心理科学进展 2014, Vol. 22, No. 3, 549–557
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2014.00549

迫选式人格测验的传统计分与 IRT 计分模型*

王 珊[1] 骆 方[1,2] 刘红云[1,2]

([1] 北京师范大学心理学院, 应用实验心理北京市重点实验室, 北京 10875)
([2] 中国基础教育质量评价与提升协同创新中心, 北京 100875)

摘 要 迫选测验的传统计分方式会产生自模式数据, 不能进行传统的信效度检验、因素分析和方差分析等。近年来研究者提出了一些基于项目反应理论的计分模型, 如瑟斯顿 IRT 模型和 MUPP 模型等, 它们可以规避自模式数据的弊端。瑟斯顿 IRT 模型方便进行参数估计, 模型定义灵活; 而 MUPP 模型的拓展性较差, 参数估计的方法有待提高。另一方面, 已有研究者基于 MUPP 模型开发了一些抗作假的迫选测验, 而瑟斯顿 IRT 模型距离这种应用还比较远。此外, 两个模型的适用性和有效性都有待更多的实证研究来检验。

关键词 迫选测验; 自模式数据; 瑟斯顿 IRT 模型; MUPP 模型
分类号 B841

传统的人格测验多是 Likert 式量表, 这种测验形式容易产生各种反应偏差, 如默许反应(acquiescence response)、光环效应(halo effect)、印象管理(impression management, 在人才测评领域又被称为“作假”) 等(Cheung & Chan, 2002; Morrison & Bies, 1991)。自 20 世纪三四十年代起, 研究者开始采用迫选式人格测验来应对 Likert 式测验的反应偏差。尤其是在开发迫选测验时, 加入“称许性匹配”这一关键步骤, 即评定每一个陈述所描述的行为受社会称许的程度, 然后将称许性相同或者接近的陈述匹配成题组, 这种迫选测验就具有抗作假的性能。基于这些优点, 组织在人才选拔中经常使用迫选测验, 但是在研究领域却鲜有研究者使用迫选测验作为研究工具。究其原因在于测量学家很早就意识到迫选测验传统的计分方式存在严重问题(Bowen, Martin, & Hunt, 2002; Greer & Dunlap, 1997; Loo, 1999; Meade, 2004; Tenopyr, 1988)。随着现代测量理论的迅猛发展, 近十年来测量学家提出了一些精细的计分模型, 并据此开发了一些新型的迫选测验(Elosua, 2007; SHL, 2006; Stark, Drasgow, & Chernyshenko, 2008)。这些努力为迫选测验的开发和应用带来了新的曙光。下面将重点介绍迫选测验的传统计分方式与问题以及 IRT 计分模型, 并就 IRT 计分模型的优缺点及迫选测验的应用前景展开讨论。

收稿日期: 2013-04-27
* 本研究得到教育部人文社会科学研究青年基金项目(11YJC190016)、国家自然科学基金项目(31100759)、北京市与中央在京高校共建项目(019-105812)资助。
通讯作者: 骆方, E-mail: luof@bnu.edu.cn

1 传统计分方式和自模式数据

1.1 迫选测验的题目形式与计分

由两个陈述匹配成一个题组是迫选测验最常见的题目形式, 每个陈述可分别测量不同的人格特质, 被试的特质得分就是他选择测量该特质的陈述的个数。例如, 一个题组:

A. 我是一个热情而友善的人。(外向性)
B. 我的物品经常保持整洁干净。(条理性)

要求被试选择最符合自己的一项。这两个陈述分别测量外向性和条理性, 如果被试选择“A”, 则外向性维度计 1 分, 条理性维度计 0 分; 如果被试选择“B”, 则外向性维度计 0 分, 条理性维度计 1 分。

迫选测验允许两个以上的陈述匹配成组。如可将测量三种不同特质的陈述配成一组, 要求被试选出最符合自己的一项和最不符合自己的一项:

A. 我是一个热情而友善的人。(外向性)
B. 我的物品经常保持整洁干净。(条理性)

C. 抗拒诱惑对我来说没太大困难。(情绪稳定性)

被试选择最符合的一项计 2 分(如选择“A”, 则外向性维度计 2 分), 最不符合的记 0 分(如选择“C”, 则情绪稳定性维度计 0 分), 未选择项记 1 分(条理性维度计 1 分)。可见, 无论被试怎样作答, 每个题组的满分都是 3 分, 因而整个测验的满分对所有被试而言也是个常数。

