心理科学进展 2014, Vol. 22, No. 8, 1350–1362
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2014.01350

·研究方法(Research Method)·

项目反应理论中模型–资料拟合检验常用统计量*

单昕彤 谭辉晔 刘 永 吴方文 涂冬波

(江西师范大学心理学院, 南昌 330027)

摘 要 项目反应理论(IRT)模型依据项目与被试的特征预测被试的作答表现, 是常用的心理测量模型。但IRT的有效运用依赖于所选用IRT模型与实际数据资料相符合的程度(即模型–资料拟合度, goodness of fit)。只有当所采用IRT分析模型与实际数据资料拟合较好时, IRT的优点和功能才能真正发挥出来(Orlando & Thissen, 2000)。而当所采用IRT模型与资料不拟合或选择了错误的模型, 则会导致如参数估计、测验等值及项目功能差异分析等具有较大误差(Kang, Cohen & Sung, 2009), 给实际工作带来不良影响。因此, 在使用IRT分析时, 应首先充分考察及检验所选用模型与实际数据是否相匹配/相拟合(McKinley & Mills, 1985)。IRT领域中常用模型–资料拟合检验统计量可从项目拟合、测验拟合两个角度进行阐述并比较, 这是心理、教育测量领域的重要主题, 也是测验分析过程中较易忽视的环节, 目前还未见此类公开发表的文章。未来的研究可以在各统计量的实证比较研究以及在认知诊断领域的拓展方面有所发展。

关键词 项目反应理论; 模型–资料拟合检验; 项目拟合; 测验拟合

分类号 B841

1 引言

项目反应理论(IRT)模型是教育与心理测量领域常用的统计模型(Hambleton, Swaminathan & Rogers, 1991)。它提供了一个解决各种测量问题的有效框架：题库建设, 测验构建, 测验等值, 计算机自适应测验等等(Lu, 2006)。相比经典测验理论(CTT), IRT模型提供了许多优势(Lu, 2006), 比如项目与能力参数的不变性, 但获得这些优势的程度取决于所选模型本身的适当程度(Stone & Zhang, 2003)。为了有效应用IRT模型, 常常要做模型–资料拟合(model-data fit)检验(Lu, 2006)。评价模型–资料拟合一般既要评价测验拟合(test fit), 又要评价项目拟合(item fit)。因为测验拟合检验评估的是测验整体水平上选用的模型与实际资料是否相吻合; 而项目拟合检验则在项目水平上评估模型与实际资料是否相吻合, 可用于筛选测验中的个别项目。

项目拟合检验一般是绝对拟合检验, 此类方法通常考察某个特定的模型与资料的拟合情况, 计算待选用模型与实际资料的某种统计量, 通过对统计量进行假设检验从而评价模型–资料的拟合。如基于卡方检验的方法, 基于预测分布的方法; 测验拟合检验一般是基于模型比较的相对拟合检验, 此类方法通常是对两个或多个模型分别进行模型–资料拟合, 计算相关统计量, 并对这几个模型的统计量进行比较, 选择一个相对与资料拟合较优的模型。如AIC (Akaike's information criterion; Akaike, 1974), BIC (Bayesian information criterion; Schwarz, 1978), DIC (deviance information criterion; Spiegelhalter, Best & Carlin, 1998)。一般情况下, 可以先采用相对拟合检验将备选模型缩减至一个或少数几个, 再采用绝对拟合检验具体考察待选模型是否能够较好地拟合数据资料(宋

收稿日期：2013-12-07

* 国家自然科学基金(31100756, 31300876, 31360237, 31160203), 教育部人文社科项目(11YJC190002), 江西省社科规划重点项目(13JY01), 高等院校博士点基金项目(编号：20123604120001), 江西教育科学规划(12YB088, 13YB029), 江西省教育厅科技计划项目(编号: GJJ13266)和江西师范大学青年英才培育资助计划等资助。

通讯作者：涂冬波, E-mail: tudongbo@aliyun.com.cn

丽红, 2012)。

2　模型–资料项目拟合检验统计量

当前, IRT 领域项目拟合检验统计量主要包括卡方类统计量、后验预测模型检查统计量和基于信息矩阵检验的统计量。

2.1　卡方类统计量

基于卡方检验的拟合统计量的基本思想是, 在选用模型为恰当模型的假设下, 按照一定的标准(如被试能力值或测验观察总分)将被试分为若干类, 并计算各类中项目得某分的观察频数与期望频数的差异量, 求取卡方值。由于这些统计量服从或近似服从卡方分布, 因此对其作卡方检验来进行模型–资料拟合评价(宋丽红, 2012)。

2.1.1　χ^2、G^2

皮尔逊 χ^2 和似然比 G^2 统计量是当前 IRT 领域使用最普遍的模型–资料拟合检验统计量, 两者都以卡方分析为指导, 有着共同的特征。其基本方法是, 首先根据估计能力值将能力量尺划分为若干个组, 然后计算每个能力组在每个题目上得特定分数的被试的观察频率和期望频率, 最后计算拟合统计量并判断其显著性(Lu, 2006)。

以二值计分项目为例, 皮尔逊 χ^2 统计量有如下形式：

$$\chi^2=\sum_{j=1}^{J}N_j\frac{(O_j-E_j)^2}{E_j(1-E_j)} \quad (1)$$

j 为能力量尺中共 J 个能力组中的第 j 个组; O_j 为第 j 组被试答对某项目的观察频率, E_j 为期望频率, N_j 是第 j 个能力组中的被试人数, 自由度等于 J-m, m 是 IRT 模型中的项目参数个数。

Bock 于 1972 年提出的 χ^2 统计量和 Yen 于 1981 年提出的 Yen 统计量(通常指 Q_1)是皮尔逊 χ^2 统计量较早的变体(Lu, 2006)。Bock 将能力量尺划分为 J 个组, 其中各组被试数基本一致, E_j 是通过类 j 中能力估计值的中位数计算而来的。Yen 则将能力量尺划分为 10 组, E_j 是通过类 j 中所有被试的能力值计算而来的, $E_j=\frac{1}{N_j}\sum_{k\in j}^{N_j}P(\theta_k)$, $P(\theta_k)$ 表示第 k 个被试基于模型的期望答对概率, 被试 k 属于第 j 组, N_j 是第 j 个能力组中的被试人数(Yen, 1981)。

Mckinley 和 Mills 于 1985 年提出似然比 G^2 统计量, 其计算与 Q_1 相似。将能力量尺划分为 10 个大小相等的组, 计算每组在每个题目上得特定分数的被试的观察频率和期望频率, 对于二值计分项目而言, G^2 如下所示：

$$G^2=2\sum_{j=1}^{J}N_j[O_j\ln\frac{O_j}{E_j}+(1-O_j)\ln\frac{1-O_j}{1-E_j}] \quad (2)$$

传统的卡方统计量的自由度虽然一般常用 J-m (二值计分的项目), 但其确定仍有歧义。有学者认为自由度可能在 J-m 到 J 之间(Lu, 2006); 另外, Lu (2006)的博士论文中提到, m 可能不止包含项目参数, 还应包含能力参数。因为, 传统卡方统计量根据被试能力分组, 但能力估计的不准确性(即误差)会导致每个能力分组内的观察频数不准确(Orlando & Thissen, 2003), 进而使拟合统计量的分布及其自由度不准确(Orlando & Thissen, 2000)。以下提出的卡方类统计量都在一定程度上改善了传统卡方统计量的分布及其自由度不准确的问题。比如 χ^{2*}、G^{2*}通过模拟产生样本分布并观察统计量在其中的位置以确定拟合检验的显著性水平; S-χ^2 和 S-G^2 中观察频率的计算是基于观察分数直接可得的, 解决了观察频率依赖能力的估计值的问题。

