

# 参数估计误差对多级评分题型测验等值的影响\*

王少杰<sup>1</sup>, 张敏强<sup>2</sup>, 黄菲菲<sup>3</sup>, 刘颖<sup>4</sup>

(1. 广东第二师范学院教育学院, 广州 510303; 2. 华南师范大学心理学院, 广州 510631;  
3. 广东技术师范大学教育科学学院, 广州 510665; 4. 广东第二师范学院教师教育学院, 广州 510303)

**摘要:**信息量加权特征曲线方法在二级评分题型测验等值中表现优异, 但未有研究探讨参数估计误差的作用。本文将其扩展到多级评分题型。通过模拟研究探讨参数估计误差、考生能力差异、题目数量对等值的影响。采用特征曲线与误差类指标评估等值表现。结果发现测验信息量加权特征曲线方法略优于传统方法, 其他方法与传统方法相当。参数估计误差与考生能力差异越小, 题目数量越大, 测验等值表现越优。偏差与方差权衡现象为测验等值提供新方向。

**关键词:**参数估计误差; 多级评分题型; 测验等值; 信息量加权; 特征曲线方法

**中图分类号:**B841.2

**文献标志码:**A

**文章编号:**1003–5184(2024)06–0550–09

## 1 引言

当考查相同内容的不同测验施测于不同考生群体时, 测验之间分数往往无法直接比较。测验等值是解决此问题的常用方法, 是指通过调整不同测验形式上的分数, 使其能够相互替代的统计过程 (Kolen & Brennan, 2014)。基于项目反应理论 (Item Response Theory, IRT) 的测验等值方法具有较高精确性与稳定性, 使其成为目前应用最广的方法之一 (Barrett & van der Linden, 2019; He & Cui, 2020; Manna & Gu, 2019; von Davier et al., 2019)。其中, 表现较为优异的方法主要包括 Haebara 方法与 Stocking–Lord 方法 (Haebara, 1980; Stocking & Lord, 1983), 它们又分别被称为项目特征曲线方法 (Item Characteristic Curve method, ICC) 与测验特征曲线方法 (Test Characteristic Curve method, TCC)。

但是, 特征曲线方法只考虑测验间项目或测验特征曲线相似性, 忽略参数估计误差, 从而会影响后续等值结果 (Barrett & van der Linden, 2019; Trierweiler et al., 2017) 以及相关决策的准确性与公平性。为此, Wang 等人 (2022) 将 IRT 信息量作为加权指标, 在二级评分题型测验等值中, 提出一类信息量加权特征曲线方法。经理论推导、模拟与实践验证, 该类方法具有优异表现, 可降低参数估计误差对测验等值的不良影响。但是, 该类方法在多级评分题型测验等值情境中的表现, 还未得到验证。同时, 鉴于该类方法理论初衷为降低参数估计误差的不利

影响, 探讨其在不同参数估计误差的测验等值情境中的表现, 具有理论意义。

另一方面, 国内多级评分题型测验等值领域的成果缺乏深入探讨。例如, 戴海崎 (2000) 推导过基于等级反应模型 (Graded Response Model, GRM) 的 ICC 方法公式。周骏等人 (2005) 在经济专业资格考试中, 使用过基于 GRM 的 ICC 等值方法。王菲等人 (2013) 采用 GRM 拟合语言考试数据, 比较过包括 ICC 在内的 6 种等值方法。众多国内大型考试均具有多级评分题, 亟需系统验证多级评分题型测验等值的表现与影响因素。

因此, 本文拟将信息量加权特征曲线方法扩展到多级评分题型测验等值情境中, 提出三种新的测验等值方法, 并基于此探讨参数估计误差等因素对多级评分题型测验等值的影响。按照逻辑顺序, 将首先介绍和扩展相应理论与方法, 然后阐述研究设计, 基于此开展模拟研究。最后, 围绕研究结果, 展开讨论并得出结论与建议。

## 2 理论基础

### 2.1 传统特征曲线方法

本文采用 GRM 拟合多级评分数据。能力为  $\theta_i$  的考生在题目  $j$  (共  $l$  题) 得到第  $k$  个等级 (共  $m$  等级) 分数的概率 (亦称运算特征曲线; Operating Characteristic Curve, OCC) 为  $p_{ijk}(\theta_i; a_j, b_j) = \frac{1}{1 + e^{-Da_j(\theta_i - b_{jk})}} - \frac{1}{1 + e^{-Da_j[\theta_i - b_{j(k+1)}]}}$ , 其中,  $D$  为常数

\* 基金项目: 广州市哲学社会科学共建项目 (2023GZCJ169)。

通信作者: 王少杰, E-mail: wang021112@126.com。

(本文为1),  $a_j$  为区分度,  $b_j$  为步骤难度向量, 分别为  $b_2, \dots, b_k, \dots, b_m$ 。当  $k = 1$  时,  $p_{ijk}(\theta_i; a_j, b_j) = 1 - \frac{1}{1 + e^{-Da_j(\theta_i - b_{j2})}}$ ; 当  $k = m$  时,  $p_{ijk}(\theta_i; a_j, b_j) = \frac{1}{1 + e^{-Da_j(\theta_i - b_{jm})}}$  (Hori et al., 2022)。

在多级评分题型测验等值中, 特征曲线方法可分为运算特征曲线方法 (Operating Characteristic Curve method, OCC)、ICC 方法与 TCC 方法 (Zhang, 2021a, 2021b)。根据锚题 (锚测验) 等值前与等值后的特征曲线的不变性, 构造出相应的损失函数:  $OCC = \sum_i w_i \sum_{j:V} \sum_k (p_{ijk,old} - p_{ijk,linked})^2$ ,  $ICC = \sum_i w_i \sum_{j:V} (\sum_k W_{jk} p_{ijk,old} - \sum_k W_{jk} p_{ijk,linked})^2$ ,  $TCC = \sum_i w_i (\sum_{j:V} \sum_k W_{jk} p_{ijk,old} - \sum_{j:V} \sum_k W_{jk} p_{ijk,linked})^2$ , 其中,  $w_i = 1$ , 将测验 Y 量纲上与测验 X 经等值转换到测验 Y 量纲上的参数分别标记为 old 与 linked, V 为锚测验。

