心理学报 2026, Vol. 58, No. 5, 995−1014 https://doi.org/10.3724/SP.J.1041.2026.0995
Acta Psychologica Sinica https://cstr.cn/32110.14.2026.0995

认知诊断评估中主效应 DIF 与交互式 DIF 检测方法开发：基于递归分割视角*

刘 凯[1,2] 郭治辰[1] 王 琴[1] 汪大勋[1] 蔡 艳[1] 涂冬波[1,3,4]

([1]江西师范大学心理学院, 南昌 330022) ([2]辽宁师范大学心理学院, 大连 116029)
([3]江西省哲学社会科学实验室—江西师范大学数据科学与智能化心理测评及服务实验室;
[4]智能信息处理与情感计算江西省重点实验室, 南昌 330022)

摘 要 在认知诊断评估中, 项目功能差异(DIF)检测是评估其测验公平性以及测量效度的重要技术手段。然而, 现有的认知诊断 DIF 检测方法局限于单一协变量视角下的主效应 DIF 检测, 对于由多个协变量交互作用引发的交互式 DIF 尚缺乏有效的检测手段。针对这一局限, 研究借鉴递归分割技术的核心思想, 提出了一种能够在认知诊断评估中同时检测主效应 DIF 和交互式 DIF 的新方法(记为 ISRPM)。模拟研究结果表明, ISRPM 不仅在主效应 DIF 检测中的整体表现与传统方法大体相当, 更重要的是, 其在交互式 DIF 检测方面的性能表现优于传统方法。实证研究则进一步支持了该方法的可用性, 结果显示, ISRPM 与传统 DIF 检测方法在检测结果上具有较高一致性, 并在识别交互式 DIF 方面展现出潜在优势。总体而言, ISRPM 的提出有望进一步提升认知诊断 DIF 检测精度, 并促进认知诊断评估在心理与教育测评实践中的推广与应用。

关键词 认知诊断评估, 项目功能差异, 主效应 DIF, 交互式 DIF, 递归分割技术

分类号 B841

1 引言

在当今迅速演变的信息社会中, 认知诊断(Cognitive Diagnosis, CD)的兴起为心理与教育测量领域带来了显著而实质性的变革(DiBello et al., 2006; Leighton & Gierl, 2007; Nichols et al., 1995)。不同于传统的测验分数评价体系, 认知诊断不仅关注个体能力的总体水平, 更着重揭示个体内部心理加工过程和认知结构, 从而更好地实现测验促进个体发展的核心目标(涂冬波 等, 2019)。在心理评估领域, 认知诊断不仅能够用于评估个体的认知功能状态, 还能准确识别其症状特征, 从而为临床医生实施精准治疗与早期干预提供重要数据支持(de la Torre et al., 2018; Tan et al., 2023; Templin & Henson, 2006)。在教育测量领域, 传统的学业成就测验通常依据总分或等级评定学生能力, 而认知诊断则更加关注学习过程本身, 它通过精准定位个体在不同认知成分上的优势与不足, 为教育工作者制定有针对性的教学策略与知识补救方案提供有效参考(Rupp et al., 2010)。总体而言, 在信息技术迅速发展的时代背景下, 认知诊断不仅拓宽了心理与教育测量的研究视角, 也为实施个性化教学与精准心理治疗提供了有效的技术支撑。

近年来, 认知诊断凭借其在提供精细粒度诊断信息方面的独特优势, 已成为国内外心理与教育测量领域的研究前沿, 并被广泛应用于心理与教育测验开发(如: 王欣瑜, 2019; Li et al., 2020; Wang et al., 2019; Xi et al., 2020)。在认知诊断测验的开发

收稿日期: 2024-04-26

* 国家自然科学基金项目(32300942, 62467002, 62167004, 32160203)、江西省科技创新基地计划—智能信息处理与情感计算江西省重点实验室(20242BCC32021)资助。

郭治辰和王琴为本文共同第一作者。

通信作者: 蔡艳, E-mail: cy1979123@aliyun.com; 汪大勋, E-mail: wangda.xun@163.com; 涂冬波, E-mail: tudongbo@aliyun.com

过程中, 测验开发者尤为关注的一个问题是测量结果是否对特定群体产生系统性偏差, 从而导致该群体在测验中获得不当的优势或处于不利地位。该问题实质上涉及到对于测验公平性的评估。在心理测量学框架下, 与测验公平性密切相关的概念主要包括测量不变性(measurement invariance, MI; Meredith, 1993)和项目功能差异(differential item functioning, DIF; Holland & Wainer, 1993)。测量不变性指的是测验在不同被试群体(如性别和文化背景)之间保持一致的测量特性。当测验在不同群体间的测量特性存在系统性差异时, 即表明该测验存在测量非不变性(measurement noninvariance, MN)。而当这种非不变性表现在题目水平上时, 则意味着该题目存在DIF。在认知诊断框架下, DIF 通常被定义为: 在属性掌握模式相同的条件下, 来自不同群体的被试在同一题目上的正确作答概率存在系统性差异(Hou et al., 2014; Li, 2008)。已有研究表明, DIF 的存在不仅会削弱认知诊断测验的测量公平性, 还可能降低测量效度(Hou et al., 2014)。此外, DIF 还可能会导致项目参数的有偏估计, 从而引起被试属性掌握模式的误判, 最终导致带有一定误导性的评估结果(Paulsen et al., 2020)。因此, 在认知诊断测验的开发与验证阶段进行 DIF 分析, 已成为心理测量学研究者普遍认可的关键环节(Hou et al., 2014; Li & Wang, 2015; Ma et al., 2021)。这一过程不仅是测验质量控制的重要组成部分, 也是确保测验公平性与测量效度的必要条件。

目前, 国内外研究者已提出多种适用于认知诊断评估的 DIF 检测方法, 这些方法主要分为两类: 参数化和非参数化方法。非参数化方法具有对样本容量要求低、操作简便直观以及易于理解的优点, 然而, 相比于参数化方法, 其检测精度通常较低。此类方法的典型代表为 Zhang (2006)开发的 Mantel-Haenszel 法以及 SIBTEST 法。参数化方法在使用过程中需要估计特定认知诊断模型的参数, 虽然其操作相对复杂且计算成本较高, 但在 DIF 检测结果的准确性方面表现更佳。典型的参数化方法包括: Wald 检验法(Hou et al., 2014; Li, 2008; Ma et al., 2021)、Logistic 回归法(王卓然 等, 2014)以及似然比检验法(Ma et al., 2021)。值得一提的是, 上述参数化方法的检测性能已在模拟研究中得到充分验证, 并为认知诊断测验的 DIF 分析提供了可靠的理论与技术支持。鉴于参数化方法在 DIF 检测中的精度优势, 近年来其相比于非参数化方法得到了更多的关注。因此, 本文聚焦于参数化认知诊断 DIF 检测方法的开发。

尽管现有的认知诊断 DIF 检测方法在性能方面整体表现良好, 但它们仍存在一定的局限性。具体来讲, 这些方法通常仅能独立评估单个协变量是否会引发 DIF, 而未充分考虑到多个协变量间的交互作用也可能会导致 DIF 的出现。例如, 性别与户籍的交互作用可能影响被试在特定题目上的作答模式, 从而导致 DIF 的出现。为了明确区分由多个协变量的交互作用引发的 DIF 与仅由单一协变量引起的 DIF, 本文将前者定义为“交互式 DIF (interactive DIF)”, 而将后者定义为“主效应 DIF (main effect DIF)”。已有研究表明, 交互式 DIF 可能普遍存在于心理与教育测验中, 并进一步增加测量偏差来源的复杂性。例如, Bauer (2017)在评估青少年犯罪行为的心理测验中发现, 由协变量之间的交互作用所导致的测量偏差较为明显。类似地, Tutz 和 Berger (2016)在智力结构测验中发现, 某些题目的 DIF 可能源于性别和年龄的交互作用。因此, 我们有理由推测, 交互式 DIF 也极有可能存在于认知诊断测验中, 并可能对测验公平性和测量效度产生不利影响。

近些年来, 众多研究者强调了交互式 DIF 检测在测验质量分析中的重要性及其理论与实践价值(Belzak, 2023; Strobl et al., 2015; Tutz & Berger, 2016)。首先, 交互式 DIF 检测有助于更加全面地揭示导致测验产生功能差异的复杂来源(Tay et al., 2015), 从而为测验开发者提供更具参考价值的题目修订依据。Collins (1990)在其研究中指出, 个体身份是多个人口学特征交叉作用的结果。这一观点提示, 在探讨测量偏差时, 除了考虑单一人口学变量的主效应之外, 还应兼顾变量之间可能存在的交互作用。基于这一认识, 与仅从单一协变量视角出发的主效应 DIF 检测相比, 开展交互式 DIF 检测有助于识别出更多潜在的测量偏差来源, 从而为测验开发者提供更精准、更具针对性的题目修订依据。其次, 交互式 DIF 的隐蔽性使其检测变得尤为必要。与主效应 DIF 相比, 交互式 DIF 更难被传统方法识别。这主要是因为传统 DIF 检测方法通常假设不同的人口学变量之间相互独立, 所以它们只能识别由单一协变量引起的 DIF 形式, 而难以揭示由多个协变量交互作用导致的测量偏差。在测验开发阶段, 研究者往往基于单一人口学变量(如性别、种族)预先假设测量偏差的可能来源, 而忽视了变量间交互作用对测试公平性可能产生的影响。这种忽视可

能导致部分测量偏差被遗漏, 从而影响测验的公平性。鉴于此, 准确识别交互式 DIF 有助于进一步提升认知诊断评估结果的公平性。最后, 交互式 DIF 的识别对于进一步提升测量效度同样具有重要价值。交互式 DIF 的存在不仅可能影响被试在特定题目上的作答表现, 还可能阻碍对其属性掌握模式的准确估计。如果未能识别由协变量间交互作用引发的 DIF, 认知诊断评估结果的质量将受到削弱。综上所述, 交互式 DIF 检测对于全面揭示测量偏差来源、保障测验公平性以及提升测量效度均具有重要价值。然而, 现有的认知诊断 DIF 检测方法在识别交互式 DIF 方面仍存在明显不足, 这对于测验公平性与测量效度的有效保障构成了挑战。因此, 在认知诊断框架下开发能够同时识别主效应 DIF 与交互式 DIF 的方法, 不仅有助于完善认知诊断 DIF 检测的理论与方法体系, 而且对于促进认知诊断评估在实践中的合理应用具有重要价值。本研究正是基于这一研究需求, 旨在提出一种能够同时识别主效应 DIF 与交互式 DIF 的方法, 从而为认知诊断测验的公平性评估提供更全面的技术支持。

随着现代数据处理技术的持续发展, 数据挖掘(Data Mining, DM)技术在心理与教育测量领域得到了广泛应用。Belzak 和 Bauer (2020)指出, 在进行 DIF 分析时, 识别导致 DIF 的协变量过程与回归建模中的变量选择(variable selection)具有高度相似性, 这为将变量选择方法引入 DIF 检测提供了理论基础, 也为 DIF 检测方法的改进与创新提供了新的思路。与传统的 DIF 检测方法相比, DM 技术具有效率高、灵活性强以及能够同时处理多个协变量等优势, 在识别复杂 DIF 形式方面展现出较大的应用潜力(Belzak, 2023)。基于变量选择方法在 DIF 检测中的应用潜力, 研究者开始将项目反应理论(Item Response Theory, IRT)与变量选择技术相结合, 开发出了一系列能够识别复杂 DIF 形式的新方法(如 Bollmann et al., 2018; Strobl et al., 2015; Tutz & Berger, 2016), 为认知诊断情境下开展交互式 DIF 检测提供了重要的方法借鉴。其中, 递归分割(Recursive Partitioning, RP)是这类方法中最具代表性的变量选择技术。该技术的基本原理是将预测变量所覆盖的特征空间递归地划分为若干子区域, 并在每个区域内拟合一个相对简单的模型(Hothorn et al., 2006)。通过持续执行数据分割与建模, RP 方法能够直观地揭示协变量的主效应及其交互作用与测验项目参数之间的关系, 从而为探索测量偏差的复杂来源提供有效的技术支持。根据 Strobl 等人(2015)、Tutz 和 Berger (2016)以及 Bollmann 等人(2018)的研究, RP 技术在 DIF 分析中的优势具体体现在以下三个方面: (1)它突破了传统方法需在分析前人为划分目标组和参照组的限制, 能够以数据驱动的方式自动识别出最优的协变量分组标准, 从而降低因分组设置不当而导致潜在 DIF 被漏检的风险。(2)传统方法通常难以考察协变量之间的交互作用, 因此在交互式 DIF 检测中存在较大局限; 相比之下, RP 技术不仅能够揭示多个协变量之间的复杂交互作用, 还能进一步评估其对测量偏差的影响, 从而有助于提升交互式 DIF 的检测精度。(3) RP 技术能够灵活处理多种类型的协变量, 包括连续型、多类别型、有序型及二分类变量, 从而拓宽了 DIF 检测方法的适用范围。综合来看, RP 技术以其灵活、直观且稳健的特性, 为 DIF 检测提供了有效框架, 并在处理协变量间交互作用的复杂情境中展现出独特优势。

