心理学探新 2025,Vol.45,No.1,63-71
PSYCHOLOGICAL EXPLORATION

同时校准与固定参数校准 MWU 策略在垂直量尺化构念漂移下的性能探讨*

陈子豪[1,2],黎光明[1]
(1.华南师范大学心理学院,心理应用研究中心,广州 510631;2.广州市干部和人才健康管理中心,广州 510530)

摘 要:旨在探究垂直量尺化数据违背跨年级同构性假设,即出现构念漂移时,同时校准与固定参数校准多先验更新策略在模型-数据不匹配条件下的性能表现,进而给出垂直量尺化中处理构念漂移的三步法。结果显示:(1)构念漂移在不同数据-模型匹配条件影响不同;(2)相较同时校准,固定参数校准多先验策略能够更有效减少构念漂移造成的误差;(3)实践中,应该综合构念漂移程度和样本量选择估计模型及校准方法。

关键词:垂直量尺化;双因子模型;构念漂移;固定参数校准;项目反应理论
中图分类号:B841.2 **文献标志码**:A **文章编号**:1003-5184(2025)01-0063-09

1 引言

学生学业水平发展情况,是国内外教育质量监测和提升的一项重点工作。垂直量尺化(vertical scaling)技术可以将难度不同,但所测构念相近的测验分数置于统一的量尺上,进而测量年级间学业成绩的增长水平(grade-to-grade growth)(Kolen & Brennan,2014)。目前主流采用基于项目反应理论(item response theory,IRT;Baker & Kim,2004)的方法来实现垂直量尺化,因此参数估计模型及校准方法是目前垂直量尺化中最受关注的热点之一。

当涉及到的年级跨度较大(4 个年级)时,不同年级的测验较难维持相同的内容结构,容易违背同构性(construct invariance)假设(Reckase & Martineau,2004)。Martineau(2004)将违背同构性的程度称为构念漂移(construct shift)。不同学科测验的特点导致构念漂移程度有很大差异。有实证研究显示某涉及 6 个年级组(3~8 年级)的数学测验出现较强的构念漂移(Martineau,2004),而某涉及 8 个年级组(3~10 年级)的大型阅读测验则能够满足同构性假设(Wang & Jiao,2009)。Martineau(2004)发现构念漂移显著扭曲了垂直量尺输出的结果,可能导致多条线性增长轨迹结合成一条非线性的轨迹,或者多条非线性轨迹结合成一条线性轨迹。使用非线性的单维 IRT(unidimensionalitem response theory,UIRT)模型拟合构念漂移的数据,垂直量尺上的分数也会被显著扭曲(Martineau,2006)。Li 和 Lissitz(2012)操纵了构念漂移的程度,发现群体均值的估计误差随着构念漂移增大而增大。他们使用双因子模型(bifactor model;Cai et al.,2011)对构念漂移进行拟合,将内容结构维度与一般能力维度分离,进而以一般能力维度得分作为量尺分的基础。近年来,Strachan 等人(2020a,2020b)结合 Ip 和 Chen(2012)的研究提出了投射性 IRT(projective item response theory,PIRT)方法,首先选择多维 IRT(multidimensional item response theory,MIRT;Reckase,2009)模型进行参数估计,再将不同维度参数通过投射转换到同一个维度上,该方法的实际效果依赖于多维模型的适当选择。

构念漂移对参数校准方法有何影响,目前研究尚不充分。Li(2011)和 Eastwood(2014)用双因子模型拟合构念漂移,但仅用了同时校准(concurrent calibration,CC)。Koepfler(2012)使用了两种双因子模型,但受限于软件算法,分别校准(separate calibration,SC)、同时校准及固定参数校准(fixed parameter calibration,FPC),仅在 UIRT 模型下进行了比较,且使用的是实证数据,无法比较构念漂移对校准方法的影响。近年来水平等值研究中,Kim(2018)比较了双因子模型下同时校准和分别校准

* **基金项目**:广东省哲学社会科学规划 2024 年度学科共建项目(GD24XXL03),广州市干部和人才健康管理中心课题(JGZX20230304)。
通信作者:黎光明,E-mail:Lgm2004100@sina.com。

的表现,结果显示同时校准等值结果更精确,项目参数和分数分布返真性更好。Kim(2019)进一步比较了 FPC 的单次(one prior weights updating,OWU)和多次先验更新(multiple prior weights updating,MWU)策略,并和分别校准对比,结果显示在等组和非等组条件下,MWU 的参数返真性都要优于 OWU,且整体上与 Haebara 法相当。在多组的非等组锚测验设计(non - equivalent groups anchor test,NEAT)下,Kim 和 Kolen(2019)研究显示,FPC - MWU 在能力参数和项目参数的估计准确性近似于同时校准,甚至在某些条件下更精确。

FPC - MWU 被视为一种改进的同时校准(Kim,2006),但其优势目前尚未凸显,它既能够对参数进行同时估计和校准,又能利用已知的项目参数多次更新潜变量先验权重,有理由推测它可以修正构念漂移所导致的参数估计误差。如果使用不适当的模型拟合构念漂移的数据,将会涉及模型 - 数据不匹配(misfit)的问题,同时校准在不匹配情况下,会出现量尺收缩(scale shrinkage)(Bolt,2014;Briggs & Weeks,2009;Camilli et al.,1993;Yen,1985)。综上,有必要探究在构念漂移及模型 - 数据不匹配条件下,FPC - MWU 和同时校准的参数校准性能差异,以期对垂直量尺化的实践给出一定的参考建议。

