心理学探新 2025,Vol.45,No.1,72-77
PSYCHOLOGICAL EXPLORATION

无锚题非等组设计下的等值实现*

董圣鸿[1],秦春影[1,2],游晓锋[2],喻晓锋[1]
(1.江西师范大学心理学院,南昌 330022;2.南昌师范学院数学与信息科学学院,南昌 330032)

摘 要:很多测验在设计的时候没有为针对等值作专门的考虑,或者是由于测验自身的高风险特性,导致测验天然地不符合常规的等值设计,但是仍然需要对不同测验版本的测验分数或考生的能力进行比较,即存在等值的必要性。本文聚焦无锚设计(即不存在锚题,也不存在等组)下的等值问题,对已有典型研究中涉及到的方法和技术进行评价,包括构造拟等组的方法,基于共同认知结构的方法等,目的是厘清不同方法的使用条件和优缺点,并对未来的研究方向进行了展望。

关键词:等值;锚题;拟等组;认知结构
中图分类号:B841.2 **文献标志码**:A **文章编号**:1003-5184(2025)01-0072-06

1 引言

测验链接(linking;Kolen & Brennan,2014)是指将一个测验上的分数转换到另一个测验分数所在量尺上,它是一种统计过程,用于确保不同版本或不同形式的测验分数可以进行比较或转换。测验链接的方法可以分为三大类:预测(predicting)、量尺校准(scale calibration)和等值(equating)。等值(equating)是其中的一项重要技术,它通过一定的测验设计,基于选定的统计方法,实现不同测验版本中的分数或参数的转换,从而实现不同测验版本间分数或参数的可比,在测量实践中应用广泛,因而受到测量和统计研究人员的关注(戴步云,罗照盛,2012;刘玥,刘红云,2015a,2015b;Kolen & Brennan,2014;Lord,1980;van der Linden,2000;von davier et al.,2003)。根据 Lord(1980)的观点,能够进行等值的两个测验需要满足如下的条件,分别是(1)共同认知结构(same construct);(2)信度等价性(equal reliability);(3)对称性(symmetry);(4)群体不变性(population invariance);(5)公平性(equity)。

当测验的不同版本之间无锚题存在,并且参加这两个测验版本的考生能力分布情况并不清楚,也就是下文将要提到的无锚设计,对于这种情况下的测验等值研究相对较少,典型的有 Haberman(2015),Longford(2015),Xin 和 Zhang(2015)等研究进行了尝试,其它的还有一些涉及到的研究,如杨志明(2015)对一年多考背景下分数等值的意义和方法进行展望,但是没有对方法进行具体的介绍;刘玥和刘红云(2015b)采用构造锚测验的方法对无锚设计下的等值探索,这种方法在实际应用中不容易实现;李雪莹(2016)基于认知诊断测验中高阶模型,基于 IRT 中的项目特征曲线法实现属性难度和区分度的转换;朱殷睿(2023)考虑加入事后锚题的方式实现无锚设计下的等值,这种方式对于高风险测验的可行性较低。

在实际的应用中,尤其是在一些高风险的测验应用中,不同的测验版本间采用无锚设计的场景还是很多的,比如我们国家的高考,不同年份间的测试题目是完全不相同的,考生的能力分布也存在差异。如何实现这样的考试实践中分数的比较,需要深入研究和探索,因为这对于掌握考生能力水平的发展、测验的编制和优化等至关重要(杨钰萍,2019)。为了方便介绍,下文将测验的不同版本之间不存在共同题,且分别参加这些测验版本的考生能力分布是否相同并不确定的测验设计记为无锚测验非等组(non-equivalent groups without anchor test,NEWAT)设计,简称为无锚设计。相对于典型的非等组锚测验(non-equivalent groups with anchor test,NEAT)设计,NEWAT 设计下的不同测验版本之间既没有共同的考生,也没有共同的题目,从而无法采用常规的方法(如经典测量理论下的链式等值,项目反应理论下的同时校准法等)来实现测验版本之间的等值。

针对 NEWAT 设计,有研究者提出采用考生的人口学信息进行等值,但是它的公平性受到质疑(Holland & Dorans,2006)。Haberman(2015)基于最小判别信息调整(minimum discriminant information

