文章编号:1000-5862(2019)01-0052-07

锚题比例与年级离散度对垂直等值的影响

黎光明,梁正妍
(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:采用锚题设计,选择年级 1 作为参照基准,通过蒙特卡洛模拟方法,考察锚题比例与年级离散度对垂直等值的影响. 研究结果表明:与基准年级的距离影响垂直等值效果,越靠近基准年级,估计精度越好;从整体而言,垂直等值锚题比例设为 30%,等值效果最好;垂直等值锚题比例的设定受年级离散度影响,2 者存在交互作用,锚题比例设为"变"值更好.

关键词:锚题比例;年级离散度;垂直等值;测验等值;项目反应理论
中图分类号:B 841 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2019.01.10

0 引言

大型测验经常需要将某一测验配备多个版本,以供多次施测,并最终使参与不同测验版本的被试成绩可以进行比较. 对于测验编制者而言,最希望看到的是不同版本的测验在同一批被试上所得的测验分数是可以相互比较的. 但是,如果不同版本的测验分数之间存在差异,这就会引起评价的不公正. 为了避免这种不公正,一种方法就是找到不同版本测验分数之间的转换关系,将不同版本测验的分数转换到同一量尺上. 在测量学上,这种方法被称为测验等值(Test equating)[1].

通常来说,在 2 种情况下需要进行等值转换:(i)测验难度与被试能力水平大体相同的情况,如将考生在高考 A、B 卷上得分数进行转换,这种等值转换被称为"水平等值"(Horizontal scale);(ii)测验难度与被试能力水平均有较大差异,如想了解不同年级学生的知识水平之间的情况,这种等值转换被称为"垂直等值"(Vertical scaling)[2-3]. 在"水平等值"中,待等值测验的难度近似相同,且学生在他们得分分布假设可比,适用于对一个测验的多个题本进行等值;而在"垂直等值"中,待等值测验的难度不同,且学生在他们得分分布假设不可比,适用于对来自不同水平(比如不同年级)的学生进行分数比较.

一般地,水平等值只能实现针对同一特质且各版本试卷难度相差不大情况下考生测验分数的等值转换,但无法实现不同年级水平学生测验分数的比较,这是因为测验所涉及的知识内容、测验难度和测验对象的能力水平均存在较大差异. 然而,对于学校以及家长们来说,他们有时更希望了解学生在整个学习和成长过程中,随着年级的增长、知识的掌握,其能力究竟是怎样的一个发展趋势[4]. 这对于国家而言,可以根据其客观的发展趋势更好地制定教育改革的方针政策. 垂直等值能解决这一问题. 垂直等值可以将各个不同水平测验分数转换到同一分数量尺上,使其能够相互比较[5-6]. 为了说明学生随着时间变化在学业上的发展或进步程度,可以使用垂直等值技术将各个不同年级水平的测验分数进行链接[7].

为了将各个年级学生的学业水平进行垂直等值,锚题设计(Common item design)是最常用的等值设计. 锚题设计的一个示例如图 1 所示[8-9].

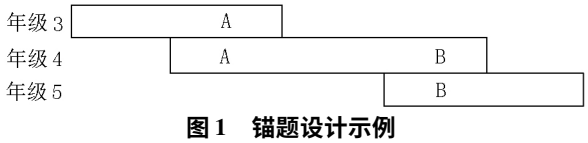


图 1 锚题设计示例

收稿日期:2018-07-13
基金项目:国家自然科学基金(31470050),教育部人文社会科学研究规划基金(18YJA190006),广东省哲学社会科学"十三五"规划一般课题(GD17CXL01),广东省 2015 年度高等教育教学改革课题(粤教高函〔2015〕173 号)和广州市哲学社会科学"十三五"规划一般课题(2017GZYB111)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

在图 1 的锚题设计示例中,年级 4 有部分题与年级 3 相同,有部分题与年级 5 相同,这就是“锚题”(Anchor item) . “锚题”作为等值的关键一环,是整个测验的“微缩版”. 通常,图 1 中联系年级 3 与年级 5 的纽带——年级 4 锚题,若占整个测验的百分比越大,则所求的等值系数越准确. 然而,在实际应用和题库的建设中,锚题使用的越多,会出现以下 2 个问题: (i) 锚题使用得越多,题库的整体题量也就要求越大,从而使得题库建设的成本增加; (ii) 锚题使用得越多,曝光的锚题比例就越多,测验的安全性和保密性就越低. 因此,年级跨度怎样“设定”一个合适范围内的锚题比例,对垂直等值的结果解释是有意义的. 然而,就目前有关垂直等值锚题比例相关的研究文献,还存在以下问题:

(i) 年级之间测验等值的锚题比例到底应该定为多少才是合适的呢? 仍缺乏相关依据. 目前,大多数有关锚题比例的研究文献多集中于“水平等值”文献[3,10] . 例如,蔡艳等[10] 在水平等值中发现,在强调等值系数的返真性时,锚题比例可选择 ≥ 18. 33%; 在强调测验分数系统或能力参数系统的等值精度时,锚题比例可选择 ≥16. 67%,甚至是 ≥ 14. 29% . 通常地,PISA、NAEP 和 TIMSS 等国际大规模教育评估项目的锚题比例至少是 20%,由“锚题”所构成的子测验是有代表性的,几乎是整个测验的“微缩”. 然而,有关年级之间垂直等值锚题比例到底应该定为多少才是合适的呢? 仍缺乏相对成熟的研究文献给出一个基本定论.

