心理学探新 2023,Vol.43,No.1,68-76
PSYCHOLOGICAL EXPLORATION

不同基准年级下锚题难度范围与年级离散程度对垂直量尺化的影响*

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

摘 要:使用3PLM和蒙特卡洛法生成数据,基于非等组锚题设计,考察不同基准年级下锚题难度范围与年级离散程度对垂直量尺化的影响。结果发现:(1)基准年级的选择会影响垂直量尺化的精度。(2)锚题设计下垂直量尺化的转换不宜超过两个年级。(3)不同基准年级下,年级离散程度越小,估计精度越好。(4)不同基准年级下,对锚题难度范围的选择应有所不同。(5)年级离散程度与锚题难度范围之间存在交互效应。

关键词:垂直量尺化;基准年级;锚题难度范围;年级离散程度;测验等值
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2023)01-0068-09

1 引言

垂直量尺化(vertical scaling),又名垂直量表化、垂直等值(vertical equating),是指在某个特质领域内,在纵向发展的不同水平(如年龄、年级)群体之间,建立关于群体或者个体特质水平发展状况的评价参照体系的过程(漆书青,戴海崎,1992)。垂直量尺化广泛应用于TIMSS、PISA等国际大型测验(罗照盛,2012;Kolen & Brennan,2013)。当不同测验之间的难度水平相差较大且受测者的能力水平存在一定差距时,不再满足测验等值(testing equating)中的水平等值(horizontal equating)条件,宜使用垂直量尺化,其能够有效评价和预估个体或群体动态发展水平和趋势,进而为教育发展与评估提供相关建议。

垂直量尺化主要是用来描述某一群体的纵向发展水平,而水平等值需要建立各个平行测验之间的确切关系,相较而言,垂直量尺化的流程无需进行最后的测验等值步骤,却需将非平行测验转换到同一量尺(叶昶成,2015)。对于能力不同层次的测验群体,比如小学1~4年级学生,必须选定一个年级作为分数转化的基准,将其他年级的分数转化到该年级上形成一个统一分数量尺,一般称为垂直量尺(vertical scale)或发展性量尺(developmental scale)。由于分数转化的算法是一个逐步叠加的过程,比如从3年级转化到2年级再转化到作为基准的1年级,所以当前年级距离基准年级越远,转化的次数越多,受到等值方法误差的影响可能就越大,垂直量尺化精度可能就越低(郭小军,2014;梁正妍,2017)。因此,在实践中通常选用处于中间的年级作为基准来减少误差。

在构建垂直量尺的最初阶段,需要选定相应模型拟合被试的真实作答情况,针对二级计分的测验,通常会选择逻辑斯蒂克模型(Petersen et al.,1989)。郭小军(2014)采用两参数logistic模型(2PLM)进行数据模拟,发现基准年级的选取和年级离散程度对垂直量尺化效果产生较大影响。梁正妍(2017)同样采用2PLM,探究年级离散程度与锚题比例对于垂直量尺化的影响,发现两者存在交互作用。

近年来,国内外对于垂直量尺化的研究不断深入,尝试从多角度、多方面来研究影响垂直量尺化精度的诸多因素,如基于不同模型、不同等值设计、不同题型种类、不同题目难度、是否包含题组、不同测验规模大小、不同被试群体差异、不同参数估计方法等(陈丽,2014;Sari & Kelecioğlu,2016;Carlson,2017)。在垂直量尺化设计中,锚题设计因其实际可操作性而应用最为广泛(王烨晖,边玉芳,2010)。基于这种设计,有学者探讨了基准年级、年级离散程度对垂直量尺化的影响,发现基准年级的合理选取

* **基金项目**:广东省自然科学基金面上项目(2021A1515012516),广东省普通高校特色创新类项目(哲学社会科学)(粤教科函[2021]7号,2021WTSCX020)。
通讯作者:黎光明,E-mail:Lgm2004100@sina.com。

可以有效降低垂直量尺化的误差，同时，其与年级离散程度存在密切关系（郭小军，2014）。梁正妍（2017）对垂直量尺化中不同年级离散程度下锚题比例的选取进行了深入研究，发现锚题比例与年级离散程度有显著的交互作用。

锚题的代表性对测验等值的影响一直以来受到众多研究者关注，其中典型的有锚题难度范围的代表性（叶萌，辛涛，2015）。叶萌和辛涛（2015）对垂直量尺化中锚题代表性问题进行了详细的阐述，提出了锚题难度范围的设定会影响垂直量尺化的精度，其将垂直量尺化中锚题难度范围设置为三种水平，结果发现不同锚题难度范围对垂直量尺化性能和参数返真都有影响，但其没有在不同年级离散程度的群体中进行探讨，未能发现年级离散程度和基准年级的设置对锚题难度范围选取的影响。在锚题设计中，如何选取难度合适的题目构成锚题是实践中的关键问题和难点。由于锚题处于低年级测验的结尾和高年级测验的开始位置，如果锚题选取不当，则会出现项目参数漂移（item parameter drift），即同样的题目在两个位置发挥不同的作用，从而降低垂直量尺化的精度（Wells et al.，2002）。过往的研究和实践中锚题是从低年级测验中随机抽取的，不能保证难度的代表性。对于如何设置锚题难度范围这一问题，Sinharay 和 Holland（2006，2007）研究发现，在题目难度和测验特征的关系中没有表明微型锚测验（锚题与总测验难度范围相匹配）是理想的锚测验，其设置了三种难度范围的锚测验，分别是微型锚测验、midi 锚测验（在内容上对总测验具有代表性，但只包括中等难度的题目）以及半 midi 测验（难度范围小于微型测验，但大于 midi 测验），结果显示 midi 锚测验和总测验的相关稳定性高于微型锚测验和总测验的相关稳定性，midi 锚的性能和微型锚的性能一样，后续的研究也验证了这一结论（Liu et al.，2011）。Chin 等（2006）在垂直量尺化中将锚题难度范围设定为小中大三个等级，结果发现不同锚题难度范围对垂直量尺化性能和参数返真都有影响，难度范围扩大会使这两种分析结果更准确。可见，在实际应用中设置锚题难度范围的标准是十分重要的。

