心理学探新 2022,Vol.42,No.3,269-278
PSYCHOLOGICAL EXPLORATION

融合反应时的多级评分 IRT 模型开发及其应用研究

汪大勋[1],郭莹莹[2]
(1.江西师范大学,南昌 330022;2.海亮教育科技集团,杭州 310052)

摘 要:当前大多数融合反应时的 IRT 模型仅适用于 0-1 评分数据资料,极大的限制了 IRT 反应时模型在实际中的应用。本文在传统的二级计分反应时 IRT 模型基础上,拟开发一种多级评分反应时模型。在层次建模框架下,分别采用拓广分部评分模型(GPCM)和对数正态模型构建融合反应时的多级评分 IRT 模型(本文记为 JRT-GPCM),并采用全息贝叶斯 MCMC 算法实现新模型的参数估计。为验证新开发的 JRT-GPCM 模型的可行性及其在实践中的应用,本文开展了两项研究:研究 1 为模拟实验研究,研究 2 为新模型在大五人格-神经质分量表中的应用。研究 1 结果表明,JRT-GPCM 模型的估计精度较高,且具有较好的稳健性。研究 2 表明,被试的潜在特质与作答速度具有一定的正相关,且本研究结果支持 Ferrando 和 Lorenzo-Seva(2007)提出的"距离-困难度假设",即当被试的潜在特质与项目的难度阈限距离越远,那么被试会花费更多的时间对项目进行作答。总之,本研究为拓展反应时信息在心理测量及教育中的应用提供新的方法支持。

关键词:项目反应理论;GPCM 模型;JRT-GPCM 模型;MCMC 算法
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2022)03-0269-10

1 引言

反应时(response Time,RT)是指测验实施中,个体完成每个任务或题目所花费的时间。在传统的心理与教育测验领域,多数任务均采取传统的纸笔作答形式,反应时这一重要的信息很难收集。但随着计算机测试技术的发展,收集反应时已经成为许多大型测试的常规工作。例如,国际学生评估(PISA)从 2012 年开始使用计算机测试并记录反应时数据。反应时可以显示被试的作答速度信息,比如当被试在低风险测试中动机较低时,他们可能以更加快速的方式对项目做出反应(Entink,van der Linden,& Fox,2009;Locke,1965;Logan,Medford,& Hughes,2011),或者有知识经验的被试花费反应时间更少(Qian,Staniewska,Reckase,& Woo,2016)。若只是关注被试的项目作答反应信息,就会丢失反应时信息因此将反应时的信息作为附加信息与项目反应信息联合使用,可以得到对被试参数更加准确的估计(Zhan,2017)。

长期以来,人们一直对作答速度和精度之间的关系很感兴趣,测量领域已经研究了几十年(Entink,Fox,& van der Linden,2009;Luce,1986;van der Linden,2006,2007;Wang & Hanson,2005)。总的来说,在速度和准确度之间的权衡反映了被试的作答速度和准确率之间存在负相关的关系,这表明在测验过程中,有些个体倾向花较多时间提高作答准确率,有些个体倾向于快速反应,这将会降低准确率(Fox & Marianti,2016)。而在实际测验中,个体在时间限制下,通常会选择以稳定的速度对整个测验进行作答(van der Linden,2007c)。

在 RT 建模方法中(Lee & Chen,2011;van der Linden,2009),van der Linden(2009)提出的层次建模框架下解释反应速度与精度之间的关系是最灵活的建模思路之一。相关研究表明(Suh,2010),与其他反应时建模方法相比,层次框架模型无论是在模拟数据还是实际数据中都能得到更好的结果。在这个框架中,反应时和项目作答反应分别在第一层建模,被试速度和能力参数、基于项目反应的项目参数和基于项目反应时的项目参数之间的方差协方差在更高的一级建模。该框架充分地对以往反应时建模进行了推广,将已有的项目反应和反应时模型采用"即插即用"的方法插入层次框架模型(Fox & Marianti,2016;Klein Entink,Fox,et al.,2009;Klein Entink,van der Linden,et al.,2009;Meng,Tao,& Chang,2015;Molenaar,Tuerlinckx,& van der Maas,2015;Wang,Chang,& Douglas,2013;Wang & Xu,2015)。