Cattell (1944)指出, 在迫选测验这种传统的计分方式下, 产生的测验分数为自模式数据(ipsative data), 这种数据的总分是一个常数; 而 Likert 量表分数的总分不固定, 是一种常模数据(normative data)。

1.2 自模式数据的性质及其应用局限性

Meade (2004)对迫选测验所产生的自模式数据的测量学性质给出了数学描述。假设一个两两配对的迫选测验, 测量两种特质(a 和 b), 那么在经典测验理论(CTT)框架下, 被试的测验总分可以用以下公式来表示：

$$C = T_a + T_b + e_a + e_b \quad (1)$$

其中, C= 题组个数(是个常数); T_a和 T_b为两个维度的真分数, e_a和 e_b为测量误差。

由于测验总分 C 是个常数, 当被试的真分数(T_a 和 T_b)比较大时, 误差值可能为负值; 当被试的真分数比较小时, 误差值比较大。显然, 在迫选测验中, 真分数与误差存在负相关。这违背了 CTT 的基本假设。

同时, 特质 a 的观测分数可表示为：

$$X_a = T_a + e_a \quad (2)$$

由公式(1), 可得：

$$X_a = C - (T_b + e_b) \quad (3)$$

则公式(2)+(3)除以 2, 得到：

$$X_a = \frac{C + T_a - T_b + e_a - e_b}{2} \quad (4)$$

因此, 特质 a 的测验分数是维度 a、b 的真分数与 a、b 的测量误差的函数。

当迫选测验测量多个特质时, 第 j 个特质的分数为：

$$X_j = \frac{C + T_j + e_j - \sum_{1}^{s \neq j}(T + e)}{2} \quad (5)$$

上述数学推导说明不同特质的测验分数之间存在着相互制约关系, 这同样违背了 CTT 的基本假设。对于自模式数据, 测量学家和统计学家反对采用与常模数据相同的方式来处理和分析(如 Clemans, 1966; Closs, 1996; Cornwell & Dunlap, 1994; Johnson, Wood, & Blinkhorn, 1988; Loo, 1999; Meade, 2004)。具体的原因如下：

(1)迫选测验的分数不能代表被试人格特质的绝对水平, 只能反映个体内部人格特质的相对水平。如公式(4)中测验分数 X_a 反映的是 T_a–T_b 的性质, 而不仅是特质 a 的分数高低。因而, 被试在一些人格特质上得分高, 在另一些特质上得分必然会低, 这使得测验分数的意义难以解释。

(2)因子(特质)分数的协方差矩阵的每行/列的和为零。因子相关之和为负值, 且趋向于 $-1/(m-1)$, 其中 m 为变量个数(Clemans, 1966; Hicks, 1970)。因而, 因子之间的相关是不可解释的, 任何典型的因素分析方法都不适用于分析迫选测验的分数。比如, 自模式数据通常会产生两极化的因子, 即一个题目在两个因子上有较大的载荷, 而且方向相反。Cornwell 和 Dunlap (1994)认为, 应使用称名变量分析技术(multi-nomial statistical techniques)来分析自模式数据, 这样可以解决两极化因子的问题; 但是当分量表数目较大时, 这种方法就变得难以操纵。

(3)由于测验分数不仅测量了某一特质的真分数, 还会受到其他特质的真分数和误差分数的污染, 经典测验理论的典型假设—— 某一给定特质的观察分数的期望值等于真分数—— 就不能成立, 传统的信度估计方法都不再适用。Tenopyr (1988)认为, 由于项目水平的误差存在相关, 迫选测验的内部一致性系数会被高估, 但另一方面这种系数的膨胀会随着每个分量表题目的增多而大幅降低。

(4) Clemans (1966)指出由于测验分数存在内部关联性, 使得测验分数与效标之间的协方差之和为零。显然, 如果测验分数满足方差齐性假设, 那么效度(测验分数与效标的相关系数)的和为零。因而, 不宜解释迫选测验与外在效标的预测效度或者同时效度的意义, 因为一些特质的效度系数为正, 一些特质的效度系数必然为负。同样, 迫选测验的分数也不适宜进行回归分析、方差分析等, 它会增加犯 I 类错误的概率, 同时也会影响统计检验力(Greer & Dunlap, 1997)。

针对自模式数据的问题, 研究者也提出了一

些解决建议，如：

(1) Saville 和 Willson (1991)和 Bartram (1996)通过模拟研究提出，测量的人格特质数量应达到30，且不存在相关或者至少不存在中等程度以上的正相关时，其信度、内部相关等接近常模数据。