在测验等值中, 求损失函数最小值所对应的等值系数, 这便是特征曲线方法求解等值系数的思路。由于损失函数较为复杂, 很难找到它的解析解。所以, 本研究利用损失函数的非负性, 采用 R 软件 (R Core Team, 2019) 中的最优化函数求解, 具体为拟牛顿算法 (BFGS 算法)。

## 2.2 信息量加权特征曲线方法

根据 Wang 等人 (2022) 在二级评分题型测验等值中提出的信息量加权特征曲线方法思路, 将其扩展到多级评分题型。在 GRM 中, 除题目与测验信息量之外, 还可得到步骤难度的信息量, 即等级信息量。因此, 可依次得到三种新的测验等值方法, 分别为等级信息量加权特征曲线方法 (Category information Weighted Characteristic Curve method, CWCC)、项目信息量加权特征曲线方法 (Item information Weighted Characteristic Curve method, IWCC) 与测验信息量加权特征曲线方法 (Test information Weighted Characteristic Curve method, TWCC)。它们分别是在传统特征曲线方法 (OCC 方法、ICC 方法与 TCC 方法) 基础上, 按照等级信息量、项目信息量与测验信息量对损失函数加权处理得到。与传统特征曲线方法相同, 信息量加权特征曲线方法也是根据锚题 (锚测验) 等值前与等值后的特征曲线的不变性, 构造出相应的损失函数:  $CWCC = \sum_i w_i \sum_{j:V} \sum_k (p_{ijk,old}$

$$- p_{ijk,linked})^2 (CIF_{ijk,old} + CIF_{ijk,linked}), IWCC = \sum_i w_i \sum_{j:V} (\sum_k W_{jk} p_{ijk,old} - \sum_k W_{jk} p_{ijk,linked})^2 (IIF_{ij,old} + IIF_{ij,linked}), TWCC = \sum_i w_i (\sum_{j:V} \sum_k W_{jk} p_{ijk,old} - \sum_{j:V} \sum_k W_{jk} p_{ijk,linked})^2 (TIF_{i,old} + TIF_{i,linked}),$$

其中,  $CIF_{ijk}(\theta_i; a_j, b_j) = \frac{[p'_{ijk}(\theta_i; a_j, b_j)]^2}{p_{ijk}(\theta_i; a_j, b_j)}, IIF_{ij}(\theta_i; a_j, b_j) = \sum_k \frac{[p'_{ijk}(\theta_i; a_j, b_j)]^2}{p_{ijk}(\theta_i; a_j, b_j)}, TIF(\theta_i; a_j, b_1, \dots, b_j, \dots, b_m) = \sum_j \sum_k \frac{[p'_{ijk}(\theta_i; a_j, b_j)]^2}{p_{ijk}(\theta_i; a_j, b_j)}, p'_{ijk}(\theta_i; a_j, b_j)$  为  $p_{ijk}(\theta_i; a_j, b_j)$  对能力参数  $\theta$  的一阶导数。

在多级评分题型测验等值情境中, 相较于传统特征曲线方法, 信息量加权特征曲线方法的损失函数更为复杂, 更难找到损失函数最小值所对应的等值系数的解析解。所以, 本研究利用损失函数的非负性, 采用 R 软件 (R Core Team, 2019) 中的最优化函数求解, 具体为拟牛顿算法 (BFGS 算法)。

## 2.3 IRT 参数估计误差

IRT 参数估计误差水平存在多种设置与操纵方式。一种为直接模拟分布情况。例如, Li 和 Lissitz (2004) 采用 delta 法计算出参数估计方差协方差矩阵, 然后在相应多元分布中, 随机抽取符合特定估计误差的参数值。该方式最接近现实分布情况, 但需要计算方差协方差矩阵, 操作复杂。而 Kaskowitz 和 De Ayala (2001) 将二参数与三参数 Logistic 模型的低、中与高水平参数估计误差分别设定为 (0.11、0.25、0.63) 和 (0.30、0.91、2.73), 然后再构造随机分布并抽取具有误差的参数估计值。但该方法忽略了不同参数估计误差存在较大差异。

另一方面, IRT 参数估计误差主要来源于考生样本抽样; 考生样本量越大, 参数估计越准确 (Kolen & Brennan, 2014)。因此, 其他条件不变, 可将考生人数作为误差水平间接指标。例如, Li 和 Lissitz (2004) 基于不同考生人数 (1000、2000、4000), 计算出渐近协方差矩阵并将其与参数估计误差对应。综上, 通过变化考生样本量间接操纵参数估计误差, 简单且易于理解。同时, 在 IRT 等值量尺转换中, 计算等值系数并不涉及考生能力, 变化考生人数仅影响因考生样本抽样导致的参数估计误差, 并不存在其他因素对测验等值的间接影响, 可排除额外因素干扰, 操纵效果较为明显。综上, 本文拟采用此方式, 通过变化考生人数, 间接操纵参数估计误差。但该

方式的合理性与有效性,需通过预研究验证。

### 3 预研究

#### 3.1 目的

采用蒙特卡洛模拟,验证本文参数估计误差操纵方式的合理性与有效性,即,通过变化考生人数改变 IRT 参数估计误差,从而保证正式研究的严谨性与可靠性。

#### 3.2 自变量

预研究自变量为考生人数,为 500、1000、2000 与 4000。1000、2000 与 4000 人分别代表 IRT 测验等值研究的小、中与大样本情境(De Ayala et al., 2018; Marcq & Andersson, 2022),分别对应参数估计误差较大、适中与较小。纳入考生人数 500 这一水平,以探索在误差较为极端情境中测验等值表现(Kolen, 2020; König et al., 2020; Peabody, 2020)。