然而,目前国内外几乎所有融合反应时的 IRT 的建模研究都是基于 0~1 评分作答数据开展的。

但在实际测验中，测验资料往往是丰富多样，既有 0～1 评分的数据资料，也有多级评分的数据资料（如心理测验中常用的 Likert 量表，教育测验中的开放型试题），还有 0～1 评分和多级评分混合的数据资料，这时传统的仅适用于 0～1 评分的反应时 IRT 模型就不适用了，因此开发研究适用于多级评分的反应时 IRT 模型显得十分必要。

鉴此，本研究拟采用一种同时利用多级计分项目反应和反应时的联合建模方式，开发融合反应时的多级评分 IRT 模型，一方面弥补国内外这一领域研究不足，另一方面为教育与心理测验充分利用反应时信息提供新的方法。

2 融合反应时的多级评 IRT 模型开发

2.1 模型开发

2.1.1 反应时对数正态模型

van der Linden（2006）基于层次框架建模思路，开发了对数正态模型反应时模型，如公式 1：

$$f(t_i;\tau,\alpha_i,\beta_i) = \frac{\alpha_i}{t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[\alpha_i(lnt_i-(\beta_i-\tau))]^2\right\} \quad (1)$$

公式 1 也可以改写为：

$$\log(T_{ni}) = \xi_i - \tau_n + \varepsilon_{ni}, \varepsilon_{ni} \sim (0,\sigma^2_{\varepsilon_i}) \quad (2)$$

在（2）式中，T_{ni} 表示第 n 个被试在第 i 题上的观察反应时间；τ_n 为被试 n 的加工速度参数，表示被试在测试中的平均加工速度；ξ_i 为时间强度参数，表示被试在第 i 题所需的平均时间；ε_{ni} 为正态分布误差项，这表明反应时模型属于对数正态族，即 $\log(T_{ni}) \sim N(\xi_i-\tau_n,\sigma^2_{\varepsilon_i})$。

2.1.2 拓广分部评分模型（简称 GPCM）

1992 年 Muraki 在分部评分模型（partial Credit Model，PCM）（Masters，1982）的基础上，把项目区分度加入到项目反应函数中，得出拓广分部评分模型，如下式：

$$P(X_{ni}=t) = \frac{\exp(\sum_{v=0}^{t}\alpha_i(\theta_n-\delta_{iv}))}{\sum_{h=0}^{m}\exp(\sum_{v=0}^{h}\alpha_i(\theta_n-\delta_{iv}))} \quad (3)$$

上式为拓广分部评分模型的项目反应函数，同时限定 $\sum_{v=0}^{0}\alpha_j(\theta-\delta_{jv})=0$。公式（3）中，$P(X_{ni}=t)$ 表示第 n 个被试在项目 i 上得 t 分的概率，m 是项目的满分值，α_i 为第 i 个题目的区分度参数，δ_{iv} 为被试在第 i 个题目第 v 步的难度参数。在国内，周民元等人（2005）对 GPCM 模型的项目参数估计程序进行了研究。

2.1.3 融合反应时信息的多级评分 IRT 模型（简称 JRT－GPCM）开发

受层次建模框架（van der Linden，2006）的启发，本研究采用反应时对数正态模型和 GPCM 模型，开发多级评分反应时模型（JRT－GPCM）。在 JRT－GPCM 中，Y_{ni} 和 $\log(T_{ni})$ 分别在第一层建模；考虑项目参数之间的依赖关系和被试参数之间的依赖关系的方差和协方差结构，则分别在更高的层次上建模。

根据分层建模框架，JRT－GPCM 模型的项目参数假设服从多元正态分布，其各自的均值向量和方差协方差矩阵如下：

$$\psi_i = \begin{pmatrix} b_{1_i} \\ b_{2_i} \\ \vdots \\ b_{k_i} \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{b1} \\ \mu_{b2} \\ \vdots \\ \mu_{bk} \\ \xi_i \end{pmatrix}, \sum item\right) \quad (4)$$

如公式（4）所示，项目难度参数和项目时间强度参数之间的协方差遵循分层建模框架的思想。然而，更重要的是它可以反应项目难度参数与时间强度参数之间的关系。残差方差 $\sigma^2_{\varepsilon_i}$ 被认为是独立分布的，因此不包括在 ψ_i 内。

同样，JRT－GPCM 模型的被试能力参数和被试加工速度参数服从二元正态分布：

$$\Theta_n = \begin{pmatrix} \theta_n \\ \tau_n \end{pmatrix} \sim \left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \sum person\right), \sum person = \begin{pmatrix} \sigma^2_\theta & \rho_{\theta\tau}\rho_\theta\rho_\tau \\ \rho_{\theta\tau}\rho_\theta\rho_\tau & \sigma^2_\tau \end{pmatrix} \quad (5)$$

