IRT 展开模型及对非累积反应机制的检测

郭庆科[1],苗金凤[2],王 昭[1]

(1.辽宁师范大学 心理学系,大连 116029;2.辽宁师范大学 海华学院,大连 116029)

摘 要:被试回答人格测验题目时并不是特质水平越高其得分率越高,这称为非累积反应机制。广义等级展开模型 GGUM 就是针对这一机制提出来的。使用 EPQ 和五因素人格问卷发现 GGUM 比累积 IRT 模型有更好的模型拟合度和测量精度。研究结果表明 GGUM 有其合理性,且有助于反应心理过程机制的深入探讨。

关键词:非累积 IRT 模型;广义等级展开模型 GGUM;模型拟合

中图分类号:B841.2 文献标识码:A 文章编号:1003 - 5184(2006)01 - 0066 - 04

1 前言

1.1 非累积反应机制与 IRT 展开模型

1980 年以来人格测量呈现了复兴的趋势,原因是人格测验在人才选拔中有不可替代的作用[1]。人格测量中能否采用项目反应理论 IRT 就很值得研究。这一问题的关键是 IRT 模型能否拟合数据。在人格测量中也假设被试的特质水平越高则其在测题上的得分也越高,或做出肯定反应的概率越大。大多数测量模型都采用了这种假设,包括经典真分数模型、因素分析模型,IRT 的两参数 Logistic 模型 2PLM、Samejima 等级反应模型 SGRM 等。这种假设在非认知测验中就不一定成立了。

Thurstone (1959)最早对这两种测量模型做了区分[2]。前一种模型体现了一种累积反应机制(cumulative mechanism),即当项目难度与被试的心理特质表示在同一量表连续体上时,则特质水平越高于项目特性的被试在该项目上得高分的概率就越大。这类测量模型可称为累积模型或优势(dominance)模型。但在态度等非认知类测量中这一模型就不一定适用。态度强的人对一个中等态度强度的项目做出肯定反应的概率可能低于一个中等态度强度的人。Thurstone 的学生 Coombs(1964)进一步指出了态度测量中的单峰(single model)特征。即若一项目是测量单维度态度的,且将态度语的强度与被试态度表示在同一量尺上时,则随着被试态度强度的升高,他们对项目肯定回答的概率也增大,当被试态度强度与态度语的强度相同时,这一概率达到最大值,这一点称为理想点(ideal point)[3],超过这一理想点后,随着被试态度强度的提高其对该项目的肯定反应概率反而会逐渐下降。比如有一个关于死刑态度的项目"死刑很可怕,但对那些令人发指的犯罪分子却是必需的",对这一项目做肯定回答的往往是对死刑持中等赞成态度的人,而不是持极赞成和极反对态度的人。

比如对上面关于死刑的题目如果选项有赞成和反对两个,在非累积模型中则隐含 4 条函数曲线,一条代表被试态度比项目所表达的强度弱而他赞同这一项目(agree from below)的概率,一条代表被试实际态度更弱而他反对这一项目(disagree from below)的概率,另两条是赞同而实际态度更强(agree from above)和反对而实际态度更强(disagree from above),4 条曲线可合并为赞同和反对两条非单调曲线。由于在非累积 IRT 模型中对表达两种态度的概率进行分解,故称为展开模型(unfolding model)。展开模型所揭示的反应机制不仅在态度测量中存在,在人格测量中也会有同样的现象,用累积 IRT 模型分析显然不适当。

展开模型则是在近年来才受到重视[3~4]。Roberts 提出的广义等级展开模型(generalized graded unfolding model,GGUM)是为多级评分资料设计的,同时也可用于分析两级评分资料[5~6]。Roberts 编制的计算机程序 GGUM(目前版本为 2004)可用于对展开模型进行参数估计,研究表明 GGUM 的估计结果是可靠的[7~8]。

1.2 IRT 中模型 - 数据的拟合检验

IRT 中模型 - 数据的拟合检验方法主要有两种,一是统计法,二是拟合图分析法(Drasgow 等,1995)[9],项目拟合 χ^2 检验则是最常用的统计检验

方法。但 χ^2 统计量对样本量敏感且检测不出特定条件下的不拟合。为克服这一缺点人们又提出了调整的 χ^2。不管实际样本大小,调整的 χ^2 都以 3000 人为期望样本基准,摆脱了对样本量的依赖,调整的 χ^2 与自由度的比在 3 以上的项目就认为是不拟合的。Wollenberg (1982)发现单个项目的 χ^2 检验对单维性的违反不敏感,进而提出了两项目对(pairs)和三项目组(triples)χ^2 检验。两项目对和三项目组对模型与数据的歪曲起放大作用,可检测出异常反应模式[9]。

