

应用项目反应理论对《中国士兵人格问卷》的项目分析

杨业兵 苗丹民 田建全 肖利军 孙 菡 洪 霞

(第四军医大学航空航天医学系心理学教研室, 西安 710032)

摘 要 采用项目反应理论(IRT)对《中国士兵人格问卷》进行项目分析。计算机呈现中国士兵人格问卷(CSPQ)对 100,523 名适龄男性青年进行测验,随机抽取 2676 名任一维度标准分均低于 70 的定为合格组;将任一维度大于 70 分并经专业人员访谈不合格的 274 名定为不合格组;从精神病院抽取男性年龄相当的 221 名缓解期精神分裂症患者定为精神病组,并完成 CSPQ 测验。运用基于 IRT 的双参数 Logistic 模型进行分析;结果发现,区分度参数超过区间(0.30,4.00)的条目删除前后,被试的能力值与标准分均存在显著相关;精神病组的测验分数经 IRT 分析,图形曲线与不合格组有高度吻合。研究结果说明,在测验精度基本相同的条件下,应用 IRT 可以减少施测条目,提高测验效率,可在一定程度上更精确地区分被试的特质水平。

关键词 人格测验,项目信息函数,双参数逻辑斯谛模型。

分类号 B848;B841

1 引言

项目反应理论(item response theory, 简称 IRT)是相对于经典测验理论(classical testing theory, 简称 CTT)而言的一种新的测验理论,具有许多优点:第一,项目与特质参数具有相对不变性;第二,项目难度水平和被试特质水平的取值,都被定义在同度量系统上,可以更方便地进行比较;第三,定义了经典理论中没有的项目和测验信息函数;第四,提出了测验编制的新指导原则,即选用项目难度水平与被试水平相匹配的原则(项目信息量最大原则),来组成测验^[1]。项目反应模型一般由某种数学函数(Mathematical Function),和这种函数建立其上的一批或一组假设所构成。项目反应理论的常用模型主要有:单参数逻辑斯谛模型(1PL)、双参数逻辑斯谛模型(2PL)、三参数逻辑斯谛模型(3PL),其主要假设为单维性假设、局部独立性假设和项目特性函数的具体形式。IRT 描述了一组可能性模型,其中每个项目都包含一组参数(例如,三参数模型中为难度、区分度、猜测系数)。项目参数与被试能力水平共同决定了被试答对该项目的概率。如果被试水平相对高于给定项目的难度,那他答对的概率就大。相反,如果被试水平低于给定项目难度,他正确做答

的概率就小,如果被试水平与项目难度相等,那在不考虑猜测系数的模型中,他正确做答的概率就等于 0.50^[2],这种关系可以形象地通过项目特性曲线(ICC)来表示,其中被试特质水平为横坐标,正确做答概率为纵坐标,项目的参数决定了曲线的形状。

目前国外大型测验(如 ASVAB、GRE、TOEFL 等)的编制、项目分析、计分等环节,均以 IRT 作为主要理论依据。基于 IRT 的计算机自适应测验(Computerized Adaptive Testing, CAT)也使得测验更加便于实施和推广。国外已有很多研究将 IRT 应用到了人格测验的领域中^[3-8],但主要集中于对原有的量表进行修订,完全依据 IRT 编制的人格量表还鲜见报道^[7]。

由于人格测验相对于成就测验有其自身的特点,应用何种模型依然是研究者争论的焦点之一。现阶段在人格测验中多应用双参数 Logistic 模型(2PL)或其多级计分条目中的推广——等级反应模型进行分析^[7],即假定在人格测验中不存在猜测因素。但 Reise 等将三参数模型(即存在猜测的模型)同双参数模型进行了对比研究,结果发现两者的拟合度差别非常小,相关达到 0.99 以上,同时他们也发现由于人格测验的独特性,条目应答存在上限和下限的现象,即使达到了某个特质的极端也不

可能完全不符合或完全符合条目所描述的内容,因此提出了四参数模型(4PL)^[8]。这些都是在应用 Logistic 模型基础上进行的研究。随着非参项目反应理论(non-parametric item response theory)的发展,也有研究指出由于人格测验的条目不一定服从 Logistic 模型的形式,应用更加灵活的非参统计方法可能更符合实际情况^[6,9]。但 Stephen 等提出有些人格测验是根据理想点的理论(Ideal Point Response Processes)^[10]来编制的,所以应用这种模型来分析人格测验的结果可能更恰当^[11]。国内也有研究者将 IRT 与 CTT 在人格测验中的应用进行了对比,结果发现在测验结果分析上,IRT 方法与 CTT 计分之间有一定的差异,IRT 所得的结果更精确^[12,13]。本研究采用现阶段广泛采用的双参数 Logistic 模型,分析《中国士兵人格问卷》(CSPQ)^[14]的 6 个维度各条目的参数和信息函数,为 IRT 应用于人格测验提供实证依据。