2.1.2　χ^{2*}、G^{2*}

考虑到依据能力分组产生的分组内观察频数的不准确性, 有些学者提出 χ^{2*}、G^{2*}统计量(Stone et al., 2003)。基本方法是利用能力估计的后验分布获得每个能力分组中观察频数的伪数 r_{jk} (或伪观察分数分布 pseudo-observed score distribution; Stone, 2000), 在此基础上计算期望频数, 具体如下：

用若干个积分结点(quadrature points)划分能力量尺, 在具体的能力区间上, 对于一个项目的作答反应为 j 分的频数为：

$$r_{jk}=\sum_{N}x_{jn}P(x_n\mid X_k)A(X_k)/P(x_n) \quad (3)$$

N 是被试数; x_{jn} 在被试 n 得 j 分时为 1, 否则为 0; k 是第 k 个结点; X_k 是第 k 个结点的能力值; r_{jk} 是能力为 X_k 且得 j 分的被试人数; $P(x_n|X_k)$是第 n 个被试在一组项目上反应模式为 x_n 的条件概率; $A(X_k)$ 是第 k 个结点相对应的权重; $P(x_n)\cong\sum_k P(x_n\mid X_k)A(X_k)$。

将 r_{jk} 在分数 j 上加和, 就得到能力取 X_k 时的后验预期人数 $r_{.k}$。据已估计出的项目参数值, 可求得能力为 X_k 的被试得 j 分的概率, 用其乘以 $r_{.k}$,

即得 e_{jk}, 指能力为 X_k, 得分为 j 的基于模型的预期人数, 即期望频数(Stone, 2000)。

χ^{2*}与 G^{2*}的伪数不是划分真实的被试得来, 而是将能力区间按结点划分, 计算后验概率而来, 与能力的先验分布和模型参数有关, 那么不同分组间的观察数据共同受到能力的先验分布及模型参数的影响, 其值并不相互独立, 且根据概率划分被试, 那么同一被试就可能不止被划分在一个能力区间内, 这就导致统计量不服从卡方分布(Stone, 2000), 因此, 可以使用蒙特卡洛重抽样方法根据某一具体的总体产生一个样本分布, 以此作为 χ^{2*}、G^{2*}的检验标准, 判断其显著性水平。

该方法的优点是解决了项目拟合检验时被试能力估计不准确的问题, 观察被试的频数是由后验概率计算的伪数, 而非根据能力的估计值划分实际的被试得来。但其缺点有二：第一, 由于一些能力水平分组下的期望频数可能为 0 或者非常小, 因此有必要在能力量尺一定的范围内计算该统计量(e.g., -2<θ<2; Stone, 2000); 第二, χ^{2*}、G^{2*}仍然是基于被试能力量尺来分组的, 结点的选取仍存在一定的随意性。

2.1.3 *S*-χ^2、*S*-G^2

Orlando 等人于 2000 年在其论文中指出皮尔逊 χ^2 和似然比 G^2 统计量依赖于被试能力组的划分, 这种划分具有随意性, 不同的划分产生不同的观察频率, 以至可能得出截然不同的检验结果, 而一个好的统计量应该是根据作答数据而不是 θ 估计值来划分被试。由此他们提出了 *S*-χ^2 和 *S*-G^2 统计量, 对于二值计分的项目, 其公式为：

$$S-\chi_i^2=\sum_{h=1}^{n-1}N_h\frac{(O_{ih}-E_{ih})^2}{E_{ih}(1-E_{ih})} \quad (4)$$

$$S-G_i^2=2\sum_{h}^{n-1}N_h[O_{ih}\ln\frac{O_{ih}}{E_{ih}}+(1-O_{ih})\ln\frac{1-O_{ih}}{1-E_{ih}}] \quad (5)$$

公式 4 和 5 中, n 是测验的项目总数; 观察频率 O_{ih} 是通过观察数据计算的; E_{ih} 是测验得分恰为 h 分的被试答对项目 i 的期望频率, 因为计算 *S*-χ^2、*S*-G^2 时, 分母部分 E_{ih} 不能为 0 或 1, 而当 h 为满分时, 答对第 i 题的概率为 1, h 为 0 分时答对第 i 题的概率为 0, 因此, $h=1,2,\cdots n-1$。E_{ih} 的计算方法如下：

$$E_{ih}=\frac{\int P_i S_{h-1}^{*i}\varphi(\theta)\partial\theta}{\int S_h\varphi(\theta)\partial\theta} \quad (6)$$

E_{ih} 表示测验得分为 h 分的被试中第 i 题答对的人数占总人数的比例。公式(6)的分母包括第 i 题答错且得 h 分的被试和第 i 题答对且得 h 分的被试; 而分子只包括第 i 题答对且得 h 分的被试, 也就是说在除第 i 题的其余题目上共得 h-1 分。S_h 是能力为 θ 的被试恰得 h 分的似然, S_{h-1}^{*i} 是当第 i 个项目被移除时, 恰得 h-1 分的似然, P_i 是被试在项目 i 上的正确作答的概率, $\phi(\theta)$ 是 θ 的总体分布。S_h 和 S_{h-1}^{*i} 可用递归算法(recursive algorithm)计算, 即先估计仅有一个项目的似然, 再估计增加一个项目后的似然, 直至包括了所有的项目(Orlando et al., 2000)。例如, 对于第 i 个项目, 恰得 h 分的似然为：

$$S_h=P_iS_{h-1}^{*}+(1-P_i)S_h^{*} \quad (7)$$

S_h^*是被试在前面 i-1 个项目上已得 h 分的似然, S_{h-1}^*是被试在前面 i-1 个项目上得 h-1 分的似然。其计算的具体步骤可详见 Orlando 等人(2000)的文献。

Orlando 等人(2000)的研究表明, 在测验长度和样本量增加时, *S*-χ^2 的表现比 Q_1 更好; 而对于短测验(10 个项目) *S*-χ^2 的表现也比 Q_1 好得多(Orlando et al., 2003)。

该方法的优点是：(1)具有不同得分的被试在某项目上正确作答的观察频率是观察可得的, 分组是依据测验得分而非依据能力值; (2)期望频率是用似然来计算而不是简单的用 p 来代替; (3)该方法可以拓广到多值计分模型(Kang & Chen, 2008)。该方法的缺点是在拟合评价中由于对 θ 进行积分而导致了一些信息的丢失(Orlando et al., 2003)。

2.1.4 校正**(adjusted)**的 χ^2/df

与 χ^{2*}、G^{2*}, *S*-χ^2、*S*-G^2 类似, Drasgow, Levine, Tsien, Williams 和 Mead 于 1995 年也对基于能力分组的缺陷提出了改善, 开发出校正(adjusted)的 χ^2/df 统计量。其计算不要求划分能力区间, 期望频数是依据能力的假设分布获得的(LaHuis, Clark & O'Brien, 2011)。由于卡方统计量对样本大小很敏感, Drasgow,等人(1995)将其进行校准后得到校正的 χ^2/df (以下记为 $Adj\chi^2/df$)：

$$Adj\chi^2/df=[\frac{3000(\chi_i^2-df)}{N}+df]/df \quad (8)$$

df 表示自由度, N 表示样本量。该统计量可能

不服从卡方分布, 因此 Drasgow 等人(1995)建议用 3 作为临界值, 当 $Adj\chi^2/df$ 的值小于 3 时, 就判断模型–资料拟合。鉴于卡方统计量对样本大小的敏感, 上式中 $\frac{3000(\chi_i^2-df)}{N}+df$ 部分是 Drasgow 等人(1995)将由观察数据计算而得的 χ_i^2 调整为样本大小固定为 3000 的新的卡方值。χ_i^2 的计算与皮尔逊 χ^2 类似：

$$\chi_i^2=\sum\nolimits_k\frac{[O_{ik}-E_{ik}]^2}{E_{ik}} \tag{9}$$

O_{ik} 表示项目 i 上得 k 分的观察频数; E_{ik} 表示相应的期望频数, 由估计的项目参数和能力的假设分布计算而来, 其表达式如下：

$$E_{ik}=N\int P(k|\theta)f(\theta)d\theta \tag{10}$$

$P(k|\theta)$ 表示能力为 θ 的被试在项目 i 上得 k 分的概率; $f(\theta)$ 表示 θ 的概率密度函数, 能力的分布一般假设为标准正态分布(N (0,1))。