到目前为止, 基于 RP 技术的 DIF 检测方法开发主要在项目反应理论(Item Response Theory, IRT)框架下开展。这类方法可大致分为两类: 全局水平的 RP 方法和项目水平的 RP 方法。全局水平的方法通过检验协变量所覆盖空间范围内的参数不稳定性(parameter instability), 来判断单个协变量的主效应以及多个协变量之间的交互作用是否会导致测量偏差。然而, 此类方法仅能识别出导致 DIF 的协变量, 而无法进一步确定具体存在 DIF 的题目。在已有研究中, Rasch Trees (Strobl et al., 2015)和多级计分的 Rasch Trees (Komboz et al., 2016)成为了此类方法的典型代表。与全局水平方法相比, 项目水平 RP 方法不仅能识别诱发 DIF 的协变量, 还能定位出具体存在 DIF 的题目, 因此在 DIF 检测中表现出更高的灵活性与实用性。此类方法的典型代表包括基于 Rasch 模型的 item-focused trees (IFT; Tutz & Berger, 2016)以及基于分部评分模型的 item-focused trees (PCM-IFT; Bollmann et al., 2018)。目前, 项目水平的 RP 方法因其在识别协变量和定位 DIF 题目方面的双重优势而受到更多关注。这具体体现在: 此类方法可在单次分析中同时处理多个协变量(Finch et al., 2015), 并在题目水平上进一步探索这些协变量之间的交互作用是否诱发功能差异。

尽管已有研究证实 RP 技术在主效应 DIF 和交互式 DIF 检测中具有较大的研究潜力与应用价值, 但现有研究工作仅在 IRT 框架下验证了其在 DIF 检

测中的有效性。更重要的是, 目前尚无公开发表的研究系统探讨认知诊断测验中的交互式 DIF 检测。值得注意的是, IRT 与认知诊断理论在模型假设、测量目标和数据分析方法等方面均存在明显区别, 这使得直接将 RP 技术应用于认知诊断情境面临新的挑战。因此, 如何将 RP 技术拓展至认知诊断评估中的主效应 DIF 与交互式 DIF 检测, 并验证其能否在认知诊断评估中保持既有的检测精度与适用性, 仍是有待深入探讨的重要问题。基于此, 本研究在借鉴 RP 技术核心思想的基础上, 开发并验证了一种适用于认知诊断测验的新型 DIF 检测方法。该方法旨在为认知诊断评估中主效应 DIF 和交互式 DIF 的识别提供有效的技术支持, 从而进一步完善认知诊断的测验公平性评估体系, 并最终推动认知诊断技术在心理与教育测量领域的深入应用。

2 基于项目水平的序列递归分割法(ISRPM)的开发

本文的主要目的是在认知诊断框架下开发出一种新颖的主效应 DIF 及交互式 DIF 检测方法, 即基于项目水平的序列递归分割法(Item-based Sequential Recursive Partitioning Method, ISRPM)。该方法将递归分割技术与认知诊断模型相结合, 将依据模型参数估计构建的 DIF 检验统计量视为协变量分割标准, 通过比较不同协变量分割方案的分割效果以识别最优分割方法, 从而为每个题目生成一棵能够反映其 DIF 表现形式的递归分割树。具体而言, ISRPM 在树的每一层比较候选协变量的潜在分割方案所对应的统计量, 选择能使组间项目参数差异最大化的协变量及分割点, 并据此递归划分被试群体、逐层扩展树结构, 最终揭示哪些协变量以及以何种作用方式可能诱导 DIF 的产生。下文首先简要介绍研究所采用的认知诊断模型及其 DIF 定义, 随后阐述 ISRPM 的操作步骤与理论基础。

2.1 拓广 DINA (G-DINA)模型

在认知诊断框架下, 认知诊断模型(Cognitive Diagnosis Model, CDM)是一类充分融合认知变量的心理计量模型。作为认知诊断评估的核心技术环节, CDM 的质量直接决定了认知诊断结果的有效性(涂冬波 等, 2019)。目前, 研究者已经开发出多种具有良好诊断性能且能够适用于不同测验情境与理论假设的 CDM。本研究采用拓广 DINA (generalized deterministic input, noisy “and” gate, G-DINA)模型。该模型由 de la Torre (2011)在决定型输入、噪音与门模型(deterministic input, noisy “and” gate, DINA)的基础上进行拓展而提出的一种广义计量模型。

在 G-DINA 模型中, 被试 i 完成对于项目 j 的作答后会被细分为 $2^{K_j^*}$ 种分类, 其中 K_j^* 为项目 j 所测量的属性个数。为了便于介绍其数学表达式, 可以假设前 K_j^* 个属性是正确作答项目 j 时所必须掌握的属性, $\boldsymbol{\alpha}_{vj}^*$ 表示第 v 种被试属性掌握模式, 其中 $v=1,\cdots,2^{K_j^*}$。那么, 在 G-DINA 模型框架下, 被试 i 在项目 j 上正确作答的条件概率可以表示为

$$f[P(\boldsymbol{\alpha}_{vj}^*)]=\delta_{j0}+\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{vk}+\sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{vk}\alpha_{vk'}+\cdots+\delta_{j12\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{vk} \quad (1)$$

其中, $P(\boldsymbol{\alpha}_{vj}^*)=P(X_{ij}=1|\boldsymbol{\alpha}_{vj}^*)$; $f[\cdot]$ 为连接函数(link function), 根据所采用连接函数的不同, G-DINA 模型有着不同的表示形式, 常用的连接函数有三种: 一致性、log 和 logit 连接函数; δ_{j0} 为项目 j 的截距项, 它是当被试未掌握测量的所有属性时而答对项目 j 的概率, 该值一般为非负值; δ_{jk} 为项目 j 上属性 k 的主效应, 一般为非负值, 该值越大说明掌握该属性对于答对项目 j 的贡献越大; α_{vk} 表示第 v 种属性掌握模式代表的被试在属性 k 上的掌握情况, 掌握该属性记为 1, 否则为 0; $\delta_{jkk'}$ 为项目 j 上属性 k 和属性 k' 的交互效应; $\delta_{j12\cdots K_j^*}$ 为项目 j 测量的所有属性之间的交互效应。需要强调的是, 为了与以往同类研究(如: 孙小坚 等, 2022; Hou, 2013; Ma et al., 2021)保持一致, 本研究采用了一致性连接函数。

2.2 G-DINA 模型下的 DIF 定义

不同于 IRT 框架下的 DIF 检测, CDM 中的 DIF 需要被重新定义。这是因为 CDM 提供的是被试在离散属性上的掌握情况, 而不是将被试定位在潜在特质水平这一连续体上。根据 Hou 等人(2014)的观点, 在 G-DINA 模型框架下, DIF 可被表示为以下公式:

$$\Delta_{j\alpha_v}=P(X_j=1|\alpha_v)_F-P(X_j=1|\alpha_v)_R\neq 0, \quad (2)$$

其中, $P(X_j=1|\alpha_v)_F=P(\alpha_v)_F$ 和 $P(X_j=1|\alpha_v)_R=P(\alpha_v)_R$ 分别表示目标组(focal group, F 组)和参照组(reference group, R 组)中属性掌握模式为 α_v 的被试在项目 j 上正确作答的概率; $\Delta_{j\alpha_v}=0$ 表示项目 j 不存在 DIF, 反之, 则存在 DIF。

2.3 基于项目水平的序列递归分割法(ISRPM)的开发: 基本原理与检测步骤

本节将详细介绍 ISRPM 的操作流程及其关键技术细节。ISRPM 将递归分割技术与 G-DINA 模型

充分融合, 通过递归分割协变量的方式实现对主效应 DIF 与交互式 DIF 的识别。该方法的基本思路如下: 对于每个题目, 在给定一组候选协变量后, ISRPM 首先会识别出各协变量所有可能的分割点, 并依据这些分割点将被试作答数据划分为若干子样本(如不同性别的作答数据集)。随后, 在各子样本上分别拟合 G-DINA 模型并估计模型参数。接下来, 将所得参数代入预设的分割标准(即 DIF 检验统计量)中进行计算, 并将统计量最大值对应的分割方案(即协变量及其分割点)确定为首个最优分割变量及其分割点, 依据该方案执行初始分割。完成首轮分割后, ISRPM 会在新生成的子样本中重复上述过程, 在每一层中依次确定最优分割方案并执行数据分割, 直至满足预设的变量搜索终止规则。最终, ISRPM 为每个被判定为存在 DIF 的题目生成相应的递归分割树图, 以可视化的形式展示该题目在哪些协变量上以何种方式表现出 DIF。下文将以单个题目 j 为例, 详细介绍 ISRPM 的主要操作步骤及其关键技术细节。

步骤 1: 确定感兴趣的协变量及其潜在分割点

在进行认知诊断 DIF 检测之前, 首先需要明确所有感兴趣的协变量及其潜在分割点。ISRPM 的目标是为每个存在功能差异的题目构建出递归分割树, 其中, 树的根节点包括全部感兴趣的协变量及其变量水平, 这些变量水平共同构成了协变量所覆盖的完整特征空间, 而子节点则对应该特征空间中的子集。从被试分组逻辑上看, 这种基于特征空间逐步细分的结构与传统 DIF 检测中依据协变量水平进行被试分组的思想是一致的。在传统的 DIF 检测中, 研究者会根据每个感兴趣的协变量及其变量水平将被试划分为目标组(Focal group, F 组)和参照组(Reference group, R 组)。通常, 数字 1 和 2 分别代表目标组和参照组, 它们对应于同一人口学特征的不同特征水平。假设 $\boldsymbol{x}^T=(x_1,\cdots,x_m)$ 表示与被试相关的 m 个协变量, 且每个协变量都包含两个水平。对于二分协变量 x_m 而言, 可以根据其两个水平定义出对应的子集 A_1 和 A_2, 并可表示为

$$A_1=A\cap\{x_m=1\}\text{和}A_2=A\cap\{x_m=2\} \tag{3}$$

其中, A 代表所有协变量共同覆盖的整个特征空间, $x_m=1$ 和 $x_m=2$ 分别对应由协变量 x_m 所定义出的目标组和参照组。由于该变量仅包含两个水平, 因此它的潜在分割点有且仅有一个。

步骤 2: 基于选定协变量分割标准的首个最佳分割协变量搜索

在确定感兴趣的协变量及其潜在分割点后, ISRPM 需对各潜在分割方案对应的分组数据集进行参数估计。为了确保估计结果可靠, 本研究采用经广泛验证的 G-DINA 模型完成这一估计过程。在获得参数估计结果后, 关键任务是确定合适的协变量分割标准以搜索出首个最佳分割协变量, 并据此判断当前题目是否在该变量上表现出 DIF。根据 Strobl 等人(2009)的研究, 递归分割框架中常用的分割标准主要包括两类: (1)基于不纯度测量(impurity measures)的分割, 如基尼指数(Gini Index)或香农熵(Shannon Entropy); (2)基于检验统计量的分割, 如对数似然统计量(Log-likelihood test statistic)。前者通过衡量节点内样本的同质性实现分割, 而后者通过对参数差异进行统计检验, 根据组间差异判断是否需要进一步划分节点。在参数化认知诊断 DIF 检测的情境下, 基于检验统计量的分割更契合研究目标, 因此本研究将此类方法作为 ISRPM 的协变量分割标准。

在搜索首个最佳分割协变量时, ISRPM 会针对每个候选协变量 x_m 的所有潜在分割点生成候选子节点, 这些节点构成了递归分割树首层的左侧和右侧潜在子节点。基于协变量 x_m 的所有特征水平, 可分别获得由其潜在分割点定义的目标组与参照组数据集, 并依据公式(4)在各组上拟合 G-DINA 模型以估计对应的项目参数。

$$\begin{cases} P(\boldsymbol{\alpha}_{vj}^*)=\text{node}_{\delta_{j0}}+\sum\limits_{k=1}^{K_j^*}\text{node}_{\delta_{jk}}\alpha_{vk}+ \\ \quad\sum\limits_{k'=k+1}^{K_j^*}\sum\limits_{k=1}^{K_j^*-1}\text{node}_{\delta_{jkk'}}\alpha_{vk}\alpha_{vk'}+\cdots+\text{node}_{\delta_{j12\cdots K_j^*}}\prod\limits_{k=1}^{K_j^*}\alpha_{vk} \\ \text{node}_{\delta_{j0}}=[\delta_{lj0}\cdot I(x_m=1)+\delta_{rj0}\cdot I(x_m=2)] \\ \text{node}_{\delta_{jk}}=[\delta_{ljk}\cdot I(x_m=1)+\delta_{rjk}\cdot I(x_m=2)] \\ \text{node}_{\delta_{jkk'}}=[\delta_{ljkk'}\cdot I(x_m=1)+\delta_{rjkk'}\cdot I(x_m=2)] \\ \vdots \\ \text{node}_{\delta_{j12\cdots K_j^*}}=[\delta_{lj12\cdots K_j^*}\cdot I(x_m=1)+\delta_{rj12\cdots K_j^*}\cdot I(x_m=2)] \end{cases} \tag{4}$$