前人对于垂直量尺化影响因素的研究较为深入，分别从被试数量、题目数量、年级数量、基准年级、年级离散程度、锚题比例、难度范围等方面对垂直量尺化的影响进行了较为深入的探讨。但是，前人的相关研究仍然存在以下问题：一是多采用两参数 logistic 模型（2PLM）来估计项目参数和能力参数，未能估计猜测参数 c，实际上，相比 2PLM，三参数 logistic 模型（3PLM）更加适用于可猜测作答的选择题等客观题型（戴海琦，张峰，2018），使用范围增大，可能将提高垂直量尺化的精确性；二是对于锚题的代表性研究，未能同时关注“锚题难度范围”和“基准年级选取”，缺乏探讨对于不同基准年级下锚题难度范围与年级离散程度对垂直量尺化的影响，这对于锚题设计下的垂直量尺化研究来说，是可以深入分析的方向。

基于此，本文对锚题的选取提出了更高的要求，以不同锚题难度范围和不同基准年级“联合作用”为突破点，使用 3PLM，探讨不同基准年级下锚题难度范围与年级离散程度对垂直量尺化的影响。

2　方法

2.1　研究设计

采用 2 × 3 × 3 三因素随机实验设计，自变量 1 为基准年级（边缘年级，中间年级）；自变量 2 为锚题难度范围（较小[μ_{low}，μ_{high}]、中等[$\mu_{low}-\sigma$，$\mu_{high}+\sigma$]、较大[$\mu_{low}-2\sigma$，$\mu_{high}+2\sigma$]）；自变量 3 为年级离散程度（相邻年级间效应 ES 大小：0.5、1.0、1.5）。因变量为等值精度指标 Bias 和 RMSE（Briggs & Peck，2015；Briggs & Dadey，2015）。

（1）基准年级。对于基准年级的选择一般有两种，即边缘年级（低年级或高年级）和中间年级。本文设定了四个年级，对于基准年级，边缘年级为 1 年级，中间年级为 2 年级。采用非等组锚题设计，锚题为相邻年级共用的题目。

（2）锚题难度范围。依据 Chin 等（2006）选取标准，分别选取锚题难度范围较小（两个相邻年级能力均值之间），锚题难度范围中等（低于低年级群体能力均值一个标准差和高于高年级群体能力均值一个标准差之间），锚题难度范围较大（低于低年级群体能力均值两个标准差和高于高年级群体能力均值两个标准差之间），作为锚题难度范围的指标。

（3）年级离散程度。垂直量尺化的结果一般从三个角度进行评价，即跨年级增长（grade - to - grade growth）、跨年级变异（grade - to - grade variability），以及年级间的离散程度（separation of grade distribution）。其中，年级间的离散程度应用最为广泛，是指两个相邻年级的量尺分数分布的重叠程度，俗称为“年级离散程度”。多数研究使用效应大小

(Effect Size,ES)这一统计量来表示年级离散程度(Yen,1986),其计算公式为:

$$ES = \frac{\hat{\mu}(Y)_{upper} - \hat{\mu}(Y)_{lower}}{\sqrt{\frac{\hat{\sigma}^2(Y)_{upper} + \hat{\sigma}^2(Y)_{lower}}{2}}} \quad (1)$$

在公式(1)中,$\hat{\mu}(Y)_{upper}$、$\hat{\sigma}^2(Y)_{upper}$表示高年级能力水平的均值和方差,$\hat{\mu}(Y)_{lower}$、$\hat{\sigma}^2(Y)_{lower}$表示低年级能力水平的均值和方差。随着 ES 的上升,年级间的增长趋势增大。对年级离散程度的选择包含年级离散程度较小(ES = 0.5),年级离散程度中等(ES = 1.0),年级离散程度较大(ES = 1.5)。

(4)固定变量。蔡艳等(2009)通过固定被试数和测验长度,得出当测验长度为 100 时锚题比例最低可达 14.29%。熊建华等(2010)提出当测验长度为 600、300、200、100 题时,相应比例可以降低到 1/15、1/12、1/10、1/5。参考前人研究,本文锚题比例固定为 20%。题目数固定为 100,年级人数固定为 1000。

2.2 模拟流程

分别以低年级和中间年级作为参照基准,使用自编 R3.0 程序,基于蒙特卡洛模拟法,采用三参数 logistic 模型获得四个不同年级组被试在本年级上的作答矩阵。模拟四个年级上各 100 道题目的项目参数以及各年级 1000 名被试的能力参数。使用 BILOG - MG 软件进行同时估计(Yildirim,2014),计算不同锚题难度范围以及不同年级离散程度下 4 个年级的偏差 Bias 和返真性参数 RMSE。

以低年级为基准年级为例,具体模拟过程见图 1。

2.3 评价指标

常用的垂直量尺化评价指标为 Bias 和 RMSE。

(1)Bias,即平均偏差,是考察真值与估计值之间偏差的一个指标,其主要用于检测研究中是否含有系统误差,以及偏差的方向性问题。Bias 值为正,代表低估,Bias 值为负,代表高估。

$$Bias = \frac{\sum_{i=1}^{n}\sum_{j=1}^{R}(\tau_{ij} - \hat{\tau}_{ij})}{n \times R} \quad (2)$$

