文章编号:1000-5862(2017)05-0454-08

非等组锚题设计下 IRT 等值方法比较及其应用

黎光明,王小婷
(华南师范大学心理学院,心理应用研究中心,广东 广州 510631)

摘要:总结了基于非等组锚题设计下的两大类 IRT 等值方法:同时参数标定和分别参数标定. 分别参数标定包含了线性参数转换和固定参数标定,以等值精度为评价标准对这 3 类等值方法的效果和适用条件进行归纳并做出相应的评析,为测验工作者选择合适的等值方法进行项目参数和测验等值提供参考依据.

关键词:项目反应理论;测验等值;非等组锚题设计
中图分类号:TP 841 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2017.05.03

0 引言

测验等值,是指测量同一心理特质的不同测验分数或试题参数,通过一定的数学模型,转换成同一单位系统中的量数,以利于相互比较的方法[1]. 测验等值在题库建设和教育评价中必不可少. IRT 框架下实施等值,不仅理论完善,前提条件较容易满足,而且等值关系式也十分简洁. 测验等值有不同的等值设计,如单组设计(single-group design)、随机等组设计(counterbalanced random-group design)、平衡单组设计(single group design with counterbalance)、非等组锚测验设计(Non-Equivalent groups with Anchor Test,NEAT)和共同被试组设计(common group design)(也叫锚人设计)等[2].

非等组锚测验设计是目前实际应用中最广泛的等值设计,因为相对于其它等值设计,这种设计更为有效、易行. 在实际情况中常有这样的情况:无法采集一个被试样本,让被试接受 2 个不同形式的测验施测,又难以获得 2 个总体分布相同的被试样本来分别接受 2 个测验的施测. 十八届三中全会审议通过的《中共中央关于全面深化改革若干重大问题的决定》,明确了未来高考改革方向:探索全国统考减少科目、不分文理科、外语等科目社会化考试一年多考. 高考“一年多考”最需要解决且最难的问题是:不同次考试的多份试卷分数是否“等值”,即需要将不同次的高考试卷实现等值,然而不能找到一个被试样本同时施测 2 次的高考试题. 在第 1 次抽取被试,第 2 次高考试卷不可能提前施测;在第 2 次抽取被试,第 1 次高考试卷已经“曝光”,测试已经不准确了. 因此只能在不同次考试分别抽取被试样本,分别施测当次考卷,但是又无充分理由证明 2 次考生总体分布是相同的,这种情况下只能采用非等组锚题设计. NEAT 设计对被试样本的要求没有像单组设计和等组设计般严格,而且锚题相对于被试样本来说,不管是题目的获取过程,还是测试的过程,都是比较容易控制的,因此 NEAT 设计的应用更加广泛[2-3]. NEAT 是大型测验中最常用的等值设计之一,如托福、GRE、SAT 等著名考试均采用非等组锚题设计对多次考试进行等值.

在计算机自适应测验(CAT)题库建设中,一般把题库中原有的测验称为基准测验(base form)或旧测验,其项目参数均在同一量尺上. 对于新编制的项目,题库建设者一般将题库中原有的部分题目作为锚题,与新编制的项目,即新题,合并组成目标测验(target form)或新测验,通过 NEAT 设计将目标测验的独立项目等值到基准测验的量尺上去,这样就能将新题的项目参数统一到原来题库的量尺上去. 具体来说,非等组锚测验设计就是将 2 个不同的测验,如基准测验和目标测验,分别施测于不同的被试样本组,但这 2 个测验中分别都包含一组相同的题目,即锚题(Anchor Items),用来作为进行等值转换的中

收稿日期:2017-02-18
基金项目:国家自然科学基金(1470050),广东省哲学社会科学发展“十三五”规划 2017 年度一般项目(GD17CXL01),广州市哲学社会科学发展“十三五”规划项目(2017GZYB111),广东省 2015 度高等教育教学改革项目(粤教高函〔2015〕173 号)和华南师范大学 2014 年度校级高等教育教学研究和改革项目(教学〔2014〕52 号)资助项目.
作者简介:黎光明(1977-),男,江西广昌人,副教授,博士,主要从事心理统计与测量的研究. E-mail:Lgm2004100@sina.com

介，2 个测验中的非锚题题目叫做独立项目（Unique Items）. NEAT 设计如图 1 所示[4].

