

不同铆测验设计下多维 IRT 等值方法的比较*

刘 玥¹ 刘红云²

(¹四川省教育科学研究所, 成都 610225) (²北京师范大学心理学院, 北京 100875)

摘 要 实际应用中测验往往具有多维结构, 如果仍采用单维 IRT 方法进行等值, 会得到不准确的结果。因此对于多维结构的测验, 需要使用多维 IRT 等值方法来实现参数的转换。基于共同题设计, 文章通过模拟研究的方法, 考察了不同铆测验设计下几种多维 IRT 等值方法的表现, 同时考虑了测验长度、两个维度题目数量的比例、铆测验长度、铆测验的选择策略、两个维度之间的相关和等值群体的能力水平差异六个因素的影响。所比较的多维 IRT 等值方法有: 均值/均值(MM)方法, 均值/标准差(MS)方法, Stoking-Lord (SL)方法, Haebara (HB)方法, 最小平方(LS)方法。结果显示: (1) SL, HB 和 LS 方法得到的等值误差均方根最小, 且在各条件下表现较为稳定。(2) MM 和 MS 方法在非等组条件下呈现出很大的误差均方根。(3)铆测验设计对 SL, HB 和 LS 方法的等值结果没有显著影响。(4)在两个维度之间的相关较高, 测验长度和铆测验长度较长, 等值群体的能力水平没有差异的条件下, SL, HB 和 LS 方法得到的等值误差均方根最小。

关键词 测验等值; 多维 IRT; 均值/均值方法; 均值/标准差方法; Stoking-Lord 方法; Haebara 方法

分类号 B841

1 引言

在教育测验中, 常常会出现考核同一个内容的多个测验形式, 为了实现这些测验分数之间的比较, 会用到测验等值的方法。根据等值所依据的原理, 一般可以分为经典测验理论(CTT)下的等值方法和项目反应理论(IRT)下的等值方法(Kolen & Brennan, 2004)。其中, IRT 等值方法又以其理论完善, 等值关系简洁, 且题目参数具有跨群体不变性等优势, 而受到广泛关注。在大多数的标准测验中, 考生的能力参数和题目参数是由单维 IRT 模型(UIRT)估计得到的, 因此, 基于单维 IRT 模型的等值方法已经得到了较为深入的研究(Kolen & Brennan, 2004)。

然而, 在现实情境里, 测验通常会包含多维结构(Ackerman, 1994; DeMars, 2006; Reckase, 1985)。在很多大型的教育评价项目中, 对于一个概括性的能力的测量, 通常会包含关于多个子学科(子能力)的题目。例如在关于学生科学能力的测验中, 可能

包含分别来自于物理, 地理, 生物等多个学科的题目。这时, 传统 IRT 理论的单维性假设很容易遭到违背。基于单维 IRT 假设的参数估计和 IRT 等值结果会出现一定的偏差(Reckase, 2009; Brossman, 2010)。因此, 许多研究者已经逐渐开展了对基于多维项目反应理论(MIRT)下等值方法的研究和探索。

迄今, 很多研究已经将单维 IRT 下的等值方法推广到多维结构中。这些方法主要有多维均值/均值方法, 均值/标准差方法(Yao, 2011), IRT 相等函数方法, Stoking-Lord (测验特征函数)方法, Haebara (项目特征函数)方法, 直接方法(Oshima, Davey & Lee, 2000), LL 方法(Li & Lissitz, 2000), Min 的方法(Min, 2003), NOP 方法(Reckase & Martineau, 2004)和同时等值的方法(Simon, 2008)等。这些方法和单维 IRT 等值方法的主要区别是, 多维 IRT 等值不仅需要调整不同测验量尺原点和单位大小的差异, 还要进行量尺旋转和维度相关调整等一系列过程(Reckase, 2009)。这些方法之间的主要区别在于它们计算参数转换矩阵的原理和方法不同。一些研究

收稿日期: 2012-09-24

* 国家自然科学基金(31100759); 全国教育科学“十二五”规划教育部重点课题(GFA111001)资助。

通讯作者: 刘红云, E-mail: hylu@bnu.edu.cn

还基于题目参数的返真性对多维 IRT 等值方法进行了比较(Davey, Oshima, & Lee, 1996; Li & Lissitz, 2000; Oshima et al., 2000; Yao & Boughton, 2009)。总的来说, 大多数研究结果表明, Stoking-Lord (SL) 方法和 Haebara (HB) 方法较为稳定, 并且得到的等值结果更加准确(Davey et al., 1996; Oshima et al., 2000; Yao, 2011; Simon & Davison, 2008)。但是 Yon (2007) 的研究也证明, NOP 方法对题目区分度的等值结果比多维 Stoking-Lord 方法准确。Brossman (2010) 在提出多维 IRT 真分数等值和观察分数等值方法后发现, 与单维 IRT 等值的规律相似, 多维 IRT 真分数等值和观察分数等值表现出很高的一致性, 仅在低分段和满分附近出现一定的差异。Simon (2008) 对多维 IRT 同时等值和分别等值方法进行了比较研究, 他们建议, 当两组被试的能力水平差异不大且多维测验各维度间的相关较高时, 同时等值的方法要优于分别等值。总的来说, 关于多维 IRT 等值方法的比较还没有得出较为一致的结论。

进行多维 IRT 等值, 数据结构必须符合以下两个条件之一(Angoff, 1982; Davey et al., 1996): (1) 测验之间包含相同的题目(铆测验); (2) 有一批相同的考生同时参加了两个测验。在实践中, 尤其是大规模测验中, 铆测验设计应用得最为广泛。铆测验本身对等值结果有显著的影响。有学者认为, 铆测验的选择必须尽可能代表整个测验的特征(Cook & Petersen, 1987; Klein & Jarjoura, 1985; Marco, Petersen, & Stewart, 1983)。Kolen 和 Brennan (2004) 曾建议, 铆测验应当包含足够多的题目, 经验的法则是铆题量至少为测验总题量的 20%。Sinharay 和 Holland (2007) 认为, 铆题应当具有内容代表性, 并且难度参数的均值与总测验相等, 只是变异性略小。针对多维 IRT 等值, Yao 和 Boughton (2009) 对同时含有多项选择题和结构性试题测验的多维 IRT 等值进行了考察, 结果发现在铆测验中使用简单结构的题目(题目只属于其中一个维度)或与结构性试题分值相等的多项选择题, 能够提高等值结果的准确性。Yao (2011) 进而提出了一些在多维 IRT 等值中选择铆测验的标准, 其中包括: 铆测验的测验特征函数应当与整个测验接近, 铆测验中各维度题目的数量应当与总测验中各维度题目的数量成比例等。其研究结果证明, 按照所提出的标准选择铆测验, 即使每个维度只有 2 道铆题的情况下, 多维 Stoking-Lord 方法也能得到较好的维度分和总分的

等值结果。

尽管铆测验的选择在等值中至关重要, 但是在多维 IRT 等值中, 很少有研究关注在不同铆测验设计下各种多维 IRT 等值方法的比较(Yao & Boughton, 2009)。另外, 以往的研究主要关注题目参数的等值结果, 很少有研究对能力参数的返真性进行比较。为了充分考察在不同的铆测验设计下, 各等值方法的表现, 从而为实践中多维 IRT 等值总结铆测验的设计原则和选择等值方法提供依据, 本文在不同的铆测验设计下, 采用模拟研究的方法对五种应用较广的多维 IRT 等值方法——均值/均值(MM)方法, 均值/标准差(MS)方法, 多维 Stoking-Lord (SL) 方法, 多维 Haebara (HB) 方法, 最小平方(LS)方法进行比较, 旨在考察铆测验中题目数量和铆测验构成方式对不同等值方法的影响, 主要解决以下几个问题: (1) 在多维 IRT 等值中, 五种常用等值方法的优劣; (2) 铆测验的长度对等值结果的影响; (3) 使用不同的策略选择铆测验(铆测验中各维度题目数量之比与总测验中各维度题目数量之比相同, 铆测验中各维度题目数量相同)是否会对等值结果有影响。同时, 还考虑了测验维度之间的相关和两个群体能力差异大小对等值结果的影响。