(2)RMSE,即均方根误差(Root Mean Square Error),是真值与观测值偏差的平方和观测次数 n 比值的平方根。均方根误差对一组测量中的特大或特小误差反映非常敏感,所以能够很好地反映出估计的精度。RMSE 是对一组测量数据可靠性的估计。RMSE 越小,测量的可靠性越大,估计精度就越高。

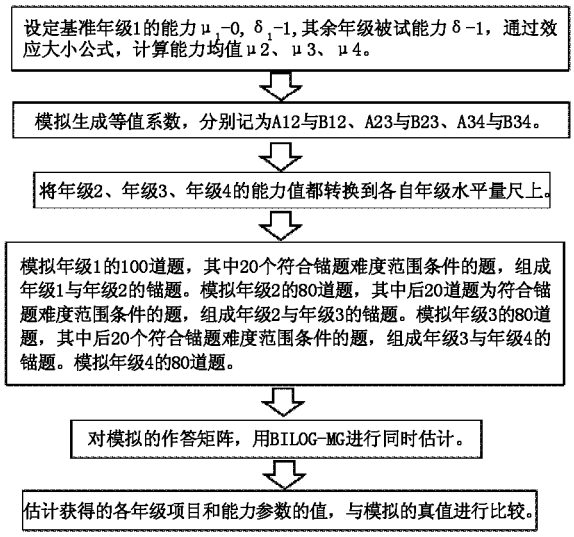

图 1 基准年级为低年级时模拟流程图

$$RMSE = \frac{1}{R}\sum_{j=1}^{R}\sqrt{\frac{1}{R}\sum_{i=1}^{n}(\tau_{ij} - \hat{\tau}_{ij})^2} \quad (3)$$

在公式(2)和公式(3)中,i 表示试题,j 表示被试,n 表示试题数量,R 表示模拟次数,$\hat{\tau}_{ij}$ 表示估计值,τ_{ij} 表示真值。

3 结果

3.1 各条件下 Bias 结果

图 2 和图 3 为各条件下项目参数、被试能力参数 Bias 折线图。

由图 2a ~ 图 2d 可知,基准年级为低年级时,对于各年级项目参数而言,会出现普遍高估的情况,对于被试能力参数来说,会出现普遍低估的情况。随着与基准年级距离的逐渐增大,各参数的估计精度也逐渐下降,且在年级 4 上的表现尤为明显。究其因,是由于锚题设计下垂直量尺化通过等值公式进行累加转换,离基准年级越远,转换的次数就越多,其估计的误差就越大。由此可知,锚题设计下垂直量尺化的转换不宜超过两个年级,此结果与郭小军(2014)的研究结果相似。

由图 3a ~ 图 3d 可知,基准年级为中间年级时,对于区分度参数 a 和被试能力参数 θ 而言,其 Bias 值时正时负,说明对参数的估计会出现忽高忽低的情况。对于难度参数 b 和猜测度参数 c 则倾向于高估,这种情况可能与基准年级的改变有关。与此同时,以中间年级为基准年级时,除区分度参数外,对其他参数的估计,年级 1 产生的 Bias 值始终大于年级 3。这两个年级的锚题均从年级 2 上选取,在垂直量尺化过程中的转换次数也相同,唯一区别在于:对于年级 1 来说,其锚题是从比自身高的年级上选