(ii) 年级之间垂直等值的锚题比例是否随年级的增长而发生一定的改变? 仍不得而知. 由于相邻年级测验难度与被试能力分布更为相似,测验之间重叠的知识内容较多,2 者之间的链接能够提供更多的信息. 但是,随着年级跨度的增大,测验的难度和被试的能力差异变大,测验间重叠的内容变少,使得年级之间的链接“强度”减小[11-12] . 依此看来,年级之间垂直等值的锚题比例设定可能会随着年级的增长而发生改变,即年级之间垂直等值的锚题比例可能不是一成不变的,可能会与年级离散度一同影响垂直等值. 年级离散度是指不同年级学生能力之间的相异程度,年级离散度越大,不同年级学生能力之间相异越大,反之亦然.

为了解决上述 2 个问题,本文固定以低年级为参照基准,旨在探究不同锚题比例与年级离散度对垂直等值的影响,以期为更好地进行测验的垂直等值提供借鉴和参考.

1　研究方法

1.1　研究设计

采用非等组锚题设计,选择以年级 1 作为参照基准,通过蒙特卡洛模拟方法,模拟不同年级的数据. 每个年级包括 100 道题的项目参数和 1 000 名被试的能力参数,并假定不同年级之间的离散度是一致的,相邻年级的离散度用效应大小(Effect Size, E_S) 来表示[13]

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

其中 $\hat{\mu}(Y)_{upper}$、$\hat{\sigma}^2(Y)_{upper}$ 分别表示高年级能力水平的均值和方差,$\hat{\mu}(Y)_{lower}$、$\hat{\sigma}^2(Y)_{lower}$ 分别表示低年级能力水平的均值和方差. 随着 E_S 的上升,年级间的增长趋势增大.

因为试题均为 0/1 计分,所以模型选用 2 参数逻辑斯蒂克模型(two parameters logistic model) [14] . 参考相关文献[15-17],研究设计设置为 3 × 3,其中第 1 个“3”为自变量 1(锚题比例) ,有 3 个水平: 15%、30% 和 45%; 第 2 个“3”为自变量 2(年级离散度) ,用效应大小表示,有 3 个水平: 0. 5、1. 0 和 1. 5.

1.2　研究过程

首先,使用 R 软件模拟数据,模拟项目参数、能力参数与作答矩阵; 其次,采用 BILOG-MG 对数据作答矩阵进行参数估计,所使用的等值方法为同时标定法; 最后,比较不同锚题比例以及年级离散度对不同年级测验垂直等值返真性指标 b_{ias} 值的影响. 为了模拟数据,需要预先设置各参数的数据分布[17-18],如表 1 所示.

表 1　各参数的数据分布

E_S	a	b_1	b_2	b_3	b_4	θ_1	θ_2	θ_3	θ_4
0. 5	(1,0. 6)	(0,1)	(0. 5,1)	(1,1)	(1. 5,1)	(0,1)	(0. 5,1)	(1,1)	(1. 5,1)
1. 0	(1,0. 6)	(0,1)	(1. 0,1)	(2,1)	(3. 0,1)	(0,1)	(1. 0,1)	(2,1)	(3. 0,1)
1. 5	(1,0. 6)	(0,1)	(1. 5,1)	(3,1)	(4. 5,1)	(0,1)	(1. 5,1)	(3,1)	(4. 5,1)

在表 1 中,区分度、难度和能力值均服从正态分布,其中各参数括号内分别表示平均数和标准差,即 (μ,σ). 下面以锚题比例 30% 为例,来说明研究的基本过程.

第 1 步,设定年级 1 的参数. 年级 1 为基准年级,能力 $\mu_1=0$,$\sigma_1=1$,其余年级被试能力 $\sigma=1$,同时分别设定相邻年级的效应大小为 0.5、1.0、1.5,通过(1)式反推,可以分别计算出年级 2、年级 3、年级 4 的能力均值 μ_2、μ_3、μ_4. 等值系数 α 在 [0.7,1.5] 取值,步长为 0.1;等值系数 β 在 [-0.6,1] 取值,步长为0.2[10]. 随机选择生成的 3 对等值系数,分别为年级 1 ~ 年级 2、年级 2 ~ 年级 3、年级 3 ~ 年级 4 的等值系数,分别记为 α_{12}与β_{12}、α_{23}与β_{23}、α_{34}与β_{34}.

