

项目反应理论下的残差加权稳健 项目参数估计方法*

高旭亮 李婷婷 王 芳

(贵州师范大学心理学院, 贵州师范大学心理健康教育与咨询中心, 贵阳 550025)

摘 要 异常作答在心理与教育测评中普遍存在, 易导致项目参数估计偏差, 并削弱统计推断的准确性与解释力。为应对这一问题, 本研究在项目反应理论框架下提出了新的加权参数估计方法 RRW, 并通过模拟实验与实证研究对其有效性进行了检验。实验结果显示, RRW 方法在多种异常作答情境下均表现出较低的 RMSE 和 Bias, 尤其在异常比例条件下, 项目参数估计精度在部分情境下较传统方法和 I_{τ}^* 加权方法提升幅度可超过 30%。实证分析进一步验证了 RRW 方法在真实测验数据中的优越性, 其能够在提升项目参数估计精度的同时, 提高正常作答者的能力估计准确性。综上, RRW 加权估计在复杂异常作答情境下具有显著优势与良好应用前景。

关键词 项目反应理论, 异常作答, 稳健项目参数估计

分类号 B841

1 引言

项目反应理论(Item Response Theory, IRT), 以潜在特质为核心, 是心理与教育测量领域的重要理论基础。它已广泛应用于心理问卷的修订与测量学特征分析(Gao & Chen, 2025; Eichenbaum et al., 2019; Manapat et al., 2021; Williams & Gotham, 2021; Zarate et al., 2023), 以及简版问卷的开发(Brenner et al., 2021; Luo et al., 2024; Sharp et al., 2022; Wang & Gao, 2025), 同时也被广泛用于大规模教育测评的数据分析与应用(Demirkol & Kelecioğlu, 2022; Gershon et al., 2024; Marcq & Braeken, 2025; Polat & Kelecioğlu, 2024)。此外, 在计算机化自适应测验(Computerized Adaptive Testing, CAT)的构建与应用中, IRT 作为核心理论框架提供了方法学支撑。目前, CAT 已在实际测评中得到了有效应用(例如, Boussakuk et al., 2021; Carlozzi et al., 2023; Miles et al., 2020; Phillips et al., 2021)。

然而, IRT 的有效应用主要依赖于项目参数估计的精度。换言之, 在应用过程中必须首先基于测验作答数据准确估计项目参数, 后续所有分析均以此为前提。一旦作答数据质量不高, 即混入了一定比例的异常作答(如被试因动机不足、注意力分散或其他非能力相关因素而产生的反应), 就会显著削弱项目参数估计的可靠性(Hong & Cheng, 2019; Patton et al., 2019; Rios, 2022)。常见的异常作答类型包括作弊(Meijer & Sijtsma, 2001)、粗心作答(Doval & Delicado, 2020)、测验后段的随机作答(王雪 等, 2024), 以及因缺乏考试动机、时间压力或任务疲劳而导致的全程随机作答或创造性作答等。

尤其是在当前广泛采用在线问卷施测的研究中, 异常作答现象更为普遍, 已成为影响数据质量的重要隐患(Goldammer et al., 2020; Maniaci & Rogge, 2014; Meade & Craig, 2012)。其主要原因在于在线施测缺乏现场监控, 被试易受外界干扰且作答动机不足, 从而更容易出现随意或粗心作答。

收稿日期: 2025-11-03

* 国家自然科学基金项目(32460212)、贵州省基础研究计划(自然科学)面上项目(黔科合基础 MS[2025] 261)和 2024 年贵州省高校心理健康教育专项项目(JYT-XLZX-2024-BK019)资助。

通信作者: 王芳, E-mail: wangf9817@foxmail.com

Chauliac 等人(2023)在一项高中生的大规模在线调查中,采用多种方法评估作答质量,结果发现多达 37.45%的参与者存在不同形式的异常作答。在缺乏控制机制的在线调查中,这一比例甚至可高达 46% (Brühlmann et al., 2020)。相关综述进一步指出,不同研究报告的异常作答比例大致介于 33% (Goldammer et al., 2020)至 46% (Oppenheimer et al., 2009)之间。

将异常作答直接纳入项目参数估计,往往会显著偏离真实值,降低研究结论的准确性。已有研究表明,即便异常作答比例仅为 10%~15%,也可能引发严重的测量偏差。比如, Wise 和 DeMars (2006)发现,10%的异常作答会使能力估计的 RMSE (均方根误差)增加约 20%; Meade 和 Craig (2012)指出,15%的异常作答可导致因子载荷偏差高达 25%,甚至改变因子结构。在异常比例达 20%~30%的情况下,问题更为严重,Patton 等人(2019)报告区分度参数 RMSE 增幅达 70%; Rios (2022)在快速猜测情境下观察到能力估计 RMSE 增加 45%; Voss (2024)发现异常比例大于 25% 时,IRT 模型的 RMSE 几乎翻倍。进一步研究还指出,异常作答不仅加剧参数误差,还会降低信度(Maniaci & Rogge, 2014)、扭曲因子结构(Voss, 2024)、破坏测量不变性(Deng & Rios, 2022)。

为识别异常作答者,研究者提出了基于个人拟合统计量(Person-Fit Statistics, PFS)的方法。PFS 通过对整体样本进行参数估计,结合个体作答模式与项目参数计算拟合指数,从而识别可能存在异常反应的被试(Gao et al., 2025; Karabatsos, 2003; Felt et al., 2017; 童昊 等, 2022; 喻晓锋 等, 2024),并常与整卷删除策略结合使用,即一旦某被试被判定为异常,其全部作答数据将被剔除(Maniaci & Rogge, 2014; Meade & Craig, 2012)。此外,Patton 等人(2019)提出基于 IRT 的迭代清洗方法,通过 PFS 统计量识别不认真作答者,并在剔除或调整可疑反应后反复重估项目参数。结果表明,该方法在异常比例较高时可降低参数偏差、改善标准误估计,强调了异常识别与参数估计之间的相互影响。

然而,PFS 的应用仍存在一些争议。首先,若被试仅在部分题目上出现异常反应,直接删除其整份答卷将导致其余有效数据被一并舍弃,显著降低数据利用率,在样本量有限的研究中尤为不利。其次,PFS 的检测效果高度依赖于项目参数的估计精度。由于其需基于全样本进行 IRT 模型估计,再计算个

体拟合值进行异常识别,若在计算 PFS 之前未对异常作答进行预处理,则模型估计本身可能受到严重干扰,尤其在异常作答者比例较高的情况下,项目参数估计将发生严重偏倚(Oshima, 1994; Meijer & Sijtsma, 2001),从而掩盖异常反应,降低检出概率。这种现象被称为掩蔽效应(masking effect),也是当前基于 PFS 方法识别异常作答时面临的核心挑战之一。

为有效应对异常作答对问卷分析结果的干扰,并克服 PFS 统计量在实际应用中的局限。Hong 和 Cheng (2019)提出了一种新的解决思路,即稳健边际极大似然估计方法(Robust Marginal Maximum Likelihood, RMML)。不同于依赖 PFS 值直接剔除异常被试的做法,RMML 方法在参数估计过程中引入基于 PFS 的加权机制,将个体的拟合统计量转化为权重,并将其嵌入对数似然函数中,从而实现在估计过程中动态调节各被试的贡献。