* **基金项目**:教育部教育考试院'十四五'规划支撑专项课题"高考实施过程中的科目跨年分数的转换研究"(NEEA2021050)。
通信作者:董圣鸿,E-mail:jxnudsh@163.com;秦春影,E-mail:cqin@jxnu.edu.cn。

adjustment，MDIA；Haberman，1984）将非等组的考生样本通过加权将他们转换到拟等组（pseudo－equivalent group；Kim & Lu，2018；Lu & Guo，2018），这里称加权后的考生样本为拟等组的原因是他们为不同测验版本上的题目得分比较提供了可能（Kolen & Brennan，2014）。通常情况下，锚测验会设计成与考生测验分数有较高的关联，而背景变量能起到的影响要小得多。在没有其它可替代的信息时，选择能对考生测验分数影响程度越大的背景变量是需要考虑的问题。Longford（2015）提出了倾向得分匹配（propensity score matching）的方法，通过背景变量（如人口学特征、教育经历）构建倾向得分，并进行分层匹配来形成局部等值子群体。由于传统测验等值仅关注总分分布的局限，Xin 和 Zhang（2015）构建基于认知属性的局部等值（local equating）框架，结合不同测验版本间具有相同的属性掌握模式的考生能力相同的假设（Tatsuoka，2009），提出条件等百分位等值算法，在特定认知属性组合（如｛掌握技能A，未掌握技能B｝的子群体）内实施局部等值，结果表明局部等值相较于传统全局等值（如链式等值），能减少能力分布差异导致的等值偏差。

鉴于实际应用中 NEWAT 设计的普遍性和 NEWAT 设计下等值需求的迫切性，有必要对现有 NEWAT 设计下等值有关的研究进行梳理，并对现有三个典型研究（Haberman，2015；Longford，2015；Xin & Zhang，2015）中的方法和思路进行分析和讨论，确定各方法的适用条件和优缺点，为接下来开发适应教育数字化背景下的无锚等值技术提供指导。需要强调的是，虽然前面还提到其它一些涉及无锚设计下的等值研究，但是为了突出方法的应用性和聚焦研究主题，这里只对这三个涉及具体应用的研究进行总结、分析和讨论。

2　无锚题非等组设计下的三种等值实现

无锚题非等组设计下的等值问题一直以来受到测量领域的关注，研究者们针对经典测量理论（CTT）和项目反应理论（item response theory，IRT；Baker & Kim，2004）提出了一些解决方法（Kolen & Brennan，2014），但是这些方法在实际应用中面临着困难，比如样本依赖性和跨群体参数没有可比性。IRT 因其参数不变性成为无锚等值的主要框架，核心思路是通过共同量尺标定将不同测验版本的参数映射至同一单位系统，比如同时校准（concurrent calibration）等。

刘玥和刘红云（2015b）考虑了一种构造锚测验的方法，即通过专家对测验题目进行难度排序，然后从不同版本的测验中选出难度相近的题目构建锚测验进行等值。与之不同的是，杨钰萍（2019）则从考生的角度考虑了构造共同虚拟人的方法来实现，即基于考生的现有数据生成共同的虚拟人，构造成锚被试的设计。综合来看，这两种方法各有其适用的场景，但是也存在一些局限，比如刘玥和刘红云（2015b）的方法存在操作上较复杂且误差较大，而杨钰萍（2019）的方法需要较大的样本量，小样本量下等值的效果可能不好。下面分别对前面提到的三个典型研究进行介绍。

2.1　基于最小判别信息调整（MDIA）加权的方法

MDIA（Haberman，1984）是一种非参数统计的加权方法，用于提供具有指定加权平均值的样本，旨在通过优化信息熵差异，调整样本权重以满足特定约束条件。其核心思想是在给定某些矩（如均值、方差等）条件下，寻找与原始分布差异最小的新权重分布，从而实现对数据的校准或等值。

下面对有关的符号进行说明，考虑一个包含 T 个版本的测验，每个测验版本 $t(1<t<T)$ 都只施测一次，参加测验版本 t 的考生人数为 N_t。但是不确定参加不同测验版本的考生是否满足等组关系，即参加不同版本测验的考生的能力分布可能是不同的。在版本 t 上的考生 i 的分数 $X_{it}\in[x_{1t},x_{2t}]$，其中 x_{1t} 和 x_{2t} 分别是分数的上下界限，通常 $x_{1t}=0$，$x_{2t}=100$，实际的测验中 x_{2t} 的值有可能小于 100。用一个 J 维的向量 Z_{it} 表示考生 i 和测验版本 t 的关联关系。假定涉及考生 i 的变量信息 (X_{it},Z_{it}) 是独立的，且 $1<i<N_t$，$1<t<T$，对于具体的测验版本 t，(X_{it},Z_{it}) 有共同的分布 (X_t,Z_t)，X_t 是区间 $[x_{1t},x_{2t}]$ 内的随机变量，而 Z_t 是 J 维的随机向量。

