心理学报 2016, Vol. 48, No.5, 588−598
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2016.00588

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*

刘彦楼[1] 辛 涛[1,2] 李令青[3] 田 伟[2] 刘笑笑[1]

([1]北京师范大学发展心理研究所, 北京 100875) ([2]中国基础教育质量监测协同创新中心, 北京 100875)
([3]泰山学院教师教育学院, 山东泰安 271000)

摘 要 Hou, de la Torre 和 Nandakumar (2014)提出可以使用 Wald 统计量检验 DIF, 但其结果的一类错误率存在过度膨胀的问题。本研究中提出了一个使用观察信息矩阵进行计算的改进后的 Wald 统计量。结果表明: (1)使用观察信息矩阵计算的这一改进后的 Wald 统计量在 DIF 检验中具有良好的一类错误控制率, 尤其是在项目具有较高区分能力的时候, 解决了以往研究中一类错误率过度膨胀的问题。(2)随着样本量的增加以及 DIF 量的增大, 使用观察信息矩阵计算 Wald 统计量的统计检验力也在增加。

关键词 Wald 统计量; 项目功能差异; 认知诊断模型; 观察信息矩阵; 经验交叉相乘信息矩阵
分类号 B841

1 引言

认知诊断模型可以提供关于受测者知识或技能掌握程度的细粒度的、多维诊断性反馈信息, 因此, 引起了学生、教师、心理测量学家以及认知心理学家等的关注(Greeno, 1980; Leighton & Gierl, 2007), 是当前心理测量领域研究的热点之一。迄今为止, 研究者提出了许多认知诊断模型, 这些模型可以被分为一般性的认知诊断模型框架以及特殊的认知诊断模型。一般性的认知诊断模型框架, 主要包括 von Davier (2005)的一般诊断模型(*General Diagnostic Model,* GDM)、Henson, Templin 和 Willse (2009)提出的对数线性认知诊断模型(*Log-Linear Cognitive Diagnosis Model,* LCDM)以及 de la Torre (2011)的 G-DINA 模型, 常见的特殊的认知诊断模型有决定性输入, 噪音与门模型(*Deterministic Input, Noisy And Gate,* DINA) (de la Torre & Douglas, 2004; Haertel, 1989; Junker & Sijtsma, 2001), 补偿的重参数化统一模型(*Compensatory Reparameterized Unified Model,* C-RUM) (e.g., Hartz, 2002)等。

从统计上来讲, 以上这些一般性的认知诊断模型与特殊的认知诊断模型都属于有约束的潜在类别模型(von Davier, 2009)。这些“约束”主要是通过 Q 矩阵来实现的。Q 矩阵是一个设计矩阵, 其中的元素一般是“0”与“1”, 虽然有研究(Chen & de la Torre, 2013)已经将 Q 矩阵扩展为多级的, 但在绝大多数的实际应用中仍假定其是二分的, 因此本研究仍假定 Q 矩阵是二分的。在认知诊断模型中一般将受测者的知识或技能统称为潜在属性, 简称属性。Q 矩阵的功能在于设定认知诊断测验中项目与属性之间的对应关系, Q 矩阵中元素取值为 1 代表正确作答某一项目需要某一对应的属性, 取值为 0 则代表不需要。将认知诊断模型与 Q 矩阵在项目水平上进行组合, 可以反映出研究者对于受测者在作答项目时的潜在认知过程或操作的假定。

在使用认知诊断测验对于受测者的属性掌握状况进行诊断的时候, 研究者面临的一个重要的理论及现实问题是如何进行项目功能差异(*Differential Item Functioning,* DIF)检验。因为当测验中含有功能差异的项目时, 不仅会产生测验公平性的问题, 而

收稿日期: 2015−09−17
* 国家自然科学基金面上项目(31371047); 中央高校基本科研业务费专项资金资助(SKZZX2013028)。
通讯作者: 辛涛, E-mail: xintao@bnu.edu.cn

且也会影响到受测者属性掌握模式的判别(王卓然, 边玉芳, 郭磊, 2015)。在认知诊断模型中一个被广泛接受的 DIF 定义是不同组中具有相同属性掌握模式的受测者正确作答某一项目的概率不同(Hou et al., 2014; Li, 2008)。当前研究者们提出了一些不同的方法用于检验认知诊断模型中的 DIF (Hou et al., 2014; Li, 2008; 王卓然, 郭磊, 边玉芳, 2014; Li & Wang, 2015; Zhang, 2006)。Zhang (2006)提出使用 MH 法(Holland & Thayer, 1988; Mantel & Haenszel, 1959)以及 SIBTEST 法(Shealy & Stout, 1993), 用受测者的测验总分以及属性掌握模式作为匹配变量去检验 DINA 模型中的 DIF。Zhang (2006)所提出的方法中的不足之处在于：目标组以及对照组的项目参数以及属性掌握模式参数是作为一个整体被同时估计出来的, 因此会导致其估计值不准确; 另外, MH 法以及 SIBTEST 法只能检验一致性 DIF。Hou (2013)的研究中指出逻辑斯蒂克回归法(*Logistic Regression,* LR) (Swaminathan & Rogers, 1990), MH 法以及 SIBTEST 法的统计检验力都受到测验中 DIF 项目比例的影响。Li (2008)使用改进的高阶 DINA 模型(de la Torre & Douglas, 2004)去检验 DIF, 然而, Li 研究的不足之处在于：在某些模拟条件下, 经验一类错误率(指的是在实际模拟中所观察到的一类错误)过高或者过低; 另外这一方法只适用于高阶模型而非一般性的模型。Hou 等人(2014)提出使用 Wald 统计量检验项目功能差异, 并且认为 Wald 统计量的检验方法的效果接近或者是优于 MH 以及 SIBTEST 方法, 然而, Hou 等人所提出的 Wald 统计量存在以下不足：首先是一类错误率过高, 不符合预先设置的显著性水平; 其次, 统计功效研究中, 正确拒绝率是使用的每个模拟条件下的 10,000 次重复所获得统计量的经验分布来计算的, 这使得其研究结果无法推广到一般性的模型以及实际应用中。另外, 需要指出, Hou 等人(2014)在计算 Wald 统计量时使用的是 de la Torre (2009, 2011)所提出认知诊断模型信息矩阵的计算方法。王卓然等人(2014)的研究发现尽管 Wald 方法的检验力要高于 LR 法与 MH 法, 但是也存在一类错误率膨胀的问题。Li 和 Wang (2015)比较了使用马尔可夫链蒙特卡罗(*Markov chain Monte Carlo,* MCMC)法计算项目参数时, LCDM-DIF 方法以及 Wald 方法在评价项目功能差异时的表现。Li 和 Wang 发现, 他们所使用的 LCDM-DIF 方法以及 Wald 统计量具有较好的一类错误控制率(仅有稍许的膨胀), 并且当被比较的组数为 3 时, Wald 统计量的统计功效要优于 LCDM-DIF。