其中, node 表示题目 j 在递归分割树中不同子节点下的 G-DINA 模型项目参数取值; $I(\cdot)$ 为指示函数, 如果满足条件 d, 那么 $I(d)=1$, 否则 $I(d)=0$; 下标 l 和 r 分别为英文单词 *left* 和 *right* 的缩写; $\{x_m=1\}$

和 $\{x_m = 2\}$ 分别表示由协变量 x_m 所定义的目标组和参照组; $\delta_{lj0} \cdot I(x_m = 1)$ 表示依据协变量 x_m 建立出的左侧子节点[即 $I(x_m = 1)$]处所对应的截距参数, 也可以视作由 x_m 定义出的目标组在当前题目上的截距参数, 其余依次类推。

在获得当前题目在各候选协变量 x_m 下的分组项目参数估计值后, 需要进一步计算该题目在每个协变量的各潜在分割点下对应的 DIF 检验统计量。一般而言, 可以采用任何在认知诊断 DIF 检测研究中已被验证有效的检验统计量。以往研究发现, 在认知诊断框架下, 相较于 MH、SIBTEST 及似然比检验等方法, 基于 Wald 统计量的检测方法在一类错误率控制与统计检验力方面表现更优(孙小坚等, 2022), 因此本研究选取 Wald 统计量作为协变量分割标准。Wald 统计量服从卡方分布, 其数学表达式为:

$$W_j = (\xi_{Rj} - \xi_{Fj})^T (\Sigma_{Rj} + \Sigma_{Fj})^{-1} (\xi_{Rj} - \xi_{Fj}) \quad (5)$$

其中, ξ_{Rj} 和 ξ_{Fj} 分别表示 G-DINA 模型框架下参照组(R)和目标组(F)在项目 j 上的项目参数估计向量, Σ_{Rj} 和 Σ_{Fj} 分别为参照组与目标组在项目 j 上的项目参数估计值所对应的抽样方差—协方差矩阵(Sampling Variance-Covariance Matrix)。值得注意的是, 在利用 Wald 统计量进行 DIF 检测之前, 通常需要实施题目提纯(item purification), 其目的是将不同组别的参数置于同一量尺之上, 以此确保不同群体之间的项目参数具有可比性(Magis et al., 2010)。然而, 本研究未进行题目提纯, 主要基于以下三方面理由: 首先, 从测量尺度的一致性来看, 不同于基于连续潜在特质的 IRT 模型, 认知诊断框架下各组的模型参数天然处于同一量尺, 因此对题目提纯的依赖性较低。具体而言, 本研究采用的 G-DINA 模型衡量的是被试在离散属性上的掌握情况, 而非连续能力水平, 因此无需额外调整参数尺度(Ma et al., 2021)。其次, 从提纯程序的适用性来看, 尽管题目提纯有助于提高 DIF 检测的准确性, 但其在实践中面临多重挑战: 1)提纯过程涉及多轮参数估计与 DIF 统计量计算, 并需要动态调整锚集合(anchor set), 这导致计算复杂度增加且流程繁琐(Meade & Wright, 2012); 2)题目提纯无法完全确保锚集合中无 DIF 题目(Yuan et al., 2021); 3)当测验包含多个 DIF 题目时, 提纯过程可能受到“掩盖效应”(masking effect)和“泛化效应” (swamping effect)的干扰, 从而削弱锚集合的构建准确性(Barnett & Lewis, 1994; Yuan et al., 2021)。最后, 从国际上已发表的文献来看, 绝大多数研究在使用 Wald 检验进行认知诊断 DIF 检测时未采用题目提纯程序(如 Hou, 2013; Hou et al., 2014; Li & Wang, 2015; Ma et al., 2021; Mehrazmay et al., 2021)。为了保持与前人研究的一致性, 本研究同样不进行题目提纯。

在计算得到所有候选协变量及其潜在分割点对应的 Wald 统计量后, ISRPM 将根据统计量的数值大小对协变量进行降序排列, 并将其中最大值所对应的协变量视为首个最佳分割协变量。为了便于后续步骤的说明, 本步骤假设首个最佳分割协变量为所有协变量中的第一个协变量, 记作 x_1。根据从协变量 x_1 中搜索出的最优分割点, 可将被试样本划分为两个子集: $\{x_1 = 1\}$ 和 $\{x_1 = 2\}$, 它们分别对应递归分割树首层的左侧和右侧子节点。需要指出的是, 一旦确定首个最佳分割协变量, 即可认为项目 j 在该协变量上的 DIF 检测过程已完成。因此, 在后续的协变量搜索与分割过程中, 将从候选协变量集合中移除变量 x_1, 从而避免重复分割或变量冗余。

步骤 3: 在剩余协变量中递归搜索最佳分割协变量及其最优分割点

在确定首个最佳分割协变量后, ISRPM 并不会终止对于题目 j 的 DIF 检测, 而是基于已生成的两个子节点(即 $\{x_1 = 1\}$ 和 $\{x_1 = 2\}$)继续对作答数据执行递归分割, 以逐步拓展递归分割树的层级结构。在此阶段, ISRPM 将在剩余协变量中持续搜索后续层级的最佳分割协变量及其最优分割点。具体而言, 该递归搜索过程包括以下几个部分:

首先, 基于前一步得到的最佳分割协变量的最优分割点, ISRPM 会将被试样本划分为由该分割点定义的两个特征水平所对应的两个部分。随后, 这两个水平将分别与其余协变量的潜在分割点(或水平)进行完全交叉, 从而生成新的特征水平组合。例如, 当协变量 x_1 已被确定为首个最佳分割协变量时, 其两个水平将与某一剩余协变量(如 x_2)的两个特征水平进行交叉, 从而形成 4 种新的水平组合。

其次, ISRPM 将根据这些新生成的特征组合重新定义目标组与参照组。例如, 在步骤 2 建立的右侧子节点 $\{x_1 = 2\}$ 的基础上, 将其与协变量 x_2 的两个特征水平组合后, 可得到两个新的特征组合, 分别为 $\{x_1 = 2\} \cap \{x_2 = 1\}$ 和 $\{x_1 = 2\} \cap \{x_2 = 2\}$。其中, 前者可视为新的目标组, 后者则作为新的参照组, 两者共同用于后续的模型拟合与 DIF 检验统计量计算。

接着, ISRPM 通过比较各候选分割方案对应的

Wald 统计量大小, 选出统计量最大的协变量及其对应分割点作为当前层级的最佳分割方案。例如, 如果计算结果显示步骤 2 中建立的右侧子节点(即 $\{x_1=2\}$)与协变量 x_2 的所有水平交叉组合后产生的 Wald 统计量数值最大, 则 ISRPM 将在该节点下进一步生成两个新的子节点, 分别记为 $\{x_1=2\}\cap\{x_2=1\}$ 和 $\{x_1=2\}\cap\{x_2=2\}$ 。随着分割树结构的进一步扩展, ISRPM 会依据新的分割层级对公式(4)进行更新, 从而得到下方的公式(6):

$$
\begin{cases}
P(\boldsymbol{\alpha}_{vj}^{*})=\text{node}_{\delta_{j0}}+\sum\limits_{k=1}^{K_j^{*}}\text{node}_{\delta_{jk}}\alpha_{vk}+ \\
\quad\sum\limits_{k'=k+1}^{K_j^{*}}\sum\limits_{k=1}^{K_j^{*}-1}\text{node}_{\delta_{jkk'}}\alpha_{vk}\alpha_{vk'}+\cdots+\text{node}_{\delta_{j12\cdots K_j^{*}}}\prod\limits_{k=1}^{K_j^{*}}\alpha_{vk} \\
\text{node}_{\delta_{j0}}=[\delta_{lj0}\cdot I(x_1=1)+\delta_{rj0}^{[n]}\cdot I(x_1=2)\cdot \\
\qquad I(x_2=1)+\delta_{rj0}^{[n]}\cdot I(x_1=2)\cdot I(x_2=2)] \\
\text{node}_{\delta_{jk}}=[\delta_{ljk}\cdot I(x_1=1)+\delta_{rjk}^{[n]}\cdot I(x_1=2)\cdot \\
\qquad I(x_2=1)+\delta_{rjk}^{[n]}\cdot I(x_1=2)\cdot I(x_2=2)] \\
\text{node}_{\delta_{jkk'}}=[\delta_{ljkk'}\cdot I(x_1=1)+\delta_{rjkk'}^{[n]}\cdot I(x_1=2)\cdot \\
\qquad I(x_2=1)+\delta_{rjkk'}^{[n]}\cdot I(x_1=2)\cdot I(x_2=2)] \\
\vdots \\
\text{node}_{\delta_{j12\cdots K_j^{*}}}=[\delta_{lj12\cdots K_j^{*}}\cdot I(x_1=1)+\delta_{rj12\cdots K_j^{*}}^{[n]}\cdot I(x_1=2)\cdot \\
\qquad I(x_2=1)+\delta_{rj12\cdots K_j^{*}}^{[n]}\cdot I(x_1=2)\cdot I(x_2=2)]
\end{cases}
\tag{6}
$$

其中, $[n]$ 表示新建立的子节点; $\delta_{rj0}^{[n]}$ 表示新建立的子节点所对应的截距参数, 其余以此类推。

步骤 4: 重复步骤 2 至步骤 3 直到满足分割终止规则

在通过步骤 2 与步骤 3 反复搜索最佳分割协变量及其分割点的过程中, ISRPM 会对作答数据进行多轮分割。随着分割层级的加深, 分配至各后续子节点的样本量将逐渐减少。为了确保每个子节点拥有足够的样本量, 从而尽可能保证参数估计的准确性和 DIF 检测结果的有效性, 本研究为 ISRPM 设定了以下两类分割终止规则: 1)协变量搜索空间中已不存在可用于递归分割的协变量; 2)任意子节点的样本量低于预设的最小阈值(本研究设定为 100)。当满足上述任意一项条件时, ISRPM 会立即终止当前题目的协变量搜索与递归分割过程。

步骤 5: 终止 DIF 检测并输出能够显示当前题目 DIF 形式的递归分割树图

当最佳分割协变量及其分割点搜索满足终止条件时, ISRPM 将立即停止当前题目的 DIF 检测, 并根据最终检测结果决定是否输出递归分割树图。具体而言, 对于单个题目, 其 DIF 检测结果仅有两种可能: 1)如果在变量搜索过程中未出现任何具有实质意义的协变量分割(即 Wald 统计量所对应的 p 值高于预设显著性水平), 则该题目被判定为不存在 DIF, 从而无需生成递归分割树; 2)如果至少一个协变量的分割具有实质意义(即 Wald 统计量所对应的 p 值并未超过预设显著性水平), 则该题目被判定为存在 DIF, ISRPM 将输出对应的递归分割树结构图, 该图以可视化的形式展示该题目在哪些协变量上以及以何种方式表现出 DIF。

图 1 展示了不同题目在两个二分类协变量条件下的递归分割树示意图。为了便于理解图中信息, 需要说明以下几点: a)每个子节点显示的是不同分组下 G-DINA 模型项目参数的估计值; b)箭头类型反映 Wald 统计量在统计上是否达到预设的显著性水平, 其中实线箭头表示显著, 虚线箭头表示不显著; c)箭头类型同时用于标识当前题目在相应协变量上是否存在 DIF, 实线箭头表示存在 DIF, 虚线箭头表示不存在 DIF。总体而言, 图 1 中的 4 个题目一共呈现出三种不同的 DIF 形式。题目 1 仅在变量 x_1 表现出主效应 DIF, 具体表现为实线箭头仅出现在递归分割树第一层的所有子节点上。题目 2 的 DIF 表现形式较为特殊: 首先, 若子节点 $\{x_1=1\}\cap\{x_2=1\}$ 和 $\{x_1=1\}\cap\{x_2=2\}$ 之间的差异与 $\{x_1=2\}\cap\{x_2=1\}$ 和 $\{x_1=2\}\cap\{x_2=2\}$ 之间的差异大致相当, 则说明该题目在两个协变量上只表现主效应 DIF, 且不存在交互式 DIF; 然而, 当这两个差异显著不相等时, 则说明题目 2 在两个协变量上同时存在主效应 DIF 和交互式 DIF。题目 3 仅存在交互式 DIF, 这主要体现为: 虚线箭头仅出现在递归分割树结构的第一层子节点, 而实线箭头出现在第二层所有子节点, 这表明只有对两个协变量的所有水平后进行交叉组合后才能识别出 DIF。在题目 4 中, 实线箭头出现在递归分割树第一层的所有子节点及第二层左半部分子节点, 而第二层右半部分为虚线箭头, 表明该题目同时存在主效应 DIF 和交互式 DIF。