记测验版本 t 的权重为 W_t，$W_t\geqslant 0$ 且 $\sum_{t=1}^{T}W_t=1$，则 Z_{it} 的样本均值 $\bar{z}_t$ 为

$$\bar{z}_t=N_t^{-1}\sum_{i=1}^{Nt}Z_{it},\tag{1}$$

进一步，$\bar{z}_t$ 的加权平均值 z 为

$$z=\sum_{t=1}^{T}W_t\bar{z}_t.\tag{2}$$

记测验版本 t 上被试 i 的权重 $\hat{w}_{it}$。对于测验版本 t，式 2 是权重 $\hat{w}_{it}$ 的平均值 $N_t^{-1}\sum_{i=1}^{Nt}\hat{w}_{it}$。$z$ 是链接向量的加权平均 $N_t^{-1}\sum_{i=1}^{Nt}\hat{w}_{it}Z_{it}$，$1<i<N_t$。基于这些约束，得到最小化的 KL（Kullback－Leibler，KL；Kullback&Leibler，1951）判别信息 $N_t^{-1}\sum_{i=1}^{Nt}\hat{w}_{it}\log(\hat{w}_{it})$。

需要注意的是，在实际的应用中不可能验证拟等组的样本权重是否完全令人满意，但这种利用样本权重调整对实现分数转换的目标是有帮助的，尤其是在大样本的时候。在等值应用中，一种常用的

做法是采用随机等组的方式将题目置于同一量尺上(Kolen & Brennan,2004;von Davier et al.,2004)。

Haberman(2015)以一个具体实例来评估所提出方法的表现。考虑某个英语测验,它评估的是母语非英语的考生的英语水平,包括两部分:听力和阅读,该测验共有 29 个版本。测验的两个部分是通过 NEAT(即非等组和锚检验的设计)进行设计,这样一来,这个例子中既可以采用传统的等值方法和拟等组的方法,也可以采用 MDIA 构造拟等组进行等值,并对它们的结果进行比较。实例中考生背景信息通过背景量表收集,量表涉及教育、出生日期、英语培训经历、工作状态,共 16 个类别记分题。经过编码,量表中涉及到 76 个独立的变量。

通过 MDIA 的调整,得到了量表分数的均值和分布。由于参加各个版本测验的考生人数非常多,都超过了 31000,平均人数为 54000。从结果来看,基于 MDIA 的方法在大样本时能够取得较好的等值结果。如果测验版本间没有适当的链接,也没有关于考生的重要信息,构造拟等组实现等值的做法值得尝试,即通过 MDIA 对被试进行权重调整构建拟等组从而实现不同测验版本间分数的比较。

2.2 基于背景变量匹配的方法

基于背景变量匹配实现分数转换方法的动机来源于缺失数据处理的方法,并与潜在结果框架(potential outcome framework, POF; Holland, 1986; Rubin,1974,2006)相关。Sinharay 和 Holland(2008)针对缺失数据处理应用于等值的原理进行了讨论。Dorans 和 Holland(2000)指出信度不同的测验版本之间是不能够进行等值的。基于测验版本中的“锚”,等值公式可以表示为如下的复合函数

$$G_2 1 = G_1(G_2^{-1}). \tag{3}$$

相对于 NEAT 设计中存在共同题,Longford(2015)考虑的是 NWEAT 设计,即不存在共同题,也不存在等组的被试,但是存在一系列背景变量。无论是链等值(chain equating, CE; von Davier et al., 2004)和后分层等值(post - stratification equating, PSE;von Davier et al.,2004)都不适用于 NWEAT 场景。倾向分数是将背景变量缩减到一个维度(Rosenbaum & Rubin,1983),利用考生背景变量(如性别、教育背景、社会经济状态)构建逻辑回归模型,预测考生属于两个群体的概率(倾向得分),它非常适合作为 NWEAT 下不同测验版本的“锚”。