该方法的核心思想是计算所有被试的个人拟合统计量 I_i^* 。通过在 EM 算法中引入基于 I_i^* 的加权机制,当被试作答行为异常时,其 I_i^* 值将显著偏离模型预测,从而被赋予较低的权重;反之,拟合良好的被试则获得较高权重。RMML 方法能够在无需剔除任何被试数据的前提下,有效降低异常作答对项目参数的干扰,从而提升估计的稳健性与精度。Hong 和 Cheng (2019)的研究表明,该方法能有效抑制异常作答的干扰,在多种模拟条件下均显著降低了项目参数的估计偏差。

Hong 和 Cheng (2019)提出的基于 I_i^* 的加权机制,其有效性主要得益于 I_i^* 指标在区分正常与异常被试方面的良好判别能力。查阅文献发现,童昊等人(2022)提出了基于残差的个人拟合统计量 R 指标,并在 IRT 框架下系统比较了 R 与 I_i^* 的识别性能,结果表明 R 在识别异常被试方面显著优于 I_i^* 。同样,Gao 等人(2025)开发了另一种基于残差的拟合统计量 WR ,并将其应用于认知诊断模型中,研究结果显示, WR 在识别异常作答被试方面同样表现优于 I_i^* ,进一步验证了残差型拟合指标在异常检测中的优势。

造成两类 PFS 表现差异的核心原因在于其异常信号提取机制的不同:残差型统计量逐题计算“观测和预测”差值,经方差标准化后累加,相当于以更加微观的方式捕捉被试在各题上的微小偏离,因此对细微且多样的异常作答模式尤为敏感。而 I_i^* 先将整卷反应压缩为单一对数似然比,再按测验信

息标准化, 此聚合过程易使相互抵消的局部偏离被“平均化”, 从而稀释甚至掩盖异常信号。

因此, 我们在 Hong 和 Cheng (2019)提出的框架基础上, 引入了残差型个人拟合统计量, 并据此提出了一种新的加权估计方法。理论上, 与基于 l_z^* 的加权方式相比, 残差加权能够更有效地削弱异常作答对参数估计的干扰, 从而提升项目参数估计的稳健性与精度。本文首先回顾了 Hong 和 Cheng (2019)基于 l_z^* 的加权思路, 随后阐述了基于残差加权方法的构建过程, 并通过模拟实验及真实心理问卷调查数据验证了该方法的有效性。

2 基于 l_z^* 加权的稳健项目参数估计方法

Hong 和 Cheng (2019)提出了稳健最大边际似然估计(Robust Maximum Marginal Likelihood, RMML)方法。该方法基于被试拟合统计量 l_z^* 对似然函数加权, 以减弱异常作答对项目参数估计的不利影响, 并在等级反应模型(Graded Response Model, GRM)框架下得到检验。假设 u_{ij} 表示第 i 位被试对第 j 个项目的作答反应。第 i 位被试的标准化个体拟合统计量 l_z 的计算公式如下:

$$l_{zi} = \frac{l_i - E(l_i)}{\sqrt{\text{var}(l_i)}} \quad (1)$$

其中, l_i 是第 i 位被试在 GRM 下的对数似然函数值, $E(l_i)$ 表示其期望, $\text{var}(l_i)$ 表示其对应的方差。公式(1)给出的 l_{zi} 是“理想情形”的个体拟合统计量: 若真实的能力 θ 与项目参数已知, l_{zi} 在原理上应近似服从 $N(0,1)$ 。

现实中 θ 与项目参数都要用数据估计, 导致(1)式中的期望与方差被系统性低估, 使其不再服从 $N(0,1)$, 进而降低异常识别效能。Sinharay (2016)基于渐近展开推导了校正后的统计量 l_{zi}^* , 该统计量在使用估计参数的情况下仍近似服从 $N(0,1)$ 。RMML 正是以 l_{zi}^* 作为加权依据。给定先验 $f(\theta)$, 则观测到被试 i 的作答向量 $\mathbf{u}_i$ 的边际概率可表示为:

$$P(\mathbf{u}_i | \gamma) = \int L_i(\theta | \mathbf{u}_i, \gamma) f(\theta) d\theta \quad (2)$$

$P(\mathbf{u}_i | \gamma)$ 表示观测数据 $\mathbf{u}_i$ 的概率, γ 表示项目参数矩阵, 在实际计算中, 可对 θ 的取值区间采用数值求积抽取 Q 个节点以及节点的权重 w_q , 公式(2)近似为

$$P(\mathbf{u}_i | \gamma) \approx \sum_{q=1}^Q w_q L_i(\theta_q | \mathbf{u}_i, \gamma) \quad (3)$$

全样本的(未加权)边际似然为:

$$P(\mathbf{u} | \gamma) = \prod_{i=1}^N P(\mathbf{u}_i | \gamma) \quad (4)$$

为了方便计算, 对整个数据向量 $\mathbf{u}$ 取对数, 其对数似然函数应为:

$$\log P(\mathbf{u} | \gamma) = \sum_{i=1}^N \log P(\mathbf{u}_i | \gamma) \quad (5)$$

Hong 和 Cheng (2019)提出了对上述公式(5)进行加权的思路, 即在公式(5)中对每位被试赋予相应权重 w_i , 通过 w_i 来优化正常和异常被试在项目参数估计的贡献度, 加权的公式如下:

$$\log P(\mathbf{u} | \gamma) = \sum_{i=1}^N w_i \log P(\mathbf{u}_i | \gamma) \quad (6)$$

其中 w_i 采用了基于 l_{zi}^* 的权重思路, 同时为了统一单位, 对计算得到的 l_{zi}^* 进行了归一化处理, 公式如下:

$$w_i = \frac{\Phi(l_{zi}^*)}{\sum_i \Phi(l_{zi}^*)} \quad (7)$$

其中 $\Phi(\cdot)$ 表示标准正态分布的累积分布函数, l_{zi}^* 是第 i 个被试的 l_z^* 统计量, 其主要的原理是当被试越异常, l_{zi}^* 越小, 其对应的 w_i 越接近 0。

3 基于残差加权的稳健项目参数估计方法开发

残差是回归分析中的重要概念, 是指实际观测值与模型预测值之间的差异, 即观测值减去预测值的结果。Yu 和 Cheng (2019)提出了一种加权残差指标, 该指标基于 IRT 模型中期望得分与实际得分的差异, 用于识别人格测验中的异常作答。当实际反应显著偏离模型预测时, 可能表明存在异常行为。其计算公式如下:

$$r_{ij} = \frac{|X_{ij} - E_{ij}|}{P_{ij}} \quad (8)$$

其中, X_{ij} 表示被试 i 对项目 j 的观测作答, P_{ij} 表示观测作答 X_{ij} 的概率, E_{ij} 是期望作答, M 表示最高得分, E_{ij} 的计算如下:

$$E_{ij} = \sum_{m=0}^M m P(X_j = m | \theta) \quad (9)$$

就被试 i 而言, 在题量为 J 的测验中, 其总体加权残差定义为

$$r_i = \sum_{j=1}^J \frac{|X_{ij} - E_{ij}|}{P_{ij}} \quad (10)$$

为体现“正常作答的被试权重大, 异常作答的被试权重小”的原则, 我们以公式(10)为基础, 提出了个体层面的残差稳健度权重(Residual Robustness Weight, RRW), 构建步骤如下:

首先, 对 r_i 取倒数 $1/r_i$ 。这是因为 r_i 值越小, 表示该被试的整体作答模式越接近期望作答模式, 即作答质量较高; 而 r_i 越大, 则偏离越严重。取倒数可以反转这种关系, 使拟合较好的被试获得更大的权重, 异常作答被试获得较小的权重。

