心理学探新 2015,Vol. 35,No. 1,56 - 61
PSYCHOLOGICAL EXPLORATION

多维数据 IRT 真分数等值和 IRT 观察分数等值研究

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,如果仍采用单维方法进行等值,会得到不准确的结果。研究基于随机等组设计下英语测验,使用 MCMC 方法估计题目参数,将单维 IRT 真分数等值和观察分数等值方法推广到多维。比较了四种等值方法: 单维 IRT 真分数等值和观察分数等值,多维近似单维 IRT 真分数等值和观察分数等值。结果显示,当数据符合多维结构时: (1) 基于多维测验的 IRT 真分数等值和观察分数等值方法优于单维 IRT 真分数等值和 IRT 观察分数等值方法; (2) 多维 IRT 观察分数等值略优于多维 IRT 真分数等值,但是两者之间的差异较小。

关键词: 测验等值; 多维 IRT; 真分数等值; 观察分数等值; MCMC 估计
中图分类号: B841.2 文献标识码: A 文章编号: 1003 - 5184(2015)01 - 0056 - 06

1 问题提出

在教育测量中,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。针对测验分数的等值,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 真分数等值和观察分数等值就是两种经典的实现测验分数等值的方法。它们既能与传统观察分数等值方法的目的一致,实现测验分数之间的转换,又能结合 IRT 等值的优势,使等值后的项目参数在同一量尺上,为题库建设中铆定新加入题目的参数提供了便利。IRT 真分数等值是当项目参数都被置于同一量度上之后,将两个测验的真分数通过被试的能力值 θ 进行链接(Kolen & Brennan,2004)。IRT 观察分数等值是产生两个测验的观察分数分布。然后,使用传统的等百分位等值方法来进行等值(Kolen & Brennan,2004)。但是,基于 IRT 的等值方法往往需要测验结构满足单维性的前提假设。

然而,在现实情境里,测验通常包含多维的结构。如英语测验,就能根据内容分为阅读,听力,写作等维度。这时,传统 IRT 理论的单维性假设很容易遭到违背。因此,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)。有很多研究者已经致力于开发适用于多维 IRT 的等值方法。这些方法主要有多维 IRT 相等函数方法,测验特征函数方法,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 & Davison,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)。

在单维 IRT 等值中,一些研究比较了 IRT 真分数等值与 IRT 观察分数等值(Harris & Crouse,1993; Han,Kolen,& Pohlmann,1997; Lord & Wingersky,1984; 刘玥,骆方,刘红云,2010)。尽管关于两种等值方法是否有区别存在不一致的结论,但是大多数研究证明,IRT 真分数等值与 IRT 观察分数等值的结果有极高的相似性。在多维 IRT 等值中,大多研究关注于项目参数的等值,很少有研究应用针对测验分数等值。Brossman(2010)首次将单维 IRT 真分数等值和观察分数等值推广到多维,并对这些方法进行比较。结果证明,对于存在中等程度多维的数据,几种多维 IRT 等值方法优于单维 IRT 等值。

在 Brossman(2010)的研究中,参数估计使用的是边缘极大似然估计方法。随着统计方法和计算机技术的发展,贝叶斯估计的 MCMC 算法以其估计结果的准确性得到了越来越多的应用(Yao,Lewis,& Zhang,2008)。因此,基于贝叶斯估计得到的项目参数,进行单维和多维 IRT 分数等值,其结果是否存在差异,是研究主要关心的问题。多维近似单维 IRT 真分数等值(unidimensional approximation of MIRT true score equating)和多维近似单维 IRT 观察分数等值(unidimensional approximation of MIRT observed score equating),因计算过程相对简单,等值效果较好,并且等值的项目参数与单维 IRT 的结果具有可比性,而具有较大的优势(Brossman,2010)。因此,选用这两种方法作为多维 IRT 等值方法。等百分位等值不包含多维性假设,并且在相等组设计中具有良好稳定的结果,所以等百分位等值将作为其他几种方法的比较标准(Brossman,2010)。综上,研究以实际数据为背景,基于贝叶斯估计的 MCMC 方法实