由于最初模拟的各年级能力参数均需与基准年级在同一量尺上,因此需要将年级 2、年级 3、年级 4 的能力值转换到各自年级水平量尺上,3 个年级的转换公式如下:

$$\theta_2=\alpha_{12}\theta_2^*+\beta_{12}, \tag{2}$$

$$\theta_3=(\alpha_{12}\theta_3^*+\beta_{12})\alpha_{23}+\beta_{23}, \tag{3}$$

$$\theta_4=((\alpha_{12}\theta_4^*+\beta_{12})\alpha_{23}+\beta_{23})\alpha_{34}+\beta_{34}, \tag{4}$$

在(2) ~(4)式中,θ 表示转换后的能力值,而 θ^* 表示转换前的能力值.

基准年级 1 的难度参数也采用正态分布,其均值和方差与年级 1 被试能力一致,区分度参数采用对数正态分布,即 $a_{i1}\sim \text{Lognormal}(0,1)$,年级 1 的题量为 100,随机抽取 30 个试题,组成年级 1 与年级 2 的锚题.

第 2 步,设定年级 2 的参数. 年级 2 相对年级 1 的独立测验为 70 题,其难度分布的均值与方差采用与年级 2 能力分布的均值和标准吻合,区分度也采用对数正态分布. 再将年级 1 与年级 2 的 30 道锚题参数以及 70 道独立测验的项目参数,通过类似于(2) ~(4)式的线性关系,转换到年级 2 上,同时从年级 2 的 70 道独立项目中随机抽取 30 道作为年级 2 与年级 3 的锚题.

第 3 步,设定年级 3 的参数. 先模拟出年级 3 的 70 道独立试题,再将这 70 道独立试题的参数与 30 道锚题参数转换到年级 3 上,同时从这 70 道独立试题中随机抽取 30 道作为年级 3 与年级 4 的锚题.

第 4 步,设定年级 4 的参数. 对于年级 4 的参数设定. 只需模拟出 70 道独立试题的参数,再与锚题参数一起转换到年级 4 上即可.

第 5 步,模拟各年级考生的作答矩阵. 根据设定的区分度、难度和能力等参数值,模拟各年级学生的原始分数作答矩阵.

第 6 步,计算估计值. 用 BILOG-MG 软件对模拟的原始分数作答矩阵进行参数估计,在写程序时,$N_{PARM}=2$,$N_{GROUPS}=4$,$M_{ETHOD}=2$,这几条语句是非常重要的. 估计方法有比较多种,根据相关文献[15,19],本文仅使用 EAP 方法,计算的返真性指标为 b_{ias}值.

$$b_{ias}=\hat{\tau}_i-\tau_i, \tag{5}$$

其中 $\hat{\tau}_i$ 为估计值,τ_i 为真值. 以上模拟步骤及过程各重复 100 次.

2 实验结果

2.1 不同年级离散度与锚题比例条件下垂直等值项目和能力参数值

以年级 1 为参照基准,将年级 2、3、4 各项目参数和被试能力参数均转换到年级 1 上. 不同年级离散度与锚题比例条件下垂直等值项目参数和能力参数值结果如表 2 所示.

表 2 不同年级离散度与锚题比例条件下垂直等值项目和能力参数值

	0.5					
	15%		30%		45%	
	M	S_D	M	S_D	M	S_D
a_1	0.932 7	0.597 5	0.968 2	0.611 3	0.948 1	0.594 4
a_2	0.916 0	0.584 9	0.951 0	0.605 4	0.947 8	0.600 4
a_3	0.951 9	0.602 4	0.958 7	0.609 7	0.967 4	0.607 8
a_4	0.957 7	0.620 0	0.940 5	0.604 0	0.964 5	0.607 7
b_1	0.012 4	0.995 6	0.033 7	0.967 5	-0.009 7	0.996 7
b_2	0.453 0	0.987 7	0.368 3	1.0146	0.254 4	0.996 8
b_3	0.980 0	1.025 4	0.879 0	1.003 8	0.738 6	0.992 3
b_4	1.433 0	0.995 0	1.346 1	1.028 3	1.289 9	1.015 7
θ_1	-0.002 1	0.980 2	-0.009 2	0.981 7	-0.017 5	0.987 3
θ_2	0.496 5	0.982 1	0.4918	0.983 8	0.497 6	0.986 4
θ_3	1.003 6	0.979 7	1.005 8	0.989 3	0.992 8	0.982 3
θ_4	1.492 2	0.979 4	1.510 0	0.986 6	1.502 6	0.989 1
	1.0					
	15%		30%		45%	
	M	S_D	M	S_D	M	S_D
a_1	0.936 1	0.599 3	0.968 2	0.611 3	0.942 6	0.592 8
a_2	0.921 0	0.586 7	0.951 0	0.605 4	0.955 7	0.606 5
a_3	0.948 3	0.599 4	0.958 7	0.609 7	0.965 3	0.611 5
a_4	0.954 7	0.617 8	0.940 5	0.604 0	0.961 0	0.608 5
b_1	0.007 3	0.987 1	0.033 7	0.967 5	-0.009 0	0.999 7
b_2	0.876 4	1.033 3	0.718 3	1.084 6	0.524 3	1.091 6
b_3	1.905 4	1.064 5	1.729 0	1.0762	1.518 6	1.083 7
b_4	2.859 3	1.035 5	2.696 1	1.104 5	2.573 3	1.110 2
θ_1	-0.002 7	0.979 9	-0.009 2	0.981 7	-0.019 0	0.987 4
θ_2	0.997 8	0.980 0	0.991 8	0.983 8	0.996 8	0.986 5
θ_3	2.005 5	0.979 6	2.005 8	0.989 3	1.993 0	0.985 9
θ_4	2.992 5	0.979 4	3.010 0	0.986 6	2.999 2	0.987 6