2.4　一类错误控制

在基于多个协变量进行 DIF 检测时, ISRPM 不可避免地涉及多重检验(multiple testing)问题。在统计

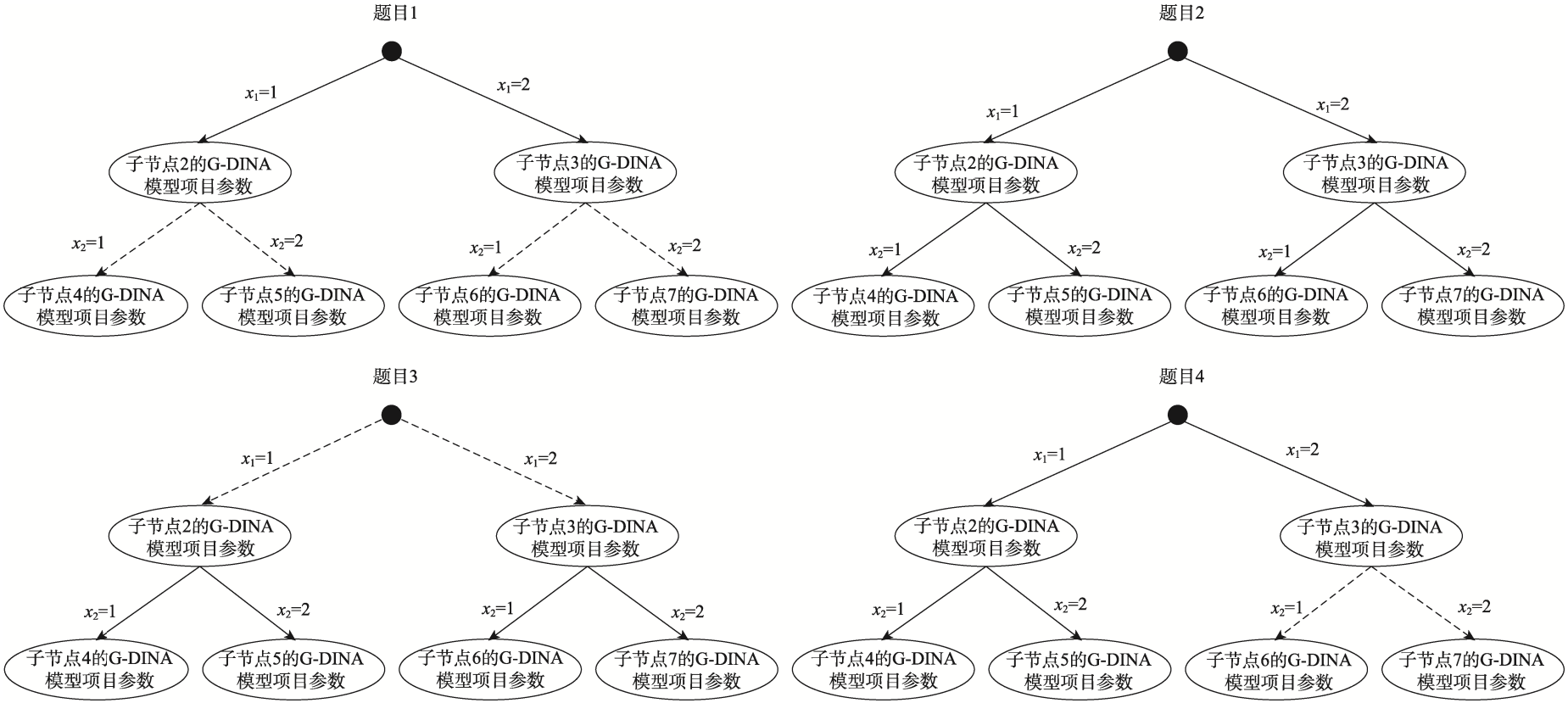

图 1　ISRPM 中不同题目在两个协变量情境下的递归分割树示意图

学文献中, 一个重要的概念是家族误差率(familywise error rate, FWER)。根据 Benjamini 和 Hochberg (1995)的定义, 家族误差率指的是在同时检验一组零假设时, 至少出现一次错误拒绝的概率。在 ISRPM 的方法学框架下, 这意味着: 对于同一题目, 在多个协变量上同时进行假设检验时, 只要存在一次错误拒绝, 即可视为该题目被错误地判定为存在 DIF, 即发生了一类错误。为了在给定全局显著性水平下有效控制 ISRPM 在 DIF 检测过程中的 FWER, 本研究遵循 Tutz 和 Berger (2016)的建议, 对每个单独检验采用更严格的显著性水平, 即通过整体显著性水平调整局部显著性水平以维持整体一类错误的控制。

具体而言, 本研究采用 Bonferroni 调整法实现一类错误控制。该方法通过将整体显著性水平 α (本研究设定为 0.05)除以同时执行的假设检验数量, 从而得到每个协变量对应的局部显著性水平。根据 Strobl 等人(2015)、Tutz 和 Berger (2016)以及 Bollmann 等人(2018)的建议, 对于单个题目而言, 在同时检测其在多个协变量上可能存在的 DIF 时, 应当按照协变量数量对整体显著性水平进行比例调整, 以抵消多重检验导致的假阳性概率上升情况, 从而确保无 DIF 题目被误判为有 DIF 的概率不超过 α。基于 Bonferroni 调整法, ISRPM 的局部显著性水平可依据以下公式计算:

$$\text{adjusted } \alpha = \alpha / m, \tag{7}$$

其中, α 指的是总体显著性水平, 本研究中设定为 α = 0.05, m 指的是被执行 DIF 检测的协变量数量。

3　Monte Carlo 模拟研究

3.1　实验设计

本模拟研究的主要目的是在多种实验条件下评估所提出的 ISPRM 方法在常用 DIF 检测方法性能评价指标上的表现, 并将其与国际上几种广泛使用的认知诊断 DIF 检测方法进行比较。参考 Ma 等人(2021)的研究设计, 本模拟研究一共设置了以下 8 个操纵变量: (1)每个组别的样本容量(含 500、1000 以及 2000 人三个水平); (2) DIF 大小(含 0.05 和 0.1 两个水平, 分别表示小 DIF 和大 DIF); (3) DIF 形式(含只有主效应 DIF、只有交互式 DIF 以及同时存在主效应 DIF 和交互式 DIF 三个水平); (4)题目质量[含高质量和中等质量两个水平, 其中高质量题目的参数 $P(\boldsymbol{\alpha}_v = \mathbf{0})$和 $1 - P(\boldsymbol{\alpha}_v = \mathbf{1})$从均匀分布 U (0.05, 0.15)中随机抽取, 而中等质量题目的对应参数则从均匀分布 U (0.15, 0.25)中随机抽取]; (5) DIF 检测方法(含 ISRPM、Wald、LR、Wald-FS、LR-FS 和 MH 法六个水平); (6)属性掌握模式分布(含均匀分布和多元正态分布两个水平); (7)属性间的相关程度(含 0 和 0.3 两个水平, 分别表示 0 相关和中等程度相关); (8)人口学协变量是否会对属性掌握模式分布产生影响(含有影响和无影响两个水平)。需要说明的是, 当属性掌握模式服从均匀分布时, 无法直接操纵属性间的相关程度以及人口学协变量对属性掌握模式分布的影响。因此, 上述两个因素的操控仅在属性掌握模式服从多元正态分布的条件下实施。具体而言, 当人口学协变量对目标组和参照组的属性掌握模式分布无影响时, 两组的属性掌握模式均从多元正态分布 MVN (0, Σ)中抽取; 而在有影响的条件下, 目标组的属性掌握模式从多元正态分布 MVN(0, Σ)中抽取, 参照组的属性掌握模式则利用多元正态分布 MVN(0.5, Σ)中抽取。

将上述 8 个操纵因素进行完全交叉组合后, 本研究最初共得到 3×2×3×2×6×2×2×2=1728 种实验条件。由于属性掌握模式分布为均匀分布这一条件无法与其中的两个操纵因素(即属性间的相关程度和人口学协变量是否会对属性掌握模式分布产生影响)有效结合, 因此剔除了与之相关的 648 种无效组合, 最终保留了 1080 种有效模拟条件。在实验过程中, 每种条件均重复模拟 100 次。所有模拟实验均在 R 环境(R Core Team, 2021)中实现。此外, 本研究还设置了以下固定条件: (1)认知属性个数, 参照 Ma 等人(2021)的研究将其固定为 5 个; (2)测验长度, 本研究固定为 30 题, 该长度在认知诊断 DIF 检测研究中较为常见(刘彦楼 等, 2016; 孙小坚 等, 2022; Hou et al., 2014; Li & Wang, 2015); (3)单个题目测量的属性数量, 参照 Ma 等人(2021)与刘彦楼等人(2016)的研究, 本研究限制为每题最多同时测量 3 个属性; (4)DIF 题目比例, 参照 Ma 等人(2021)的研究, 将 DIF 题目占总测验长度的比例固定为 20%, 即 30 题中包含 6 个 DIF 题目; (5)测验 Q 矩阵, 采用以往研究中广泛使用的 30 题 Q 矩阵(刘彦楼等, 2016; Hou et al., 2014; Ma et al., 2021), 该矩阵包含 5 个认知属性且采用了平衡设计, 其中每个属性被题目所测量的次数相等, 并且测量 1、2、3 个属性的题目数量也相等; (6)DIF 题目位置, 参照 Ma 等人(2021)的做法, 在所有的实验条件中, 将第 6、9、12、14、24 和 25 题设定为存在 DIF 的题目; (7)感兴趣的协变量数量, 参照 Strobl 等人(2015)以及 Tutz

和 Berger (2016)的做法, 本研究设定为 2 个协变量, 分别记为 x_1 和 x_2, 且每个协变量均包含两个水平。

3.2 不同 DIF 形式的作答数据模拟及 DIF 检测

本研究在 R 环境(R Core Team, 2021)下开展不同 DIF 形式的作答数据模拟与检测, 全部代码可通过在线 OSF 平台获取(链接: https://osf.io/7ykqj/?view_only=e873ab9c2d2c408385a90e945952f296)。具体操作步骤如下: (1)使用 GDINA 包(Ma & de la Torre, 2020)中的 *simGDINA* 函数生成参照组的被试参数和项目参数, 并基于此模拟参照组的作答数据。(2)进行 DIF 操纵并生成目标组被试作答数据。首先, 依据参照组被试在每个题目上的正确作答概率 $P(\alpha_v)_R$, 结合不同 DIF 大小计算目标组被试在相同题目上的正确作答概率 $P(\alpha_v)_F$。在本研究中, 不同的 DIF 形式中 $P(\alpha_v)_F$ 的计算方法有所不同, 详见表 1。随后, 利用得到的 $P(\alpha_v)_F$ 模拟目标组被试的作答数据。(3)在生成两组被试的作答数据后, 分别采用本研究考虑的 6 种 DIF 检测方法(ISRPM、Wald、LR、Wald-FS、LR-FS 与 MH)进行 DIF 检测, 并输出相应检测结果。需要说明的是, 本研究的所有检测方法均未实施题目提纯程序。

表 1 不同 DIF 形式下的 DIF 模拟方法

DIF 形式	$P(\alpha_v)_F$ 的计算方法
只有主效应 DIF	$P(\alpha_v)_R + z \cdot I(x_2 = 1)$
只有交互式 DIF	$P(\alpha_v)_R + \begin{cases} z \cdot I(\{x_1 = 1\} \cap \{x_2 = 1\}) \\ z \cdot I(\{x_1 = 2\} \cap \{x_2 = 2\}) \end{cases}$
同时存在主效应 DIF 和交互式 DIF	$P(\alpha_v)_R + z \cdot I(x_1 = 2) + z \cdot I(\{x_1 = 1\} \cap \{x_2 = 2\})$