2 多维 IRT 模型及等值方法

2.1 两参数多维 IRT 模型

研究使用两参数多维 IRT 模型估计题目和被试参数。假设有 J 名被试, 参加一个由 I 道题目组成的测验, 所有题目均为 0/1 计分。这 J 名被试对 I 道题目的反应可以形成一个 $J \times I$ 的作答反应矩阵 $U = (u_{ji})$ 。每个被试的能力参数是一个向量 $\bar{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jD})$, 其中 D 是假设的维度个数。在由两参数 Logistic 模型推广的两参数多维 IRT 模型中, 给定能力向量为 $\bar{\theta}_j$ 的被试答对题目 i 的概率可以表示为(Reckase, 2009):

$$P(U_{ij} = 1 | \bar{\theta}_j, \bar{a}_i, d_i) = \frac{e^{\bar{a}_i \bar{\theta}_j + d_i}}{1 + e^{\bar{a}_i \bar{\theta}_j + d_i}} \quad (1)$$

其中, $u_{ji}=0$ 或 1 分别表示被试 j 答错或答对题目 i 。 $\bar{a}_i = (a_{i1}, a_{i2}, \dots, a_{iD})$ 是一个含 D 个元素的区分度参数矩阵。 d_i 是第 i 题的截距参数。即:

$$\begin{aligned} \bar{a}_i \bar{\theta}_j + d_i &= a_{i1} \theta_{j1} + a_{i2} \theta_{j2} + \dots \\ &+ a_{iD} \theta_{jD} + d_i = \sum_{k=1}^D a_{ik} \theta_{jk} + d_i \end{aligned} \quad (2)$$

其中, d_i 不同于单维 IRT 模型中的难度参数。多维 IRT 的难度参数定义为:

$$b = \frac{-d}{\sqrt{aa'}} = \frac{-d}{\sqrt{\sum_{v=1}^m a_v^2}} \quad (3)$$

2.2 多维 IRT 等值

与单维 IRT 等值的原理类似, 多维 IRT 等值的目的是找到合适的转换矩阵 A, B , 从而将不同测验的题目和能力参数等值到同一量尺上。使用转换矩阵进行量尺转换的公式为:

$$\bar{a}_i = \bar{a}_i^* \bar{A}^{-1} \quad (4)$$

$$b_i = b_i^* + \bar{a}_i^* \bar{A}^{-1} \bar{B}^T \quad (5)$$

$$\bar{\theta}_j = \bar{\theta}_j^* \bar{A}^T + \bar{B} \quad (6)$$

其中, $\bar{a}_i^*$, b_i^* 和 $\bar{\theta}_j^*$ 表示新测验(new form)的区分度参数矩阵, 难度参数和能力参数矩阵, $\bar{a}_i$, b_i 和 $\bar{\theta}_j$ 表示新测验等值到基础测验(base form)上的区分度参数矩阵, 难度参数和能力参数矩阵。公式与单维 IRT 等值的区别为, A 是一个 $D \times D$ 的矩阵, 表示不同量尺坐标系之间的旋转, B 是一个含有 D 个元素的位置向量, 表示不同量尺坐标之间的校准。

下面, 介绍本研究采用的五种多维 IRT 等值方法。

MM 方法。多维 IRT 等值的 MM 方法和 MS 方法是由 Yao (2011)提出的。该方法基于简单结构的测验, 即测验中的每道题目只属于其中一个维度, 所有题目呈现项目间多维的特征。设 $\mu_a = (\mu_{a1}, \dots, \mu_{aD})$, $\mu_b = (\mu_{b1}, \dots, \mu_{bD})$, $\mu_a^* = (\mu_{a1}^*, \dots, \mu_{aD}^*)$, $\mu_b^* = (\mu_{b1}^*, \dots, \mu_{bD}^*)$ 分别为基础测验和新测验的多维 IRT 区分度参数和难度参数的均值。该方法即要实现新测验对基础测验的量尺转换。转换矩阵 $A_{D \times D}$ 和 $B_{1 \times D}$ 可以表示为:

$$A = \begin{pmatrix} a_1 & 0 & \dots & 0 \\ \dots & & & \\ \dots & & & \\ 0 & \dots & 0 & a_D \end{pmatrix}_{D \times D} \quad (7)$$

$$\bar{B}_{1 \times D} = (b_1, \dots, b_D) \quad (8)$$

与单维 IRT 等值中的 MM 方法类似, 令

$a_k = \frac{\mu_{a_k}^*}{\mu_{a_k}}$, 且有 $b_i = \mu_{b_k}^* - a_k \mu_{b_k}$ 。然后, 可以将转换

矩阵应用于公式(4)到(6)完成两套测验参数之间的量尺转换。

MS 方法。与 MM 方法类似, 设 $\sigma_b = (\sigma_{b1}, \dots, \sigma_{bD})$, $\sigma_b^* = (\sigma_{b1}^*, \dots, \sigma_{bD}^*)$ 分别表示基础测验和新测验中多维 IRT 难度参数的标准差。那么, 转换矩阵 A, B 中的各元素可以通过 $a_k = \frac{\sigma_k^*}{\sigma_k}$ 和 $b_k = \mu_k^* - a_k \mu_k$ 计算得到。

SL 方法。Oshima, Davey 和 Lee (2000)将单维 IRT 中的 Stoking-Lord 方法推广到多维结构的数据中。因此, 对于含有两个维度的多维测验, 多维 SL 等值方法求出将下面函数最小化时的转换矩阵 A, B 。

$$\sum_{q1=1}^{Q1} \sum_{q2=1}^{Q2} w_{q1q2} [T(\theta_{q1}, \theta_{q2}) - T^*(\theta_{q1}, \theta_{q2})]^2 \quad (9)$$

其中, $Q1$ 和 $Q2$ 分别表示两个维度的能力结点数。 w_q 是每个能力结点的权重。 T 表示基础测验中铆测验的测验特征曲线, T^* 表示新测验中铆测验转换到基础测验上的测验特征曲线。

HB 方法。Oshima, Davey 和 Lee (2000)将单维 IRT 中的 Haebara 方法推广到多维结构的数据中。因此, 对于含有两个维度的多维测验, 多维 Haebara 等值方法求出将下面函数最小化时的转换矩阵 A, B 。

$$\sum_{i=1}^I \sum_{q1=1}^{Q1} \sum_{q2=1}^{Q2} w_{q1q2} [E(\theta_{q1}, \theta_{q2}) - E^*(\theta_{q1}, \theta_{q2})]^2 \quad (10)$$

其中 $Q1, Q2$ 和 w_q 的定义与 SL 方法中相同。 E 表示在给定能力水平 θ_{q1} 和 θ_{q2} 的情况下, 基础测验上题目 i 的期望值; E^* 表示在给定能力水平 θ_{q1} 和 θ_{q2} 的情况下, 新测验中铆测验转换到基础测验上题目 i 的期望值。

LS 方法。在 LS 方法中, 题目参数的转换矩阵与之前的方法有一些区别, 其定义如下(Li & Lissitz, 2000):

$$\bar{a}_i = k \bar{a}_i^* T \quad (11)$$

$$\bar{\theta}_j = \frac{1}{k} (T^{-1} \bar{\theta}_j^* - \beta) \quad (12)$$

$$d_i = d_i^* + (\bar{a}_i^* T) \beta \quad (13)$$

$\bar{a}_i^*$, $\bar{\theta}_j^*$, $\bar{a}_i$, $\bar{\theta}_j$ 的意义与公式(4)到(6)中相同, d_i^* 和 d_i 分别表示新测验和新测验等值到基础测验上的截距参数。其中, k 是膨胀系数, T 是正交旋转矩阵, β 是平移向量。在多维 LS 等值方法中, 转换矩阵 T 能够使得两套测验中铆测验上的题目区分度估计值的差异平方和最小化(Li & Lissitz, 2000), 可以采用一般正交 Procrustes 旋转方法计算得到(Schönemann, 1966)。平移向量 β 和膨胀系数 k 的估计使用了最小二乘法。