$Adj\chi^2/df$ 具体有三种形式：单个项目的 $Adj\chi^2/df$、项目对的 $Adj\chi^2/df$ 以及三个项目的 $Adj\chi^2/df$。由于单个项目的 $Adj\chi^2/df$ (记为 singles χ^2)统计量是针对一个项目的边际统计量, 因此不能检验项目局部独立性(LI)假设, Drasgow 等人(1995)提出了针对项目对以及三个项目的 $Adj\chi^2/df$ (分别记为 doubles χ^2 以及 triples χ^2)统计量, 方法是, 根据项目难度将测验分成三组, 在组内及组间分别组合项目, 以计算不同的 doubles χ^2 和 triples χ^2。

以上三种形式又可分别在校准样本(calibration sample)和交叉验证样本(cross-validation sample; Ounpraseuth, Lensing, Spencer & Kodell, 2012)下求得。Drasgow 等人(1995)提出用交叉验证样本(以下简称 CV)代替校准样本来计算卡方值, 其目的是为了确保模型对数据不会过分严格的拟合(overfit), 即数据稍有变化, 原本拟合的模型就不再拟合了。

Tay 和 Drasgow (2012)的研究表明, 一般情况下, 对于 IRT 中的逻辑斯蒂克(logistic)模型, doubles χ^2 和 triples χ^2 的检验力随着样本量和测验长度的增加而增大, triples χ^2 表现最好, doubles χ^2 次之, singles χ^2 最差; 当样本量足够大时(e.g., 5,000; Tay et al., 2012), CV 的 I 型错误较校准样本小, 检验力(adjusted power)更高; 当 2PLM 去拟合 3PL 数据时, $Adj\chi^2/df$ 的检验力很低, 统计量的值常低于 3; 对于 2PL、3PL 数据多维性的检验, $Adj\chi^2/df$ 的检验力很低, 大部分值常低于 3。后两点说明简单的固定临界值 3 并不通用。

$Adj\chi^2/df$ 统计量的优点是解决了能力分组导致的一些缺陷, 调整了卡方统计量对大样本量的敏感。缺点是：(1)与 S-χ^2、S-G^2 统计量类似, $Adj\chi^2/df$ 统计量由于对 θ 进行积分可能会导致一些信息的丢失; (2)对于同一种统计量的三种变体, singles χ^2、doubles χ^2 以及 triples χ^2 统计量, 其判断不拟合的临界值或许不应都是同一个值; (3)对于不同的检验情形(如用 2PLM 拟合多维数据), 临界值 3 可能并不通用。

2.1.5　参数靴绊(parameter bootstrap)法修正的 $Adj\chi^2/df$

Tay 等人 2012 年发表的论文中指出 $Adj\chi^2/df$ 的临界值 3(Drasgow et al., 1995)并不通用, 因此, 他们提出使用参数靴绊法修正 $Adj\chi^2/df$ 来评价模型–资料拟合, 基本程序如下：

(a)用一种二值计分 IRT 模型(1PLM 或 2PLM 或 3PLM)模拟数据, 此为观察数据;

(b)假设其符合 1PLM 或 2PLM 或 3PLM, 用(a)步中的观察数据估计相应假设模型的项目参数, 求估计的模型与(a)步中的观察数据的拟合统计量, 得到观察校正卡方比率, 记为 $(Adj\chi^2/df)_o$, 下脚标字母 O 代表“观察的”。

(c)用(b)步中确定项目参数的模型模拟数据, 用此数据和(b)步的模型做模型–资料拟合检验, 得到 $(Adj\chi^2/df)_r$;

(d)将(c)步重复 m 次获得 $Adj\chi^2/df$ 的样本分布, 确定 $(Adj\chi^2/df)_o$ 在此分布中的相对位置以及样本分布中大于 $(Adj\chi^2/df)_o$ 的值所占的概率 p;

(e)将(a)至(d)步重复 r 次得到统计检验力和 I 型错误概率。这里, 检验力为严格模型与数据作拟合检验时, r 次中 p 小于 0.05 的概率; I 型错误概率为恰当的模型或更不严格的模型与数据作拟合检验时, r 次中 p 小于 0.05 的概率。

Tay 等人(2012)的研究表明：第一, Logistic 模型–资料拟合时, 参数靴绊法修正而来的 $Adj\chi^2/df$ 的表现比原来的要好, 对小样本情况的改善尤甚。但当样本量为 500 时, 表现仍然较差; 第二, 相比 $Adj\chi^2/df$, 新方法在 2PLM 和 3PLM 情况下, 对于数据多维性检验具有适当的检验力。

2.1.2 到 2.1.5 都是对传统卡方统计量(皮尔逊

χ^2 和似然比 G^2 统计量)的改进，且大多是对数据分组的改善。χ^{2*}、G^{2*}不用估计能力，而是将能力的先验分布按结点划分，用贝叶斯后验计算观察频数的伪数，由此计算期望频数；S-χ^2、S-G^2 利用测验得分来分组，用联合似然函数计算期望频率；$Adj\chi^2/df$ 则不对数据进行能力或总分分组，期望频率是基于项目参数和能力的假设分布计算的。

基于卡方检验的模型–资料拟合检验统计量——包括皮尔逊 χ^2 和似然比 G^2 统计量，χ^{2*} 和 G^{2*} 统计量，S-χ^2 和 S-G^2 统计量、$Adj\chi^2/df$ 和参数靴绊法修正的 $Adj\chi^2/df$ 统计量—— 都存在以下问题：(1)统计量 2.1.1 到 2.1.3 都必须将被试分组，这就要求有足够大的样本量，使得每个类中的理论频数不少于 5，否则，计算的统计量可能会违背卡方分布，导致检验结果不可靠；(2)卡方类统计量都须计算理论频数，但即使是大样本也难以保证理论频数不过少；(3)卡方检验对样本量很敏感，当样本量大时，卡方统计量总是倾向拒绝虚无假设，由此得到模型–数据拟合不好的结果(宋丽红, 2012)。不过，以上提到的几个统计量中，后两个统计量对于卡方检验对样本量敏感的问题做了一定的调整。

2.2 后验预测模型检查方法

Sinharay, Johnson 和 Stern (2006)指出卡方统计量只是对不拟合差异的一个总的描述，除了 doubles χ^2 和 triples χ^2 统计量能够检验项目局部独立性(LI)假设外，卡方统计量并不能检测出导致模型–资料不拟合的具体原因，如模型假设所有项目等区分度，但实际数据并非如此。而研究者却常常希望通过检验模型是否能够解释数据的各个方面(如测验是否等区分度)来评价模型–资料拟合。后验预测模型检查(posterior predictive model checking [PPMC])法就是其中之一。

