无铆题情况下测验分数等值方法探索——构造铆测验法*

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 研究旨在探索无铆题情况下，使用构造铆测验法，实现测验分数等值。研究一和研究二分别探索题目难度排序错误、铆题难度差异对构造铆测验法的影响。结果表明：（1）等组条件下，随着错误铆题比例，难度排序错误程度，铆题难度差异增大，构造铆测验法的等值误差逐渐增大，随机等组法的等值误差较为稳定；不等组条件下，构造铆测验法的等值误差均小于随机等组法；（2）对于构造铆测验法，在不等组条件下，铆测验长度越短，等值误差越大。

关键词 无铆题 铆测验 等百分位等值

1 引言

共同题设计和共同被试设计是测验等值中两种最常见的数据收集设计方式。在共同题设计中，通过被试在铆测验上的作答反应，能够得到不同测验分数之间的等值关系（Kolen & Brennan, 2004）。因此，铆测验应代表整个测验的特征 (Cook & Petersen, 1987)，同时具有内容代表性（Sinharay & Holland, 2007)。

共同题设计在大规模测验中应用最为广泛。然而，在实际的等值设计中，这种设计有时难以实现。首先，一些高利害测验往往不允许设置内铆题。例如，我国的高考竞争激烈，很难保证在正式试卷中嵌入铆题不被曝光（戴海崎 , 2003）。其次，在一些大规模测试中，由于测验本身较为复杂，也增加了铆测验的命题成本。要保证铆测验具有内容代表性，铆测验开发需耗费大量资源。最后，铆测验的安全性一直是备受关注的问题。随着互联网的广泛应用，铆题曝光的情况频频发生，这严重威胁到共同题设计的有效性。因此，有必要探索在无铆题情况下，如何寻找可替代的方法，实现测验分数等值。

Liao 和 Livingston（2012）曾利用实际数据对此问题进行探索，他们提出了三种替代性方法。方法一是构造随机等组测验法。但是，这种方法受预估难度准确性的影响很大，如果估计不准确，会造成层内题目难度差异大，无法构造出随机等组测验。方法二是构造随机等组样本法。基于与测验分数密切相关的人口学变量，采用倾向分数的方法，得到两个随机等组的样本。再利用随机等组的等值方法实现等值。然而，经过检验，这种方法的等值结果与直接随机等组法更接近，说明这种方法无法构造真正的随机等组。方法三是单一组等值法。假设被试的能力在一个月内没有变化，然后让被试在一个月时间内先后接受两个测验，根据单一组设计下的等值方法实现分数等值。但是这个方法的结果并不理想。

另外，Mislevy 等人（1993）曾提出过利用题目先验信息进行等值的方法。他们的思路是使用题目特征变量建立线性 Logistic 模型，对题目参数进行估计，从而得到题目参数的先验分布。然后在贝叶斯估计的框架下，通过期望反应曲线得到不同能力被试的期望分数，从而采用 IRT 真分数等值法实现测验分数等值。这种方法能够得到较为准确的结果，但它对题目先验信息的准确性依赖较大，并且采用贝叶斯估计的方法，计算过程较复杂，因此在实践中仍无法得到广泛应用。

综上，本研究借鉴构造随机等组测验法的思路，

* 本研究得到全国教育科学“十二五”规划教育部重点课题（GFA111001）的资助。
** 通讯作者：刘红云。 E-mail: hyliu@bnu.edu.cn
DOI:10.16719/j.cnki.1671-6981.20150632

提出一种构造随机等组铆测验的方法。并采用模拟的方法，对该方法进行了检验。旨在为无铆题情况下实现测验分数等值提供参考。

2　研究方法

2.1　构造铆测验法

该方法是通过构造铆测验，使用共同题设计下的等值方法实现测验分数等值。具体步骤如下：

（1）难度排序。采用专家评定法，对不同测验的所有题目按照难度高低进行排序。

（2）选出不同测验中难度接近的题目。根据题目难度排序结果，找出来自不同测验且难度相近的题目，作为备选铆题。

（3）构造铆测验。在考虑题目内容、题目类型等因素的基础上，利用备选铆题，构造铆测验。

（4）进行等值。采用共同题设计下的等值方法进行测验分数等值。

2.2　研究设计

由于构造铆测验法在很大程度上依赖于专家对题目难度的评定，因此，该方法产生的误差可能来源于两个方面。一是专家对题目难度的排序出现错误，二是铆题间难度差异较大。因此，实验 1 探索题目难度排序错误对构造铆测验法的影响，实验 2 探索铆题难度差异对构造铆测验法的影响。