2 双因子 IRT 模型

Gibbsons 和 Hedeker(1992)对 Holzinger 和 Swineford(1937)的模型进行了拓展,以正态肩型模型为基础,提出了适用于二级计分数据的全息双因子模型(full - information bifactor model)。该模型有两个条件限制:(1)每道题目都在一般因子上有非零负载且只在一个特殊因子上有非零负载;(2)特殊因子彼此之间以及和一般因子之间正交,即彼此是完全独立的。Cai 等人(2011)在此基础上,提出了基于 logistic 模型的两参数双因子模型(bifactor two - parameter logistic model,BF - 2PL):

$$P(\theta_i, a_j, d_j) = \frac{1}{1 + exp[-D(a_{j0}\theta_{i0} + a_{js}\theta_{is} + d_j)]} \quad (1)$$

在公式(1)中,下标 i 和 j 分别表示人和题目。D 是量尺化常数,取 1.0(logistic 度量)或 1.7(正态化);θ_{i0}表示一般因子,θ_{is}表示特殊因子。a_{j0}和 a_{js}表示一般因子和特殊因子的项目区分度参数;d_j 是截距参数,与难度参数有关。

3 FPC 算法

FPC 通过保持锚题参数跨测验/水平不变来实现所有待链接测验/水平上分数量尺的统一。在 NEAT 设计下,两个测验/水平通过三步实现量尺统一:(1)针对一个测验/水平估算其题目参数和能力参数;(2)将该测验/水平和另一测验/水平的锚题参数固定为第 1 步得到的参数值;(3)估计另一测验/水平上非锚题的项目参数和能力参数(叶萌,辛涛,2014)。

Kim(2006)提出根据 EM 算法中潜变量分布更新与否,FPC 有三种变式:无先验权重更新(no prior weights updating,NWU)、OWU 及 MWU 策略。NWU 在第一次 EM 循环中使用初始指定的潜变量先验分布,之后每一次 EM 循环都不再更新分布;OWU 则使用第 1 次 EM 循环的结果更新一次先验分布,之后不再更新;MWU 在第 2 次 EM 循环开始不断更新先验分布,直到迭代收敛为止。积分权重通过上一步循环求解的参数信息进行连续更新,将使潜变量分布的估计精度越来越高。他发现对异质性越强的多组数据,MWU 估计精确性和稳健性越高于其他两种。但权重更新也有一定风险,如果新的题目质量较差或者题目和模型的拟合较差,将可能在估计项目参数和潜变量密度函数时引入误差(Kim,2019)。

在实现 FPC 的软件上,Pack 和 Young(2005)指出 BILOG - MG 软件中用于防止重新量尺化的 NO-ADJUST 命令覆盖了更新潜变量分布的 EMPIRICAL 命令,导致算法不能在每次 EM 循环中使用上一次循环获得的潜变量分布,相当于是 FPC - NWU,且只能计算二分数据。PARSCALE 软件没有这种冲突,因此是 FPC - MWU。其支持混合题型数据,但仅支持等组的两组数据。MULTILOG 软件支持多组数据,却要求多组的方差是同质的。Kim(2006)和 Kim(2019)分别采用更为灵活的 ICL 软件和 R 语言编程实现了水平等值的 FPC - MWU,Kim 和 Kolen(2019)使用 ICL 软件实现了适用于多组异质数据的 FPC - MWU。但 ICL 和 R 语言相对来说使用难度大,语句较难理解。FlexMIRT(Cai,2017)软件兼容混合题型和多组异质数据,且使用 FPC - MWU,运算功能强大且快捷,命令语句简洁易懂,但目前尚未有相关研究采用。研究者使用 flexMIRT 编写适用于垂直量尺化的 FPC - MWU 脚本。

4 方法

4.1 研究设计

采用 NEAT 设计,双因子模型中因子与题目的

负载关系如图1所示。每一道题目分别负载到一般因子G和其对应的年级特殊因子S上。年级跨度设置4个年级,以体现年级跨度对垂直量尺化产生的影响,年级标签设置为3年级到6年级。抽取来自相邻低年级的题目作为锚题,共有3套锚题(G34、G45、G56)。

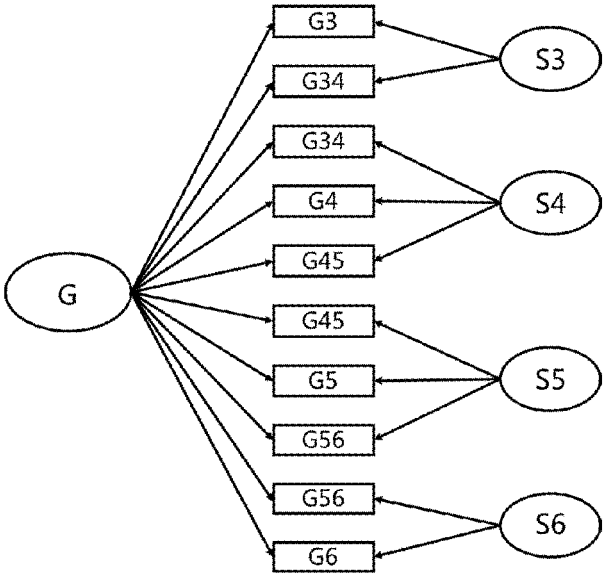

图1　基于双因子模型的NEAT设计

基准年级设定为4年级。基准年级的选择会影响量尺化的参数估计精度,一般以中间年级作为基准年级,可以最大地减少误差(Kim et al.,2009)。每个年级测验长度设为40题,符合前人研究在40~60题之间的惯例(Eastwood,2014;Meng,2007)。锚题比例为25%,即相邻年级之间有10道锚题,锚题类型为内锚。能力参数估计方法选用贝叶斯期望后验估计(expected a posterior estimation,EAP)。

