Advances in Psychology 心理学进展, 2026, 16(7), 213-224
Published Online July 2026 in Hans. https://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2026.167353

Hans汉斯

准确性与公平性视角下异常锚题对测验等值的影响

王少杰

广东第二师范学院教育学院，广东 广州

收稿日期：2026年6月3日；录用日期：2026年7月2日；发布日期：2026年7月15日

摘　要

在项目反应理论参数链接和等值过程中，锚题起着至关重要的作用。本研究在非等组锚测验设计下，采用蒙特卡洛方法，操纵异常锚题比例、区分度以及难度参数漂移程度，考查异常锚题对测验等值的影响。评估指标包括链接系数的偏差与均方根误差、等值分数的绝对偏差与加权绝对偏差以及分类准确性和一致性的绝对偏差。结果表明，在链接系数和等值分数方面，相较于传统特征曲线方法，信息量加权方法更优。此外，各方法在分类准确性和一致性方面均表现出较高可靠性。

关键词

项目反应理论，信息量加权，特征曲线方法，异常锚题，分类准确性与一致性

The Effect of Outlying Common Items on Test Equating: A Perspective on Accuracy and Equity

Shaojie Wang

School of Education, Guangdong University of Education, Guangzhou Guangdong

Received: June 3, 2026; accepted: July 2, 2026; published: July 15, 2026

Abstract

Common items are crucial in item response theory (IRT) linking and equating. Within a common item nonequivalent groups design, this study utilized Monte Carlo simulations to manipulate the proportion of outlying common items, the magnitude of a-parameter drift, and the magnitude of b-

文章引用: 王少杰(2026). 准确性与公平性视角下异常锚题对测验等值的影响. *心理学进展, 16(7),* 213-224.
DOI: 10.12677/ap.2026.167353

parameter drift, in order to examine the impact of outlying common items on test equating. Evaluation criteria included bias and root mean square error of linking constants, absolute bias of equated scores, weighted absolute bias of equated scores, and absolute bias of classification accuracy and consistency. The results for errors in linking constants and equated scores demonstrated that the information-weighted characteristic curve methods consistently outperformed traditional characteristic curve linking methods. Additionally, all IRT linking methods exhibited high fidelity in recovering classification accuracy and consistency.

Keywords

Item Response Theory, Information Weighting, Characteristic Curve Methods, Outlying Common Items, Classification Accuracy and Consistency

1. 引言

当进行不同测验或考生群体之间的题目参数或测验分数比较时，往往需要通过项目反应理论(Item Response Theory, IRT)链接(linking)过程将处于不同量纲上的项目和能力参数调整到同一量尺(Kolen & Brennan, 2014)。因此，参数链接的准确性直接关系到后续测验等值结果的有效性与可解释性。在现有研究中，Haebara 方法(Haebara, 1980)和 Stocking-Lord 方法(Stocking & Lord, 1983)等特征曲线方法(以下简称“传统方法”)被广泛应用于 IRT 参数链接，并在不同研究情境下表现出较好的稳定性和准确性(例如，Baldwin, 2011; Hu et al., 2008; Kim & Kolen, 2022)。

然而，参数链接的效果在很大程度上依赖于锚题的质量。在 IRT 参数链接过程中，锚题在两次测试中的表现可能存在差异，这一现象被称为异常锚题(outliers or outlying common items) (He & Cui, 2020; Liu & Jurich, 2022)。异常锚题会导致项目参数发生漂移(drift) (Hu et al., 2008)，进而影响链接系数的估计，并进一步降低测验等值结果的准确性(He et al., 2013; Michaelides, 2010)。因此，如何降低异常锚题对参数链接和测验等值过程的不良影响，已成为测验等值领域的重要研究议题。

针对异常锚题，目前主要形成了两类解决思路。第一类是检测并剔除异常锚题，以减少其对后续参数链接和测验等值过程的干扰。例如，Liu 和 Jurich (2022, 2023)分别提出抽样方差法(sampling variance method)和 t 检验法(t-test method)。抽样方差法通过分析新、旧测验中的项目参数差异来识别异常锚题，而 t 检验法则基于新测验的项目难度参数是否显著偏离题库基准来识别异常锚题(Liu & Jurich, 2023)。尽管这类方法能够提高参数链接和测验等值的准确性，但剔除异常锚题可能会削弱测验内容的代表性(content representativeness) (Kolen & Brennan, 2014; Moses, 2022)。

第二类方法则是稳健链接方法(robust linking methods)，即保留全部锚题，通过统计加权或稳健估计降低异常锚题的影响。例如，He 等人(2015)及 He 与 Cui (2020)提出的加权 Stocking-Lord 方法，包括面积加权法(area-weighted method)和最小绝对值法(least absolute values method)。其中，面积加权法在传统 Stocking-Lord 方法的损失函数中引入 Huber 加权，而最小绝对值法则采用新、旧测验项目特征曲线之间绝对差值的倒数作为权重。这类方法能够在保持锚题内容代表性的同时降低异常锚题的影响(Kolen & Brennan, 2014; Moses, 2022)，但其对参数链接和测验等值准确性的改善程度仍存在差异。

在此背景下，Wang 等人(2022, 2024)提出两种信息量加权特征曲线方法(information-weighted

characteristic curve methods) (以下简称“信息量加权方法”)，并证明其在链接系数返真性方面的表现优于传统方法。从原理上看，该方法同样属于稳健链接方法，因为其通过对异常锚题和估计不佳项目赋予较小权重，以减少其对参数链接过程的不良影响。然而，现有研究主要关注其链接系数估计的表现，对于其在异常锚题条件下的测验等值效果仍缺乏系统检验。换言之，当锚题发生参数漂移时，信息量加权方法能否仍然保持优势，以及其优势是否能够进一步传递到测验等值结果，尚有待深入研究。