两个模拟研究中测验题目均为 40 道，0/1 计分。采用 Rasch 模型生成作答反应矩阵。

2.2.1　实验 1 设计

首先模拟两个测验的题目难度，并将所有题目按难度产生值从小到大进行排序，以模拟实际中专家评分的过程。然后在考虑铆题难度真值的基础上，挑选相邻的来自两个测验的题目，作为铆题。规定每次构造铆测验时，总是为原始测验的题目选择从小向大数第 1 个来自新测验的题目。由于在实际中，排序的结果不一定完全准确，因此设计了两种题目难度排序错误程度，即在难度从小到大的基础上，向下挪动 1 个位置（简称 +1）和 5 个位置（简称 +5）。自变量有 5 个。（1）被试量： 2000 人、5000 人；（2）铆测验长度：5 道（占测验长度的 1/8）、10 道（占测验长度的 1/4）；（3）错误铆题比例：在 5 道铆题的情况下，分 0%、20%、60%、100% 四个水平，在 10 道铆题的情况下，分 0%、30%、50%、70%、100% 五个水平；（4）题目难度排序错误程度：+1、+5；（5）测验难度差异：相等、不等。相等的情况下，两个测验题目难度分布均为 $b \sim N(0, 1^2)$，参加两个测验的总体能力分布均为 $\theta \sim N(0, 1.5^2)$。不等的情况下，新测验题目难度分布为 $b \sim N(1, 1^2)$，原始测验为 $b \sim N(0, 1^2)$。新测验的总体能力分布为 $\theta \sim N(1, 1.5^2)$，原始测验为 $\theta \sim N(0, 1.5^2)$。

2.2.2　实验 2 设计

首先设定两个测验的题目难度符合公差为 0.1 的等差数列。在按照难度真值排序后，设计了两种铆题难度差异程度，即铆题难度差为 .15（简称 +.15）和 .25（简称 +.25）。自变量有 5 个，其中（1）（2）（3）（5）与研究一相同，新增自变量为铆题难度差异程度（+.15、+.25）。另外，测验难度差异各水平的设定与研究一不同。相等的情况下，新测验题目难度从 -2.95 到 1.95，原始测验题目难度从 -2 到 1.9，依次递增 .1。不等的情况下，新测验题目难度从 -.95 到 2.95，原始测验题目难度从 -2 到 1.9，依次递增 .1。参加原始测验、新测验的总体能力分布与研究一相同。

3.1　研究步骤

3.1.1 数据生成

使用 R 语言产生每种条件下的反应数据，每种条件下数据重复模拟 30 次。

3.1.2　分数等值

比较的等值方法有两种：（1）构造铆测验法；（2）随机等组法，将两组被试视为随机组进行等百分位等值。

3.1.3　评价标准

模拟生成接受原始测验的被试在新测验上的作答反应，并结合他们在原始测验上的作答，进行随机等组的等百分位等值，作为比较标准。

首先，对于随机等组设计下的等百分位等值，计算等值分数的标准误，以考察该方法的稳定性。

然后，从四个方面评价各方法的等值分数与标准的差异。

偏差（*BIAS*）考察了各条件下，等值分数是否有定向的偏差。计算公式如下：

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

其中，表示各方法得到的等值分数，表示标准得到的等值分数，R 表示题目数量，N 表示重复的次数。

绝对偏差（*MAE*）和误差均方根（*RMSE*）考

察了各条件下，各等值方法的等值分数与标准的差异大小。计算公式如下：

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{R}\sum_{r=1}^{R}|\hat{\varsigma}_r - \varsigma_r| \quad (2)$$

$$RMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\varsigma}_r - \varsigma_r)^2} \quad (3)$$

公式中各参数表示的意义与公式 (1) 相同。

等值后分数与标准的相关计算了各条件下 , 各方法得到的等值分数与标准的相关。

3 结果

3.1 等值标准误

所有条件下，随机等组法与标准的等值标准误均较小（.007~.002），并随着样本量增加，等值标准误均减小。

3.2 实验 1 结果

3.2.1 偏差

在不等组的条件下，随机等组法的偏差明显大于构造铆测验法，并且随机等组法得到的等值分数均偏高。在等组的条件下，随机等组法的偏差小于构造铆测验法。

3.2.2 绝对偏差和误差均方根

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。

结果显示，等值方法的主效应显著（F（1，36）= 2317.53，p < .001，η^2 = .985），总体来说，构造铆测验法的误差均方根小于随机等组法。交互作用分析表明，等值方法与测验难度差异（F（1，36）= 4634.07，p < .001，η^2 = .992）的交互作用显著。在等组的条件下，两种方法的误差均方根差异相近，在不等组的条件下，随机等组法的误差均方根明显大于构造铆测验法。

对构造铆测验法得到的等值分数误差均方根进行方差分析。结果显示，铆测验长度（F（1，7）= 29.46，p <.001，η^2 = .808）、错误铆题比例（F（5，7）= 189.64，p <.001，η^2=. 993）、题目难度排序错误程度（F（1，7）= 1185.93，p < .001，η^2= .994）、测验难度差异（F（1，7）=2845.70，p <.001，η^2 = .998）的主效应显著，铆测验长度越长，构造铆测验法误差越小，错误铆题比例、题目难度排序错误程度、测验难度差异越大，构造铆测验法误差越大。而样本量的主效应不显著（F（1，7）= .07，p > .1）。