(2)数据分析时去掉一个项目、部分因子，或者改变计分权重等，使测验总分不再是一个常数(Hicks, 1970; White & Young, 1998)。

(3)针对自模式数据直接进行因素分析所存在的问题, Chan 等人提出在验证性因素分析框架下，可以通过一个转换矩阵，将自模式因子模型转换为常模因子模型，并且运行一个两阶段的参数估计程序，可以得到较好的模型拟合结果(Chan, 2003; Chan & Bentler, 1998)。

自模式数据大大限制了迫选测验的发展和应用 (Christiansen, Burns, & Montgomery, 2005; Converse et al., 2010; Goffin, Jang, & Skinne, 2011; Heggestad, Morrison, Reeve, & McCloy, 2006)。虽然测量和统计学家对自模式数据的性质进行了深入探讨，并且提出了一些补救的方法，但是只要迫选测验的计分方式没有改变，迫选测验的分数就会是自模式数据，测量学家的各种努力只是在尽量降低自模式数据的影响，使其分析方法和解释更加合理，并不能从根本上避免自模式数据的问题。

2　迫选式人格测验的 IRT 计分模型

进入 21 世纪以后，测量学家基于被试对迫选测验的作答过程建构了一些项目反应理论(IRT)模型，这些精细的模型摒弃了传统的计分方式，可以估计被试潜在的人格特质水平(特质分数是常模数据)，从而规避了自模式数据带来的一系列问题。目前有较大影响的 IRT 模型是由 Brown 等人提出的瑟斯顿 IRT 模型(Thurstonian IRT Model)和 Stark 等人提出的 MUPP 模型(Multi-Unidimensional Pairwise-Preference Model, MUPP)。

2.1　瑟斯顿 IRT 模型

Thurstone (1927)提出了著名的瑟斯顿比较判断法则(Thurstone's Law of Comparative Judgment)来描述被试进行迫选时的反应过程。他认为对于配对的两个陈述 i 和 k，被试独立反应各产生一个效应值(utility) t_i 和 t_k。假设

$$y_l^* = t_i - t_k + \varepsilon \tag{6},$$

$$y_l = \begin{cases} 1, & \text{if} \quad y_l^* \geq 0, \\ 0, & \text{if} \quad y_l^* \leq 0. \end{cases} \tag{7},$$

即如果 t_i 大于 t_k，则被试倾向于选择陈述 i，否则选择陈述 k。Thurstone 认为应该允许判断误差 ε 的存在。他假设 t_i、t_k 和 ε 都服从正态分布，并且相互独立。

Brown 在此基础上提出了瑟斯顿 IRT 模型(Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010)。假设效应值 t 为所测潜在特质 η 的线性函数，那么被试对题目 l 的两个陈述产生的效应值 t_i 和 t_k，可以表述为潜在特质 η_a 和 η_b 的线性函数：

$$\begin{cases} t_i = \mu_{t_i} + \lambda_i \eta_{\mathrm{a}} + e_i \\ t_k = \mu_{t_k} + \lambda_k \eta_{\mathrm{b}} + e_k \end{cases} \tag{8}$$

则 y_l^* 也有一个新的表达式：

$$\begin{aligned} y_l^* &= t_i - t_k = \mu_{t_i} - \mu_{t_k} + \lambda_i \eta_a - \lambda_k \eta_b \\ &+ e_i - e_k = \gamma_l + \lambda_i \eta_a - \lambda_k \eta_b + e_i - e_k \end{aligned} \tag{9}$$

y_l^* 是一个连续变量，被试的反应 y_l 是(0, 1)分类变量。假设潜在特质、误差服从正态分布，则 y_l^* 也服从正态分布，项目反应函数(Item Characteristic Function, ICF)满足正态拱形模型(Normal Ogive Model)，即被试选择陈述 i 而不是陈述 k 的条件概率为：

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi\left(\frac{\gamma_l + \lambda_i \eta_a - \lambda_k \eta_b}{\sqrt{\psi_i^2 + \psi_k^2}} \right) \tag{10}$$

也可以表述为：

$$\Pr(y_l = 1 | \eta_a, \eta_b) = \Phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b \right) \tag{11}$$

这是一个 0-1 记分的双参数多维 IRT 模型。η 为被试的潜在特质分数，与 Likert 量表所测的潜在特质具有相同的性质，服从多元正态分布，没有测量误差，这些特点都优于传统计分。