现参数估计，比较了四种等值方法：单维 IRT 真分数等值，单维 IRT 观察分数等值，多维近似单维 IRT 真分数等值，多维近似单维 IRT 观察分数等值。研究丰富了多维 IRT 的等值方法，为实际中针对测验分数的等值方法的选择提供了参考。

2　研究方法

2.1　实验数据

研究采用 2007 年国家教育质量分析评估大型初中英语抽样测试的数据。该英语测验分为 A，B 卷。测试采用相等组等值设计，即同一所参加测试的学生随机分为两组，一组测试 A 卷，一组测试 B 卷。因此估计出的两套测验的项目参数在同一量尺上，项目参数不需要进行量尺转换。每套测验均由听力和阅读两个部分组成，共 40 题。根据测验内容，可以假设题目分别属于两个维度。所有题目均为 0/1 计分，测验总分为原始分。

完成测验 A 的有 3242 名考生，完成测验 B 的有 3308 名考生。研究要进行测验 B 到测验 A 的分数等值。

2.2　等值方法

2.2.1　多维 IRT 真分数等值

多维 IRT 真分数等值主要是通过将多维 IRT 的参数估计结果合成单维参数，从而采用与单维 IRT 真分数等值类似的过程完成，因此又称为多维近似单维 IRT 等值方法(Brossman，2010)。

首先，进行多维两参数 Logistic 模型的参数估计。

然后，计算每个维度的权重。

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{a}_{jk}}{\sqrt{\sum_{k=1}^{\delta} \left(\sum_{j=1}^{N} \hat{a}_{jk} \right)^2}} \tag{1}$$

$k = 1, 2, \cdots, K$，K 表示维度数，$\overset{\circ}{\alpha}_k$ 表示第 k 个维度的权重，$\overset{\circ}{\alpha}_k$ 表示系数的线性组合中的第 k 个值，$\hat{a}_{jk}$ 表示第 k 个维度上第 j 道题目的区分度，N 表示题目总数，δ 表示维度总数。

利用权重合成多维近似单维项目参数。

$$\hat{a}_{\alpha j} = \left(1 + \hat{\sigma}_{\alpha j}^2\right)^{-\frac{1}{2}} \hat{a}_j^T \hat{\Sigma} \hat{\alpha}_{jk} \tag{2}$$

$$\hat{d}_{\alpha j} = \left(1 + \hat{\sigma}_{\alpha j}^2\right)^{-\frac{1}{2}} \hat{d}_j \tag{3}$$

$$\hat{b}_{\alpha j} = \frac{-\hat{d}_{\alpha j}}{\hat{a}_{\alpha j}} \tag{4}$$

$$\hat{\sigma}_{\alpha j}^2 = \hat{a}_j^T \hat{\Sigma} \hat{a}_j - \left(\hat{a}_j^T \hat{\Sigma} \hat{\alpha}_{jk}\right)^2 \tag{5}$$

$\hat{\alpha}_j$ 是区分度参数估计值，$\hat{\alpha}_{jk}$ 表示第 j 道题目所对应的维度权重 $\hat{\alpha}_k$，$\hat{a}_j^T$ 是区分度参数估计值，$\hat{d}_j$ 是难度参数估计值，$\hat{\Sigma}$ 是能力估计的协方差矩阵。

然后根据下面的公式将正态肩形模型系统中的上述参数转换到 Logistic 模型中(Lord，1980)。

$$p_{\alpha ij}(\theta_{\alpha i}) = \hat{c} + (1 - \hat{c}_j)\, \Phi\left(\hat{a}_{\alpha j}(\theta_{\alpha i} - \hat{b}_{\alpha j})\right) \approx \hat{c}_j + (1 - \hat{c}_j) \frac{e^{1.7 \hat{a}_{\alpha j}(\theta_{\alpha i} - \hat{b}_{\alpha j})}}{1 + e^{1.7 \hat{a}_{\alpha j}(\theta_{\alpha i} - \hat{b}_{\alpha j})}} \tag{6}$$