通过以上文献综述我们可以发现, 尽管研究者们一致地认为 Wald 统计量在检验 DIF 时有着高的统计检验力, 但是不同的研究对于 Wald 统计量的一类错误控制率的表现却有着不同的结果。澄清不同的方法构建的 Wald 统计量为什么在一类错误控制率的表现不同这个问题, 不仅在理论上具有重要意义, 而且对于测验实践也有重要意义。Hou 等人(2014)以及王卓然等人(2014)所使用 Wald 统计量, 均是基于 de la Torre (2009, 2011)所提出的项目参数的经验交叉相乘信息矩阵而构建的, 而非基于全部的模型参数(即模型中所有自由估计的参数)。然而, 相关研究指出(Tian, Cai, Thissen, & Xin, 2013; Paek & Cai, 2013)通过对信息矩阵求逆计算误差—协方差矩阵时, 信息矩阵应该包括全部的模型参数, 而非仅仅是项目参数; 并且研究发现当模型的参数是通过 EM (Expectation-Maximization)方法(de la Torre, 2009, 2011)所估计获得时, 应该通过对观察信息矩阵(基于样本观测数据所计算的信息矩阵, 有些研究中也将其简称为观察矩阵)求逆的方法计算误差—协方差矩阵(Kenward & Molenberghs, 1998; Louis, 1982)。已有研究发现在项目反应理论中观察信息矩阵的逆可以很好的渐近误差—协方差矩阵(Paek & Cai, 2013)。

针对以往研究中 Wald 统计量构建方法的局限, 解决在认知诊断模型中更加准确地估计 Wald 统计量这一重大理论问题, 促进认知诊断测验在实践中的运用, 本研究拟将观察信息矩阵的计算方法引入到认知诊断模型中, 期望获得一个好的误差—协方差矩阵的估计方法, 从而改进 Wald 统计量在检验 DIF 时的表现。研究包括主要包括以下 3 个部分：首先, 介绍用于检验认知诊断模型中 DIF 的 Wald 统计量的构建, 重点强调误差—协方差矩阵在构建中所起的重要作用; 其次, 介绍认知诊断模型中经验交叉相乘信息矩阵以及观察信息矩阵的计算方法; 第三, 采用模拟的方法, 探索本研究所提出的改进后的 Wald 统计量在计算 DIF 时的一类错误控制率以及统计检验力的表现, 并且与通过经验交叉相乘信息矩阵而构建的 Wald 统计量所获得的结果进行比较; 为了更好的说明本研究中的研究结果, 我们也将本研究的结果与其他采用相同实验设计的研究的结果(如, Hou et al., 2014; Li & Wang, 2015)进行了直接的比较。

2 改进的 Wald 统计量的计算方法

在本研究中，我们将使用 LCDM 作为例子，说明在认知诊断模型中如何应用改进后的 Wald 统计量进行 DIF 检验。LCDM 是一个广义的认知诊断模型，对于其中的参数进行约束，便可以获得一些特殊的模型，如 DINA 以及 C-RUM 等(Henson et al., 2009)。

我们首先假定，在认知诊断测验中共有 J 个项目以及 K 个属性，并且为了方便与以往研究结果进行比较(Hou et al., 2014; Li & Wang, 2015)设定属性之间不存在层级关系，因此，Q 矩阵的维度为 $J\times K$，所有可能的属性掌握模式为 $L=2^K$。为行文方便，我们以加粗的英文或希腊字母表示向量以及矩阵，设 $\boldsymbol{\beta}=(\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_j',\ldots,\boldsymbol{\beta}_J')'$ 为模型中所包含的所有项目参数，其中 $\boldsymbol{\beta}_j=(\lambda_{j,0},\boldsymbol{\lambda}_j')'$，$\lambda_{j,0}$ 为项目 j 的截距参数，$\boldsymbol{\lambda}_j$ 为项目 j 的主效应以及交互效应参数向量。在测验中共有 N 个受测者，X_{ij} 为受测者 $i(i=1,\ldots,I)$ 在项目 j 上的作答，$\boldsymbol{\alpha}_i$ 为受测者 i 的属性掌握模式。根据 LCDM，属性掌握模式为 $\boldsymbol{\alpha}_i$ 的受测者，答对项目 j 的概率可以表示如下：

$$P_j(\boldsymbol{\alpha}_i)=P(X_{ij}=1\,|\,\boldsymbol{\alpha}_i)=\frac{\exp\left[\lambda_{j,0}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\boldsymbol{q}_j)\right]}{1+\exp\left[\lambda_{j,0}+\boldsymbol{\lambda}_j'\mathbf{h}(\boldsymbol{\alpha}_i,\boldsymbol{q}_j)\right]} \tag{1}$$

映射函数 $\mathbf{h}$ 是用来设定 $\boldsymbol{\alpha}_i$ 与 $\boldsymbol{q}_j$ 之间线性关系的(Henson et al., 2009)，其中向量 $\boldsymbol{q}_j$ 代表项目 j 所对应的 Q 矩阵元素。对于饱和 LCDM 模型进行不同的约束，可以获得一些特殊的认知诊断模型。在本研究中，我们将 LCDM 中的 $\boldsymbol{\lambda}_j$ 参数的主效应以及低阶交互效应全部设置为 0，只保留最高阶交互效应项，即 $\boldsymbol{\beta}_j=(\lambda_{j,0},\lambda_{j,Kj,(1,\ldots,Kj)})'$，以获得 DINA 模型。LCDM 与 DINA 模型之间的等价关系，可以表示如下，

$$g_j=\frac{\exp\left[\lambda_{j,0}\right]}{1+\exp\left[\lambda_{j,0}\right]} \tag{2}$$

$$1-s_j=\frac{\exp\left[\lambda_{j,0}+\lambda_{j,K_j,(1,\ldots,K_j)}\right]}{1+\exp\left[\lambda_{j,0}+\lambda_{j,K_j,(1,\ldots,K_j)}\right]} \tag{3}$$

其中，g_j 为项目 j 的猜测参数，代表受测者没有掌握项目所要求的全部属性，只能靠猜测答对这一项目；s_j 为这一项目的滑动参数，代表即使受测者掌握了项目所要求的全部属性，但由于受测者的粗心等因素的存在，仍可能在项目 j 上做出错误的反应。$K_j=\sum_{k=1}^{K}q_{jk}$ 为作答项目 j 所需的属性数量。即，DINA 模型中只包含饱和 LCDM 的截距 $\lambda_{j,0}$ 与最高阶交互效应 $\lambda_{j,K_j,(1,\ldots,K_j)}$ 这两个参数。