此外，现有研究对测验等值结果的评价主要集中于链接系数和等值分数的准确性(如 Baldwin & Clauser, 2022; van der Linden, 2022)，而较少关注其对分类决策结果的潜在影响。对于教育考试和资格认证测验而言，分类准确性(Classification Accuracy)和一致性(Classification Consistency)是衡量测验公平性和决策质量的重要指标(Park et al., 2023; Setzer et al., 2023)。其中，分类准确性反映了基于观测临界分数(observed cut scores)进行分类的结果与基于真实临界分数(true cut scores)进行分类的结果之间的一致程度(Kim & Lee, 2019; Lee, 2010)；分类一致性则衡量了考生在测验多次重复测量过程中是否被稳定划分到相同类别中(Kim & Lee, 2019; Lee, 2010)。虽然 Yi 等人(2007)提出了一种估计两个等值测验分类准确性和一致性的方法，但该研究并未考察 IRT 参数链接与测验等值误差如何进一步影响分类结果，也未考虑异常锚题的作用。因此，目前尚不清楚异常锚题是否不仅影响测验等值的准确性，还会进一步影响基于等值结果的分类决策公平性。

综上所述，现有研究仍存在两个重要不足：其一，缺乏关于信息量加权方法在异常锚题条件下测验等值表现的系统证据；其二，异常锚题对分类准确性和分类一致性的影响尚未得到充分关注。为弥补上述不足，本研究在非等组锚测验设计(common-item nonequivalent groups design)下，采用蒙特卡洛模拟，同时从测验等值准确性和分类决策公平性两个视角，系统考察异常锚题对 IRT 参数链接和测验等值过程的影响，并比较传统方法与信息量加权方法在不同异常条件下的表现。本文首先介绍蒙特卡洛模拟研究方案，然后报告结果，最后总结研究发现并对未来研究方向提出建议。

2. 研究设计

2.1. 操纵变量

本研究操控的自变量包括异常锚题比例、区分度参数漂移程度以及难度参数漂移程度。具体而言，异常锚题占锚题总量的比例分别设置为 5%、10%和 15%，从而模拟三种不同程度的锚题异常情况(He & Cui, 2020; Liu & Jurich, 2023)。区分度和难度参数的漂移程度均分别设定为无、较小和适中。

根据以往研究，较多锚题数量或较大考生样本量有助于提升 IRT 参数链接和测验等值的表现(Diao & Keller, 2020; Goodman et al., 2020; Wang et al., 2022)，且规律较为一致。因此，为重点探讨异常锚题对测验等值的影响，本研究设定参加新、旧测验的每个组考生人数均为 2000 人，每个测验均包含 100 道题目，其中 20%为内锚锚题(Manna & Gu, 2019)，即共有 20 道锚题。

在三参数逻辑斯蒂克模型(three-parameter logistic model, 3PL)中，项目的区分度(a)、难度(b)和猜测系数(c)随机抽取自 $a \sim U(0.4,1.4)$、$b \sim N(0,1)$、$c \sim Beta(8,24)$ (Kim, 2020; Kim et al., 2019; Zhang, 2021)。在区分度参数较小和适中漂移条件下，分别从 $U(-0.3,0.3)$ 和 $U(-0.5,0.5)$ 中随机抽取数值，并将其加到新测验 X 的区分度数值上(He & Cui, 2020; Liu & Jurich, 2022)；在区分度参数无漂移条件下，该随机值设为 0。模拟研究中锚题难度参数的漂移设定方式与区分度参数一致，不再赘述。考虑到研究结果的可解释性以及以往研究经验(例如，He & Cui, 2020; He et al., 2015; Liu & Jurich, 2022, 2023)，本研究未将猜测系数的漂移程度纳入自变量。参加新、旧测验的考生能力参数均服从 $N(0,1)$。

需要注意，在区分度和难度参数均无漂移的情况下，若仍存在异常锚题，其结果无实际意义。因此，本研究共设定 24 种模拟重复条件，即 3 种异常锚题比例 × 8 种区分度和难度参数漂移组合。

2.2. 参数链接方法

在 3PL 模型下，考生正确回答某道题目的概率由项目特征曲线(Item Characteristic Curve, ICC)决定，其数学表达式为

$$P\left(\theta_i;a_j,b_j,c_j\right)=c_j+\frac{1-c_j}{1+\mathrm{e}^{-a_j\left(\theta_i-b_j\right)}}, \tag{1}$$

其中，θ_i 为考生的能力参数，a_j 为项目区分度参数，b_j 为项目难度参数，c_j 为猜测系数。

在 IRT 参数链接过程中，传统 Haebara 方法和 Stocking-Lord 方法也被称为项目特征曲线方法(item characteristic curve method, ICC)和测验特征曲线方法(test characteristic curve method, TCC)。它们的损失函数分别为：

$$\mathrm{ICCcrit}=\sum_i w_i\sum_j\left[\mathrm{ICC}\left(\theta_{Yi};a_{Yj},b_{Yj},c_{Yj}\right)-\mathrm{ICC}\left(\theta_{Yi};\frac{a_{Xj}}{A},Ab_{Xj}+B,c_{Xj}\right)\right]^2, \tag{2}$$

$$\mathrm{TCCcrit}=\sum_i w_i\left[\mathrm{TCC}\left(\theta_{Yi};a_Y,b_Y,c_Y\right)-\mathrm{TCC}\left(\theta_{Yi};\frac{a_X}{A},Ab_X+B,c_X\right)\right]^2, \tag{3}$$

其中，w_i 为考生 j 的权重，$\mathrm{ICC}\left(\theta_{Yi};a_{Yj},b_{Yj},c_{Yj}\right)$ 和 $\mathrm{ICC}\left(\theta_{Yi};\frac{a_{Xj}}{A},Ab_{Xj}+B,c_{Xj}\right)$ 为旧测验 Y 和新测验 X 的项目特征曲线，$\mathrm{TCC}\left(\theta_{Yi};a_Y,b_Y,c_Y\right)$ 和 $\mathrm{TCC}\left(\theta_{Yi};\frac{a_X}{A},Ab_X+B,c_X\right)$ 为旧测验 Y 和新测验 X 的测验特征曲线，A 和 B 为待估计的 IRT 参数链接系数。