交互作用分析表明，铆测验长度与题目难度排序错误程度（F（1，7）= 8.69，p < .05，η^2= .554）、错误铆题比例与题目难度排序错误程度（F（5，7）= 93.20，p < .001，η^2= .985）、题目难度排序错误程度与测验难度差异（F（1，7）= 30.27，p < .001，η^2= .812）、铆测验长度与测验难度差异（F（1，7）= 256.30，p < .001，η^2=.973）、测验难度差异与错误铆题比例（F（5，7）= 7.40，p < .05，η^2= .841）的交互作用显著。如图 1 所示：（1）当题目难度排序错误不存在时，5 道铆题的等值误差大于 10 道铆题，随着题目难度排序错误程度增加，10 道铆题的等值误差逐渐增大，当题目难度排序错误程度为 +5 时，两种铆测验长度的等值误差近似；（2）当题目难度排序错误程度为 +1 时，构造铆测验法的等值误差几乎不受错误铆题比例增加的影响，当排序错误程度为 +5 时，构造铆测验法的等值误差随着错误铆题比例的增加有明显的上升；（3）在不等组的条件下，构造铆测验法的等值误差随着题目难度排序错误程度平缓上升，在等组条件下，构造铆测验法的等值误差随题目难度排序错误程度的增加而明显上升；（4）在等组条件下，构造铆测验法的等值误差随着铆测验长度增加而增加，在不等组条件下，构造铆测验法的等值误差随着铆测验长度增加而减少；（5）在测验难度差异不同的条件下，构造铆测验法的等值误差随着错误铆题比例的增加呈现出不同的增长趋势，在不等组条件下，错误铆题比例为 20% 和 60% 的情况下误差较大，这主要因为在实验设计中，只有 5 道铆题的条件下才存在错误铆题比例为 20% 和 60% 的条件，因此结合铆测验长度与测验难度差异的交互作用可以推测，较大的误差可能是由于铆题长度太短造成的。

图 2 展示了各条件下不同方法等值分数的误差均方根。

从图中可以看出：（1）对于构造铆测验法，当错误铆题比例增加至 50%-60% 时，等值误差有明显的增加；（2）对于构造铆测验法，当铆题错误程度为 +5 时，即使错误铆题比例较小（20%-30%），其等值误差也大于铆题错误程度为 +1 的所有情况；(3) 总的来说，在等组条件下，如果错误铆题比例为 0%，两种等值方法表现相同，随着错误铆题比例，题目难度排序错误程度增大，构造铆测验法的等值误差逐渐大于随机等组法，在不等组条件下，构造铆测验法的等值误差始终小于随机等组法。

3.2.3　等值分数与标准的相关

两种方法在各条件下与标准的相关均较高，在 .985~1.000 之间，等值后的分数与标准具有较高的一致性。

3.3　实验 2 结果

3.3.1　偏差

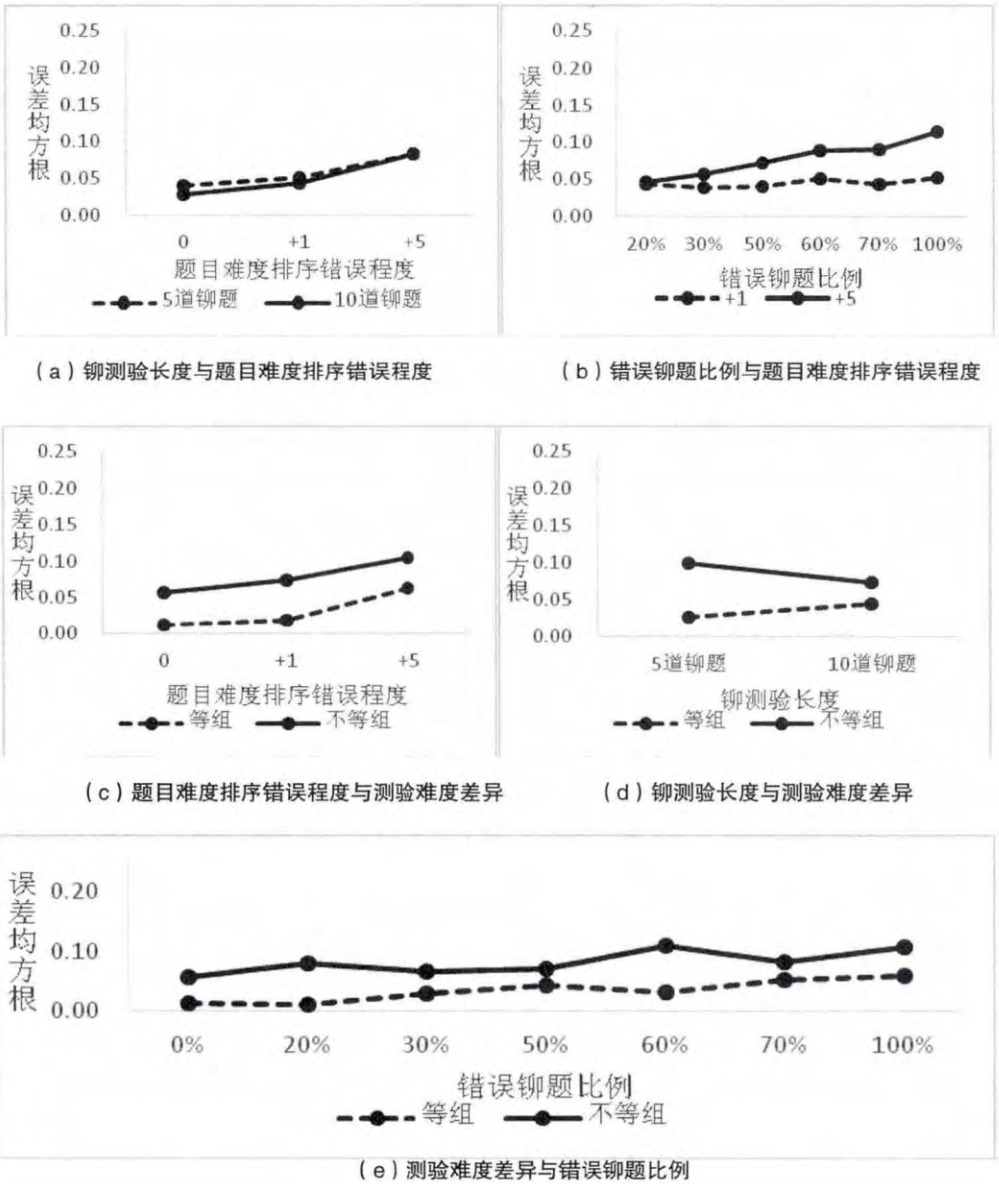
实验 2 偏差的结果与实验 1 有类似的趋势，但构造铆测验法得到的偏差比实验 1 略小。