PPMC 方法是贝叶斯模型常用的模型–资料拟合检验方法。该方法的主要思想是比较观察数据与模型预测数据(replicated data)在差异度量(discrepancy measures; Gelman, Meng & Stern, 1996)上的差异大小，该差异度量(如优势比)要求对模型不拟合情况是灵敏的(宋丽红, 2012)。如果观察数据的差异度量的分布与预测数据的差异度量的分布之间存在较大的差异，则显示该模型不能解释该数据，也就是说模型–资料不拟合。因为预测数据是根据拟合模型产生的，若拟合模型与观察数据拟合，则预测数据与观察数据应存在某些相似性(如满足 LI 假设，这种具体方面的检验通过不同的差异度量统计量描述)。PPMC 法主要通过绘图表示观察数据与预测数据的差异度量的分布情况，并且采用后验预测概率值(posterior predictive p-value)，即 PPP 值作为量化指标。PPP 值即计算预测数据的差异度量大于观察数据的差异度量的概率。当 PPP 值接近 0 或 1 时，说明模型不拟合，而较不极端的 PPP 值，如小于 0.05 或大于 0.95 的值，也预示着模型不拟合。PPMC 法的基本过程如下(Sinharay et al., 2006)：

(a)根据观察数据产生 N 批参数(蒙特卡洛马尔科夫链法)。

(b)根据(a)中生成的 N 批参数模拟 N 批预测数据。

(c)计算观察数据以及预测数据的差异度量。

(d)根据观察数据与预测数据的差异度量绘图并计算 PPP 值。

其中，$D(y,\omega)$ 表示观察数据的差异度量，$D(y^{rep},\omega)$ 表示预测数据的差异度量。是模型中的所有参数，包括项目参数和能力参数；y 是观察数据(实证研究中是观察数据；模拟研究中是用给定的 ω 模拟的作答数据)；$p(\omega|y)$ 代表模型参数 ω 基于观察数据 y 的后验分布，$p(\omega|y)\propto p(y|\omega)p(\omega)$，$p(y|\omega)$ 是基于观察数据的似然函数，其中 ω 是根据 y 估计的，$p(\omega)$ 是参数的先验分布。在 PPMC 方法中可以用蒙特卡洛马尔科夫链(Markov Chain Monte Carlo [MCMC])算法生成 $p(\omega|y)$ (Gelman et al., 1996; Sinharay et al., 2006)；y^{rep} 是假设要检验的模型模拟的后验预测数据；$P(y^{rep}|\omega)$ 中 ω 的是从模型参数的后验分布中抽取的值；$P(y^{rep}|y)$代表预测数据的后验预测分布。后验预测数据的生成可用以下公式表示：

$$p(y^{rep}|y)=\int p(y^{rep}|\omega)p(\omega|y)d\omega \qquad (11)$$

Sinharay 等人(2006)认为可用以下三个统计量具体描述差异度量："观察分数的分布"(observed score distribution)可以描述能力的实际分布与假设分布的拟合情况(属于测验整体拟合统计量，详见文章第 3 部分)；"二列相关系数"(biserial correlation coefficient)能够检验模型是否缺少数据所必需的区分度参数；"项目对的关联度

量” (measures of association among the item pairs)用于检验 LI 假设。

2.2.1　二列相关系数

常用于检验模型是否缺少数据所必需的区分度参数(slope parameter; Sinharay et al., 2006), 如用 Rasch 模型去估计项目区分度参数不相等的测验数据。

二列相关系数计算的是个人总分与二值计分项目的作答结果之间的相关。若实际的观察数据的项目间区分度参数不等, 则个人总分与二值计分项目的作答之间的相关不止由题目的难度决定, 还受区分度的影响, 因此该统计量能描述项目是否等区分度。

2.2.2　项目对的关联度量

此类统计量常用于检验 LI 假设, 主要通过下面介绍的“优势比”和“MH 统计量”来反应。

优势比(odds ratio; Chen & Thissen, 1997)：

$$OR = \frac{n_{11}n_{00}}{n_{10}n_{01}} \tag{12}$$

其中, $n_{kk'}$为在第一个项目上得 k 分, 在第二个项目上得 k’分的人数, 对于二值计分项目, k,k’=0,1。Chen 等人(1997)认为 OR 可以用来检验 IRT 局部独立性(LI)假设, 如果 LI 假设满足, 观察数据的 OR 与预测数据的 OR 基本相等。然而, LI 假设由于种种原因可能无法满足, 如速度测验或是心理测验的非单维性(e.g., Chen et al., 1997)。若 LI 假设不满足, 单维 IRT 模型下, 测量同一种特质的项目的观察的 OR 大于预测的 OR; 测量不同特质的项目的观察的 OR 小于预测的 OR。

MH 统计量(Mazor, Clauser & Hambleton, 1992)：

$$MH = \frac{\sum_r n_{11r}n_{00r}/n_r}{\sum_r n_{10r}n_{01r}/n_r} \tag{13}$$

定义 $OR_r = \frac{n_{11r}n_{00r}}{n_{10r}n_{01r}}$, 表示基于剩余分数的项目对的优势比, 剩余分数是指除项目对外, 在其余项目上总的得分。n_r为剩余分数为 r 的人数。$N_{kk'r}$是指在测验中除项目对 k,k’之外的所有项目上的总分为 r, 而在第一个项目上得 k 分, 第二个项目上得 k’分的人数。类似 OR, 若 LI 假设不满足, 单维 IRT 模型下, 测量同一种特质的项目的观察的 MH 大于预测的 MH; 测量不同特质的项目的观察的 MH 小于预测的 MH。

PPMC 方法有几个优点：(1)不要求统计量有特定的分布形式, 可通过程序提供一个模拟的分布; (2)经典的拟合统计量(如卡方统计量)都是一个点估计, 而 PPMC 方法通过使用参数的后验分布, 从而克服了参数估计的不准确性; (3)经典的拟合检验无法切实解决复杂模型的拟合问题, 而 PPMC 方法随着贝叶斯计算的发展, 可以让研究者去拟合更富实际或更复杂的模型; (4)PPMC 方法用多种差异度量统计量来描述模型的不同方面—— 能够解释 IRT 模型不同形式间的差异(如 Rasch 模型与 2PLM 的差异), 以及其他一些不拟合的原因, 如“项目对的关联度量”可以检验 LI 假设—— 研究者可根据实际情况进行选择(Sinharay et al., 2006)。

但其也存在一些问题：应用 PPMC 方法的关键在于差异度量统计量的选择, 要检验模型–资料拟合的哪些方面, 怎样去检验至关重要, 只有模型拟合数据的各个方面, 研究者才能自信的做出推断, 而当模型在某一方面并不拟合数据时, 该如何判断尚不清楚(Sinharay et al., 2006)。

2.3　基于信息矩阵的检验

卡方类统计量无法检验特定形式的不拟合, PPMC 方法对有些形式的不拟合也无法检验, 如 ICC 单调性。所以 Ranger 和 Kuhn (2012)提出用信息矩阵检验方法评价模型–资料拟合。

信息矩阵检验由 White 于 1982 年提出, 后由 Ranger 等人(2012)引入到 IRT 模型–资料拟合检验中, 通过比较信息矩阵不同形式间的差异进行拟合检验。信息矩阵可以用负的黑塞矩阵(Hessian matrix; 即对数似然函数的二阶导矩阵)的期望表示, 也可以用分数向量的外积的期望(the expectation of the outer product of the score vector)来表示, 也就是分数向量的协方差矩阵。只有正确的选定模型, 负的黑塞矩阵与协方差矩阵才相等, 因此, 这两个矩阵间的差异程度就代表了模型–资料拟合程度。该方法通过选择不同的信息矩阵元素构建出近似服从卡方分布的统计量, 其基本步骤如下(Ranger et al., 2012)：

(a)根据选定模型估计项目参数;