#### 3.3 评价指标

偏差(Bias)、标准误(Standard Error, SE)与均方根误差(Root Mean Square Error, RMSE),分别代表系统误差(systematic error)、随机误差(random error)与总误差(total error)。计算所有题目误差指标

$$\text{均值, } Bias_{average}(\hat{\lambda}) = \frac{\sum_{j=1}^{N_{item}} Bias(\hat{\lambda}_j)}{N_{item}}, SE_{average}(\hat{\lambda}) = \frac{\sum_{j=1}^{N_{item}} SE(\hat{\lambda}_j)}{N_{item}} \text{ 与 } RMSE_{average}(\hat{\lambda}) = \frac{\sum_{j=1}^{N_{item}} RMSE(\hat{\lambda}_j)}{N_{item}}, \text{ 其}$$

$$\text{中, } Bias(\hat{\lambda}_j) = \frac{1}{R} \sum_{r=1}^R \hat{\lambda}_{jr} - \lambda_j, SE(\hat{\lambda}_j) = \frac{1}{R} \sum_{r=1}^R \left[ \hat{\lambda}_{jr} - \frac{1}{R} \sum_{r=1}^R \hat{\lambda}_{jr} \right]^2, RMSE(\hat{\lambda}_j) = \sqrt{Bias^2(\hat{\lambda}_j) + SE^2(\hat{\lambda}_j)}, \hat{\lambda}_{jr} \text{ 为在第 } r \text{ 次重复中, 题目 } j \text{ 参数估计值, } \lambda_j \text{ 为其真值; } R \text{ 为重复次数 } 5000, N_{item} \text{ 为题目数量 } 50。 \text{ 指标值越小代表参数估计误差越小。}$$

#### 3.4 研究流程

除模拟重复次数(5000 次)与参数设定外,预研究与正式研究流程类似,不再赘述,可参考“4.4 研究流程”部分。为方便误差计算, $a$  在  $[1.02, 2]$  范围内,每隔 0.02 取一值; $b_2$  在  $[-1, -0.02]$  范围内,每隔 0.02 取一值; $b_3$  在  $[0.02, 1]$  范围内,每隔 0.02 取一值。

#### 3.5 结果

总体而言,GRM 参数估计误差较小。题目参数  $a$ 、 $b_2$  与  $b_3$  误差均在可接受范围内。除 Bias 外,SE 与 RMSE 均随考生人数的增加而逐渐降低。相较于 SE 与 RMSE,考生人数变化对 Bias 指标基本无影响(除  $a$  外)。绝大部分 RMSE 来源于 SE,即,在 GRM 参数估计中,随机误差影响最大(对总误差贡献最大)。因此,将考生人数作为 GRM 参数估计误差间接指标,较为有效与合理。

表 1 参数估计误差

考生人数	Bias			SE			RMSE		
	$a$	$b_2$	$b_3$	$a$	$b_2$	$b_3$	$a$	$b_2$	$b_3$
500	-0.0047	-0.0069	0.0076	0.1401	0.0984	0.0954	0.1402	0.0987	0.0958
1000	-0.0102	-0.0037	0.0088	0.0980	0.0690	0.0670	0.0985	0.0692	0.0677
2000	-0.0124	-0.0030	0.0084	0.0692	0.0487	0.0473	0.0703	0.0489	0.0482
4000	-0.0130	-0.0039	0.0064	0.0489	0.0343	0.0335	0.0506	0.0346	0.0342

### 4 正式研究

#### 4.1 目的

基于预研究结论,通过蒙特卡洛模拟,探讨参数估计误差等因素对多级评分题型测验等值的影响,并验证信息量加权特征曲线方法的表现。

#### 4.2 自变量

(1)考生能力差异:分为无、较小与适中。参加测验 Y 的考生  $\theta \sim N(0, 1)$ , 参加测验 X 的考生分别为  $\theta \sim N(0, 1)$ 、 $\theta \sim N(0.2, 1.1^2)$  与  $\theta \sim N(0.5, 1.2^2)$ , 代表两组考生无、较小与适中三种能力差异(Andersson, 2018)。

(2)参数估计误差:分为较小、适中、较大与极端。通过变化考生人数达到间接操纵 IRT 参数估计误差目的。四种误差水平分别对应考生人数 4000、2000、1000 与 500(De Ayala et al., 2018)。

(3)题目数量:分为 25、50 与 100。结合相关文献与国内教育考试情况,25、50 与 100 分别代表题目数量较少、适中与较多情境(De Ayala et al., 2018; Wallin et al., 2021)。锚题占题目总数的 20%(Kolen & Brennan, 2014; Manna & Gu, 2019),均为内锚。

(4)测验等值方法:分为传统方法与信息量加

权特征曲线方法两类,前者包括 OCC 方法、ICC 方法与 TCC 方法,后者包括 CWCC 方法、IWCC 方法与 TWCC 方法。

因此,本研究操纵自变量水平总数为  $3 \times 4 \times 3 \times 6 = 216$ 。

#### 4.3 评价指标

评价指标包含测验特征曲线类与误差类。测验特征曲线类指标可整体评估测验等值一般表现,误差类指标可单独反映测验等值方法对等值系数的返真性。

(1)测验特征曲线类:该类指标从测验 X 等值后与等值前的测验特征曲线差异角度考虑,为相对测验特征曲线平方差(Relative Squared TCC Difference, RSTD)(Kim,2010;Kim & Kolen,2007)。

$$RSTD = \frac{\sum_{r=1}^R \sum_{i=1}^I [(TCC_{r,A}(\theta_i) - TCC_{r,B}(\theta_i))^2 f(\theta_i)]}{R \times N_{score}^2}, \text{ 其}$$

中,  $TCC_{r,B}(\theta_i)$  与  $TCC_{r,A}(\theta_i)$  分别为在第  $r$  次重复中,基于测验 X 等值前(真值)、等值后(经等值转换的估计值)的题目参数计算出的测验特征曲线;  $\theta_i$  在  $[-4, 4]$  范围内每隔 0.05 取一值,  $I = 161$ , 且  $f(\theta_i) = 1$ ;  $N_{score}$  为测验满分,分别为 50、100 与 200。表现良好的测验等值方法, RSTD 应接近或等于 0 (Lee & Ban,2009;Trierweiler et al.,2017)。

(2)误差类:第二类评价指标包括 Bias、SE 与 RMSE(Manna & Gu,2019)。

$$Bias(\hat{\lambda}) = \frac{1}{R} \sum_{r=1}^R \hat{\lambda}_r - \lambda, SE(\hat{\lambda}) = \sqrt{\frac{1}{R} \sum_{r=1}^R \left[ \hat{\lambda}_r - \frac{1}{R} \sum_{r=1}^R \hat{\lambda}_r \right]^2}, RMSE(\hat{\lambda})$$

$= \sqrt{Bias^2(\hat{\lambda}) + SE^2(\hat{\lambda})}$ , 其中,  $\hat{\lambda}_r$  为在第  $r$  次重复中,等值系数 A(或 B)的估计值,  $\lambda$  为其真值。参加测验 X 的考生  $\theta \sim N(0, 1)$ 、 $\theta \sim N(0.2, 1.1^2)$  与  $\theta \sim N(0.5, 1.2^2)$  时,等值系数真值分别为  $A = 1$  与  $B = 0$ 、 $A = 1.1$  与  $B = 0.2$ 、 $A = 1.2$  与  $B = 0.5$ 。该指标绝对值越小代表测验等值方法表现越好。