然后, 为消除测验长度 J 对权重尺度的影响, 需要对 $1/r_i$ 除以题目数 J , 以获得与测验长度无关的平均残差偏离程度。同时, 为了减弱极端 r_i 所带来的数值不稳定性, 同时保持与 $1/r_i$ 单调一致的排序, 我们对上述结果取平方根, 使权重落在更平滑的尺度范围, 并回归到与残差相同的量纲。

基于上述思路, 我们提出残差稳健度权重(Residual Robustness Weight, RRW)。对于被试 i , 在题量为 J 的测验中, 其原始的 RRW_i^{raw} 定义为:

$$RRW_i^{raw} = \sqrt{\frac{1/r_i}{J}} \quad (11)$$

公式(11)的平方根变换属于次线性变换(sub-linear transformation)。当残差指标 r_i 取值较大时, 其增长幅度在平方根函数作用下被压缩, 且边际增速随 x 的增大而递减。因此, 相较于线性变换, 平方根函数对大残差的放大效应更弱, 从而有效降低极端异常值在权重计算中的影响强度。在计算初始的 RRW_i^{raw} 后, 为减少极端值对权重分布的影响, 并使权重尺度具有可比性, 我们对其进行秩转换(rank transformation), 并以秩均值进行归一化处理, 得到调整后的权重:

$$RRW_i = \frac{\text{rank}(RRW_i^{raw})}{\text{mean}(\text{rank}(RRW_i^{raw}))} \quad (12)$$

其中, 分子为第 i 个被试的 RRW_i^{raw} 秩值, 分母为所有被试 RRW_i^{raw} 秩值的均值。公式(12)采用秩变换(rank transformation), 将原始连续权重映射为相对排序信息。秩统计量仅依赖观测值的相对位置, 而不受具体数值幅度影响, 因此具有分布无关特性。对于极端值而言, 无论其数值多大, 在秩空间中仅占据一个有限的排序位置, 从而有效阻断了其通过数值尺度对权重分布产生过度拉伸的可能。

该权重随着残差偏离程度的增大而单调减小, 且在不同测验长度之间具有良好的可比性。因此, 原边际对数似然函数中的个体权重可替换为式(12)所定义的新权重, 更新的公式如下:

$$\log P(\mathbf{u} | \gamma) = \sum_{i=1}^N w_i \log P(\mathbf{u}_i | \gamma), w_i = RRW_i \quad (13)$$

4 模拟实验

4.1 实验设计

为评估本文所提出的基于残差加权的稳健项目参数估计方法的性能, 并与现有的 I_2^* 加权方法进行比较, 我们设计了一项模拟实验。实验采用四因素实验设计, 各实验因素及其水平如下: (1)测验长度: 30 题、50 题; (2)失拟被试类型: 作弊、粗心、随机作答, 幸运猜测, 混合异常类型; (3)失拟被试比例: 0%、20%、30%; (4)异常作答题目比例: 20%、30%, 表示失拟被试在部分题目上的异常作答比例。各实验条件均独立运行 100 次, 并以重复结果的均值进行汇总。模拟与估计程序均在 R 环境中实现, 其中项目参数估计使用 mirt 程序包 (Chalmers, 2012)。

参考 Hong 和 Cheng (2019) 的实验设计, 本研究的模拟实验采用等级反应模型(Graded Response Model, GRM), 并设定各题均为 6 点计分(对应 5 个阈值参数)。模拟生成 1000 名被试, 其能力参数 θ 从标准正态分布 $N(0,1)$ 中生成; 项目的难度参数 (b_k) 同样从标准正态分布 $N(0,1)$ 中抽取, 并按升序排列以确保阈值单调递增; 区分度参数 (a) 由对数正态分布 $\ln N(0,1)$ 生成, 为避免出现极端取值, 限制 a 参数的下限为 0.3, 上限为 3。

第一种失拟类型为作弊(De La Torre & Deng, 2008), 其定义为低能力被试通过作弊、抄袭等方式在高难度项目上获得不符合其真实能力的高分。本研究将能力值低于第 37.5 百分位的被试界定为低能力被试, 能力值高于第 62.5 百分位的被试界定为高能力被试。在模拟异常数据时, 首先根据设定比例随机选取异常被试, 再在其得分为 0 或 1 分的项目中随机抽取指定比例, 并将这些作答分数修改为该题目的满分值。

第二种失拟类型为粗心(Doval & Delicado, 2020), 其定义为高能力被试在低难度项目上因疏忽作答错误而获得 0 分。在真实测验情境中, 粗心作答通常与时间压力密切相关, 因此更易发生于测验后半阶段。在模拟异常数据时, 首先依据预设比例随机选取异常被试; 随后, 在其作答序列的后半部分中, 从原始得分大于等于 3 分的题目中按设定比例随机抽取若干题项, 并将其作答得分统一替换为 0 分, 以此构造粗心作答数据。

表 1 粗心作答情境下不同估计方法精度比较

J	异常被试	异常题目	估计方法	Bias		RMSE	
				Bias a	Bias b	RMSE a	RMSE b
30	0%	0%	传统估计法	0.007	-0.012	0.094	0.137
			$I_{\frac{1}{2}}^*$ 加权估计	0.278	-0.249	0.363	0.404
			RRW 加权估计	0.223	-0.259	0.303	0.409
	20%	20%	传统估计法	-0.011	-0.251	0.139	0.335
			$I_{\frac{1}{2}}^*$ 加权估计	0.313	-0.402	0.426	0.497
			RRW 加权估计	0.263	-0.395	0.373	0.508
	30%	30%	传统估计法	0.000	-0.338	0.144	0.396
			$I_{\frac{1}{2}}^*$ 加权估计	0.404	-0.465	0.507	0.766
			RRW 加权估计	0.375	-0.495	0.472	0.804
	20%	30%	传统估计法	0.154	-0.481	0.250	0.615
			$I_{\frac{1}{2}}^*$ 加权估计	0.477	-0.737	0.614	0.953
			RRW 加权估计	0.467	-0.761	0.607	1.009
	30%	30%	传统估计法	0.237	-0.686	0.341	0.799
			$I_{\frac{1}{2}}^*$ 加权估计				
			RRW 加权估计				
50	20%	20%	传统估计法	0.262	-0.269	0.345	0.412
			$I_{\frac{1}{2}}^*$ 加权估计	0.211	-0.271	0.287	0.408
			RRW 加权估计	-0.003	-0.242	0.129	0.325
	20%	30%	传统估计法	0.310	-0.407	0.422	0.509
			$I_{\frac{1}{2}}^*$ 加权估计	0.267	-0.393	0.375	0.505
			RRW 加权估计	0.033	-0.333	0.145	0.391
	30%	20%	传统估计法	0.379	-0.44	0.475	0.672
			$I_{\frac{1}{2}}^*$ 加权估计	0.355	-0.456	0.446	0.683
			RRW 加权估计	0.157	-0.444	0.235	0.549
	30%	30%	传统估计法	0.486	-0.720	0.613	0.918
			$I_{\frac{1}{2}}^*$ 加权估计	0.470	-0.732	0.601	0.953
			RRW 加权估计	0.254	-0.652	0.349	0.749

注: J = 测验长度; 传统估计法 = 未加权的常规参数估计方法; RMSE b = 难度参数的平均 RMSE。

第三类失拟类型为随机作答。在数据生成过程中, 依据预设比例随机抽取异常被试, 并在其作答中按比例随机选取题目; 对所选题目, 从其可能的得分类别中随机抽取得分以替换原始作答。例如, 若题目满分为 M , 则从 $\{0, 1, 2, \dots, M\}$ 中随机抽取得分类别, 从而构造随机作答数据。