取的，对于年级 3 来说，其锚题是从比自身低的年级上选取的。因此，根据 Bias 结果，这表明在垂直量尺化中，从高年级选取锚题会比从低年级选取锚题产生更大误差。

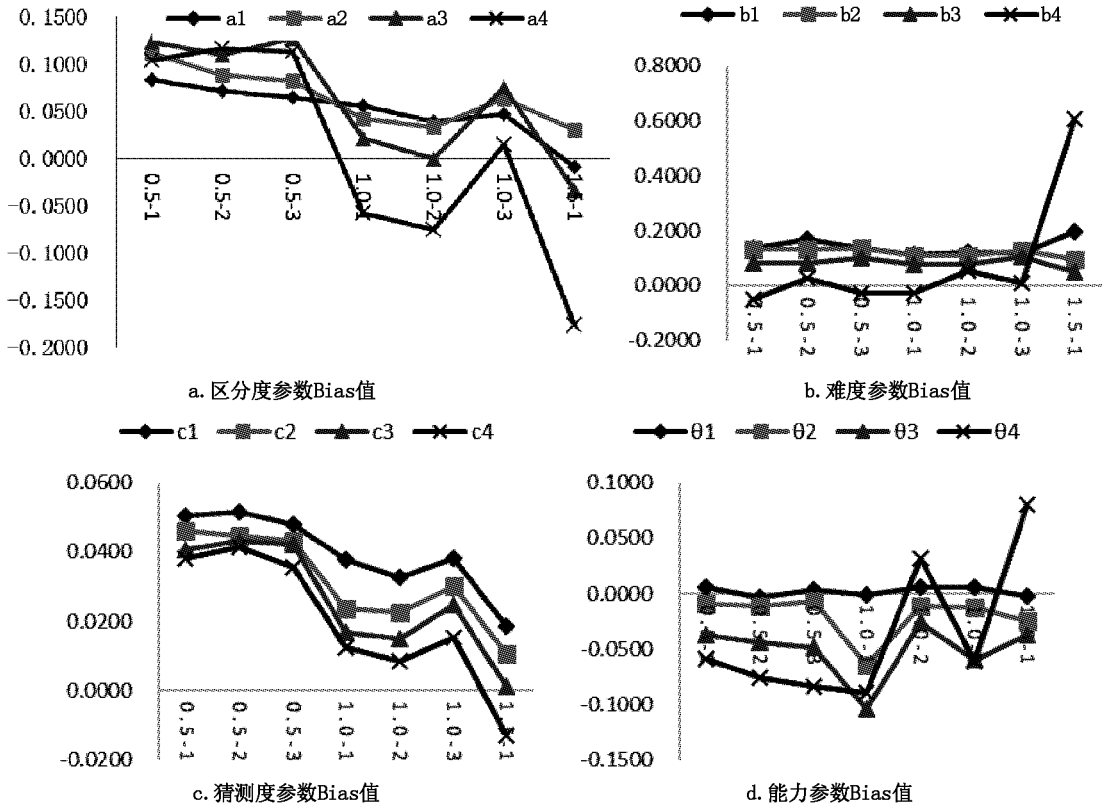

图 2　基准年级为低年级时不同条件下各年级项目与能力参数 Bias 值折线图

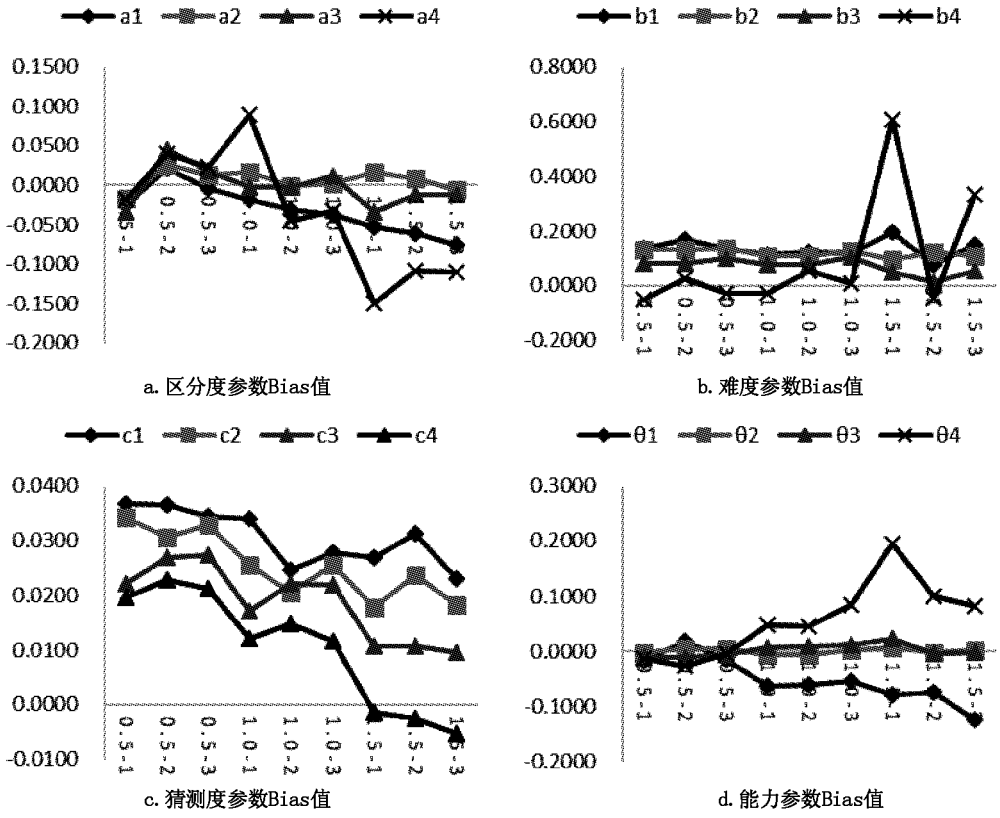

图 3　基准年级为中间年级时不同条件下各年级项目与能力参数 Bias 值折线图

综合图 2 和图 3,对比发现,基准年级为中间年级时,各参数的 Bias 的绝对值明显小于基准年级为低年级时,说明以中间年级为基准进行的垂直量尺化,将会产生更小的估计误差。