3.3.2　误差均方根

结果显示，等值方法的主效应显著（F（1，36）= 6529.60，p < .001，η^2= .995），总体来说，构造铆测验法的误差均方根小于随机等组法。交互作用分析表明，等值方法与测验难度差异（F（1，36）= 8817.72，p < .001，η^2= .996）的交互作用显著。在等组的条件下，两种方法的误差均方根差异相近，在不等组的条件下，随机等组法的误差均方根明显大于构造铆测验法。

对构造铆测验法得到的误差均方根进行方差分析，结果显示，铆测验长度（F（1，7）= 27.12，p < .001，η^2= .795）、错误铆题比例（F（5，7）=20.17，p <.001，η^2= .935）、铆题难度差异程度（F（1，7）= 41.85，p <.001，η^2= .857）、测验难度差异（F（1，7）= 1849.51，p < .001，η^2= .996）的主效应显著，铆测验长度越长，构造铆测验法误差越小，错误铆题比例、铆题难度差异程度、测验难度差异越大，构造铆测验法误差越大。而样本量的主效应不显著（F（1，7）= .08，p > .1）。

交互作用分析表明，铆测验长度与测验难度差

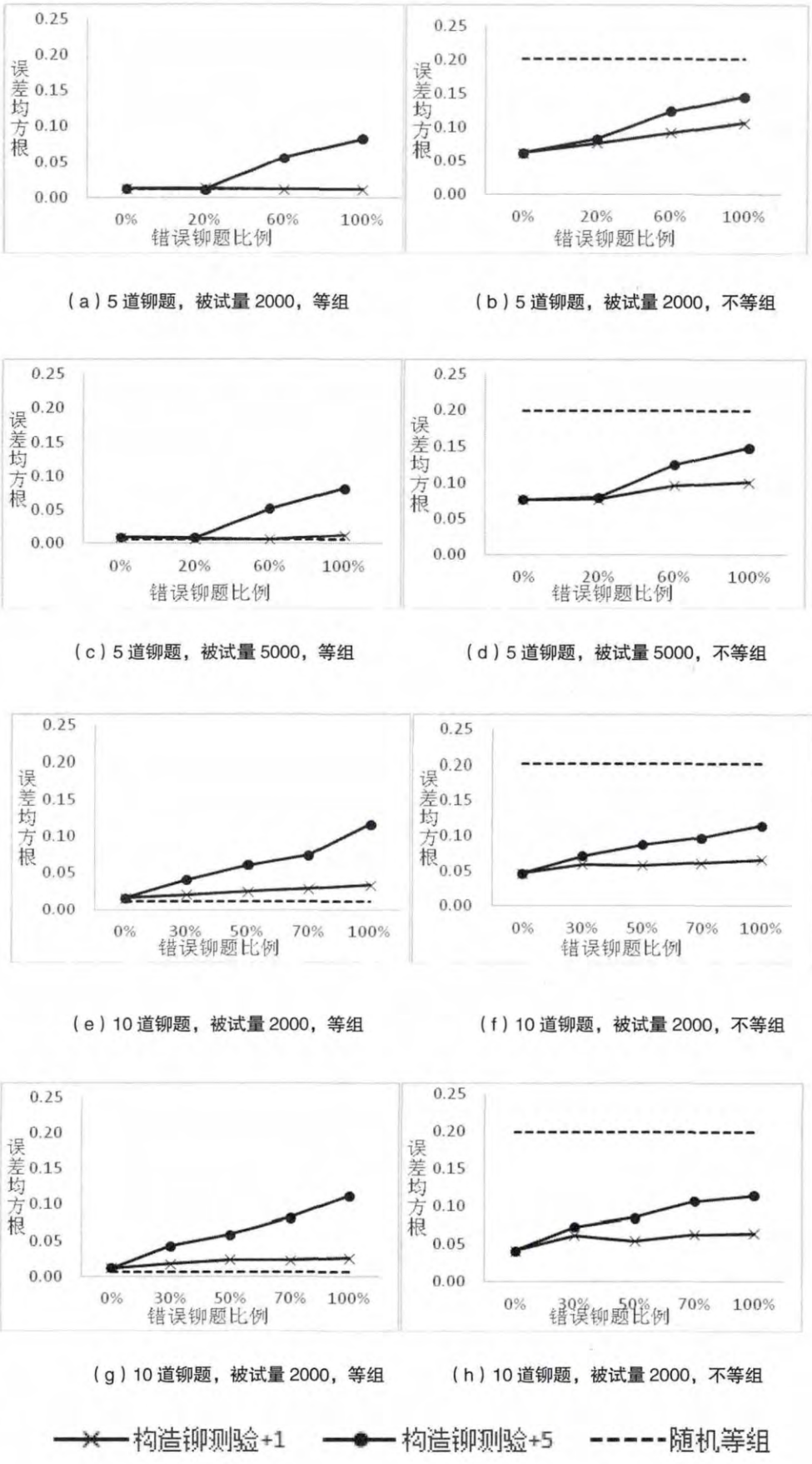


（a）铆测验长度与题目难度排序错误程度　（b）错误铆题比例与题目难度排序错误程度

（c）题目难度排序错误程度与测验难度差异　（d）铆测验长度与测验难度差异

（e）测验难度差异与错误铆题比例

图 1　实验 1 各因素交互作用



(a) 5 道铆题，被试量 2000，等组　(b) 5 道铆题，被试量 2000，不等组

(c) 5 道铆题，被试量 5000，等组　(d) 5 道铆题，被试量 5000，不等组