操纵变量共有三个,各变量水平设置如表1所示。构念漂移数据通过BF-2PL生成。为模拟模型-数据不匹配的条件,采用U2PL模型估计具有构念漂移的数据。使用BF-2PL进行参数估计作为模型-数据匹配条件。构念漂移代表特殊因子的离散程度,通过改变特殊因子的方差来操纵,参照Li(2011)设置三个水平。由于一般因子的方差固定为1.0,特殊因子方差为0.25相对较小,方差为0.5和1.0时,特殊因子会对一般因子的估计产生更大的干扰。样本量表示每个年级的考生数量。参考Kim和Kolen(2019)在多组FPC的估计中设置了500和2000两种样本量,分别代表小样本,中样本和中等偏大样本。

表1　操纵变量

变量	水平设置
模型-数据匹配	BF-2PL(匹配),U2PL(不匹配)
构念漂移程度	0.25,0.50,1.00
样本量	500,1000,2000

4.2　数据生成

使用自编R语言程序,模拟生成被试作答数据,所有题目均为0/1计分。共有3(构念漂移)×3(样本量)=9种条件,每种条件重复生成和估计500次,以获得稳定的估计。

参考Kim(2018)将一般因子的项目区分度参数a_{j0}~log-Normal(0,0.25),限制在(0.5,2)的范围内。特殊因子项目区分度参数a_{js}固定为常数1.7,便于模型识别和自由估计特殊因子方差。D取1.7。

对于非锚题,其难度参数b_j的分布与相应年级的一般因子能力分布相同,如对于4年级题目,b_j~$N(0,1)$。截距参数d_j通过公式(2)得到,限制其取值范围[-3,3]。对于锚题,G34的难度参数服从$U(-1,0.5)$,G45的难度参数服从$U(-0.5,1)$,G56的难度参数服从$U(0,1.5)$,以保证锚题在两个相邻年级的难度适中。

$$d_j = -b_j\sqrt{a_{j0}^2 + a_{js}^2} \tag{2}$$

对于能力参数,将基准年级群体能力分布设为标准正态分布$N(0,1)$,其余年级能力均值差距为0.5,标准差固定为1(Gotzmann,2011;Li,2011)。由于特殊因子的方差是操纵变量,将特殊因子均值固定为0,方差为构念漂移程度,4个年级的特殊因子分布设为相同,以构念漂移程度为0.25时为例,具体设置见表2。

表2　能力参数设置

年级水平	一般因子	特殊因子
	θ_{i0}	$\theta_{i3}/\theta_{i4}/\theta_{i5}/\theta_{i6}$
3年级	$N(-0.5,1^2)$	$N(0,0.5^2)$
4年级	$N(0,1^2)$	$N(0,0.5^2)$
5年级	$N(0.5,1^2)$	$N(0,0.5^2)$
6年级	$N(1.0,1^2)$	$N(0,0.5^2)$

4.3　参数估计及校准

采用flexMIRT 3.5软件编写脚本,以实现多组同时校准和FPC-MWU。参考Li(2012)的IRTPRO软件语句获得双因子模型的估计语句(flexMIRT和

IRTPRO 同源,是后者的升级版本)。具体实现方式为,将三套锚题的项目参数分别固定为一套常数,由于没有现成题库,这套常数从锚题的分布函数中随机抽取,在脚本中通过 FIX 和 VALUE 命令赋值。通过 EmpHist = Yes 以及 FREE 命令确保先验权重得到重复更新。

4.4 评价指标

为了评价参数估计方法对于能力项目参数的估计准确性,计算偏差(Bias)和误差均方根(root mean square error,RMSE)。Bias 是误差的总体偏差方向,正值表示偏高估,负值表示偏低估。RMSE 是总体误差的评价指标,代表误差的绝对大小。其数值越小,估计的精度越高。两者计算公式如下:

$$Bias = \frac{\sum_{i=1}^{R}(\hat{\beta}_r - \beta)}{R} \quad (3)$$

$$RMSE = \frac{\sum_{i=1}^{R}(\hat{\beta}_r - \beta)^2}{R} \quad (4)$$

在公式(3)和公式(4)中,$\hat{\beta}_r$ 为第 r 次估计时参数的估计值,β 为参数真值,R 为重复次数。

5 结果

5.1 构念漂移的影响

表 3 和表 4 分别给出了模型 - 数据匹配(BF - 2PL 模型)及模型 - 数据不匹配(U2PL 模型)条件下,两种参数校准方法能力参数估计值的 Bias 和 RMSE。总体来看,随着构念漂移增大,能力参数的 Bias 和 RMSE 增大,估计精确性下降,跨年级的估计稳定性(各年级的误差水平)没有显著变化。对于项目参数,使用 BF - 2PL 模型时,随着构念漂移增大,区分度和难度参数的 Bias 和 RMSE 水平降低,估计精确性提升。区分度参数的跨年级估计稳定性也随之提高。而使用 U2PL 模型时,随着构念漂移增大,项目参数的估计精确性下降,且距离基准年级越远,估计误差越大。区分度参数从被无偏估计到偏向于被高估。