注: z 表示 DIF 大小, 本研究中取 0.05 和 0.1, 分别代表小 DIF 和大 DIF。

3.3 评价指标

以往的认知诊断 DIF 研究普遍采用真阳性率(true positive rate, TPR)和假阳性率(false positive rate, FPR)评估 DIF 检测方法的性能。TPR 相当于统计检验力, 即正确识别出 DIF 题目的比例; FPR 相当于一类错误率, 即错误地将不存在 DIF 的题目识别为存在 DIF 的比例。然而, 仅在题目水平上计算这两个指标难以全面反映所提出方法的 DIF 检测性能, 也难以与传统方法进行合理比较。这是因为以往研究通常仅考虑由单个协变量引发的 DIF, 此时的 TPR 与 FPR 仅能评估检测方法在单一协变量上的检测性能表现。与之不同的是, 本研究进一步考虑了多个协变量同时导致 DIF 出现的情况。在这种情境下, 评价检测方法的性能不仅应关注其能否正确识别出存在 DIF 的题目, 还应考察其在识别出 DIF 题目后, 能否进一步准确定位出导致 DIF 的具体协变量。因此, 除了在题目水平上计算 TPR 与 FPR 之外, 本研究还在题目与协变量结合的水平上计算这两个指标, 从而更全面地评估 DIF 检测方法的整体性能。

令 $\boldsymbol{\xi}_j^T = (\xi_{j1}, \cdots, \xi_{jm})$ 表示项目 j 在 m 个协变量上的 DIF 检测结果向量, 其中 $\xi_{jm} = 1$ 表示题目 j 在协变量 m 上有 DIF, 而 $\xi_{jm} = 0$ 则表示没有 DIF。如果向量 $\boldsymbol{\xi}_j^T$ 中的任何一个元素为 1, 则说明题目 j 是一个有 DIF 的题目; 如果该向量中的所有元素都为 0, 即 $\boldsymbol{\xi}_j^T = (0, \cdots, 0)$, 那么说明题目 j 是一个不存在 DIF 的题目。基于此, 各评价指标的计算方式如下:

(1)题目水平的 TPR, 记为 TPR_I,

$$TPR_I = \frac{1}{\#\{j : \boldsymbol{\xi}_j \neq 0\}} \sum_{j:\boldsymbol{\xi}_j \neq 0} I(\hat{\boldsymbol{\xi}}_j \neq 0) \tag{8}$$

(2)题目水平的 FPR, 记为 FPR_I,

$$FPR_I = \frac{1}{\#\{j : \boldsymbol{\xi}_j = 0\}} \sum_{j:\boldsymbol{\xi}_j = 0} I(\hat{\boldsymbol{\xi}}_j \neq 0) \tag{9}$$

(3)题目与协变量结合水平的 TPR, 记为 TPR_IC,

$$TPR_{IC} = \frac{1}{\#\{j, m : \xi_{jm} \neq 0\}} \sum_{j,m:\xi_{jm} \neq 0} I(\hat{\xi}_{jm} \neq 0) \tag{10}$$

(4)题目与协变量结合水平的 FPR, 记为 FPR_IC,

$$FPR_{IC} = \frac{1}{\#\{j, m : \xi_{jm} = 0\}} \sum_{j,m:\xi_{jm} = 0} I(\hat{\xi}_{jm} \neq 0) \tag{11}$$

上述公式中, $I(\cdot)$ 表示指示函数, 如果满足条件 d, 那么 $I(d) = 1$ 否则 $I(d) = 0$。

3.4 实验结果

受篇幅限制, 正文部分仅报告了在属性掌握模式服从均匀分布的条件下, 各认知诊断 DIF 检测方法在不同实验条件下的统计检验力与一类错误率的总体表现, 相关结果见图 2 和图 3; 对应的完整数值结果列于网络版附录中的附表 1 和附表 2。进一步地, 当属性掌握模式分布服从多元正态分布时, 各检测方法在不同实验条件下的统计检验力与一类错误率结果则统一呈现于在线附录(https://osf.io/7ykqj/?view_only=e873ab9c2d2c408385a90e945952f296)中的附表 S1 至 S8。总体来看, ISRPM 在多数实验条件下均表现出可接受的检测性能, 并且在整体上的表现优于已有的 Wald、LR、FS-Wald、FS-LR 和 MH 方法。接下来的部分将进一步分析不同操纵变量对统计检验力与一类错误率的影响。

图 2　各认知诊断 DIF 检测方法在不同实验条件下的统计检验力结果

图 3　各认知诊断 DIF 检测方法在不同实验条件下的一类错误率结果

3.4.1　统计检验力

图 2 展示了不同实验条件下, 各认知诊断 DIF 检测方法在题目水平以及题目与协变量结合水平上的统计检验力结果。然而, 当一个题目的 DIF 由多个协变量的主效应和/或交互作用共同引起时, 仅依据题目水平的 TPR_I 指标可能无法全面反映方法的检测能力。具体而言, TPR_I 指标仅衡量检测方法能否正确识别出存在 DIF 的题目, 而未进一步考虑这些 DIF 由哪些协变量导致。换言之, 即使某种方法仅识别出部分 DIF 来源, TPR_I 仍会将其视为正确识别, 从而可能在一定程度上高估该方法的检测性能。相较之下, TPR_{IC} 指标不仅考察 DIF 题目是否被正确识别, 还进一步评估检测方法是否能够准确定位引发 DIF 的协变量。因此, TPR_{IC} 提供了更细致的评估视角, 更能反映检测方法在多个协变量 DIF 情境下的检测性能。

从图 2 的结果可以看出, 当测验中存在交互式 DIF 时, ISRPM 的统计检验力在大多数条件下的表现优于已有的 Wald、LR、FS-Wald、FS-LR 和 MH 方法。随着每组样本量的增加, 所有方法的统计检验力均明显提升, 这说明样本量是影响 DIF 检测性能的一个重要因素。当每组样本量为 2000 时, ISRPM 在统计检验力上的整体表现优于其他 5 种传统方法, 尤其是在仅存在交互式 DIF 的条件下, 其优势更加明显。然而, 在每组样本量为 500 的条件下, ISRPM 的统计检验力受到一定限制。具体而言, 在此样本量条件下, 当测验中仅存在主效应 DIF 或同时存在主效应 DIF 与交互式 DIF 时, 尽管所有方法的统计检验力均偏低, 但 ISRPM 的整体表现不如其余 5 种方法。这一结果可能与 ISRPM 的递归分割过程中子节点被分配到的样本量过小有关, 这使得参数估计精度下降, 进而影响其 DIF 检测的准确性。例如, 在每组 500 人且包含两个二分类协变量的条件下, ISRPM 在第一层分割中分配至每个子节点的样本量为 500; 若进入第二层分割, 4 个子节点被分配至的样本量将降至 250。在此情况下, 样本量不足可能导致参数估计误差增大, 进而影响检测结果的准确性。值得注意的是, 在相同样本量条件下, 当 DIF 仅由协变量间的交互作用引起时, ISRPM 的检测结果相对更优。这可能与该方法能够在分析过程中考虑多个协变量的交互作用有关, 从而在识别交互式 DIF 时展现出更高的敏感性。

随着 DIF 效应的增大, 所有方法的统计检验力均有明显提升。例如, 如附表 1 所示, 当测验中仅存在主效应 DIF 且每组样本量为 1000 时, ISRPM 在小 DIF 条件下的 TPR_I 值介于 0.53 至 0.59 之间, 而在大 DIF 条件下, 这一指标上升至 0.98。同样, Wald、LR、FS-Wald、FS-LR 和 MH 方法在小 DIF 条件下的 TPR_I 范围分别为 0.62~0.69、0.67~0.76、0.64~0.73、0.67~0.75 和 0.48~0.49, 而大 DIF 条件下的分布范围则分别为 0.99~0.99、1~1、0.99~0.99、1~1 和 0.96~0.96。从 DIF 形式来看, ISRPM 在交互式 DIF 的检测上表现出相对更高的统计检验力。例如, 附表 1 显示, 当测验中仅存在交互式 DIF 且每组样本量为 100 时, ISRPM 法的 TPR_I 范围为 0.22~ 0.81, 而 Wald、LR、FS-Wald、FS-LR 和 MH 方法的 TPR_I 范围分别为 0.05~0.08、0.08~0.10、0.06~ 0.08、0.08~0.11 以及 0.07~0.09。在仅存在主效应 DIF 的情况下, ISRPM 在较大样本量条件下的表现与其他方法接近。然而, 当每组样本量降至 500 时, ISRPM 在统计检验力方面的整体表现不如其他 5 种传统方法。整体而言, 在每组样本量为 1000 及以上的条件下, ISRPM 在主效应 DIF 的检测上的整体表现与传统方法相当, 同时在涉及交互式 DIF 的情境中展现出更高的统计检验力。此外, 随着题目质量的提高, 各方法的统计检验力均呈现出上升趋势。最后, 从在线附录部分的表 S1、S3、S5 和表 S7 的结果来看, 属性间的相关程度以及人口学协变量是否会对属性掌握模式分布产生影响这两个变量均未对 ISRPM 的统计检验力产生明显的影响。

3.4.2　一类错误率

Ma 等人(2021)指出, 在名义显著性水平为 α 且模拟研究重复次数为 n 的前提下, 由于抽样误差的存在, 实际观察到的一类错误率并不一定与名义水平完全一致, 但有 95%的概率落在区间 $\alpha \pm 1.96 \times \sqrt{\alpha(1-\alpha)/n}$ 之内。对应到本研究(α=0.05, n=100), 该区间为[0.007, 0.093]。若某一检测方法的观测一类错误率位于此范围内, 则可认为其控制效果合理。如图 3 所示, ISRPM 在绝大多数实验条件下对一类错误率的控制较为稳定, 整体结果均位于合理区间范围内。进一步比较发现, 在绝大多数实验条件下, ISRPM 的一类错误率低于其余 5 种传统方法, 这种结果表明在同时涉及多个协变量进行 DIF 检测的情境中, 新方法对于一类错误的控制效果总体上优于传统方法。例如, 附表 2 显示, 当测验中仅存在主效应 DIF 时, ISRPM 的 FPR_I 值范围为 0.01~0.07, 这说明 ISRPM 对于一类错误的控制效果符合

预期; 而 Wald、LR、FS-Wald、FS-LR 和 MH 法的 FPR_I 分布范围分别为 0.04~0.11、0.11~0.17、0.06~0.10、0.11~0.14 和 0.09~0.30, 其中部分条件下的结果高于合理区间上限(即 0.093)。整体来看, ISRPM 在一类错误率控制方面表现出较高的稳定性, 不同的 DIF 形式与 DIF 大小对各检测方法的一类错误率均未表现出明显影响。此外, 各方法在高质量题目条件下的一类错误率整体上低于中等质量题目条件, 这可能与高质量题目下参数估计的精度较高有关。最后, 在线附录中表 S2、S4、S6 和表 S8 呈现的研究结果表明, 属性间的相关程度以及人口学协变量是否影响属性掌握模式分布这两个因素均不会对各方法的一类错误率产生明显的影响。

4 实证研究

4.1 数据介绍

本研究的实证分析以 Xi 等人(2020)编制的分裂型人格问卷认知诊断版本(A Diagnostic Classification Version of the Schizotypal Personality Questionnaire; DC-SPQ)为例, 用以展示所提出的 ISRPM 在真实测验数据中的可用性, 并将其与国际上已有的 5 种认知诊断 DIF 检测方法进行结果对比。DC-SPQ 包含 74 个题目, 每个题目均采用二级计分法。本研究采用了 Xi 等人(2020)所提供的 DC-SPQ 的测验 Q 矩阵(见原始文献中的表 A1), 该矩阵共包含 9 个属性。Xi 等人(2020)提供的测试数据来源于来自中国 3 个城市 7 所大学的 980 名大学生, 被试平均年龄为 20.5 岁(SD = 1.79)。在 DIF 检测中, 本研究选取常用的性别与户籍作为协变量。其中, 性别分布方面, 女性被试占比为 62.3% (N = 611); 户籍分布方面, 43.1% (N = 422)的被试来自城市地区。

4.2 DIF 检测结果

Xi 等人(2020)在对 DC-SPQ 进行 DIF 分析时采用了 Wald 检验法, 并且将整体显著性水平设定为 0.01。为了更全面地比较不同方法的检测表现, 本研究分别在显著性水平为 0.05 和 0.01 的条件下, 使用所提出的 ISRPM 及 5 种现有的认知诊断 DIF 检测方法(Wald、MH、LR、FS-Wald 和 FS-LR)对 DC-SPQ 进行了分析。需要强调的是, 在实施 ISRPM 时, 所用到的最佳分割协变量及其分割点的搜索终止规则与模拟研究保持一致。

表 2 展示了两种整体显著性水平(0.05 与 0.01)下不同方法对 DC-SPQ 的 DIF 检测结果。总体来看, ISRPM 与其他 5 种传统方法的检测结果一致性较高。例如, 在总体显著性水平为 0.05 的条件下, 第 24、25、26、27 和第 40 题均被所有方法识别为可能存在由性别引起的主效应 DIF。然而, 第 35 题的检测结果在不同方法之间表现出一定差别: ISRPM 将其识别为可能存在由性别与户籍的交互作用引起的交互式 DIF; LR 与 MH 方法将其识别为可能仅