这时，多维近似单维能力也可以表示为各个维度能力参数的线性组合。

$$\theta_{\alpha i} = \sum_{k=1}^{\delta} \alpha_k \theta_k \tag{7}$$

最后，利用多维近似单维 IRT 题目参数，就能实现多维近似单维 IRT 真分数等值。

2.2.2　多维 IRT 观察分数等值

多维 IRT 观测分数等值通过将多维测验中每个维度能力的结点值转换到单维能力结点值，然后采用与单维 IRT 观测分数等值类似的过程完成，又称为多维近似单维 IRT 观察分数等值，该方法需要得到被试能力的边缘分布(Brossman，2010)。可以按照下面的方法求出被试能力分布的结点与权重*。

第一步，根据标准多元正态分布求出每个维度的结点和整体的权重。这个过程可以通过 R 语句编程实现。例如，多维 IRT 能力的结点与权重可以表示为：

θ_1	θ_2	*density*
−4.00	−4.00	[*density*]
−4.00	−3.58	[*density*]
……	……	……
−4.00	4.00	[*density*]
−3.58	−4.00	[*density*]
−3.58	−3.58	[*density*]
……	……	……
4.00	4.00	[*density*]

(8)

第二步，将每个维度的结点值乘以线性转换系数 α 并求和，得到近似单维结点值。

第三步，将上一步得到的结果按照结点从小到大进行排序，得到下面的矩阵：

θ_α	*Density*
$\theta_{11}\alpha_1 + \theta_{12}\alpha_2$	*density*
$\theta_{21}\alpha_1 + \theta_{22}\alpha_2$	*density*
……	……
$\theta_{n\delta 1}\alpha_1 + \theta_{n\delta 2}\alpha_2$	*density*

(9)

然后，按顺序合成结点与权重，结点数与单维 IRT 观察分数中保持一致。其中，每个区间结点之

---

* 结点与权重：将连续的能力分布看做基于有限数量的能力值的离散分布，其中能力值称为结点，与之相对应的密度称为权重。结点与权重可以表示能力的后验分布。这是进行 IRT 观察分数等值需要用到的条件。

和作为区间的结点，每个区间的权重之平均数作为区间的权重。这一步骤是为了使得到的结点和权重更加稳定。

最后，使用上面得到的参数、结点和权重进行多维近似单维 IRT 观察分数等值。

2.3 研究步骤

2.3.1 维度分析

使用 DETECT 软件(Stout，Habing，& Douglas，1996)对测验的维度进行非参数方法的分析，检验测验是否存在多维结构。

2.3.2 参数估计

采用 BMIRT 程序(Yao，Lewis，& Zhang，2008)，分别完成单维两参数 Logistic 模型和多维两参数 Logistic 模型对数据的拟合。

2.3.3 分数等值

研究采用的分数等值方法主要有三类，分别是：单维 IRT 真分数等值和 IRT 观察分数等值、多维近似单维 IRT 真分数等值 IRT 观察分数等值，以及等百分位等值。

(1) 单维 IRT 真分数等值和 IRT 观察分数等值

根据标准正态分布，使用 R 语句求出两组被试能力的结点与权重。最后，使用 PIE 程序(Hanson & Zeng，1995)，完成 IRT 真分数等值和观察分数等值。

(2) 多维近似单维 IRT 真分数等值和 IRT 观察分数等值

先求出多维近似单维各题目参数，以及能力分布的结点和权重。然后使用 PIE 程序(Hanson & Zeng，1995)，完成多维近似单维 IRT 真分数等值和观察分数等值。

(3) 等百分位等值

使用 RAGE－RGEQUATE(Zeng，Kolen，Hanson，Cui，& Chien，2004)完成等百分位等值和平滑。选择 S＝0.01 后平滑的结果作为最终的等百分位等值结果。

2.3.4 评价标准

(1) DETECT 结果

根据 DETECT 探索性分析结果，能够大致估计多维 IRT 等值是否能有较好的表现。如果 DETECT 的分类与测验本身的结构较一致，说明每个维度内的题目几乎指向同一个方向，这就为多维 IRT 等值提供了很好的基础。

(2) 等值标准误

等值标准误表示了等百分位等值中的随机误差。Equating Error 程序(Kolen & Brennan，2004)使用 Bootstrap 方法计算等值标准误。用等百分位等值的分数加减等值标准误，能得到等百分位等值 68% 的置信区间。如果某种等值方法的结果大部分落在了等百分位等值标准误置信区间之外，说明这种方法的结果与等百分位等值显著不同。