5.2 样本量的影响

由表 3 和表 4 可见,随着样本量增大,能力参数的 Bias 和 RMSE 水平随之有所降低,估计精确性提高,但跨年级的估计误差水平没有显著变化。对于项目参数,随着样本量增大,不论在 BF - 2PL 模型还是 U2PL 模型中,参数的 Bias 和 RMSE 水平均随之降低;区分度参数在 BF - 2PL 模型下跨年级的误差水平随之降低,难度参数则在 U2PL 模型下跨年级的估计稳定性随之提高。

表 3 一般能力参数的 Bias 和 RMSE(BF - 2PL)

样本量	构念漂移	方法	Grade3		Grade4		Grade5		Grade6	
			Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
500	0.25	CC	-0.01	0.57	0.00	0.55	0.00	0.55	0.01	0.57
		FPC	0.00	0.56	-0.01	0.53	0.00	0.57	0.01	0.63
	0.50	CC	0.00	0.64	0.00	0.62	0.00	0.62	0.00	0.64
		FPC	0.00	0.57	0.00	0.55	0.00	0.58	0.00	0.64
	1.00	CC	0.01	0.72	0.00	0.68	0.00	0.68	-0.01	0.73
		FPC	0.00	0.60	0.00	0.59	0.00	0.60	0.00	0.64
1000	0.25	CC	0.00	0.62	0.00	0.59	0.00	0.59	-0.01	0.62
		FPC	-0.01	0.52	-0.01	0.50	0.00	0.53	0.01	0.57
	0.50	CC	0.01	0.62	0.00	0.60	0.00	0.60	-0.01	0.62
		FPC	0.00	0.54	-0.01	0.52	0.00	0.54	0.00	0.57
	1.00	CC	0.01	0.71	0.00	0.66	0.00	0.67	-0.01	0.71
		FPC	0.00	0.58	0.00	0.57	0.00	0.58	0.00	0.60
2000	0.25	CC	0.00	0.51	0.00	0.50	0.00	0.50	0.00	0.51
		FPC	-0.01	0.50	-0.01	0.49	0.00	0.50	0.01	0.53
	0.50	CC	0.01	0.60	0.00	0.58	0.00	0.58	-0.01	0.60
		FPC	0.00	0.52	-0.01	0.51	0.00	0.52	0.01	0.55
	1.00	CC	0.01	0.68	0.00	0.66	0.00	0.66	-0.01	0.69
		FPC	0.00	0.57	0.00	0.56	0.00	0.57	0.00	0.59

表4　能力参数Bias和RMSE(U2PL)

样本量	构念漂移	方法	Grade3		Grade4		Grade5		Grade6	
			Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
500	0.25	CC	0.00	0.51	0.00	0.51	0.00	0.51	0.01	0.51
		FPC	-0.02	0.50	-0.01	0.48	0.01	0.49	0.01	0.51
	0.50	CC	0.01	0.64	0.00	0.64	-0.01	0.64	-0.03	0.64
		FPC	-0.03	0.64	-0.01	0.61	0.01	0.61	0.00	0.65
	1.00	CC	0.01	0.81	-0.01	0.81	-0.03	0.81	-0.07	0.81
		FPC	-0.03	0.82	-0.01	0.82	0.02	0.80	-0.03	0.80
1000	0.25	CC	0.00	0.49	0.00	0.49	0.00	0.49	-0.01	0.49
		FPC	-0.01	0.49	-0.01	0.48	0.01	0.48	0.00	0.50
	0.50	CC	0.01	0.63	-0.01	0.63	-0.02	0.62	-0.05	0.63
		FPC	-0.02	0.63	-0.01	0.61	0.01	0.61	-0.01	0.63
	1.00	CC	0.02	0.79	-0.01	0.79	-0.04	0.79	-0.10	0.79
		FPC	-0.03	0.81	-0.01	0.81	0.01	0.79	-0.04	0.79
2000	0.25	CC	0.00	0.49	0.00	0.49	-0.01	0.49	-0.03	0.49
		FPC	-0.01	0.49	-0.01	0.48	0.01	0.48	-0.01	0.49
	0.50	CC	0.01	0.62	-0.01	0.62	-0.03	0.62	-0.06	0.62
		FPC	-0.02	0.63	-0.01	0.61	0.01	0.61	-0.02	0.62
	1.00	CC	0.02	0.78	-0.01	0.79	-0.05	0.78	-0.11	0.78
		FPC	-0.02	0.80	-0.01	0.81	0.02	0.78	-0.04	0.77

5.3　同时校准和FPC－MWU的性能比较

为了直观体现在不同条件下，两种参数校准方法(CC和FPC－MWU)的性能差异，图2～图4分别呈现了它们在区分度参数、难度参数和能力参数的Bias和RMSE折线图。