表 2 整体显著性水平为 0.05(0.01)时, DC-SPQ 各题目在不同 DIF 检测方法下的检测结果

题目	ISRPM				Wald		MH		LR		FS-Wald		FS-LR	
	首层变量	1^{st}	2^{nd}-L	2^{nd}-R	性别	地区	性别	地区	性别	地区	性别	地区	性别	地区
1	性别	×(×)	×(×)	×(×)	×(×)	×(×)	√(√)	×(×)	√(×)	×(×)	√(×)	×(×)	√(√)	×(×)
2	性别	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)
3	性别	√(√)	×(×)	×(×)	√(√)	√(×)	√(×)	×(×)	×(×)	√(√)	×(×)	√(×)	√(×)	√(×)
5	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	×(×)	×(×)	√(×)
7	性别	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	√(√)	×(×)
9	户籍	√(√)	×(×)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(×)	×(×)	√(√)
12	性别	√(√)	×(×)	×(×)	√(√)	×(×)	×(×)	√(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)
14	性别	√(×)	×(×)	×(×)	√(×)	√(×)	√(√)	√(√)	√(√)	√(√)	×(×)	√(√)	√(×)	√(√)
15	户籍	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(×)	×(×)	√(×)
16	户籍	√(×)	×(×)	×(×)	×(×)	√(×)	√(√)	√(×)	√(√)	√(√)	√(×)	√(√)	√(√)	√(√)
18	性别	√(×)	×(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)
20	性别	√(√)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)
22	性别	√(√)	×(×)	×(×)	√(√)	×(×)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)
24	性别	√(×)	×(×)	×(×)	√(×)	×(×)	√(×)	√(×)	√(√)	×(×)	√(×)	×(×)	√(√)	×(×)
25	性别	√(√)	×(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)
26	性别	√(√)	×(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)

续表

题目	ISRPM				Wald		MH		LR		FS-Wald		FS-LR	
	首层变量	1st	2nd-L	2nd-R	性别	地区	性别	地区	性别	地区	性别	地区	性别	地区
27	性别	√(√)	×(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)
28	性别	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)
29	性别	×(×)	×(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(×)	×(×)	×(×)	×(×)	×(×)	×(×)
30	性别	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)
31	户籍	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	√(√)	×(×)	×(×)	×(×)	√(×)
34	性别	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)
35	性别	×(×)	×(×)	√(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	×(×)	×(×)	×(×)	×(×)
38	性别	√(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	×(×)	×(×)
40	性别	√(√)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	√(×)	√(√)	×(×)	√(√)	√(×)
43	性别	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)
44	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)
46	性别	√(√)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)
48	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	×(×)
49	性别	√(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(×)	×(×)	√(√)	×(×)
50	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	√(×)
52	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	×(×)	×(×)	√(×)
54	性别	√(√)	×(×)	×(×)	√(√)	×(×)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	√(×)
55	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	×(×)
59	户籍	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(×)	×(×)	×(×)	×(×)	√(×)
61	户籍	√(√)	×(×)	×(×)	×(×)	√(√)	×(×)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)
62	性别	√(×)	×(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(×)	×(×)	√(√)	×(×)
64	性别	×(×)	×(×)	×(×)	×(×)	×(×)	√(√)	×(×)	√(×)	×(×)	√(×)	×(×)	√(×)	×(×)
65	户籍	√(×)	×(×)	×(×)	×(×)	√(×)	×(×)	√(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)
69	性别	√(×)	×(×)	×(×)	√(×)	×(×)	×(×)	√(√)	√(√)	×(×)	√(×)	×(×)	√(√)	×(×)
71	性别	√(√)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	√(×)	×(×)	×(×)	√(×)	×(×)
72	性别	×(×)	×(×)	×(×)	×(×)	×(×)	×(×)	√(√)	√(×)	×(×)	√(√)	×(×)	√(×)	×(×)
74	性别	√(√)	×(×)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)	√(√)	×(×)

注: 此表仅呈现至少有一种方法侦查有 DIF 的试题结果。其中, “首层变量”指的是 ISRPM 在当前题目中自动搜索出的首个最佳分割变量, 即递归分割树结构中的首层分割变量, 不同的题目会依据每个变量分割效果的不同可能显示出不同的首层最佳分割变量; “√”表示当前方法将该题目标记为存在 DIF, 而“×”表示当前方法将该题目标记为不存在 DIF。括号外的结果表示在总体显著性水平为 0.05 条件下的检测结果; 括号内的结果表示在总体显著性水平为 0.01 条件下的检测结果。

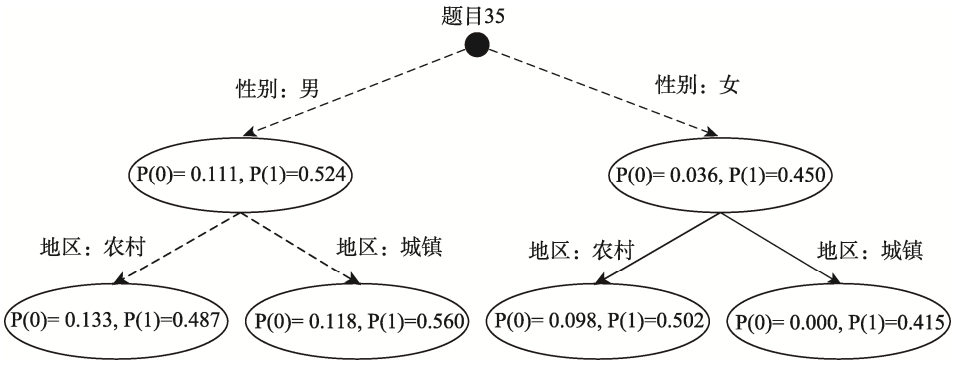

图 4　题目 35 中基于 ISRPM 输出的递归分割树图

表现出由性别引发的主效应 DIF; 而 Wald、FS-Wald 和 FS-LR 方法均未在该题中检测出任何的 DIF 形式。图 4 展示了 ISRPM 对第 35 题输出的递归分割树结构。从图中可以看到: 第一层中的两条针对性

别的路径均为虚线, 这说明表明此题目可能未在性别上表现出主效应 DIF; 在第二层分割中, 左侧分支仍为虚线箭头, 而右侧分支则为实线箭头, 这表明来自不同户籍地区的女性在该题上的作答模式存在明显差异, 而男性在户籍变量上的组间差异不明显。这一结果提示, 第 35 题有可能存在由性别与户籍的交互作用所导致的交互式 DIF。尽管对 DIF 检测结果的事后解释仍存在一定程度的不确定性, 但实证分析结果表明, ISRPM 在识别交互式 DIF 方面表现出一定的潜在优势。需要指出的是, 对 DIF 检测结果的解释不应仅依赖统计检验。一种更为合理的策略是结合统计分析与专家判断：首先利用统计方法筛选出可能存在 DIF 的题目, 随后由相关内容领域专家对这些题目的内容与表述进行严格审查。通过这种定量与定性分析相结合的方式, 测验开发者能够更全面地评估检测结果的实质意义, 从而为测验修订与质量控制提供更有依据的参考。

5 讨论与结论

5.1 讨论

正如《教育与心理测验标准》(AERA et al., 2014)中所强调的那样, 为了确保测验结果的公平性和准确性, 对测验进行项目功能差异(DIF)分析已成为心理与教育测验开发过程中评估测验质量与效度的重要环节。因此, 在认知诊断测验开发与验证过程中, DIF 检测仍然是不可或缺的质量评估步骤。然而, 现有的认知诊断 DIF 检测方法通常仅能在单次分析中针对单一协变量开展检测, 从而局限于针对主效应 DIF 的识别。尽管这些方法在识别主效应 DIF 方面表现可靠, 但难以有效检测由多个协变量交互作用引起的交互式 DIF, 这直接会影响认知诊断评估的测量公平性和效度。为了解决这一问题, 本研究将认知诊断模型与数据挖掘中的递归分割技术相结合, 提出了一种新的认知诊断 DIF 检测方法, 即基于项目水平的序列递归分割法(ISRPM)。该方法旨在为认知诊断测验中进行主效应 DIF 和交互式 DIF 检测提供更加有效的方法学支持。总的来说, 本研究的主要贡献在于所提出的 ISRPM 实现了在认知诊断 DIF 检测过程中同时处理多个协变量, 并能够有效地检测出主效应 DIF 和交互式 DIF, 从而有望进一步提升认知诊断 DIF 检测方法的检测精度。

为了评估 ISRPM 的检测性能, 本研究进行了一系列 Monte Carlo 模拟实验, 主要考察以下两个重要方面: (1)每个组别的样本容量、DIF 大小、DIF 形式、题目质量、属性间的相关程度以及人口学协变量是否会对属性掌握模式分布产生影响等因素对新方法检测性能的影响; (2)新方法与国际上已有的 5 种认知诊断 DIF 检测方法(Wald、LR、FS-Wald、FS-LR 和 MH)进行综合性能比较的结果。此外, 为了展示并验证新方法在真实测试环境中的可行性, 本研究以认知诊断版本的分裂型人格问卷为例进行了实证分析, 并将其与国际上已有的 5 种 DIF 检测方法的结果进行了比较。研究结果表明: (1)每个组别的样本量、DIF 大小以及项目质量均是影响 ISRPM 的 DIF 检测性能的重要因素; (2)当题目中存在交互式 DIF 时, 所提出的 ISRPM 在统计检验力和一类错误率中的表现整体上优于传统的 Wald、LR、FS-Wald、FS-LR 以及 MH 方法; (3)在仅存在主效应 DIF 的情境下, 当样本量达到每组 1000 人及以上时, ISRPM 的检测性能与其他 5 种 DIF 检测方法大体相当, 而当每组样本量降至 500 时, ISRPM 的性能表现受到一定限制; (4)属性间的相关程度以及人口学协变量对属性掌握模式分布的影响均未对 ISRPM 的检测性能产生明显影响。

通过模拟研究与实证分析, 可以总结出 ISRPM 具有以下几点优势: (1)它有望进一步提升认知诊断测验的 DIF 检测精度。模拟实验结果表明, 与现有检测方法相比, ISRPM 在识别 DIF 题目以及导致 DIF 的相关协变量方面整体上有着更好的表现。(2)能够同时识别主效应 DIF 和交互式 DIF。新方法这种能够同时处理交互式 DIF 的特性有望帮助研究者更加全面地分析导致测验出现测量偏差的复杂人口学来源。(3)有望为测验开发者提供更为详细的 DIF 诊断信息。与传统方法相比, ISRPM 不仅能够准确识别出存在 DIF 的题目, 还能够揭示不同协变量的主效应及其交互作用对 DIF 的影响, 从而为测验开发者提出更有针对性的题目修改建议。

综合模拟研究与实证研究的结果, 本文对 ISRPM 的潜在使用者提出以下建议: (1)关于认知诊断模型的选择。本研究通过模拟与实证分析验证了 ISRPM 在饱和认知诊断模型(即 G-DINA 模型)下的适用性。需要指出的是, 该方法同样可扩展至 G-DINA 模型的若干简化形式, 如 DINA、DINO 和 ACDM 等常用模型, 因此研究者可根据研究目标和数据特征灵活选用。(2)关于协变量数量和类型。模拟研究结果显示, ISRPM 能够支持同时对两个及以上的协变量进行 DIF 检测。在研究涉及多个协变