(e) 10 道铆题，被试量 2000，等组　(f) 10 道铆题，被试量 2000，不等组

(g) 10 道铆题，被试量 2000，等组　(h) 10 道铆题，被试量 2000，不等组

图 2　实验 1 各条件下不同方法等值分数误差均方根

异（F（1，7）=74.44，p <.001，η^2=.914）的交互作用显著。对于构造铆测验法，在等组的条件下，5 道铆题和 10 道铆题的等值误差近似，在不等组的条件下，10 道铆题的等值误差小于 5 道铆题的情况。

图 3 展示了各条件下不同方法等值分数的误差均方根。

从图中可以看出：（1）对于构造铆测验法，当错误铆题比例增加至 60%-70% 时，等值误差有明显的增加；（2）总的来说，在等组条件下，如果错误铆题比例为 0%，两种等值方法表现相同，随着错误铆题比例，铆题难度差异程度增大，构造铆测验法的等值误差逐渐大于随机等组法，在不等组条件下，构造铆测验法的等值误差始终小于随机等组法。

3.3.3　等值分数与标准的相关

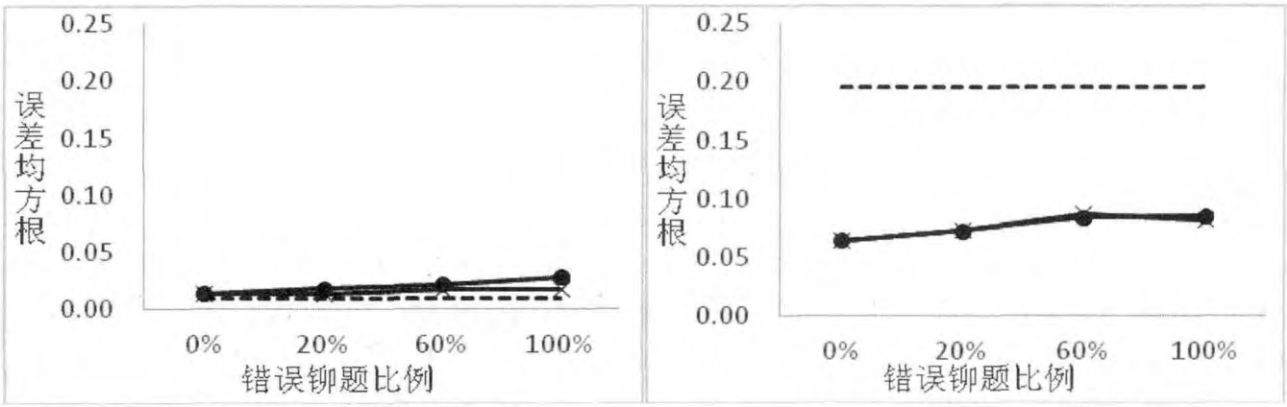
两种方法在各条件下与标准的相关均较高，在 .986-1.000 之间，等值后的分数与标准具有较高的一致性。

4　讨论

4.1　构造铆测验的共同题目等值法与随机等组法的比较

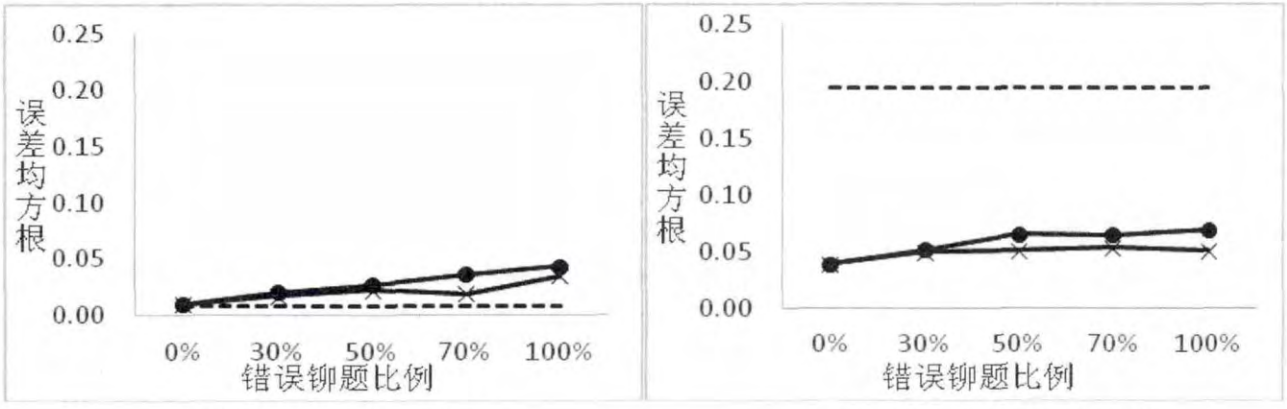
随机等组设计中，接受不同测验的群体应来自同一个总体，并且应同时接受不同的测验。在实际的等值中，这种方法有一定的局限性。原因，一是不同群体的能力分布往往存在差异，很难证明接受不同测验的群体是真正的随机等组；二是很难保证不同测验施测的条件完全相同，例如，跨年度等值。当两组群体的能力分布有差异时，这种方法就无法使用。