根据 Wang 等人(2022, 2024)的研究，信息量加权特征曲线方法将项目信息量和测验信息量作为传统特征曲线方法损失函数成分的权重系数，可以根据损失函数的估计误差分别赋予其不同权重，从而降低异常参数对参数链接和测验等值过程的影响。具体而言，项目信息量加权特征曲线方法(item-information-weighted characteristic curve method, IWCC)和测验信息量加权特征曲线方法(test-information-weighted characteristic curve method, TWCC)的损失函数分别如下：

$$\begin{aligned}\mathrm{IWCCcrit}=\sum_i w_i\sum_V\Bigg\{&\left[\mathrm{ICC}\left(\theta_{Yi};a_{Yj},b_{Yj},c_{Yj}\right)-\mathrm{ICC}\left(\theta_{Yi};\frac{a_{Xj}}{A},Ab_{Xj}+B,c_{Xj}\right)\right]^2\\&\times\left[\mathrm{IIF}\left(\theta_{Yi};a_{Yj},b_{Yj},c_{Yj}\right)+\mathrm{IIF}\left(\theta_{Yi};\frac{a_{Xj}}{A},Ab_{Xj}+B,c_{Xj}\right)\right]\Bigg\},\end{aligned} \tag{4}$$

$$\begin{aligned}\mathrm{TWCCcrit}=\sum_i w_i\Bigg\{&\left[\mathrm{TCC}\left(\theta_{Yi};a_Y,b_Y,c_Y\right)-\mathrm{TCC}\left(\theta_{Yi};\frac{a_X}{A},Ab_X+B,c_X\right)\right]^2\\&\times\left[\mathrm{TIF}\left(\theta_{Yi};a_Y,b_Y,c_Y\right)+\mathrm{TIF}\left(\theta_{Yi};\frac{a_X}{A},Ab_X+B,c_X\right)\right]\Bigg\},\end{aligned} \tag{5}$$

其中，$\mathrm{IIF}\left(\theta_{Yi};a_{Yj},b_{Yj},c_{Yj}\right)$ 和 $\mathrm{IIF}\left(\theta_{Yi};\frac{a_{Xj}}{A},Ab_{Xj}+B,c_{Xj}\right)$ 分别为测验 Y 和测验 X 的项目信息函数，$\mathrm{TIF}\left(\theta_{Yi};a_Y,b_Y,c_Y\right)$ 和 $\mathrm{TIF}\left(\theta_{Yi};\frac{a_X}{A},Ab_X+B,c_X\right)$ 分别为测验 Y 和测验 X 的测验信息函数。

因此，本研究纳入的参数链接方法包括 ICC 方法、TCC 方法、IWCC 方法、TWCC 方法。需要说明，从实践应用的普遍性、信息量加权方法的逻辑基础以及研究重点(异常锚题)三个角度考虑，模拟研究未纳

入其他参数链接方法。

2.3. 评价指标

本研究使用三类指标，分别从参数链接系数、测验等值分数和分类误差角度评价测验等值的表现。

2.3.1. 参数链接系数

参数链接系数评价指标为信息量加权方法较相应传统方法在偏差(bias)和均方根误差(RMSE)上的相对改善(relative improvement, RI)，该指标定义为 $\mathrm{RI}\left(\hat{\lambda}\right)=\frac{\left|Cr_{traditional}\left(\hat{\lambda}\right)\right|-\left|Cr_{new}\left(\hat{\lambda}\right)\right|}{\left|Cr_{traditional}\left(\hat{\lambda}\right)\right|}$，其中，$Cr_{traditional}\left(\hat{\lambda}\right)$ 和 $Cr_{new}\left(\hat{\lambda}\right)$ 为采用传统方法和信息量加权方法所得的参数链接系数偏差或均方根误差，$\mathrm{Bias}\left(\hat{\lambda}\right)=\frac{1}{R}\sum_{r=1}^{R}\hat{\lambda}_r-\lambda$，$\mathrm{RMSE}\left(\hat{\lambda}\right)=\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\lambda}_r-\lambda\right)^2}$，$R$ 为各实验条件的模拟重复次数(本研究为 500)，λ 为参数链接系数的真值($A=1$，$B=0$)。RI 值大于 0 且越接近 1，表明信息量加权方法相较于传统方法的优势越明显。Kim 与 Kolen (2006)认为，该类指标可作为一种类似于效应量的统计值。同时，在 IRT 测验等值研究中，研究者更多关注结果与结论的现实意义，即，其对教育考试政策制定与人才选拔等的实际影响，少有研究者重点探究相应结果的统计显著性(例如，He et al., 2015; Kolen & Brennan, 2014; Manna & Gu, 2019)。因此，本文对测验等值方法的表现差异未作统计检验。

2.3.2. 测验等值分数

如果待等值两测验具有相同的长度，且难度相似，那么自身等值(identity equating；即测验 X 的分数在等值转换前后保持不变)在数学上等同于线性等值方法或平均值等值方法(Peabody, 2020)，并且可降低测验等值误差(Wolkowitz & Wright, 2019)。因此，本研究采用自身等值作为评价测验等值分数的比较标准(Fellinghauer et al., 2023; Li, 2012)。

测验等值分数评价指标包括测验等值绝对偏差(absolute bias of equating scores, ABE)与测验等值加权绝对偏差(weighted absolute bias of equating scores, WABE)，它们分别定义为 $\mathrm{ABE}\left(x_i\right)=\frac{1}{R}\sum_{r=1}^{R}\left|x_{ir,e}-x_{i,t}\right|$，$\mathrm{WABE}\left(x\right)=w_i\times\mathrm{ABE}\left(x_i\right)$，其中，$x_{ir,e}$ 和 $x_{i,t}$ 分别为第 r 次模拟下的测验等值分数估计值及其真值，w_i 为分数 x_i 的实际频次。ABE 和 WABE 越小，代表测验等值分数的误差越低。