3.2 各条件下 RMSE 结果

图 4 和图 5 为各条件下项目参数、被试能力参数 RMSE 折线图。

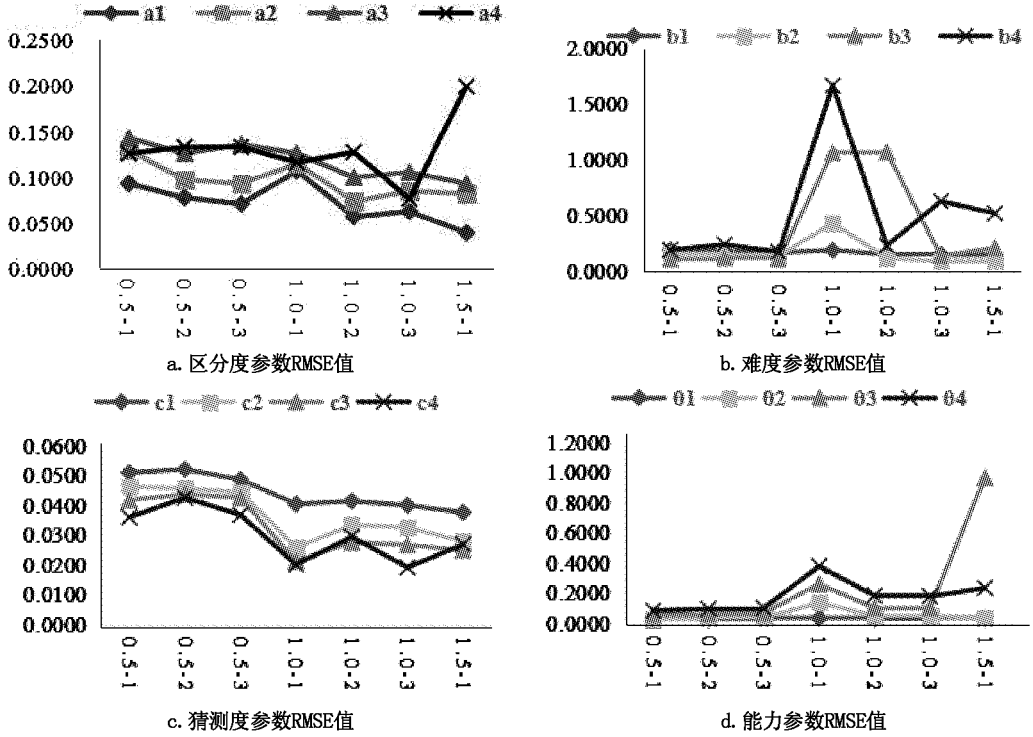

图 4 基准年级为低年级时不同条件下各年级项目与能力参数 RMSE 值折线图

基准年级为低年级时,对于区分度参数 a(图 4a),年级离散程度为 0.5 和 1.0 时参数估计精度差别不大,且均好于年级离散程度为 1.5。对于猜测度参数 c(图 4c),三种年级离散程度下的参数估计精度相差不大。然而,对于难度参数 b(图 4b)和被试能力参数 θ(图 4d),均当离散程度为 0.5 时,估计精度最好;年级离散程度为 1.0 和 1.5 时,在个别情况下均存在较大误差。因此,综合看来,在基准年级为低年级时,对于所有参数,年级离散程度为 0.5 时估计精度最好,年级离散程度为 1.0 时次之,年级离散程度为 1.5 时最差。

在基准年级为低年级时,对于区分度参数 a(图 4a),中等或较大锚题难度范围下的估计精度较好,较小难度范围下估计精度最差。对于难度参数 b(图 4b),较大锚题难度范围下的估计精度较好,中等范围次之,较小难度范围下估计精度最差。对于猜测度参数 c(图 4c),较小或较大难度范围下的估计精度较好,中等范围次之,但总体来说三种锚题难度范围下的猜测度参数估计结果相差不大。对于被试能力参数 θ(图 4d),中等或较大难度范围下的估计精度较好,较小范围次之。因此,综合看来,在基准年级为低年级时,较大难度范围下的参数估计精度最佳,参数返真性最好,中等范围次之,较小范围最差。

在基准年级为低年级时,对于区分度参数 a(图 4a),年级离散程度为 0.5 时,对于年级 1、2,较大锚题难度范围下的结果最佳,对于年级 3,中等范围最好,对于年级 4,较小范围最好。年级离散程度为 1.0 时,对于年级 1、2、3,中等锚题难度范围下的效果最好,对年级 4 较大范围最好。年级离散程度为 1.5 时,只有较小锚题难度范围下的结果收敛。对于难度参数 b(图 4b),年级离散程度为 0.5 时,在各年级上,三种锚题难度范围下的结果差别不大。年级离散程度为 1.0 时,年级 1、2、3 在较大锚题难度范围下表现最好,年级 4 在较小锚题难度范围下表现最好。年级离散程度为 1.5 时,只有较小难度范围下的结果收敛。对于猜测度参数 c(图 4c),在各年级离散程度与锚题难度范围下,4 个年级结果差别不大。对于被试能力参数 θ(图 4d),年级离散程度为 0.5 时,三种锚题难度范围下被试表现差别不大。年级离散程度为 1.0 时,在中等和较大锚题难度范围下,4 个年级表现均较好,较小方范围产生的误差最大。年级离散程度为 1.5 时,只有较小锚题难度范围下的结果收敛。