对于 CE 和 PSE,平滑(或者连续化)都是其中的重要内容。假设有两套单位系统 A 和 B,它们分别对应于测量 TA 和 TB,并且对应收集了背景变量。这里的背景变量是指在单位系统里,如果一个变量在该单位系统里的值不受该单位或任何其他单位所选择或处理的影响,则该变量就可以被称为背景变量。针对 NWEAT 中的连续化,Longford(2015)基于 POF 提出了基于经验的连续化方法,通过对一组背景变量得出的倾向得分对考生进行匹配,形成来自测验不同版本的拟等组,然后将这两组考生分数进行等值处理。该方法包括两步:首先第一步是从不同单位系统里根据估计的倾向分数进行配对,在倾向分数分析里,采用 logistic 回归对背景变量进行分析。这一步的目的是形成两个大小相等的组,它们具有几乎相同的背景变量分布,因此它们可以看作是随机抽取的组,即可以看成是等组。接下来第二步是采用适用于等组等值的方法对这两组进行等值。需要注意的是,抽取背景变量的目的是为了形成等组,第二步不涉及背景变量。

2.3 基于共同认知结构的观察分数等值

出于等值公平性的考虑,Lord(1980)提出了等值的公平性准则,这个准则要求对于群体的每个考生,在给定能力的条件下,在不同测验版本上的分数分布相同。等值公平性准则保证了不同测验版本上的分数可以互换地使用,这也是等值与量尺化和预测的区别(Holland & Dorans, 2006)。根据 Lord(1980),等值公平性隐含了测验版本三个方面的含义:(1)相同信度;(2)相同认知结构;(3)群体不变性(van der Linden,2010)。由于等值公平性过于严格,几乎没有方法可以满足这个标准,很多研究者认为它阻碍了等值方法的发展(Dorans & Holland, 2000;Kolen & Brennan,2004)。因此,这个标准在大多数等值方法中往往被忽视或者产用折衷的方法(Kolen & Brennan,2004)。

公平性准则表明了等值要达到的终极目标,即实现无误差的等值转换(van der Linden &Wiberg, 2010)。具体来说,对于群体内能力为 θ 的被试,在测验的两个版本 X 和 Y 上的分数不能只是用单一的转换规则,而是需要通过一系列的转换规则来完成,即

$$\varphi_\theta(x) = F_{Y|\theta}^{-1}(F_{X|\theta}(x)), \theta \in R, \tag{4}$$

其中,$F_{X|\theta}$和 $F_{Y|\theta}$分别是测验的版本 X 和 Y 上能力为 θ 的被试观察分数的累积分布。van der Linden(2010)指出,单一的转换规则不能但是系列转换规则可以满足等值公平性准则,公式 4 同时也满足对称性和群体不变性准则。

考虑到等值公平性在实际应用中难以实现,van der Linden(2000,2006,2010)引入了局部等值。在进行局部等值的过程中,最大的挑战是为每个被试的能力选择相应的指标,这个指标需要能够体现该

被试属于当前的等值群体。简单来说,需要将每位考生根据能力水平进行分组,van der Linden 和 Wiberg(2010)的研究表明即使是“粗略的分类”,也会比将考生合并成单个群体时的等值效果好。基于IRT的局部等值不依赖随机抽样,因而比常规的方法更灵活,但是由于考生能力参数和题目参数估计的时候需要“连接”来实现统一量尺,因而也不能解决NEWAT下的等值问题。

不同测验版本考察相同认知结构是等值的前提条件之一,在CTT或单维IRT下认知结构是通过真分数或潜在特质来界定的,但是它们并不足以保证这些版本的测验间可以进行等值(Xin et al.,2015),而认知诊断能够提供心理测量中的认知结构表征(Embreston,1983)。

基于属性水平的测验结构,Xin 和 Zhang(2015)对认知诊断测验中观察分数的局部等值进行了探索。为方便介绍,下面先介绍相关的符号。假设 X 和 Y 为测验的两个版本,现在需要将 X 测验上的量尺转换到 Y 上,记这两个测验上的观察分数分别为 x 和 y,它们对应的属性关联矩阵、可达矩阵和Q矩阵(Tatsuoka,1983)分别为 A_X 和 A_Y、R_X 和 R_Y、Q_X 和 Q_Y。在属性水平上,测验的认知结构不再像单维IRT上那样模糊,它包括三个方面的内容:(1)属性集合;(2)属性间的层级关系(Leighton et al.,2004);(3)测验Q矩阵。测验的两个版本考察共同认知结构需要满足的条件是它们测量了相同的属性集合,即有相同的A矩阵和R矩阵。在认知诊断测量中,相同的属性掌握模式在测验的不同版本间具有相同的含义(Tatsuoka,2009)。因此,具有相同属性掌握模式(attribute mastery pattern,AMP)的考生可以作为在测验 X 和 Y 上的“锚”,从而实现等值。据 Xin 和 Zhang(2015),共同认知结构的充分必要条件是相同的Q矩阵,充分条件是相似的Q矩阵(指相同的题目类型,即题目和属性间的组合关系相同)。