量时, ISRPM 具有良好的适用性。但需要注意的是, 随着协变量数量的增加, 为了保持新方法的稳定性与检测精度, 使用者应尽可能地增大样本容量。因此, 在涉及到多个协变量的场景下, 使用者需要平衡样本量与协变量数量之间的关系。为了进一步验证新方法在包含多分类及连续型协变量条件下的表现, 本文在在线附录中展示了两项补充实验(参见在线附录中的补充实验 1 与补充实验 2)。结果表明, ISRPM 在两个以上的协变量条件下仍然能够保持可接受的检测性能, 并且对于不同类型协变量均展示出可接受的检测效果。(3)关于样本量。模拟结果表明, 样本量是影响 ISRPM 检测性能的重要因素。当每组样本量达到 1000 及以上时, ISRPM 在大多数条件下均表现出较为理想的检测精度; 而当样本量降至 500 及以下时, 其统计检验力出现下降趋势, 且整体表现不如传统方法。因此, 在实际应用中, 本文建议应当尽可能保证每组样本量不低于 1000, 从而获得较为稳健的检测结果。(4)关于 DIF 检测方法的选择。在认知诊断测验中选择合适的 DIF 检测方法时, 应综合考虑研究目标、样本量及感兴趣的协变量数量等关键因素。ISRPM 虽然在识别交互式 DIF 方面表现出更好的效果, 但并非在所有 DIF 形式中的性能表现都优于传统方法。对于只关注主效应 DIF 的研究且每组样本量较小(如每组低于 500)时, 更推荐使用传统方法(如 Wald 检验法); 而在每组样本量较大(≥1000)或研究同时关注主效应与交互式 DIF 的情境下, 更推荐使用 ISRPM。需要提醒 的是, 当样本量较小时, 无论采用何种方法, DIF 检测结果的精度都可能受限, 因而不宜仅依据统计检验结果做出结论。(5)关于 DIF 题目的处理。在认知诊断测验中进行 DIF 检测时, 研究者不应仅依赖统计检验结果。一种更为稳健的做法是结合统计分析与专家评估: 首先利用 ISRPM 或其他方法筛选出可能存在 DIF 的题目, 然后邀请领域专家对这些题目的内容及表述进行复核。通过这种定量分析与定性判断相结合的方式, 能够更加准确地判断当前检出的 DIF 是否具有实际意义, 从而为测验修订与质量控制提供更有依据的决策支持。

5.2　认知诊断 DIF 检测方法与 CTT 和 IRT 框架下 DIF 检测方法的区别与联系

在测验公平性研究领域, DIF 检测方法主要依托于三大测量框架: 经典测量理论(CTT)、项目反应理论(IRT)和认知诊断(CD)。这三种框架下的 DIF 检测方法在核心目标、检测逻辑和基本原则上存在一定的共性, 但在理论基础、等值处理方式及能力匹配标准等方面有着明显的区别。

首先, CTT、IRT 与 CD 框架在 DIF 检测的核心目标上保持一致: 三者都旨在识别可能引起测量偏差的题目, 从而维护测验结果的公平性。其次, 无论采用哪种测量框架, DIF 检测均围绕一个共同的逻辑展开：在控制能力水平的前提下, 考察不同被试群体在同一题目上的作答表现是否存在系统性差异。最后, 这些方法均遵循“条件匹配”原则, 即在控制被试能力水平的前提下进行比较, 而非直接依据原始得分判断 DIF。具体而言, CTT 方法采用观察总分进行匹配, IRT 方法通过潜在特质匹配进行匹配, 而 CD 方法则依据被试的属性掌握模式来匹配能力。

尽管在 DIF 检测的核心目标、基本逻辑和基本原则上保持一致, CTT、IRT 和 CD 在理论基础、等值处理方式以及能力匹配标准上仍存在明显区别。首先, 从理论基础来看, CTT 方法基于经典测量理论, IRT 方法依托于项目反应理论, 而 CD 方法则构建在认知诊断理论的基础之上。其次, 从等值处理方式来看, CTT 和 IRT 框架下的 DIF 检测通常需要进行等值处理, 通过将不同组别的能力分数置于同一量尺之上, 以便进行公平的比较。与之不同的是, CD 方法中使用的认知诊断模型(CDM)天然具备参数等值的特性, 因为 CDM 主要衡量被试在离散认知属性上的掌握情况, 而并非将其能力定位在潜在特质的连续尺度上。因此, CD 框架下的 DIF 检测通常不需要进行额外的等值处理。最后, 在能力匹配标准方面, CTT 方法通过观察总分进行匹配, IRT 方法通过潜在特质分数进行匹配, 而 CD 方法则依据被试在认知属性上的掌握模式进行匹配。

5.3　研究局限性与展望

尽管本研究在认知诊断评估中的主效应 DIF 和交互式 DIF 检测方面取得了实质性进展, 但仍存在一些明显的局限性和改进空间, 主要包括以下几个方面:

(1)与现有的 Wald、LR、FS-Wald、FS-LR 以及 MH 方法类似, 本研究中提出的 ISRPM 在小样本情况下的 DIF 检测性能仍有进一步改进的空间。这意味着对 ISRPM 在小样本条件下的性能需要进一步的探讨与优化。

(2)本研究主要关注了 ISRPM 在二级计分的认知诊断模型(即 G-DINA)模型中的应用, 但递归分割技术在多级计分认知诊断模型以及多策略认知

诊断模型中的表现情况仍然是一个未知领域。未来研究可以进一步尝试将 ISRPM 拓展至在多级计分认知诊断模型和多策略模型下的 DIF 检测。

(3)与现有 IRT 框架下基于递归分割技术的 DIF 检测方法(Strobl et al., 2015; Tutz & Berger, 2016)一致, ISRPM 目前仅支持二元分割(binary split)。这可能限制其对单一协变量中多个区间中表现出 DIF 的复杂模式的识别能力。因此, 未来的研究可以尝试将 ISRPM 进一步拓展为能够执行多元分割(multivariate split)的 DIF 检测方法, 以增强其在更复杂测试情境中的应用潜力。

(4)尽管本研究通过模拟实验对 ISRPM 的性能进行了系统评估, 但仍需认识到模拟研究与实证研究之间存在一定差别。模拟研究的实验条件是经过精心设计的且能够严格控制变量, 这使其能够考察检测方法在不同测试条件下的合理性和有效性。然而, 实证数据结构通常更加复杂, 可能受到测量误差、被试特征异质性及潜在混淆因素的影响。因此, 模拟研究的结果应被视为新方法 DIF 检测性能的初步验证, 而非其在实际测验中应用效果的直接证据。未来研究应当基于更多类型、更大样本量的实证数据, 对 ISRPM 的适用性与稳健性进行进一步检验。

5.4 结论

本研究的主要结论如下:

(1)本研究提出了一种在认知诊断框架下用于同时检测主效应 DIF 与交互式 DIF 的新方法, 即基于项目水平的序列递归分割法(ISRPM)。该方法实现了递归分割技术与认知诊断模型的充分融合, 不仅能够在单次检测中同时处理多个协变量, 更能同时识别出表现出 DIF 的题目以及导致该题目出现 DIF 的协变量。

(2)就整体而言, ISRPM 具有可接受的 DIF 检测性能。在主效应 DIF 检测上, 其总体性能与现有的 Wald、LR、FS-Wald、FS-LR 及 MH 方法大体一致; 而在交互式 DIF 检测中, ISRPM 表现出相对更优的识别能力。

(3)综合模拟结果显示, 当每个组别样本量较大、DIF 效应较大且题目质量较高时, ISRPM 能够保持较为理想的检测性能。

参 考 文 献

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Hoboken: Wiley.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507−526.

Belzak, W. C. (2023). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice, 42*(1), 24−33.

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673−690.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289−300.

Bollmann, S., Berger, M., & Tutz, G. (2018). Item-focused trees for the detection of differential item functioning in partial credit models. *Educational and Psychological Measurement*, *78*(5), 781−804.

Collins, P. H. (1990). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. UnwinHyman.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179−199.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281−296.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31a review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics, 26*, 979−1030.

Finch, W. H., Hernández Finch, M. E., & French, B. F. (2015). Recursive partitioning to identify potential causes of differential item functioning in cross-national data. *International Journal of Testing, 16*(1), 21−53.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651−674.

Hou, L. (2013). *Differential item functioning assessment in cognitive diagnostic modeling: Applying the Wald test to investigate DIF in the generalized DINA model framework* (Unpublished doctoral dissertation). University of Delaware.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98−125.

Komboz, B., Strobl, C., & Zeileis, A. (2016). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement, 78*(1), 128−166.

Leighton, J. P., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and Applications*. Cambridge, UK: Cambridge University Press.

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* (Unpublished doctoral dissertation). University of Georgia.

Li, L., Zhou, X., Huang, J., Tu, D., Gao, X., Yang, Z., & Li, M.

(2020). Assessing kindergarteners' mathematics problem solving: The development of a cognitive diagnostic test. *Studies in Educational Evaluation*, *66*, 100879.

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28−54.

Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016). An improved method for differential item functioning detection in cognitive diagnosis models: An application of Wald statistic based on observed information matrix. *Acta Psychologica Sinica*, *48*(5), 588−598.

[刘彦楼, 辛涛, 李令青, 田伟, 刘笑笑. (2016). 改进的认知诊断模型项目功能差异检验方法——基于观察信息矩阵的 Wald 统计量. *心理学报, 48*(5), 588−598.]

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1−26.

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement, 45*(1), 37−53.

Magis, D., Béland, S., Tuerlinckx, F., & de Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847−862.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(5), 1016−1031.

Mehrazmay, R., Ghonsooly, B., & de la Torre, J. (2021). Detecting differential item functioning using cognitive diagnosis models: Applications of the Wald test and likelihood ratio test in a university entrance examination. *Applied Measurement in Education, 34*(4), 262–284.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525−543.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Routledge.

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement, 44*(4), 267−281.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R project.org/

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289−316.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323−348.

Sun, X., Liu, Y., Wang, S., Xin, T., Song, N., & Zhou, M. (2022). Using information matrix-based method to detect differential item functioning with multiple groups in cognitive diagnostic test. *Journal of Psychological Science, 45*(3), 710−717.

[孙小坚, 刘彦楼, 王诗梦, 辛涛, 宋乃庆, 周蔓. (2022). 认知诊断测验中基于信息矩阵的多群组 DIF 检验. *心理科学, 45*(3), 710−717.]

Tan, Z., de La Torre, J., Ma, W., Huh, D., Larimer, M. E., & Mun, E.−Y. (2023). A tutorial on cognitive diagnosis modeling for characterizing mental health symptom profiles using existing item responses. *Prevention Science, 24*(3), 480−492.

Tay, L., Huang, Q., & Vermunt, J. K. (2015). Item response theory with covariates (IRT-C): Assessing item recovery and differential item functioning for the three-parameter logistic model. *Educational and Psychological Measurement, 76*(1), 22−42.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287−305.

Tu, D., Cai, Y., Gao, X., & Wang, D. (2019). *Advanced cognitive diagnosis*. Beijing: Beijing Normal University Publishing Group.

[涂冬波, 蔡艳, 高旭亮, 汪大勋. (2019). 高级认知诊断. 北京: 北京师范大学出版社.]

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727−750.

Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a new instrument for depression with cognitive diagnosis models. *Frontiers in Psychology, 10*, 1306.

Wang, X. (2019). Development and verification of cognitive diagnostic test for cross-grade pupils' mathematics learning ability. *Chinese Exam, 8*, 71−78.

[王欣瑜. (2019). 跨年级小学数学学力认知诊断测验的开发与验证. *中国考试*, *8*, 71−78.]

Wang, Z., Guo, L., & Bian, Y. (2014). Comparison of DIF detecting methods in cognitive diagnostic test. *Acta Psychologica Sinica*, *46*(12), 1923−1932.

[王卓然, 郭磊, 边玉芳. (2014). 认知诊断测验中的项目功能差异检测方法比较. *心理学报, 46*(12), 1923−1932.]

Xi, C., Cai, Y., Peng, S., Lian, J., & Tu, D. (2020). A diagnostic classification version of schizotypal personality questionnaire using diagnostic classification models. *International Journal of Methods in Psychiatric Research, 29*(1), e1807.

Yuan, K. H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika*, *86*(2), 345−377.

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Unpublished doctoral dissertation). The University of North Carolina at Greensboro.

Development of main effect DIF and interactive DIF detection method in cognitive diagnosis assessments: A recursive partitioning-based perspective

LIU Kai[1,2], GUO Zhichen[1], WANG Qin[1], WANG Daxun[1], CAI Yan[1], TU Dongbo[1,3,4]

([1] *School of Psychology, Jiangxi Normal University, Nanchang* 330022, *China*)

([2] *College of Psychology, Liaoning Normal University, Dalian* 116029, *China*)

([3] *Jiangxi Laboratory of Philosophy and Social Sciences—Data Science and Intelligent Psychological Assessment and Service Laboratory of Jiangxi Normal University, Nanchang* 330022, *China*)

([4] *Jiangxi Provincial Key Laboratory of Intelligent Information Processing and Affective Computing, Nanchang* 330022, *China*)

Abstract

With the growing recognition of the advantages of cognitive diagnosis (CD) in psychological and educational measurement, applying the CD framework to test development has become an important research direction in the field of psychology. In the development of cognitive diagnostic assessments, detecting differential item functioning (DIF) remains a crucial quality control procedure to ensure test fairness and validity. However, existing CD-based DIF detection methods typically focus on a single covariate at a time. While these approaches are effective for identifying main effect DIF induced by a single covariate, they are limited in detecting interactive DIF caused by the interaction among multiple covariates. Such limitations may compromise the fairness and interpretability of assessment outcomes. To address this issue, the present study integrates CD modeling with recursive partitioning techniques by proposing a novel DIF detection method, namely the *Item-based Sequential Recursive Partitioning Method* (ISRPM). Building on the core principles of recursive partitioning, the ISRPM allows the simultaneous consideration of multiple covariates within a single DIF detection procedure and facilitates the identification of both main effect DIF and interactive DIF in cognitive diagnostic assessments.