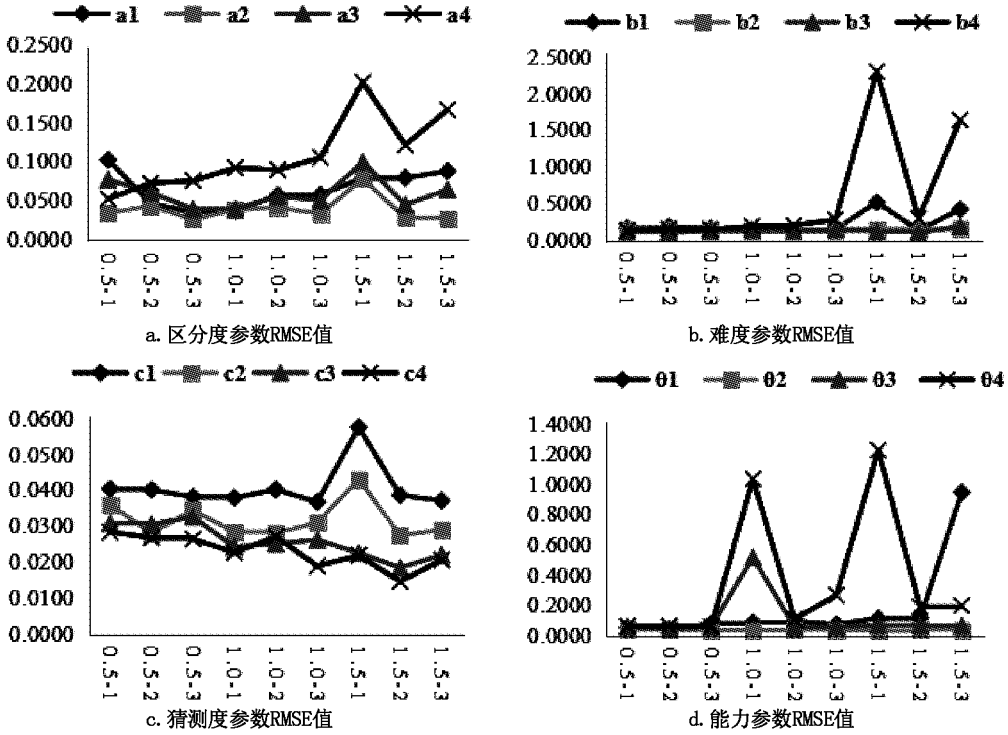

图5　基准年级为中间年级时不同条件下各年级项目与能力参数RMSE值折线图

基准年级为中间年级时，对于区分度参数a（图5a），当年级离散程度为0.5时，估计精度最好；当年级离散程度为1.0时次之，而当年级离散程度为1.5时，结果不够稳定，在各个年级上起伏较大。对于难度参数b（图5b），当年级离散程度为0.5和1.0时，估计精度均较好；而当年级离散程度为1.5时，结果不够稳定，在各个年级上起伏较大。对于猜测度参数c（图5c），年级离散程度为0.5和1.0时参数估计精度差别不大，且均好于年级离散程度为1.5时参数估计的精度。对于被试能力参数θ（图5d），当年级离散程度为0.5时，估计精度最好；当年级离散程度为1.0时个别条件下误差较大；当年级离散程度为1.5时，估计精度不如其他两种情况。因此，综合看来，在基准年级为中间年级时，年级离散程度为0.5时估计精度最好，年级离散程度为1.0时次之，年级离散程度为1.5时最差。

在基准年级为中间年级时，对于区分度参数a（图5a），中等或较大难度范围下估计精度较好，较小锚题难度范围下估计的误差不稳定。对于难度参数b（图5b），中等难度范围下估计精度较好，较大范围次之，较小难度范围下估计精度最差。对于猜测度参数c（图5c），中等及较大难度范围下估计精度较好，较小范围次之。对于被试能力参数θ（图5d），中等难度范围下的估计精度较好，较小或较大难度范围下的估计结果均在个别条件下存在较大误差。因此，综合看来，在基准年级为中间年级时，中等难度范围下的参数估计精度最佳，参数返真性最好，较大范围次之，较小范围最差。

在基准年级为中间年级时，对于区分度参数a（图5a），年级离散程度为0.5时，对于年级1、2、3，较大锚题难度范围下的结果最佳，对于年级4，较小范围最好。在年级离散程度为1.0时，对于各年级，较小难度范围下的效果最好，但中等难度范围下的效果与之差别不大。在年级离散程度为1.5时，对于各年级，中等难度范围下的效果最好。对于难度参数b（图5b），年级离散程度为0.5和1.0时，各年级在三种的估计结果差别不大。在年级离散程度为1.5时，对于各年级，中等难度范围下的结果最好。对于猜测度参数c（图5c），年级离散程度为0.5和1.0时，对于各年级，三种锚题难度范围下的误差均较小。在年级离散程度为1.5时，对于各年级，中等或较大锚题难度范围下的结果均较好。对于被试能力参数θ（图5d），年级离散程度为0.5时，各年级在三种锚题难度范围下的被试表现差别不大。在年级离散程度为1.0和1.5时，各年级在中等锚题难度范围下的表现最好。

综合图4和图5，对比发现，当基准年级为中间年级时，RMSE整体误差小于基准年级为低年级时产生的误差。

4　分析与讨论

4.1　不同基准年级对垂直量尺化参数估计的影响

当基准年级为中间年级时，Bias和RMSE的整

体误差均小于基准年级为低年级时产生的误差。这表明,基准年级的选择会影响垂直量尺化的结果。锚题设计下的垂直量尺化过程是一个累加转换的过程,换言之,由于只有相邻年级间存在锚题,因而与基准年级较远的年级需要经过层层转换,最终转换到基准年级上,而多次的转换势必造成偏差的累加,正因为存在这种“累积效应”,所以通常需要选择中间年级作为基准年级(Yen et al.,2012)。

当基准年级为中间年级时,到高年级和低年级的垂直量尺转化是方便的。如果选择以低年级或高年级作为基准年级,则有可能加大垂直量尺转化难度,显得“路途遥远”,而以基准年级选择为中间年级,显得更为方便。因此,在教育与心理实践中,为了更加关注学生的连续发展和学业上的进步模式,或为了更加关注学生在不同学年的进步表现,年级与年级之间的垂直量尺转化选择以“中间模式”较好,以防止垂直量尺化过程中产生更大的误差。

4.2 与基准年级的间隔对垂直量尺化参数估计的影响

整体而言,当年级离散程度为 0.5 时,估计精度最好,且随着年级离散程度的增大,估计精度随之下降。当年级离散程度为 1.5 时,估计精度极不稳定,甚至出现了数据无法收敛的情况。特别地,即使以中间年级作为基准年级,虽然没有出现类似以低年级为基准年级时无法收敛的情形,但在年级 4 上也出现了一些极不稳定的 Bias 和 RMSE 值,特别是在难度参数和能力参数上,形成若干 Bias 和 RMSE“极端值”。

基于非等组锚题设计,以低年级为基准年级时,在 Bias 和 RMSE 上,出现较多的无法收敛的情况,其原因是由于年级离散程度过大,年级与年级之间的“异质性”增加,不同年级学生的“能力差异”就会不断放大,造成“分数偏差”增加,最终导致难以相互转化(Ye & Xin,2014)。同样地,以中间年级为基准年级时,虽然情况有一定的好转,但也出现了较多的偏差“极值”。如果转换超过两个年级,那么垂直量尺化精度开始下降。特别地,在年级 4 上出现了若干 Bias 和 RMSE“极值”。基于此,在教育与心理实践中,如追踪监测学生的学业进展,进行垂直量尺化时,建议与基准年级间隔不宜超过 2 个年级。