2 方法

2.1 测量工具

《中国士兵人格问卷》(CSPQ)^[14]。该量表主要用于应征公民心理检测,包括 3 个人格分量表,如分离特质(Dit)、神经特质(Net)和敏感特质(Set);3 个效度分量表,如防御性(D)、求异性(S)和掩饰性(T),CSPQ 共 213 个条目。

2.2 被试及其分组方法

被试为来自全国适龄男性青年 100523 名,平均年龄 18.51 ± 2.32 岁。完成计算机 CSPQ 测验;将任一分量表标准分均低于 70 的随机抽取 2676 人定为合格组;将任一维度大于 70 分并经专业人员访谈不合格的 274 名定为不合格组。从精神病院选取治疗缓解期男性精神分裂症患者 221 人定为精神病组,平均年龄为 20.72 ± 3.60 岁。

2.3 统计方法

单维性检验运用 SPSS 11.5 软件进行主成分分析,项目反应理论模型采用双参数 Logistic 模型(2PL)。将三组被试放在一起运用 BILOG-MG 软件进行参数分析,采用适用于大样本的边际极大似然估计(MML),估出条目的区分度参数(a)和位置参数(b)。根据项目反应理论,a 太小说明题目对于被试的能力估计提供的信息量太少,但如果 a 太大也会对结果造成影响。因此结合结果,将 $a \leq 0.3$ 或 $a \geq 4$ 的条目删除,重新估计各分量表的参数及被试的能力水平。

3 结果

3.1 单维性检验

应用 SPSS 11.5 对数据进行主成分分析,结果发现各量表第一因子的负荷量都接近 20% 且超过第二因子 3 倍以上(见表 1),且第一因子拐点明显,由此可以认为数据满足单维性的要求^[15]。

表 1 各量表主要因子特征值

分量表	第一负荷量	第二负荷量	第一负荷量/第二负荷量
Dit	20.12	3.73	5.39
Net	24.02	4.16	5.77
Set	19.23	5.11	3.76
D	19.90	5.37	3.71
S	18.60	4.42	4.21
T	24.09	7.81	3.08

3.2 应用双参数 Logistic 模型(2PL)对数据进行分析

3.2.1 Dit 在 Dit 分量表 75 个条目中有 2 个条目的初始 $a < -0.15$,有 2 个条目参数分布异常而未估出,有 7 个条目 $a < 0.30$,予以删除。然后对剩下的条目进行重新估计,a 的均值为 0.73。在 $[-3.00, 3.00]$ 的全距中,正常被试组的能力值平均为 -0.72,访谈不合格组的能力值平均为 0.94,精神病组的能力值平均为 0.01。删题前在特质水平为 0.60 时信息量最大为 25.25,删题后在特质水平为 0.60 时信息量最大为 24.67(见图 1 和 2)。

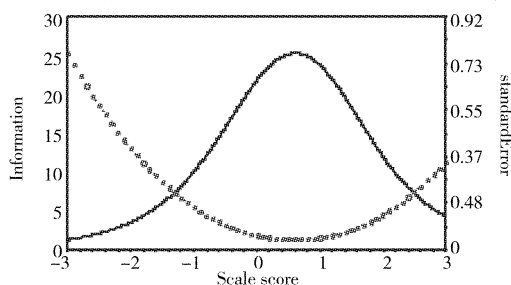


图 1 Dit 删题前测验信息曲线

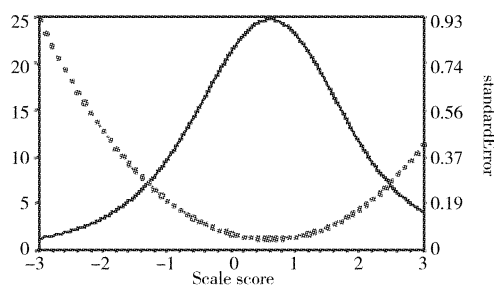


图 2 Dit 删题后测验信息曲线

3.2.2 Net 在 Net 分量表 48 个条目中有 2 个条目参数分布异常而未估出,有 3 个条目 $a < 0.30$,予以删除。然后再对剩下的条目进行重新估计,结果发现 a 的均值为 0.89。在 $[-3.00, 3.00]$ 的全距中,正常被试组的能力值平均为 -0.53,访谈不合格组的能力值平均为 0.95,精神病组的能力值平均为 0.01。删题前在特质水平为 0.30 时信息量最大为 26.17,删题后在特质水平为 0.30 时信息量最大为 26.00(见图 3 和 4)。