给定AMP,测验 X 和 Y 上观察分数的完全分布可以表示为:

$$F_{Y|\alpha}(y)=F_{\varphi(X)|\alpha}(\varphi(x)),\alpha\in\Omega. \quad (5)$$

其中 $F_{Y|\alpha}(y)$ 表示在测验 Y 上AMP为 α 的被试的观察分数分布,这个观察分数分布从测验上转换到测验上的分布为 $F_{\varphi(X)|\alpha}(\varphi(x))$。对于所有的AMP,等值的误差可以计算如下:

$$e(x;\alpha)=F_{Y|\alpha}(y)-F_{\varphi(X)|\alpha}(\varphi(x)),\alpha\in\Omega. \quad (6)$$

当控制每种AMP的等值误差为0时,对 $F_{Y|\alpha}$ 求反函数,可以解出 y。

$$y=\varphi_\alpha(x)=F^{-1}_{Y|\alpha}F_{\varphi(X)|\alpha}(\varphi(x)),\alpha\in\Omega. \quad (7)$$

因为函数 $\varphi(x)$ 单调,所以 $F_{\varphi(X)|\alpha}(\varphi(x))=F_{X|\alpha}(x)$,则可进一步得到等值转换公式8:

$$\varphi_\alpha(x)=F^{-1}_{Y|\alpha}F_{\varphi(X)|\alpha}(\varphi(x)),\alpha\in\Omega. \quad (8)$$

3　对三种NEWAT设计下等值实现的讨论

等值作为测量中的关键技术之一,旨在解决不同测验版本间的分数可比性问题。然而,传统等值方法通常依赖锚题设计或群体能力分布相同的强假设,在非理想情境下(如无锚题、群体异质性高或测试目标多维化)面临显著挑战,这一研究领域具有重要意义,传统的测验等值往往依赖于锚题或锚被试,然而在实际情境中,获取锚题或锚被试可能面临诸多困难,如成本高昂、测验内容受限等,因此无锚测验的研究为测验等值提供了新的方向。Haberman(2015)、Longford(2015)及 Xin 和 Zhang(2014)的三项研究从拟等组构建和认知诊断模型融合等角度突破传统范式,为复杂测评场景提供了创新解决方案。

Haberman(2015)针对传统等值中“等组假设”难以满足的问题,提出拟等组的构建思路,通过统计调整而非随机分组实现群体能力分布对齐。其方法的优点主要体现在:(1)拟等组的引入:拟等组并不要求严格的等组,而是在一定统计意义上具有相似性的考生群体。通过构建拟等组,旨在解决在实际测试场景中,传统等值方法受限的问题。(2)非随机设计的适应性:利用协变量(如人口学特征、前期成绩)构建倾向得分模型,加权生成拟等组,降低群体差异对等值的干扰。(3)双重稳健性:结合回归调整与概率加权,即使倾向得分模型或能力预测模型存在误设,仍可最大程度保证估计结果的稳健性。这种方法具有很强的实践意义,即它可以为无法实施随机分组的大规模考试(如国家级的高风险考试)提供可行性方案。但是这种方法也存在一定的局限,比如协变量选择高度依赖领域知识,若遗漏关键变量(如未考虑进来的学习动机等),构造的拟等组的有效性可能会受损。另外在一些复杂场景下,如考生群体特征随时间快速变化且存在多种交互因素时,方法的有效性可能受到挑战,这也是在实际应用中使用该方法需要考虑的方面。

Longford(2015)考虑降低传统等值方法对锚题的依赖,提出一种无锚设计的等值方案,从研究方法上看,该方法在一定程度上突破了传统范式,尝试从测验的内在结构中挖掘信息,以实现不同测验版本之间的等值转换。这一方法的优势在于摆脱了对锚题的依赖,使得在无法设置锚题的情况下,也能进行测验分数的比较。例如,在一些特殊的测验场景中,如不同年份、不同地区的测验,难以找到合适的锚