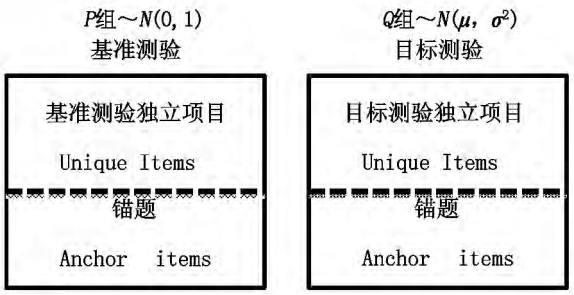

图 1　非等组锚测验设计模式

作为 2 个独立测验进行等值转换的中介，锚题要具有代表性，能够代表整个测验，作为整个测验的浓缩版（Mini-Test）；锚题的题型应该尽量涵盖测验中的所有题型；难度指标全距应该足够宽，区分度指标应该至少在中等水平以上；对于锚题数量的要求，包含 40 道题目或以上的测验中，锚题量至少应为测验总题量的 20%[5]，这样基准测验和目标测验就会有较高的相关. 相关越高，链接能力就越强，保证等值结果的稳定性，从而越有利于对测验等值关系的认识[6-7].

1　IRT 等值方法

进行 IRT 测验等值，需要进行等值设计、数据收集、等值模型选取、模型参数估计、量表化、测验等值、等值结果评价等 7 个步骤. 其中一个很重要的步骤——量表化，就是将从不同测验估计出的项目参数等值或标定到同一个量尺上去. 在非等组锚测验设计下，基于 IRT 的项目参数等值或标定的方法，即 IRT 等值方法，主要有两类：分别参数标定和同时参数标定，其中分别参数标定包含线性参数转换和固定参数标定[8].

1.1　分别参数标定

1.1.1　线性参数转换（Linking Separate Calibration，LSC）　线性参数转换对基准测验和目标测验的项目参数分别估计，使用锚题作为链接，通过线性转换关系，将新测验组即目标测验组，标定到旧测验组基准测验的量尺上去. 目标测验上项目参数的计算过程如下：a_i、b_i、c_i 分别表示新测验项目 i 的区分度、难度、猜测参数，转换后的参数分别表示为：$a_i^* = a_i/A$，$b_i^* = A \times b_i + B$，$c_i^* = c_i$. 等值系数 A 和 B，可以通过数学方法估计出来，包括矩估计法（Moment methods）[9-10] 和特征曲线法（Characteristic curve method）[11-12].

矩估计法主要包括了平均数-平均数法（Mean/Mean Method，MM）[9]、平均数-标准差法（Mean/Sigma Method，MS）[10]、稳健的平均数-标准差法（Robust Mean and Sigma Method）[13] 及稳健迭代加权平均数-标准差法[12].

特征曲线法，根据其等值准则的不同，主要有 Haebara 法（Hcrit，又称项目特征曲线法）[11]、Stocking-Lord 法（SLcrit，又称测验特征曲线法）[12]. 其他等值准则还有对称相对熵准则（Symmetric Relative Entropy criterion，SREcrit）[14]、Haebara 加权准则（Weighted criterion，Wcrit）[15]、绝对值等值准则（Absolute Value equating method）[16] 和余弦等值准则（Cosine criterion，COScrit）[17] 等.

此外，还有回归方法. 但是，由于回归方程并不是对称的，跟测验等值的基本要求不合，因此实践中较少使用回归法.

1.1.2　固定参数标定（Fixed Item Parameter Calibration，FIPC）　固定参数标定结合了线性参数转换和同时参数标定的特点，又被称为锚题估计法（item anchoring estimation method）[8,18]. FIPC 对基准测验和目标测验分别估计项目参数. 先估计基准测验上锚题的参数，在进行目标测验的参数估计时把锚题参数固定为已经得到的值，这样就使得目标测验的参数自动与基准测验位于一个量表中. 该方法具有较好的灵活性，适用于不同的等值设计，并且 FIPC 相对于其它的等值方法更简单、省时[19].

1.2　同时参数标定（concurrent calibration，CC）

在 NEAT 设计下，同时参数标定是将 2 个测验的数据合并，看成同一个测验，将一组被试未作答的在另一个测验中独立项目上的反应当作缺失值，在单次标定程序中就能同时估计出基准测验和目标测验的项目参数，不需要求取等值系数. 由于基准测验和目标测验都含有锚题，CC 得到的 2 个测验的项目参数就在同一个量尺上，但是该量尺是基于基准组和目标组所有被试的水平，并不是在基准测验的量尺上[20-22].