#### 4.4 研究流程

除参数估计(mirt包;Chalmers,2012)外,其余均采用 R 软件(R Core Team,2019)自编程序完成。

主要分为:(1)随机抽取参加测验 X 与 Y 的考生能力  $\theta$ 。(2)随机抽取测验 X 与 Y 的独立题与锚题的参数值,  $a \sim U(1, 2)$ ,  $b_2 \sim N(-0.5, 1)$ ,  $b_3 \sim N(0.5, 1)$ , 且同一题目  $b_2 < b_3$ 。(3)生成考生作答矩阵。(4)采用测验等值方法计算等值系数 A 与 B。(5)重复上述过程 1000 次,然后计算评价指标。

#### 4.5 正式研究结果

首先,在三种不同考生能力差异情境中,测验等值趋势与比较的结果一致;而在同等条件下,考生能力差异越大,测验等值误差越大。因此,下文以能力差异适中的情境为例阐述。

从 RSTD 角度,各测验等值方法均可较好地匹配测验等值前与等值后的特征曲线形态(保留四位小数,均为 0,未呈现),表现较为优异。

图 1 为等值系数的 Bias。对于等值系数 A,三种传统方法与三种信息量加权方法的结果差异不大,仅有 TWCC 方法表现略优于 TCC 方法。而在三种信息量加权特征曲线方法中, TWCC 方法表现亦较优,但无较大差异。当参数估计误差变大时, Bias 绝对值也随之增加。题目数量对测验等值方法的影响,因参数估计误差水平的不同而存在一定差异。具体而言,当参数估计误差较小与适中时,扩大题目数量基本不影响测验等值方法的表现;当参数估计误差较大时,扩大题目数量使得 Bias 先增加,后平稳;而当存在较极端的参数估计误差时,扩大题目数量可显著降低 Bias。对于等值系数 B,传统方法与信息量加权方法的结果同样差异不大。但参数估计误差与题目数量对结果的影响,表现出一定差异。具体而言,当参数估计误差变大时, Bias 绝对值稍有降低。当参数估计误差较小时,扩大题目数量基本不影响各测验等值方法的表现;当参数估计误差适中时,扩大题目数量使得 Bias 先增加,后减少;当参数估计误差较大与极端时,扩大题目数量使得 Bias 先减少,后增加。此外,当参数估计误差较小、适中与较大时,等值系数 B 的误差值略大于 A。而当存在极端的参数估计误差时,等值系数 A 的误差值略大于 B。明显存在对等值系数 A 与 B 的高估现象。

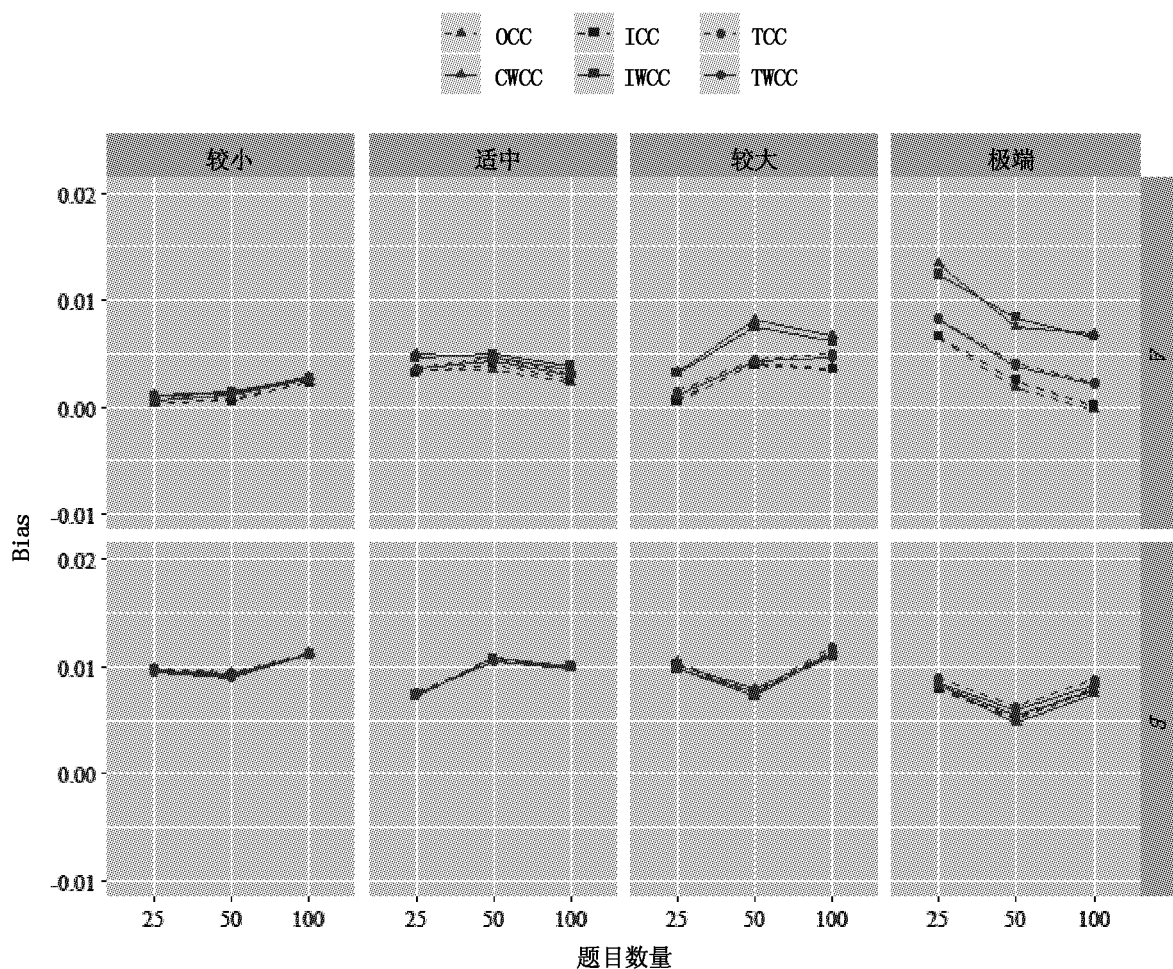


图1 Bias(能力差异适中)