2.3.3. 分类误差

本研究采用 Rudner 方法(Rudner, 2005)计算分类准确性和一致性，该方法基于考生潜在能力 θ 进行分类。此外，研究还使用 Lee 方法(Lee, 2010)进行对比，结果表明两种方法计算的分类准确性和一致性结果基本相同。Rudner 方法假设考生 i 的能力 θ_i 及其标准误服从正态分布。当测验临界分数(cut score)为 θ_{cut} 时，如果考生 i 的能力 θ_i 大于 θ_{cut}，其分类准确性估计值为正态分布中 θ_{cut} 右侧面积的比例；如果考生 i 的能力 θ_i 未达到 θ_{cut}，其分类准确性估计值为正态分布中 θ_{cut} 左侧面积的比例。分类一致性的计算方式与此类似，即分别计算上述两个区域面积的平方和。

本研究设定 $\theta_{cut}=1.1$ (Furter & Dwyer, 2020)，从而计算分类准确性的绝对偏差(absolute bias of classification accuracy, ABA)和分类一致性的绝对偏差(absolute bias of classification consistency, ABC)：$\mathrm{ABA}=\frac{1}{R}\sum_{r=1}^{R}\left|CA_{r,e}-CA_{r,t}\right|$，$\mathrm{ABC}=\frac{1}{R}\sum_{r=1}^{R}\left|CC_{r,e}-CC_{r,t}\right|$，其中，$CA_{r,e}$ 和 $CC_{r,e}$ 为分类准确性和一致性的估计值，$CA_{r,t}$ 和 $CC_{r,t}$ 为其真值(采用模拟获得的项目参数真值计算得到)。ABA 和 ABC 值越小，代表相应测验等值结果的分类误差越小。

2.3.4. 软件

所有模拟均使用 R 4.4.0 (R Core Team, 2024)完成。IRT 参数校准使用 mirt 软件包(Chalmers, 2012)，分类准确性和一致性的计算使用 cacIRT 软件包(Lathrop, 2014)。

3. 研究结果

3.1. 测验等值的准确性

表 1 呈现的是链接系数估计偏差(bias)的相对改善(RI)结果。该表在不同异常锚题比例、区分度和难度参数漂移程度条件下，比较信息量加权方法(IWCC 方法和 TWCC 方法)与其对应传统方法(ICC 方法和 TCC 方法)的差异。需要注意的是，在本研究中，将 IWCC 方法与 ICC 方法、TWCC 方法与 TCC 方法进行对比，表中加粗数值表示 RI 为负数。整体而言，IWCC 方法和 TWCC 方法的偏差(系统误差)低于相应传统方法(大多数 RI 值为正)。此外，所有负值 RI 均出现在 TWCC 方法结果中，这主要与链接系数的异常情况相关。此外，无论异常锚题比例、区分度和难度参数漂移程度如何变化，信息量加权方法相较于传统方法的优势始终保持稳定。该结果表明，对信息量较低或参数估计不稳定的锚题赋予较小权重，有助于降低异常锚题对链接系数估计的系统性影响，从而提高参数链接的准确性。

Table 1. RI for the bias of IRT linking coefficient estimates
表 1. IRT 链接系数估计偏差的 RI

区分度参数漂移程度	难度参数漂移程度	5%异常锚题				10%异常锚题				15%异常锚题			
		链接系数 *A*		链接系数 *B*		链接系数 *A*		链接系数 *B*		链接系数 *A*		链接系数 *B*	
		IWCC	TWCC	IWCC	TWCC	IWCC	TWCC	IWCC	TWCC	IWCC	TWCC	IWCC	TWCC
无	较小	0.410	0.001	0.690	0.741	0.369	0.030	0.720	0.568	0.389	0.273	0.716	0.523
无	适中	0.382	0.455	0.703	0.619	0.377	0.219	0.770	**−0.532**	0.359	0.279	0.756	0.911
较小	无	0.398	0.128	0.750	0.978	0.442	0.266	0.817	0.095	0.321	**−1.992**	0.760	0.513
较小	较小	0.432	**−11.779**	0.742	**−66.791**	0.395	0.105	0.670	0.249	0.333	0.155	0.700	**−11.785**
较小	适中	0.364	0.091	0.690	**−0.204**	0.402	0.323	0.803	0.810	0.386	0.213	0.821	**−0.255**
适中	无	0.364	0.489	0.773	0.674	0.314	0.112	0.733	0.857	0.309	0.458	0.801	0.898
适中	较小	0.351	0.054	0.724	0.493	0.351	0.175	0.778	**−1.772**	0.315	0.165	0.797	**−0.284**
适中	适中	0.334	0.927	0.798	0.892	0.345	0.139	0.678	0.242	0.246	0.284	0.791	**−0.057**

注：IWCC = 项目信息量加权特征曲线方法，TWCC = 测验信息量加权特征曲线方法。RI 为负值代表 IWCC 方法或 TWCC 方法的偏差超过其对应传统方法。

表 2 呈现的是链接系数均方根误差(RMSE)的相对改善(RI)结果。其解读方式与表 1 相似，不再赘述。整体而言，IWCC 方法和 TWCC 方法的表现均优于相应传统方法，所有 RI 均为正值，这与表 1 中偏差的 RI 结果形成对比。值得注意的是，均方根误差的 RI 值低于偏差的 RI 值，这表明信息量加权的思路对 IRT 参数链接过程中的系统误差影响更大，而对随机误差的影响较小。此外，与偏差的 RI 结果类似，信息量加权方法相较于传统方法的优越性在不同模拟条件下保持稳定。同时，表 2 倒数第二行的平均数统计量表明，随着异常锚题比例的增加，RI 呈现略微下降趋势。这说明当异常锚题数量持续增加时，即使采用信息量加权策略，也难以完全抵消参数漂移带来的累积影响，但其整体优势仍然存在。