1.3　IRT 参数估计程序

随着计算机软件和硬件的发展，IRT 参数估计程序也得到了有效的发展. 从早期只适用于 0-1 计分的 LOGIST[23]、BILOG[24] 到现在能同时处理多级计分的 MULTILOG[25]、PARSCALE[26] 等. 但是同一种等值方法，在不同软件中参数估计的方法可能会有所不同，如联合极大似然估计、边缘极大似然估计、贝叶斯估计、EM 算法等. 即使是同一种参数估

计方法,在不同软件中估计的结果会存在差异. 如在 FIPC 等值方法中,不同软件中 EM 循环的次数或先验能力分布的更新次数可能会有所不同,又如 BILOG-MG 不更新先验信息,而 PARSCALE 中更新先验信息,两者得出的等值结果有所区别. 因此,在研究等值方法的差异时,应该要注意分离出不同参数估计方法的差异.

2 IRT 等值方法的比较

2.1 线性参数转换(LSC)不同方法的比较

矩估计法(包括平均数-平均数法、平均数-标准差法、稳健的平均数-标准差法、稳健迭代加权平均数-标准差法)均未能完整利用项目参数信息,并容易受到奇异值的影响. 而特征曲线法能够有效利用项目参数信息,比矩估计法更加优良[5,11-12]. 矩估计法,一些研究者认为平均数-标准差方法较好,因为 b 参数的估计结果比 a 参数更稳定,如在 H. Ogasawara 等[27]的研究中,平均数-标准差法比平均数-平均数法等值结果更稳定. 然而,F. B. Baker 等[28]研究得出平均数-平均数法更稳定. 矩估计法中何种方法最佳,仍存在争议.

对于特征曲线法的研究,不少学者发展了许多不同的等值准则[9-12,14-15,17,28],其中最为常用的是 Haebara 法和 Stocking-Lord 法. F. B. Baker 等[28]和 H. Ogasawara[29-30]的研究中均得出 Stocking-Lord 法在求取等值系数的各方法中的精确性最高,在 IRT 真分数等值时优于 Haebara 法. 等值准则的选取会影响等值的效果,关于不同等值准则的比较,需要考虑不同的 IRT 模型、不同题型、样本量、锚题和被试能力分布差异等因素. 线性参数转换不同方法的优缺点比较如表 1 所示.

表 1 线性参数转换不同方法的优缺点比较

求取等值系数的方法		优点	缺点
矩估计法	平均数-平均数法,平均数-标准差法	保证了变换的对称性,且方法比较简单	未考虑两套参数的估计精度的差异,等值精度容易受到极端值的影响
	稳健的平均数-标准差法	考虑了估计精度的不同,修补了平均数-标准差法的不足	忽略了使用项目区分度之间的关系,等值精度容易受到极端值的影响
	稳健迭代加权平均数-标准差法	考虑了估计精度的不同,极端值的影响	忽略了使用项目区分度之间的关系
特征曲线法	Haebara 法	能够有效利用项目参数信息,比矩估计法优良	需要迭代,算法较复杂
	Stocking-Lord 法	能够有效利用项目参数信息,比矩估计法优良,IRT 真分数等值时优于 Haebara 法	需要迭代,算法较复杂

2.2 固定参数标定(FIPC)不同方法的比较

在过去,由于估计方法的难以实现,固定参数标定法未能充分表现出其优势. 近年来,随着估计方法的发展,如采用 EM 算法实现了边际极大似然估计(MMLE)、联合极大似然估计(JMLE)和贝叶斯估计等. 随着参数估计软件的发展,如 BILOG-MG 和 PARSCALE 等,研究者开始研究 FIPC 方法的性能[31-33].

S. Kim[32]比较了 5 种 IRT 固定参数标定方法,这 5 种方法的区别在于更新先验能力分布的次数和 EM 循环的使用次数不同,分别为没有先验能力分布的更新和 EM 循环次数为 1(no prior weights updating and one EM cycle,NWU-OEM)、多次 EM 循环(multiple EM cycle,NWU-MEM)、1 次先验能力分布的更新和 1 次 EM 循环(one prior weight updating and one EM cycle,OWU-OEM)、多次 EM 循环(multiple EM cycles,MWU-MEM)、多次更新先验能力分布并使用多次 EM 循环(multiple weights updating and multiple EM cycles,MWU-MEM),结果表明在目标组被试不同能力分布下:$N(0,1)$,$N(0.5,1.2^2)$,$N(1,1.4^2)$,只有 MWU-MEM 具有良好的参数估计精度[32].