第四类失拟类型为幸运猜测作答, 用于刻画低能力被试通过随机猜测而获得高分的情形。在数据生成过程中, 依据预设比例抽取低能力被试, 并在其作答中从原始得分为 0 或 1 的题目中随机选取一定比例的项目; 假设题目满分为 M , 对所选题目, 从高分得分类别 $\{M-2, M-1, M\}$ 中按照给定概率分布 $\{0.2, 0.4, 0.4\}$ 进行随机抽样, 以替换原始得分, 从而生成幸运猜测作答数据。

第五种失拟类型为混合异常作答, 即同时包含前述四类异常作答模式。在数据生成过程中, 将预设的异常被试按比例均匀分配至作弊、粗心、随机作答与幸运猜测四类。例如, 当异常被试比例为 20% 时, 则每类异常作答各占 5%。对于每一类异常被试, 按设定比例随机抽取题目生成异常作答。

4.2 评价标准

偏差 (Bias) 和均方根误差 (root mean square error, RMSE) 作为评价指标:

$$Bias = \frac{1}{L} \frac{1}{J} \sum_{l=1}^L \sum_{j=1}^J (O_j - \hat{O}_j^{(l)}) \quad (14)$$

$$RMSE = \sqrt{\frac{1}{L} \frac{1}{J} \sum_{l=1}^L \sum_{j=1}^J (O_j - \hat{O}_j^{(l)})^2} \quad (15)$$

其中, O_j 为第 j 个参数的真实值, $\hat{O}_j^{(l)}$ 为该参数在第 l 次模拟中的估计结果, J 和 L 分别表示参数个数(或测验题目数)及各实验条件的重复次数。Bias 越接近 0, 表明估计偏差越小; RMSE 越小, 表明参数估计精度越高。

4.3 实验结果

表 1 至表 5 汇总了不同异常作答情境下各估计方法的 RMSE 与 Bias 结果。由于每个题目包含 5 个难度(阈值)参数, 我们在初始结果整理中保留了全部难度参数的估计精度。但考虑到在正文中呈现所有参数会导致表格过于庞大、影响直观性, 因此正文仅报告平均难度参数的估计精度。完整结果可向通讯作者索取。需要说明的是, 为了更清晰地比较各估计方法的表现, 表 1 至表 5 中的异常被试和异常题目的比例均为 0, 代表理想化的标准作答情形。尽管该情形在实际应用中较为少见, 但为便于与

异常作答情境进行比较, 仍将其结果作为基准条件。

从表 1 可以看出, 当被试存在异常作答时, 即使在较低比例条件下, 传统估计方法的误差亦不容忽视。例如, 在测验长度为 30 题、异常被试比例与异常题目比例均为 20%的条件下, 区分度参数 a 和难度参数 b 的 RMSE 已分别达到 0.363 和 0.404。随着异常被试比例和异常题目比例的进一步增加, 传统估计方法的 RMSE 与 Bias 均显著上升。表明传统估计方法对异常作答敏感, 易受失拟信息干扰并产生系统性偏差。

与传统方法相比, I_2^* 加权主要改善区分度参数估计, 但对难度参数作用有限, 甚至可能增加误差; 例如, 在测验长度为 30 题、异常比例均为 30%时, 传统方法与 I_2^* 加权方法在 b 参数上的 RMSE 分别为 0.953 和 1.009, 表明该方法在难度参数估计上的稳健性不足。

表 2 作弊作答情境下不同估计方法精度比较

J	异常被试	异常题目	估计方法	Bias		RMSE	
				Bias a	Bias b	RMSE a	RMSE b
30	0%	0%	传统估计法	0.007	0.012	0.094	0.137
			传统估计法	0.269	0.252	0.357	0.400
			I_2^* 加权估计	0.205	0.250	0.284	0.399
			RRW 加权估计	-0.014	0.255	0.141	0.342
	20%	30%	传统估计法	0.290	0.408	0.391	0.502
			I_2^* 加权估计	0.245	0.397	0.346	0.509
			RRW 加权估计	-0.007	0.341	0.154	0.400
			传统估计法	0.404	0.474	0.503	0.751
			I_2^* 加权估计	0.373	0.493	0.465	0.765
			RRW 加权估计	0.169	0.498	0.264	0.640
	30%	20%	传统估计法	0.433	0.710	0.553	0.863
			I_2^* 加权估计	0.423	0.713	0.548	0.899
			RRW 加权估计	0.206	0.667	0.302	0.745
50	20%	20%	传统估计法	0.275	0.254	0.360	0.426
			I_2^* 加权估计	0.209	0.251	0.284	0.415
			RRW 加权估计	0.027	0.229	0.146	0.325
			传统估计法	0.298	0.406	0.417	0.504
	20%	30%	I_2^* 加权估计	0.253	0.389	0.366	0.500
			RRW 加权估计	0.032	0.320	0.151	0.382
			传统估计法	0.396	0.470	0.492	0.765
			I_2^* 加权估计	0.366	0.485	0.457	0.798
			RRW 加权估计	0.178	0.488	0.269	0.649
	30%	20%	传统估计法	0.461	0.732	0.591	0.950
			I_2^* 加权估计	0.440	0.724	0.572	0.944
			RRW 加权估计	0.240	0.651	0.341	0.773

注: J = 测验长度; 传统估计法 = 未加权的常规参数估计方法; RMSE b = 难度参数的平均 RMSE。

相比之下, RRW 加权估计方法在表 1 所列各类条件下均表现出更强的稳健性。在区分度参数和难度参数的估计上整体优于传统方法和 I_z^* 加权方法, 且优势随异常比例增加而更加明显。以测验长度为 30 题为例, 在异常比例均为 20% 时, RRW 方法将区分度参数的 RMSE 从 0.363 降至 0.139, 难度参数由 0.404 降至 0.335; 当异常比例提高至 30% 时, 其区分度参数 RMSE 为 0.341, 显著低于传统方法(0.614)和 I_z^* 方法(0.607), 难度参数 RMSE 为 0.799, 同样优于传统方法(0.953)和 I_z^* 方法(1.009)。

在测验长度为 50 题的条件下, RRW 方法仍保持一致优势。例如, 在异常比例均为 20% 时, 其区分度参数和难度参数的 RMSE 分别为 0.129 和 0.325; 在异常比例为 30% 时, 分别为 0.349 和 0.749, 均明显低于其他方法。

从表 2 可以看出, 在作弊作答情境下, 三种估

计方法的变化趋势与粗心作答情境基本一致。与传统估计方法相比, I_z^* 加权方法在一定程度上降低了区分度参数 a 的估计误差, 但在大多数条件下会增加难度参数 b 的估计误差。例如, 在测验长度为 50 题、异常被试比例为 30%、异常题目比例为 20% 的条件下, 传统方法在 a 和 b 参数上的 RMSE 分别为 0.492 和 0.765, 而 I_z^* 加权方法分别为 0.457 和 0.798, 表现出对区分度参数的改善但对难度参数的负面影响。

相比之下, RRW 加权估计方法表现出更优的估计精度。在测验长度为 30 题、异常比例均为 20% 时, 其区分度参数和难度参数的 RMSE 分别为 0.141 和 0.342; 当异常比例增加至 30% 时, 分别为 0.302 和 0.745, 均显著低于传统方法和 I_z^* 加权方法。在测验长度为 50 题的条件下, RRW 方法同样表现出稳定优势。