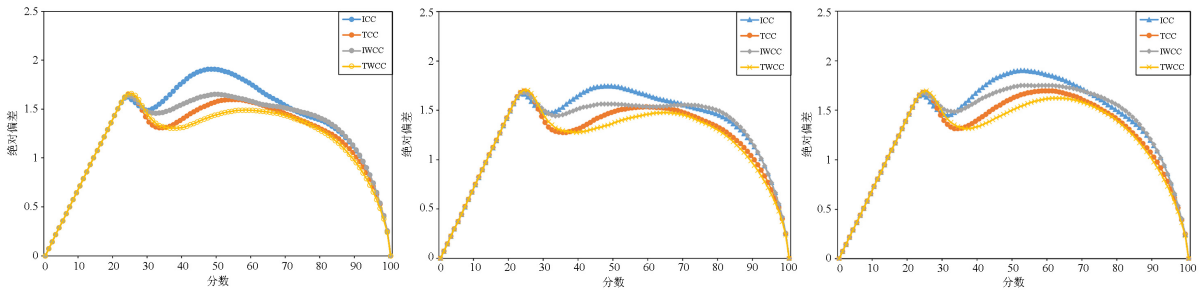
图 1 为测验等值分数的绝对偏差，左、中、右分别表示异常锚题比例分别为 5%、10%和 15%，区分度参数无漂移且难度参数较小漂移的情境。整体而言，测验等值分数的绝对偏差在异常锚题比例分别为

5%、10%和 15%的条件下呈现一致的趋势。在分数低于 25 和高于 95 的范围内，四种 IRT 参数链接方法的表现不存在明显差异。在分数 25 和 50 附近，IRT 参数链接方法绝对偏差曲线均出现两个峰值。在 30~70 分范围内，四种方法的绝对偏差从小到大依次为：TWCC 方法、TCC 方法、IWCC 方法和 ICC 方法。其他条件下的结果与图 1 相似，未呈现。上述结果表明，异常锚题对测验等值的影响主要集中在考生人数较多、分数转换较为频繁的中间分数段，而信息量加权方法能够在这一关键区间提供更准确的分数转换结果。

Table 2. RI for the RMSE of IRT linking coefficient estimates
表 2. IRT 链接系数估计均方根误差的 RI

区分度参数漂移程度	难度参数漂移程度	5%异常锚题				10%异常锚题				15%异常锚题			
		链接系数 *A*		链接系数 *B*		链接系数 *A*		链接系数 *B*		链接系数 *A*		链接系数 *B*	
		IWCC	TWCC	IWCC	TWCC	IWCC	TWCC	IWCC	TWCC	IWCC	TWCC	IWCC	TWCC
无	较小	0.122	0.210	0.242	0.210	0.109	0.204	0.216	0.219	0.122	0.201	0.264	0.262
无	适中	0.133	0.204	0.192	0.187	0.099	0.206	0.161	0.208	0.086	0.192	0.193	0.211
较小	无	0.155	0.201	0.206	0.199	0.158	0.209	0.206	0.192	0.086	0.180	0.255	0.219
较小	较小	0.143	0.215	0.237	0.236	0.093	0.211	0.258	0.217	0.119	0.173	0.190	0.224
较小	适中	0.083	0.192	0.218	0.205	0.135	0.168	0.183	0.199	0.111	0.214	0.157	0.212
适中	无	0.125	0.210	0.257	0.241	0.059	0.195	0.255	0.233	0.044	0.165	0.226	0.225
适中	较小	0.130	0.204	0.229	0.200	0.091	0.176	0.209	0.216	0.138	0.168	0.219	0.159
适中	适中	0.071	0.193	0.201	0.210	0.083	0.209	0.201	0.200	0.026	0.185	0.198	0.193
M		0.120	0.204	0.223	0.211	0.103	0.197	0.211	0.210	0.092	0.185	0.213	0.213
SD		0.027	0.007	0.021	0.018	0.029	0.015	0.031	0.012	0.037	0.016	0.033	0.027

注：IWCC = 项目信息量加权特征曲线方法，TWCC = 测验信息量加权特征曲线方法。



注：左、中、右分别表示异常锚题比例分别为 5%、10%和 15%，区分度参数无漂移且难度参数较小漂移的情境。ICC = 项目特征曲线方法，TCC = 测验特征曲线方法，IWCC = 项目信息量加权特征曲线方法，TWCC = 测验信息量加权特征曲线方法。

Figure 1. ABE of equated score estimates
图 1. 测验等值分数估计的绝对偏差

表 3 呈现的是测验等值分数的加权绝对偏差，下划线数值代表信息量加权方法的表现优于相应传统方法。在整个分数范围内，四种 IRT 链接方法均能提供可靠的链接常数估计，以供后续测验等值过程使用。平均而言，IWCC 方法和 TWCC 方法的加权绝对偏差相较于对应传统方法分别降低 0.06 和 0.09。此外，在不同异常锚题比例、区分度和难度参数漂移程度条件下，加权绝对偏差未呈现明显的规律性变化。这表明信息量加权方法的优势具有较好的稳定性，其表现并不依赖于特定的异常锚题情境。

Table 3. WABE of equated score estimates
表 3. 测验等值分数估计的加权绝对偏差

区分度参数漂移程度	难度参数漂移程度	5%异常锚题				10%异常锚题				15%异常锚题			
		ICC	TCC	IWCC	TWCC	ICC	TCC	IWCC	TWCC	ICC	TCC	IWCC	TWCC
无	较小	1.642	1.463	1.525	1.398	1.594	1.439	1.523	1.382	1.729	1.563	1.665	1.503
无	适中	1.587	1.436	1.508	1.391	1.650	1.474	1.574	1.443	1.584	1.448	1.520	1.390
较小	无	1.668	1.549	1.583	1.501	1.671	1.536	1.572	1.468	1.676	1.504	1.580	1.430
较小	较小	1.631	1.513	1.555	1.469	1.649	1.506	1.546	1.450	1.651	1.505	1.568	1.456
较小	适中	1.692	1.554	1.625	1.504	1.669	1.534	1.572	1.470	1.666	1.520	1.599	1.442
适中	无	1.629	1.503	1.538	1.439	1.646	1.503	1.554	1.436	1.664	1.529	1.590	1.470
适中	较小	1.711	1.522	1.581	1.476	1.622	1.490	1.538	1.437	1.632	1.500	1.563	1.438
适中	适中	1.666	1.549	1.607	1.510	1.695	1.547	1.636	1.495	1.707	1.584	1.617	1.510
M		1.653	1.511	1.565	1.461	1.650	1.504	1.564	1.448	1.664	1.519	1.588	1.455
SD		0.037	0.040	0.038	0.044	0.029	0.034	0.032	0.031	0.042	0.039	0.040	0.037