3 模拟设计

3.1 混合实验设计

在研究中, 由于五种多维 IRT 等值方法均会应

用于每组需等值的模拟数据,因此,这是一个混合实验设计。组内变量为五种等值方法,组间变量为各种模拟条件。组间变量有 6 个。(1)测验长度:测验中所有的题目个数,分别取 40 和 80 两种情况。(2)两个维度题目数量的比例:有两个水平,分别为 1:1(两个维度题目数量相等)和 3:1(两个维度题目数量不等)。(3)铆测验长度:取三个水平,铆测验题目数量分别为全卷的 1/20, 1/5 和 1/3。(4)铆测验的选择策略:考虑了两种方法,即在每个维度中选取相等数量的题目组成铆测验和根据全卷中每个维度题目的数量,按比例选取铆题组成铆测验。(5)两个维度之间的相关:取三个水平 $r=0, 0.5, 0.9$, 分别代表无相关,中等程度相关和高相关。(6)等值群体的能力水平差异:分为两种情况,两组群体能力水平无差异(等组等值)和两组群体能力水平有差异(非等组等值)。在等组等值的情况下,两组被试的能力服从均值为 0, 标准差为 1 的多元正态分布;在非等组等值的情况下,参加基础测验组的被试能力服从均值为 -0.5, 标准差为 1 的多元正态分布,参加新测验组的被试能力仍服从均值为 0, 标准差为 1 的多元正态分布。组间变量(1)到(4)交叉形成了 17 种不同的铆测验设计,具体见表 1。每种铆测验设计下,考虑两个维度之间的相关和等值群体能力的差异,模拟设计共有 $17 \times 2 \times 3 = 102$ 种。

3.2 数据生成

使用 R 语言自编程序产生每种条件下的反应数据。每种条件下包括在基础测验和新测验上,各 3000 名被试对题目的反应,所有的题目均为 0/1 计分。产生数据所依据的模型为多维两参数 Logistic 模型,模型包含两个维度,并且测验符合简单结构。在各条件下,模拟数据的参数分布参照前人文献及实证研究经验。其中 $a_i \sim N(0.8, 0.2^2)$, $b_i \sim N(0, 1^2)$, 该分布与 ETS 进行 SAT 测验的观察分数的边缘分布匹配,故可以与实际数据相比较(Bradlow, Wainer & Wang, 1999)。每种设计条件下数据重复模拟 30 次。

3.3 参数估计和等值

在每种条件下,使用 BMIRT (Yao, 2003)软件,基于贝叶斯框架下的 MCMC 算法完成多维两参数 Logistic 模型和数据的拟合,估计出题目参数和能力参数。在参数估计中,区分度参数,难度参数和能力参数的先验分布采用软件例子中的默认值,分别为 $\mu_a = 1.5, \sigma_a^2 = 1.5, \mu_b = 0, \sigma_b^2 = 1.5, \mu_\theta = 0, \sigma_\theta^2 = 1$ 。

表 1 不同铆测验设计

铆测验设计	全卷题目总数		铆测验题目总数		
	维度 1	维度 2	维度 1	维度 2	维度 1&2
1	20	20	1	1	2
2	20	20	4	4	8
3	20	20	7	7	14
4	30	10	1	1	2
5	30	10	4	4	8
6	30	10	7	7	14
7	30	10	6	2	8
8	30	10	10	4	14
9	40	40	2	2	4
10	40	40	8	8	16
11	40	40	14	14	28
12	60	20	2	2	4
13	60	20	8	8	16
14	60	20	14	14	28
15	60	20	3	1	4
16	60	20	12	4	16
17	60	20	20	8	28

按照等值研究的步骤,研究采用了两步等值法(Kim & Cohen, 1998; Kim & Cohen, 2002)。第一步将新测验上的题目参数和能力参数转换到基础测验的量尺上。第二步将第一步中所有等值后的参数转换到真值的量尺上。在每种等值方法的条件下,第一步和第二步使用的方法完全相同。例如,在第一步中使用了 MM 方法将新测验的参数等值到基础测验的量尺上,在第二步中,以两套测验的所有题目为铆题,再次使用 MM 方法将等值后的参数转换到产生值的量尺上。

对于 MM 和 MS 方法,通过程序 LinkMIRT (Yao, 2004)计算转换矩阵 A 和 B。对于 SL, HB 和 LS 方法,调用 R 中的软件包“Plink”完成等值(Weeks, 2010)。根据前面介绍的两步方法,每次等值时,等值方法需重复使用两次以得到最后的参数结果。

3.4 评价标准

本研究从四个方面评价各参数等值后的返真性:(1)偏差(Bias),绝对偏差(MAE),误差均方根(RMSE),等值后结果与真值的相关(Correlation)。

偏差的意义是总体考察各条件下,各参数等值后结果是否有定向的偏差。其计算公式如下:

$$BIAS = \frac{1}{N} \sum_{n=1}^N \frac{1}{R} \sum_{r=1}^R (\hat{\varsigma}_r - \varsigma_r) \quad (14)$$

其中, $\hat{\zeta}_r$ 表示等值后的各参数值, ζ_r 表示参数的真实值, R 表示全卷的题目数量, N 表示重复的次数。

绝对偏差和误差均方根考察了各条件下, 等值后的各参数值与真实值的差异的大小。这两个值越小, 说明等值后各参数值与模拟的真实值越接近, 等值效果越好。其计算公式如下:

$$MAE = \frac{1}{N} \sum_{n=1}^N \frac{1}{R} \sum_{r=1}^R |\hat{\zeta}_r - \zeta_r| \quad (15)$$

$$RMSE = \frac{1}{N} \sum_{n=1}^N \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\zeta}_r - \zeta_r)^2} \quad (16)$$

公式中各参数表示的意义与公式(14)相同。

等值后结果与真值的相关计算了各条件下, 等值后各参数值与真实值的积差相关。它的意义是考察等值后参数结果与真实值的一致性水平。

4 结果

在参数估计过程中, 所有设计因子的条件下模型均成功收敛。下面, 从 4 个评价标准的角度呈现研究结果。

4.1 偏差

表 2 列出了各条件下各参数等值后的偏差。对于区分度参数, 不同条件下得到的结果基本一致, MM, SL, HB 和 LS 方法的偏差总体较小, 而 MS 方法偏差较大。对于截距参数, 总体来看, SL, HB 和 LS 方法的偏差小于 MM 和 MS 方法; 随着测验长度增加, 各方法下截距参数等值偏差减小; 两个维度的题目数量比例、铆测验长度、铆测验选择策略和测验维度之间的相关对截距参数等值偏差几乎没有明显影响; 等值群体能力水平的差异对 SL, HB 和 LS 方法没有影响, 对 MM 和 MS 方法有一定影响。对于能力参数, SL, HB 和 LS 方法几乎没有偏差, 明显小于 MM 和 MS 方法, 且几乎不受其他因素的影响。MM 和 MS 方法容易受其他因素的影响, 如在非等组等值的情况下, MM 和 MS 方法的偏差很大。

4.2 绝对偏差和误差均方根

由于绝对偏差和误差均方根具有较高的一致性, 因此在本节的结果中, 都参照误差均方根加以比较。

4.2.1 等值方法的比较

多维 IRT 等值方法对参数估计的精度有很大的

影响。SL, HB 和 LS 方法得到的等值后各参数精度较高, 且这三种多维 IRT 等值方法几乎不受其他因素的影响, 在各条件下等值结果相对稳定。MM 方法对区分度参数等值的误差也较小, 但是, MM 方法对截距参数和能力参数的估计误差较大, MS 方法对所有参数等值的误差都较大。

具体来说, 对于题目参数和能力参数, 在所有模拟因子的条件下, 等值方法对误差均方根的影响仅依赖于等值群体的能力水平差异而不同。图 1 表示了等值群体的能力水平差异的不同水平下, 五种等值方法得到的区分度参数、截距参数、能力参数的误差均方根。

从图 1 中可以看出, 在等组等值的条件下, SL, HB 和 LS 方法的结果略优于 MM 和 MS 方法; 随着等值群体的能力水平差异增大, SL, HB 和 LS 方法的结果保持稳定, 而 MM 和 MS 方法的结果出现了较大的偏差。尤其对于截距参数和能力参数, MM 和 MS 方法在非等组条件下的误差明显增大。

综上, 从五种等值方法的比较结果可以看出, SL, HB 和 LS 方法得到误差均方根显著小于 MM 和 MS 方法, 在实际应用中应当作为首选方法。因此, 在下面对铆测验设计的考察中, 仅选取 SL、HB 和 LS 三种等值方法。