题,此时该方法便具有实际应用价值。然而,该方法也存在一些局限性。首先,它需要一套足够丰富的背景变量,并且假设考生分配机制是合适的。也就是说,分数必须条件地独立于考生组,但这是一个不可检测的假设。其次模型的假设条件较为严格,对数据的质量和分布要求较高,需要大样本支持分层匹配。在实际应用中,数据往往难以完全满足这些假设,这可能导致等值结果的偏差。最后是选取的样本可能无法完全代表所有的考生群体。如果样本存在偏差,例如只选取了特定地区、特定教育背景或特定能力水平的考生,那么得出的等值结果可能不具有广泛的适用性,无法推广到更广泛的考生总体中。

Xin 和 Zhang(2015)首次将认知诊断模型与观察分等值结合,提出局部等值框架。这个方法至少有两个方面的贡献,一是通过 CDMs 估计考生在细粒度认知属性上的掌握状态,替代传统单维能力参数,使等值过程反映知识结构的差异。动态权重调整:二是基于认知属性掌握概率,为不同能力层次的考生分配差异化等值权重,提升低分段与高分段等值精度。将方法应用到在英语测试中,局部等值较传统方法(如链式等值)减少约 15% 的等值误差。Xin 和 Zhang(2015)可能存在的局限在于该方法需基于 Q 矩阵(题目 - 属性关联关系),若 Q 矩阵存在不合适(彭亚风 等,2018)或误设(de la Torre,2008;de la Torre & Chiu,2016;Qin et al. ,2024)或属性层级关系定义不正确(颜玉枝,2022;Yan et al. ,2025),可能导致等值系统偏差。另外,基于 CDM 的局部等值方法需要基于一些特定的模型假设,这些假设在实际中可能并不完全符合情况。例如,对认知诊断模型的假设可能过于简化了实际中考生的认知过程和作答行为,实际情况可能更为复杂多样,从而影响等值结果的准确性。并且,该研究主要集中在理论推导和模拟验证上,缺乏实际数据的验证研究,还需要进一步验证其在不同考试环境、不同学科领域等条件下的适用性和稳定性。

综合来看,针对 NEWAT 设计下的等值问题已经受到众多研究者们的重视,前面的这些方法在具体使用的时候需要结合方法本身的特点和测验的性质,从而选择合适的方法。在教育评价改革的当下,继续深入探索和研究适合教育数字化改革背景下测量实践的等值方法亟需开展。

4 进一步的研究

无锚题非等组等值研究正从“理想假设”走向“复杂现实”,虽然已有研究方法在各自的实验条件下已经取得了较好的结果,但是针对无锚(题目或被试)设计下的分数比较或测量参数转换仍然有许多工作需要开展。首先是未来研究需要进一步优化拟等组的构建方法,探索如何降低模型依赖,提高方法的稳健性。例如挖掘更多潜在的群体相似性特征,以更灵活地应对复杂多变的考生群体(Kim & Lu,2018;Lu & Guo,2018)。同时,可以将拟等组概念拓展到更多教育测量场景,如形成性评价、自适应测试等领域,进一步验证和拓展其应用范围。第四是需要开展跨学科的研究,与计算机、统计学等学科相结合,从不同的角度探讨测验等值问题,为无锚测验的发展提供更坚实的理论基础。最后是需要加强实证研究(Albano & Wiberg,2019),将无锚测验的方法应用到更多的实际测验场景中,比如我国高考不同科目分数的跨年比较等,积累实践经验,验证其有效性和可靠性。

其次是可以进一步优化现有的模型和方法,通过改进算法(Filonczuk & Cheng,2025;Hong & Cheng,2019;童昊 等,2022),降低对数据分布的严格要求,提高模型的稳健性。例如,可以结合机器学习中的一些技术,如深度学习,对数据进行更深入的挖掘和分析,从而更准确地实现测验等值(Jiang et al. ,2023)。第三是未来需探索非参数或半参数模型的潜力,例如结合核函数估计或贝叶斯分层模型,以进一步降低对锚题的依赖。同时,需关注群体差异的动态补偿机制,如引入协变量(如被试背景特征)修正群体能力分布偏差,提升不等组设计下的等值稳健性(He & Cui,2019;Leoncio et al. ,2022)。最后是随着在线化的测验越来越受到重视,需开发适合在线测验和小样本场景下的等值技术(Babcock & Hodge,2019),研究基于数据增强(如合成数据生成)或迁移学习的等值方法,以降低实际应用门槛。

参考文献

戴步云,罗照盛.(2012).题目难度分布和样本容量对两种 CTT 等值结果的影响.*心理学探新*,*32*(3),246 - 251.