共同题设计对被试群体的分布、被试接受测验的条件等没有严格要求，但是在测验设计上需要构造铆测验，并且对于铆测验的统计假设也较为严格。构造铆测验是难点。铆测验应当是整体测验的缩小版本，尤其当接受不同测验的群体能力分布差异较大时，铆测验在内容和统计上的代表性更加重要。另外，铆题在不同测验中的位置应相同。研究所构造的铆测验与真正的铆测验有三个方面的差别。一是“铆测验”中的题目只能保证难度值近似，而真正的铆测验需要题目所有参数完全相同；二是“铆测验”对整个测验内容的代表性只能通过在难度排序接近的题目中选择铆题时尽量平衡题目所属的内容维度来实现，可能很难严格满足内容代表性的要



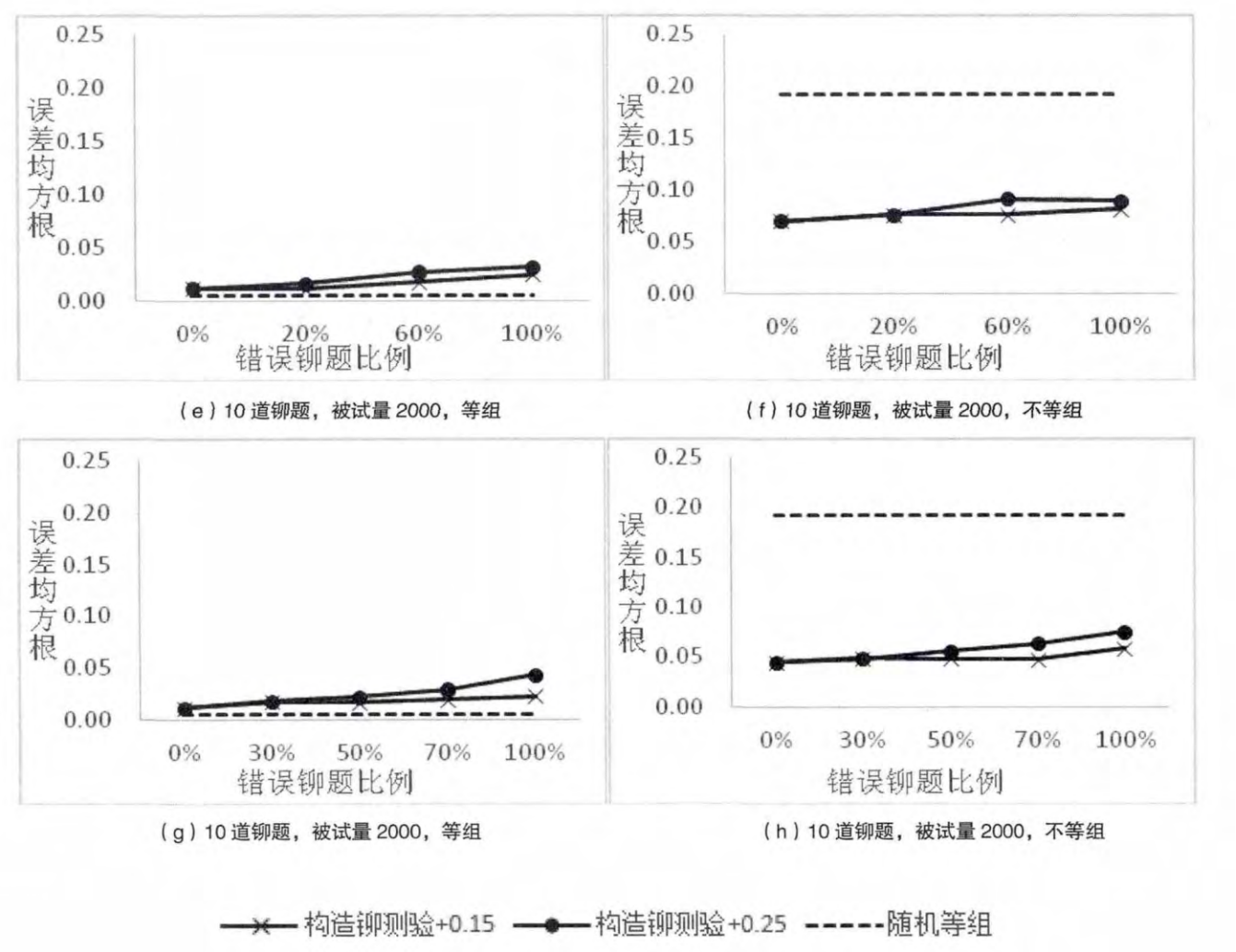
（a）5 道铆题，被试量 2000，等组

（b）5 道铆题，被试量 2000，不等组



（c）5 道铆题，被试量 5000，等组

（d）5 道铆题，被试量 5000，不等组



（e）10 道铆题，被试量 2000，等组　（f）10 道铆题，被试量 2000，不等组

（g）10 道铆题，被试量 2000，等组　（h）10 道铆题，被试量 2000，不等组

图 3 实验 2 各条件下不同方法等值分数误差均方根

求；三是很难保证“铆题”出现的位置在不同测验间是相同的。然而，即便如此，当接受不同测验的群体能力分布存在显著差异时，构造铆测验法在各条件下都优于随机等组法，且与标准相比等值误差是可以接受的。说明在模拟的条件下，使用随机等组法所带来的误差比构造铆测验法带来的误差更大。