T. Kang 等[4]得出了与 S. Kim[32]一致的结果,在其研究中,比较了 FIPC 的 2 种方法,无先验信息更新的 FIPC-BMG 方法和有多次更新先验能力分布的 FIPC-PSL 方法,这 2 种方法分别在 BILOG-MG 和

PARSCALE 中实现. 结果显示,相对于有先验信息更新的 FIPC-PSL 方法,FIPC-BMG 方法更有可能低估了均值和标准差的真值,且 FIPC-BMG 方法有一定的系统误差[4].

综上所述,在使用 FIPC 并选择使用 EM 算法时,应该选用 MWU-MEM 的固定参数标定法,也就是多次更新先验能力分布并使用多次 EM 循环,这样即使在目标组被试能力分布与基准组差异较大时,这种方法的等值精度仍较高.

2.3　同时参数标定(CC)和线性参数转换(LSC)方法的比较

有关 CC 和 LSC 的研究有很多,对于两者的比较也有较多学者进行了相关研究. 以下列举了部分具有代表性的研究结果,并对前人的研究结果进行了评析和总结.

N. S. Petersen 等[34] 及 M. S. Wingersky 等[35] 均得出 CC 比 LSC 的等值效果更好. 但是,这些研究都是在 LOGIST 程序下进行的,这个程序使用了联合极大似然估计法对项目参数进行估计.

S. H. Kim 等[20] 用 BILOG 和 MULTILOG 模拟比较了 CC 和 LSC,他们的研究包含了 4 个锚题数量水平(5,10,25,50),测验总题目数为 50,并包含了等组和非等组的设计来模拟水平和垂直等值的情况. 他们采用了均方根误差(Root Mean Square Difference,RMSD)和欧式距离均值(Mean Euclidean Distances,MED)来评价不同条件下 CC 和 LSC 中的特征曲线法的等值效果. 研究结果指出,当锚题数量较小时,LSC 要优于 CC,相对能得到更精确的结果; 当锚题数量较大时,2 种方法得到了类似的等值结果[20]. 但是,A. B. Hanson 等[8] 指出,该研究的不足在于,LSC 用的是 BILOG 软件,而 CC 用的是 MULTILOG,因此,在 NEAT 设计下这 2 种等值方法的差异与软件的差异混淆[8]. BILOG 中也可以实现 CC,但是在 NEAT 下,BILOG 不能精确估计出非等组被试指定的不同能力分布.

A. B. Hanson 等[8] 在非等组锚题设计下用模拟研究的方法比较了 CC 和 LSC,这两类方法均在 BILOG-MG 和 MULTILOG 中实现,以避免混淆方法差异和软件差异. LSC 具体共有 4 种方法: 2 种项目特征曲线法(Stocking-Lord 法和 Haebara 法)及 2 种矩估计法(MM 和 MS). 该模拟研究考虑了 4 种影响因素: 等组和非等组被试群体、CC 和 LSC、锚题数量大小、样本量大小. 评价等值效果的指标采用了基于 IRT 真分数等值的平均偏移均方差(Mean Squared Errors,MSE)和基于加权的和未加权的项目特征曲线的 MSE. A. B. Hanson等[8] 的研究结果指出,总体看来,CC 要比 LSC 产生的误差小,且不同的被试群体非等组被试下的等值误差要明显大于等组被试. 随着锚题数目的增多,等值误差呈减少的趋势. 在其它因素不变的条件下,样本量太小会增大等值误差[8].

J. S. Kim 等[21] 在多级 IRT 模型下比较了 CC 和 LSC. 在他们的模拟研究中同样考虑了样本量、锚题数量、等组和非等组 3 个影响等值效果的因素. 采用了 MULTILOG 软件进行多级 IRT 模型的参数估计. 采用项目参数和能力参数的 RMSE 值来评估不同等值方法对参数真值的修复程度. 该研究得出,CC 比 LSC 产生的等值误差要小,尽管这个差异非常小[36].

A. A. Béguin 等[37-38] 的研究中,用单维 IRT 模型处理多维 IRT 模型下产生的数据,来比较 CC 和 LSC. 同时,他们还在多维 IRT 模型下用 CC 进行测验等值,探讨了在忽略数据多维性的条件下,等值精确性是否有差异,并考虑了等组和非等组、潜在特质的方差大小这 2 个因素. 该研究采用了基于等值结果估计的分数分布和模拟产生的分数分布的差异来评估等值精度. 结果显示,在单维 IRT 模型下,多维的数据会影响 CC 和 LSC 的表现. 在等组和非等组的不同条件下,CC 和 LSC 的等值精度是不一样的. 在非等组群体条件下,基于单维 IRT 模型下的等值方法的表现相对于多维 IRT 等值方法,明显受到了多维数据的影响. 研究还表明,在等组群体条件下,CC 产生的等值误差比 LSC 小[3-38].