表 3 随机作答情境下不同估计方法精度比较

J	异常被试	异常题目	估计方法	Bias		RMSE	
				Bias a	Bias b	RMSE a	RMSE b
30	0%	0%	传统估计法	0.007	0.012	0.094	0.137
			I_z^* 加权估计	0.032	-0.002	0.123	0.215
			RRW 加权估计	-0.177	-0.010	0.248	0.206
			传统估计法	0.157	-0.004	0.252	0.219
	20%	30%	I_z^* 加权估计	0.050	-0.015	0.137	0.215
			RRW 加权估计	-0.163	-0.003	0.236	0.207
			传统估计法	0.163	-0.010	0.257	0.227
			I_z^* 加权估计	0.078	-0.024	0.168	0.221
	30%	20%	RRW 加权估计	-0.151	-0.002	0.209	0.211
			传统估计法	0.224	0.011	0.347	0.283
			I_z^* 加权估计	0.087	0.007	0.175	0.256
			RRW 加权估计	-0.078	0.014	0.173	0.252
50	20%	20%	传统估计法	0.125	-0.014	0.208	0.235
			I_z^* 加权估计	0.052	-0.008	0.127	0.245
			RRW 加权估计	-0.149	-0.016	0.202	0.210
			传统估计法	0.158	-0.002	0.248	0.213
	20%	30%	I_z^* 加权估计	0.044	0.000	0.117	0.211
			RRW 加权估计	-0.140	-0.010	0.198	0.206
			传统估计法	0.161	0.015	0.262	0.220
			I_z^* 加权估计	0.074	0.012	0.155	0.231
	30%	20%	RRW 加权估计	-0.114	0.011	0.178	0.211
			传统估计法	0.223	0.004	0.350	0.271
			I_z^* 加权估计	0.096	0.008	0.177	0.242
			RRW 加权估计	-0.053	-0.009	0.154	0.220

注: J = 测验长度; 传统估计法 = 未加权的常规参数估计方法; RMSE b = 难度参数的平均 RMSE。

从表 3 可以看出,在随机作答情境下,各方法的估计误差随异常被试比例和异常题目比例的增加而上升,但整体误差水平低于粗心和作弊情境。与传统估计方法相比, I_z^* 加权方法在区分度参数 a 上表现出更优的估计精度(RMSE 更小),但在难度参数 b 上优势不明显,部分条件下甚至略有增加。

RRW 加权估计方法在多数条件下表现出较好的稳健性。在异常比例较低或中等(如 20%)时,其区分度参数的估计精度略逊于 I_z^* 方法,但在难度参数上通常优于 I_z^* 方法。随着异常比例的增加,RRW 方法的优势逐渐显现。

值得注意的是,在异常被试比例与异常题目比例均为 30%的条件下,RRW 方法在区分度参数和难度参数上的 RMSE 均达到最优,明显优于传统方法和 I_z^* 加权方法。此外,随着测验长度由 30 题增加至 50 题,RRW 方法的估计结果保持稳定,误差增长幅

度较小。

从表 4 可以看出,在幸运猜测作答情境下,与传统估计方法相比, I_z^* 加权方法在区分度参数 a 和难度参数 b 上均表现出一定程度的改进。例如,在测验长度为 30 题、异常被试比例与异常题目比例均为 30%的条件下,传统方法在 a 和 b 参数上的 RMSE 分别为 0.452 和 0.833,而 I_z^* 加权方法分别降低至 0.355 和 0.740。然而,RRW 加权估计方法表现出更优的估计精度。在相同条件下,其区分度参数和难度参数的 RMSE 分别为 0.222 和 0.705,均明显低于传统方法和 I_z^* 加权方法。随着异常被试比例和异常题目比例的增加,各方法的估计误差均呈上升趋势,但 RRW 方法的误差增长幅度相对较小,表现出更好的稳健性。当测验长度由 30 题增加至 50 题时,RRW 方法在不同条件下仍保持较低的 RMSE 水平。

表 4 幸运猜测作答情境下不同估计方法精度比较

J	异常被试	异常题目	估计方法	Bias		RMSE	
				Bias a	Bias b	RMSE a	RMSE b
30	0%	0%	传统估计法	0.007	0.012	0.094	0.137
			I_z^* 加权估计	0.143	0.227	0.202	0.389
	20%	20%	RRW 加权估计	-0.038	0.252	0.162	0.351
			传统估计法	0.229	0.294	0.313	0.424
			I_z^* 加权估计	0.134	0.265	0.196	0.375
	30%	30%	RRW 加权估计	-0.056	0.295	0.141	0.358
			传统估计法	0.320	0.339	0.414	0.607
			I_z^* 加权估计	0.265	0.406	0.337	0.611
	30%	30%	RRW 加权估计	0.089	0.433	0.205	0.554
			传统估计法	0.359	0.560	0.452	0.833
			I_z^* 加权估计	0.284	0.538	0.355	0.740
			RRW 加权估计	0.134	0.588	0.222	0.705
50	20%	20%	传统估计法	0.213	0.186	0.291	0.364
			I_z^* 加权估计	0.140	0.188	0.202	0.335
			RRW 加权估计	-0.041	0.192	0.148	0.293
	30%	30%	传统估计法	0.241	0.329	0.328	0.477
			I_z^* 加权估计	0.152	0.277	0.207	0.397
			RRW 加权估计	-0.003	0.300	0.128	0.368
	30%	30%	传统估计法	0.307	0.341	0.391	0.622
			I_z^* 加权估计	0.255	0.388	0.318	0.581
			RRW 加权估计	0.100	0.399	0.192	0.508
	30%	30%	传统估计法	0.390	0.569	0.493	0.878
			I_z^* 加权估计	0.293	0.517	0.365	0.711
			RRW 加权估计	0.179	0.604	0.269	0.720

注: J = 测验长度; 传统估计法 = 未加权的常规参数估计方法; RMSE b = 难度参数的平均 RMSE。

表 5 混合异常作答情境下不同估计方法精度比较

<i>J</i>	异常被试	异常题目	估计方法	Bias		RMSE	
				Bias <i>a</i>	Bias <i>b</i>	RMSE <i>a</i>	RMSE <i>b</i>
30	0%	0%	传统估计法	0.007	0.012	0.094	0.137
			传统估计法	0.202	0.043	0.279	0.209
			I_z^* 加权估计	0.125	0.041	0.194	0.216
			RRW 加权估计	-0.101	0.050	0.163	0.191
	20%	30%	传统估计法	0.221	0.089	0.325	0.204
			I_z^* 加权估计	0.137	0.077	0.225	0.225
			RRW 加权估计	-0.090	0.085	0.155	0.198
			传统估计法	0.290	0.080	0.386	0.280
	30%	20%	I_z^* 加权估计	0.218	0.081	0.303	0.269
			RRW 加权估计	0.002	0.084	0.143	0.206
			传统估计法	0.294	0.125	0.415	0.250
			I_z^* 加权估计	0.235	0.109	0.343	0.252
50	20%	20%	RRW 加权估计	0.000	0.118	0.154	0.212
			传统估计法	0.191	0.057	0.269	0.221
			I_z^* 加权估计	0.128	0.053	0.198	0.231
			RRW 加权估计	-0.077	0.057	0.143	0.191
	20%	30%	传统估计法	0.208	0.075	0.315	0.200
			I_z^* 加权估计	0.148	0.059	0.235	0.216
			RRW 加权估计	-0.072	0.054	0.141	0.185
			传统估计法	0.281	0.083	0.372	0.288
	30%	20%	I_z^* 加权估计	0.212	0.077	0.290	0.276
			RRW 加权估计	0.026	0.089	0.141	0.213
			传统估计法	0.292	0.131	0.419	0.262
			I_z^* 加权估计	0.241	0.107	0.351	0.260
	30%	30%	RRW 加权估计	0.027	0.121	0.165	0.213