4.2.2 铆测验设计对题目参数等值结果的影响

区分度参数。如图 2 所示, 对于区分度参数, 测验长度和铆测验长度对区分度参数估计的精度有较大的影响, 即测验长度越长, 铆测验长度越长, 误差均方根越小。同时, 测验长度和铆测验长度的交互作用对参数估计的精度也有较大的影响, 在测验长度较短的条件下, 铆测验长度仅为全卷的 1/20 时得到的误差均方根远大于其他条件。在相同的测验长度和铆测验长度下, 铆测验选择策略²对参数等值结果没有影响。另外, 根据附表 2~附表 4, 两个维度题目数量的比例对区分度参数等值结果没有显著影响。两个维度之间的相关和等值群体的能力水平差异对区分度参数等值结果没有显著影响。并且, 在维度之间的相关或等值群体的能力水平差异不同的条件下, 测验长度、铆测验长度和铆测验选择策略对参数等值结果的影响不存在差异。

² “铆测验选择策略”这一因子依赖于“两个维度题目数量的比例”。即, 当“两个维度题目数量的比例”相等时, 两种选择策略形成的条件是相同的; 当“两个维度题目数量的比例”不同时, 两种选择策略的比较才有意义。因此, 本研究中对“铆测验选择策略”的考察均针对“两个维度题目数量的比例”不同的情境。

表2 各条件下各参数偏差

组间变量		区分度参数					截距参数					维度 1 能力参数					维度 2 能力参数				
		MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS
相关	0	0.00	0.02	0.00	-0.01	0.01	0.09	0.08	0.11	0.11	0.11	-0.08	-0.08	0.00	0.00	0.00	-0.08	-0.11	0.00	0.00	0.00
	0.5	0.00	0.05	0.00	-0.01	0.01	0.09	0.08	0.11	0.11	0.11	-0.10	-0.10	0.00	0.00	0.00	-0.13	-0.15	0.00	0.00	0.00
	0.9	0.00	0.05	0.00	0.00	0.01	0.09	0.07	0.11	0.11	0.11	-0.09	-0.09	0.00	0.00	0.00	-0.11	-0.13	0.00	0.00	0.00
测验长度	40	0.00	0.04	0.00	-0.01	0.01	0.31	0.37	0.24	0.24	0.24	-0.09	-0.07	0.00	0.00	0.00	-0.11	-0.15	0.00	0.00	0.00
	80	0.00	0.04	0.00	0.00	0.01	-0.10	-0.12	0.00	0.00	0.00	-0.10	-0.10	0.00	0.00	0.00	-0.11	-0.12	0.00	0.00	0.00
两个维度题目数量比例	1:1	0.00	-0.01	0.00	-0.01	0.01	-0.10	-0.12	0.00	0.00	0.00	-0.10	-0.11	0.00	0.00	0.00	-0.09	-0.09	0.00	0.00	0.00
	1:3	0.00	0.06	0.00	0.00	0.01	0.19	0.18	0.17	0.17	0.17	-0.09	-0.08	0.00	0.00	0.00	-0.12	-0.15	0.00	0.00	0.00
	1/20	0.00	0.00	0.00	-0.01	0.01	0.07	-0.10	0.08	0.08	0.08	-0.10	-0.09	0.00	0.00	0.00	-0.09	-0.15	0.00	0.00	0.00
铆测验长度	1/5	0.00	0.06	0.00	0.00	0.01	0.11	0.15	0.13	0.13	0.13	-0.07	-0.06	0.00	0.00	0.00	-0.11	-0.11	0.00	0.00	0.00
	1/3	0.00	0.04	0.00	0.00	0.01	0.08	0.10	0.13	0.13	0.13	-0.11	-0.11	0.00	0.00	0.00	-0.12	-0.13	0.00	0.00	0.00
铆测验选择策略	数量相等	0.00	0.02	0.00	-0.01	0.01	0.06	0.02	0.10	0.10	0.10	-0.11	-0.12	0.00	0.00	0.00	-0.09	-0.10	0.00	0.00	0.00
	按比例	0.00	0.08	0.00	0.00	0.01	0.16	0.19	0.15	0.15	0.15	-0.05	-0.03	0.00	0.00	0.00	-0.15	-0.19	0.00	0.00	0.00
等值群体的能力水平差异	0	0.00	0.02	0.00	0.00	0.01	0.12	0.12	0.11	0.11	0.11	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.5	0.00	0.06	0.00	-0.01	0.01	0.06	0.03	0.11	0.11	0.11	-0.18	-0.16	0.00	0.00	0.00	-0.22	-0.26	0.00	0.00	0.00