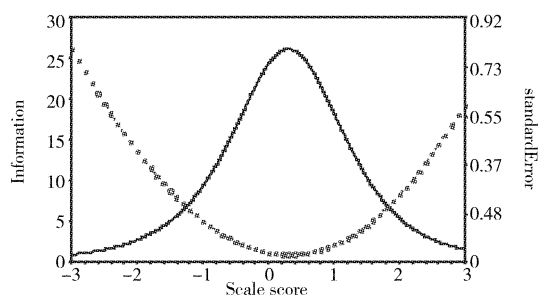


图 3 Net 删题前测验信息曲线

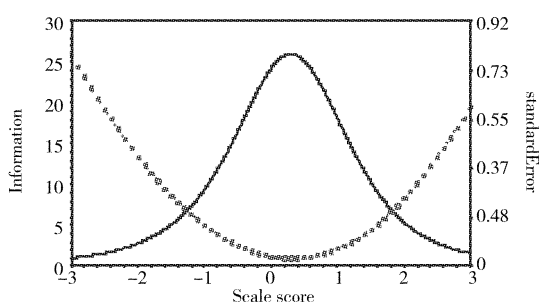


图 4 Net 删题后测验信息曲线

3.2.3 Set 在 Set 分量表 40 个条目中有 6 个条目的初始 $a < -0.15$,有 5 个条目参数分布异常而未估出,有 3 个条目 $a < 0.30$,予以删除。然后对剩下的条目进行重新估计, a 的均值为 1.57。在 $[-3.00, 3.00]$ 的全距中,正常被试组的能力值平均为 -0.63,访谈不合格组的能力值平均为 0.81,精神病组的能力值平均为 0.01。删题前在特质水平为 1.00 时信息量最大为 16.00,删题后在特质水平为 1.00 时信息量最大为 16.32(见图 5 和 6)。

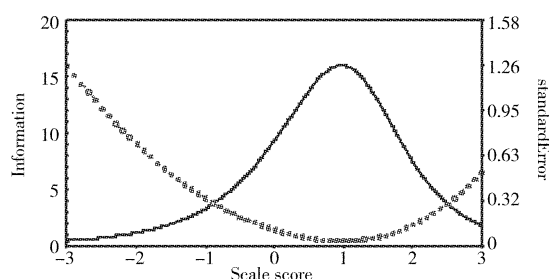


图 5 Set 删题前测验信息曲线

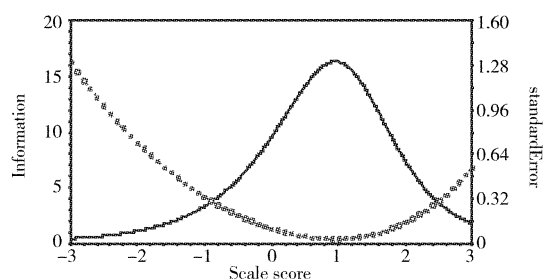


图 6 Set 删题后测验信息曲线

3.2.4 D D 分量表 29 个条目的 a 都在适当范围内, a 的均值为 1.58。在 $[-3.00, 3.00]$ 的全距中,正常被试组的能力值平均为 0.70,访谈不合格组的能力值平均为 0.10,精神病组的能力值平均为 0.19。在特质水平 0.60 处信息量最大为 59.16(见图 7)。

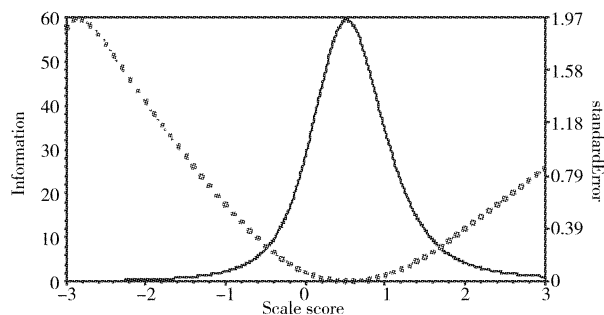


图 7 D 测验信息曲线

3.2.5 S 在 S 分量表 59 个条目中有 3 个条目参数分布异常而未估出,有 1 个条目 $a < 0.30$,予以删除。然后对剩下的条目进行重新估计, a 的均值为 1.04。在 $[-3.00, 3.00]$ 的全距中,正常被试组的能力值平均为 -0.89,访谈不合格组的能力值平均为 0.81,精神病组的能力值平均为 0.01。删题前在特质水平为 1.20 时信息量最大为 16.74,删题后在特质水平为 1.20 时信息量最大为 17.19(见图 8 和 9)。