注: *J* = 测验长度; 传统估计法 = 未加权的常规参数估计方法; RMSE *b* = 难度参数的平均 RMSE。

从表 5 可以看出, 在混合异常作答情境下, I_z^* 加权估计方法相较于传统方法在区分度参数 *a* 的估计上具有一定改善, 但对难度参数 *b* 的提升效果较为有限, 甚至在部分条件下会出现精度下降。例如, 在测验长度为 30 题、异常被试比例为 20%、异常作答比例为 30% 的条件下, I_z^* 方法的 *b* 参数 RMSE 为 0.225, 高于传统方法的 0.204, 表明其在难度参数估计上的稳健性不足。

相比之下, RRW 加权估计方法在所有异常作答条件下均表现出更优的稳健性, 能够同时有效降低区分度参数和难度参数的估计误差, 其 RMSE 在三种方法中始终最小。例如, 在测验长度为 30 题、异常被试比例为 30%、异常作答比例为 20% 的条件下, 传统方法的 *a* 和 *b* 参数 RMSE 分别为 0.386 和 0.280, I_z^* 方法分别降至 0.303 和 0.269, 而 RRW 方法进一步降低至 0.143 和 0.206, 显示出明显的优势。

5 实证研究

为进一步验证新方法的实际应用效果, 我们基于真实问卷调查数据开展了实证研究。数据来源于一次面向研究生群体的在线心理健康调查, 测量工具为学业压力认知量表(Perceptions of Academic Stress, PAS), 该量表由 Bedewy 和 Gabriel (2015) 编制, 共含 18 个题目, 是用于评估学生在学业情境下感知压力的常用自陈式问卷。采用 Likert 5 点计分, 得分越高表示感知的学业压力水平越大。

为确保数据质量, 问卷中设置了测谎项目, 对部分已作答题目在测验末尾进行重复呈现。若被试在前后两次作答中表现出不一致, 则将其判定为存在不认真作答行为, 通常表现为随机作答(random responding), 并将其归类为异常被试; 若前后作答一致, 将其界定为正常作答被试。该方法基于作答

一致性原则对被试作答质量进行筛查,能够在一定程度上识别低动机或随意作答行为,从而为后续实证分析提供相对可靠的数据基础。在 3500 份在线调查数据中剔除了 735 份不认真作答的答卷,占总样本的 21%。在比较三种估计方法时,将认真作答的 2765 份答卷估计的项目参数作为参考值(等同于模拟研究中的“真值”),然后使用包含异常作答数据的全样本,分别采用 3 种方法进行项目参数估计,并将结果与参考值比较,计算 Bias 与 RMSE 两项指标。

此外,本研究将 2765 名正常作答者在无异常数据干扰条件下的能力估计值作为真值,可视为理想作答情境下的参数估计结果。随后,在包含异常作答的全样本中重新进行能力估计,以检验异常作答对估计精度的影响。由于异常作答会在反应数据中引入额外噪音,从而干扰项目参数和能力参数的准确估计,传统方法在此情境下通常表现出精度下降。相比之下,加权参数估计方法通过降低异常作答在估计过程中的权重,能够有效削弱其不利影响并提升能力估计的准确性。该对比情境不仅更贴近实际测评中数据结构复杂、作答质量参差的特点,也为验证 RRW 方法的稳健性提供了有力实证依据。

表 6 汇总了 3 种参数估计方法的项目参数的 RMSE,从表 6 可知,RRW 加权估计在所有参数(a 参数及 $b1\sim b4$ 难度参数)上的 RMSE 均显著低于传统估计法与 I_z^* 加权估计,尤其是 a 参数 RMSE 仅为 0.101,相较传统方法(0.231)下降了 56.3%,相较 I_z^* 加权(0.184)下降了 45.1%。

对于难度参数而言,RRW 加权估计在 $b1\sim b4$ 的 RMSE 中均显著低于 I_z^* 加权方法;与传统方法相比,除 $b1$ 难度参数外,RRW 在 $b2\sim b4$ 的 RMSE 仍表现出明显优势。就难度参数的平均 RMSE 而言,表 6

最后一列显示,RRW 的数值为 0.187,较传统方法(0.226)下降约 17.3%,较 I_z^* 加权方法(0.296)下降约 36.8%,显示出更优的稳健性。

表 7 汇总了 3 种参数估计方法的项目参数的 Bias,从表 7 可知,RRW 方法的难度参数平均 Bias (-0.007)接近于零,几乎消除了系统性偏倚,而传统方法(-0.096)和 I_z^* 加权(-0.078)均存在一定程度的偏差。对于 a 参数,RRW 的 Bias 为-0.087,虽然存在一定负偏差,但与 RMSE 同时下降的趋势一致,反映了其在控制误差的同时保持较低偏差的优势。

表 8 汇总了三种参数估计方法在能力参数估计中的 Bias 与 RMSE 表现。总体来看,RRW 加权估计方法在 RMSE 指标上表现最佳,显示出更高的估计精度; I_z^* 加权方法次之;传统估计方法的 RMSE 最大,估计精度相对较低。

此外,本研究通过测谎题识别和筛选异常作答,但该方法难以完全捕捉所有异常行为,仍可能存在残余异常反应。这些未被检测出的异常数据可能为能力参数估计引入额外噪音,从而削弱不同估计方法之间的性能差异。

6 讨论和未来研究方向

IRT 在心理与教育测评数据分析中已被广泛应用,其有效性依赖于项目参数能否被准确估计。然而,当测验数据中混入异常作答时,参数估计的精度将受到严重干扰,进而削弱 IRT 的应用价值。在实际应用中,异常作答普遍存在,尤其在缺乏监控的在线问卷调查中更为突出且严重。这类现象可能导致项目参数估计出现显著偏差,并使基于这些参数的后续分析明显背离真实情况,从而削弱研究结论的准确性与解释力。

表 6 三种参数估计方法 RMSE 的比较

方法	RMSE a	RMSE $b1$	RMSE $b2$	RMSE $b3$	RMSE $b4$	RMSE b
传统估计法	0.231	0.129	0.198	0.243	0.303	0.226
I_z^* 加权估计	0.184	0.373	0.289	0.221	0.278	0.296
RRW 加权估计	0.101	0.264	0.133	0.135	0.185	0.187

注: RMSE $b = 4$ 个难度参数的平均 RMSE。

表 7 三种参数估计方法 Bias 的比较

方法	Bias a	Bias $b1$	Bias $b2$	Bias $b3$	Bias $b4$	Bias b
传统估计法	0.188	-0.010	-0.033	-0.133	-0.213	-0.096
I_z^* 加权估计	0.144	-0.129	-0.035	-0.052	-0.098	-0.078
RRW 加权估计	-0.087	-0.183	-0.019	0.043	0.140	-0.007

注: Bias $b = 4$ 个难度参数的平均 Bias。

表 8 三种参数估计方法下正常作答被试能力估计精度比较

方法	Bias	RMSE
传统估计法	0.022	0.163
I_z^* 加权估计	0.037	0.148
RRW 加权估计	0.042	0.129

为应对异常作答对 IRT 项目参数估计精度的不利影响, 本研究提出了一种新的加权估计方法 RRW, 并通过模拟实验和实证数据检验其有效性。研究表明, 异常作答会显著干扰项目参数估计, 且随着异常被试比例与异常题目比例的增加, 传统估计方法的 RMSE 与 Bias 均呈现明显上升趋势, 表现出对失拟数据的高度敏感性。