表 2(续)

	1.5					
	15%		30%		45%	
	M	S_D	M	S_D	M	S_D
a_1	0.927 7	0.595 0	0.967 3	0.610 0	0.942 6	0.592 8
a_2	0.920 3	0.587 2	0.957 6	0.605 6	0.955 7	0.606 5
a_3	0.956 6	0.602 1	0.960 2	0.608 9	0.965 3	0.611 5
a_4	0.960 0	0.620 1	0.941 9	0.606 5	0.961 0	0.608 5
b_1	0.006 0	0.985 8	0.967 3	0.610 0	−0.009 0	0.999 7
b_2	1.295 7	1.113 1	1.077 2	1.195 1	0.799 3	1.223 7
b_3	2.829 8	1.141 0	2.580 4	1.193 3	2.293 6	1.220 2
b_4	4.293 0	1.106 0	4.036 1	1.219 6	3.848 3	1.248 2
θ_1	−0.003 3	0.978 3	−0.007 9	0.981 3	−0.019 0	0.987 4
θ_2	1.497 4	0.981 3	1.491 5	0.984 1	1.496 8	0.986 5
θ_3	3.003 5	0.982 8	3.005 8	0.989 1	2.993 0	0.985 9
θ_4	4.493 5	0.980 2	4.509 5	0.984 9	4.499 2	0.987 6

在表 2 中,a_1 为年级 1 的区分度,a_2 为年级 2 的区分度,a_3 为年级 3 的区分度,a_4 为年级 4 的区分度; b_1 为年级 1 的难度,b_2 为年级 2 的难度,b_3 为年级 3 的难度,b_4 为年级 4 的难度; θ_1 为年级 1 的被试能力,θ_2 为年级 2 的被试能力,θ_3 为年级 3 的被试能力,θ_4 为年级 4 的被试能力; M 为平均数,S_D 为标准差.

从表 2 可以看出,各年级区分度参数均值均接近 1,标准差趋近 0.6,这与表 1 预设的数据分布特征值是吻合的. 在表 2 中,各年级难度参数的标准差均接近于 1,但在效应大小为 1.5 时,难度参数的标准差除年级 1 外,年级 2 ~ 年级 4 的难度参数标准差偏差大,且随着锚题比例的上升,偏差逐渐增大,与实际情况是相符合的,即随着年级离散度的增加,年级之间差距越来越大,这时如果锚题越多,那么垂直等值的效果就可能会越来越差.

2.2　不同年级离散度与锚题比例条件下垂直等值各参数估计值的 b_{ias} 结果

不同年级离散度与锚题比例条件下垂直等值各参数估计值的 b_{ias} 结果如表 3 所示.

在表 3 中,$\hat{\tau}_i$ 为估计值,b_{ias} 是依据表 2 的 τ_i 和表 3 的 $\hat{\tau}_i$ 根据(5)式计算所得. 当效应大小为 1.5、锚题比例为 45% 时,进行参数估计,数据不收敛,用“NA”表示,此部分结果不纳入后续数据分析.

为了更直观地展现结果,根据表 3 的 b_{ias} 值,分别绘制了不同离散度与锚题比例组合下各年级区分度、难度和能力参数垂直等值的 b_{ias} 值的折线图,如图 2 所示.

在图 2 中,纵坐标为 b_{ias} 值,横坐标为年级离散度与锚题比例组合,如 0.5-15%,表示年级离散度为 0.5、锚题比例为 15% 的 2 者组合,其余依次类推.

为了更直观地展现结果,根据表 3 的 b_{ias} 值,分别绘制了不同锚题比例区分度、难度和能力参数垂直等值的 b_{ias} 值的折线图,如图 3 所示.