LCDM 假定在给定属性掌握模式 $\boldsymbol{\alpha}_i$ 的条件下，受测者 i 在各个项目上的作答是独立的，其反应向量 $\mathbf{X}_i$ 的似然函数，可以表示如下，

$$L(\mathbf{X}_i|\boldsymbol{\alpha}_i)=\prod_{j=1}^{J}P_j(\boldsymbol{\alpha}_i)^{X_{ij}}\left[1-P_j(\boldsymbol{\alpha}_i)\right]^{1-X_{ij}} \tag{4}$$

反应向量 $\mathbf{X}_i$ 的边际概率，可以表示为：

$$L(\mathbf{X}_i)=\sum_{l=1}^{L}L(\mathbf{X}_i|\boldsymbol{\alpha}_l)\,p(\boldsymbol{\alpha}_l) \tag{5}$$

在公式(5)中，$p(\boldsymbol{\alpha}_l)$ 是属性掌握模式 $\boldsymbol{\alpha}_l$ 的概率，在 LCDM 中，所有属性掌握模式的概率之和为 1。为满足这一约束，本研究参考 Rupp, Templin 和 Henson (2010)所使用的概念，设 $\boldsymbol{\eta}=(\eta_1,\ldots,\eta_L)'$ 为模型的结构参数(*structural parameters*)，用以描述任一受测者来自特定属性掌握模式的概率，使用以下表达式，

$$p(\boldsymbol{\alpha}_l)=\frac{\exp(\eta_l)}{\sum_{l=1}^{L}\exp(\eta_l)} \tag{6}$$

并且对结构参数 $\boldsymbol{\eta}$ 施加约束，固定其中任一参数为 0，一般而言，可以选择固定最后一个模型参数 η_L 为 0。

再进一步假定，受测者之间的作答都是独立的，因此所有受测者作答 $\mathbf{X}$ 的似然函数为可以用如下公式来表示，

$$L(\mathbf{X})=\prod_{i=1}^{N}L(\mathbf{X}_i) \tag{7}$$

对公式(7)取对数似然函数，对参数求一阶导数，然后通过 EM 算法，可以求得所有自由估计的模型参数。详细的推导过程，感兴趣的读者可以参考 de la Torre (2009, 2011)，虽然在这两篇文献中，分别是以 DINA 以及 G-DINA 为例进行推导的，但其原理是一致的。通过公式(6)与(7)可以发现，认知诊断模型中除了有项目参数 $\boldsymbol{\beta}$ 外，还有结构参数 $\boldsymbol{\eta}$，因此，可以将认知诊断模型的模型参数表述为 $\boldsymbol{\gamma}=(\boldsymbol{\beta}',\boldsymbol{\eta}')'$。饱和 LCDM 中自由估计的个数为 $F=\sum_{j=1}^{J}2^{K_j}+(L-1)$，其特例 DINA 模型中自由估计的模型参数个数为 $F=2J+(L-1)$。因此，构建的 LCDM 信息矩阵(或者是其特例 DINA 等模型)，应该是 $F\times F$ 维的，它

不仅要包含项目参数, 而且还要包含结构参数, 而非 de la Torre (2009, 2011)所认为的仅项目参数 $\boldsymbol{\beta}$ 需要被考虑。

信息矩阵的逆就是模型的渐近方差—协方差矩阵。设向量 $\hat{\boldsymbol{\beta}}_j=(\hat{\lambda}_{j,0},\hat{\boldsymbol{\lambda}}_{j,K_j,(1,\ldots,K_j)})'$ 为 DINA 模型中项目 j 的参数估计值, $\hat{\boldsymbol{\Sigma}}_j$ 为项目 j 的渐近方差—协方差矩阵, 是项目 j 参数在模型的方差—协方差矩阵中的对应部分。据此, 可以在项目水平上构建用于 DIF 检验的 Wald 统计量(Li & Wang, 2015),

$$W_j=(\hat{\boldsymbol{\beta}}_{Rj}-\hat{\boldsymbol{\beta}}_{Fj})'(\hat{\boldsymbol{\Sigma}}_{Rj}+\hat{\boldsymbol{\Sigma}}_{Fj})^{-1}(\hat{\boldsymbol{\beta}}_{Rj}-\hat{\boldsymbol{\beta}}_{Fj}) \tag{8}$$

在虚无假设下, $\mathrm{H}_0:\hat{\boldsymbol{\beta}}_{Rj}-\hat{\boldsymbol{\beta}}_{Fj}=0$, 当为 DINA 模型时 Wald 统计量 W_j 服从自由度为 2 的渐近卡方分布。$\hat{\boldsymbol{\beta}}_{Rj}$ 为对照组中项目 j 参数, $\hat{\boldsymbol{\beta}}_{Fj}$ 为目标组中的项目参数; $\hat{\boldsymbol{\Sigma}}_{Rj}$ 与 $\hat{\boldsymbol{\Sigma}}_{Fj}$ 分别为对照组与目标组中项目 j 参数估计值所对应的渐近方差—协方差矩阵。项目 j 的渐近方差—协方差矩阵 $\hat{\boldsymbol{\Sigma}}_j$ 可以表达如下,

$$\hat{\boldsymbol{\Sigma}}_j=\begin{pmatrix}\hat{\sigma}^2_{\lambda_{j,0}} & \hat{\sigma}_{\lambda_{j,0},\boldsymbol{\lambda}_{j,K_j,(1,\ldots,K_j)}}\\ \hat{\sigma}_{\lambda_{j,0},\boldsymbol{\lambda}_{j,K_j,(1,\ldots,K_j)}} & \hat{\sigma}^2_{\boldsymbol{\lambda}_{j,K_j,(1,\ldots,K_j)}}\end{pmatrix} \tag{9}$$

从公式(8)可以发现方差—协方差矩阵估计的准确性, 对于 Wald 统计量会产生重大的影响, 这也就是说 LCDM 中信息矩阵的估计会对 Wald 统计量的计算产生重大影响。

EM 算法(Dempster, Laird, & Rubin, 1977)对于心理测量学产生了非常大的影响, 它将复杂的计算非完整数据似然函数最大值问题转换为较为简单的一系列伪完整数据问题, 在认知诊断模型分析软件中得到了广泛的应用。然而, 在通过 EM 算法计算参数时, 信息矩阵(或者是其逆方差—协方差矩阵)并不是伴随产生的, 因此, 需要去进行专门的计算。研究发现, 当使用期望—最大化算法去计算模型的极大似然估计值时, 使用观察信息矩阵能够很好的去渐近模型的方差—协方差矩阵(Louis, 1982), 感兴趣的研究者可以参考 Kenward 和 Molenberghs (1998)的研究。对于 LCDM 而言, 包含所有自由估计参数的经验交叉相乘信息矩阵的公式可以表达如下:

$$\mathcal{I}_{XPD}=\frac{\partial\ell(\boldsymbol{X})}{\partial\boldsymbol{\gamma}}\cdot\frac{\partial\ell(\boldsymbol{X})}{\partial\boldsymbol{\gamma}'} \tag{10}$$