(b)根据(a)中的项目参数构建在所有项目 G 上特定反应模式为 U 的边际似然函数;

$$P(U;\alpha) = \int\prod_{j=1}^{G} P(u_j \mid \theta;\alpha_j) f(\theta) d\theta \tag{14}$$

其中，$P(u_j \mid \theta;\alpha_j)$ 是能力水平为 θ，作答模式为 U 的被试在项目 j 上的答对概率；α_j 是第 j 题的项目参数，向量 $\alpha=[\alpha_1,\ldots,\alpha_G]$ 是 G 个项目的项目参数；$f(\theta)$ 是能力的先验分布。

(c)令 $f(U;\alpha)$ 为在所有项目 G 上作答模式为 U 的多项分布函数，其中项目参数 α 是由边际概率 $P(U;\alpha)$ 决定的；

(d)对 $f(U;\alpha)$ 取对数得 $L(U;\alpha)=ln[f(U;\alpha)]$，并求取其对项目参数的一阶导矩阵 $g(U;\alpha)$ 和二阶导矩阵(黑塞矩阵)$H(U;\alpha)$；

$$g(U;\alpha)=\partial L/\partial\alpha \quad (15)$$

$$H(U;\alpha)=\partial L/\partial\alpha\partial\alpha' \quad (16)$$

(e)根据一阶导和二阶导构建一个对角矩阵并求其期望，即：

$$E[D(U;\alpha)]=E[g(U;\alpha)g(U;\alpha)^t+H(U;\alpha)]=0 \quad (17)$$

上式可以转换为：

$$\frac{1}{N}\sum_{i=1}^{N}D(u_i;\hat{\alpha})=$$

$$\frac{1}{N}\sum_{i=1}^{N}[g(u_i;\hat{\alpha})g(u_i;\hat{\alpha})^t]+\frac{1}{N}\sum_{i=1}^{N}[H(u_i;\hat{\alpha})] \quad (18)$$

其中，$\hat{\alpha}$ 是项目参数的极大似然估计，N 表示被试数。令 $d(u_i;\hat{\alpha})$ 为 $D(u_i;\hat{\alpha})$ 中所有非重复元素，$r_n=\frac{1}{n\sum_n d(x_i;\hat{\alpha})}$。$\sqrt{n}r_n$ 的协方差阵如下：

$$\Sigma=E[d(U;\alpha)d(U;\alpha)^t]-E[d(U;\alpha)g(U;\alpha)^t]$$

$$(E[g(U;\alpha)g(U;\alpha)^t])^{-1}E[g(U;\alpha)d(U;\alpha)^t] \quad (19)$$

(f)构建基于信息矩阵的模型–资料拟合检验统计量：

$$T=nr_n^t\Sigma^{-1}r_n \quad (20)$$

该统计量服从自由度为 r_n 中的元素个数的近似卡方分布。

T 又可构建为 8 个分统计量(Ranger et al., 2012)，其中 4 个是项目拟合统计量(另 4 个是测验拟合统计量，详见文章第 3 部分)：

2.3.1 IM_3、IM_4

这两个统计量用于检验 ICC 单调性假设。

令 $r_n=diag[1/n\sum_n D(x_n;\hat{\alpha})]$，$T=r_n'\Sigma^{-1}r_n$。

T 统计量服从自由度为 m 的近似卡方分布，m 是对角线元素的个数，即项目参数的个数；$D(x_n;\hat{\alpha})$ 详见公式 18。

IM_3 中 r_n 是抽取信息矩阵中所要检验的某个项目的对角元素中基于截距参数 α_{0j} 的元素计算的，IM_4 中 r_n 是抽取相应的基于区分度参数 α_{1j} 的元素计算的。

2.3.2 IM_7、IM_8

IM_3、IM_4 虽对 ICC 单调性假设的检验力较好，但在 LI 假设检验的模拟研究中检验力很低，因此有人提出 IM_7、IM_8。需取矩阵中相邻的某对项目的非对角元(off-diagonal elements)来计算 r_n，进而构建 LI 假设检验统计量 IM_7、IM_8。计算 IM_7 需取信息矩阵非冗余元素的非对角元素中与截距参数 α_{0j}、α_{0j+1} 相关的元素；计算 IM_8 则相应地取与区分度参数 α_{1j}、α_{1j+1} 相关的元素。

以上 4 个统计量均服从自由度为 1 的近似卡方分布。

信息矩阵检验的优点是：第一，提供了一个评价 IRT 模型–资料拟合的灵活框架，研究者可以根据具体的情况选择合适的统计量检验某个具体的假设。这对于以往的拟合检验只能检验出总体不拟合，而不能指出不拟合的来源而言是一个很大的进步，该方法与 PPMC 方法能够检验的不拟合来源不同，二者可相互补充；第二，该方法只要求计算对数边际似然函数的一阶导和二阶导，这些量值在参数估计中经常是需要的，所以检验就变得简单易计算(Ranger et al., 2012)；第三，它既可应用于单维二值计分模型也可扩展到多维或多值计分模型(Ranger et al., 2012)。

当然，该方法也存在一些不足：第一，计算很耗时，有些特定的情况下(比如项目参数是极端值)，其结果不稳定有时甚至无法算出；第二，需要相当大的样本。不过这在 IRT 指导下有时不成问题，因为：(1)复杂的 IRT 中校准样本常常很大，比如有学者提出比较简单的 2PLM 也需要 500 个被试(Ranger et al.,2012)；(2)对于小样本检验，该方法的表现主要依赖于协方差矩阵的计算方法，相比基于样本平均数，基于期望计算协方差矩阵有了相当大的改善；(3)即使是在小样本检验中，该方法的表现仍比卡方类统计量要好(Ranger et al., 2012)。

3 模型–资料测验拟合检验统计量

虽然项目拟合统计量可以检验测验中每个项目与模型的拟合，从而指导项目的选择，但现实情况中研究者往往不知道用哪种模型去拟合资料，此时若用项目拟合统计量逐一检验可能的模型与

资料的拟合就不太现实, 并且盲目地检验测验中所有项目的拟合也不合理。

此外, 项目拟合统计量一般是绝对拟合统计量, Maydeu-Olivares 和 Cai (2006)认为, 评价模型绝对拟合存在一定困难, 因此研究者常常只评价模型相对拟合, 例如比较严格模型和宽泛模型与资料的拟合情况(Maydeu-Olivares et al., 2006); 并且, 当有几个模型都能拟合数据时, 相对拟合统计量可以帮助选择更合适的模型。相对拟合统计量的特点是没有绝对的划界点, 一般是对两个或多个模型进行比较, 从而选择一个拟合较优或最优的模型(宋丽红, 2012)。

3.1 卡方类

3.1.1 $G^2(dif)$

Maydeu-Olivares 等人 2006 的论文中提出似然比之差 $G^2(dif)$ 统计量是分类数据分析中比较两个模型的拟合程度时最常用的统计量, 其公式如下:

$$G^2(dif) = G_0^2 - G_1^2 \tag{21}$$

G_0^2 代表 M_0 模型的绝对拟合统计量; G_1^2 代表 M_1 模型的绝对拟合统计量; M_0 与 M_1 是嵌套(nested)模型, 其中, M_0 是更严格(constrained)的模型, M_1 是更宽泛的模型, 即 M_0 是 M_1 的特例(如 M_0 是 1PLM, M_1 是 2PLM)。$G^2(dif)$ 统计量在大样本下近似服从卡方分布, 其自由度是两个绝对统计量的自由度之差, 即 q_1- q_0, q_0、q_1 分别是模型 M_0 、 M_1 中的参数个数。由于 G_0^2 始终大于 G_1^2, $G^2(dif)$始终为正值, 当其值很大时(大于卡方分布指定的显著性水平的临界值), 相对拟合检验结果为拒绝虚无假设, 接受相比 M_0, M_1 能更好的拟合数据的结论。