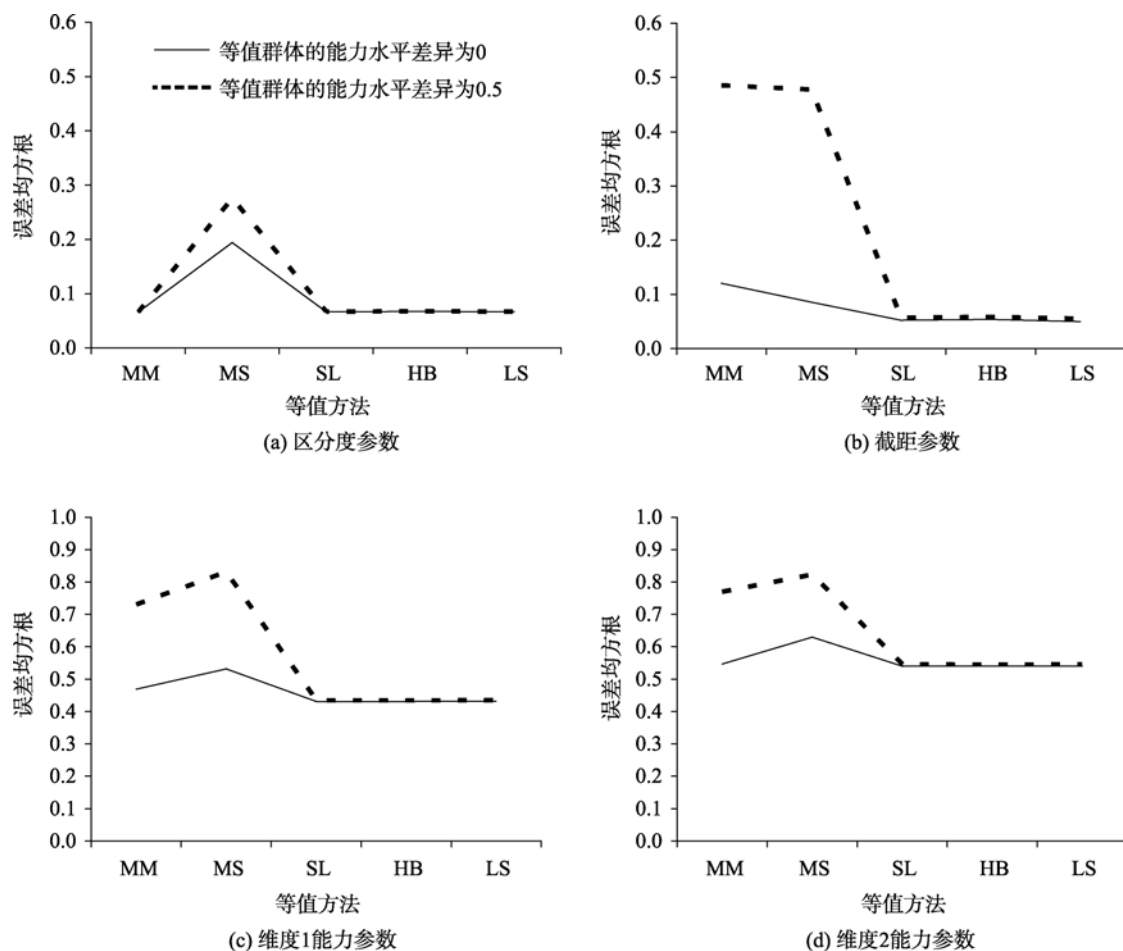


图 1 等值方法与等值群体的能力水平差异交互作用

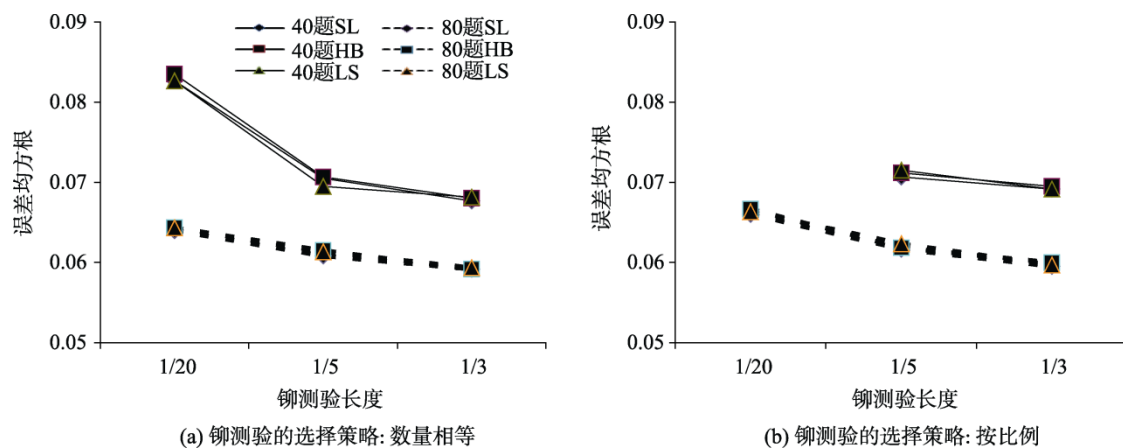


图 2 柳测验设计对区分度参数误差均方根的影响

截距参数。柳测验设计对截距参数的误差均方根的影响与区分度参数类似。如图 3 所示, 测验长度越长, 柳测验长度越长, 截距参数估计的误差均方根越小。另外, 根据附表 1, 附表 2, 附表 4, 两个维度题目数量的比例, 柳测验选择策略和维度之间的相关对截距参数的等值结果没有显著影响。不同的是,

等值群体的能力水平差异对截距参数的等值结果有一定影响, 等值群体的能力水平差异越大, 误差均方根越大。在等值群体的能力水平差异不同的条件下, 柳测验设计对截距参数等值结果的影响是相同的。

4.2.3 柳测验设计对能力参数等值结果的影响

如图 4 所示, 首先, 题目数量较多的维度(维度

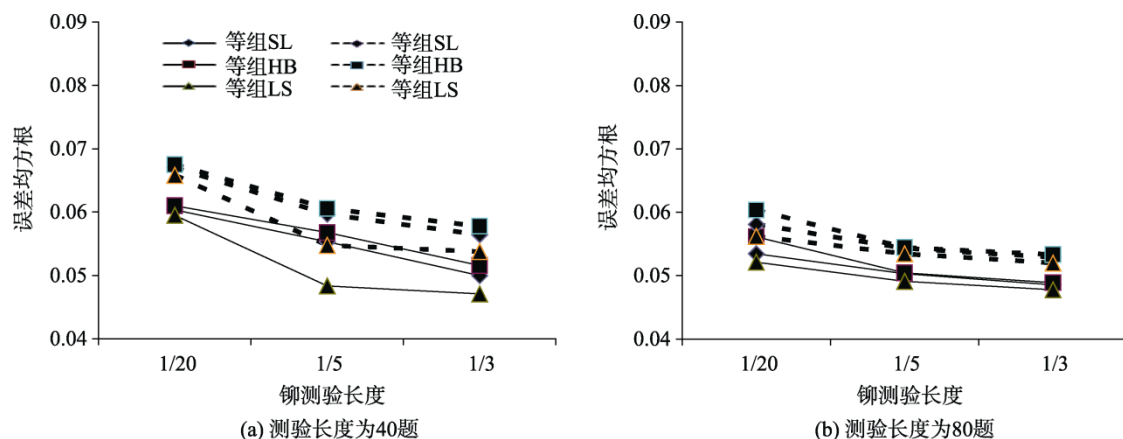


图 3 铆测验设计对截距参数误差均方根的影响

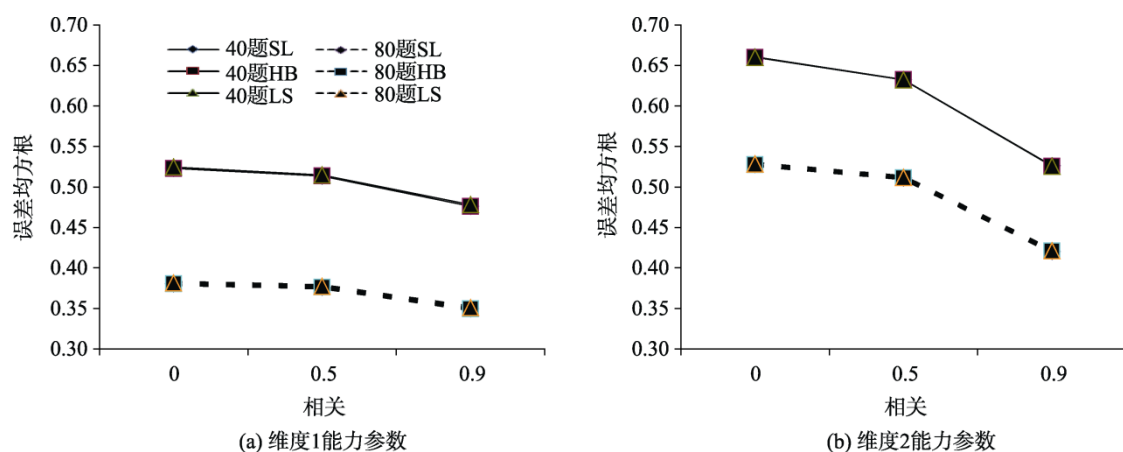


图 4 测验长度和相关对能力参数误差均方根的影响

1), 其能力参数的等值误差均方根要小于题目数量较少的维度(维度 2)。其次, 对于两个维度的能力参数, 测验长度和两个维度之间的相关的影响显著, 即测验长度越长, 相关越高, 误差均方根越小。并且, 对于题目数量较少的维度, 相关对能力参数估计结果的影响更大。测验长度和相关还呈现出交互作用, 当测验长度较短时, 随着相关增加, 误差均方根减小的幅度较大(维度 1 约 0.047, 维度 2 约 0.134); 当测验长度较长时, 随着相关增加, 误差均方根减小的幅度较小(维度 1 约 0.031, 维度 2 约 0.107)。

根据附表 4, 两个维度题目数量的比例对能力参数等值结果有显著影响, 对于维度 1 能力参数, 两个维度题目数量相等时的误差均方根要大于数量不等的条件, 对于维度 2 能力参数, 则与之相反。这是与每个维度题目数量相关的。在该因子的两个水平下, 其他因子对能力参数等值结果的影响具有很强的一致性, 并且, 仅当两个维度题目数量的比

例不同时, 才能进行铆测验选择策略的比较。因此, 为了仔细考察铆测验选题策略对能力参数等值的影响, 本部分以两个维度题目数量不等的条件为例进行分析。表 3 列出了两个维度题目数量比例不同的条件下, SL 方法得到的两个维度能力参数的误差均方根(HB 和 LS 方法的结果与之类似)。

从表 3 中可以看出, 当两个维度题目数量不等时, 两种选题策略下的误差均方根在数值上非常接近。随着铆测验长度的增加, 误差均方根有一定的减少, 尤其对于维度 2 能力参数, 在测验长度较长的条件下, 该现象较为明显。等值群体的能力水平差异对能力参数估计结果没有影响。

4.3 等值后结果与真值的相关

在各种条件下, SL, HB 和 LS 方法等值后各参数值与真实值的相关(区分度参数: 0.923~0.998, 截距参数: 0.997~1.000, 能力参数: 0.697~0.987)大于 MM 和 MS 方法(区分度参数: 0.479~0.998, 截距参数: 0.829~1.000, 能力参数: 0.542~0.943), 且更