瑟斯顿 IRT 模型可以采用有限信息估计方法比如未加权最小二乘法(Unweighted Least Squares, ULS) 或者对角加权最小二乘法(Diagonally Weighted Least Squares, DWLS) 等(Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009)来进行题目参数估计。对于被试特质水平 η 的估计可以采用极大后验估计法(Maximum a Posteriori, MAP)。

Brown 和 Maydeu-Olivares (2011)通过模拟研究探讨了影响瑟斯顿 IRT 模型的拟合度、模型参数和潜在特质估计精度的一些因素。结果表明，迫选测验满足以下条件才能实现瑟斯顿 IRT 模型的良好估计：

(1)用来配对的陈述的区分度高，即要选择测量属性优良的陈述匹配成题组；

(2)测评的特质较多，而且相关较低；

(3)瑟斯顿 IRT 模型允许两个以上的陈述配对，而且多个陈述的配对优于两两配对；

(4)最好有一半的题组是正向陈述和负向陈述配对，比如“A. 我是一个很乐观的人。”是正向陈述，而“B. 我每天心情都不好。” 是负向陈述。如果都是正向陈述配正向陈述，则题组数量的增多不会增加估计精度。例如，当测量特质为 5 个、每个题组包含 4 个陈述、且含有正向和负向陈述配对的题组时，仅需 15 个题组，模型参数(阈值、载荷、特质间相关和残差方差)的相对偏差(relative bias)在 0.01 左右，估计的特质分数与真值的平均相关为 0.94，信度为 0.89。如果没有正负配对的题组，则同样条件下特质间相关的相对偏差增大到 0.095，估计的特质分数与真值的平均相关为 0.87，信度为 0.75。

可见含有正向和负向陈述配对的题组对瑟斯顿 IRT 模型实现精确估计非常重要，但是如果迫选测验满足了该条件就很难具备抗作假的性能了，因为在高利害的选拔情境下，被试如果有意作弊是不会选择负向陈述的。因而，Brown 和 Maydeu-Olivares (2012, 2013)指出编制适用于瑟斯顿 IRT 模型的抗作假迫选测验是一件非常有挑战性的工作，还需要进一步的深入研究。目前，对于用以应对 Likert 量表常出现的一致性偏差、光环效应等的迫选测验，可直接使用瑟斯顿 IRT 模型进行分析。

Brown 和 Maydeu-Olivares (2013) 收集了实测数据来比较 Likert 式和迫选式的顾客沟通风格问卷 (Customer Contact Styles Questionnaire, CCSQ)的信度和效度。该问卷测量了 16 个维度，共 128 个题目。研究者将来自不同维度的四个陈述组成一个题组，要求被试在作答时对题组中的每个陈述均进行 Likert 式的五点量表评定，然后选择出“最符合”和“最不符合”自己的陈述各一项。这种作答方式可以同时获得 Likert 式和迫选式的作答数据，对每种数据都同时进行 CTT 和 IRT 分析，其中 Likert 式数据采用的 IRT 模型是等级反应模型(Graded Response Model)，迫选式数据采用的是瑟斯顿 IRT 模型。结果表明，对于迫选式数据如果采用瑟斯顿 IRT 方法估计的经验信度为 0.72~0.89，中数为 0.80，大于 CTT 框架下估计的信度(α 系数为 0.57~0.80，中数为 0.72)，但是它们都低于 Likert 式数据的信度，这是由于被试对迫选测验的作答反应较少造成的。例如被试对一个题组的迫选反应，只需选出“最符合”和“最不符合”两项就可以了，而对题组中的四个陈述都需作 Likert 式反应。对于 Likert 式数据，采用 CTT 和 IRT 方法分析，得到的特质(维度)间的相关均为正值，平均相关分别为 0.22 和 0.21；对于迫选式数据，采用瑟斯顿 IRT 模型估计的特质相关也均为正值，平均相关为 0.12，与 Likert 式数据的特质相关比较接近；而采用传统计分得到的维度间的相关有正有负，平均相关为-0.07，这是由于它们是自模式数据造成的。效标关联效度(效标为基于工作表现给予的奖金)的检验也发现，采用迫选测验传统计分方式计算的维度分数与效标的相关有正有负(其中有两个维度存在显著的负相关)，而采用瑟斯顿 IRT 模型估计的特质分数与效标的相关与 Likert 问卷的效标关联效度非常接近。