表 3　不同年级离散度与锚题比例条件下垂直等值各参数估计值的 b_{ias}

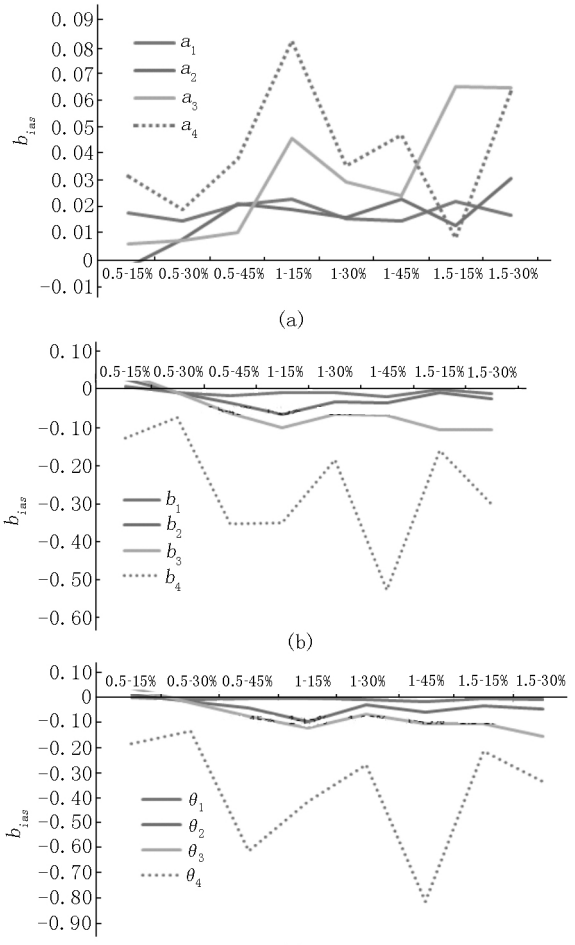
	0.5					
	15%		30%		45%	
	$\hat{\tau}_i$	b_{ias}	$\hat{\tau}_i$	b_{ias}	$\hat{\tau}_i$	b_{ias}
a_1	0.950 1	0.017 4	0.982 5	0.014 3	0.968 5	0.020 4
a_2	0.913 1	−0.002 9	0.958 5	0.007 5	0.968 7	0.020 9
a_3	0.957 6	0.005 7	0.965 8	0.007 1	0.977 4	0.010 0
a_4	0.989 2	0.031 5	0.959 4	0.018 9	1.002 4	0.037 9
b_1	0.018 9	0.006 5	0.023 0	−0.010 7	−0.027 7	−0.018 0
b_2	0.478 5	0.025 5	0.356 8	−0.011 5	0.217 7	−0.036 7
b_3	1.021 3	0.041 3	0.868 3	−0.010 7	0.674 7	−0.063 9
b_4	1.303 6	−0.129 4	1.271 6	−0.074 5	0.936 0	−0.353 9
θ_1	−0.004 2	−0.002 1	−0.018 4	−0.009 2	−0.021 0	−0.003 5
θ_2	0.510 0	0.013 5	0.474 7	−0.017 1	0.454 6	−0.043 0
θ_3	1.040 6	0.037 0	0.983 6	−0.022 2	0.914 8	−0.078 0
θ_4	1.305 6	−0.186 6	1.375 8	−0.134 2	0.890 1	−0.612 5
	1.0					
	15%		30%		45%	
	$\hat{\tau}_i$	b_{ias}	$\hat{\tau}_i$	b_{ias}	$\hat{\tau}_i$	b_{ias}
a_1	0.958 8	0.022 7	0.983 4	0.015 2	0.957 3	0.014 7
a_2	0.940 0	0.019 0	0.966 7	0.015 7	0.978 2	0.022 5
a_3	0.993 6	0.045 3	0.988 0	0.029 3	0.989 3	0.024 0
a_4	1.036 6	0.081 9	0.975 5	0.035 0	1.007 9	0.046 9
b_1	−0.003 2	−0.010 5	0.022 5	−0.011 2	−0.030 7	−0.021 7
b_2	0.809 0	−0.067 4	0.684 5	−0.033 8	0.487 0	−0.037 3
b_3	1.802 8	−0.102 6	1.662 0	−0.067 0	1.448 3	−0.070 3
b_4	2.508 8	−0.350 5	2.510 7	−0.185 4	2.044 5	−0.528 8
θ_1	−0.005 4	−0.002 7	−0.018 4	−0.009 2	−0.038 0	−0.019 0
θ_2	0.899 8	−0.098 0	0.959 9	−0.031 9	0.938 0	−0.058 8
θ_3	1.881 6	−0.123 9	1.940 0	−0.065 8	1.888 4	−0.104 6
θ_4	2.574 3	−0.418 2	2.738 7	−0.271 3	2.184 8	−0.814 4
	1.0					
	15%		30%		45%	
	$\hat{\tau}_i$	b_{ias}	$\hat{\tau}_i$	b_{ias}	$\hat{\tau}_i$	b_{ias}
a_1	0.949 4	0.021 7	0.984 1	0.016 8	NA	−0.942 6
a_2	0.933 0	0.012 7	0.987 9	0.030 3	NA	−0.955 7
a_3	1.021 5	0.064 9	1.024 5	0.064 3	NA	−0.965 3
a_4	0.968 0	0.008 0	1.004 9	0.063 0	NA	−0.961 0
b_1	0.004 5	−0.001 5	0.954 3	−0.013 0	NA	0.009 0
b_2	1.285 2	−0.010 5	1.051 3	−0.025 9	NA	−0.799 3
b_3	2.721 1	−0.108 7	2.471 9	−0.108 5	NA	−2.293 6
b_4	4.131 2	−0.161 8	3.733 3	−0.302 8	NA	−3.848 3
θ_1	−0.006 6	−0.003 3	−0.015 8	−0.007 9	NA	0.019 0
θ_2	1.465 3	−0.032 1	1.446 0	−0.045 5	NA	−1.496 8
θ_3	2.896 8	−0.106 7	2.847 7	−0.158 1	NA	−2.993 0
θ_4	4.276 7	−0.216 8	4.170 7	−0.338 8	NA	−4.499 2