在实际的等值情境中，等值设计的选择需要考虑测验的施测情况、测验的开发情况和所需的统计假设等方面的复杂程度（Kolen & Brennan, 2004）。这些都是先验的，即先进行数据收集设计，再开发测验，施测，等值。而研究所探讨的是在测验已经施测情况下的等值方法，是一种“后验”方法。实际的等值过程应当首选“先验”方法，根据预先的设计实现等值。如果在已有数据的条件下只能使用“后验”方法，也应当结合测验本身、施测情况和等值群体的分布等因素选择合适的等值方法。

4.2 实际数据中使用构造铆测验法的建议

（1）探索影响题目难度的因素

为帮助专家得到更加准确的题目难度排序结果，有必要探索影响题目难度的因素，指导专家在估计题目难度时综合考虑这些影响因素。一些学者很早就开始了这类研究。例如，Scheuneman（1991）等人发现，GRE 的心理测验和国家教师考试的阅读测验部分中，题目难度变异的 65% 可以被与题目可读性、语义内容、认知要求和知识要求等相关的变量解释。Heilman 等人（2008）研究证明，使用词汇和语法的特征建立回归模型，能有效预测阅读题难度。影响题目难度的因素有一定的共性，但是对于不同的测验，影响因素及其影响程度又可能有所差异。因此，未来可以基于已有研究结果，探索题目难度的影响因素。

（2）加强对专家的培训

研究证明，对专家进行有针对性的培训能够有效提高题目难度估计的准确性。如，在 Chalifour 和

Powers（1989）对 GRE 逻辑推理题的研究中，就发现训练有素的命题专家对题目难度的预测能够解释难度值实际变异的 72%。题目难度和区分度参数的人工赋值也是题库建设中的一种常用方法（董圣鸿等 , 2005）。在实际中，基于题目难度的影响因素，对专家进行培训，有助于减小题目难度估计的误差。

（3）要求专家估计题目难度值

在构造铆测验法中，铆题的选择是以所有题目难度排序结果为基础。但是，即使难度排序后相邻的题目，其难度值的差异也可能较大。实验 2 正是考虑了所构造的成对的铆题难度差异大小的情况，其结果证明，铆题难度差异越大，构造铆测验法的等值误差越大。因此，在实际操作中，最好要求专家估计题目难度的具体数值，以便筛选出最接近的题目作为铆题。

（4）控制专家评定一致性程度

为保证专家评分的一致性程度，有必要对专家估计的难度值进行评分者信度检验。对于评分者信度较低的题目，要求专家在反复思考、充分讨论的基础上进行重新估计。另外，还可以应用多面 Rasch 模型对难度估计结果进行分析，剔除不同专家宽严度对难度估计值的影响，使其更贴近真实值。

（5）对构造铆题质量进行评估

为保证专家所选定的铆题具有较强的相似性，可以利用两次测验的实际数据进行验证。例如，分别基于两次测试的实际作答反应估计测验题目的难度参数，假设在两套测验中专家构造的铆测验分别为 A 和 A'，将 A 和 A' 中的所有铆题分别按照难度排序。如果所有被认为近似的题目在 A 和 A' 中的难度位次均相等，那么可以推测专家评定法构造的铆测验较为可靠；如果难度位次不同，则可以认为这两道题目可能存在一定的差异，不能成为近似铆题。

（6）增加铆题的数量

根据本研究的结果，在不等组条件下，构造铆测验法的等值误差随着铆测验长度增加而减少。因此，如果能够判断需要等值的两个测验在难度上存在较大差异，那么在使用构造铆测验法时，最好多选择一些近似的“铆题”，以减小该方法可能带来的误差。

（7）在命题阶段设计近似的题目

构造铆测验法是当测验已经形成并施测之后所采取的方法，是一种“后验”的方法，受已有测验题目特征的限制，具有一定的局限性。在实际应用中，如果无法在测验中使用真正的铆题，最好能够在测验设计阶段，有意识地编制或筛选在题目特征上与原始测验中某些题目近似的题目，使之作为近似的“铆题”，实现测验等值。

4.3　有待进一步研究的问题

研究采用模拟研究的方法，对所有题目难度排序、错误铆题比例、错误铆题程度、铆题难度差异等设定都较为理想。在实际中，由于专家主观排序的不确定性，可能产生更多的排序错误情况，为构造铆测验法带来更多的随机误差。因此，这种方法是否适用，还有待使用实践中的数据进行进一步验证。

另外，由于专家评定法具有较强的主观性、随意性，可以借鉴董圣鸿等人（2005）提出的人工赋值与参数赋值相结合的思路，将专家评定法与使用影响因素估计题目参数的统计方法相结合，得到更为匹配的铆题。另外，在董圣鸿等人（2005）的研究中，还对区分度的人工赋值进行了初探，今后可以结合两参数模型，进行构造铆测验法的模拟研究，并探讨在实际中如何同时考虑难度、区分度参数的近似性，寻找到更加接近的铆题。