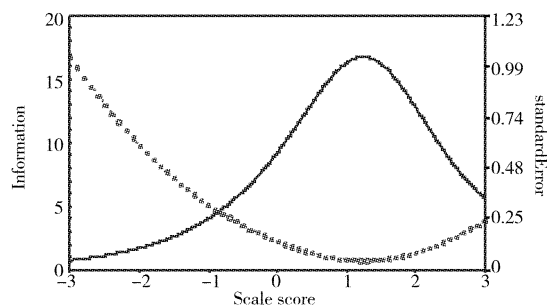


图 8 S 删题前测验信息曲线

3.2.6 T 在 T 分量表 15 个条目中有 4 个条目的 $a > 4$,予以删除。然后对剩下的条目进行重新估计,

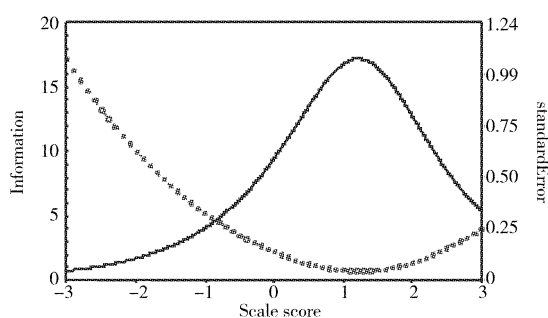


图9 S删题后测验信息曲线

a 的均值为 1.84。在 $[-3.00, 3.00]$ 的全距中, 正常被试组的能力值平均为 0.79, 访谈不合格组的能力值平均为 0.15, 精神病组的能力值平均为 0.05。删题前在特质水平为 0.70 时信息量最大为 34.88, 删题后在特质水平为 1.1 时信息量最大为 9.19 (见图 10 和 11)。

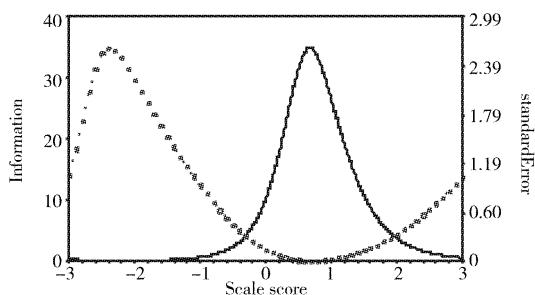


图10 T删题前测验信息曲线

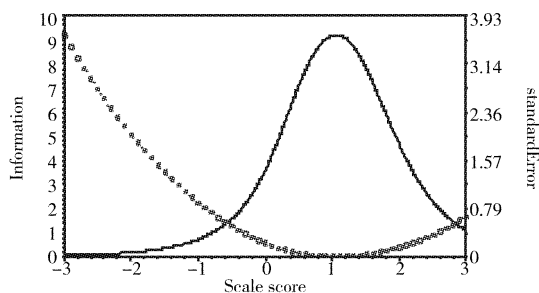


图11 T删题后测验信息曲线

3.3 原始计分与应用项目反应理论估计被试得分之间的相关分析

将 $a \leq 0.3$ 或 $a \geq 4$ 的条目删掉前后, 应用 BILOG-MG 软件对被试能力值进行估计, 然后应用 SPSS11.5 计算能力估计值与原始分数的相关, 结果发现各分量表删题前后估得的分数与原始分均有显著相关, 且相关系数变化不大。Dit 分量表删题前后相关系数分别为 0.96、0.96; Net 分量表删题前后相关系数分别为 0.97、0.96; Set 分量表删题前后相关系数分别为 0.82、0.82; D 分量表相关系数为 0.96;

S 分量表删题前后相关系数分别为 0.97、0.95; T 分量表删题前后相关系数分别为 0.96、0.93。

3.4 三组人群在两种计分方法中的分布比较

将三个临床量表 (Dit、Net、Set) 原始分数转换为符合标准正态分布的 z 分数, 同时将删题后用 IRT 估得的分数也转换到标准正态分布上, 比较三组人群分数分布的变化, 结果见图 12。