模型－数据匹配条件(BF－2PL)下，对于一般因子区分度参数，FPC－MWU在构念漂移为0.25，样本量为500～1000时，Bias和RMSE都要明显小于CC，但在其他条件组合下，两种校准方法误差水平近似。以RMSE作为因变量，构念漂移与样本量为被试内变量，进行三因素方差分析以检验校准方法的组间差异。总体上两种参数校准方法的RMSE无显著差异($p=0.184$，$\eta^2=0.033$)；在特殊因子区分度参数上，FPC－MWU的RMSE显著更小($p<0.001$，$\eta^2=0.640$)；对于难度参数的估计，CC略偏高估，且离基准年级越远Bias越大；CC的RMSE显著大于FPC－MWU($p<0.001$，$\eta^2=0.247$)；对于一般能力参数，CC出现轻微的量尺收缩，离基准年级越远，越有偏低估的趋势，并且CC的RMSE显著更大($p<0.001$，$\eta^2=0.718$)。

模型－数据不匹配条件(U2PL)下，对于区分度参数，两种校准方法都偏高估，但CC的Bias幅度更大。FPC－MWU的RMSE显著更小($p=0.011$，$\eta^2=0.114$)；在难度参数上，FPC－MWU是无偏估计，而CC离基准年级越远的年级Bias越大，RMSE也显著更高($p<0.001$，$\eta^2=0.945$)；对能力参数，CC出现明显的低估，即出现了一定程度的量尺收缩，但在RMSE上两种校准方法没有显著差异($p=0.135$，$\eta^2=0.041$)。

6　讨论

参数估计模型和校准方法的选择，一直是垂直量尺化的重要一环。构念漂移如何影响参数估计和校准方法的性能表现，至今相关研究仍较缺乏。目前只有Li(2011)探究了构念漂移对CC的影响，发现在模型－数据不匹配时，CC对群体能力参数和区分度参数的估计误差随之增大，而与难度相关的截距参数基本不受影响。研究者通过模拟研究，对两种参数校准方法的效果进行了较为系统的探讨。

6.1　FPC－MWU有效减少了构念漂移带来的误差

在模型－数据匹配条件下，两种方法对一般能力参数的估计精确性均随着构念漂移的增大而下降，但FPC－MWU相较CC受影响的程度更小，估计精度更高，在各年级的误差水平都相对一致。在模型－数据不匹配条件下，两种方法受构念漂移的影响更大。对能力参数，CC同样在构念漂移的影响下出现明显的低估，而FPC－MWU在各年级的偏差都相对更小。对于区分度参数，CC受构念漂移影响出现更大幅度的高估。在难度参数上，CC出现了明显的低估而FPC－MWU能够一致保持在无偏的状态。同时，FPC－MWU的项目参数估计精确性即使