注：ICC = 项目特征曲线方法，TCC = 测验特征曲线方法，IWCC = 项目信息量加权特征曲线方法，TWCC = 测验信息量加权特征曲线方法。下划线数值代表 IWCC 方法或 TWCC 方法的表现优于其对应传统方法。

3.2. 测验等值的公平性

测验等值的公平性体现为等值结果的分类准确性与一致性的绝对偏差。表 4 和表 5 分别呈现的是分类准确性和一致性的绝对偏差。整体而言，在所有模拟条件下，分类准确性与一致性指标估计值与其真值之间的偏差较小，表明测验等值后的分数在分类准确性与一致性方面具有较高的返真性。此外，异常锚题比例、区分度和难度参数漂移程度对分类准确性与一致性的影响较小。该结果说明，尽管异常锚题会影响参数链接和分数等值的精度，但其误差尚未显著传递至分类决策层面。因此，在本研究设定的条件下，异常锚题对测验等值公平性的影响相对有限，分类结果总体保持较高稳定性。

Table 4. ABA of classification accuracy estimates
表 4. 分类准确性估计的绝对偏差

区分度参数漂移程度	难度参数漂移程度	5%异常锚题				10%异常锚题				15%异常锚题			
		ICC	TCC	IWCC	TWCC	ICC	TCC	IWCC	TWCC	ICC	TCC	IWCC	TWCC
无	较小	0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001
无	适中	0.002	0.000	0.001	0.001	0.002	0.001	0.000	0.001	0.002	0.001	0.001	0.001
较小	无	0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001
较小	较小	0.002	0.001	0.001	0.001	0.002	0.000	0.001	0.001	0.002	0.001	0.000	0.001
较小	适中	0.002	0.001	0.001	0.001	0.003	0.001	0.002	0.001	0.002	0.001	0.003	0.001
适中	无	0.002	0.001	0.001	0.001	0.003	0.000	0.002	0.001	0.002	0.001	0.002	0.001
适中	较小	0.002	0.001	0.002	0.001	0.001	0.001	0.001	0.000	0.002	0.001	0.002	0.000
适中	适中	0.002	0.000	0.001	0.000	0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001
M		0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.002	.001	0.001	0.001
SD		0.000	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000

注：ICC = 项目特征曲线方法，TCC = 测验特征曲线方法，IWCC = 项目信息量加权特征曲线方法，TWCC = 测验信息量加权特征曲线方法。由于保留三位小数，部分数值为 0。

Table 5. ABA of classification consistency estimates
表 5. 分类一致性估计的绝对偏差

区分度参数漂移程度	难度参数漂移程度	5%异常锚题				10%异常锚题				15%异常锚题			
		ICC	TCC	IWCC	TWCC	ICC	TCC	IWCC	TWCC	ICC	TCC	IWCC	TWCC
无	较小	0.003	0.001	0.001	0.001	0.003	0.001	0.001	0.001	0.003	0.001	0.001	0.001
无	适中	0.003	0.000	0.001	0.001	0.002	0.001	0.001	0.001	0.003	0.001	0.001	0.001
较小	无	0.003	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.003	0.001	0.001	0.001
较小	较小	0.002	0.001	0.002	0.001	0.003	0.001	0.001	0.001	0.002	0.001	0.001	0.001
较小	适中	0.003	0.001	0.001	0.001	0.004	0.001	0.003	0.001	0.003	0.001	0.004	0.001
适中	无	0.002	0.001	0.001	0.001	0.003	0.000	0.003	0.001	0.002	0.002	0.003	0.000
适中	较小	0.003	0.001	0.002	0.001	0.002	0.001	0.002	0.001	0.002	0.001	0.003	0.000
适中	适中	0.003	0.000	0.002	0.000	0.003	0.001	0.001	0.001	0.003	0.001	0.001	0.001
M		0.003	0.001	0.001	0.001	0.003	0.001	0.002	0.001	0.001	0.001	0.002	0.001
SD		0.000	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000

注：ICC = 项目特征曲线方法，TCC = 测验特征曲线方法，IWCC = 项目信息量加权特征曲线方法，TWCC = 测验信息量加权特征曲线方法。由于保留三位小数，部分数值为 0。

4. 总结与讨论

在测验等值过程中，异常锚题可能对 IRT 参数链接与分数等值过程产生不利影响，这一直是学术研究者和实践工作者关注的焦点。本研究采用蒙特卡罗模拟方法，通过操控异常锚题比例以及区分度和难度参数漂移程度，考查不同 IRT 链接方法在测验等值准确性与公平性方面的表现。

本研究发现，信息量加权方法在链接系数和测验等值分数估计方面整体优于传统方法，且这一优势在不同异常锚题比例和项目参数漂移程度条件下均保持相对稳定。该结果进一步支持了信息量加权思想在 IRT 参数链接中的应用价值。传统方法默认所有锚题对链接过程具有相同贡献，而信息量加权方法则根据项目的信息量和参数估计误差动态调整各锚题的权重，从而降低异常锚题和估计不稳定项目对链接结果的干扰。以往研究表明，在 IRT 参数链接方法中引入信息量(或误差)加权策略，能够有效降低参数链接误差(Barrett & van der Linden, 2019; He & Cui, 2020; He et al., 2015; van der Linden & Barrett, 2016; Wang et al., 2022, 2024)。本研究进一步表明，即使在存在异常锚题及项目参数漂移的情况下，信息量加权方法仍能够保持较好的稳健性，说明其不仅适用于理想测验条件，也具有一定的实践应用潜力。