相比之下, I_z^* 加权方法在一定程度上能够改善区分度参数的估计精度, 但其对难度参数的提升效果有限, 甚至在部分情境下可能引入额外误差, 反映出其稳健性在多级计分模型中的局限性。RRW 加权估计方法在多数异常作答情境下均表现出更好的稳健性, 能够有效降低区分度和难度参数的估计误差。尤其是在异常比例较高的情境下, RRW 方法的优势更加突出, 表明其在复杂失拟条件下具有更强的抗干扰能力。此外, 随着测验长度的增加, RRW 方法的估计结果保持较好的稳定性, 进一步体现了其良好的适用性。

基于真实的心理问卷调查数据进行分析发现, 在 3 种对比的参数估计方法中, RRW 加权估计在项目参数估计误差方面表现良好。同时, 对正常作答被试的能力估计精度比较结果同样表明, RRW 方法优于其他方法。这表明, 在真实测验情境中, 异常作答被试的混入可能导致正常作答者的能力估计产生偏差, 而采用 RRW 加权估计能够有效削弱这种不利影响, 其作用机制在于先提升项目参数的估计精度, 进而提高能力参数的估计精度。

总之, 模拟与实证研究的结果均表明, RRW 加权估计方法在应对复杂异常作答情境时, 能够有效降低其对项目参数估计的不利影响, 展现出良好的应用前景。尽管结果令人鼓舞, 但该方法仍存在一定局限性。

首先, 本研究在多级计分的 GRM 模型框架下验证了 RRW 方法的有效性。理论上, RRW 加权方法依据被试作答行为的异常程度, 对异常反应与正常反应赋予差异化权重, 从而在估计过程中削弱异常反应的影响, 降低其对项目参数估计的干扰。尽管如此, 本文的研究仍主要局限于 GRM 模型框架。

考虑到不同 IRT 模型在参数结构、项目反应函数形式及测量假设等方面存在差异, 后续有必要将 RRW 方法进一步拓展至其他 IRT 模型, 如 GPCM、PCM、NRM、多维 IRT 模型以及 Rasch 类模型, 从而在更为多样化的模型设定与测量情境下, 系统考察其方法表现、稳健性及适用范围。

其次, 反应时间是问卷作答中的关键过程性数据, 能够较好地反映被试的实际作答行为。研究表明, 反应时间能够显著区分正常作答与异常作答 (Choe et al., 2018; Marianti et al., 2014; Man & Harring, 2023; Gorney et al., 2024)。此外, 基于响应时间的混合 IRT 模型已被提出, 用于识别考试过程中被试采用的多种策略, 并有效识别异常作答行为 (Huang et al., 2025)。与此同时, 有研究采用贝叶斯变化点方法, 通过监测响应时间突变来检测异常作答 (Lu et al., 2024; Zhu et al., 2023)。目前, RRW 方法仅依赖作答结果, 后续可考虑将反应时间、作答路径等过程性指标引入权重计算, 构建多源信息融合的加权策略, 以更高效地识别并校正异常作答的影响。

参 考 文 献

- Bedewy, D., & Gabriel, A. (2015). Examining perceptions of academic stress and its sources among university students: The Perception of Academic Stress Scale. *Health Psychology Open*, 2(2), 2055102915596714.
- Boussakuk, M., Bouchboua, A., El Ghazi, M., El Bakkali, M., & Fattah, M. (2021). Designing and developing e-assessment delivery system under IMS QTI ver.2.2 specification. *International Journal of Emerging Technologies in Learning*, 16(1), 219–233.
- Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, 2, 100022.
- Brenner, R. E., Colvin, K. F., Hammer, J. H., & Vogel, D. L. (2021). Using item response theory to develop revised (SSOSH-7) and ultra-brief (SSOSH-3) self-stigma of seeking help scales. *Assessment*, 28(5), 1488–1499.
- Carlozzi, N. E., Kallen, M. A., Troost, J. P., Bragg, A., Martin-Howard, J., Moldovan, I., ... Mitchell, S. E. (2023). Development and calibration data for the Medication Adherence Item Bank: A new computer adaptive test for persons with type 2 diabetes mellitus. *Quality of Life Research*, 32(3), 813–826.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Chauliac, M., Willems, J., Gijbels, D., & Donche, V. (2023). The prevalence of careless response behaviour and its consequences on data quality in self-report questionnaires on student learning. *Frontiers in Education*, 8, 1197324.
- Choe, E. M., Zhang, J., & Chang, H. H. (2018). Sequential detection of compromised items using response times in

- computerized adaptive testing. *Psychometrika*, 83(3), 650–673.
- Demirkol, S., & Kelecioğlu, H. (2022). Investigating the effect of item position on person and item parameters: PISA 2015 Turkey sample. *Journal of Measurement and Evaluation in Education and Psychology*, 13(1), 69–85.
- Deng, J., & Rios, J. A. (2022). Investigating the effect of differential rapid guessing on population invariance in equating. *Applied Psychological Measurement*, 46(7), 589–604.
- De La Torre, J., & Deng, W. (2008). Improving person - fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, 45(2), 159–177.
- Doval, E., & Delicado, P. (2020). Identifying and classifying aberrant response patterns through functional data analysis. *Journal of Educational and Behavioral Statistics*, 45(6), 719–749.
- Eichenbaum, A. E., Marcus, D. K., & French, B. F. (2019). Item response theory analysis of the Psychopathic Personality Inventory-Revised. *Assessment*, 26(6), 1046–1058.
- Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, 8, 863.
- Gao, X., & Chen, Y. (2025). Comparison of resilience measures in Chinese adolescents: Based on item response theory. *Journal of Personality Assessment*, 107(5), 611–620.
- Gao, X., Hou, M., Wang, F., & Zhou, J. (2025). A new person-fit statistic for the detection of aberrant responses in polytomous cognitive diagnostic models. *Behavior Research Methods*, 57(5), 138.
- Gershon, S. A. K., Anghel, E., & Alexandron, G. (2024). An evaluation of assessment stability in a massive open online course using item response theory. *Education and Information Technologies*, 29(3), 2625–2643.
- Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, 31(4), 101384.
- Gorney, K., Sinharay, S., & Liu, X. (2024). Using item scores and response times in person - fit assessment. *British Journal of Mathematical and Statistical Psychology*, 77(1), 151–168.
- Hong, M. R., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods*, 51(2), 573–588.
- Huang, S., Luo, J., & Jeon, M. (2025). A response time-based mixture item response theory model for dynamic item-response strategies. *Behavior Research Methods*, 57(1), 54.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277–298.
- Lu, J., Wang, C., Zhang, J., & Wang, X. (2024). A sequential Bayesian changepoint detection procedure for aberrant behaviours in computerized testing. *British Journal of Mathematical and Statistical Psychology*, 77(1), 31–54.
- Luo, Y., Chen, Q., Chen, J., & Zhan, P. (2024). Development and validation of two shortened anxiety sensitive index-3 scales based on item response theory. *Humanities and Social Sciences Communications*, 11(1), 1078.
- Man, K., & Haring, J. R. (2023). Detecting preknowledge cheating via innovative measures: A mixture hierarchical model for jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, 83(5), 1059–1080.
- Manapat, P. D., Edwards, M. C., MacKinnon, D. P., Poldrack, R. A., & Marsch, L. A. (2021). A psychometric analysis of the Brief Self-Control Scale. *Assessment*, 28(2), 395–412.
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83.
- Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39(6), 426–451.
- Marcq, K., & Braeken, J. (2025). From framework to functionality: A cross-country analysis of PISA 2018 reading assessment framework's item features as determinants of item difficulty. *Large-scale Assessments in Education*, 13(1), 26.
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.
- Miles, M., Nadarajah, V., Jauregui, J. J., Dubina, A. G., Smuda, M. P., Bennett, C. H., ... Henn, R. F., III. (2020). Evaluation of the PROMIS physical function computer adaptive test in patients undergoing knee surgery. *The Journal of Knee Surgery*, 33(8), 810–817.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45, 867–872.
- Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31(3), 200–219.
- Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, 44(3), 309–341.
- Phillips, J. L., Freedman, M. K., Simon, J. I., & Beredjikian, P. K. (2021). The PROMIS upper extremity computer adaptive test correlates with previously validated metrics in patients with carpal tunnel syndrome. *Hand*, 16(2), 164–169.
- Polat, M., & Kelecioğlu, H. (2024). The comparison of PISA 2015–2018 mathematics trend items based on item response times. *Journal of Measurement and Evaluation in Education and Psychology*, 15(3), 183–192.
- Rios, J. A. (2022). A comparison of robust likelihood estimators to mitigate bias from rapid guessing. *Applied Psychological Measurement*, 46(3), 236–249.
- Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, 81(4), 992–1013.
- Tong, H., Yu, X., Qin, C., Peng, Y., & Zhong, X. (2022). Detection of aberrant response patterns using a residual-based statistic in testing with polytomous items. *Acta Psychologica Sinica*, 54(9), 1122–1136.
- [童昊, 喻晓峰, 秦春影, 彭亚风, 钟小缘. (2022). 多级计分测验中基于残差统计量的被试拟合研究. *心理学报*, 54(9), 1122–1136.]
- Sharp, C., Steinberg, L., McLaren, V., Weir, S., Ha, C., & Fonagy, P. (2022). Refinement of the reflective function questionnaire for youth (RFQY) scale B using item response theory. *Assessment*, 29(6), 1204–1215.
- Voss, N. M. (2024). The effects of careless responding on the fit of confirmatory factor analysis and item response theory