T. Kerkee 等[39] 用真实数据比较了垂直等值下 Stocking-Lord 法和 CC,结果显示,CC 下会有更多无法收敛的项目,并且他们还发现,LSC 在每个年级上都比 CC 有更好的拟合性.

A. Sayaka 等[40] 研究认为项目特征曲线表现得更好,CC 和 LSC 这 2 种方法在锚题数量较小的时候等值的效果都不够好. 当测验试题总数和锚题数量增大时,等值效果会变好些.

尽管有学者对同时参数标定法和分别参数估计法进行了对比研究,但是很难得出一个结论说哪种方法更优越. 上述模拟研究和实证研究结论的不同,可能是由于他们各自的研究中有许多不同因素造成的,如不同的数据类型、样本数量、等值准则、估计方法、参数估计程序等. 因此,这些研究的结论都不能充分地证明同时参数标定法比分别参数标定法能产生更精确的等值结果. 但是,测验工作者仍然可以从

这些研究中得到启发,根据不同的条件选用不同的项目参数标定方法,以最大程度提高等值的精确性.

2.4 同时参数标定(CC)、线性参数转换(LSC)和固定参数标定(FIPC)这3种等值方法的比较

目前,关于CC、LSC和FIPC这3种等值方法的比较研究比较少. N. S. Petersen 等[34]进行了一项传统等值方法和IRT等值方法的比较研究,比较了CC、FIPC和特征曲线法. 研究结果显示:CC得到了最稳健的等值结果[34]. Li Yuanhua 等[18]比较了FIPC和特征曲线法,参数修复结果显示:FIPC和特征曲线法都得到了稳健且精确的等值结果.

Zhang Zhonghua 等[41]在NEAT设计下用模拟研究的方法比较了这3种等值方法. 该研究考虑了4种影响因素:目标组群体能力分布分别为$N(0,1)$和$N(1,1)$,基准组均为$N(0,1)$、锚题数量(10,20,40)、样本量(200,500,1 000)、锚题的平均难度($\bar{b}bbb$)($\bar{b}bbb$表示所有锚题的难度均值). 结果显示:当基准组和目标组的能力分布没有差异时,CC、LSC和FIPC这3种方法对被试能力真值的修复得到了相似的结果. 虽然在某些条件下,CC产生的等值误差更大[41],这个差异可能是由于对区分度参数的修复程度不一造成的,这与S. H. Kim 等[20]的结果是一致的. S. H. Kim 等[20]发现,CC比LSC在对区分度参数的修复上表现不够好,尤其是在水平等值的条件下. 但是,随着样本量的增大,CC和LSC的差异减小了[20] A. B. Hanson 等[8]也发现矩估计法和特征曲线法之间的差异要大于特征曲线法和CC之间的差异. 当锚题的平均难度值小于整个测验的难度值时(除了样本量为200的情况),CC优于特征曲线法[8]. 因此,可以说在多数情况下,同时参数标定在非等组锚题设计下的等值效果要优于特征曲线法,S. H. Kim 等也得到了同样的结果[36].

T. Kang 等[4]对这3种方法进行了比较,模拟研究中考查的因素有:被试数量、能力分布、锚题的数目. 模拟结果的评价采用了潜在能力参数分布(underlying ability distributions)、项目特征曲线(item characteristic curves)和测验特征曲线(test characteristic curves)的返真修复(recovery)程度. FIPC可以在BILOG-MG和PARSCALE程序中实现,但是这2种程序实现FIPC的方法有所不同. CC和LSC的等值结果可以直接比较,在所有条件下这2种方法得到的结果返真性都较好. 在2种FIPC程序中,只有在合理使用PARSCALE程序时,才能得到和前2种方法相似的项目参数链接结果[4].

王菲等[42]在等级记分模型下采用实测数据对这3种方法进行了比较,等值效果的比较采用了RMSD和REMSD[50]为评价标准,结果得出分测验1以平均数-平均数法的等值效果最好,分测验2则以FIPC为佳. 该研究的参数估计是在PARSCALE软件中进行,其他程序使用了Visual Foxpro 6.0自行编写[42].