在图 3 中,纵坐标为 b_{ias} 值,横坐标分别为年级离散度与不同年级区分度、不同年级难度、不同年级被试能力的组合. 例如,对于图 3(a),0.5-a_1,表示年级离散度为 0.5、年级 1 区分度的两者组合; 对于图 3(b),1-b_2,表示年级离散度为 1、年级 2 难度的两者组合; 其余依次类推.

3 分析与讨论

3.1 垂直等值项目和能力参数值分析

在表 2 中,难度参数均值随年级的上升而增加,这是符合逻辑的,即一般情况下,随着年级的上升,题目的难度设置越来越高. 但是,在效应大小相同的情况下,各年级难度均值均随着锚题比例的上升而下降. 在表 2 中,基准年级(年级 1)被试能力的均值趋近于 0,其他年级的被试的能力均值逐渐上升,这反映出各年级的被试能力是逐渐增加的,这与 A. A. Sari 等[4]的研究结果一致.



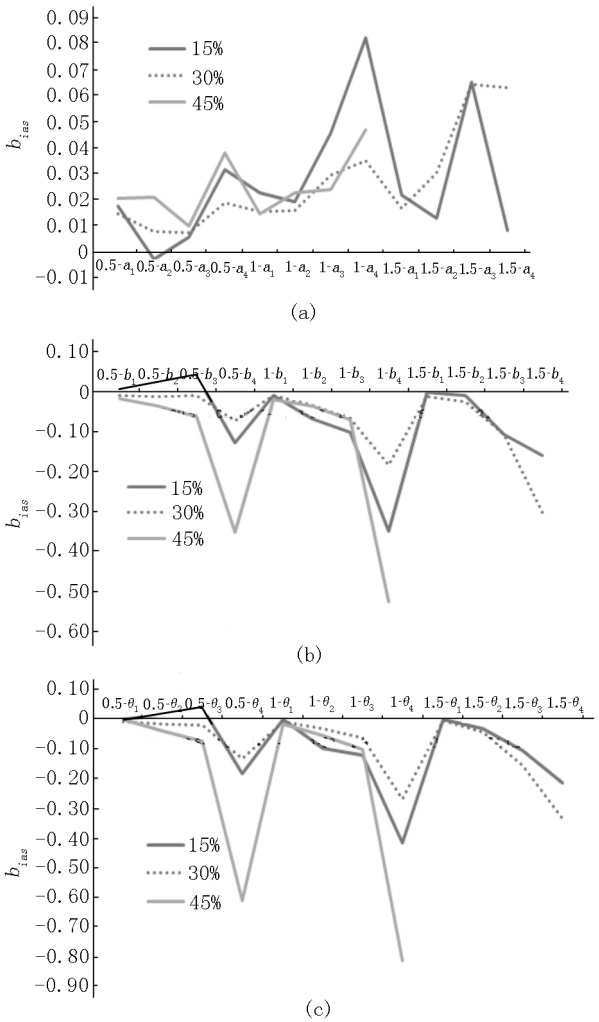
年级离散度与锚题比例组合

图 2 各年级区分度参数 b_{ias} 值

表 2 结果表明,各项参数值(即真值)与设定的数据分布特征值(见表 1)是基本吻合的,这说明数据模拟方法是可信的,这为后续分析生成被试原始分数作答矩阵,再估计项目的区分度、难度及被试的能力奠定了基础.

3.2 不同年级垂直等值 b_{ias} 结果分析

从表 3 可知,当效应大小为 1.5、锚题比例为 45% 时,数据不收敛,这表明垂直等值随着年级增长、锚题越多,年级之间学生能力的共同题目效果越来越差(直至导致数据不能收敛),这是符合逻辑的. 通常,年级 1 的学生与年级 4 的学生能力相差较为悬殊,如果这时设置过多的共同题(锚题),垂直等值的效果反而较差,这直接导致了垂直等值难以进行,最终失败(即如本文表 3 中出现的“NA”).