表 3 两个维度题目数量比例不同条件下 SL 方法使用两种铆测验选题策略得到的能力参数误差均方根

参数	等值群体的能力水平差异	测验长度	铆测验长度	相关					
				0		0.5		0.9	
				数量相等	按比例	数量相等	按比例	数量相等	按比例
维度 1 能力参数	0	40	1/20	0.496	—	0.486	—	0.467	—
			1/5	0.492	0.488	0.485	0.481	0.462	0.459
			1/3	0.489	0.491	0.482	0.482	0.462	0.461
		80	1/20	0.353	0.351	0.350	0.351	0.333	0.333
			1/5	0.351	0.355	0.348	0.350	0.331	0.337
			1/3	0.351	0.358	0.347	0.348	0.332	0.335
	0.5	40	1/20	0.490	—	0.489	—	0.467	—
			1/5	0.489	0.487	0.485	0.484	0.465	0.461
			1/3	0.489	0.490	0.483	0.484	0.462	0.462
		80	1/20	0.354	0.353	0.353	0.351	0.337	0.336
			1/5	0.354	0.357	0.352	0.354	0.335	0.338
			1/3	0.352	0.358	0.351	0.356	0.334	0.338
维度 2 能力参数	0	40	1/20	0.714	—	0.678	—	0.547	—
			1/5	0.713	0.715	0.676	0.679	0.542	0.545
			1/3	0.714	0.707	0.675	0.674	0.543	0.542
		80	1/20	0.575	0.579	0.552	0.555	0.438	0.446
			1/5	0.566	0.573	0.547	0.55	0.435	0.44
			1/3	0.566	0.565	0.541	0.543	0.434	0.435
	0.5	40	1/20	0.719	—	0.684	—	0.549	—
			1/5	0.718	0.721	0.684	0.688	0.547	0.549
			1/3	0.720	0.716	0.686	0.682	0.547	0.547
		80	1/20	0.581	0.585	0.506	0.566	0.446	0.448
			1/5	0.575	0.579	0.554	0.560	0.441	0.444
			1/3	0.570	0.570	0.550	0.554	0.438	0.441

加稳定。说明 SL, HB 和 LS 方法的最后得到的等值结果与真实值存在较高的一致性。MM 和 MS 方法得到的相关系数变异较大, 在一些条件下, 它们的结果甚至不可接受。对于 SL, HB 和 LS 方法, 铆测验设计对区分度参数和能力参数相关值的影响与误差均方根的结果类似, 而截距参数的相关在各条件下普遍较高(在 0.998 左右), 不受其他模拟因子的影响。

5 讨论与结论

本研究基于含有两个维度的多维补偿性 IRT 模型模拟数据, 比较了五种多维 IRT 等值方法在各铆测验设计下的表现, 研究结果可以为多维 IRT 等值方法的比较及其影响因素的探索提供更详细的信息。

首先, 通过对等值方法比较可以看出, SL, HB 和 LS 方法得到的等值结果比 MM 和 MS 方法准确,

特别是在非等组等值的情况下, 其优势更为明显。在各模拟条件下, SL, HB 和 LS 方法三种等值方法得到的结果较为稳定, 这不仅与之前单维 IRT 等值的研究结果相呼应(Hanson & Béguin, 2002), 也与前人关于多维 IRT 等值的已有研究结果一致(Davey et al., 1996; Oshima et al., 2000; Yao, 2011)。SL, HB 和 LS 方法的优势可能仍在于它们在计算转换矩阵时, 同时考虑了所有的题目参数。然而, 除了 MM 方法对区分度参数等值的误差均方根较小之外, MM 和 MS 方法在大多数情况下表现很差。其中, MS 方法的误差均方根最大, 这可能一方面由于该方法在等值过程中只考虑了难度参数, 而另一方面, 计算转换矩阵时需要用到难度参数的标准差, 这个统计量本身不如平均数稳定, 并且当铆题数量较小时, 其标准差也极不准确。在应用 LinkMIRT 软件进行 MM 和 MS 方法等值时, 计算出的转换矩阵 A 中的元素有时很大。这与其他方法的结果存在显著

差异。回顾 MM 和 MS 方法的原理, 可知转换矩阵 A 为对角矩阵, 即对角线之外的元素均为 0。但是我们在估计参数时, 没有加入维度之间独立的假设, 所以在计算转换矩阵时理应考虑维度之间的相关。这两种方法没有能够将维度之间的关系纳入转换矩阵的计算中, 这可能是它们的结果出现异常值的一个原因。因此, 在实际应用中最好选用 SL, HB 和 LS 方法进行多维 IRT 等值。

其次, 在选择 SL, HB 和 LS 方法的前提下考虑不同铆测验设计的影响。对于第一个研究问题, 铆测验长度对区分度参数, 截距参数都有一定的影响。当测验长度较短时, 如果铆测验长度仅为全卷的 $1/20$, 得到的结果最差。随着铆测验长度增加, 参数等值后的误差均方根减小。但当铆测验长度增加到全卷的 $1/5$ 以后, 等值误差均方根没有明显的减小。在单维 IRT 中, Harris 和 Crouse (1993) 建议铆测验长度不需要太长, 经验的标准是铆测验至少需包含 8 道题目或者占全卷的 20% (Sykes, 1997)。在本研究中, 对于 SL, HB 和 LS 方法, 虽然铆测验长度对等值误差均方根的影响在统计检验上达到了显著水平, 但是该变量在数值上的影响较为微弱。这与 Yao (2011) 的研究结果类似, 在多维 IRT 等值中, 如果选择了较好的等值方法, 等值结果准确性将几乎不受铆测验长度影响。这可能是在数据模拟与模型拟合过程中, 均采用了多维补偿性模型的缘故。该结果对于实际应用具有重要的意义, 一旦选择了良好的多维 IRT 等值方法, 只需要较少数量的铆题就能得到较准确的等值结果。这样既能够减少编制铆题时花费的人力物力, 还能在一定程度上缩短测验长度, 提高效率。对于第二个研究问题, 不同的铆测验选择方法对 SL, HB 和 LS 方法没有显著影响。在构建铆测验时, 基本要求是它应当是整个测验的一个“微缩版本” (Kolen & Brennan, 2004)。很多研究者关注了铆测验在内容上的代表性 (Cook & Petersen, 1987)。他们认为, 使用具有内容代表性的铆测验能够去除等值偏差, 并且增加估计的等值系数的精确性。因此本研究假设, 如果按照两个维度题目数量的比例选择铆题, 会得到较好的等值结果。研究结果没有证明之前的假设, 这其实也没有违背对铆测验内容代表性的要求。因为在研究设计中, 每个维度的铆题都符合简单结构。可以推测, 如果某些维度不含铆题, 或者含有的铆题均同时属于多个维度 (项目内多维), 那么将会违背铆测验具有内容代表性的原则, 从而导致有偏差的

等值结果。另外, 铆测验选择策略的问题, 对于单个维度来说就转化为铆测验长度的问题, 只要对于每个维度来说, 简单结构的铆题数量达到了一定的比例, 就能得到较稳定的等值结果。

最后, 在大多数关于多维 IRT 等值的研究中, 都只考察了等值方法对题目参数的返真性 (Oshima et al., 2000; Yao & Boughton, 2009; Simon, 2008)。但是, 有些情况下等值的目的是将参加不同测验考生的能力水平调整到同一量尺上。因此, 本研究还包括了对各种等值方法下, 能力参数返真性的比较。对于 SL, HB 和 LS 方法, 两个维度能力参数的等值结果普遍受到相关和测验长度的影响。另外, 题目数量少的维度能力参数等值结果较差, 并且容易受到铆测验长度和等值群体的能力水平差异的影响。说明对于题目数量少的维度, 更应该注意对铆测验的选择。总的来说, 即使对于这三种方法, 能力参数等值后的误差均方根也较大。为了探索原因, 以铆测验设计 9~11 为例 (在此情况下测验长度最长, 且两个维度的题目数量相等, 参数估计准确性应当最高), 计算了 BMIRT 对能力参数估计值的误差均方根。结果显示, SL, HB 和 LS 方法能力参数估计的误差均方根在 0.375 到 0.436 之间, 而参数的等值误差均方根在 0.375 到 0.439 之间。因此, 能力参数等值后的误差很可能大部分来源于参数估计的误差。迄今, 很多研究者已经对多维 IRT 模型的参数估计进行了深入的研究 (de la Torre, 2009; Zhang & Stone, 2004; Bolt & Lall, 2003; Yao, 2003), 并编写了一些估计软件。其中大部分能得到较为准确的题目参数估计结果, 而能力参数估计结果并不理想。因此, 采用合适的估计方法减小模型对能力参数估计的误差, 可能是提高多维 IRT 能力参数等值准确性的一个有效途径。