Chernyshenko 等(2001)对第 5 版 16PF 和 50 题的大五量表的研究发现,IRT 模型在人格测量中的拟合性要远远差于认知测量。其中 2PLM 拟合最好,但也有相当多的项目不拟合,SGRM 拟合最差,所有分量表中均有不拟合的现象。Chernyshenko 认为这是因为人格测量的题目也是让被试表达出自己的态度,也存在与态度测量中相同的非累积的反应机制,因此用展开模型可能更适合[10]。

展开模型的提出为研究被试反应的内部心理机制提供了很好的方法,符合心理测量的发展方向。本研究的目的是探讨 GGUM 是否与数据拟合更好,是否有更好的心理测量学性能等。在此基础上探讨如何提高人格测量的有效性。2PLM 被证明适合分析 2 级评分的自陈量表式资料,SGRM 则适合分析等级反应资料[11],因此本研究结果很有代表性。

2　研究方法

研究使用了五因素问卷 NEO – FFI(以下简称 FFI)和艾森克人格问卷 EPQ 两个有代表性的人格量表。FFI 是大五人格测验 NEO – PI 的简化版。测量神经质 N、外倾性 E、开放性 O、宜人性 A 和认真性 C 五项特质,每个维度 12 题,采用 5 级评分[12]。EPQ 有 3 个人格量表(即内 – 外向 E、情绪性 N、精神质 P),研究中使用了 EPQ 中除说谎量表 L 量表外的 65 道题目。

辽宁师范大学本科生 491 人(男 126,女 365)和辽宁朝阳二中 500 名高中生(男 253,女 247)参加了 EPQ 测试,辽师大学生 514 人(男 121,女 393)和朝阳二中 503 名高中生(男 228 人,女 275 人)参加了 FFI 测试。

参数估计软件为 BILOG[13],MULTILOG[14] 和 GGUM2004。

3　结果

3.1　模型 – 数据拟合度对比

χ^2 统计量对样本量敏感,因此检验时使用了调整的 χ^2 与自由度的比值。这一比值在 3 以上时就说明项目拟合不好,在 5 以上时说明不拟合相当严重。EPQ 和 FFI 的拟合检验结果见表 1。表中列出了各分量表项目调整的 χ^2 与自由度的比值的平均数与标准差,及这一比值分布在 3 以上的项目数。

表 1　各分量表项目调整 χ^2 与自由度的比值的平均数与标准差

		GGUM			2PLM 或 SGRM		
		平均数	标准差	χ^2 大于 3 的次数	平均数	标准差	χ^2 大于 3 的次数
EPQ 的 E 量表	Singles	0	0	0	0.06	0.27	0
	Doubles	0.49	1.29	2	0.60	1.45	2
	Triples	0.65	1.62	1	0.69	1.25	1
EPQ 的 N 量表	Singles	0.64	1.37	2	0	0	0
	Doubles	0.71	1.25	3	0.47	1.25	2
	Triples	0.76	1.07	0	0.56	1.01	0
EPQ 的 P 量表	Singles	0	0	0	0.32	1.23	1
	Doubles	0.17	0.45	0	1.22	2.82	3
	Triples	0.37	0.62	0	2.02	2.55	3
EFI 的 N 量表	Singles	0.06	0.19	0	1.95	0.66	2
	Doubles	5.59	2.73	9	8.65	2.36	11

	Triples	4.12	1.53	3	7.43	2.87	3
EFI 的 E 量表	Singles	0	0	0	2.05	0.57	1
	Doubles	6.37	3.98	9	9.98	4.08	12
	Triples	3.22	1.98	2	8.55	2.66	4
EFI 的 O 量表	Singles	0.59	1.42	1	1.98	1.63	2
	Doubles	8.62	13.00	7	12.75	3.87	12
	Triples	8.17	9.47	3	10.22	4.41	4
EFI 的 A 量表	Singles	0	0	0	1.28	0.57	1
	Doubles	5.34	2.36	10	8.07	3.24	11
	Triples	4.22	1.74	4	6.98	2.84	4
EFI 的 C 量表	Singles	0	0	0	1.84	1.72	1
	Doubles	5.24	3.37	8	7.08	5.47	8
	Triples	3.90	1.17	3	6.35	3.42	4

注：Singles 指单项目，Doubles 指两项目对，Triples 指三项目组

从表中可见无论与 2PLM 相比还是与 SGRM 相比，GGUM 都是拟合最好的。2PLM 基本不存在大的拟合问题，SGRM 则问题较多。SGRM 的问题突出表现为两项目对和三项目组的不拟合，说明项目间的关系没有被 SGRM 充分地解释，这一问题在 GGUM 中得到了改善。