年级离散度与不同年级区分度、难度、被试能力组合

图 3 不同锚题比例区分度参数 b_{ias} 值

从图 2 可看出,随着年级的增长,垂直等值的 b_{ias} 值越来越大. 其中,年级 1 和年级 2 的 b_{ias} 值折线基本平行,而年级 3 随着效应值的增加,b_{ias} 值也逐渐变大,年级 4 随着效应值的增加,估计精度大幅度下降. 这充分体现出,离基准年级越远,估计精度随之下降. 年级 1、年级 2 和年级 3 之间的 b_{ias} 值相差还相对较小,但到了年级 4 时,b_{ias} 的值已经大得难以接受了(可从图 2 的虚线看出). 因此,若选择以低年级为基准年级时,在高年级测验分数向低年级转换过程中,若连续累积转换的次数超过 2 次,则会有较大的偏差. 反之亦然. 因此,若高年级向低年级转换,则建议与基准年级间隔不宜超过 2 个年级. 否

则,造成的垂直等值 b_{ias} 值是不可接受的.

3.3　不同锚题比例垂直等值 b_{ias} 结果分析

从图 3(a) 可看出,区分度参数对各锚题比例的估计精度在年级 1 与年级 2 上相差并不大,但是在年级 3 与年级 4 上各锚题比例的估计精度却相差较大. 对于年级 3,区分度参数锚题比例 30% 和 45% 的估计精度相差不大,但要优于锚题比例 15%. 对于年级 4,不同锚题比例估计精度相差较大,相较而言,锚题比例 15% 时效果更好些,但与锚题比例 30% 相比,没有太大差异. 因此,从总体趋势看,锚题比例 30% 时可以更好地保证得到较佳的区分度估计.

从图 3(b) 可看出,难度参数在不同锚题比例下估计精度的差异要比区分度参数明显,即 b_{ias} 的值呈现出 30% <15% <45%,且当锚题比例为 30% 时,其估计精度明显优于锚题比例为 45% 和 15% 的估计精度,即便是年级 4 本身的估计精度要比其它 3 个年级差很多,但依旧也是锚题比例为 30% 时估计精度最佳.

从图 3(c) 可看出,能力参数在不同锚题比例下估计精度存在差异,分化明显. 当效应大小为 1.5 时,锚题比例 45% 出现了数据不收敛的情况,即在年级间学生能力增长较大的情况下,锚题比例的设定是不宜过大的,45% 是不合适的. 比较 30% 和 15% 锚题比例在能力参数上产生的等值误差,前者明显小于后者.

整体而言,当锚题比例为 30% 时,垂直等值的精度相对最高,产生的等值误差相对最小.

3.4　年级离散度与锚题比例的交互作用

按上所述,从整体而言,锚题比例为 30% 时垂直等值效果相对最佳,但综合图 3(a) ~图 3(c) 不难发现,当年级离散度为 1.5 时,锚题比例为 30% 的等值精度都小于 15% 的等值精度. 由于“局部”等值精度存在差异,因此不得不考虑不同年级离散度与锚题比例的“交互作用”.

年级离散度可以反映不同年级学生水平之间的差异. 当年级离散度为 1.5 时,不同年级学生水平之间的差异相对较大,可能造成垂直等值的估计精度极不稳定,甚至有可能出现数据无法收敛(上文已提及并作了分析) . 因此,在离散度相对较大的情形下,锚题比例设置过高是不适合的. 这是因为,过多的锚题反而会对垂直等值造成一种严重的“负担”.

即便依照上述分析得出从整体上锚题比例为 30% 最好,但对于局部而言,其效果仍然值得商榷或考究. 实际上,从图 3 揭示出,当离散度为 1.5 时,只有锚题比例为 15%,估计精度才有所回归,表现出不是“渐行渐远”的状态. 因此,当离散度为 1.5 时,锚题比例不宜设置过高,这是在充分考虑不同年级学生实际情况下所形成的,是与锚题比例整体结果设定(30%) 不相抵触的,是考虑了年级离散度与锚题比例所形成的一种“交互效应”后果.

总之,当年级离散度为 0.5 和 1.0 时,锚题比例为 30% 时等值精度的效果最好. 但是,当年级离散度为 1.5 时,锚题比例在 15% 时估计精度最佳.

3.5　不足与展望

本研究还存在着一些不足,是后续研究有待作出进一步探讨的,如下: (i) 仅考虑以低年级作为基准,没有对以中年级或高年级作为基准作进一步的分析; (ii) 所考虑的锚题参数分布在不同年级之间不发生变化,这不同于 A. A. Sari 等[4] 的做法; (iii) 假定的年级离散度和测验长度在不同年级之间是相同的,并不发生变化,这不同于熊建华等[20] 的做法; (iv) 项目参数校准/估计方法仅限于同时标定法,未考虑分别标定、固定共同题参数标定等其它方法[21]; (v) 返真性指标仅考虑 b_{ias} 值,限于篇幅没有讨论 RMSE 的结果.

4　结论

1) 与基准年级的距离影响垂直等值效果,越靠近基准年级,估计精度越好. 选择以低年级为基准年级时,在高年级测验分数向低年级转换过程中,若连续累积转换的次数超过 2 次,则会有较大的垂直等值偏差. 因此,若是高年级向低年级转换,则建议与基准年级间隔不宜超过 2 个年级.