4 讨论与结论

IRT 作为新兴的测验理论相对于 CTT 具有诸多优势, 但由于各种原因在人格测验中应用尚未占据主要地位。许多问题还有待进一步研究, 包括模型的选择, 条目的增加与删除等。本研究应用 IRT 对 CSPQ 进行了分析, 并为 IRT 在人格测验中的应用提供了实证依据。

4.1 项目区分度参数及测验信息量

基于 IRT 的双参数 Logistic 模型有两个主要参数, 区分度参数 (a) 和位置参数 (在智力测验中称为难度参数 b)。项目所能提供的最大信息量与 a 的平方成正比, 因此 a 越大, 项目所能提供的最大信息量越大, a 越小项目所能提供的最大信息量越小。但当 a 过大时也会造成测验的偏差。在人格测验中尚未见到 a 选择范围的报道, 在本研究中结合研究结果采用 (0.3, 4) 的区间进行选择, 将超过此区间的条目进行删除。由结果部分可以看出删除 a 较小的条目对测验的信息量以及被试能力估计值的影响均较小, 但删除 a 过大的条目后测验的信息量明显减小 (如 T 分量表), 这一点可以通过以下项目信息量计算公式看出:

$$I_j(\theta) = \frac{a_j^2 e^{a_j(\theta - b_j)}}{[1 + e^{a_j(\theta - b_j)}]^2}$$

其中 $I_j(\theta)$ 表示第 j 个项目在能力值 θ 上的信息量, a_j 表示第 j 个项目的区分度参数, b_j 表示第 j 个项目的难度参数。

但删题前后能力估计值与原始分数的相关并未明显改变。说明删掉这些条目之后对于被试能力值的估计影响不大, 表明模型的稳健。但删题之后对项目的参数重新估计发现, 其他项目的参数稍有调整, 有些项目 a 有所增大, 使测验信息量有所变化, 如 Set 分量表中第 24 题在删题前 $a = 2.41$, 最大信息量为 1.44, 在删题后 $a = 2.44$, 最大信息量为 1.48 (如图 13)。由于测验信息量是项目信息量的累加和, 因此当增大的项目信息量不足以弥补删题损失的信息量时测验的信息量减小了 (如 Net、Dit、T、D