$\ell(\boldsymbol{X})=\log L(\mathbf{X})$ 是观察到的受测者反应模式似然函数的对数。对于 $\mathcal{I}_{XPD}$ 求逆, 便可以获得模型的渐近方差—协方差矩阵, $\hat{\boldsymbol{\Sigma}}_j$ 为项目 j 在渐近方差—协方差矩阵中的对应部分, 分别获得 $\hat{\boldsymbol{\Sigma}}_{Rj}$ 与 $\hat{\boldsymbol{\Sigma}}_{Fj}$ 后, 将其代入公式(11)中可以获得 Wald 统计量的值。观察信息矩阵的公式可以表达如下(Kenward & Molenberghs, 1998; Louis, 1982):

$$\mathcal{I}_{Obs}=\frac{\partial^2\ell(\boldsymbol{X})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} \tag{11}$$

对于 $\mathcal{I}_{Obs}$ 求逆, 可以获得模型的渐近方差—协方差矩阵的估计值, $\hat{\boldsymbol{\Sigma}}_j$ 同样为项目 j 在渐近方差—协方差矩阵中的对应部分, 获得 $\hat{\boldsymbol{\Sigma}}_{Rj}$ 与 $\hat{\boldsymbol{\Sigma}}_{Fj}$ 后, 可以获得使用观察信息矩阵方法计算的 Wald 统计量的值。

3　方法

3.1　研究设计

采用 Monte Carlo 的方法进行研究, 受测者的作答反应、模型的参数估计以及 Wald 统计量的计算均采用 *R* 语言(R Core Team, 2015)编程实现。每种实验条件均重复 1000 次, 以获得稳定的结果。为了便于与以往研究结果进行直接的比较, 本研究所采用 Hou 等人(2014)所设计的实验条件, 这些实验条件也被 Li 和 Wang (2015)所采用。与 Hou 等人(2014)研究不同的是, 本研究中 Wald 统计量的计算是通过包含全部模型参数的观察信息矩阵或者是经验交叉相乘信息矩阵所计算获得的。

本研究中所采用 Q 矩阵中包含 30 个测验项目, 5 个属性, 并且限制每个项目所包含的属性数量最多为 3。Q 矩阵采用平衡设计, 每个属性被项目所测量的次数相等, 同样使包含 1、2、3 个属性的项目数量也相等即包含 1、2、3 个属性的项目分别有 10 个。具体的 Q 矩阵设计见表 1。

为方便与以往研究结果进行直接对比, 本研究设计中的数据生成模型也同样采用 DINA 模型, 对照组中的猜测以及滑动参数设置为相等, 且有三个水平: 0.1, 0.2 以及 0.3, 猜测以及滑动参数值设置的越小, 说明项目越能够区分出受测者是否掌握了所测的属性(Templin & Henson, 2006)。DIF 类型有两个水平: 一致性 DIF 以及非一致性 DIF。一致性 DIF 指的是对于某一个组而言, 正确作答某个项目的概率在所有可能的属性掌握模式下均一致性地高或者是低; 非一致性 DIF 指的是正确作答某个项目的概率在一些属性掌握模式下高, 在另外一些属性掌握模式下低, 或者是相反, 即正确作答的概率具有非一致性。DIF 大小有两个水平: 0.05 与 0.1, 当项目参数值为 0.1 时仅考虑了 0.05 这一水平的 DIF 大小, 以防项目参数值等于 0。样本大小有两个水平: 500 与 1000。在认知诊断模型中样本的大小会对模

表 1 Q 矩阵

项目	属性 1	属性 2	属性 3	属性 4	属性 5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	1	0	0	0
7	1	0	0	0	1
8	0	1	1	0	0
9	0	0	1	1	0
10	0	0	0	1	1
11	1	1	1	0	0
12	1	1	0	0	1
13	1	0	0	1	1
14	0	1	1	1	0
15	0	0	1	1	1
16	1	0	0	0	0
17	0	1	0	0	0
18	0	0	1	0	0
19	0	0	0	1	0
20	0	0	0	0	1
21	1	0	1	0	0
22	1	0	0	1	0
23	0	1	0	1	0
24	0	1	0	0	1
25	0	0	1	0	1
26	1	1	0	1	0
27	1	0	1	1	0
28	1	0	1	0	1
29	0	1	1	0	1
30	0	1	0	1	1

型参数估计值的精确性产生影响, 进而也会影响到Wald统计量的计算, 因此, 样本大小也是一个需要考虑的重要因素。

3.2 评价指标

本研究中所采用的评价指标为经验一类错误率以及统计检验力。经验一类错误率是通过 1000 次模拟中, 错误地检验出每个项目出现 DIF 的百分比, 然后参照以往研究结果的呈现方式(Hou et al., 2014), 分别对包含一个、两个以及三个属性的项目求平均。统计检验力指的是在这 1000 次循环中正确拒绝原假设的比例。当认知诊断测验中不存在 DIF 时, 如果我们所构建 Wald 统计量是渐近卡方分布的, 那么它观察到的一类错误率应该符合预先设置的理论上的一类错误控制率, 如 0.05; 如果在认知诊断测验中存在 DIF, 那么 Wald 统计量正确拒绝的比例越高, 说明它能够检验出 DIF 项目的能力越强。

4 研究结果

4.1 经验一类错误率

表 2 呈现了各个实验条件下的使用观察信息矩阵估计方法的 Wald 统计量获得的平均经验一类错误率。计算一类错误控制率所使用的参照分布为自由度为 2 的卡方分布。通过表 2 可以发现当项目的猜测以及滑动参数都比较小的时候, 即项目能够较为有效的区分受测者是否掌握所测属性的时, 一类错误控制率能够很好的接近预先设置的显著性水平。随着样本量的增大, 一类错误控制率的表现也越好。另外, 不论是包含一个、两个还是三个属性的项目, 其观察一类错误率均能较好的接近 0.05 这一显著性水平。另外需要指出的是, 尽管在当样本量较小($N = 500$)且项目的猜测参数以及滑动参数较大的情况下($g_j = s_j = 0.3$), 平均的经验一类错误率表现较差, 但根据 Bradley (1978)的健壮宽松准则(当显著性水平为 0.05 时经验一类错误控制率在 0.025 与 0.075 之间), 仍然可以认为是得到了较好的控制。可以发现, 本研究中所提出的改进的 Wald 统计量计算方法所获得的结果并不存在过度膨胀的现象, 这与 Hou 等人(2014)以及王卓然等人(2015)的结果恰好相反, 说明本研究中所提出的 Wald 统计量的计算方法明显优于以上两个研究所使用的 Wald 统计量的计算方法。通过比较表 2 与表 3 中的一类错误控制率可以发现基于观察信息矩阵计算的 Wald 统计量的表现要优于基于经验交叉相乘信息矩阵而计算的 Wald 统计量。基于经验交叉相乘矩阵而获得的 Wald 统计量的一类错误控制率较为保守, 但是表 3 的结果同样显示包含一个、两个以及三个属性的项目的一类错误控制率仍大致相等。