(3) 重要的差异(Differences That Matter)

Dorans 等(2003)提出了一种重要的差异(Differences That Matter)评价标准。他们认为，在特定分数点上，等值结果之间的差异大于 0.5 倍原始分数，则为重要的差异。在研究中，用这个标准衡量某等值方法与等百分位等值的差异。

3 结果

3.1 两测验描述性统计汇总

表 1 是测验 A，B 的描述统计。

表 1 两测验的描述统计量汇总

	人数	满分	最小值	最大值	均值	中数	标准差	峰度	偏度	信度
A 卷	3242	40	4	40	25.393	26	8.426	－0.249	－0.919	0.902
B 卷	3308	40	5	40	25.907	26	9.178	－0.168	－1.135	0.922

从表中可以看出，两套试卷上的分数分布略呈负偏态，说明这两套测验较为容易。参加测验的人数都达到了 3000 以上，保证了单维和多维 IRT 参数估计都能得到较准确的结果。

3.2 维度分析结果

DETECT 维度分析在探索性分析和验证性分析两种模式下，分别提供三种指标。DETECT 值说明测验在多大程度上符合多维结构。小于 0.2 表示单维结构，0.2 到 0.4 表示弱至中等程度的多维，0.4 到 1.0 表示中至强程度的多维。IDN 指数表示测验在多大程度上符合简单结构。接近 1 表示数据较好地拟合了简单结构模型。r 比值显示了分析结果稳定性的程度。接近 1 表示得到的结果较为稳定(Zhang & Stout，1999)。表 2 是对 A，B 两套测验进行维度分析的结果。

表 2 两测验 DETECT 结果

	A 卷		B 卷	
	探索性	验证性	探索性	验证性
DETECT 指数	0.320	0.312	0.287	0.287
IDN 指数	0.703	0.704	0.655	0.655
r 比值	0.660	0.644	0.613	0.613

通过 DETECT 指数可以看出，两套测验存在着弱至中等强度的多维结构。两套测验的 IDN 指数说明数据基本符合简单结构。r 比值证明得到的结果较为稳定。

在 DETECT 的探索性分析模式下，将得到的题目维度分类信息与测验的先验维度分类设定进行比较，总的来说两种方法的分类是很一致的。可以推测，多维 IRT 等值能够得到较好的结果。

3.3　参数估计结果

表 3 是两套测验项目参数估计结果的描述统计。

表 3　两测验题目参数估计结果的描述统计

	A 卷				B 卷			
	区分度		难度		区分度		难度	
	单维	多维	单维	多维	单维	多维	单维	多维
均值	0.839	0.556	−0.859	−0.831	0.984	0.610	−1.029	−1.016
标准差	0.433	0.348	1.102	1.103	0.410	0.312	0.914	1.006
相关	0.877		0.920		0.728		0.934	

可以看出，对于区分度参数，多维方法得到的均值和标准差小于单维方法。而对于难度参数，两种方法得到的均值和标准差是相近的。同时，两种方法在各参数估计结果上的相关较高，在难度参数上两种方法的估计结果更加接近。