该统计量的流行源于 Haberman (1977)的研究, 当 M_1 拟合数据时,$G^2(dif)$检验的结果在小样本和类别很多的情况下也有效。然而, 若 M_1 不拟合, 大样本下的 $G^2(dif)$不再服从近似卡方分布, 其分布的形态取决于 M_1 与数据不拟合的程度。

$G^2(dif)$统计量的优点是: (1)易于计算, 有一些现成的电脑软件可以分别计算 G_0^2 和 G_1^2 (e.g., BILOG、PARSCALE、MULTILOG; Du Toit, 2003); (2)Haberman (1977)的研究表明, 该统计量在类别很多的情况下的检验结果仍是可信的。其缺点是: (1)当最不严格的模型(M_1)与数据并不拟合时(如 M_1 是 2PLM, 产生数据的模型是 3PLM), 其统计推断判断 M_1 拟合数据并不正确, 实际上 M_0 与 M_1 均不拟合数据。犯这种误导性错误的原因在于, 这种情况下, 该统计量并不服从近似卡方分布; (2)该统计量只能考察两个模型的相对拟合度, 当实际需要比较两个以上的模型时, 该统计量便不再适用(Maydeu-Olivares et al., 2006)。

3.1.2 似然比 G^2

似然比(likelihood ratio [LR]; Mckinley et al., 1985)初用于项目绝对拟合检验, 当用于测验相对拟合检验时, 可用于嵌套模型的比较(如 1PLM、2PLM、3PLM), 其显著性检验能够帮助研究者选择最适合数据的模型。一般情况下, 当模型是嵌套的时候, 参数更多的模型比参数更少的模型更适合数据(Kang & Cohen, 2007)。

以上两个相对拟合统计量都是基于卡方思想的统计量, 因此具有卡方类统计量共有的缺点。

3.2 后验预测模型检查方法

该方法在文章第 2 部分已有描述, 其中, 观察分数的分布虽是绝对拟合统计量, 但也是测验整体拟合统计量, 因此放在此处描述。该统计量可以检验能力的实际分布与假设分布的拟合情况。

Ferrando 和 Lorenzo-seva (2001)认为, 比较观察的和预测的分数分布可以检验模型–资料拟合。Beguin 和 Glas (2001)提出

$$\chi_{NC}^2 = \sum_j \frac{[NC_j - E(NC_j)]^2}{E(NC_j)},$$

其中, NC_j 表示答对 j 个项目的观察人数, j=0,1, 2,…J; 令 $NC = (NC_0, NC_1, \cdots NC_J)'$; $E(NC_j)$ 是预期答对 j 个项目的人数。PPMC 方法通过多次模拟产生 χ_{NC}^2 的一个样本分布, 通过比较观察数据与预测数据的 χ_{NC}^2 分布, 检验能力的假设分布是否适合。

3.3 基于信息矩阵的检验

该方法虽是绝对拟合检验方法, 但其既可作项目拟合又可作测验拟合, 这是该方法的独特之处。此法在文章第 2 部分已有描述, 其中有 4 个分统计量是测验拟合统计量。

3.3.1 IM_1、IM_2

与 IM_3、IM_4 类似, 这两个统计量用于检验 ICC 单调性假设。IM_1 中 r_n 是抽取信息矩阵中所有项目的对角元素中基于截距参数的元素计算的; IM_2 中 r_n 是抽取相应的基于区分度参数的元素计

算的。这两个统计量均近似服从自由度为 $m/2$ (以 2PLM 为例)的卡方分布。

3.3.2 ***IM*$_5$、*IM*$_6$**

与 IM_7、IM_8 类似，这两个统计量用于检验 LI 假设。计算 IM_5 需取信息矩阵中所有项目的非冗余元素的非对角元素中与截距参数相关的元素，计算 IM_6 则相应地取与区分度参数相关的元素。这两个统计量均近似服从自由度为 $m/2-1$ (以 2PLM 为例)的卡方分布。

这类统计量的优缺点详见文章第 2 部分。

3.4 基于信息量的统计量

该方法是当前 IRT 领域中相对拟合检验的主要方法，用于从待选模型中选择更合适的模型，这一方面基于模型–资料拟合，一方面也要考虑模型的复杂度。有些 IRT 模型虽然互不相同但却可适用于同一类型的数据，因此有必要评价哪个模型更好或更拟合，研究者应选择能够拟合数据的最简洁模型，基于该原则选择的模型较少引起矛盾、歧义以及冗余。卡方类统计量并不能满足这一点，因此，可以选择基于信息量的统计量，如 AIC、DIC、BF、CVLL 等统计量(Kang et al., 2007)。

3.4.1 偏差

偏差(deviance, -2Log-Likelihood [-2LL])统计量是该方法最早开发的统计量，其值越小，表示拟合程度越好。其计算公式如下：

$$-2LL = -2\ln(L(X \mid \hat{\beta})) \tag{22}$$

$\hat{\beta}$ 为模型参数，X 为被试观察得分，$L(X \mid \hat{\beta})$ 为似然函数。

3.4.2 Akaike 信息量标准

AIC (Akaike, 1974)在-2LL 的基础上进一步考虑了模型参数的影响，公式如下：

$$AIC = d + 2p \tag{23}$$

AIC 有两个成分，第一个成分是 3.4.1 部分所述偏差 $d = -2LL$，对于一个模型而言，偏差 d 越小，越合适。另一个成分是 $2 \times p$，这个 p 是模型中要估计的参数个数，它是对模型参数过多的惩罚。

AIC 的优点是：在偏差统计量的基础上，考虑了模型参数的影响，参数多的模型将受到惩罚。但其也有缺点，有学者(Forster, 2004; Janssen & De Boeck, 1999)批评 AIC 没有在其计算中直接考虑样本量，所以 AIC 没有渐近一致性(asymptotically consistent)，AIC 检验的结果应是选择该值最小的模型，但在样本量很大时，AIC 却趋于选择饱和模型(saturated models)，也就是参数更多的模型。

3.4.3 贝叶斯信息量标准

BIC (Schwarz, 1978)克服了 AIC 以上的缺点，该统计量一样对复杂模型进行了惩罚，且惩罚力度比 AIC 大。该统计量将样本量 N 考虑在内，对复杂模型的惩罚利用样本量进行了加权。BIC 值越小，表示模型相对拟合越好。其计算公式如下：

$$BIC = \mathrm{d} + p \times \ln N \tag{24}$$

其中，N 为被试样本量。当样本量大时，与 AIC 所选择的模型相比，BIC 倾向于选择更简单的模型，也就是参数更少(更不饱和)的模型。

尽管 AIC 和 BIC 不能提供显著性检验，却能给出选择不同模型的相对差异，若模型参数的极大似然估计可得，AIC 和 BIC 是合适的，但若不可得，则可用贝叶斯估计代替(MCMC 方法获得)，此时可以用 DIC, PsBF, CVLL 统计量(Kang et al., 2007)。

3.4.4 偏离信息量标准

DIC (Spiegelhalter et al., 1998)适用于通过 MCMC 算法进行参数估计的模型间的比较，它与 AIC, BIC 一样，也包括模型拟合和模型复杂度惩罚两部分，其计算公式如下：

$$DIC = \overline{D(\theta)} + p_D = D(\bar{\theta}) + 2p_D = 2\overline{D(\theta)} - D(\bar{\theta}) \tag{25}$$