综上,CC、LSC和FIPC这3种等值方法的优缺点如表2所示.

表2 不同等值方法的优缺点比较

等值方法		优点	缺点
同时参数标定(CC)		在单次标定程序中就能将项目参数标定到同一量尺上,能在多数IRT软件中实现,且程序运行耗时少;等值效果不易受被试群体能力差异的影响	当锚题数量较小时,等值误差较大;项目参数等值到基准组和目标组合并后的量尺上,而不是基准组的量尺
线性参数转换(LSC)	矩估计法	方法简单,耗时短	未能有效利用项目参数信息;等值精度易受被试群体能力差异的影响
	特征曲线法	能够有效利用项目参数信息,比矩估计法优良	求取等值系数需要迭代,程序运行耗时长;等值精度易受被试群体能力差异的影响
固定参数标定(FIPC)		标定灵活,适用于不同的等值设计	参数估计复杂,且不同的IRT软件中实现FIPC的具体步骤有所不同,BILOG-MG中FIPC程序有一定的系统误差

3 总结与展望

3.1 等值方法选择

两大类等值方法并没有好坏之分,而是各自有不同的适用条件,应根据实际情况选择合适的等值方法,以尽量减少等值误差,提高等值精度.

以下总结最为常见的0-1记分题型,且采用3参数逻辑斯蒂模型(3PLM)时,不同的条件下采用哪种等值方法能达到最佳等值效果.

1) 当锚题数量为中等或较大水平时,群体能力分布没有差异或差异较小时,CC 和 LSC 的等值效果均较好,测验工作者可根据实际需要等值到哪个量尺上来选择不同的方法: (i) 当需要等值到基准组被试群体的量尺上时,可选用 LSC; (ii) 当需要等值到基准组和目标组合并后的被试群体的量尺上,应选用 CC.

2) 当锚题数量为中等或较大水平时,若群体能力分布差异较大时,采用 CC 等值效果更佳,若需转换到基准测验量尺上,可先采用 CC 方法估计出项目参数,再采用矩估计法转换到基准测验量尺上.

3) 当锚题数量为较小水平时,使用 LSC 中的特征曲线法时等值效果较好.

4) 当构建大型题库时,采用 FIPC 更为灵活、有效、省时.

5) 样本量越大,不同等值方法的差异越小,当样本量较大时(一般为 3 000 左右) ,不同等值方法的等值精度均较高,且差异较小,测验工作者可灵活选择等值方法,若对等值样本量没有信心,可参照上述 4 条选择合适的等值方法.

3.2　研究展望

对不同的等值方法的比较,普遍考虑的因素有: 等值方法(包含等值准则) 、样本量、锚题数目和被试群体能力水平差异. 近年来已有学者开始考虑其他因素,如不同题型(0-1 记分,多级记分,混合题型,题组题型) 、单维或多维 IRT 和不同的模型等. 例如,Yao Lihua 等[43] 比较了含有混合题型的测验,锚测验的构成对等值结果的影响; Tian Feng 在基于 3PLM 和 GPC(广义分布评分模型) 的混合模型下,得出了同时参数标定比线性参数转换中的 SLcrit 方法的等值精确性更高的结果[43]; S. H. Kim 等在基于 3PLM 和 GPC 的混合模型下,得出特征曲线法要优于矩估计法,而 Haebera 法又略优于 SL 法[44]; Yao Lihua 等针对混合题型测验提出了用多维分部评分模型进行等值[45]. 然而国内许多研究对等值方法的探讨和比较都是在 0-1 记分项目的题型下进行的,关于多级记分题型,国内虽有实证研究[42] ,但是该研究对不同等值方法的探讨和比较都是基于同一个模型——等级记分模型之下进行的,未能涉及其他已有的多级记分模型,基于不同模型之下等值方法的比较仍是一个有待研究的内容.

以上研究都是在直接等值(direct equating) 的条件下进行的,对于间接等值(indirect equating) 下不同等值方法的比较目前只有 Li Deping 等[46] 进行了模拟研究,结果显示特征曲线法(SLcrit 和 Haebera) 的表现优于矩估计法(MM 和 MS) . 然而该研究仅比较了间接等值下 LSC 下的不同方法,缺少和 FIPC、CC 方法的比较.