Li 和 Wang (2015)在 MCMC 框架下采用 LCDM-DIF 以及 Wald 统计量对于 DIF 检验方法进行了研究, 在其研究一中同样采用了 Hou 等人(2014)的研究设计, 因此本研究的研究结果同样也是可以直接与 Li 等人的结果进行比较。通过对比研究结果可以发现, 本研究中所提出基于观察信息矩阵计算的 Wald 统计量与 Li 等人的研究中所使用的 LCDM-DIF 以及 Wald 统计量均具有较好的一类错误控制率。一个非常有意思的现象是在本研究中的一些实验条件下(见表 2)Wald 统计量一类错误率有细微的保守而 Li 等人研究结果中的 LCDM-DIF

表 2　基于观察信息矩阵的平均的经验一类错误率($\alpha = 0.05$)

项目参数值	$N = 500$			$N = 1000$		
	$K_j = 1$	$K_j = 2$	$K_j = 3$	$K_j = 1$	$K_j = 2$	$K_j = 3$
0.1	0.043	0.046	0.051	0.047	0.048	0.048
0.2	0.044	0.047	0.041	0.053	0.050	0.047
0.3	0.032	0.039	0.037	0.039	0.046	0.042

表 3　基于经验交叉相乘信息矩阵的平均的经验一类错误率($\alpha = 0.05$)

项目参数值	$N = 500$			$N = 1000$		
	$K_j = 1$	$K_j = 2$	$K_j = 3$	$K_j = 1$	$K_j = 2$	$K_j = 3$
0.1	0.023	0.023	0.034	0.036	0.036	0.037
0.2	0.026	0.022	0.019	0.039	0.037	0.034
0.3	0.024	0.024	0.020	0.036	0.038	0.031

以及 Wald 统计量在某些实验条件中一类错误率却有稍许膨胀。从公式(8)中可以发现 Wald 统计量的准确性, 依赖于模型参数估计值的准确性。当受测者数量较少(如 $N = 500$ 时)或者是模型中的“噪音”过大时(如项目的猜测与滑动参数均为 0.3 时), 模型参数估计值的准确性会受到相对较大的影响, 因此, 在本研究的 $N = 500$ 以及 $g_j = s_j = 0.3$ 这两种条件下 Wald 统计量一类错误率有细微的保守。

4.2　统计检验力

表 4 中呈现的是当认知诊断测验中存在一致性 DIF 时的考察一个、两个以及三个属性项目在 1000 次循环中的基于观察信息矩阵计算的 Wald 统计量的平均经验拒绝比例, 所使用的参照分布同样为自由度为 2 的卡方分布。从表 4 中可以看出, 随着 DIF 的增大, Wald 统计量的统计检验力也会随之增大, 并且当项目的猜测以及滑动参数都为 0.2 的时候, 总平均的拒绝率要大于同为 0.3 时的项目参数值的条件。这是由于同项目参数值 0.3 相比, DIF 大小为 0.1 时, 这一值对于项目参数值 0.2 而言相对更大。随着样本量的增加, Wald 统计量的统计检验力也在变大, 即样本量的大小对用于检验 DIF 的 Wald 统计量而言也是一个重要因素。因为随着样本量的增加, 模型参数估计值的准确性也会增加, 进而会使得参数估计值的标准误变小, 因此, 在对照组与目标组项目参数差异相等的情况下, 更倾向于获得一个大的 Wald 统计量的值。另外, 通过观察平均值可以发现, 当目标组具有负向的 DIF 时, 同正向 DIF 相比, Wald 统计量的统计检验力更大。比较表 4 与表 5, 可以发现基于观察信息矩阵的 Wald 统计量的统计检验力均要明显优于基于经验交叉相乘信息矩阵的 Wald 统计量的统计检验力。这也说明基于经验交叉相乘信息矩阵的 Wald 统计量存在保守的问题。

表 6 中呈现的是非一致性 DIF 条件下采用观察信息矩阵的 Wald 统计量的 1000 次模拟结果, 计算统计检验力所使用的参照分布同样为自由度为 2 的卡方分布。从表 6 中同样可以发现随着 DIF 的增大,

表 4　基于观察信息矩阵的一致性 DIF 的平均经验统计检验力($\alpha = 0.05$)

项目参数	DIF 大小	DIF 类型		$N = 500$				$N = 1000$			
		Δ_{g_j}	Δ_{s_j}	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均
0.10	0.05	+	+	0.579	0.567	0.565	0.57	0.890	0.889	0.883	0.887
		–	–	0.583	0.699	0.778	0.686	0.897	0.951	0.968	0.939
0.20	0.05	+	+	0.268	0.321	0.361	0.317	0.518	0.599	0.639	0.585
		–	–	0.284	0.358	0.419	0.354	0.529	0.645	0.724	0.633
	0.10	+	+	0.851	0.888	0.905	0.881	0.990	0.998	0.998	0.995
		–	–	0.852	0.943	0.975	0.923	0.992	1.000	1.000	0.997
0.30	0.05	+	+	0.118	0.171	0.224	0.171	0.237	0.347	0.465	0.350
		–	–	0.115	0.193	0.257	0.188	0.241	0.370	0.488	0.366
	0.01	+	+	0.428	0.592	0.718	0.579	0.795	0.913	0.968	0.892
		–	–	0.447	0.663	0.815	0.642	0.810	0.943	0.988	0.913

注：+表示正向 DIF; –表示负向 DIF。

表 5 基于经验交叉相乘信息矩阵的一致性 DIF 的平均经验统计检验力($\alpha = 0.05$)