值得注意的是，信息量加权方法相较于传统方法的优势并未随着锚题异常程度的增加而显著扩大。这表明，虽然信息量加权策略能够缓解异常锚题带来的负面影响，但并不能完全消除参数漂移所导致的测量误差。因此，在实际测验等值过程中，提高链接方法的稳健性固然重要，但保证锚题质量仍然是获得高质量等值结果的重要前提。此外，TWCC 方法在部分情境下出现负向 RI 值，表明其表现存在一定波动。这主要是由于信息量加权方法将传统方法的损失函数与信息量相乘，使其目标函数复杂程度显著提高，从而增加了最优化过程中出现局部最优解或无解情况的可能性。该现象在结构更为复杂的 TWCC 方法中表现得更为明显，也提示未来研究在构建稳健链接方法时需要进一步关注估计精度与计算稳定性之间的平衡。

在测验等值的公平性方面，无论是信息量加权方法还是相应传统方法，在分类一致性和准确性评价指标上均表现出较高的准确性。该发现与以往研究一致(Lathrop & Cheng, 2013; Wyse & Hao, 2012)。值得

注意的是，计算分类误差的 Rudner 方法和 Lee 方法均基于 IRT，但它们实现的方式不同：Rudner 方法采用考生的潜在能力量尺(Rudner, 2005)，而 Lee 方法则使用观测分数量尺(Lee, 2010)。无论采用何种方法，传统方法和信息量加权方法在链接系数估计方面均表现出较高的准确性和稳定性(Kolen & Brennan, 2014; Wang et al., 2022)。因此，基于 IRT 参数链接结果计算的分类一致性和准确性，与采用模拟真值计算的分类一致性和准确性基本相同。换言之，分类一致性和准确性的返真性不仅表明各方法能够支持较为稳定的分类决策，也说明参数链接和分数等值过程中产生的误差尚未显著传递到分类决策层面。从实践应用角度来看，这意味着在本研究设定的条件下，异常锚题虽然会影响参数链接和分数转换的精度，但对最终分类结果的影响相对有限。

尽管本研究为理解异常锚题对测验等值准确性与公平性的影响提供了新的证据，但仍存在一定局限。首先，本研究主要聚焦于异常锚题，而考生能力分布同样是影响 IRT 参数估计和测验等值结果的重要因素。因此，本研究尚未考察异常能力分布与异常锚题共同存在时的作用机制。未来研究可进一步探讨二者对参数链接、分数等值以及分类决策的联合影响，以更加全面地揭示测验等值误差的来源。其次，本研究设定的异常锚题比例和参数漂移程度相对较低，而在实际应用中，较大比例的异常锚题、极端参数漂移程度及多个临界分数(cut scores)均较为常见。因此，未来研究可扩展至更复杂的异常锚题比例、参数漂移程度和多临界分数情境，以进一步检验不同链接方法的稳健性并提升研究结论的实践适用性。再次，本研究以自身等值作为比较基准，未来研究可采用更精确的方法作为参照标准，以提高研究结论的外部效度。最后，本研究将分类准确性和一致性作为测验等值结果的评价指标，而未将其直接纳入 IRT 参数链接和测验等值过程。未来研究可在 IRT 参数链接和测验等值过程中直接纳入分类准确性与一致性，以进一步提升其实际应用价值，并促进测验等值研究从参数层面向决策层面的拓展。

基金项目

本研究获得广东省哲学社会科学规划青年项目(GD25YJY21)资助。

参考文献

Baldwin, P. (2011). A Strategy for Developing a Common Metric in Item Response Theory When Parameter Posterior Distributions Are Known. *Journal of Educational Measurement, 48,* 1-11. https://doi.org/10.1111/j.1745-3984.2010.00127.x

Baldwin, P., & Clauser, B. E. (2022). Historical Perspectives on Score Comparability Issues Raised by Innovations in Testing. *Journal of Educational Measurement, 59,* 140-160. https://doi.org/10.1111/jedm.12318

Barrett, M. D., & van der Linden, W. J. (2019). Estimating Linking Functions for Response Model Parameters. *Journal of Educational and Behavioral Statistics, 44,* 180-209. https://doi.org/10.3102/1076998618808576

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software, 48,* 1-29. https://doi.org/10.18637/jss.v048.i06

Diao, H., & Keller, L. (2020). Investigating Repeater Effects on Small Sample Equating: Include or Exclude? *Applied Measurement in Education, 33,* 54-66. https://doi.org/10.1080/08957347.2019.1674302

Fellinghauer, C., Debelak, R., & Strobl, C. (2023). What Affects the Quality of Score Transformations? Potential Issues in True-Score Equating Using the Partial Credit Model. *Educational and Psychological Measurement, 83,* 1249-1290. https://doi.org/10.1177/00131644221143051

Furter, R. T., & Dwyer, A. C. (2020). Investigating the Classification Accuracy of Rasch and Nominal Weights Mean Equating with Very Small Samples. *Applied Measurement in Education, 33,* 44-53. https://doi.org/10.1080/08957347.2019.1674307

Goodman, J. T., Dallas, A. D., & Fan, F. (2020). Equating with Small and Unbalanced Samples. *Applied Measurement in Education, 33,* 34-43. https://doi.org/10.1080/08957347.2019.1674311

Haebara, T. (1980). Equating Logistic Ability Scales by a Weighted Least Squares Method. *Japanese Psychological Research, 22,* 144-149. https://doi.org/10.4992/psycholres1954.22.144

He, Y., & Cui, Z. (2020). Evaluating Robust Scale Transformation Methods with Multiple Outlying Common Items under IRT True Score Equating. *Applied Psychological Measurement, 44,* 296-310. https://doi.org/10.1177/0146621619886050

He, Y., Cui, Z., & Osterlind, S. J. (2015). New Robust Scale Transformation Methods in the Presence of Outlying Common Items. *Applied Psychological Measurement, 39,* 613-626. https://doi.org/10.1177/0146621615587003

He, Y., Cui, Z., Fang, Y., & Chen, H. (2013). Using a Linear Regression Method to Detect Outliers in IRT Common Item Equating. *Applied Psychological Measurement, 37,* 522-540. https://doi.org/10.1177/0146621613483207

Hu, H. Q., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of Irt-Based Equating Methods in the Presence of Outlier Common Items. *Applied Psychological Measurement, 32,* 311-333. https://doi.org/10.1177/0146621606292215