3.2 测量精度对比

模型有效性的另一个衡量标准是模型的测量精度，即 GGUM 与 2PLM 和 SGRM 相比是否提供更大的信息量。表 2 中列出了 3 个模型在 $\theta=-2$，$\theta=-1$，$\theta=0$，$\theta=1$，$\theta=2$ 五个能力点上的平均信息量，从中可以看出模型的测量精度。

表 2 3 个模型五个能力点上的平均信息量

	EPQ 分量表				FFI 分量表			
	E	N	P	N	E	O	A	C
GGUM	3.63	6.05	2.66	7.76	3.02	3.21	4.10	2.67
2PLM	2.51	2.25	1.39					
SGRM				4.71	3.87	2.36	2.61	3.11

从表中可以看出 GGUM 在多数情况下比累积 IRT 模型即 2PLM 和 SGRM 提供更多的信息量（只有 FFI 中的 E 量表和 C 量表除外）。由于 GGUM 引入了更复杂的数学函数，充分利用了项目的信息，使得对能力的估计更为准确。

3.3 效标关联效度的对比

用同伴提名法搜集了 EPQ 和 FFI 的效标。向参加测试的大学生详细讲解 EPQ 和 FFI 各因素的含义，然后让他们从自己班里（测试是以班级为单位的）选出与各因素高分特征最相近的三个人。然后统计出被试者被提名的次数。研究中将被提名的次数低于 2 次者编码为 0，被提名的次数在 2 次或 2 次以上者编码为 1。再计算与 GGUM、2PLM、SGRM 能力估计值的点二列相关。结果见表 3。

表 3 提名次数与 EPQ 和 FFI 的点二列相关

	EPQ（$N=290$）			FFI（$N=221$）			
	E 量表	P 量表	N 量表	N 量表	E 量表	O 量表	A 量表	C 量表
2PLM	0.36**	0.22**	0.05	0.09	0.25*	0.08	0.13	0.14*
GGUM	0.35**	0.21**	0.05	0.07	0.32**	0.13	0.09	0.16*

可以看出用三种 IRT 模型分析时两人格量表的效标关联效度基本没有差异。另外，用 GGUM 估计出的 E、N、P 三分量表能力值与这三个量表 CTT 总分的相关为 0.914、0.590、0.473，平均数低于用 2PLM 估计

的能力值与 CTT 总分的相关(分别为0.921、0.865、0.632)。同样用 GGUM 估计出的 N、E、O、A、C 五个分量表能力值与 CTT 总分的相关为 0.794、0.560、0.952、0.319、0.961,也低于 SGRM 与 CTT 总分的相关(分别为 0.935、0.912、0.904、0.886、0.911)。

4　讨论

Chernyshenko 等(2001)对人格测量中的模型拟合问题提出了两点建议,一是采用更复杂的模型,如 Levine(1984)的极大似然公式记分模型 (MFS),MFS 对项目反应函数的形式没有特殊的规定,因此适应性更强。二是采用非累积 IRT 模型[10]。本研究证实了 Chernyshenko 的假设,即使用 GGUM 改善了人格测验的拟合度和测量精度。结果证实了 GGUM 的合理性,同时也说明人格测验中的非累积反应机制是存在的。虽然没有改善测量的效度,但由于 GGUM 改善了测量精度,它在项目功能差异分析和测验等值中会更有优势。

以往研究和本研究都发现有的量表使用累积 IRT 模型也是可行的,说明被试对有些题目的反应符合累积反应机制。其中的道理可能是:这些题目是更为极端的题目,比如项目"我喜欢热闹的聚会"就比"我喜欢和朋友聊天"代表了更强的外向特质。而被试对极端项目的回答是服从累积反应机制的。比如对"我喜欢热闹的聚会"这一项目只存在随着外向程度的提高而赞同概率越大这一情况,其他情况都不会发生(即概率是 0)。

但只使用极端的题目并不符合心理测量学原理。极端的题目选答率不高,其区分度会高吗?中等特质水平的被试者能否得到有效的测量?而通常的认识是中等难度的题目要多。但中等选答率的题目又可能存在非累积的反应机制,使题目得分不能很好地体现被试特质水平。这些问题如能得到解决将非常有利于人格测量的发展。

研究初步证明人格测量中可能存在非累积反应机制。但 GGUM 也存在不拟合的现象,而且并没有改善测量的效度,这表明 GGUM 也没有完全揭示人格测量的反应机制。以上这些问题都需要进一步深入研究。

参考文献

1　Barrick M R, Mount M K. The Big Five personality dimensions and job performance: A meta – analysis. Personnel Psychology, 1991, 44: 1 – 26.

2　Andrich D. A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. British Journal of Mathematical and Statistical Psychology, 1996, 49: 347 – 365.

3　Andrich D. the application of an unfolding model of the PIRT type to the measurement of attitude. Applied Psychological Measurement, 1998, 12: 33 – 51.