4.3 不同年级离散程度下的参数估计分析

年级离散程度为 0.5 时,垂直量尺化精度最好,年级离散程度为 1.0 时次之,年级离散程度为 1.5 时,垂直量尺化精度最差。在 3PLM 下,不同基准年级下,年级离散程度越小,估计精度越好,这与前人使用 2PLM 得出的结果是一致的(梁正妍,2017;郭小军,2014)。年级离散程度越大,对于距离基准年级越远的年级,垂直量尺化精度越低。因此,在实际的教育与心理测量实践中,年级离散程度不宜过大,否则会影响垂直量尺化的精度。

实际上,在使用项目反应理论方法进行垂直量尺化时,需要满足两大潜在假设:一是年级内的测验单维性(Unidimensionality);二是年级间测验同构性(Construct invariance)(Li,2011;Martineau,2004)。单维性是指每道题目只测量单一的潜在能力特质;同构性是指不同年级水平或难度水平的测验维持相同的内容结构(Li & Lissitz,2012;Reckase & Martineau,2004)。在垂直量尺化实践中,普遍涉及到 3~6 个年级的多组学生和试题。在试题层面,所考察的知识点跨度较大,因此跨年级同构性假设一般较难满足。Martineau(2006)最初用结构漂移(Construct shift)一词来描述违背同构性假设的现象。以数学测验为例,在进行垂直量尺化时,3 年级题目和 6 年级题目虽然都是考察数学知识,但是具体考察的知识点可能是很不一样的。

4.4 不同锚题难度范围下的参数估计分析

随着与基准年级距离的逐渐增大,各参数的估计精度也逐渐下降,在年级 4 上表现尤为明显,说明锚题设计下的累加转换确实会降低估计精度。当基准年级为低年级时,为了提高垂直量尺化的估计精度,就需要有较大的锚题难度范围,这是因为锚题难度范围较大,低一年级学生的题目与上一年级学生的题目重叠可能更多,那么这对于低一年级的学生而言,这是有利的,从而使得垂直量尺化更为顺利。实际上,作为低一级年级的学生可能缺乏能力来完成上一年级学生的题目,但若锚题难度范围不断变大,则较有可能完成上一年级学生的题目(Lao,2015)。但是,当基准年级为中间年级时,因为其要向两边扩充,既要“冲上”(向年级 3 和年级 4 转化),又要“冲下”(向年级 1 转化),所以难度范围的设置不能过于“宽”,也不能过于“窄”,中等的锚题难度范围则相对更好。

由此可见,在不同的基准年级下进行垂直量尺化,对于锚题难度范围的选取要求不同。当基准年级为低年级时,建议选取较大的锚题难度范围;当基准年级为中间年级时,建议选取中等的锚题难度范围。但是,从高年级选取锚题会比从低年级选取锚题产生更大误差,这是因为高级学生做低年级学生题目是相对容易的,但低年级学生做高年级学生题目则是相对困难的。基于此,在教育与心理实践中,为了比较不同年级学生学业能力的发展轨迹,设置有利于低年级学生的锚题难度范围,对垂直量尺化

可能更为受益。

4.5　不同基准年级下年级离散程度与锚题难度范围的交互效应

以低年级作为基准年级，在年级离散程度为0.5时，年级1、2更适合锚题难度较大范围，年级3更适合锚题难度中等范围，年级4更适合锚题难度较小范围。在年级离散程度为1.0时，年级1、2、3选择锚题难度中等或较大范围均可，年级4的表现则非常不稳定，无法选取最佳锚题难度范围。在年级离散程度为1.5时，只有选取锚题难度较小范围，才能进行可靠的垂直量尺化。若需要4个年级统一选择相同的锚题难度范围，则建议：当基准年级为低年级时，年级离散程度为0.5时，选择锚题难度较大范围；当年级离散程度为1.0时，选择锚题难度中等或较大范围；年级离散程度为1.5时，选择锚题难度较小范围。

以中间年级作为基准年级，在年级离散程度为0.5时，年级1、2、3更适合锚题难度较大范围，年级4更适合锚题难度较小范围。在年级离散程度为1.0时，各年级选择锚题难度较小或中等范围均可。在年级离散程度为1.5时，各年级更适合锚题难度中等范围。若需要4个年级统一选择相同的锚题难度范围，则建议：当基准年级为中间年级时，年级离散程度为0.5时，选择锚题难度较大范围；当年级离散程度为1.0时，选择锚题难度较小或中等范围；当年级离散程度为1.5时，选择锚题难度中等范围。

综上所述，基于不同基准年级，年级离散程度与锚题难度范围存在交互效应（见表1）。

表1　不同基准年级下年级离散程度与锚题难度范围的交互效应

基准年级	年级离散程度	锚题难度范围选择
低年级	ES=0.5	较大范围
	ES=1.0	中等或较大范围
	ES=1.5	较小范围
中间年级	ES=0.5	较大范围
	ES=1.0	较小或中等范围
	ES=1.5	中等范围

5　结论

（1）基准年级的选择会影响垂直量尺化的精度。选择以中间年级为基准进行垂直量尺化，将会使得垂直量尺化的结果保持在一个较好的精度。

（2）锚题设计下垂直量尺化的转换不宜超过两个年级。如果转换超过两个年级，那么垂直量尺化精度开始下降。特别地，在年级4上出现了若干Bias和RMSE“极值”，建议与基准年级间隔不宜超过2个年级。

（3）不同基准年级下，年级离散程度越小，估计精度越好。年级离散程度为0.5时，垂直量尺化精度最好，年级离散程度为1.0时次之，年级离散程度为1.5时，精度最差。

（4）不同基准年级下，对锚题难度范围的选择应有所不同。当基准年级为低年级时，锚题难度较大范围时垂直量尺化精度最好。当基准年级为中间年级时，锚题难度中等范围时垂直量尺化精度最好。

（5）年级离散程度与锚题难度范围之间存在交互效应。在不同基准年级与不同年级离散程度下，对于锚题难度范围的选择应该有所不同。

参考文献

蔡艳，丁树良，涂冬波.（2009）.锚题比例对等值精度的影响.*心理学探新*，*29*（2），56－59.