对于不同等值方法得到的等值函数,还可以通过求取等值函数均值的方法得到新的等值函数,以减少等值误差,提高等值稳定性. 这个方法最开始是由 Angoff 提出,Angoff 对同一个线性等值函数进行多次估计,得到不同的估计函数,对这些函数进行平均可能会得到一个更合适的等值函数[47] ,且已有不少学者对求取等值函数均值进行了研究[48-50] . 目前还未有学者专门将求取等值函数均值的方法,与 CC、LSC 和 FIPC 进行比较研究,未来的研究方向可对此进行相关研究.

此外,虽然多数等值模拟研究的结果会采用真值的修复程度 Bias、RMSD 等指标,但是等值效果的评价标准问题一直是等值研究中的难点,不同的研究采用的评价标准不完全一致,确定或者寻找一种评价等值研究的一致评价标准是值得进一步研究的课题.

4　参考文献

[1] 张敏强,胡晖. 略论测验等值的理论、方法和应用 [J] . 华南师范大学学报: 社会科学版,1988(4) : 113-118.

[2] 漆书青,戴海琦. 项目反应理论及其应用研究 [M] . 南昌: 江西高校出版社,1992.

[3] 漆书青,戴海琦,丁树良. 现代教育与心理测量学原理 [M] . 北京: 高等教育出版社,2002.

[4] Kang Taehoon,Petersen N S. Linking item parameters to a base scale [J] . Asia Pacific Education Review,2012,13(2) : 311-321.

[5] Kolen M J,Brennan R L. Test equating,linking,and scaling: methods and practices [M] . New York: Springer Verla,2004.

[6] 罗照盛. 项目反应理论基础 [M] . 北京: 北京师范大学出版社,2012.

[7] Kolen M J,Brennan R L. Test equating scaling and lingking: method and practices [M] . 3ed. New York: Springer Verlag,2014.

[8] Hanson A B,Beguin A A. Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design [J] . Applied Psychological Measurement,2002,26(1) : 3-24.

[9] Loyd B H,Hoover H D. Vertical equating using the rasch model [J] . Journal of Educational Measurement,1980,17

(3):179-193

[10] Marco G L. Item characteristic curves solutions to three intractable testing problems [J]. Journal of Educational Measurement,1977,14(2):139-160

[11] Haebara T. Equating logistic ability scale by weighted least squares method [J]. Japanese Psychological Research, 1980,22(3):144-149.

[12] Stocking M L,Lord F M. Developing a common metric in item response theory [J]. Applied Psychological Measurement,1983,7(2):201-210.

[13] Linn R L,Levine M V,Hastings C N,et al. Item Bias in a test of reading comprehension [J]. Applied Psychological Measurement,1981,5(2):159-173.

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

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

[16] 程德巧. 绝对值等值准则及求解算法的应用 [D]. 南昌:江西师范大学,2005.

[17] 吴锐,丁树良,甘登文. 一种新的项目反应理论等值准则:余弦准则 [J]. 江西师范大学学报:自然科学版,2008,32(2):224-245.

[18] Li Yuanhua,Tam H P,Tompkins L J. A comparison of using the fixed common pre-calibrated parameter method and the matched characteristic curve method for linking multiple-test items [J]. International Journal of Testing, 2004,4(3):267-293.

[19] Paek I, Young M J. Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data [J]. Applied Measurement in Education,2005,18(2):199-215.

[20] Kim S H,Cohen A S. A comparison of linking and concurrent calibration under item response theory [J]. Applied Psychological Measurement,1996,22(2):131-143.

[21] Kim J S,Hanson B A. Test equating under the multiple-choice model [J]. Applied Psychological Measurement, 2002,26(3):255-270.

[22] Wingersky M S,Lord F M. An investigation of methods for reducing sampling error in certain IRT procedures [J]. Applied Psychological Measurement,1983,8:347-364.

[23] Wingersky M S, Barton P, Lord F. LOGIS [EB/OL]. [2017-01-06]. http://www.ets.org.

[24] Mislevy B,Bock D. BILOG [EB/OL]. [2017-01-09]. http://www.ssicentral.com.

[25] Thissen D. MULTILOG: Multiple categorical item analysis and test scoring using item response theory(Version 6.0) [EB/OL]. [2017-01-10]. http://www.chegg.com.

[26] Muraki E,Bock R D. PARSCALE: IRT analysis and scoring of rating scale data [J]. Science,2014,343(6169):350.

[27] Ogasawara H. Asymptotic standard errors of IRT equating coefficients using moments [J]. Economic Review,2000, 51(1):1-23.

[28] Baker F B,Al-Karni A. A comparison of two procedures for computing IRT equating coefficients [J]. Journal of Educational Measurement,1991,28(2):147-162.