Brown 和 Maydeu-Olivares 认为，这些结果可以说明采用瑟斯顿 IRT 模型计分优于传统计分，模型估计的特质分数已不具有自模式数据的性质，可以在迫选测验的计分和信效度检验中应用。但是，迫选测验是否优于 Likert 式测验，或者在某些测试条件下(如选拔情境、临床筛查等)哪种形式的测验效度更佳，都还需要更多的研究来检验。

2.2 Stark 的 MUPP 模型

Stark, Chernyshenko 和 Drasgow (2005)提出了 MUPP 模型。瑟斯顿 IRT 模型和 MUPP 模型都描述了被试发生迫选行为时的反应过程，但是它们的根基—— 关于被试对单个的人格陈述如何反应的理解—— 却是完全不同的。瑟斯顿 IRT 模型假设它是一种优势模型(Dominance Model)，即被试的特质水平越高，被试的反应概率越大。例如，对于陈述“我大多数的时候心情都很好”，被试的乐观性越高，越容易同意这句话。MUPP 模型则假设被试对单个陈述的反应是一种展开模型(Unfolding Model)，即被试距离题目 i 的位置参数

(x_i)越近，越容易同意；越远，越容易不同意。例如，对于一个中等程度的陈述“我认为到一些国家旅游还不错”，如果被试不热爱旅游，或者是个旅游发烧友都不会同意这句话。

Stark 基于展开模型建构了 MUPP 模型，该模型认为，假设配对的两个陈述 s 和 t 分别测量特质 θ_{ds}和θ_{dt}，被试对两个陈述 s 和 t 独立反应，则其选择 s 而不选择 t 的反应概率等于

$$P_{(s>t)i}\left(\theta_{d_s},\theta_{d_t}\right)=\frac{P_{st}\{1,0\}}{P_{st}\{1,0\}+P_{st}\{0,1\}}\approx \frac{P_s\{1\}P_t\{0\}}{P_s\{1\}P_t\{0\}+P_s\{0\}P_t\{1\}} \tag{12}$$

其中，$P_s(1)$是指被试同意陈述 s 的概率，$P_s(0)$是指被试不同意陈述 s 的概率；$P_t(1)$是指被试同意陈述 t 的概率，$P_t(0)$是指被试不同意陈述 t 的概率。它们都是被试对单个陈述 s 或者 t 的反应概率，被定义为广义等级展开模型(Generalized Graded Unfolding Model, GGUM)，即

$$P_s(0)=\frac{1+\exp(\alpha_s\left[3(\theta_{d_s}-\delta_s)\right])}{\gamma_s} \tag{13}$$

和

$$P_s(1)=\frac{\exp(\alpha_s\left[(\theta_{d_s}-\delta_s)-\tau_{s1}\right])+\exp(\alpha_s\left[2(\theta_{d_s}-\delta_s)-\tau_{s1}\right])}{\gamma_s} \tag{14}$$

MUPP 模型中的参数估计不是基于被试对迫选题目的反应数据，而是先让被试对单个的陈述(采用 Likert 评定方式)作答，然后基于这个数据进行单维的 GGUM 模型的参数估计。参数估计结束后，把 α、δ、γ 和 τ 的估计值代入公式(12)，用该模型来拟合被试对迫选题目的反应数据，采用 Bayes 估计求出被试的潜在特质 θ，该特质分数服从多元正态分布，不再是自模式数据。

作为一个工业与组织心理学领域的研究者，Stark 比 Brown 更看重迫选测验在组织中的应用价值，特别强调迫选测验的抗作假性能。他提出了基于 MUPP 模型的抗作假的迫选测验的编制过程，其具体步骤如下：(1)针对每个特质，编写大量的陈述；(2)要求一组评委评定每个陈述的社会称许性；(3)让一组被试作答每一个陈述(采用 Likert 评定方式)，定义其反应模型为展开模型，采用 GGUM2000 软件(Roberts, Donoghue, & Laughlin, 2000)，估计每个陈述的位置参数；(4)创建抗作假的题组。匹配陈述的条件是：(1)社会称许性近似；(2)配对的陈述来自不同的维度，但在所有配对中要有 10%~20%的单维配对(即来自同一维度的陈述两两配对)；(3)用于配对的两个陈述的位置参数不同，如果位置参数接近，被试会随机反应。Stark 等人(2005)基于该编制过程生成了多套迫选测验，然后模拟被试的反应，再采用 MUPP 模型估计被试的特质分数。结果发现，当陈述个数为 40 (两两配对成 20 个题组)时，估计的特质分数与真值的平均相关为 0.90；如果陈述个数为 20，且只含有 10%的单维配对时，相关为 0.76；而陈述个数为 80，且含有 20%的单维配对时，相关为 0.94。

