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM	(TM)	TM
0.4	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM	TM

附表 2 实验 1 当随机误差取值(σ_1=1/20, σ_2=1/10)的结果

B	A															
	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
–0.4	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
–0.3	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
–0.2	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	TM
–0.1	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
0	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
0.1	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
0.2	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
0.3	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)
0.4	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)	(TM)

附表 3　实验 10 当随机误差取值(σ_1=1/10, σ_2=1/6)的结果

B	A															
	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
–0.4	SL	SL	SL	SL	SL	LC	SQR	W	SQR	SQR	W	H	H	H	H	H
–0.3	SL	SL	SL	SL	SL	SQR	W	SQR	W	SQR	SQR	SQR	SQR	H	SQR	SRE
–0.2	SL	SL	SL	SL	SL	W	W	W	SQR	SQR	SQR	SQR	H	H	H	H
–0.1	SL	SL	SL	SL	SL	SQR	SL	LC	SQR	SQR	H	H	SRE	H	H	H
0	SL	SL	SL	SL	SL	SL	SL	SQR	SQR	SQR	SQR	SRE	W	SQR	H	H
0.1	SL	LC	SL	SL	SL	SL	SQR	W	W	SQR	W	W	SQR	H	H	H
0.2	SL	SL	SL	SL	SL	SL	SQR	SQR	W	SRE	W	SQR	H	H	SQR	H
0.3	SL	SL	SL	SL	SL	SL	SL	W	W	SQR	SQR	H	SQR	H	H	H
0.4	SL	SL	SL	SL	SL	SL	SQR	W	SQR	SQR	H	H	H	H	H	H

附表 4　实验 10 当随机误差取值(σ_1=1/20, σ_2=1/10)的结果

B	A															
	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
–0.4	SL	SL	SL	SL	SL	SL	SL	SL	SQR	SQR	SQR	H	SQR	H	SQR	H
–0.3	SL	SL	SL	SL	SL	SL	SL	LC	W	SQR	SQR	SQR	W	SRE	SRE	SQR
–0.2	SL	SL	SL	SL	SL	W	SL	SQR	SL	W	SQR	SQR	H	H	H	H
–0.1	SL	SL	SL	SL	SL	SL	SL	SL	SQR	SQR	SQR	SQR	SRE	SRE	SQR	SRE
0	SL	SL	SL	SL	SL	SL	W	W	SQR	W	SRE	SQR	SQR	W	SQR	SQR
0.1	SL	SL	SL	SL	SL	SL	SL	SQR	W	SQR	W	SQR	H	SRE	H	H
0.2	SL	SL	SL	SL	SL	SL	SL	W	SL	W	SQR	SQR	SQR	H	SRE	H
0.3	SL	SL	SL	SL	SL	SL	SQR	SQR	SQR	W	SRE	SRE	SQR	SQR	H	SQR
0.4	SL	SL	SL	SL	SL	SL	LC	SL	W	SQR	W	SQR	W	SRE	SRE	SQR

Test Equating with Testlets

WU Rui[1,2], DING Shu-Liang[1], GAN Deng-Wen[1]

([1]*Computer Information Engineer College, JiangXi Normal University, Nanchang, Jiangxi* 330022, *China*)

([2]*East China JiaoTong University Library, Nanchang, Jiangxi* 330013, *China*)

Abstract

The research on test equating is very important for fairness of examnation, item banking, teaching quality assessing and computerized adaptive test. Along with the development of research on examination, testlets have appeared in different examnations increasingly, such as reading comprehension, mathematics, map etc. How to equate tests composed of testlets is a problem we are facing. When item response theory (IRT) models are applied in test equating, strong statistical assumptions—local independence (LI)—must be met. However, previous studies have shown that local independence is likely to be violated when testlets are contained in test. Hence, when equating tests composed of testlets, that local dependence is ignored can lead to distortion of

equating coefficients using standard IRT model.

In order to solve this problem, we use a testlets-based model—2 Parameters Testlet Model (2PTM), which derives from IRT 2 Parameters Logistic Model by adding random-effect parameters associated with each testlet. Local dependence is considered in 2PTM. IRT characteristic curve equating methods and specific procedures for calculating equating coefficients were presented in this paper. In terms of the recovery of estimating the equating coefficients and based on Wilcoxon sign-rank test, a lot of experiments was done using Monte Carlo simulation method. The effectiveness of equating tests containing testlets was investigated under the several conditions, including the accuracy of the estimation of item parameters (AEIP), the number of examinees and the degree of local dependence. The findings of equating tests made up of testlets using 2PTM were compared with standard IRT model—2PLM, which not account for local dependence among items from a common testlet.

Results suggest that 2PTM is better than 2PLM in recovery and have significant differences mostly, so 2PTM is suitable for equating tests based testlets. In addition, the findings of using six different equating criterions for 2PTM were also compared with each other. The results showed that, generally speaking, when the value of the coefficient A is between 0.5 and 0.9, the performance of SLcrit is the best, SQRcrit is proper for $0.9<A<1.5$ and Hcrit is proper for $1.5<A<2.0$. The higher AEIP is, the better SQRcrit and SLcrit perform. Hcrit and SQRcrit are proper for large testlet effect. LCcrit, Wcrit and SREcrit are rarely better than others.

Key words test equating; testlets; item response theory; Monte Carlo simulation; equating criterion