项目参数	DIF 大小	DIF 类型		$N = 500$				$N = 1000$			
		Δ_{g_j}	Δ_{s_j}	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均
0.10	0.05	+	+	0.469	0.461	0.465	0.465	0.871	0.863	0.860	0.865
		–	–	0.476	0.621	0.730	0.609	0.873	0.941	0.963	0.925
0.20	0.05	+	+	0.197	0.233	0.270	0.233	0.478	0.553	0.598	0.543
		–	–	0.205	0.274	0.323	0.267	0.488	0.603	0.684	0.592
	0.10	+	+	0.790	0.830	0.851	0.824	0.988	0.996	0.997	0.994
		–	–	0.796	0.913	0.957	0.889	0.989	0.999	1.000	0.996
0.30	0.05	+	+	0.099	0.133	0.167	0.133	0.222	0.323	0.429	0.325
		–	–	0.093	0.150	0.196	0.146	0.225	0.343	0.452	0.340
	0.01	+	+	0.406	0.544	0.651	0.534	0.786	0.903	0.962	0.883
		–	–	0.396	0.607	0.758	0.587	0.797	0.935	0.985	0.906

表 6 基于观察信息矩阵的非一致性 DIF 的平均经验统计检验力($\alpha = 0.05$)

项目参数	DIF 大小	DIF 类型		$N = 500$				$N = 1000$			
		Δ_{g_j}	Δ_{s_j}	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均
0.10	0.05	+	–	0.698	0.748	0.791	0.746	0.959	0.969	0.976	0.968
		–	+	0.441	0.516	0.562	0.506	0.764	0.822	0.857	0.814
		+	0	0.362	0.172	0.083	0.206	0.694	0.392	0.196	0.427
		0	+	0.245	0.418	0.520	0.394	0.470	0.730	0.821	0.674
		–	0	0.251	0.140	0.089	0.160	0.491	0.260	0.155	0.302
		0	–	0.360	0.619	0.748	0.576	0.693	0.914	0.960	0.855
0.20	0.05	+	–	0.305	0.370	0.418	0.365	0.561	0.652	0.732	0.648
		–	+	0.216	0.276	0.334	0.275	0.383	0.503	0.603	0.496
		+	0	0.175	0.107	0.072	0.118	0.325	0.177	0.113	0.205
		0	+	0.133	0.238	0.308	0.226	0.233	0.449	0.580	0.420
		–	0	0.147	0.092	0.066	0.102	0.256	0.147	0.096	0.166
		0	–	0.160	0.298	0.383	0.281	0.298	0.545	0.670	0.504
	0.10	+	–	0.923	0.964	0.982	0.956	0.991	0.998	1.000	0.996
		–	+	0.534	0.697	0.834	0.688	0.832	0.948	0.989	0.923
		+	0	0.631	0.349	0.170	0.384	0.920	0.656	0.385	0.654
		0	+	0.382	0.706	0.842	0.643	0.675	0.953	0.991	0.873
		–	0	0.413	0.230	0.138	0.260	0.725	0.433	0.245	0.468
		0	–	0.604	0.904	0.970	0.826	0.903	0.998	1.000	0.967
0.30	0.05	+	–	0.144	0.205	0.265	0.204	0.254	0.370	0.484	0.369
		–	+	0.125	0.166	0.214	0.168	0.148	0.239	0.360	0.249
		+	0	0.088	0.068	0.052	0.069	0.154	0.099	0.069	0.107
		0	+	0.067	0.146	0.208	0.140	0.097	0.248	0.383	0.243
		–	0	0.068	0.062	0.048	0.059	0.116	0.082	0.064	0.087
		0	–	0.084	0.168	0.247	0.167	0.132	0.303	0.462	0.299
	0.10	+	–	0.530	0.708	0.836	0.691	0.847	0.948	0.989	0.928
		–	+	0.360	0.481	0.610	0.484	0.420	0.588	0.791	0.600
		+	0	0.299	0.166	0.100	0.189	0.568	0.318	0.192	0.359
		0	+	0.181	0.408	0.601	0.397	0.255	0.650	0.878	0.594
		–	0	0.195	0.129	0.088	0.137	0.304	0.160	0.105	0.190
		0	–	0.271	0.608	0.798	0.559	0.501	0.893	0.982	0.792

表 7　基于经验交叉相乘信息矩阵的非一致性 DIF 的平均经验统计检验力($\alpha = 0.05$)

项目参数	DIF 大小	DIF 类型		$N = 500$				$N = 1000$			
		Δ_{g_j}	Δ_{s_j}	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均	$K_j = 1$	$K_j = 2$	$K_j = 3$	总平均
0.10	0.05	+	–	0.592	0.669	0.740	0.667	0.947	0.961	0.972	0.960
		–	+	0.343	0.414	0.476	0.411	0.730	0.787	0.833	0.783
		+	0	0.258	0.097	0.049	0.134	0.646	0.335	0.148	0.376
		0	+	0.177	0.332	0.439	0.316	0.429	0.693	0.794	0.639
		–	0	0.168	0.088	0.052	0.103	0.438	0.218	0.124	0.260
		0	–	0.266	0.541	0.706	0.504	0.653	0.898	0.953	0.835
0.20	0.05	+	–	0.227	0.278	0.326	0.277	0.517	0.611	0.695	0.608
		–	+	0.160	0.206	0.245	0.204	0.345	0.461	0.560	0.455
		+	0	0.115	0.061	0.035	0.070	0.284	0.144	0.084	0.171
		0	+	0.093	0.173	0.227	0.164	0.202	0.407	0.538	0.382
		–	0	0.096	0.052	0.032	0.060	0.219	0.117	0.074	0.137
		0	–	0.109	0.221	0.298	0.209	0.263	0.500	0.631	0.465
	0.10	+	–	0.883	0.941	0.967	0.930	0.998	0.999	1.000	0.999
		–	+	0.480	0.635	0.779	0.631	0.818	0.940	0.986	0.914
		+	0	0.531	0.237	0.092	0.287	0.903	0.605	0.324	0.611
		0	+	0.314	0.628	0.778	0.573	0.647	0.944	0.987	0.859
		–	0	0.324	0.148	0.073	0.182	0.684	0.376	0.195	0.418
		0	–	0.522	0.864	0.950	0.779	0.885	0.998	0.999	0.961
0.30	0.05	+	–	0.110	0.155	0.193	0.153	0.236	0.339	0.443	0.339
		–	+	0.110	0.139	0.169	0.139	0.168	0.244	0.348	0.253
		+	0	0.064	0.042	0.030	0.045	0.139	0.083	0.054	0.092
		0	+	0.062	0.113	0.157	0.111	0.098	0.236	0.355	0.230
		–	0	0.057	0.042	0.026	0.042	0.113	0.073	0.052	0.079
		0	–	0.066	0.127	0.187	0.126	0.123	0.277	0.424	0.275
	0.10	+	–	0.478	0.645	0.780	0.635	0.835	0.939	0.985	0.920
		–	+	0.360	0.462	0.576	0.466	0.470	0.641	0.828	0.646
		+	0	0.247	0.114	0.055	0.139	0.540	0.284	0.154	0.326
		0	+	0.170	0.381	0.561	0.370	0.283	0.675	0.889	0.616
		–	0	0.168	0.095	0.052	0.105	0.326	0.166	0.096	0.196
		0	–	0.232	0.544	0.739	0.505	0.483	0.880	0.977	0.780