李雪莹.(2016).*基于锚属性非等组设计的认知诊断等值方法研究:属性特征曲线等值法*(硕士学位论文).江西师范大学.

刘玥,刘红云.(2015a).多维数据 IRT 真分数等值和 IRT 观察分数等值研究.*心理学探新*,*35*(1),56 - 61.

刘玥,刘红云.(2015b).无锄题情况下测验分数等值方法探索:构造锄测验法.*心理科学*,*38*(6),1504 - 1512.

彭亚风,罗照盛,李喻骏,高椿雷.(2018).不同认知结构被试的测验设计模式.*心理学报*,*50*(1),130 - 140.

童昊,喻晓锋,秦春影,彭亚风,钟小缘.(2022).多级计分测验中基于残差统计量的被试拟合研究.*心理学报*,*54*(9),1122 - 1136.

王少杰,张敏强,黄菲菲,刘颖.(2024).参数估计误差对多级评分题型测验等值的影响.*心理学探新*,*44*(6),550-558,565.

王一波,杨涛,辛涛.(2017).无锚题测验等值设计方法研究进展.*考试研究*,(3),48-54.

颜玉枝.(2022).*排序理论在认知诊断中辅助标定属性层级关系:基于属性关联强度矩阵*(硕士学位论文).江西师范大学.

杨钰萍.(2019).*共同总体假设下基于虚拟人的测验等值研究*(硕士学位论文).江西财经大学.

杨志明.(2015).一年多考背景下分数等值的意义和方法.*教育测量与评价(理论版)*,(12),58-61.

朱殷睿.(2023).*事后引入多连接组的多群组无锚题测验等值研究*(硕士学位论文).浙江师范大学.

Albano,A. D. ,& Wiberg,M. (2019). Linking With External Covariates: Examining Accuracy by Anchor Type, Test Length, Ability Difference, and Sample Size. *Applied Psychological Measurement*,*43*(8),597-610.

Babcock,B. ,& Hodge,K. J. (2019). Raschversus classical equating in the context of small sample sizes. *Educational and Psychological Measurement*,*80*(3),499-521.

de la Torre,J. (2008). An empirically based method of Q-matrix validation for the DINA model:Development and applications. *Journal of Educational Measurement*, *45* (4), 343-362.

de la Torre,J. ,& Chiu,C. -Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*,*81*(2),253-273.

Filonczuk,A. ,& Cheng,Y. (2025). Robust estimation of the latent trait in graded response models. *Behavior Research Methods*,*57*(1),55.

Haberman,S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*,*12*(3),971-988.

Haberman,S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*,*40*(3),254-273.

He,Y. ,& Cui,Z. (2019). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*, *44* (4),296-310.

Hong,M. R. ,& Cheng,Y. (2019). Robust maximum marginal likelihood(RMML) estimation for item response theory models. *Behavior Research Methods*,*51*(2),573-588.

Jiang,Z. ,Han,Y. ,Xu,L. ,Shi,D. ,Liu,R. ,Ouyang,J. ,& Cai,F. (2023). The NEAT equating via chaining random forests in the context of small sample sizes: A machine-learning method. *Educational Psychological Measurement*,*83*(5),984-1006.

Kim,S. ,& Lu,R. (2018). The pseudo-equivalent groups approach as an alternative to common-item equating. *ETS Research Report Series*,*2018*(RR-18-02),1-13.

Kullback,S. ,& Leibler,R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*,*22*,79-86.

Kolen,M. J. ,& Brennan,R. L. (2014). *Test equating*,*scaling*, *and linking*:*Methods and practices*(3rd ed.). New York,NY: Springer Press.

Leighton,J. P. ,Gierl,M. J. ,& Hunka,S. M. (2004). The attribute hierarchy method for cognitive assessment:A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*,*41*(3),205-237.

Leôncio,W. ,Wiberg,M. ,& Battauz,M. (2022). Evaluating equating transformations in IRT observed-score and kernel equating methods. *Applied Psychological Measurement*,*47*(2), 123-140.

Lord,F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ:Erlbaum.

Lord,F. M. ,& Wingersky,M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equating". *Applied Psychological Measurement*,*8*(4),452-461.