- models. *Behavior Research Methods*, 56(2), 577–599.
- Wang, F., & Gao, X. (2025). Psychometric properties of the short mood and feelings scale among Chinese adolescents using item response theory. *Frontiers in Psychiatry*, 16, 1561728.
- Wang, X., Luo, F., Cai, Y., & Tu, D. (2024). Detection of back random responding in forced-choice questionnaires: The change point analysis method. *Journal of Psychological Science*, 47(6), 1507–1518.
- [王雪, 罗芬, 蔡艳, 涂冬波. (2024). 迫选测验中后期随机作答的侦查: 基于变点分析法. *心理科学*, 47(6), 1507–1518.]
- Williams, Z. J., & Gotham, K. O. (2021). Assessing general and autism-relevant quality of life in autistic adults: A psychometric investigation using item response theory. *Autism Research*, 14(8), 1633–1644.
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort - moderated IRT model. *Journal of Educational Measurement*, 43(1), 19–38.
- Yu, X., Tang, Q., Qin, C., & Li, Y. (2024). Person-fit in cognitive diagnostic assessment. *Journal of Psychological Science*, 47(3), 744–751.
- [喻晓锋, 唐茜, 秦春影, 李喻骏. (2024). 认知诊断评价中的被试拟合研究. *心理科学*, 47(3), 744–751.]
- Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*, 24(5), 658–674.
- Zarate, D., Hobson, B. A., March, E., Griffiths, M. D., & Stavropoulos, V. (2023). Psychometric properties of the Bergen Social Media Addiction Scale: An analysis using item response theory. *Addictive Behaviors Reports*, 17, 100473.
- Zhu, H., Jiao, H., Gao, W., & Meng, X. (2023). Bayesian change-point analysis approach to detecting aberrant test-taking behavior using response times. *Journal of Educational and Behavioral Statistics*, 48(4), 490–520.

A residual-based weighted robust item parameter estimation method for aberrant responses

GAO Xuliang, LI Tingting, WANG Fang

(School of Psychology Guizhou Normal University; Mental Health Education and Counseling Center, Guizhou Normal University, Guiyang 550025, China)

Abstract

Item Response Theory (IRT) models are widely used in psychological and educational measurement. Their effectiveness, however, can be undermined by data quality issues, particularly aberrant responding, such as cheating, careless errors, rapid guessing, and hybrid forms involving multiple aberrant response behaviors. Such responses are especially prevalent in online assessments, where the absence of supervision, external distractions, and low participant motivation can introduce substantial measurement noise. Empirical evidence indicates that even 10~15% proportion of aberrant responses can meaningfully distort parameter estimates, while higher proportions (20~30% or more) can lead to severe bias in item and ability parameter estimates, reduce reliability, distort factor structures, and compromise measurement invariance test.

Person-fit statistics (PFS) are widely used to identify aberrant respondents. However, their dependence on full-sample parameter estimates makes them susceptible to the masking effect, in which aberrant responses bias item parameter estimation and, in turn, obscure person misfit, particularly when the proportion of aberrant responses is high. To overcome this limitation, a robust marginal maximum likelihood method has been proposed that incorporates PFS-derived weights into the likelihood function, thereby down-weighting aberrant respondents during parameter estimation. Using the l_z^* statistic, this method can effectively reduce bias while retaining all response data.

Building on this framework, the present study introduces a new robust weighting method, namely Residual-Based Robust Weighting (RRW), which replaces l_z^* with a residual-based person-fit statistic (PFS). Residual-based statistics quantify item-level deviations between observed and expected responses, standardized by their variances and aggregated across items. This finer-grained approach improves sensitivity to subtle and heterogeneous aberrant patterns compared to the aggregated likelihood ratio underlying l_z^* . As a result, RRW assigns lower weights to respondents whose residual patterns deviate substantially from model expectations, while preserving the full contribution of respondents whose responses are consistent with the model.

We evaluated the proposed RRW method through simulation studies and an empirical example, comparing its performance with unweighted estimation and the I_z^* -based RMML approach. Results showed that aberrant responding substantially distorts item parameter estimation, with the unweighted method exhibiting the largest increases in RMSE and bias as aberrance levels rise. The I_z^* -based method improved discrimination estimates to some extent but showed limited effectiveness for threshold parameters. In contrast, RRW consistently achieved lower estimation errors for both discrimination and threshold parameters across most conditions, with its advantages becoming more pronounced under higher aberrance levels. Moreover, RRW maintained stable performance as test length increased, demonstrating strong robustness and applicability in complex testing scenarios.

Empirical results further validated RRW's superiority: it reduced item parameter estimation error to a certain extent and enhanced ability estimates for non-aberrant respondents, demonstrating its effectiveness in mitigating the adverse impact of aberrant responses in real testing contexts. By improving item parameter accuracy, RRW indirectly increases the precision of ability estimation, providing a practical and robust alternative for IRT applications in complex testing environments. Overall, evidence from both simulation and empirical analyses indicates that the RRW weighted estimation method effectively reduces the negative effects of complex aberrant response patterns on item parameter estimation and shows strong potential for practical implementation.

Keywords item response theory, aberrant responding, robust item parameter estimation