Chernyshenko 等人(2009)通过实测数据，比较了测量秩序、自我控制和社交性三种人格特质的 Likert 式测验和基于 MUPP 模型开发的抗作假迫选测验的信效度。采用 MUPP 模型分析迫选测验的作答数据，三种特质的信度分别为 0.75, 0.66 和 0.73，高于 Likert 式测验的信度(α 系数分别为 0.74, 0.54 和 0.62)。多质多法分析的结果显示，对于所测的同一种特质，采用两种测验形式估计的特质分数的相关大多在 0.7 以上，而且不同测验形式同种特质间的相关要远大于相同测验形式不同特质间的相关。两种测验的效标关联效度 (效标为健康和学习行为) 非常接近。此外，Stark, Chernyshenko, Drasgow 和 White (2012) 基于 MUPP 模型开发了计算机自适应的迫选测验，实测数据的检验结果表明迫选测验与效标之间的相关符合理论假设，效标效度良好。

2.3　瑟斯顿 IRT 模型和 MUPP 模型的比较

无论是理论模型的建构，还是实际应用的可行性，两个模型都有很大的差异，也各有一些不足的地方。

首先，瑟斯顿 IRT 模型和 MUPP 模型关于被试对人格题目的反应方式的理解不同。那么，人们对单个的人格陈述是怎样反应的，是优势模型，还是展开模型更合理，至今仍是测量学家争论的热点话题。传统的人格测验题目都是能够代表潜在特质的典型陈述，比如，“我是个快乐的、对什么事情都兴致勃勃的人”，测量“乐观性”的题目一般称许性都会比较高。这些题目适合于优势模型，即被试选择“同意”，代表乐观性高，选择“不同意”，代表乐观性低。展开模型适用于题目位置比

较分散的测验，很多题目的社会称许性就会不明显，因而更适用于编制抗作假的测验。

其次，目前的 MUPP 模型只适用于两两配对的题目，还没有将其应用到一个题组包含多个陈述的题目形式中，应用范围较窄。此外, MUPP 模型的题目参数是基于被试对单个陈述的反应数据估计出来的，还需要对模型进行改进，提出基于迫选反应数据的题目参数估计方法。相反，瑟斯顿 IRT 模型适用于各种迫选的题目形式，比如多于两个的陈述配成题组，多个陈述排序等题型。而且，瑟斯顿 IRT 模型存在更大的拓展空间，例如可以在模型中增加参数，定义潜变量之间的关系，引入预测变量和结果变量等，使其适用于多样化的迫选反应数据(如存在不同子群体或者多个层级的数据)，满足更多的分析需求(如探究特质和外源变量之间的关系)。

第三, Stark 等基于 MUPP 模型给出了开发抗作假的迫选测验的操作步骤，模拟研究显示 MUPP 模型能够较好地拟合这种抗作假的迫选测验的反应数据，但是还缺少大量的在选拔情境中的实证数据的检验，这限制了该模型的实际应用和推广。瑟斯顿 IRT 模型是针对一般性的迫选测验提出的反应模型，然而它能否较好地拟合抗作假迫选测验的作答数据，目前还没有任何的模拟研究和实证研究来证明。实际上，直接把瑟斯顿 IRT 模型应用到抗作假的迫选测验上很难得到良好的估计，必须基于瑟斯顿 IRT 模型的性质来改进迫选测验的编制方法，增加陈述匹配的条件，这样产生的抗作假的迫选测验的反应数据才适用以瑟斯顿 IRT 模型进行分析。

总之，瑟斯顿 IRT 模型是一个具有较好的测量学属性的模型，它把瑟斯顿比较判断法则定义为一个普通的多维 IRT 模型，采用大众化的软件 Mplus (Muthén & Muthén, 2013)就可以实现参数估计，非常有利于普及。但是，还需要进一步探讨该如何基于瑟斯顿 IRT 模型开发抗作假的迫选测验，使其有益于解决人格测验在应聘情境中使用的窘境。Stark 等人已经提出了基于 MUPP 模型的抗作假迫选测验的开发和计分程序，但是MUPP模型的拓展性较差，参数估计的方法有待提高，而且模型复杂，参数估计程序不易获得和使用。可见，两个模型都有待进一步的理论和实证研究。