Wald 统计量的统计检验力也在增大。随着样本量的增加, Wald 统计量的统计检验力同样是在增大的。而且在 DIF 大小相同条件下, 当项目的猜测以及滑动参数相对较小时, Wald 统计量的统计检验力会相对较大。比较表 6 与表 7 同样可以发现, 在非一致性 DIF 条件下, 采用观察信息矩阵计算的 Wald 统计量的统计检验力均高于采用经验交叉相乘信息矩阵而计算获得的 Wald 统计量的统计检验力。

5　讨论

认知诊断模型能够提够关于受测者属性掌握模式的较为详尽的诊断性信息, 它不仅能为老师的教以及学生的学提供有针对性的建议, 而且也有助于教育者深入理解受测者的认知心理。在使用这一模型来解释受测者的作答之前, 研究者需要确定认知诊断测验项目的参数对于所有受测者都是不变的, 否则会对受测者的属性掌握模式的估计带来不良的影响(王卓然等, 2015), 进而导致错误的诊断性信息。DIF 检验可以用以确认不同组的受测者在同一个项目的作答上是否存在差异, 即除了属性掌握模式外, 受测者所在的组会影响到他们对于项目的反应。为保证测验的效度, 在使用认知诊断模型来拟合受测者的作答数据前, 需要进行 DIF 检验。先前研究者发现 Wald 统计量在检验 DIF 时, 有着许多其他统计量所不具备的优点, 然而前人研究中对于 Wald 统计量在检验 DIF 时的一类错误率的表

现, 存在明显的结论冲突。如, Hou 等人(2014)以及王卓然等人(2014)的模拟研究发现 Wald 统计量会存在一类错误控制率膨胀的问题, Li 和 Wang (2015)的模拟研究却发现, 其研究中所用的 LCDM-DIF 以及 Wald 统计量在使用 MCMC 计算时有着良好的一类错误控制率。本研究采用 Hou 等人以及 Li 等人研究中所使用的同等条件通过模拟发现, 这些差异主要是由于 Wald 统计量计算方法的差异引起的。因此, 我们认为本研究提出的改进的 Wald 统计量的计算方法解决了 DIF 研究中一直困扰研究者的 Wald 统计量在检验 DIF 时的一类错误率的表现不同这一重要问题, 具有重大的理论意义。

5.1 Wald 统计量在检验 DIF 时的一类错误控制率

在模型正确设定的前提下, 如果统计量能够很好的服从渐近分布, 那么, 它的一类错误控制率应该能够较好的接近预先设定好的显著性水平。本研究中所提出改进的 Wald 统计量的计算方法具有这一特征, 从结果中可以发现, 本研究的一类错误控制率均较好地接近预先设定的 0.05 这一显著性水平。因此, 我们认为在 Hou 等人(2014)以及王卓然等人(2014)研究中所产生的 Wald 统计量一类错误膨胀的问题, 是由于不恰当的信息矩阵估计方法而引起的。本研究的这一结果明确地解释了为什么 Wald 统计量在不同研究中有不同表现的问题, 对于认知诊断模型的理论发展有一定的推动作用。另外, 相对于 MCMC 参数估计方法, MMLE/EM 具有运算量小、耗时短等优点, 本研究所提出的改进的 Wald 统计量正是基于 MMLE/EM, 因此, 本研究不仅具有重大的理论意义, 而且对于认知诊断实践也具有重要的现实意义。

5.2 Wald 统计量在检验 DIF 时的统计检验力

当确认统计量的一类错误控制率能够较好的接近预先设定的显著性水平后, 接下来所要考虑的另外一个重要问题是当认知诊断测验中的项目中存在 DIF 时, 这一统计量能否有效地拒绝不存在 DIF 的原假设而选择备择假设。通过表 4 与表 6 中的结果, 可以发现在样本量较大时(N = 1000), 改进后的 Wald 统计量在检验 DIF 时的统计检验力均明显的高于样本量比较小时(N = 500)的统计检验力。因此, 本研究建议在应用 Wald 统计量进行 DIF 检验的时候, 如果想要达到较高的统计检验力, 应保证较大的样本量。因为 Hou 等人(2014)发现, 其研究中所采用的 Wald 统计量计算方式, 会导致一类错误率膨胀, 因此, 在计算统计检验力的时候, 她们采用了两种方式进行。第一种方式是直接用 Wald 统计量的理论分布即自由度为 2 的卡方分布的理论值来计算, 由于其开发的 Wald 统计量的计算方式的一类错误率膨胀会使得原本不存在 DIF 的项目被误判为存在 DIF, 因此计算结果不够可靠; 她们所采用的第二种方式是计算当不存在 DIF 项目时 Wald 统计量在每种实验条件组合下 10,000 次模拟的经验分布, 然后通过获得的显著性水平的临界值, 来计算 Wald 统计量的统计检验力, 这种计算方式虽然保证了模拟实验结果具有较高的可靠性, 但是不具备现实的可操作性, 因此, 对于其研究目的而言只能算是一种不完整解决的方案。因为 Hou 等人(2014)的第二种计算方式具有较高的理论上的结果可靠性, 因此可以作为研究结果的一个参考。通过研究结果对照我们发现, 本研究所采用的自由度为 2 的卡方分布理论值所计算获得的研究结果与 Hou 等人(2014)的第二种计算方式所获结果具有很高的一致性, 这也能够间接的表明, 本研究所使用的改进后的 Wald 统计量计算方式具有准确性及可靠性的特点。

5.3 以后的研究方向

由于本研究关注的重点在于, 在 EM 算法框架下提出一个恰当的 Wald 统计量的计算方式, 用以准确有效地来检验认知诊断测验中可能存在的 DIF 项目, 澄清以往研究中所用 de la Torre (2009, 2011) 所提出的信息矩阵方法计算 Wald 统计量时所产生的令人困惑的结果。因此, 本研究仅采用了 Hou 等人(2014)的研究设计, 通过结果对比的方式来证明本研究所提出的改进的 Wald 统计量在检验 DIF 时具有准确性可靠性等特点。具体而言, 研究者可以就以下几方面进行后续研究: 首先, 样本大小对于 Wald 统计量有重要影响, 因此, 后续研究中可以使用本研究中所用 Wald 统计量考察这一因素对于 DIF 的影响; 其次, 目前的研究中普遍采用 DINA 或者是高阶 DINA 作为例证模型, 本研究出于结果比较的因素考虑, 也是以 DINA 模型为例, 在其他认知诊断模型中 Wald 统计量用以检验 DIF 时的表现, 也是一个非常有意思的研究方向。由于本研究所采用的是对于 LCDM 模型进行约束而获得的 DINA 模型, 因此, 可以很方便的进行扩展; 第三, 本研究所采用的项目数量为 30, 且受测者组的数量为 2, 在不同项目数量下以及不同的受测者组数量数下, Wald 统计量的表现也值得研究者关注; 第四, 在认知诊断模型中, 除了 Wald 统计量可以进行 DIF 检验之外, 还有一些其他的统计量也可以进行 DIF 检验(Li, 2008;