注:图的横行代表测验等值系数,纵列代表参数估计误差,分别为较小、适中、较大与极端;子图的横轴代表题目数量,纵轴代表 Bias。

图2为等值系数的SE。相较于Bias(图1),SE(随机误差)均较大。对于等值系数A与B,TWCC方法的表现略优于TCC方法,其他方法的表现差异不大。而在三种信息量加权特征曲线方法中,TWCC方法的表现较优。当参数估计误差变大时,SE也随之增加。扩大题目数量,可降低SE,此现象在等值系数A中最为明显。此外,等值系数A与B的SE差异并不明显。

结合上述系统误差(Bias,图1)与随机误差(SE,图2)结果,可推测,由于SE较大,RMSE(总误差)主要取决于SE。图3为等值系数的RMSE。对于等值系数A与B,TWCC方法的表现优于TCC方法,其他方法的结果差异不大。而在三种信息量加权方法中,TWCC方法的表现亦较优。当参数估计误差变大时,RMSE也随之增加。扩大题目数量,可降低RMSE,这在等值系数A中最为明显。此外,等

值系数A与B的RMSE差异较小。

## 5 讨论

### 5.1 信息量加权特征曲线方法的表现

TWCC方法表现略优于TCC方法,其他方法表现基本相当。而在三种信息量加权特征曲线方法中,TWCC方法表现略优。根据信息量加权特征曲线方法的理论阐释与预研究,参数估计误差影响等值系数计算。因此,将反映参数估计误差的指标(信息函数)纳入损失函数中,可改善测验等值方法的表现(Barrett & van der Linden,2019)。以上结果证明将信息量加权的思路从二级扩展到多级评分题型是可行的。但另一方面,传统方法表现亦可圈可点,尤其是在RSTD指标上,达到了比较高的精度(全部为0)。从误差类指标角度,传统方法估计值与真值的差异同样较小。因此,本研究也验证了传统方法的稳健性。

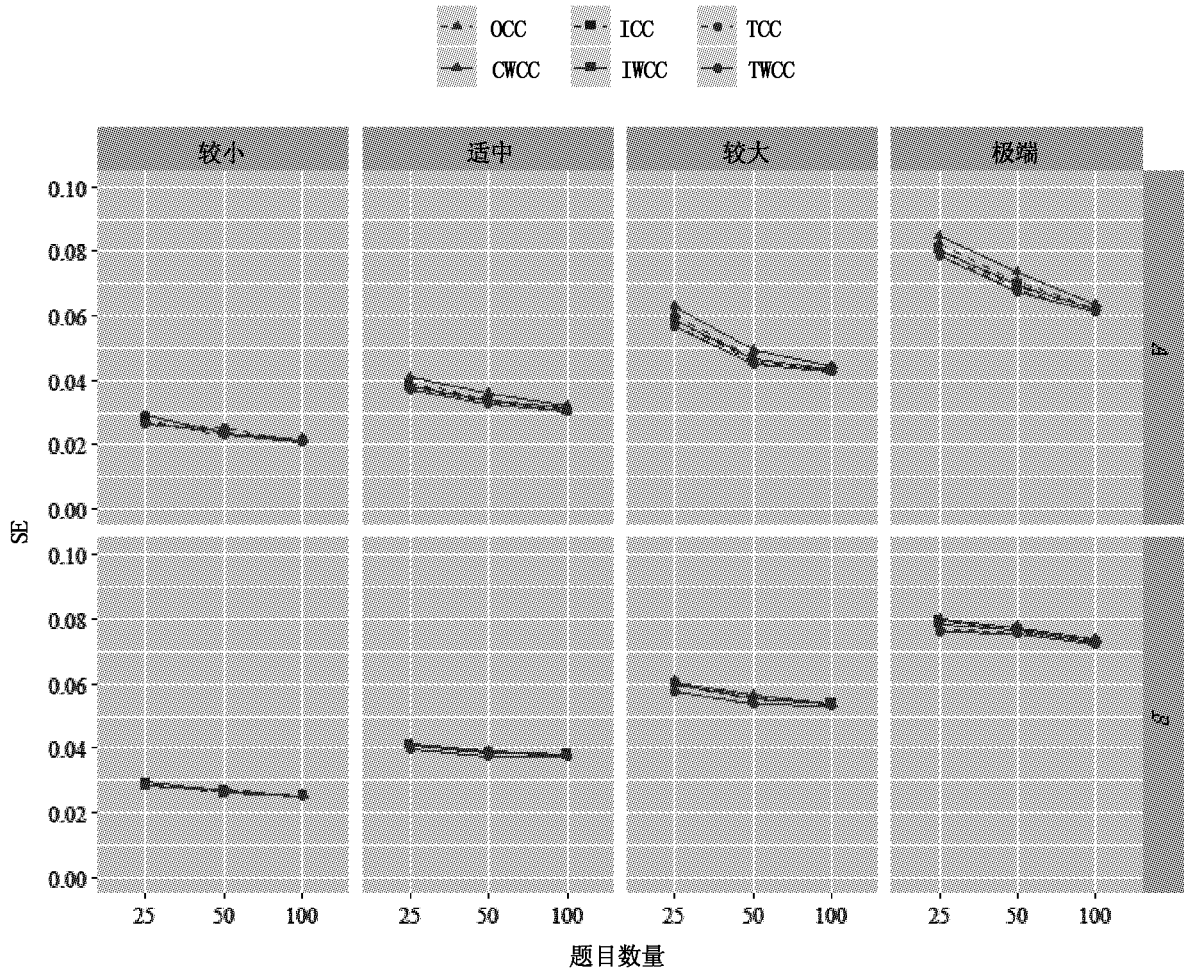


图2 SE(能力差异适中)

注:图的横行代表测验等值系数,纵列代表参数估计误差,分别为较小、适中、较大与极端;子图的横轴代表题目数量,纵轴代表SE。

需注意,传统方法与信息量加权特征曲线方法在 RSTD 指标上均为0,并不代表两类方法的表现不存在差异。出现该现象的关键并不在于两类方法的区别,而在于其共性。RSTD 代表测验 X 等值后与等值前的测验特征曲线差异。而特征曲线方法正是使测验等值前与后的特征曲线差异最小化(Hacbara,1980;Stocking & Lord,1983)。因此,它们在该指标上不存在差异,恰好验证这类测验方法优异表现。相关研究也同样发现该现象(Lee & Ban,2009;Trierweiler et al.,2017)。