分量表),当增大的项目信息量超过删题损失的信息量时测验信息量就会有所增加(如 Set、S 分量表),表明测验有所改善。而且删掉这些条目能减

少施测时间,提高测验的效率,尤其在征兵这种大型人事测验中其重要性更可见一斑。

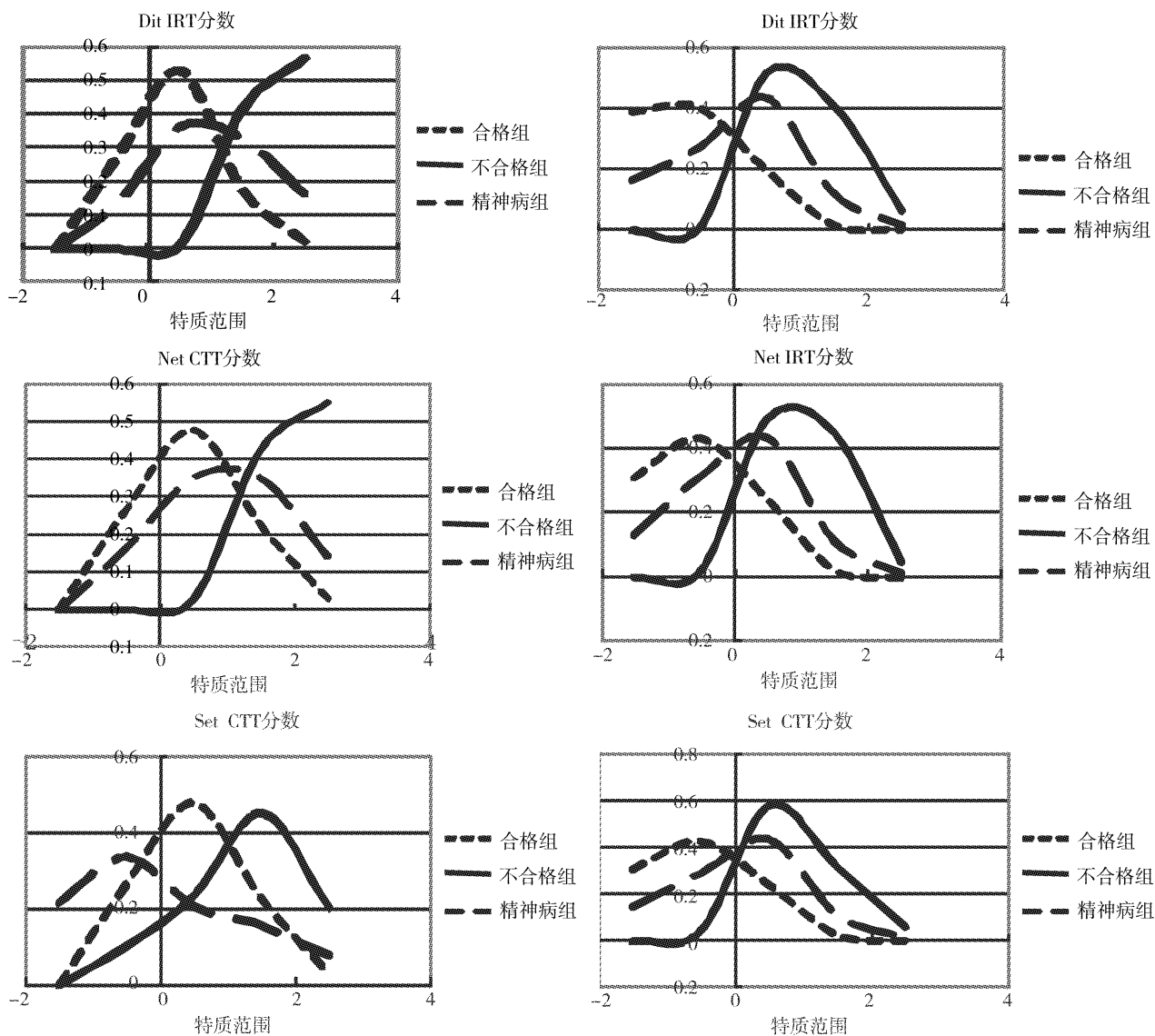


图 12 两种计分方法的比较

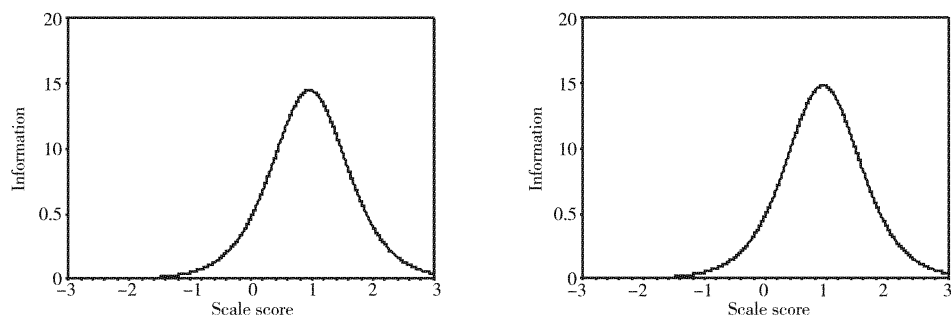


图 13 删题前后信息量变化图

4.2 被删条目分析

在本研究中,a 超出区间(0.30,4.00)的条目大多数为反向计分题,同 James 等的^[16]的研究结果相似。其中 T 是全部由反向计分题构成的,在删题前该量表在低能力范围的信息量很小,误差很大(如图 10),在删题后虽然测验的最大信息量有所下降,但信息函数在 $[-3.00, 3.00]$ 的全距中测验精度都得到了-定的保证,而且在特质水平稍高于平均水平处具有最大信息量,因此对于本测验的目标人群具有更理想的信度(如图 11)。这种情况的出现可能由于转换成反向计分以后被试对条目内容的理解会变得模糊;也可能从正面叙述内容会使条目在特质维度上的位置变得极端化;而且条目本身的尺度不明确也会对 a 造成影响;同时条目的计分方向可能也会对 a 的计算造成影响。正如郭庆科等对于人格测验中题目正反向陈述的研究所示,具体原因尚需要从测验心理机制上进行进一步探讨^[17]。

同国外的许多研究相似^[18,19],原有量表虽然在 CTT 中经过大量的研究验证了其条目的区分度,量表的结构效度、内容效度等指标在可接受范围内,但由于 IRT 运用不同的计算方法,将每个条目对测验所能提供的信息量进行了再区分,各个条目的计分因其参数的不同而不同,因而较 CTT 计分方法更科学、合理。但由于适用于人格测验的 IRT 模型还存在争议,而在不同模型中被删掉的题可能不同,因此本研究结合大量实证数据探讨了 IRT 应用于人格测验的可行性,为进一步的研究提供了实证基础。