在极端的构念漂移程度(1.0)下也能相对保持稳定。

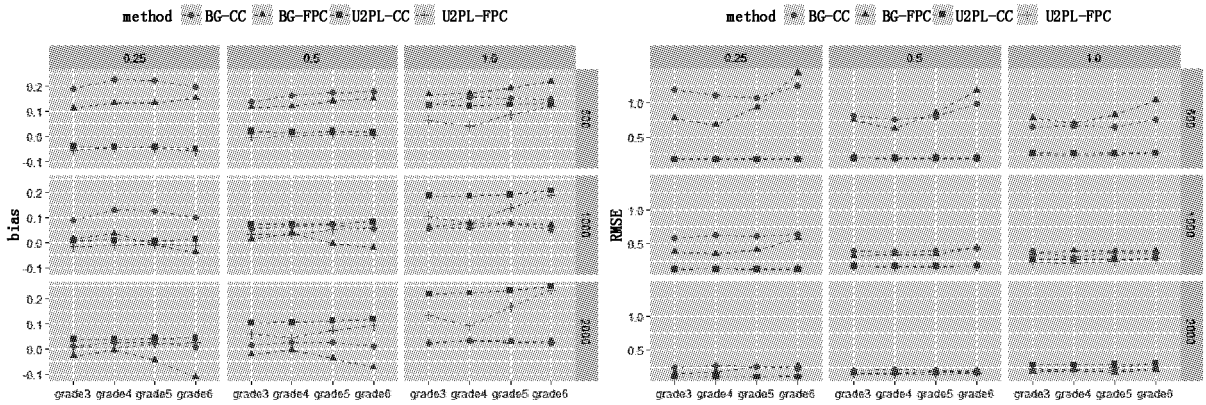

图 2 区分度参数 Bias(左)对比与 RMSE(右)对比

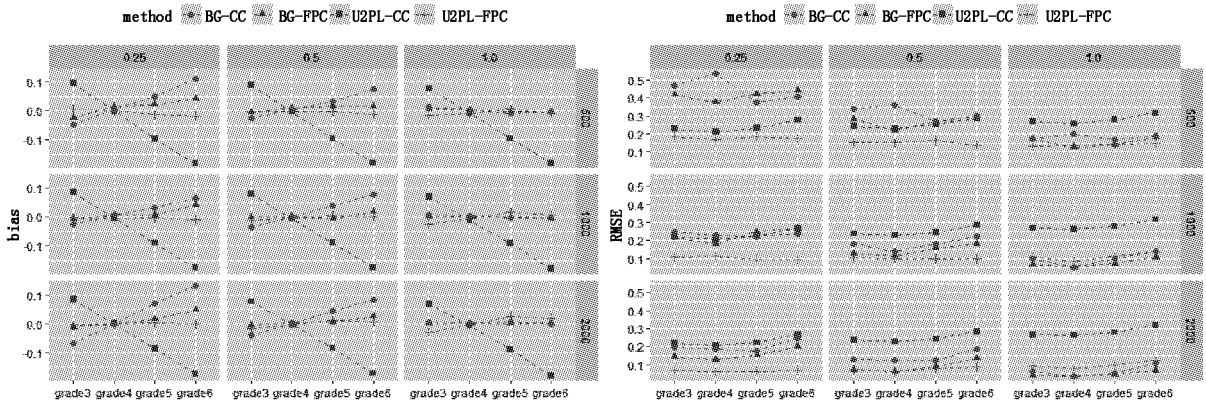

图 3 难度参数 Bias(左)对比与 RMSE(右)对比

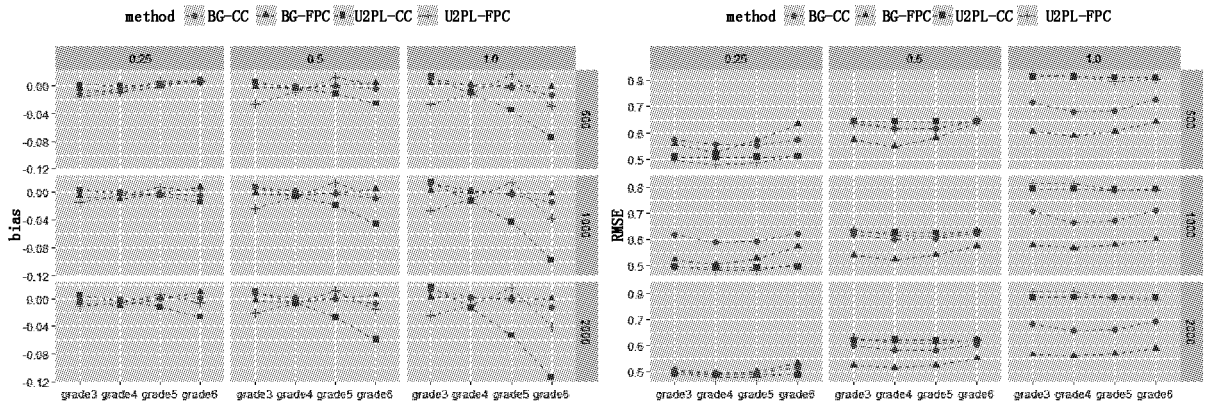

图 4 能力参数 Bias(左)对比与 RMSE(右)对比

研究结果进一步证实了 FPC - MWU 在构念漂移下具有更好的表现。这种优势的原因可能有两点,一是 FPC - MWU 在 EM 循环中进行了多次潜变量分布更新,这确保了潜变量估计不会因为构念漂移(特殊因子方差)的增大出现较大偏差;二是 FPC - MWU 通过已知的固定参数,确立了准确的量尺,而 CC 则是在基准年级上建立量尺,即使不同年级的参数是同时估计的,也需要通过线性转换的过程将参数转换到基准年级的量尺上,这必然会引入误差。研究设计虽然没有纳入分别校准进行比较,但可以合理推测分别校准多次链接的过程会产生较大误差。在未考虑构念漂移和多组校准的背景下,已有研究表明在水平等值时,分别校准并未优于 CC 和 FPC - MWU(Kang & Peterson,2012;Kim,2018;Kim,2019)。同样,FPC - OWU 和 FPC - NWU 也没有被纳入比较考虑,因为前人研究发现这两者并未

优于同时校准和FPC－MWU(Gotzmann,2011;Kim et al.,2009)。

6.2　同时校准出现量尺收缩和基准年级距离效应

通过让U2PL模型估计具有构念漂移的数据,构造出模型－数据不匹配条件。结果发现CC出现了一定程度的量尺收缩,即能力参数偏向于被低估,能力的全距小于设定的范围。前人发现使用UIRT模型估计多维属性的数据,CC也会出现量尺收缩(Bolt,2014;Briggs & Weeks,2009;Yen,1985),研究结果补充了CC在构念漂移造成的模型－数据不匹配下的性能表现。此外,CC在难度参数和能力参数的估计上,出现了典型的基准年级距离效应(Briggs & Weeks,2009;Gotzmann,2011)。即对基准年级的估计最准确,而在距离越远的年级,估计误差越大。

以上两种现象与CC内在的校准逻辑有很大关系,基准年级一般被设定服从0/1标准正态分布,其他年级参数必须要线性转换到这一量尺上,转换系数通过两个相邻年级的参数属性计算得出,距离基准年级越远,需要叠加的转换系数就越多,进而累积更多的误差。因此使用CC时,一般选择中间位置的年级作为基准,以获得最小的转换次数(Kim et al.,2009)。结果中FPC－MWU并没有明显出现以上两种现象,推测与其固定了与基准年级相邻的两套锚题参数有关。锚题的参数是从生成数据的模拟程序中随机抽取的,在生成作答矩阵时也同样使用了这一套参数,这模拟了题库的使用场景,基准年级的锚题从题库中抽取。其好处是量尺是准确的,无需进行估计产生误差,进而系数转换时积累的误差也会更少。可见,在题库的背景下进行垂直量尺化,FPC－MWU是更佳的选择。