5.2 影响因素

5.2.1 参数估计误差与题目数量

首先,纳入 IRT 参数估计误差(信息量)对损失函数进行加权处理,是信息量加权特征曲线方法的

优势。探讨参数估计误差对测验等值的影响兼具理论与实践意义。研究发现,参数估计误差变大,各测验等值方法表现明显恶化(误差增加),且信息量加权特征曲线方法较传统方法的优势并不存在明显提升。从考生人数角度解读此结论,即,考生人数减少,各测验等值方法误差增加。IRT 参数估计是基于样本推测总体的统计推断过程,测验等值方法 SE 的主要来源,便是样本代表性(Andersson,2018;Kolen & Brennan,2014;Manna & Gu,2019)。因此,不考虑其他因素,较大样本量(较小参数估计误差)是测验等值结果稳定性(即较小 SE)的充分条件(王少杰等,2022;Kim,2006;Kolen & Brennan,2014;Liu,2020)。

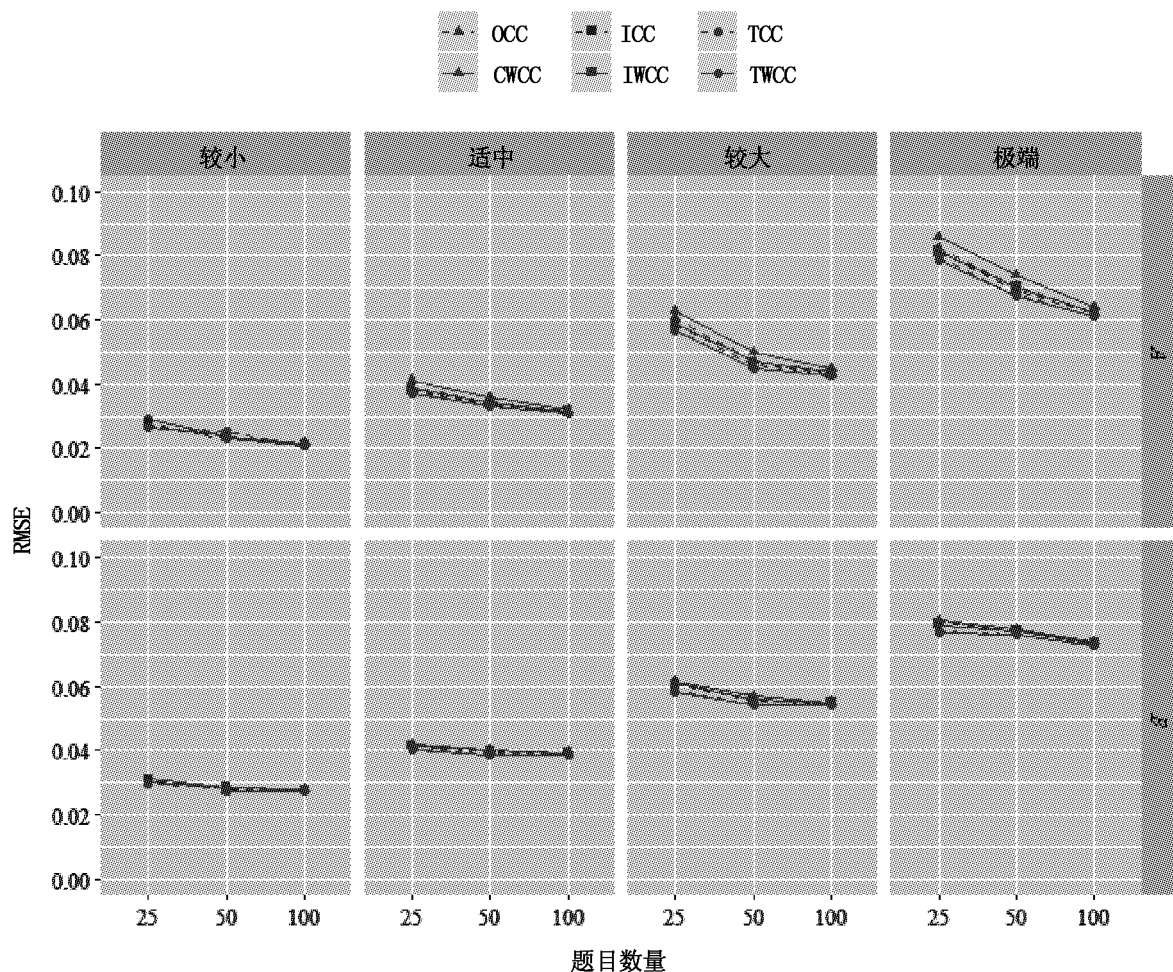


图3 RMSE(能力差异适中)

注:图的横行代表测验等值系数,纵列代表参数估计误差,分别为较小、适中、较大与极端;子图的横轴代表题目数量,纵轴代表RMSE。

虽然本文并未探讨由题目抽样导致的参数估计误差,但由于题目数量对测验等值方法影响较大,故将其纳入模拟研究。结果发现,增加题目数量基本不影响测验等值系统误差,却可使其随机误差与总误差显著降低。一方面,不考虑其他因素,测验包含题目数量越多,信度越高,而信度是反映测量稳定性(误差)的关键指标,因此增加题目数量可降低随机误差。另一方面,IRT测验等值方法需依据锚测验题目参数计算等值系数。而题目数量增加,所提供的参数信息也相应增加。因此,增加题目数量也可降低测验等值结果的不确定性(随机误差)(Kang & Petersen, 2012; Kaskowitz & De Ayala, 2001; Kim, 2006)。

同时,参数估计误差与题目数量对测验等值的影响,也存在边际递减效应。随着参数估计误差的降低,测验等值表现得到一定程度改善。此时,如果使参数估计误差或题目数量再次降低或增加相同比

例,测验等值方法的改善程度却在明显缩小(De Ayala et al., 2018)。由于本研究等值误差均较小(Roberts et al., 2003),题目数量在改善其表现方面作用较小(Kim, 2010)。“过犹不及”可较为恰当概括该现象。它提醒我们,在IRT测验等值实践中,应谨慎设定符合要求的考生人数与题目量。当低于该要求时,测验等值结果误差较大,结论可靠性较差;而当明显超过该要求时,虽然准确性与稳定性均有一定程度改善,但付出成本明显大于收益。