4.3 两种计分方法下三组人群分数分布的比较

本研究中由于条件所限合格组和不合格组是使用计算机呈现进行测验,精神病组完成的是纸笔测验。但回顾 Roper 等的文献发现,其检测结果仅在个别维度上存在差异,在大多数情况下两种方式不会对测验结果造成影响^[20]。由图 12 可以看出,三个临床量表的得分在两种计分方法下,合格组与不合格组之间的分数差别没有明显的变化。但精神病组的分布变化较明显。在 CTT 分数下,精神病组大多与合格组相近,但 IRT 计分下,精神病组的分布曲线向不合格组移动,且三组的分布趋势也更加明显。在 IRT 计分下由于每个条目的 a 不同,对于估计被试能力水平的贡献不同,即信息量不同,而在 CTT 中每个条目的计分都相同,所以 IRT 计分能更好的区分出不同人群各自的特质水平,正如本研究的结果所示。

可见 IRT 作为新的测验理论在人格量表的分析

中是可以发挥其优势的,但人格测验自身的特点,如条目的相对固定,作答动机问题等,确实是在应用 IRT 时应该注意的方面。本研究运用实证数据对其进行了探讨,但由于条件的限制没能与更多的 IRT 模型进行对比,尚不能得出何种模型更适合人格测验的实际情况,这也是下一步研究的方向。而且在运用 BILOG - MG 分析人格数据的过程中,迭代收敛较慢,有些条目的模型资料拟合度也较差,说明双参数 Logistic 模型应用在人格测验数据中还存在-定的问题。但相对于 CTT 而言,IRT 能更有效地评定被试的特质水平,能提供条目及测验的信息量,对于量表的编制与修订是非常有帮助的。而且可以去-除原量表中信息量较小的条目,减少条目本身对被试的影响,在相似测验精度要求下所需条目更少,可提高测验效率。

参 考 文 献

- 1 Qi S Q, Dai H Q. Item Response Theory and its Applications, 1992. 14 ~ 15
(漆书青,戴海崎.项目反应理论及其应用研究,1992.14~15)
- 2 Qi S Q, Dai H Q, Ding S L. Principles of modern education and psychological measurement (in Chinese). Beijing higher education Publishing Press, 2002. 86 ~ 90
(漆书青,戴海崎,丁树良.现代教育与心理测量学原理.北京:高等教育出版社,2002.86~90)
- 3 Chernyshenko O S, Stark S, Chan K, Drasgow F, Williams B. Fitting item response theory models to two personality inventories: Issues and insights. Multivariate Behavioral Research, 2001, 36: 523 ~ 562
- 4 Harvey R J, Murry W D. Scoring the Myers - Briggs Type Indicator: Empirical comparison of preference score versus latent - trait analyses. Journal of Personality Assessment, 1994, 62: 116 ~ 129
- 5 Reise S P, Waller N G. Fitting the two - parameter model to personality data: The parameterization of the Multidimensionality Personality Questionnaire. Applied Psychological Measurement, 1990, (15): 45 ~ 58
- 6 Santor D A, Ramsay J O, Zuroff D C. Nonparametric item analysis of the Beck Depression Inventory: Evaluating gender item bias and response option weights. Psychological Assessment, 1994, (6): 255 ~ 270
- 7 Reiseand S, Henson J. A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales. Journal of Personality Assessment, 2003, 81(2): 93 ~ 103
- 8 Reise S P, Waller N G. How many IRT parameters does it take to model psychopathology items? Psychological Methods, 2003, (8): 164 ~ 184
- 9 Meijer R, Baneke J. Analyzing Psychopathology Items: A Case for Nonparametric Item Response Theory Modeling. Psychological Methods, 2004, (9): 354 ~ 368
- 10 Thurstone L L. A law of comparative judgment. Psychological Review, 1927, 34: 273 ~ 286
- 11 Stark S, Chernyshenko O. Examining Assumptions About Item