3.4　等值标准误

研究中，等百分位等值标准误均值为 0.285，说明等百分位等值包含的随机误差较小。图 1 和图 2 分别呈现了各等值方法与等百分位等值置信区间的关系。

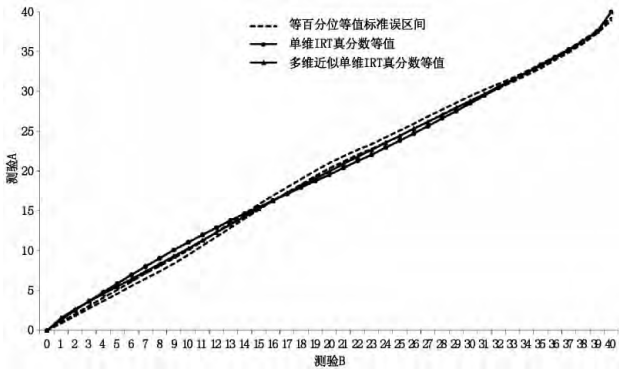

图 1　两种真分数等值方法和等百分位等值比较

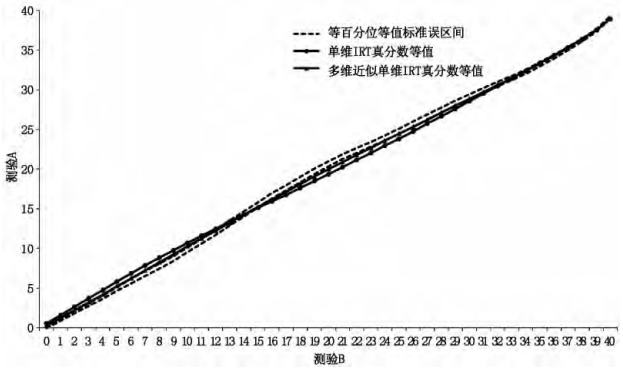

图 2　两种观察分数等值方法和等百分位等值比较

从图中可以看出，各等值方法与等百分位等值的趋势较为一致（相关达到 0.998 以上）。其中，多维 IRT 等值方法与等百分位等值更加相似，而单维 IRT 等值方法在很多分数点上的结果远超过了等百分位等值的标准误区间。

3.5　重要的差异

图 3 表示相同分数点上单维 IRT 真分数等值、单维 IRT 观测分数等值、多维近似单维 IRT 真分数等值、多维近似 IRT 观测分数等值与等百分位等值结果的差异。

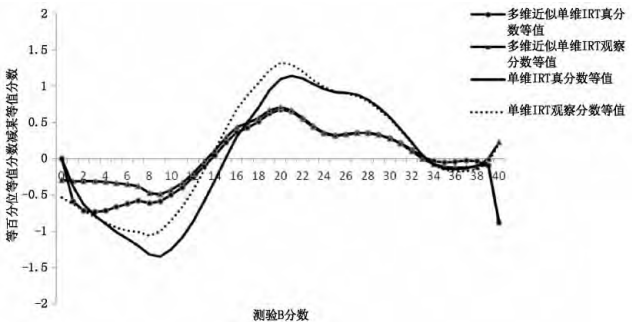

图 3　四种等值方法与等百分位等值结果的差异

根据定义，超过纵坐标上 [−0.5，0.5] 这个区间的结果与等百分位等值存在重要的差异。从图中可以看出，多维 IRT 等值方法所包含的重要差异的分数点较单维 IRT 等值方法少。并且，两种多维 IRT 等值结果非常接近，仅在低分段和高分段出现了较大的差异。另外，多维近似单维 IRT 观察分数等值与等百分位等值结果差异绝对值的均值和标准差最小，说明针对这批实际数据，这种方法与等百分位等值的结果最为接近。

4　讨论

4.1　关于单维 IRT 和多维 IRT 等值方法的比较

对单维和多维 IRT 等值方法进行比较，首先，这两类方法得到的等值趋势是一致的。这是因为两类方法对题目参数估计结果具有较高的相似性，而得到题目参数之后，单维和多维 IRT 等值的过程也是类似的。

其次，对于真分数等值和观察分数等值方法，单维和多维 IRT 等值方法之间的差异较大。这主要是由于两类方法的前提假设和模型定义不同，尽管在题目参数估计中，单维的方法和近似多维方法结果的相关很高，但参数估计的大小存在差异，这就可能导致两类等值方法的差异。将两类等值方法与等百分位等值的结果做比较，发现在随机等组设计下，多维 IRT 等值的结果与等百分位等值的结果更加接

近。这是由于根据维度分析的结果，该英语测验存在着弱至中等强度的多维结构，违背了传统 IRT 的单维性假设。而多维 IRT 等值是建立在多维性的假设下，所以这类等值方法所包含的系统误差较小，其结果也与等百分位等值更为相似。另外，在所比较的四种等值方法中，多维 IRT 观察分数等值与等百分位等值的结果最为接近。一方面归因于这种方法是建立在多维 IRT 的结构下; 另一方面是由于观察分数等值的方法与等百分位等值都利用了被试分布的信息，所以，以等百分位等值作为比较标准，可以认为在四种等值方法中，多维 IRT 观察分数等值的方法所包含的误差最小，得到的结果最准确。