在IRT测验等值中,随机误差明显大于系统误差。这使得相较于系统误差,从参数估计随机误差入手,可更好地提升测验等值表现。该现象在模拟研究中得到验证。信息量加权特征曲线方法试图降低参数估计随机误差对测验等值的影响,从而改善其表现。但另一方面,正如信度为效度的必要条件,降低参数估计随机误差使得测验等值稳定性明显提高的同时,也可一定程度改善测验等值表现准确性。



此外,在预研究中,操纵题目数量在使得参数估计随机误差明显变化的同时,也使得其系统误差出现变化。综上,本文在描述参数估计误差时,并未简单将其局限为随机误差,而是将其一般化为参数估计误差。

### 5.2.2 考生能力差异

正因为参加测验 X 与 Y 的考生存在能力差异,在 IRT 测验等值中,将处于不同量尺的参数进行链接处理尤为重要。考生能力越接近,测验等值结果相似性越高 (Sinharay & Holland, 2010)。假如两组考生能力分布相同或相似 (例如单组与随机组设计),无需测验等值处理,各参数与分数便可直接比较 (Kolen & Brennan, 2014)。本研究表明,随着两组考生能力差异变大,测验等值误差稍微增加,但并未影响结果,即,各测验等值方法在具有不同考生能力差异的情境中表现相似。这可能是因为本研究分别设定  $\theta \sim N(0,1)$ 、 $\theta \sim N(0.2,1.1^2)$  与  $\theta \sim N(0.5,1.2^2)$  三种能力情境,对应能力差异 (平均值差值分别为 0、0.2 与 0.5,标准差差值分别为 0、0.1 与 0.2) 较为典型,并未涉及极端情况。例如, Lee 与 Ban (2009) 探讨过  $\theta \sim N(0,1)$  分别与  $\theta \sim N(0,1)$ 、 $\theta \sim N(0.5,1)$ 、 $\theta \sim N(0.8,1)$  和  $\theta \sim N(1,1)$  的差异组合 (平均值差值最大为 1)。Kim 和 Kolen (2007) 模拟过  $\theta \sim N(0,1)$  与几种不同偏态分布。结果均发现,当两组考生能力分布的形态与差异变大时,IRT 测验等值表现变差。

### 5.3 偏差与方差的权衡

在本研究中,测验等值 Bias 绝对值明显小于 SE,因此后者对 RMSE 贡献最大,即,降低系统误差是以提高随机误差为代价。随着模型复杂度提高,偏差不断降低,而方差却在持续增加,即“偏差与方差的权衡” (Bias - Variance tradeoff) 现象 (Mohr et al., 2018)。当模型较为简单时,会出现其与数据的欠拟合 (underfitting); 而当模型较为复杂时,则会出现其与数据的过拟合 (overfitting)。偏差与方差的权衡同样也可用于解释测验等值表现。考虑锚题参数的矩估计方法,只涉及平均数与标准差的运算;同时考虑项目 (或测验) 特征曲线的传统方法,需求特征曲线差异的最优解;同时考虑 (运算、项目或测验) 特征曲线与 (等级、项目或测验) 信息量的信息量加权特征曲线方法,要解同时包含特征曲线差异与信息量二者的乘积最优化问题。求解等值系数的函数愈发复杂,模型复杂度不断增加。因此,不论是传统方法,还是信息量加权特征曲线方法,其 SE 均大于相应 Bias 绝对值。根据偏差与方差权衡规律,

欲寻找损失函数最优解,需要在刻画损失函数复杂程度的区间上寻找一点,使得该点所对应的测验等值总误差最小 (Kim, 2010)。基于此,在测验等值的方差与偏差权衡中,通过适当增加模型复杂程度,信息量加权特征曲线方法可作为较可靠的局部最优解。同时,信息量加权特征曲线方法亦存在改善空间,以进一步降低其测验等值误差。

## 6 结论

本文将信息量加权特征曲线方法扩展到多级评分题型测验等值情境,并通过模拟研究,验证参数估计误差等因素对测验等值的影响。结果发现,新提出的 TWCC 方法略优于 TCC 方法, CWCC、IWCC 方法与传统特征曲线方法表现相当。在影响因素方面,参数估计误差与考生能力差异越小,题目数量越大,测验等值表现越优。尤其是参数估计误差,在测验等值中发挥重要作用,为提升等值表现提供新颖视角。本研究亦发现偏差与方差的权衡现象,可基于此探讨如何更好提升与服务测验等值的理论与实践。

## 参考文献

- 戴海崎. (2000). 等级反应模型项目特征曲线法等值研究. *心理学探新*, 20(3), 49 - 53.
- 王非, 任杰, 张泉慧, 曹文静. (2013). 等级记分模型下几种等值方法的比较研究. *中国考试*, (6), 10 - 17.
- 王少杰, 张敏强, 黄菲菲, 黄丽芳, 袁琪婷. (2022). 项目反应理论观察分数核等值的影响因素. *心理科学*, 45(4), 988 - 997.
- 周骏, 欧东明, 徐淑媛, 戴海琦, 漆书青. (2005). 等级反应模型下项目特征曲线等值法在大型考试中的应用. *心理学报*, 37(6), 126 - 132.
- Andersson, B. (2018). Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Applied Psychological Measurement*, 42(3), 192 - 205.
- Barrett, M. D., & van der Linden, W. J. (2019). Estimating linking functions for response model parameters. *Journal of Educational and Behavioral Statistics*, 44(2), 180 - 209.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1 - 29.
- De Ayala, R. J., Smith, B., & Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Applied Psychological Measurement*, 42(2), 155 - 168.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22(3), 144 - 149.