其中，$D(\theta) = -2\log p(y \mid \theta) + 2\log f(y)$，$p(y \mid \theta)$ 是似然函数，是观察数据 y 分布的先验；$\overline{D(\theta)}$ 是贝叶斯后验平均偏差即偏差的后验均值(the posterior mean of the deviance)；$D(\bar{\theta})$ 是贝叶斯后验均值偏差即后验均值偏差(the deviance of the posterior means)；$p_D = \overline{D(\theta)} - D(\bar{\theta})$ 是模型中有效参数个数，是模型复杂度的一个统计量(1PL, 2PL, 3PL)。DIC 可以认为是 AIC 统计量的一个推广，其值越小越好。

3.4.5 贝叶斯因子

BF (Bayes factor; Spiegelhalter & Smith, 1982)适用于两个模型的比较，通过计算模型 A 对模型 B 的后验概率(贝叶斯后验分布)的比值除以模型 A 对模型 B 的先验概率的比值来比较不同模型的拟合优度。

$$BF = posteriorodds / priorodds = \frac{P(data \mid ModelA)}{P(data \mid ModelB)} \tag{26}$$

BF 计算得到的值若大于 1 则支持模型 A，若

小于 1 则支持模型 B。

该统计量的不足是, 选择比较的两个模型中必须有一个是真实模型(true model; Bolt, Cohen, & Wollack, 2001), 也就是说必须有一个是与数据确实拟合的模型, 满足这个条件的 BF 检验才是可靠的。当两个模型中有一个是真实模型时, Schwarz (1978)认为 BIC 是 BF 的近似值(Ghosh & Samanta, 2001)。根据 Western (1999)的观点, 两个 BIC 之间的差异, 即 BIC(模型A) - BIC(模型B), 与 $-2\log(BF)$ 的值相当接近。

3.4.6 交叉验证对数似然估计

CVLL (cross-validation log-likelihood estimates; Bolt, et al., 2001; Geisser & Eddy, 1979; Gelfand & Dey, 1994)是基于贝叶斯算法的统计量, 可以同时比较两个及两个以上的模型, 适合于模型的选择, CVLL 值最大的模型最好, 其计算程序如下：

(a)从施测样本中随机抽取 1000 个被试的作答数据作为校准样本 Y_{cal};

(b)利用 Y_{cal} 将模型参数先验分布更新到后验分布, 根据需要检验的模型以及 Y_{cal} 估计模型参数;

(c)从(a)步中剩余的样本中再随机抽取 1000 个被试的作答数据作为交叉验证样本 Y_{cv}, 并计算 Y_{cv} 基于模型的似然 $P(Y_{cv} \mid Model)$;

(d)模型的 $CVLL = \ln P(Y_{cv} \mid Model)$

$$P(Y_{cv} \mid Model) = \int P(Y_{cv} \mid \theta_i, Y_{cal}, Model) f_\theta(\theta_i \mid Y_{cal}, Model) d\theta_i \quad (27)$$

上式中, $P(Y_{cv} \mid \theta_i, Y_{cal}, Model)$ 代表条件似然, 是基于模型和 Y_{cv} 计算的, 但模型中的参数是根据 Y_{cal} 估计而来的; $f_\theta(\theta_i \mid Y_{cal}, Model)$ 是能力的条件后验分布(贝叶斯后验), 是将能力的先验分布利用条件似然 $P(Y_{cv} \mid \theta_i, Y_{cal}, Model)$ 更新的。

Kang 等人(2007)的研究表明, CVLL 在模型选择上比 AIC、BIC 和 DIC 精确。

3.4.7 伪贝叶斯因子

两个模型的 CVLL 可以计算 PsBF (pseudo-Bayes factor; Bolt et al., 2001; Geisser et al., 1979; Gelfand et al., 1994), 以决定选择哪一个模型, 其公式如下：

$$PsBF = \exp(CVLL_A - CVLL_B) \quad (28)$$

当 PsBF 大于 1 时, 模型 A 更好; 当 PsBF 小于 1 时, 模型 B 更好。

基于信息量的统计量有一个缺点, 即 AIC、BIC 等多个统计量有时存在结果不一致、不精确的问题, 由于研究者并不知道实际的数据符合哪种模型, 因此在模型选择上存在一定的困难。

4 小结与展望

IRT 模型–资料拟合检验在众多学者多年的努力下已有长足的发展, 但 IRT 中关于选择哪个模型才合适的问题仍没有完全解决。其主要的困难在于, 被试真实的能力是不能获得的, 只能通过项目作答表现出来, 这就导致用作答数据计算而来的统计量与用真实的能力值计算而来的统计量之间存在一定的差异, 从而影响了模型与资料的拟合(Lu, 2006)。本文详述了各种常用的模型–资料拟合检验统计量以及各自的优缺点, 以期对模型–资料拟合检验有一个较综合的考量。绝对拟合统计量虽然主要是针对项目的, 但 χ^2、G^2, PPMC 方法, 基于信息矩阵的检验方法也可检验测验拟合; 相对拟合统计量则主要是针对测验的, 但 $G^2(dif)$ 也可检验项目拟合。卡方类统计量由于计算简便, 易于理解, 仍是最常用的拟合检验统计量, 其中靴绊法修正的 $Adj\chi^2/df$ 是对其较新的调整, 但卡方类统计量仍存在一些无法克服的缺点, 并且多是针对项目的绝对拟合统计量。由此一些依据非卡方思想的拟合检验方法, 比如 PPMC 方法、基于信息矩阵的检验方法就得以开发。尽管有些统计量可能在理解上存在困难(如基于信息矩阵的检验), 并缺乏统一的应用软件, 但其克服了一些卡方类统计量的缺点, 并且有些统计量还能够描述具体不拟合的缘由(如 PPMC 方法可以检验 LI 假设, IM_1、IM_2 可以检验 ICC 单调性假设), 这比卡方类统计量只能笼统的说明模型–资料不拟合是一个很大的进步; 还有些统计量则可以用于针对测验的相对拟合检验, 如 AIC、DIC、CVLL。

应该注意的是模型–资料绝对拟合检验不同于模型选择, 其目的是确定模型是否充分。一个模型拟合数据(或者更精确的说法是未发现模型不拟合数据)并不能说明这个模型就是最好的, 或合适的模型。实际上通过绝对拟合检验的方法可能会发现在某一(如满足 LI 假设)甚至多个方面有几个模型都拟合数据(Sinharay, 2006), 要评价个人选择的模型是否真正适合分析实践的数据还要从多个方面评价模型–资料拟合(Lu,

2006)，并将绝对拟合统计量与相对拟合统计量结合应用，相对拟合用于比较待选模型，绝对拟合用于评价模型与资料的实际拟合情况。当然，模型与资料的完全拟合几乎是不可能的，但 IRT 模型往往具有一定的稳健性(McKinley et al., 1985)，在具体的情况下，仍可使用。下面以表格的形式给本文中所涉及的统计量一个总体的描述(见表 1)。

评价模型–资料拟合对于判断基于 IRT 模型做出的推断是否有效至关重要，从发展的角度，本文对模型–资料拟合检验的未来研究方向作出以下三点展望。

第一，未来研究可以用 Monte Carlo 或实证研究来综合比较各统计量的优劣，探讨在不同的使用条件下(如 ICC 单调性检验或 LI 检验)，哪个统计量才是最好的，并且开发出各种较优的拟合统计量的统一应用软件，为实际使用者提供借鉴。