Sünbül & Sünbül, 2015, July), 虽然目前研究表明, Wald 统计量在检验 DIF 时, 具有一些其他统计量所不具有的优点, 但是, 在另外的应用情景中, 这些 DIF 检验方法的优缺点, 仍然值得研究者的关注。

6 结论

本研究中所提出的改进的 Wald 统计量的计算方法, 在认知诊断测验中不存在 DIF 项目时, 有着良好的一类错误控制率, 能够较为准确地接近预先设定的显著性水平, 即当认知诊断模型为 DINA 时, 改进的 Wald 统计量服从自由度为 2 的卡方分布; 在认知诊断测验中存在 DIF 时, 改进的 Wald 统计量能够准确有效的鉴别出存在 DIF 的项目。本研究同样发现样本量对于 Wald 统计量的一类错误控制率及统计检验力存在重要影响。另外, 我们建议认知诊断模型的研究者与使用者, 当采用 EM 算法进行参数估计时, 在确认认知诊断模型正确设定后, 使用本研究中所使用观察信息矩阵的方法计算项目参数的标准误。

参 考 文 献

Bradley J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

Chen, J. S., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*, 419–437.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

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

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Greeno, J. G. (1980). Trends in the theory of knowledge for problem solving. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 9–23). Hillsdale, NJ: Erlbaum.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctorial dissertation). Department of Statistics, University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum.

Hou, L. K., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis modeling: Applying Wald test to investigate DIF for DINA model. *Journal of Educational Measurement, 51*, 98–125.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kenward, M. G. & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13*, 236–247.

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

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

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226–233.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Paek, I., & Cai, L. (2013). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement, 74*, 58–76.

R Core Team (2015). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. Retrieved July 2, 2015, from http://www.R-project.org

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

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Sünbül, Ö., & Sünbül, S. Ö. (2015, July). *Evaluating performance of differential item functioning detection methods for DIF data in DINA model*. Paper presented at the meeting of the annual meeting of the International Meeting of the Psychometric Society, Beijing, China.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.

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

Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement, 73*, 412–439.

von Davier, M. (2005). *A general diagnostic model applied to language testing data (ETS Research Report RR-05-16)*. Princeton: Educational Testing Service.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives, 7*, 67–74.

Wang, Z. R., Bian, Y. F., & Guo, L. (2015). The impact of DIF

on estimating accuracy of cognitive diagnostic test. *Psychological Exploration, 35*, 272–278.
[王卓然, 边玉芳, 郭磊. (2015). 项目功能差异对于认知诊断测验估计准确性的影响. *心理学探新, 35*, 272–278.]
Wang, Z. R., Guo, L., & Bian, Y. F. (2014). Comparison of DIF detecting methods in cognitive diagnostic test. *Acta Psychologica Sinica, 46*, 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*, 1923–1932.]
Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Unpublished doctorial dissertation). University of North Carolina at Greensboro.

An improved method for differential item functioning detection in cognitive diagnosis models: An application of Wald statistic based on observed information matrix

LIU Yanlou[1]; XIN Tao[1,2]; LI Lingqing[3]; TIAN Wei[2]; LIU Xiaoxiao[1]

([1] *Institute of Developmental Psychology, Beijing Normal University, Beijing* 100875, *China*)

([2] *National Innovation Center for Assessment of Basic Education Quality, Beijing* 100875, *China*)

([3] *School of Teacher Education, Taishan University, Taian* 271000, *China*)

Abstract

In cognitive diagnostic models (CDMs), differential item functioning (DIF) refers to the probabilities of success of an item being different for examinees with the same attribute mastery pattern in the groups. The detection of DIF is an important step to ensure the fairness and validity of results from CDMs for all groups. Hou et al. (2014) proposed that the Wald statistic can be used to detect DIF in CDMs. Unfortunately, their results revealed that the Wald statistic based on the information matrix estimation method developed by de la Torre (2009, 2011) yielded inflated Type I error rates. However, Li and Wang (2015) found that the Type I error rates of the Wald statistic in which MCMC algorithms were implemented were slightly inflated in their study under the same conditions. In this study, we proposed an improved Wald statistic based on the observed information matrix for DIF assessment. As a general demonstration, we took the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) as an example.

In this simulation study, in order to compare the results with previous studies (e.g., Hou et al.,2014; Li & Wang, 2015), we followed the simulation design used by Hou et al. (2014), except that we implemented the observed or cross-product (XPD) information matrix in the Wald statistic computation. Parameters set in the studies were: the test length at 30, the number of attributes at 5, and the maximum number of required attributes for an item at 3. Binary item response data were generated from the DINA model. Three sets of true item parameter values were considered $(g_j = s_j = .1, .2, \text{or} .3)$ for the reference group. Two DIF sizes: .05 and .10, and two types of DIF: uniform and nonuniform, were manipulated. Two sample sizes were considered, 500 and 1,000. Each condition was replicated 1000 times, and the estimation code was written in R (R Core Team, 2015).

The simulation results showed that: (1) for the relatively discriminating items, Wald statistic had accurate Type I error control when the observed information matrix was used in its computation. However, when the slip and guessing parameters were large $(s_j = g_j = 0.3)$, the Type I error control was slightly conservative. (2) When the XPD information matrix was used for the computation of the Wald statistic, the Type I error control was conservative; that is, the performance of the observed information matrix was better than the XPD information matrix. (3) The number of attributes required for success on the item did not have a notable impact on the Type I error control of Wald statistic, irrespective of whether the observed or the XPD information matrix was used for the statistic. (4) The power rates of Wald statistic for detecting DIF increased as the sample size increased.

We conclude that our improved Wald statistic provided follows asymptotically a chi-square distribution with degrees of freedom equal to 2, for DINA model. The improved Wald statistic is a useful and powerful tool for DIF detection in CDMs.

Key words Wald statistic; differential item functioning; cognitive diagnosis model; observed information matrix; cross-product information matrix