最后，在整个分数的量尺上，单维 IRT 等值与多维 IRT 等值方法的差异并不一致，在一些分数点上单维 IRT 和多维 IRT 等值方法的差异较小，而在一些分数点上两种方法的差异较大。产生这种现象的原因可能是，在不同的分数点上，测验多维性结构对分数的影响是不同的，即，当测验测量的结构为多维时，在不同分数点上，考生在两个测验上分数的差异，所代表的意义可能不同。例如，在这两套英语试卷上，可能对于低分段的考生，他们分数的差异更大程度上来自于听力，对于高分段的考生，他们分数的差异更大程度上来自于阅读，而中等分数考生的差异同时来自于这两个方面。也就是说，在整个分数段上，可能一些分数体现了更多的多维性结构，而一些分数则显得更接近单维性结构。因此，在多维性结构较强的分数点上，单维 IRT 和多维 IRT 等值方法的差异就较大，而在单维性结构较强的分数点上，这两种方法的结果就更接近。

4.2 关于 IRT 真分数等值和 IRT 观察分数等值方法的比较

IRT 真分数等值和 IRT 观察分数等值的原理不同。真分数等值是将两个测验上的真分数进行链接，而观察分数等值旨在使用统计的方法对观察分数的分布进行调整，从而使得两个测验上观察分数的分布尽量相似。对四种等值方法比较可以看出，真分数等值和观察分数等值的差异较小，而单维和多维等值方法的结果差异相对较大。这与 Brossman (2010) 的研究结果是相似的。Kolen 和 Brennan (1995) 曾经指出，单维 IRT 真分数等值和 IRT 观察分数等值的结果非常接近，它们最大的区别可能会出现在满分附近，或者是 C 参数估计之和的分数之下。在刘玥等人(2010) 的研究中，也对单维 IRT 真分数等值和 IRT 观察分数等值进行了比较，发现两种方法得到的等值结果基本相等，差别较大的部分出现在被等值测验的低分数段。

在研究中，真分数等值和观察分数等值表现出很高的一致性，尤其在中高分数段，两种方法得到的等值结果几乎相同。而在低分段和满分附近，两种等值方法表现出了一定的差异。这说明单维 IRT 真分数等值和 IRT 观察分数等值的规律，也能延伸到多维 IRT 的体系中。另外，相对于真分数等值的方法，观察分数等值的方法与等百分位等值的结果更加接近，这是因为等百分位等值的过程从原理上说就是观察分数等值。

4.3 实际数据中进行真分数等值和观察分数等值的建议

在实际数据中，要进行 IRT 真分数等值和 IRT 观察分数等值，首先最好使用多维分析的软件，对测验的多维性及其具体结构进行检验。如果测验符合单维性结构，则可以选用传统的单维 IRT 真分数和观察分数等值的方法; 如果测验符合多维性结构，但是测验对维度的先验分类和软件探索性分析的结果不一致，则可以考虑通过一些探索性的方法重新划分维度，再进行维度检验; 如果测验符合多维性结构，测验对维度的先验分类和软件探索性分析的结果也一致，则选用多维 IRT 真分数等值和 IRT 观察分数等值的方法能得到较好的结果。

4.4 有待进一步研究的问题

由于研究采用了实际数据，所得到的等值结果只适用于该等值情境。因此研究得到的结论具有一定的局限性。并且研究中各等值方法的比较标准为等百分位等值的结果，但是这种等值方法本身也包含了等值误差，也不能准确地反映两套测验之间真实关系，因此使用它的结果作为比较标准是带有偏差的。

另外，目前针对 IRT 真分数等值和 IRT 观察分数等值的研究大部分是针对实际数据的，模拟研究还较少，没有得到广泛认可的等值评价标准，因此在今后的研究中可以探索如何对这两种等值方法的比较进行模拟研究。从而使得研究结论更具有推广性，为方法的比较和选择提供参考。

最后，研究使用的多维 IRT 分数等值方法，只能实现总分的等值，不能进行维度分数的转换。今后可以出于实际应用的考虑，对多维数据维度分数等值进一步探索。

5 结论

5.1 在研究设置的等值情境下，四种等值方法和等百分位等值具有相似的趋势。

5.2 当测验存在弱至中等程度的多维结构时，基于多维测验的 IRT 真分数等值和观察分数等值方法优于单维 IRT 真分数等值和 IRT 观察分数等值方法; 多维 IRT 观察分数等值略优于多维 IRT 真分数等值，但是两者之间的差异较小。

5.3 在实际情况下，最好先对测验的维度结构进行分析，再根据分析结果选择合适的等值方法。如果

测验确实存在多维结构，最好选用多维 IRT 的等值方法以减小系统误差。

参考文献

刘玥，骆方，刘红云.（2010）. IRT 真分数等值和 IRT 观察分数等值的对比研究.*心理科学*，*33*（3），676－680.