- Responding in Personality Assessment: Should Ideal Point Methods Be Considered for Scale Development and Scoring? *Journal of Applied Psychology*, 2006, 91: 25 ~ 39
- 12 Zhu N N, Zhang H C. The Comparison Between CTT and IRT in Results of Personality Test, *Psychological Exploration*, 2003, 23: 48 ~ 51
(朱宁宁, 张厚粲. CTT 与 IRT 方法对人格测验结果处理的比较研究. *心理学探新*, 2003, 23: 48 ~ 51)
- 13 Guo Q K, Chen Y M, Meng Q M. Applicability of Item Response Theory in Self-report Questionnaire. *Acta Psychologica Sinica*, 2005, 37(2): 275 ~ 279
(郭庆科, 陈英敏, 孟庆茂. 自陈量表式测验应用 IRT 的可行性. *心理学报*, 2005, 37(2): 275 ~ 279)
- 14 Xiao L J, Miao D M, Xiao W, et al. Personality Measurement for Recruits Selection. *Acta Psychologica Sinica*, 2007, 39(2): 200 ~ 210
(肖利军, 苗丹民, 肖玮, 武圣君, 李红政. 应征公民心理选拔的人格评估. *心理学报*, 2007, 39(2): 200 ~ 210)
- 15 Linden W, Hambleton R K. *Handbook of Modern Item Response Theory*. Springer - Verlag New York Inc, 1997
- 16 Stansbury J, Ried D. Unidimensionality and Bandwidth in the Center for Epidemiologic Studies Depression (CES - D) Scale. *Journal of Personality Assessment*, 2006, 86: 10 ~ 22
- 17 Guo Q K, Han D, Wang Z, Shi K. Effects of Positively and Negatively Worded Items in Personality Measurement. *Acta Psychologica Sinica*, 2006, 38(4): 626 ~ 632
(郭庆科, 韩丹, 王昭, 时勤. 人格测验中题目正反向陈述的效应. *心理学报*, 2007, 38(4): 626 ~ 632)
- 18 Ferrando P J. The accuracy of the E, N and P trait estimates: an empirical study using the EPQ - R. *Personality and Individual Differences*, 2003, 34: 665 ~ 679
- 19 Ferrando P J. Person Reliability in Personality Measurement: An Item Response Theory Analysis. *Applied Psychological Measurement*, 2004, 28: 126 ~ 140
- 20 Roper B L, Ben - Porath Y S, Butcher J N. Comparability and validity of computerized adaptive testing with the MMPI2. *Journal of Personality Assessment*, 1995, 65: 358 ~ 371

Item Analysis of the Chinese Soldier Personality Questionnaire Using Item Response Theory

YANG Ye-Bing, MIAO Dan-Min, TIAN Jian-Quan, XIAO Li-Jun, SUN Han, HONG Xia
(*Department of Psychology, School of Aerospace Medicine, Fourth Military Medical University, Xi'an 710032, China*)

Abstract

With the development of psychological and educational measurement, item response theory (IRT) is being increasingly used in psychological and educational testing areas. There have been many large - scale tests based on IRT, such as TOEFL and GRE. Further, there have also been many studies on the applications of IRT in personality questionnaires. However, many problems still remain, including model selection and determining the appropriate numbers of parameters. It is vital to conduct empirical studies in order to discuss the probability of applying IRT to this area. The Chinese Soldier Personality Questionnaire (CSPQ) was used in this study in order to discuss the possibility of applying IRT to personality questionnaires with large samples.

In the study, 100523 young individuals were tested using the computerized CSPQ. Subjects whose standard score in every subtest was less than 70 were considered to be normal, while the others were disqualified. Further, structural interviews were conducted by psychological experts in order to confirm the symptoms of those disqualified. If they indeed exhibited some symptoms, they were placed in the disqualified group. Additionally, 221 schizophrenia patients were tested using the same questionnaire, and they were placed in the schizophrenia group. From among the normal subjects, 2676 were randomly selected and placed in the normal group. The SPSS 11.5 software was used to test the unidimensionality of every subtest. Moreover, the BILOG - MG software was used to analyze the data with the two - parameter logistic model (2PL) based on IRT.

Items with discriminating parameters beyond the range of (0.3, 4) were deleted as they provided little information that was relevant to testing. Moreover, before and after deleting these items, there was no noticeable change in the maximum test information and the correlations between IRT scores and standard scores. In the case of the standard scores, the distributions of the schizophrenia group were similar to those of the normal group, while in the case of the IRT scores, the distributions of the schizophrenia group moved towards those of the disqualified group.

As shown in this study, IRT can improve the efficiency of testing with nearly the same precision. Further, IRT is more suitable for identifying the differences between varied groups, which is another advantage of IRT analysis. Therefore, as a new testing theory, IRT can and should be used in the analysis of personality questionnaires. However, personality questionnaires are not the same as cognitive tests, and different models of IRT (for example, nonparametric models) have not been compared in this study. Therefore, the best model for personality questionnaires should be discussed further.

Key words personality testing, item information function, two - parameter logistic model.