现在, 多维 IRT 模型在教育测量领域得到了越来越多的应用。多维 IRT 等值自然也成为了实际中急需实现的方法。本研究探讨了五种常用的多维 IRT 等值方法在不同铆测验设计下的表现, 对多维测验中多维 IRT 等值方法的选择和铆测验的设计具有一定的实践意义。第一, 在编制多维 IRT 测验时, 应尽量保证每个维度都有充足的题目数量, 维度之间具有较高的相关。第二, 在设计多维 IRT 等值的铆题时, 应当保证每个维度都有一定数量的铆题, 并且尽可能使用简单结构的铆题。另外, 各维度铆题的分布不一定需按照全卷各维度题目数量的比例, 但需要达到每个维度铆测验长度比例的最低标

准。第三,在等值方法上选择使用 SL, HB 和 LS 方法,尽量避免 MM 和 MS 方法。本研究还存在一定的局限性。如研究数据均模拟 0/1 计分的题目,并且被试的能力分布也满足多元正态分布,在实际中会出现等级评分的题目甚至混合类型的题目,并且被试群体的能力也会呈现出各种分布形态,有必要在以后的研究中对这些更加复杂的情况加以讨论。此外,研究仅讨论了对于维度能力的等值,实际中可能还需要实现对总能力的等值。因此结合 Yao (2011)提出的方法,在各条件下对多维 IRT 维度能力和总能力等值的结果进行比较,是值得进一步研究的问题。

本研究得到的主要结论如下:

第一,在本研究设置的各个模拟条件下,多维 SL, HB 和 LS 等值方法得到的等值结果比 MM 和 MS 方法准确,并且这三种方法表现更加稳定。在等值群体的能力水平存在差异的情况下,多维 MM 和 MS 等值方法得到的等值结果误差很大。因此,在等值方法选择上,推荐使用多维 SL, HB 和 LS 等值方法。

第二,对于多维 SL, HB 和 LS 等值方法:(1)测验长度对各参数的等值结果影响显著,测验长度越长,等值结果的误差均方根越小;(2)铆测验长度对各参数的等值结果有一定影响,铆测验长度越长,等值误差均方根越小;(3)铆测验选择策略对各参数等值结果没有显著的影响;(4)维度之间的相关对区分度参数和截距参数的等值结果没有影响,对于能力参数,随着相关的增加等值误差均方根减小;(5)等值群体的能力水平差异仅对截距参数有一定影响,在非等组的条件下得到的等值误差均方根大于等组的条件。

参 考 文 献

- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4), 255-278.
- Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. *Test Equating*, 55, 69.
- Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27(6), 395-414.
- Bradlow, E. T., Wainer, H., & Wang, X. H. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153-168.
- Brossman, B. G. (2010). Observed score and true score equating procedures for multidimensional item response theory. University of Iowa.
- Cook, L. L., & Paterson, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, 11(3), 225-244.
- Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, 20(4), 405-416.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, 33(8), 620-639.
- DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to test let-based tests. *Journal of Educational Measurement*, 43(2), 145-168.
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3-24.
- Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6(3), 195-240.
- Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22(2), 131-143.
- Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, 26(1), 25-41.
- Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement*, 22(3), 197-206.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer Verlag.
- Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, 24(2), 115-138.
- Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). *A large scale evaluation of linear and curvilinear score equating models Volume I* (RM-83-2). Princeton, NJ: Educational Testing Service.
- Min, K. S. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations*. Michigan State University, Department of Counseling, Educational Psychology, and Special Education.
- Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, 37(4), 357-373.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401-412.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer Verlag.
- Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Committee on Test Design for K-12 Science Achievement. Washington, DC.
- Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1), 1-10.
- Simon, M. K. (2008). *Comparison of concurrent and separate multidimensional IRT linking of item parameters*.

- University of Minnesota.
- Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed. *Journal of Educational Measurement*, 44(3), 249–275.
- Sykes, R. C. (1997). *Guidelines for the selection of anchor items for mixed (or single) item format tests*. Monterey, CA: CTB/McGraw-Hill.
- Weeks, J. P. (2010). Plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software*, 35(12), 1–33.
- Yao, L. H. (2003). *BMIRT: Bayesian multivariate item response theory*. Monterey, CA: CTB/McGraw-Hill.
- Yao, L. H. (2004). *LinkMIRT: Linking of multivariate item response model*. Monterey, CA: Defense Manpower Data Center.
- Yao, L. H. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement*, 35(1), 48–66.
- Yao, L. H., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, 46(2), 177–197.
- Yon, H. (2007). Multidimensional item response theory (MIRT) approaches to vertical scaling. Michigan State University.
- Zhang, B., & Stone, C. (2004). *Direct and indirect estimation of three-parameter compensatory multidimensional item response models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Comparison of MIRT Linking Methods for Different Common Item Designs

LIU Yue¹; LIU Hongyun²

(¹ Sichuan Educational Assessment and Evaluation Centre, Sichuan Institute Of Education Sciences, Chengdu 610225, China)

(² School of Psychology, Beijing Normal University, Beijing 100875, China)

Abstract

A great number of educational assessments usually measure more than one trait (Ackerman, 1992; DeMars, 2006; Reckase, 1985). In order to adjust scores on these different test forms, multidimensional item response theory (MIRT) and its linking procedures should be developed. So far, some researchers have already extended UIRT linking methods to the multidimensional structure (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Yon, 2006). There were numerous studies comparing MIRT linking methods in the literature. However, although choosing anchor items was of great importance in common item designs, a few of studies compared MIRT linking methods under different common item designs. It was still in doubt that, how we could choose the common items across different MIRT linking methods. The purpose of this study was to compare five MIRT linking methods under two kinds of common item choosing strategies in various situations.

The study was a mixed measure design of simulation conditions (between-factors) and linking methods (within-factor). There were six between-factors: (1) 2 test lengths (40 items and 80 items); (2) 2 levels of the proportion of the number of items in one dimension to another (1:1 and 1:3); (3) 3 anchor lengths (1/20, 1/5 to 1/3 of total test); (4) 2 strategies of choosing common items (averagely choosing the items in all dimensions or choosing according to the proportions of items in every dimension); (5) 3 correlations between two ability dimensions ($r=0, 0.5, 0.9$); (6) 2 levels of equivalent/non-equivalent ability levels between two populations. The five MIRT linking methods we investigated were: Mean/Mean (MM) method, Mean/Sigma (MS) method, Stoking-Lord's (SL) method, Haebara's (HB) method and Least Square (LS) method. Under each condition, the number of examinees was fixed as $I=2000$, and 30 replications were generated. BMIRT (Yao, 2003) was applied to estimate item and ability parameters using an MCMC method. Based on the previous studies about equating (Kim & Cohen, 1998; Kim & Cohen, 2002), a two-step of linking was applied. The first step was to transform the scale of parameters in the new test onto the base test, and the second step was to transform the scale of all the simulated items onto the generating scale. In each step, the transformation matrices were produced by LinkMIRT (Yao, 2004) and the R package called "Plink" (Weeks, 2010). Finally, the recovery of parameters was evaluated by four criteria: bias, mean absolute error, root mean square error, correlation between the parameters after

equating and true values.

To compare the five MIRT equating methods, the results showed that: the RMSE for parameters under SL, HB and LS methods were smaller and more stable in different situations; however, the RMSE for parameters in MM and MS methods were significantly large, especially in non-equivalent group conditions. Therefore, the latter results were displayed for the SL, HB and LS methods. It was found that these methods were not affected by the common item design. It meant that in multidimensional linking, if the number of common items was more than 5% of the total test, the RMSE became acceptable. Meanwhile, the strategy of choosing common items didn't have significant influence on the linking results of the three methods across different conditions of test structure. Moreover, for other simulation factors: as test length increased, the RMSE of these methods decreased; as the correlations between two ability dimensions increased, the RMSE of ability parameter decreased; the difference of ability levels between two populations had smaller effect on these methods, that only for intercept parameter, the non-equivalent group condition produced larger error.