5　结论

（1）总的来说，构造铆测验法的等值误差小于随机等组法。在等组条件下，如果没有错误铆题，两种等值方法表现相同，随着错误铆题比例，题目难度排序错误程度，铆题难度差异程度的增大，构造铆测验法的等值误差逐渐大于随机等组法；在不等组条件下，构造铆测验法的等值误差始终小于随机等组法。

（2）对于构造铆测验法，错误铆题比例越大，题目难度排序错误程度越大，铆题难度差异程度越大，等值误差越大。在测验难度差异大的情况下，铆测验长度越短，等值误差越大。

（3）在实际情况下，如果测验难度差异较大（或参加测验的被试总体能力差异较大），并且难以实现真正的共同题设计，建议使用构造铆测验法取代简单的随机等组法，但应注意专家排序和匹配铆题的准确性。

参考文献

董圣鸿 , 漆书青 , 戴海琦 , 丁树良 . (2005). 题目难度、区分度参数人工赋值方法的研究 . *考试研究 , 1,* 24-32.

戴海崎 . (2003). 高考等值试验的几个重要问题研究 . *湖北招生考试 , 4,* 7-9.

Chalifour, C. L., & Powers, D. E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations.

Journal of Educational Measurement, 26, 120-132.

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,* 225-244.

Heilman, M., Collins-Thompson, K., & Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. *Paper presented at the Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications.* Columbus, USA.

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

Liao, C., & Livingston, S. A. (2012). A search for alternatives to common-item equating. *Paper presented at the Annual Meeting of the National Council on Measurement in Education* (*NCME*). Vancouver, British Columbia.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

Scheuneman, J. (1991). Effects of prose complexity on achievement test item difficulty. *ETS Research Report Series, 2,* 1-53.

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,* 249-275.

A Search for Alternatives to Test Equating with no Common Items

Liu Yue[1] *Liu Hongyun*[2]

([1] Sichuan Institute Of Education Sciences, Chengdu, 610225)

([2] School of Psychology, Beijing Normal University, Beijing, 100875)

Abstract The non-equivalent groups with anchor test (NEAT) design is widely used in large scale educational assessments. However, in practical, security problems in some regions make it difficult to re-use items; it is also impossible to use inner anchor sets in high-stakes tests. Besides, the criteria for selecting anchor set are hard to achieve in some situations. The purpose of this study is to find a practical alternative for score equating when NEAT design is not conducted in large-scale assessments, and to evaluate the new method in various conditions.

This new approach is called Assembling Anchor Sets (AAS) approach. First, items from different tests are ordered by the predicted difficulty by experienced experts. Then, items of nearly equal difficulty are chosen as common items, considering the representativeness of anchor tests at the same time. Last, equating methods under NEAT design should be applied.

The study is a mixed measure design of simulation conditions and score equating methods. Two equating procedures are compared. One is to build anchor sets using the AAS approach and imply the equipercentile equating methods under the NEAT design; the other is to treat the two simulated samples as random groups, and use the equipercentile equating methods under random groups design. There are two simulations in the research. Simulation 1 aims to explore the effect of the error of experts' judgments on the AAS approach. There are five simulation conditions: (1) number of examines (2000 and 5000); (2) anchor length (5 items, 1/8 of total test;10 items,1/4 of total test); (3) proportion of mistakenly predicted items (four levels in the condition of 5 common items, 0%, 20%, 60%, 100%; five levels in the condition of 10 common items, 0%, 30%, 50%, 70%, 100%); (4) errors of predicted item difficulty (+1 and +5); (5) difficulty levels of total tests (equivalent and non-equivalent).

Simulation 2 aims to find out the influence of the differences of nearly equal difficulty items on the AAS approach. There are five simulation conditions, four of which are the same as the factors (1) (2) (3) (5) in simulation 1. The other one is the differences between the difficulty of common items (+0.15 and +0.25).

In both simulations, test length is fixed to 40 items, and 30 replications are generated. Data are generated according to Rasch model using R program. The equipercentile equating methods are conducted by R package called "Equate". The responses of the examinees to the original test on the new test are also simulated. Therefore, it can be referred as the examinees taking both the original and the new tests. Then, the equipercentile equating method under random groups design can be applied to equate the scores on the new test with the original test. These are the true scores in the study. Finally, the two equating methods are evaluated by four criteria: bias, mean absolute error, root mean square error, and correlation between the scores after equating and true scores.

The results show that: (1) As the proportion of mistakenly predicted items, the errors of predicted item difficulty, and the differences between the difficulty of common items increase, and the equating error of AAS approach increases; (2) In the conditions of equivalent groups, when the proportion of mistakenly predicted items is large, the errors of predicted item difficulty are serious, and the differences between the difficulty of common items are obvious, the AAS approach is worse than the method under random groups design; in the conditions of non-equivalent groups, the AAS approach is always better than random groups methods; (3) In the conditions of non-equivalent groups, as anchor length increases, the equating error of AAS approach decreases. In conclusion, it is highly recommended to use the AAS approach in non-equivalent conditions. Further research should focus on developing some practical ways to increase the accuracy of predicted difficulty of items.

Key words no common items, anchor set, equipercentile equating