另一发现是当样本量在500～1000和构念漂移程度在0.25时,U2PL模型的估计精确性反而高于BF－2PL模型。这一现象有两点解释:第一,双因子模型作为一种MIRT模型,为了获得稳定和准确的参数估计结果,必须要有一定的样本量,而样本量设置与维度数量成正比(Reckase,1997)。在Li(2011)的研究中,其最小样本量1000时BF－2PL模型相对U2PL的优势也不够明显,有学者建议MIRT的样本量至少为2000(Yao & Boughton,2007)。可见样本量在使用双因子模型之前是必须考虑的。第二,构念漂移程度小于一定量级,数据本身可视为单维性,使用BF－2PL模型的拟合度反而更差,Ip(2010)和Carlson(2017)的研究提出了相同的见解。Li(2011)建议如果构念漂移小于0.25,使用U2PL模型足够精确。可见构念漂移值在0.25以下时,其引入的误差很小,可以不必专门处理。如果有意探讨BF－2PL模型在构念漂移量级小于0.25时估计性能的提升,或许可以考虑结合双因子模型和PIRT方法进行参数估计。PIRT方法本质上等价于放宽了局部独立性假设的UIRT(Ip,2010),相比单纯使用双因子模型,它还考虑到了特殊因子的参数信息。

6.3　FlexMIRT脚本有效性

使用flexMIRT实现基于BF－2PL模型的FPC－MWU,相较于ICL和R语言,脚本更简单易懂,易于上手操作。该软件采用降维算法,运算速度快捷(Cai,2017)。为了论证该脚本的有效性,对比近似的研究结果:Kim和Kolen(2019)采用ICL软件编写多组FPC－MWU,数据生成和估计均采用U3PL模型,群体能力均值Bias水平为0.01～0.02,RMSE水平为0.03～0.08(样本量500～2000);Li(2011)采用IRTPRO实现双因子模型的CC,群体能力均值Bias水平在－0.04～0.01之间,RMSE水平在0.56～0.79(样本量1000～4000);在模型－数据匹配条件下,FPC－MWU的Bias水平为0.01,RMSE水平为0.49～0.64,CC的Bias水平为±0.01,RMSE水平为0.50～0.73(样本量500～2000)。可见估计误差与前人同类研究是近似水平。

6.4　处理构念漂移三步法

综合前人的研究成果,提出处理垂直量尺化中构念漂移的三步法:

第一步,测定数据的构念漂移量。在垂直量尺化中,应该始终假定存在构念漂移。由于在实际应用中不清楚数据的构念漂移量,此时可先采用限制性双因子模型拟合数据。改进Li(2011)比较多个限制程度由少到多的泛双因子模型拟合系数的做法,建议直接选择限制性最小的固定斜率双因子模型,将特殊因子的斜率设定为1或一般因子区分度参数的均值,以便估计出特殊因子方差,即构念漂移量。

第二步,根据测得的构念漂移量以及样本量大小,选择适当的估计模型。如果各年级的构念漂移程度小于0.25,则可以选择UIRT模型垂直量尺化。反之则要再考虑样本量,如果样本量小于1000,双因子模型对项目参数的估计会出现较大误差,如果此时偏重能力参数的准确性,可以继续选择双因子

模型(如表 5)。如果偏重项目参数的准确性,则考虑选择 UIRT 模型(如表 6)。如果样本量大于 1000,则建议选择双因子模型。

表 5 垂直量尺化模型及校准方法选择参考(偏重能力参数准确性)

样本量	构念漂移		
	0.25	0.50	1.00
500	UIRT 模型 + MWU	双因子模型 + MWU	双因子模型 + MWU
1000	UIRT 模型 + MWU	双因子模型 + MWU	双因子模型 + MWU
2000	UIRT 模型 + MWU/CC	双因子模型 + MWU	双因子模型 + MWU

表 6 垂直量尺化模型及校准方法选择参考(偏重项目参数准确性)

样本量	构念漂移		
	0.25	0.50	1.00
500	UIRT 模型 + MWU	UIRT 模型 + MWU	UIRT 模型 + MWU
1000	UIRT 模型 + MWU	UIRT 模型 + MWU	UIRT 模型 + MWU
2000	UIRT 模型 + MWU/CC	双因子模型 + MWU	双因子模型 + MWU

第三步,选择参数校准方法。无论第二步选择何种模型,首选 FPC - MWU 进行参数估计,在个别条件下也可选择 CC。

7 结论

第一,相较同时校准,FPC - MWU 更少受到构念漂移造成的影响,多次更新潜变量分布对于构念漂移造成的误差是有效的修正。

第二,在模型 - 数据错误匹配与构念漂移的作用下,同时校准出现了典型的量尺收缩和基准年级距离效应,而 FPC - MWU 的估计准确性能够相对保持稳定。

第三,实践中应结合构念漂移程度和样本量,选择估计模型和参数校准方法。双因子模型适用的前提是构念漂移量 0.25 以上和样本量 1000 以上,否则 UIRT 模型的精确性已足够。

参考文献

叶萌,辛涛.(2014).垂直量尺化中的参数标定方法及其性能比较.*心理科学进展*,*22*(10),1669 - 1678.

Baker,F. B.,& Kim,S. - H.(Eds.).(2004). *Item Response Theory:Parameter Estimation Techniques*,*Second Edition*(2nd ed.). CRC Press.