Longford,N. T. (2015). Equating without an anchor for nonequivalent groups of examinees. *Journal of Educational and Behavioral Statistics*,*40*(3),227-253.

Lu,R. ,& Guo,H. (2018). A simulation study to compare nonequivalent groups with anchor test equating and pseudo-equivalent group linking. *ETS Research Report Series*,*2018*(RR-18-08),1-16.

Qin,C. Y. ,Dong,S. H. ,& Yu,X. F. (2024). Exploration of polytomous-attribute Q-matrix validation in cognitive diagnostic assessment. *Knowledge-Based Systems*,*292*,111577.

Tatsuoka,K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*,*20*(4),345-354.

Tatsuoka,K. K. (2009). *Cognitive assessment*:*An introduction to the rule space method*. New York,NY:Routledge.

vonDavier,A. A. ,Holland,P. W. ,& Thayer,D. T. (2004). *The kernel method of test equating*. New York,NY:Springer.

van der Linden,W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*,*65*(4),437-456.

van der Linden,W. J. (2006b). Equating error in observed-score equating. *Applied Psychological Measurement*,*30*(5), 355-378.

van der Linden,W. J. (2010). Local observed-score equating. In A. A. vonDavier (Ed.), *Statistical models for equating*, *scaling and linking*(pp. 201-223). New York,NY:Springer.

van der Linden,W. J. ,& Wiberg,M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement*,*34*(8),620-640.

Xin,T. ,& Zhang,J. H. (2015). Local equating of cognitively diagnostic modeled observed scores. *Applied Psychological Measurement*,*39*(1),44-61.

(下转第86页)

Wong, C. S., & Law, K. S. (2002). The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. *The Leadership Quarterly*, *13* (3), 243 – 274.

Ye, J., Yeung, D. Y., Liu, E. S., & Rochelle, T. L. (2019). Sequential mediating effects of provided and received social support on trait emotional intelligence and subjective happiness: A longitudinal examination in Hong Kong Chinese university students. *International Journal of Psychology*, *54* (4), 478 – 486.

Validity and Reliability of the Rotterdam Emotional Intelligence Scale among Chinese Employees

Li Chunxuan[1], Li Qirong[1,2]

(1. School of Business and Management, Jilin University, Changchun 130012;
2. JLU Research Institute of Innovation and Entrepreneurship, Jilin University, Changchun 130012)

Abstract: In the context of Chinese culture, the Rotterdam Emotional Intelligence Scale (REIS) was revised to improve sample representativeness. To test the reliability, validity, and measurement invariance of the scale, employees from both sides of the Taiwan Strait and three regions were selected as samples. 1021 employees from the three regions across the Taiwan Strait were tested using the Wong Law Emotional Intelligence Scale and Work Performance Scale as the performance indicators. The results showed that the Chinese version of REIS had good reliability and validity in the context of Chinese culture, and had the measurement invariance across gender, age, seniority, region groups, and the time points. Therefore, the Chinese version of REIS can be used as a suitable tool for the research of emotional intelligence.

Key words: Rotterdam emotional intelligence scale; reliability; validity; measurement invariance

(上接第 77 页)

Yan, Y. Z., Dong, S. H., & Yu, X. F. (2025). Using ordering theory to learn attribute hierarchies from examinees' attribute profiles. *Journal of Educational and Behavioral Statistics*. Advanced Online. Doi. 10.3102/10769986241280389.

Equating Under the Design of Non – equivalent Groups without Anchor Items

Dong Shenghong[1], Qin Chunying[1,2], You Xiaofeng[2], Yu Xiaofeng[1]

(1. School of Psychology, Jiangxi Normal University, Nanchang 330022;
2. School of Mathematics and Information Science, Nanchang Normal University, Nanchang 330032)

Abstract: In many testing contexts, the design of assessments often lacks intentional consideration for equating purposes, or the high – stakes nature of assessments inherently precludes conventional equating designs (e. g., those requiring anchor items or equivalent groups). Nevertheless, there remains a compelling need to compare scores across different test forms or evaluate examinee abilities under such constraints, thereby necessitating equating solutions. This study focuses on the equating problem under anchorless designs (i. e., scenarios devoid of anchor items and equivalent groups). We critically evaluate existing methodologies and techniques in the literature, clarify the applicability and limitations of different approaches, and propose future research directions.

Key words: equating; anchor items; non – equivalent group; construct