2) 从不同锚题比例的区分度、难度和能力参数的垂直等值 b_{ias} 的结果分析反映出,就整体而言,当锚题比例为 30% 时,垂直等值的精度相对最高,产生的等值误差相对最小. 因此,在不考虑其它条件影响的情况下,进行垂直等值,锚题比例建议设为 30%.

3) 垂直等值锚题比例的设定受年级离散度影响,两者存在交互作用,锚题比例设为“变”值更好. 当效应大小为 0.5 和 1.0 时,各年级估计精度锚题比例为 30% 时效果最好; 当效应大小为 1.5 时,各年级估计精度锚题比例为 15% 时效果最好. 因此,不同年级锚题比例的设定还应该考虑年级离散度的影响,锚题比例设为“变”值更好.

5　参考文献

[1]　漆书青,戴海崎,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.
[2] 王怡,唐文清,刘晶,等. IRT 与 MIRT 在测验垂直等值中的应用 [J]. 心理科学进展,2014,22(5):881-888.
[3] 叶萌,辛涛. 测验链接中的锚题代表性研究 [J]. 心理科学,2015,38(1):209-215.
[4] Sari A A, Kelecioglu H. Assessment of achievement and growth by vertical scaling: comparison of vertical scaling methods [J]. Journal of Educational Sciences Research, 2016,6(2):25-38.
[5] Kolen M J, Brennan R L. Test equating, scaling, and linking: methods and practices [M]. 2nd ed. New York: Springer-Verlag,2004.
[6] Kolen M J, Brennan R L. Test equating scaling and linking: method and practices [M]. 3rd ed. New York: Springer-Verlag,2013.
[7] 王烨晖,边玉芳,辛涛. 垂直等值的应用及最新发展述评 [J]. 心理学探新,2011,31(5):472-476.
[8] Duong M Q. Evaluating equating results in the non-equivalent groups with anchor test design using equipercentile and equity criteria [EB/OL]. https://eric.ed.gov/?id=ED529395.
[9] 叶昶成. 不同垂直等化设计下可能值方法估计效果值探讨 [D]. 台湾:台中教育大学,2015.
[10] 蔡艳,丁树良,涂冬波. 锚题比例对等值精度的影响 [J]. 心理学探新,2009,29(2):86-89.
[11] Martineau J A. The effects of construct shift on growth and accountability models. Unpublished doctoral dissertation [D]. Michigan: Michigan State University,2004.
[12] Martineau J A. A distorting value added, the use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability [J]. Journal of Educational & Behavioral Statistics,2006,31(1):35-62.
[13] Yen W M. The choice of scale for educational measurement: an IRT perspective [J]. Journal of Educational Measurement,1986,23(4):299-325.
[14] 罗照盛. 项目反应理论基础 [M]. 北京:北京师范大学出版社,2012.
[15] Dorans N J, Pommerich M, Holland P W. Linking and aligning scores and scales [M]. New York: Springer Verlag, 2007:253-273.
[16] 郭小军. 不同参照基准与年级离散程度对垂直等值的影响研究 [D]. 南昌:江西师范大学,2014.
[17] Tong Y, Kolen M J. Comparisons of methodologies and results in vertical scaling for educational achievement tests [J]. Applied Measurement in Education, 2007, 20(2): 227-253.
[18] Kolen M J. Equating and vertical scaling: research questions [C]. The Annual Meeting of the National Council on Measurement in Education,2003.
[19] Yildirim H H. Findings from an empirical vertical scaling study with BILOG-MG [EB/OL]. [2018-02-11]. https://www.researchgate.net/publication/288239514_Findings_from_an_empirical_vertical_scaling_study_with_BILOG-MG.
[20] 熊建华,叶新蓉,丁树良,等. 等值设计中锚题比例研究 [J]. 第三届国际教育技术与训练大会,2010,7:197-200.
[21] 叶萌,辛涛. 垂直量尺化中的参数估计方法及其性能比较 [J]. 心理科学进展,2014,22(10):1669-1678.

The Effect of the Ratio of Anchor Items to Total Test and the Separation of Grade Distributions on the Precision of Vertical Scaling

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

Abstract: The paper aimed at the effect of the ratio of anchor items to total test and the separation of grade distributions on the precision of vertical scaling. The paper used common-item design and chose grade 1 as base grade, by Monte Carlo simulation method simulated the different ratio of anchor items to total test and separation of grade distributions studied the influence of vertical scaling. The result shows that: (i) Selection of base grade affected the precision of vertical scaling. The closer to the base grade, the better the accuracy of the estimate is. (ii) When there is a 30% ratio of anchor items to total test, the precision of vertical scaling is best. (iii) The ratio of anchor items to total test is affect by the separation of grade distributions. There is an interaction between the two, and the ratio of anchor questions is set to "change".

Key words: ratio of anchor items to total test; separation of grade distribution; vertical scaling; test equating; item response theory

(责任编辑:冉小晓)