4　Luo G. A. joint maximum likelihood estimation procedure for the hyperbolic cosine model for single – stimulus responses. Applied Psychological Measurement, 2000, 24: 33 – 49.

5　Roberts J S, Donoghue J R, Laughlin J E. A general item response theory model for unfolding unidimensional polytomous responses. Applied Psychological Measurement, 2000, 24(1): 3 – 32.

6　Roberts J S. GGUM2000: Estimation of parameters in the generalized graded unfolding model. Applied Psychological Measurement, 2001: 25, 38.

7　Roberts J S, Donoghue J R, Laughlin J E. Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. Applied Psychological Measurement, 2002, 26: 192 – 207.

8　DeMars C E. Type I error rates for the generalized graded unfolding model fit indices. Applied Psychological Measurement, 2004, 28: 48 – 71.

9　Drasgow F, Levine M V, Tsien S, Williams B A, Mead A D. Fitting polytomous IRT models to multiple – choice tests. Applied Psychological Measurement, 1995, 19: 143 – 165.

10　Chernyshenko O S, Stark S, Chan K Y, Drasgow K, Williams B. Fitting Item Response Theory Models to Two Personality Inventories: Issues and Insights. Multivariate Behavioral Research, 2001, 36 (4): 523 – 562.

11　郭庆科,周晶. Likert 量表分析中不同 IRT 模型的有效性.心理学探新,2004,24(3): 67 – 70.

12　许淑莲,吴志平,吴振云等.成年人心理幸福感的年龄差异研究.中国心理卫生杂志,2003,17(3): 147 – 151.

13　Mislevy R, Bock R D. PC BILOG 3: Item analysis and test scoring with binary logistic models (2nd. ed.). Chicago IL: Scientific Software, Inc, 1990.

14　Thissen D. MULTILOG user's guide (Version 6.0). Mooresville, IN: Scientific Software, 1991.

(下转第 78 页)

3 Reise S P, Widaman K F, Pugh R H. Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance. Psychological Bulletin, 1993, 114(3), 552 – 566.

4 Vadenberg R J, Lance C E. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. organizational research. Organizational Research Methods, 2000, 3, 4 – 69.

5 Byrne B M, Shavelson R J. Adolescent self – concept: Testing the assumption of equivalence structural across gender. American Educational Research Journal, 1987, 24, 365 – 385.

6 Byrne B M. Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum. 1998.

7 Flowers C P, Oshima T C, Raju N S. A description and demonstration of the polytomous – DFIT framework. Applied Psychological Measurement, 1999, 23(4), 309 – 326.

8 Raju N S, van der Linden W, Fleer P. An IRT – based internal measurement of test bias with applications for differential item funtioning. Applied Psychological Measurement, 1995, 19, 353 – 368.

9 Muraki E, Bock R D. PARSCALE: IRT based test scoring and item analysis for graded open – ended exercises and performance tasks. Chicago, IL: Scientific Software. 1997.

10 International. Raju, Backer F B. EQUATE 2.1: Computer program for equating two metrics in item response theory. [Compter program]. Madison: University of Wisconsin, Laboratory of Experimental Design. 1995.

11 N. S. DFIT5P: A Fortran program for calculating DIF/DTF [Compter program]. Chicago: Illinois Institute of Technology. 1999.

12 漆书青,等. 现代教育与心理测量学. 南昌: 江西教育出版社. 1998. 278.

Differential Item Functioning: A Comparison of Methods Based on CFA and IRT

Luo Fang, Zhang Houcan

(Psychology College, Beijing Normal University, Beijing 100875)

Abstract: Currently, there are several methods for identify Differential Item Functioning, which are in confirmatory factor analysis and item response theory fields. The major purpose of this article is to offer a comparison of these two methods with a special emphasis on their methodological similarities and differences, for polytomous unidimensional case.

key words: confirmatory factor analysis; item response theory; Differential Item Functioning

(上接第 69 页)

Unfolding IRT Model and the Non – cumulative Response Mechanism in Personality Tests

Guo Qingke[1], Miao Jinfeng[2], Wang Weili[1]

(1. Department of Psychology, Laoning Normal University, Dalian 116029;
2. HaiHua College, Liaoning Normal University, Dalian 116029)

Abstract: People with high trait level will not necessarily get high scores when responding to non – cognitive items, this is called non – cumulative response mechanism. Generalized Graded Unfolding Model (GGUM) is an IRT model developed to solve this problem in personality. In this study EPQ and NEO – FFI were administered to 991 and 1017 students, the results show that GGUM can fit the data better and provide more information than cumulative IRT models (2PLM and SGRM). The study also suggests that the issues of model fit and response mechanism in personality should be further studied.

Key words: Non – cumulative IRT Model; Generalized Graded Unfolding Model; Model – Data Fit