In conclusion, SL, HB and LS methods generally performed better than the other two methods across all conditions, so it was highly recommended to use these methods in practical. The performances of SL, HB and LS methods were similar under different common item designs, which was amazing for MIRT linking. Once an appropriate method was chosen, shorter anchor set could be applied, as developing good common items for multidimensional tests was quite time-consuming. Meanwhile, the common items could be chosen either according to the proportions of items in every dimension or averagely in all dimensions. This might be more convenient for practitioners as well. Lastly, as test length had significant effect on the accuracy of equated parameters, it was suggested to make sure the test was comprised of enough items in every dimension before conducting an MIRT linking.

Key words test equating; Multidimensional Item Response Theory; Mean/Mean (MM) method; Mean/Sigma (MS) method; Stoking-Lord's (SL) method; Haebara's (HB) method; Least Square (LS) method

附表 1 题目数量比例不同条件下各铆测验选题策略的误差均方根

测验长度	铆测验长度	选题策略	区分度参数					截距参数					维度 1 能力参数					维度 2 能力参数				
			MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS
40	1/5	数量相等	0.068	0.243	0.071	0.071	0.070	0.333	0.308	0.059	0.060	0.051	0.665	0.809	0.480	0.480	0.480	0.753	0.865	0.647	0.646	0.645
		按比例	0.070	0.251	0.071	0.071	0.072	0.306	0.304	0.055	0.055	0.051	0.651	0.795	0.477	0.477	0.479	0.746	0.872	0.650	0.648	0.648
	1/3	数量相等	0.067	0.195	0.068	0.068	0.068	0.312	0.301	0.054	0.055	0.051	0.639	0.714	0.478	0.478	0.480	0.759	0.820	0.648	0.648	0.647
		按比例	0.068	0.227	0.069	0.070	0.069	0.311	0.300	0.053	0.054	0.050	0.654	0.735	0.478	0.479	0.480	0.740	0.853	0.645	0.644	0.643
80	1/20	数量相等	0.064	0.223	0.064	0.064	0.064	0.257	0.251	0.055	0.056	0.053	0.486	0.632	0.347	0.347	0.348	0.634	0.692	0.525	0.525	0.526
		按比例	0.066	0.315	0.066	0.067	0.066	0.286	0.262	0.056	0.057	0.055	0.514	0.655	0.345	0.346	0.347	0.633	0.749	0.530	0.529	0.530
	1/5	数量相等	0.060	0.199	0.061	0.062	0.061	0.278	0.261	0.052	0.051	0.052	0.499	0.632	0.345	0.345	0.347	0.633	0.676	0.520	0.520	0.520
		按比例	0.061	0.245	0.062	0.062	0.062	0.287	0.283	0.054	0.054	0.052	0.527	0.608	0.349	0.349	0.350	0.627	0.593	0.524	0.524	0.524
	1/3	数量相等	0.058	0.141	0.059	0.059	0.059	0.287	0.280	0.050	0.051	0.050	0.500	0.557	0.345	0.345	0.346	0.634	0.650	0.517	0.517	0.517
		按比例	0.059	0.150	0.060	0.060	0.060	0.294	0.294	0.052	0.052	0.051	0.521	0.570	0.350	0.350	0.351	0.622	0.656	0.518	0.518	0.519

附表 2 各维度间相关条件下各参数误差均方根

测验长度	维度间相关	区分度参数					截距参数					维度 1 能力参数					维度 2 能力参数				
		MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS
40	0	0.072	0.217	0.073	0.073	0.072	0.290	0.253	0.057	0.057	0.053	0.669	0.740	0.523	0.523	0.525	0.731	0.796	0.661	0.660	0.660
	0.5	0.071	0.229	0.072	0.072	0.072	0.371	0.342	0.058	0.060	0.054	0.735	0.848	0.513	0.514	0.515	0.766	0.868	0.633	0.632	0.632
	0.9	0.071	0.221	0.073	0.073	0.072	0.333	0.303	0.058	0.059	0.055	0.670	0.760	0.476	0.476	0.478	0.650	0.761	0.526	0.526	0.526
80	0	0.061	0.235	0.062	0.062	0.062	0.238	0.231	0.052	0.053	0.051	0.489	0.573	0.380	0.380	0.381	0.609	0.652	0.528	0.528	0.528
	0.5	0.062	0.251	0.062	0.063	0.063	0.312	0.304	0.054	0.055	0.053	0.555	0.656	0.376	0.376	0.378	0.657	0.694	0.511	0.511	0.512
	0.9	0.061	0.244	0.062	0.062	0.062	0.285	0.276	0.053	0.054	0.052	0.510	0.613	0.350	0.350	0.351	0.558	0.614	0.421	0.421	0.421

附表 3 各能力差异条件下各参数误差均方根

测验长度	能力差异	区分度参数					截距参数					维度 1 能力参数					维度 2 能力参数				
		MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS
40	0	0.071	0.220	0.072	0.073	0.072	0.153	0.097	0.055	0.056	0.051	0.569	0.646	0.503	0.504	0.505	0.608	0.721	0.604	0.604	0.604
	0.5	0.071	0.224	0.072	0.073	0.072	0.509	0.501	0.060	0.061	0.057	0.814	0.919	0.505	0.505	0.507	0.823	0.896	0.609	0.608	0.608
80	0	0.061	0.177	0.062	0.062	0.062	0.091	0.078	0.051	0.052	0.050	0.380	0.455	0.367	0.367	0.368	0.493	0.570	0.483	0.483	0.484
	0.5	0.062	0.310	0.062	0.063	0.063	0.465	0.463	0.055	0.056	0.054	0.656	0.773	0.371	0.371	0.372	0.723	0.761	0.490	0.490	0.490

附表 4 数量相等选题策略下各维度题目数量的误差均方根

测验长度	测验长度	维度题目数量	区分度参数					截距参数					维度 1 能力参数					维度 2 能力参数				
			MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS	MM	MS	SL	HB	LS
40	1/20	1:1	0.080	—	0.081	0.082	0.079	0.368	—	0.063	0.063	0.062	0.788	—	0.548	0.547	0.550	0.654	—	0.541	0.540	0.543
		3:1	0.082	—	0.083	0.084	0.083	0.422	—	0.065	0.065	0.064	0.757	—	0.483	0.483	0.487	0.746	—	0.649	0.649	0.648
	1/5	1:1	0.069	0.229	0.071	0.071	0.070	0.300	0.284	0.058	0.061	0.053	0.695	0.848	0.546	0.546	0.546	0.663	0.731	0.540	0.539	0.540
		3:1	0.068	0.243	0.071	0.071	0.070	0.333	0.308	0.059	0.060	0.051	0.665	0.809	0.480	0.480	0.480	0.753	0.865	0.647	0.646	0.645
	1/3	1:1	0.066	0.189	0.067	0.067	0.066	0.297	0.299	0.053	0.055	0.051	0.685	0.795	0.545	0.545	0.546	0.664	0.711	0.534	0.534	0.535
		3:1	0.067	0.195	0.068	0.068	0.068	0.312	0.301	0.054	0.055	0.051	0.639	0.714	0.478	0.478	0.480	0.759	0.820	0.648	0.648	0.647
80	1/20	1:1	0.066	0.600	0.066	0.066	0.067	0.288	0.292	0.056	0.062	0.054	0.534	0.707	0.414	0.413	0.414	0.596	0.643	0.415	0.418	0.417
		3:1	0.064	0.223	0.064	0.064	0.064	0.257	0.251	0.055	0.056	0.053	0.486	0.632	0.347	0.347	0.348	0.634	0.692	0.525	0.525	0.526
	1/5	1:1	0.060	0.158	0.061	0.061	0.061	0.258	0.252	0.052	0.053	0.050	0.538	0.584	0.413	0.412	0.413	0.544	0.601	0.414	0.414	0.415
		3:1	0.060	0.199	0.061	0.062	0.061	0.278	0.261	0.052	0.051	0.052	0.499	0.632	0.345	0.345	0.347	0.633	0.676	0.520	0.520	0.520
	1/3	1:1	0.059	0.161	0.060	0.060	0.060	0.270	0.258	0.050	0.051	0.049	0.544	0.581	0.413	0.413	0.414	0.549	0.637	0.414	0.415	0.415
		3:1	0.058	0.141	0.059	0.059	0.059	0.287	0.280	0.050	0.051	0.050	0.500	0.557	0.345	0.345	0.346	0.634	0.650	0.517	0.517	0.517