Kim, K. Y. (2020). Two IRT Fixed Parameter Calibration Methods for the Bifactor Model. *Journal of Educational Measurement, 57,* 29-50. https://doi.org/10.1111/jedm.12230

Kim, K. Y., Lim, E., & Lee, W. (2019). A Comparison of the Relative Performance of Four IRT Models on Equating Passage-Based Tests. *International Journal of Testing, 19,* 248-269. https://doi.org/10.1080/15305058.2018.1530239

Kim, S. Y., & Lee, W. (2019). Classification Consistency and Accuracy for Mixed-Format Tests. *Applied Measurement in Education, 32,* 97-115. https://doi.org/10.1080/08957347.2019.1577246

Kim, S., & Kolen, M. J. (2006). Robustness to Format Effects of IRT Linking Methods for Mixed-Format Tests. *Applied Measurement in Education, 19,* 357-381. https://doi.org/10.1207/s15324818ame1904_7

Kim, S., & Kolen, M. J. (2022). Scale Linking for the Testlet Item Response Theory Model. *Applied Psychological Measurement, 46,* 79-97. https://doi.org/10.1177/01466216211063234

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices.* Springer. https://doi.org/10.1007/978-1-4939-0317-7

Lathrop, Q. N. (2014). R Package cacIRT: Estimation of Classification Accuracy and Consistency Under Item Response Theory. *Applied Psychological Measurement, 38,* 581-582. https://doi.org/10.1177/0146621614536465

Lathrop, Q. N., & Cheng, Y. (2013). Two Approaches to Estimation of Classification Accuracy Rate under Item Response Theory. *Applied Psychological Measurement, 37,* 226-241. https://doi.org/10.1177/0146621612471888

Lee, W. (2010). Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory. *Journal of Educational Measurement, 47,* 1-17. https://doi.org/10.1111/j.1745-3984.2009.00096.x

Li, Y. (2012). Examining the Impact of Drifted Polytomous Anchor Items on Test Characteristic Curve (TCC) Linking and Irt True Score Equating. *ETS Research Report Series, 2012,* i-22. https://doi.org/10.1002/j.2333-8504.2012.tb02291.x

Liu, C., & Jurich, D. (2022). Application of Sampling Variance of Item Response Theory Parameter Estimates in Detecting Outliers in Common Item Equating. *Applied Psychological Measurement, 46,* 529-547. https://doi.org/10.1177/01466216221108122

Liu, C., & Jurich, D. (2023). Outlier Detection Using T-Test in Rasch IRT Equating under NEAT Design. *Applied Psychological Measurement, 47,* 34-47. https://doi.org/10.1177/01466216221124045

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, 2019,* Article ID: 12244. https://doi.org/10.1002/ets2.12244

Michaelides, M. P. (2010). Sensitivity of Equated Aggregate Scores to the Treatment of Misbehaving Common Items. *Applied Psychological Measurement, 34,* 365-369. https://doi.org/10.1177/0146621609359626

Moses, T. (2022). Linking and Comparability across Conditions of Measurement: Established Frameworks and Proposed Updates. *Journal of Educational Measurement, 59,* 231-250. https://doi.org/10.1111/jedm.12322

Park, S., Kim, K. Y., & Lee, W. (2023). Estimating Classification Accuracy and Consistency Indices for Multiple Measures with the Simple Structure MIRT Model. *Journal of Educational Measurement, 60,* 106-125. https://doi.org/10.1111/jedm.12338

Peabody, M. R. (2020). Some Methods and Evaluation for Linking and Equating with Small Samples. *Applied Measurement in Education, 33,* 3-9. https://doi.org/10.1080/08957347.2019.1674304

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Rudner, L. M. (2005). Expected Classification Accuracy. *Practical Assessment Research & Evaluation, 10,* 1-4.

Setzer, J. C., Cheng, Y., & Liu, C. (2023). Classification Accuracy and Consistency of Compensatory Composite Test Scores. *Journal of Educational Measurement, 60,* 501-519. https://doi.org/10.1111/jedm.12357

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement, 7,* 201-210. https://doi.org/10.1177/014662168300700208

Van der Linden, W. J. (2022). What Is Actually Equated in "Test Equating"? A Didactic Note. *Journal of Educational and Behavioral Statistics, 47,* 353-362. https://doi.org/10.3102/10769986211072308

van der Linden, W. J., & Barrett, M. D. (2016). Linking Item Response Model Parameters. *Psychometrika, 81,* 650-673.

https://doi.org/10.1007/s11336-015-9469-6

Wang, S., Lee, W., Zhang, M., & Yuan, L. (2024). IRT Characteristic Curve Linking Methods Weighted by Information for Mixed-Format Tests. *Applied Measurement in Education, 37*, 377-390. https://doi.org/10.1080/08957347.2024.2424547

Wang, S., Zhang, M., Lee, W., Huang, F., Li, Z., Li, Y. et al. (2022). Two IRT Characteristic Curve Linking Methods Weighted by Information. *Journal of Educational Measurement, 59*, 423-441. https://doi.org/10.1111/jedm.12315

Wolkowitz, A. A., & Wright, K. D. (2019). Effectiveness of Equating at the Passing Score for Exams with Small Sample Sizes. *Journal of Educational Measurement, 56*, 361-390. https://doi.org/10.1111/jedm.12212

Wyse, A. E., & Hao, S. (2012). An Evaluation of Item Response Theory Classification Accuracy and Consistency Indices. *Applied Psychological Measurement, 36*, 602-624. https://doi.org/10.1177/0146621612451522

Yi, H. S., Kim, S., & Brennan, R. L. (2007). A Method for Estimating Classification Consistency Indices for Two Equated Forms. *Applied Psychological Measurement, 31*, 275-291. https://doi.org/10.1177/0146621606294209

Zhang, Z. (2021). Asymptotic Standard Errors of Generalized Partial Credit Model True Score Equating Using Characteristic Curve Methods. *Applied Psychological Measurement, 45*, 331-345. https://doi.org/10.1177/01466216211013101