[29] Ogasawara H. Item response theory true score equating and their standard errors [J]. Journal of Educational Behavioral Statistics,2001,26(1):31-50.

[30] Ogasawara H. Least square estimations of item response theory linking coefficients [J]. Applied Psychological Measurement,2001,25(4):3-21.

[31] Ban Jae-chun,Hanson B A,Wang Tianyou,et al. A comparative study of on-line pretest-item calibration/scaling methods in computerized adaptive testing [J]. Journal of Educational Measurement,2001,38(3):191-212.

[32] Kim S. A comparative study of IRT fixed parameter calibration methods [J]. Journal of Educational Measurement,2006,43(4),355-381.

[33] Zhang Z H,Ni Y J. A comparison of fixed pre-calibrated parameter method,linking separate calibration and concurrent calibration for linking different groups [C]. Chicago: The Annual Meeting of American Education Research Association,2007.

[34] Petersen N S,Cook L L,Stocking M L. IRT versus conventional equating methods: a comparative study of scale stability [J]. Journal of Educational Statistics,1983,8(2):137-156.

[35] Wingersky M S,Cook L L,Eignor D R. Specifying the characteristics of linking items used for item response theory item calibration [R]. ETS Research Report,1987:87-24.

[36] Kim S H,Cohen A S. A comparison of linking and concurrent calibration under the graded response model [J]. Applied Psychological Measurement,2002,26(1):25-41.

[37] Beguin A A,Hanson B A,Glas C A W. Effect of multidimensionality on separate and concurrent estimation in IRT equating [C]. New Orleans: The National Council on Measurement in Education,2000.

[38] Beguin A A,Hanson B A. Effect of non-compensatory multidimensionality on separate and concurrent estimation in IRT observed score equating [C]. Seattle: The National Council on Measurement in Education,2001.

[39] Kerkee T,Lewis D M,Hoskens M,et al. Separate versus concurrent calibration methods in vertical scaling [C].

Chicago: The National Council on Measurement in Education,2003.

[40] Sayaka A,Shinichi M. A comparison of equating methods and linking designs for developing an item pool under item response theory [J]. Behaviormetrika,2011,38(1):1-16.

[41] Zhang Zhonghua. Comparison of different equating methods and an application to link testlet-based tests [D]. Hong Kong: Chinese University of Hong Kong,2010.

[42] 王菲,任杰,张泉慧,等. 等级记分模型下几种等值方法的比较研究 [J]. 中国考试,2013(6):10-17.

[43] Tian Feng. A comparison of equating/ling using the stocking-Lord method and concurrent calibration with mixed-format test in the non-equivalent groups common-item design under IRT [D]. Boston: Boston College,2011.

[44] Kim S H,Lee W C. An extension of four IRT linking methods for mixed-format tests [J]. Journal of Educational Measurement,2006,43(1):53-76

[45] Yao Lihua,Schwarz R D. A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests [J]. Applied Psychological Measurement,2006,30(6):469-492.

[46] Li Deping,Jiang Yanlin,Davier A A. The accuracy and consistency of a series of IRT true score equating [J]. Journal of Educational Measurement Summer,2012,49(2):167-189.

[47] Thorndike R L. Educational measurement [M]. Washington,DC: American Council on Education,1971:508-600.

[48] Kim S,von Davier A A,Haberman S. Equating with small samples [M]. Princeton,NJ: Educational Testing Service,2006

[49] von Davier A A. Statistical models for test equating,scaling,and linking [M]. New York: Springer,2011:89-107.

[50] Michela B. IRT test equating in complex linkage plans [J]. Psychometrika,2013,78(3):464-480.

The Comparison of Equating Methods in Non-Equivalent Group with Anchor Test Design Based on Item Response Theory and Its Application

LI Guangming,WANG Xiaoting

(School of Psychology,Center for Studies of Psychological Application,South China Normal University,Guangzhou Guangdong 510631,China)

Abstract: Two kinds of methods in test equating has been commented: concurrent calibration method and separate calibration method. The second kind includes linking separate calibration methods(e. g. the moment methods and the characteristic curve methods) and FIPC(Fixed Item Parameter Calibration) method. Taking equating accuracy as the criterion,the effects and suitable conditions of each method are summarized and corresponding comments are provided. The reference for users will be prouided in selecting the appropriate methods to process test equating.

Key words: IRT; test equating; non-equivalent groups with anchor test

(责任编辑: 冉小晓)