Brossman，B. G.（2010）. *Observed score and true score equating procedures for multidimensional item response theory*. University of Iowa.

Dorans，N. J.，Holland，P. W.，Thayer，D. T.，& Tateneni，K.（2003）. Invariance of score linking across gender groups for three Advanced Placement Program Examinations. In N. J. Dorans（Ed.），*Population invariance of score linking: Theory and applications to Advanced Placement Program examinations*（pp. 79－118）. Princeton，NJ: Educational Testing Service.

Han，T.，Kolen，M.，& Pohlmann，J.（1997）. A comparison among IRT true－and observed－score equatings and traditional equipercentile equating. *Applied Measurement in Education*，*10*（2），105－121.

Hanson，B.，& Zeng，L.（1995）. *PIE: A computer program for IRT equating*（Version 1.0）. Iowa City，IA: ACT.

Harris，D. J.，& Crouse，J. D.（1993）. A study of criteria used in equating. *Applied Measurement in Education*，*6*（3），195－240.

Kolen，M. J.，& Brennan，R. L.（2004）. *Test equating，scaling，and linking: Methods and practices*. 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.

Lord，F. M.（1980）. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates New Jersey.

Lord，F. M.，& Wingersky，M. S.（1984）. Comparison of IRT True－Score and Equipercentile Observed－Score "Equatings". *Applied Psychological Measurement*，*8*（4），453.

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.，Davey，T.，& Lee，K.（2000）. Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*，357－373.

Reckase，M. D.（2009）. *Multidimensional item response theory*. Springer Verlag.

Reckase，M.，& Martineau，J. A.（2004）. *The vertica lscaling of science achievement tests*. Unpublished Report. Michigan State University.

Simon，M. K.（2008）. *Comparison of concurrent and separate multidimensional IRT linking of item parameters*. University of Minnesota.

Stout，W.，Habing，B.，& Douglas，J.（1996）. Conditional covariance－based nonparametric multidimensionality assessment. *Applied Psychological Measurement*，*20*（4），331.

Yao，L.，Lewis，D.，& Zhang，L.（2008）. *An introduction to the application of BMIRT: Bayesian multivariate item response theory software*. Training Secession Presented at the Annual Meeting of the National Council on Measurement in Education－tion，Ny.

Zeng，L.，Kolen，M.，Hanson，B.，Cui，Z.，& Chien，Y.（2004）. *RAGE－RGEQUATE* [*Computer software*]. Iowa City: University of Iowa.

Zhang，J.，& Stout，W.（1999）. The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*，*64*（2），213－249.

IRT True Score Equating and Observed Score Equating for Multidimensional Data

Liu Yue[1]，Liu Hongyun[2]

（1. Sichuan Institute of Education Sciences，Chengdu 610225；2. School of Psychology，Beijing Normal University，Beijing 100875）

Abstract: In practice，tests usually measure more than one trait. In these situations，if the multidimensional structure is neglected and traditional unidimensional item response theory（UIRT）equating methods are still used，the equated parameters may be inaccurate. The research data was selected from a large－scale English test，using random－group design. Item parameters were estimated using MCMC method. UIRT true score equating and observed score equating were extended to multidimensional structure. Four equating methods were conducted and compared: UIRT true score and observed score equating，multidimensional item response theory（MIRT）true score and observed score equating. The results demonstrated that:（1）due to the multidimensional structure of our dat，MIRT equating methods performed better than UIRT methods;（2）MIRT observed score equating was slightly better than MIRT true score equating，but the results of these methods didn't show significant difference.

Key words: test equating; MIRT; true score equating; observed score equating; MCMC method