Bolt,D. M.,Deng,S.,& Lee,S.(2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*,*51*(2),141 - 162.

Briggs,D. C.,& Weeks,J. P.(2009). The impact of vertical scaling decisions on growth interpretations. *Educational and Measurement:Issues and Practice*,*28*(4),3 - 14.

Cai,L.(2017). *flexMIRT version* 3.51:*Flexible multilevel multidimensional item analysis and test scoring*[Computer software]. Chapel Hill,NC:Vector Psychometric Group.

Camilli,G.,Yamamoto,K.,& Wang,M.(1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*,*17*(4),379 - 388.

Carlson,J. E.(2017). *Unidimensional vertical scaling in multidimensional space*(pp. 1 - 28). ETS Research Report Series.

Eastwood,M.(2014). *The effects of construct shift and model - data misfit on estimates of growth using vertical scales*[Unpublished doctoral dissertation]. University of Connecticut.

Gotzmann,A. J.(2011). *Comparison of vertical scaling methods in the context of NCLB*[Unpublished doctoral dissertation]. University of Alberta.

Gibbons,R. D.,& Hedeker,D. R.(1992). Full - information item bi - factor analysis. *Psychometrika*,*57*(3),423 - 436.

Holzinger,K. J.,& Swineford,F.(1937). The bi - factor method. *Psychometrika*,*2*,41 - 54.

Ip,E. H.(2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*,*63*(2),395 - 416.

Ip,E. H.,& Chen,S. H.(2012). Projective item response model for test - independent measurement. *Applied Psychological Measurement*,*36*(7),581 - 601.

Kang,T.,& Petersen,N. S.(2012). Linking item parameters to a base scale. *Asia Pacific Education Review*,*13*(2),311 - 321.

Kim,J.,Lee,W.,Kim,D.,& Kelley,K.(2009). *Investigation of vertical scaling using the rasch model.*[Conference]. National Council on Measurement in Education,San Diego,CA,United States.

Kim,K. Y.(2018). A comparison of the separate and concurrent calibration methods for the full - information bifactor model. *Applied Psychological Measurement*,*43*(7),512 - 526.

Kim,K. Y.(2019). Two IRT fixed parameter calibration methods for the bifactor model. *Journal of Educational Measure-*

ment,*57*(1),29 - 50.

Kim,S.,& Kolen,M. J. (2019). Application of IRT fixed parameter calibration to multiple - group test data. *Applied Measurement in Education*,*32*(4),310 - 324.

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

Koepfler,J. R. (2012). *Examining the bifactor IRT model for vertical scaling in K - 12 assessment* [Unpublished doctoral dissertation]. James Madison University.

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

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

Li,Y.,& Lissitz,R. W. (2012). Exploring the full - information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*,*36*(1),3 - 20.

Martineau,J. A. (2004). *The effects of construct shift on growth and accountability models* [Doctoral dissertation, Michigan State University]. ProQuest Information & Learning.

Meng,H. (2007). *A comparison study of IRT calibration methods for mixed - format tests in vertical scaling* [Unpublished doctoral dissertation]. University of Iowa.

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.

Reckase,M. D. (2009). *Multidimensional item response theory*. Springer.

Strachan,T.,Ip,E.,Fu,Y.,Ackerman,T.,Chen,S. H.,& Willse,J. (2020a). Robustness of projective IRT to misspecification of the underlying multidimensional model. *Applied Psychological Measurement*,*44*(5),362 - 375.

Strachan,T.,Cho,U. H.,Kim,K. Y.,Willse,J. T.,Chen,S. - H.,Ip,E. H.,Ackerman,T. A.,& Weeks,J. P. (2020b). Using a projection IRT method for vertical scaling when construct shift is present. *Journal of Educational Measurement*,*58*(2),211 - 235.

Wang,S.,& Jiao,H. (2009). Construct equivalence across grades in a vertical scale for a K - 12 large - scale reading assessment. *Educational and Psychological Measurement*,*69*(5),760 - 777.

Yao,L.,& Boughton,K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*,*31*(2),83 - 105.

Yen,W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*,*50*(4),399 - 410.

Comparison of Concurrent Calibration and Fixed - Parameter Calibration with Multiple Prior Weights Updating under the Impact of Construct Shift

Chen Zihao[1,2], Li Guangming[1]

(1. School of Psychology,Center for Studies of Psychological Application,South China Normal University,Guangzhou 510631;
2. Guangzhou Cadre and Talent Health Management Center,Guangzhou 510530)

Abstract:Construct invariance assumption is violated easily when conducting vertical scaling for more than four grade - groups. Construct shift can distort the developmental scale and lower the accuracy of parameter estimation. In order to find a better estimation method,this paper compared the performance of concurrent calibration and fixed - parameter calibration with multiple prior weights updating through a simulation experiment when construct shift exists. A 2(model:matched,mismatched) × 3(construct shift:0.25,0.5,1.0) × 3(sample size:500,1000,2000) common - item design was conducted,which including four grade - groupsResults indicate that construct shift has different effects on two estimated models. The FPC - MWU strategy effectively reduces errors caused by construct shift and has fewer base grade distance effect than concurrent calibration. In conclusion,practice reference is given for choosing most appropriate model and calibration method by considering the construct shift and sample size together.

Key words:vertical scaling;bifactor model;construct shift;fixed - parameter calibration;item response theory