第二，在未来，开发新的模型–资料拟合检验统计量，使其具备现已存在的各种拟合统计量的优点，是一个值得努力的方向。

第三，随着心理测量学和认知心理学的发展，认知诊断模型(cognitive diagnosis model)受到越来越多的重视(see, e.g., Chen & de la Torre, 2013; Finkelman, Kim, Roussos, & Verschoor, 2010; Sinharay & Almond, 2007)，与 IRT 的应用类似，为了使基于不同的认知诊断模型做出的推论有效，模型与资料的拟合问题仍然非常重要(Chen, de la Torre & Zhang, 2013)，IRT 中的模型–资料拟合检验统计量是否能够完全拓广到认知诊断领域以及如何拓广仍有很大的研究价值及潜力。

限于能力及篇幅，本文尚有一些拟合检验方法没有考虑到，如残差分析、拉格朗日乘数检验、因素分析等，有待未来的研究者增添改善。

表 1 模型–资料拟合检验统计量及其特点概览

类型	统计量		特点	
项目绝对拟合统计量	卡方类	χ^2、G^2	分组依据估计的能力值，每组人数大致相同，可作测验拟合统计量	要求大样本；除后两个统计量外，对大样本量敏感；难以确保观察频数不太小；自由度确定有歧义
		χ^{2*}、G^{2*}	按结点划分能力区间，各组观察频率是依赖结点求取的后验概率	
		S–χ^2、S-G^2	按测验总分分组，观察频数和期望频数只与题目和得分有关	
		$Adj\chi^2/df$	只按项目得分分组，观察频率只与题目和得分有关，期望频数基于项目参数和能力的先验分布求取	
		参数靴绊法修正 $Adj\chi^2/df$	相比 $Adj\chi^2/df$，产生样本分布，而不用固定临界值	
	PPMC 方法	观察分数的分布	检验能力的假设分布，是测验统计量	能检验不拟合的来源
		二列相关系数	检验模型是否缺少数据所必需的区分度参数	
		项目对的关联度量	检验局部独立性假设	
	基于信息矩阵的检验	IM_1、IM_2	测验拟合，检验项目特征曲线单调性假设	能检验不拟合的来源
		IM_3、IM_4	项目拟合，检验项目特征曲线单调性假设	
		IM_5、IM_6	测验拟合，检验局部独立性假设	
		IM_7、IM_8	项目拟合，检验局部独立性假设	
	卡方类	$G^2(dif)$	比较两个模型，选择更不严格的模型，可用于项目拟合	比较嵌套模型
		G^2	初用于项目绝对拟合检验	
测验相对拟合统计量	基于信息量	–2LL	不考虑模型复杂度，基于极大似然估计，是 AIC、BIC 基础	比较非嵌套模型
		AIC	惩罚复杂模型，趋于选择更复杂模型	
		BIC	惩罚力度比 AIC 大，考虑样本量，趋于选择更简单模型	
		DIC	参数估计基于 MCMC 算法，惩罚复杂模型	
		BF	基于贝叶斯估计，比较两个模型，且其中之一须真正拟合数据	
		CVLL	参数估计基于 MCMC 算法，其值越大越好	
		PsBF	基于 CVLL，比较两个模型	

参考文献

宋丽红. (2012). *DINA 改进模型(R-DINA)的提出及三个诊断模型自动选择机制研究*. 未出版之博士学位论文 江西师范大学.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541–562.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381–409.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.

Chen, J., de la Torre, J., & Zhang Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. Retrieved August 28, 2013, from http://www.jstor.org/stable/1165285

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*(2), 143–166.

Du Toit, M. (Ed.). (2003). *IRT from SSI: Bilog-MG, multilog, parscale, testfact* (561–731). Chicago: Scientific Software International.

Ferrando, P. J., & Lorenzo-seva, U. (2001). Checking the appropriateness of item response theory models by predicting the distribution of observed scores: The program eo-fit. *Educational and Psychological Measurement, 61*(5), 895–902.

Finkelman M. D., Kim W., Roussos L., & Verschoor A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement, 34*(5), 310–326.

Forster, M. R. (2004). *Simplicity and unification in model selection*. Manuscript, University of Wisconsin–Madison. Retrieved August 22, 2013, from http://philosophy.wisc.edu/forster/520/Chapter%203.pdf

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological), 56*(3), 501–514. Retrieved September 8, 2013, from http://www. jstor.org/stable/2346123

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733–807.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*(365), 153–160. Retrieved September 8, 2013, from http://www.jstor.org/stable/2286745

Ghosh, J. K., & Samanta, T. (2001). Model selection-An overview. *Current Science, 80*(9), 1135–1144.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics, 5*(6), 1148–1169.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage. Retrieved August 8, 2013, from http://books. google.com.hk/books?hl=zh-CN&lr=&id=cmJU9SHCzecC&oi=fnd&pg=PR7&dq=Fundamentals+of+item+response+theory&ots=K_PKJ-ChZT&sig=omFp1vCThNvEqv9gPM1LWTbvv6I&redir_esc=y&hl=zh-CN&sourceid=cndr#v=onepage&q=Fundamentals%20of%20item%20response%20theory&f=false

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*(2), 245–268.

Kang, T., & Chen, T. T. (2008). Performance of the generalized $S-\chi^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391–406.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331–358.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499–518.

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods, 14*(1), 10–23.

Lu, Y. (2006). *Accessing fit of item response theory models* (Unpublished doctoral dissertation), Massachusetts Amherst University.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using $G^2(dif)$ to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*(1), 55–64.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443–451.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49–57.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied*

Psychological Measurement, 24(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S-\chi^2$: an item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298.

Ounpraseuth, S., Lensing, S. Y., Spencer H. J., & Kodell, R. L. (2012). Estimating misclassification error: A closer look at cross-validation based methods. *BMC Research Notes, 5*(1), 656.

Ranger, J., & Kuhn, J. T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49*(3), 247–268.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*(2), 429–449.

Sinharay S., & Almond R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement, 67*(2), 239–257.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298–321.

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, MRC Biostatistics Unit, Cambridge, UK.

Spiegelhalter, D. J., & Smith, A. F. M. (1982). Bayes factors for linear and loglinear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological), 44*(3), 377–387. Retrieved September 16, 2003, from http://www.jstor.org/stable/2345495.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*(1), 58–75.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331–352.

Tay, L., & Drasgow, F. (2012). Adjusting the adjusted χ^2/df ratio statistic for dichotomous item response theory analyses: does the model fit. *Educational and Psychological Measurement, 72*(3), 510–528.

Western, B. (1999). Bayesian analysis for sociologists. *Sociological Methods and Research, 28*(1), 7–34.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245–262.

Common Model-Data Fit Test Statistics in Item Response Theory

SHAN Xintong; TAN Huiye; LIU Yong; WU Fangwen; TU Dongbo

(*College of Psychology*, *Jiangxi Normal University*, *Nanchang* 330022, *China*)

Abstract: Item response theory (IRT) models are widely applied psychometric models. They can predict responses based on characteristics of items and participants. But the validity in all applications of IRT is dependent on the extent to which the selected model accurately reflects the data at hand (goodness of fit). Only when the IRT model fits the data well, can the advantages and functions of IRT emerge (Orlando & Thissen, 2000). Selection of the wrong model would lead to relatively large error in parameter estimation, test equating, the analysis of differential item functioning and so on, which would result in adverse effect (Kang, Cohen & Sung, 2009). Therefore, it is required to evaluate model-data fit before applying IRT (McKinley & Mills, 1985). Common fit statistics in the field of IRT can be expounded and compared from test fit and item fit, which is very important in the field of psychological and educational measurement and also easily neglected in test analysis process. It has not been found yet that there are any similar published articles. Directions of future research for model-data fit could emphasize simulation and empirical study of this issue. And common fit statistics in the field of cognitive diagnosis could be investigated.

Key words: Item Response Theory; Test of Model-Data Fit; Item Fit; Test Fit