陈丽.（2014）.垂直量尺化对大学英语分级教学测评体系弊端的解析.*西安外国语大学学报*，*22*（2），76－78.

戴海崎，张锋.（2018）.*心理与教育测量*（第四版）.广州：暨南大学出版社.

郭小军.（2014）.*不同参照基准与年级离散程度对垂直等值的影响研究*（硕士毕业论文）.江西师范大学，南昌.

罗照盛.（2012）.*项目反应理论基础*. 北京师范大学出版社.

梁正妍.（2017）.*年级离散程度与锚题比例对垂直量尺化精度的影响*（硕士学位论文）.华南师范大学，广州.

漆书青，戴海崎.（1992）.*项目反应理论及其应用研究*. 南昌：江西高校出版社.

王烨晖，边玉芳.（2010）.构建学业发展性量表－垂直等值的应用.*中国考试*，*22*（10），7－12.

熊建华，叶新蓉，丁树良，罗芬.（2010）.*等值设计中锚题比例研究*. Proceedings of 2010 Third International Conference on Education Technology and Training（Volume 7）.

叶萌，辛涛.（2015）.测验链接中的锚题代表性研究.*心理科学*，*38*（1），209－215.

叶昶成.（2015）.*不同垂直等化设计下可能值方法估计效果值探讨*（硕士学位论文）.台中教育大学.

Briggs，D. C.，& Dadey，N.（2015）. Making sense of common test items that do not get easier over time：Implications for vertical scale designs. *Educational Assessment*，*20*（1），1－22.

Briggs，D. C.，& Peck，F. A.（2015）. Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement Interdisciplinary Research & Perspectives*，*13*（2），75－99.

Carlson，J. E.（2017）. Unidimensional vertical scaling in multidimensional space. *Ets Research Report*，（4）.

Chin，T. Y.，Kim，W.，& Nering，M. L.（2006）. *Five statistical factors that influence IRT vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education，San Francisco，CA.

Kolen，M. J.，& Brennan，R. L.（2013）. *Test equating scaling*

and linking——method and practices (3rd ed.). Springer - Verlag New York Inc.

Lao, H. (2015). Some thoughts on using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement Interdisciplinary Research & Perspectives*, *13*(3), 195 - 199.

Li, Y. (2011). *Exploring the full - information bifactor model in vertical scaling with construct shift* (Unpublished doctoral dissertation). University of Maryland, College Park, MD.

Li, Y., & Lissitz, R. W. (2012). Exploring the full - information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, *36*, 3 - 20.

Liu, J. S., Sinharay, S., Holland, P., Feigenbaum, M., & Curley, E. (2011). Observed score equating using a mini - version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement*, *71*(2), 346 - 361.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models*. ProQuest Information & Learning. Michigan State University, East Lansing, MI.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth - based, value - added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35 - 62.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221 - 262). Washington, DC.

Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Paper commissioned by the committee on test design for K - 12 Science achievement, center for education, national research council.

Sari, A. A., & Keleciǧlu, H. (2016). Assessment of achievement and growth by vertical scaling: Comparison of vertical scaling methods. *Journal of Educational Sciences Research*, *6*(2), 25 - 38.

Sinharay, S., & Holland, P. W. (2006). *The correlation between the scores of a test and an anchor test* (ETS RR - 06 - 04). Princeton, NJ: Educational Testing Service.

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.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*(1), 77 - 87.

Ye, M., & Xin, T. (2014). Effects of item parameter drift on vertical scaling with the nonequivalent groups with anchor test (neat) design. *Educational & Psychological Measurement*, *74*(2), 227 - 235.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299 - 325.

Yen, W. M., Lall, V. F., & Lora, M. (2012). Evaluating academic progress without a vertical scale. *Ets Research Report*, (1), 1 - 55.

Yildirim, H. H. (2014). *Findings from an empirical vertical scaling study with BILOG - MG*. Education & Science.

The Influence of Difficulty Range of Anchor Items and Separation of Grade Distributions on Vertical Scaling Under Different Base Grades

Li Guangming Zhang Xiaoting

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

Abstract: In this study, we used 3PLM under common - item design, and set grade 1 and grade 2 as the base grade respectively. Setting the item parameters of 100 items of four grades and the ability parameters of 1000 subjects in each grade in different combinations of difficulty range of anchor items and the separation of grade distributions, and then simulated the response matrix by Monte Carlo method using BILOG - MG for concurrent calibration. Bias and RMSE values are calculated as accuracy criteria. This study shows that: (1) the choice of the base group affects the results of the vertical scaling. (2) The conversion of vertical scaling under anchor question design should not exceed two grades. (3) Under different base grades, the smaller the separation of grade distribution is, the better the estimation accuracy is. (4) When the base grade is the lower grade, the difficulty range of the anchor items should be wide. When the base grade is middle grade, the difficulty range of the anchor items should be medium to get a better accuracy. (5) There is an interaction between the separation of grade distributions and the difficulty range of the anchor items.

Key words: vertical scaling; base grade; difficulty range of anchor items; separation of grade distributions; test equating