- He, Y. , & Cui, Z. (2020). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*, 44(4), 296 – 310.
- Hori, K. , Fukuhara, H. , & Yamada, T. (2022). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(2), e1531.
- Kang, T. , & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review*, 13(2), 311 – 321.
- Kaskowitz, G. S. , & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement*, 25(1), 39 – 52.
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43(4), 355 – 381.
- Kim, S. (2010). An extension of least squares estimation of IRT linking coefficients for the graded response model. *Applied Psychological Measurement*, 34(7), 505 – 520.
- Kim, S. , & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, 32(4), 371 – 397.
- Kolen, M. J. (2020). Equating with small samples (commentary). *Applied Measurement in Education*, 33(1), 77 – 82.
- Kolen, M. J. , & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.
- König, C. , Spoden, C. , & Frey, A. (2020). An optimized Bayesian hierarchical two – parameter logistic model for small – sample item calibration. *Applied Psychological Measurement*, 44(4), 311 – 326.
- Lee, W. C. , & Ban, J. C. (2009). A comparison of IRT linking procedures. *Applied Measurement in Education*, 23(1), 23 – 48.
- Li, Y. H. , & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement*, 41(2), 85 – 117.
- Liu, R. (2020). Addressing score comparability in diagnostic classification models: An observed – score equating and linking approach. *Behaviormetrika*, 47(1), 55 – 80.
- Manna, V. F. , & Gu, L. (2019). Different methods of adjusting for form difficulty under the Rasch model: Impact on consistency of assessment results. *ETS Research Report Series*, (1), 1 – 18.
- Marcq, K. , & Andersson, B. (2022). Standard Errors of Kernel Equating: Accounting for Bandwidth Estimation. *Applied Psychological Measurement*, 46(3), 200 – 218.
- Mohri, M. , Rostamizadeh, A. , & Talwalkar, A. (2018). *Foundations of machine learning*. MIT Press.
- Peabody, M. R. (2020). Practical issues in linking and equating with small samples. *Applied Measurement in Education*, 33(1), 1 – 2.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Roberts, J. S. , Bao, H. , Huang, C. W. , & Gagne, P. (2003, April). *Exploring alternative characteristic curve approaches to linking parameter estimates from the generalized partial credit model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois.
- Sinharay, S. , & Holland, P. W. (2010). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, 47(3), 261 – 285.
- Stocking, M. L. , & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201 – 210.
- Trierweiler, T. J. , Lewis, C. , & Smith, R. L. (2017, July). *Reducing conditional error variance differences in IRT scaling*. Paper presented at the annual meeting of the Psychometric Society, Zurich, Switzerland.
- von Davier, M. , Yamamoto, K. , Shin, H. J. , Chen, H. , Khorramdel, L. , Weeks, J. , . . . Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000 – 2012. *Assessment in Education: Principles, Policy & Practice*, 26(4), 466 – 488.
- Wallin, G. , Häggström, J. , & Wiberg, M. (2021). How Important is the Choice of Bandwidth in Kernel Equating? *Applied Psychological Measurement*, 45(7 – 8), 518 – 535.
- Wang, S. , Zhang, M. , Lee, W. , Huang, F. , Li, Z. , Li, Y. , & Yu, S. (2022). Two IRT characteristic curve linking methods weighted by information. *Journal of Educational Measurement*, 59(4), 423 – 441.
- Zhang, Z. (2021a). Asymptotic standard errors of generalized partial credit model true score equating using characteristic curve methods. *Applied Psychological Measurement*, 45(5), 331 – 345.
- Zhang, Z. (2021b). Asymptotic standard errors of parameter scale transformation coefficients in test equating under the nominal response model. *Applied Psychological Measurement*, 45(2), 134 – 138.

(下转第 565 页)

- IEA. (2019). *PIRLS 2021 Assessment Frameworks*. Boston: TIMSS & PIRLS International Study Center.
- Marilyn, K., & Peter, K. (2014). *Literacy and Language in East Asia*. Singapore: Springer.
- OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.
- OECD. (2021). *PISA 2025 Foreign Language Assessment Framework*. Paris: OECD Publishing.
- PCAP. (2016). *Pan – Canadian Assessment Program (PCAP) 2016 Assessment Framework*. Toronto: Council of Ministers of Education, Canada.
- Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition*, 13(2), 275 – 298.
- Tseng, W. T., & Gao, X. (2021). Individual differences in second language learning: The road Ahead. *English Teaching & Learning*, 45(3), 237 – 244.

## Development, Reliability and Validity of a Questionnaire for English Reading Literacy in Middle Schools

Li Weilai<sup>1</sup>, Kang Shumin<sup>2</sup>

(1. School of Educational Sciences, Taizhou University, Taizhou 225300;

2. College of Foreign Languages, Qufu Normal University, Jining 273165)

**Abstract:** Reading literacy is an important part of the cultivation of English core literacy. The evaluation of middle school students' English reading literacy has distinct characteristics of learning stage and students' age. Referring to the national curriculum standards, this research firstly develops a questionnaire for English reading literacy in middle schools, with 6 dimensions and 14 items. The 806 data from two middle schools were used to validate the questionnaire. The questionnaire was analyzed through exploratory factor analysis, confirmatory factor analysis, reliability and validity test. The results shows that the questionnaire has high reliability and validity, which can be used as an evaluation instrument to assess students' English reading literacy at an intermediate level in China.

**Key words:** middle school English; reading literacy; evaluation questionnaire; reliability and validity

(上接第 558 页)

## The Impact of Parameter Estimation Error on IRT Linking Methods with Polytomous Items

Wang Shaojie<sup>1</sup>, Zhang Minqiang<sup>2</sup>, Huang Feifei<sup>3</sup>, Liu Ying<sup>4</sup>

(1. School of Education, Guangdong University of Education, Guangzhou 510303; 2. School of Psychology,

South China Normal University, Guangzhou 510631; 3. School of Educational Science, Guangdong

Polytechnic Normal University, Guangzhou 510665; 4. School of Teacher Education,

Guangdong University of Education, Guangzhou 510303)

**Abstract:** The information – weighted characteristic curve methods have shown excellent performance in IRT linking with dichotomous items. However, few researches explore the effect of parameter estimation error on test linking. This paper extends the information – weighted characteristic curve methods to IRT linking with polytomous items and explores the effects of parameter estimation error, ability differences, and test length on linking through simulation studies. IRT linking performance was evaluated using indices related to characteristic curves and errors. The results indicated that the information – weighted characteristic curve methods performed slightly better than the traditional characteristic curve methods, while other new methods performed as well as the traditional methods. The linking performance was better when the parameter estimation error and ability differences were smaller, and the test was longer. The bias and variance tradeoff provides a new direction for test linking and equating.

**Key words:** Parameter estimation error; polytomous item; IRT linking; information weighted; characteristic curve methods