To evaluate the performance of the proposed method, a series of Monte Carlo simulation studies were conducted focusing on two key objectives: (1) examining how factors such as sample size per group, DIF magnitude, DIF type, item quality, correlations among attributes, and the influence of demographic covariates on attribute mastery distribution affect the performance of ISRPM; and (2) comparing ISRPM with several existing DIF detection methods across varied experimental conditions. In addition, to illustrate its practical utility, ISRPM was applied to a cognitive diagnostic version of the Schizotypal Personality Questionnaire (DC-SPQ) and compared with five established DIF detection methods.

The results showed that (1) sample size, DIF magnitude, and item quality substantially influenced the performance of all methods; and (2) when items exhibited interactive DIF, ISRPM achieved higher detection accuracy than the Wald, LR, FS-Wald, FS-LR, and Mantel−Haenszel (MH) approaches. When only the main effect DIF was present, the overall performance of ISRPM was comparable to that of the existing methods.

These findings suggest that ISRPM provides a flexible and effective framework for identifying both main effect DIF and interactive DIF in cognitive diagnostic assessments, thereby contributing to methodological advancements in fairness evaluation and the broader application of CD-based measurement in psychological and educational measurement.

Keywords cognitive diagnosis assessments, differential item functioning, main effect DIF, interactive DIF, recursive partitioning

附　录

附表 1　属性掌握模式分布为均匀分布的条件下，不同认知诊断 DIF 检测方法在不同实验条件下的统计检验力结果

每个组别的样本容量	DIF 大小	DIF 形式	题目质量	TPR_I						TPR_{IC}					
				ISRPM	Wald	LR	FS-Wald	FS-LR	MH	ISRPM	Wald	LR	FS-Wald	FS-LR	MH
500	0.05	只有主效应 DIF	高	0.18	0.31	0.53	0.38	0.51	0.31	0.18	0.29	0.49	0.35	0.48	0.28
			中	0.19	0.29	0.43	0.31	0.41	0.31	0.16	0.25	0.37	0.27	0.35	0.28
		同时存在主效应 DIF 和交互式 DIF	高	0.47	0.58	0.85	0.69	0.85	0.56	0	0.03	0.10	0.04	0.10	0.06
			中	0.41	0.51	0.72	0.55	0.71	0.58	0.01	0.03	0.07	0.04	0.06	0.05
		只有交互式 DIF	高	0.02	0.03	0.12	0.06	0.13	0.08	0.01	0	0.02	0.01	0.01	0
			中	0.05	0.08	0.16	0.08	0.13	0.08	0.02	0.01	0.01	0	0.01	0
	0.1	只有主效应 DIF	高	0.68	0.75	0.96	0.82	0.96	0.76	0.68	0.75	0.95	0.82	0.96	0.75
			中	0.64	0.73	0.89	0.79	0.92	0.78	0.63	0.72	0.87	0.78	0.90	0.76
		同时存在主效应 DIF 和交互式 DIF	高	0.89	0.92	1	0.95	1	0.95	0.01	0.07	0.22	0.12	0.19	0.13
			中	0.93	0.95	0.99	0.97	1	0.97	0.05	0.13	0.23	0.15	0.20	0.17
		只有交互式 DIF	高	0.14	0.03	0.12	0.06	0.12	0.10	0.13	0	0.01	0	0.01	0
			中	0.17	0.07	0.15	0.08	0.13	0.10	0.14	0.01	0.01	0.01	0.01	0
1000	0.05	只有主效应 DIF	高	0.59	0.69	0.76	0.73	0.75	0.49	0.58	0.68	0.75	0.71	0.74	0.47
			中	0.53	0.62	0.67	0.64	0.67	0.48	0.51	0.61	0.66	0.62	0.65	0.46
		同时存在主效应 DIF 和交互式 DIF	高	0.90	0.94	0.96	0.95	0.96	0.83	0.13	0.14	0.19	0.14	0.18	0.13
			中	0.88	0.91	0.92	0.91	0.93	0.82	0.13	0.13	0.17	0.13	0.16	0.11
		只有交互式 DIF	高	0.22	0.07	0.10	0.08	0.11	0.09	0.20	0.01	0.01	0	0	0.01
			中	0.22	0.08	0.10	0.08	0.11	0.08	0.19	0	0	0	0	0.01
	0.1	只有主效应 DIF	高	0.98	0.99	1	0.99	1	0.96	0.98	0.99	1	0.99	1	0.96
			中	0.98	0.99	1	0.99	1	0.96	0.97	0.99	1	0.99	1	0.96
		同时存在主效应 DIF 和交互式 DIF	高	1	1	1	1	1	1	0.19	0.25	0.33	0.26	0.33	0.30
			中	1	1	1	1	1	1	0.25	0.30	0.36	0.29	0.34	0.32
		只有交互式 DIF	高	0.77	0.05	0.09	0.06	0.09	0.08	0.76	0.01	0.01	0.01	0.01	0
			中	0.81	0.07	0.08	0.06	0.08	0.07	0.81	0.01	0.01	0.01	0.01	0

续表

每个组别的样本容量	DIF 大小	DIF 形式	题目质量	TPR_I						TPR_{IC}					
				ISRPM	Wald	LR	FS-Wald	FS-LR	MH	ISRPM	Wald	LR	FS-Wald	FS-LR	MH
2000	0.05	只有主效应 DIF	高	0.92	0.95	0.95	0.95	0.95	0.80	0.91	0.94	0.95	0.95	0.95	0.79
			中	0.81	0.87	0.86	0.85	0.86	0.78	0.80	0.86	0.85	0.84	0.84	0.76
		同时存在主效应 DIF 和交互式 DIF	高	1	1	1	1	1	0.98	0.44	0.31	0.35	0.30	0.33	0.22
			中	0.99	1	1	1	1	0.98	0.44	0.30	0.29	0.28	0.29	0.25
		只有交互式 DIF	高	0.72	0.09	0.12	0.10	0.11	0.10	0.71	0	0.01	0	0	0
			中	0.63	0.09	0.10	0.10	0.11	0.09	0.61	0.01	0.01	0	0	0.01
	0.1	只有主效应 DIF	高	1	1	1	1	1	1	1	1	1	1	1	1
			中	1	1	1	1	1	1	1	1	1	1	1	1
		同时存在主效应 DIF 和交互式 DIF	高	1	1	1	1	1	1	0.79	0.63	0.66	0.63	0.66	0.49
			中	1	1	1	1	1	1	0.87	0.71	0.70	0.68	0.69	0.61
		只有交互式 DIF	高	1	0.08	0.10	0.09	0.09	0.08	1	0.03	0.07	0.01	0.06	0
			中	1	0.10	0.11	0.09	0.10	0.07	1	0.01	0.01	0.01	0.01	0

附表 2　属性掌握模式分布为均匀分布的条件下，不同认知诊断 DIF 检测方法在不同实验条件下的一类错误率结果

每个组别的样本容量	DIF 大小	DIF 形式	题目质量	TPR_I						TPR_{IC}					
				ISRPM	Wald	LR	FS-Wald	FS-LR	MH	ISRPM	Wald	LR	FS-Wald	FS-LR	MH
500	0.05	只有主效应 DIF	高	0.01	0.04	0.12	0.06	0.12	0.10	0.01	0.02	0.06	0.03	0.06	0.05
			中	0.06	0.09	0.17	0.10	0.14	0.09	0.05	0.05	0.09	0.05	0.07	0.04
		同时存在主效应 DIF 和交互式 DIF	高	0.01	0.03	0.11	0.06	0.11	0.11	0.01	0.02	0.06	0.03	0.06	0.06
			中	0.04	0.07	0.15	0.09	0.13	0.11	0.04	0.03	0.08	0.05	0.07	0.06
		只有交互式 DIF	高	0.02	0.03	0.12	0.06	0.11	0.09	0.02	0.02	0.06	0.03	0.06	0.04
			中	0.05	0.08	0.15	0.09	0.12	0.08	0.04	0.04	0.08	0.04	0.06	0.04
	0.1	只有主效应 DIF	高	0.01	0.04	0.12	0.06	0.11	0.13	0.01	0.02	0.06	0.03	0.06	0.07
			中	0.05	0.08	0.14	0.08	0.13	0.13	0.04	0.04	0.07	0.04	0.07	0.07
		同时存在主效应 DIF 和交互式 DIF	高	0.02	0.04	0.12	0.07	0.12	0.17	0.02	0.02	0.06	0.03	0.06	0.09
			中	0.05	0.09	0.15	0.09	0.13	0.18	0.04	0.04	0.08	0.05	0.07	0.09
		只有交互式 DIF	高	0.02	0.04	0.12	0.06	0.12	0.09	0.02	0.02	0.06	0.03	0.06	0.04
			中	0.05	0.07	0.14	0.09	0.13	0.08	0.04	0.04	0.07	0.04	0.07	0.04

续表

每个组别的样本容量	DIF 大小	DIF 形式	题目质量	TPR_I						TPR_{IC}					
				ISRPM	Wald	LR	FS-Wald	FS-LR	MH	ISRPM	Wald	LR	FS-Wald	FS-LR	MH
1000	0.05	只有主效应 DIF	高	0.03	0.06	0.11	0.08	0.11	0.11	0.03	0.03	0.05	0.04	0.05	0.06
			中	0.04	0.07	0.11	0.08	0.11	0.11	0.04	0.04	0.05	0.04	0.06	0.06
		同时存在主效应 DIF 和交互式 DIF	高	0.03	0.06	0.11	0.08	0.11	0.16	0.03	0.03	0.06	0.04	0.05	0.08
			中	0.04	0.07	0.10	0.08	0.10	0.16	0.03	0.03	0.05	0.04	0.05	0.08
		只有交互式 DIF	高	0.03	0.06	0.10	0.08	0.10	0.09	0.03	0.03	0.05	0.04	0.05	0.05
			中	0.04	0.07	0.11	0.09	0.11	0.10	0.04	0.04	0.06	0.04	0.06	0.05
	0.1	只有主效应 DIF	高	0.03	0.07	0.11	0.08	0.11	0.20	0.03	0.03	0.05	0.04	0.06	0.10
			中	0.04	0.07	0.10	0.08	0.11	0.20	0.04	0.04	0.05	0.04	0.06	0.10
		同时存在主效应 DIF 和交互式 DIF	高	0.03	0.06	0.10	0.08	0.11	0.29	0.03	0.03	0.05	0.04	0.05	0.15
			中	0.04	0.07	0.11	0.08	0.11	0.29	0.04	0.04	0.05	0.04	0.05	0.15
		只有交互式 DIF	高	0.03	0.06	0.11	0.08	0.10	0.10	0.03	0.03	0.05	0.04	0.05	0.05
			中	0.04	0.07	0.10	0.08	0.10	0.09	0.03	0.03	0.05	0.04	0.05	0.05
2000	0.05	只有主效应 DIF	高	0.05	0.09	0.11	0.09	0.11	0.14	0.04	0.04	0.06	0.05	0.05	0.07
			中	0.06	0.10	0.11	0.10	0.11	0.14	0.05	0.05	0.06	0.05	0.06	0.07
		同时存在主效应 DIF 和交互式 DIF	高	0.04	0.08	0.10	0.09	0.11	0.19	0.04	0.04	0.05	0.05	0.06	0.10
			中	0.06	0.11	0.11	0.10	0.11	0.20	0.05	0.05	0.06	0.05	0.06	0.11
		只有交互式 DIF	高	0.05	0.08	0.11	0.09	0.11	0.09	0.04	0.04	0.06	0.05	0.06	0.04
			中	0.06	0.09	0.10	0.09	0.10	0.09	0.05	0.05	0.05	0.05	0.05	0.04
	0.1	只有主效应 DIF	高	0.05	0.08	0.11	0.09	0.10	0.29	0.04	0.04	0.05	0.05	0.05	0.15
			中	0.07	0.11	0.11	0.10	0.11	0.30	0.06	0.06	0.06	0.05	0.06	0.16
		同时存在主效应 DIF 和交互式 DIF	高	0.05	0.09	0.11	0.10	0.11	0.47	0.04	0.04	0.06	0.05	0.06	0.25
			中	0.07	0.11	0.12	0.10	0.11	0.50	0.06	0.06	0.06	0.05	0.06	0.27
		只有交互式 DIF	高	0.05	0.08	0.11	0.09	0.10	0.09	0.04	0.04	0.05	0.05	0.05	0.04
			中	0.07	0.11	0.11	0.10	0.10	0.09	0.05	0.05	0.06	0.05	0.05	0.05