3 发展趋势与研究方向

以上对迫选测验的传统计分方式及其弊端进行了总结，并介绍了针对迫选测验的 IRT 计分模型的新进展。下面将主要探讨迫选测验计分模型的应用前景及发展中亟待研究者去探索和解决的问题。

3.1 基于迫选反应过程的 IRT 计分模型的出现，有望解决传统计分的问题

面对传统计分方式带来的难以解决的自模式数据的问题, IRT 模型为研究者提供了一个全新的视角和广阔的研究空间，目前已成为近几年来人格测验研究领域的一个热点问题，除了瑟斯顿 IRT 模型和 MUPP 模型，一些新的 IRT 模型将会陆续出现。但是，这些新型的计分模型都有待大量实证性研究的检验和修正，模型的适用性和有效性等问题都有待进一步探讨。

3.2 有待开展采用 IRT 模型估计的迫选测验特质分数在应聘情境中的效度研究

迫选式人格测验在应聘情境下使用的有效性研究通常采用相同的研究范式，即通过对测验作答情境的操控，分别创建一个无压力情境(假设被试会诚实作答)，和一个模拟或者真实的应聘情境(假设被试在此种情境下较易出现作假反应偏差)。然后，考察被试在不同的测验情境中，对迫选测验与 Likert 量表的作答是否存在差异。已有研究结果发现：在(模拟)应聘情境下，迫选测验受作假影响较小。例如，相比无压力情境，在应聘情境下，《个人动机评估》(Assessment of Individual Motivation, AIM) (White & Young, 1998)这种迫选测验的分数升高 0.1 个标准差；而传统的 Liket 测验《背景和生活经历评估》(Assessment of Background and Life Experiences, ABLE)分数升高了 1.5 个标准差(Stark et al., 2011)。此外，迫选测验在无压力情境与(模拟)应聘情境中都对效标有显著预测作用，但是 Likert 测验的这种预测作用仅体现在无压力情境中(Hogan, 2005; Viswesvaran, Deller, & Ones, 2007)。这些都是基于采用传统计分方式获得的迫选反应数据得出的研究结论，由于自模式数据存在很大的局限性，这些结论并不可靠。因而，采用 IRT 模型估计的迫选测验特质分数是否同样具有较强的抗作假性，以及在无压力情境和(模拟)应聘情境下其与 Likert 量表分数的关联性和差异性都有待进一步探讨。

3.3 有待开发自适应的迫选测验来解决迫选测验题量大的问题

为了实现 IRT 模型的精确估计, 迫选测验往往需要大量的题目, 这会阻碍它在实际中的应用。计算机自适应测试(Computerized Adaptive Test, CAT)是随着 IRT 技术和计算机技术发展而出现的一种新的测量手段, CAT 根据被试当前的能力估计水平不断地从题库中选择下一个最适合被试作答的题目施测, 因而使用 CAT 只需较少的项目就可以提供更加准确的能力估计。如果将迫选测验与 CAT 技术结合起来, 可以有效地克服迫选测验题量大的问题。

2011 年和 2012 年 Stark 等人开发了基于 MUPP 模型的自适应测验(Stark & Chernyshenko, 2011; Stark et al., 2012)。他们发现, 题量限制相同的情况下, 如果迫选测验采用自适应程序施测, 对被试能力估计的精度更高。但是, 被试必须完成所有的题目该适应程序才能结束, 因而不能达到减少题量的目的。此外, 这一自适应程序并没有实现真正的随机配对, 而是按照每个陈述的参数事先进行配对, 也就是题库中的题对是事先预置好的, 还不能在被试作答过程中实现陈述的自由配对。此外目前还没有开发出基于瑟斯顿 IRT 模型的自适应测验, 因而还需要大量的迫选自适应测验的研究。

迫选式人格测验可以有效降低传统 Likert 式测验的反应偏差, 有利于提高人格测评的测量精度和有效性。迫选测验的 IRT 模型, 可以规避传统计分方式带来的自模式数据的问题, 测验分数方便进行统计分析, 有利于挖掘和揭示变量之间的关系。因而迫选式人格测验不仅在人才选拔应用中能够发挥重要的作用, 而且在学术研究领域也有望得到重视和广泛应用。

参考文献

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology, 69*, 25–39.

Bowen, C. C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *International Journal of Organizational Analysis, 10*(3), 240–259.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502.

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292–303.

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30*(1), 99–121.

Chan, W., & Bentler, P. M. (1998). Covariance structure analysis of ordinal ipsative data. *Psychometrika, 63*(4), 369–399.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22*(2), 105–127.

Cheung, M. W., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*(1), 55–77.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307.

Clemans, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures* (Psychometric Monograph No. 14). Richmond, VA: Psychometric Society.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational Psychology, 69*, 41–47.

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance, 23*(4), 323–342.

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational and Organizational Psychology, 67*(2), 89–100.

Elosua, P. (2007). Assessing vocational interests in the Basque country using paired comparison design. *Journal of Vocational Behavior, 71*(1), 135–145.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625–641.

Goffin, R. D., Jang, I., & Skinner, E. (2011). Forced-choice and conventional personality assessment: Each may have unique value in pre-employment testing. *Personality and Individual Differences, 51*(7), 840–844.

Greer, T., & Dunlap, W. P. (1997). Analysis of variance with ipsative measures. *Psychological Methods, 2*(2), 200–207.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9–24.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167–184.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance, 18*(4), 331–341.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153–162.

Loo, R. (1999). Issues in factor-analyzing ipsative measures: The learning style inventory (LSI-1985) example. *Journal of Business and Psychology, 14*(1), 149–154.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*(6), 935–974.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*(4), 531–551.

Morrison, E. W., & Bies, R. J. (1991). Impression management in the feedback-seeking process: A literature review and research agenda. *The Academy of Management Review, 16*(3), 522–541.

Muthén, L. K., & Muthén, B. O. (2013). *Mplus user's guide (Version 6.1)* [Computer software and manual]. Los Angeles, CA: Muthén & Muthén.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). GGUM2000: A DOS-based program for unfolding ideal point responses [Computer program]. Department of Measurement and Statistics, University of Maryland.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology, 64*, 219–238.

SHL. (2006). *OPQ32 Technical manual.* Thames Ditton, UK: SHL Group plc.

Stark, S., & Chernyshenko, O. S. (2011). Computerized adaptive testing with the Zinnes and Griggs pairwise preference ideal point model. *International Journal of Testing, 11*(3), 231–247.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184–203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*(3), 463–487.

Stark, S., Chernyshenko, O. S., Drasgow, F., Lee, W. C., White, L. A., & Young, M. C. (2011). Optimizing prediction of attrition with the US Army's Assessment of Individual Motivation (AIM). *Military Psychology, 23*(2), 180–201.

Stark, S., Drasgow, F., & Chernyshenko, O. S. (2008, 29 Sept - 3 October). *Update on Tailored Adaptive Personality Assessment System (TAPAS): The next generation of personality assessment systems to support personnel selection and classification decisions.* Paper presented at the meeting of 50th annual conference of the International Military Testing Association, Amsterdam, Netherlands.

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology, 73*(4), 749–751.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Viswesvaran, C., Deller, J., & Ones, D. S. (2007). Personality measures in personnel selection: Some new contributions. *International Journal of Selection and Assessment, 15*(3), 354–358.

White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the meeting of annual meeting of the American Psychological Association, San Francisco, CA.

The Conventional and the IRT-based Scoring Methods of Forced-Choice Personality Tests

WANG Shan[1]; LUO Fang[1,2]; LIU Hongyun[1,2]

([1] *Beijing Key Lab of Applied Experimental Psychology*, *School of Psychology*, *Beijing Normal University*, *Beijing* 100875, *China*)

([2] *National Cooperative Innovation Center for Assessment and Improvement of Basic Education Quality*, *Beijing Normal University*, *Beijing* 100875, *China*)

Abstract: The conventional scoring method of forced-choice personality tests produces the ipsative data to which reliability and validity analysis, factor analysis, and analysis of variance (ANOVA) can not apply. In order to overcome these problems, some researchers have come up with item response theory-based (IRT-based) models for the scoring and analysis of forced-choice tests. These models, such as Thurstonian IRT model and Multi-Unidimensional Pairwise-Preference Model (MUPP), can avoid the disadvantages of ipsative data. On one hand, Thurstonian IRT model can be easily used for parameter estimation and it is flexible in model specification, while MUPP has poor expansibility and still needs improvement in parameter estimation method. On the other hand, MUPP has been used to develop forced-choice tests against faking but Thurstonian IRT model is still far from this kind of application. More empirical studies are needed to test the applicability and effectiveness of both models.

Key words: forced-choice tests; ipsative data; Thurstonian IRT model; MUPP