公式（1－5）共同构成 JRT－GPCM 模型，其层次框架如图 1。

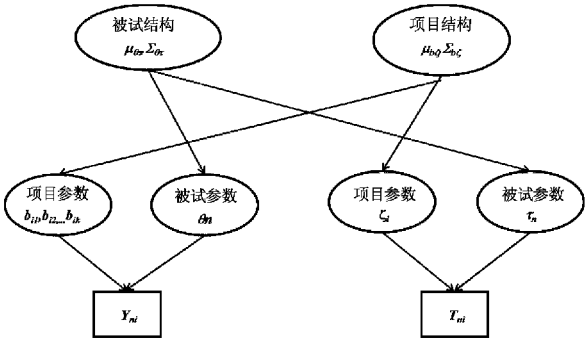

图 1 多级评分反应时分层模型

该联合建模具有两个参数识别性的问题：θ_n 和 τ_n 的可识别性，以及 τ_n 与 ξ_i 的可识别性。参照 vander Linden(2007)的研究，本文设置了三个约束条件以解决参数的可识别性问题：$\mu_\theta = 0$，$\sigma_\theta = 1$，和 $\mu_\tau = 0$。除了模型识别外，还做了两个局部独立性假设：对于给定的 τ_n，$\log(T_{ni})$ 是条件独立的，给定所有被试参数，Y_{ni} 与 $\log(T_{ni})$ 也局部独立(van der Linden，2009)。

2.2　参数估计

采用全息贝叶斯 MCMC 算法实现对新开发模型 JRT－GPCM 的参数估计。在贝叶斯估计中，模型参数的先验分布和观测数据的似然产生了模型参数的联合后验分布。

根据局部独立性假设 Y_{ni}，$\log(T_{ni})$ 有各自独立的分布，分别为：

$$Y_{ni} \sim Binomial(P(Y_{ni} = 1)), \log(T_{ni}) \sim N(\xi_i - \tau_n, \sigma^2_{\varepsilon_i})$$

假设项目参数的先验服从多元正态分布，可表示为：

$$\begin{pmatrix} b_{1_i} \\ b_{2_i} \\ \vdots \\ b_{k_i} \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{b1} \\ \mu_{b2} \\ \vdots \\ \mu_{bk} \\ \xi_i \end{pmatrix}, \sum \text{item}\right), \sigma^2_{\varepsilon_i} \sim \text{InvGamma}(1,1),$$

其中参数的先验分布分别为：

$\mu_{b_1} \sim N(0,1)$，$\mu_{b_2} \sim N(0,1)$，$\mu_{b_k} \sim N(0,1)$，$a \sim logN(0,0.27)$，$\mu_\xi \sim N(3,2)$，$\sum item \sim InvWishart(R, k+1)$。

被试参数的先验设置为：

$$\begin{pmatrix} \theta_n \\ \tau_n \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sum person\right).$$

需要注意的是为了参数的可识别性，θ_n 的方差被固定为 1，为了解决这个问题，需要对 $\sum person$ 进行转换，转换方法为克列斯基分解，如：$\sum person = \Delta_{person}\Delta'_{person}$ 对于其中的下三角矩阵，对角线上的元素有正值，而对角线下的元素没有限制。Δ'_{person} 是 Δ_{person} 的转置矩阵。因此，对于 Δ_{person} 的元素设定为 $\varphi \sim N(0,1)$、$\psi \sim Gamma(1,1)$，即 $\Delta_{person} = \begin{pmatrix} 1 & 0 \\ \varphi & \varphi \end{pmatrix}$(Zhan，Jiao，& Liao，2018)。

3　研究总体设计

为验证新开发的 JRT－GPCM 模型的可行性及其在实践中的应用，本文开展了二项研究。

研究 1：JRT－GPCM 模型参数估计精度验证。该研究 1 采用 Monte Carlo 模拟研究方式，探讨新开发模型的参数估计精度及在不同实验条件下的表现，同时验证新模型的合理性。

研究 2：JRT－GPCM 模型在人格测量中的应用。研究 2 以大五人格量表中的神经质分量表为例，一方面展示新模型在人格测量中的具体应用过程，同时还进一步探讨加入反应时信息的建模方法在心理学中的应用。

4　研究 1：JRT－GPCM 模型参数估计精度验证

4.1　实验设计及研究方法

采用 2×2 双因素实验设计，其中因素一为被试数(分别为 1000 人和 2000 人两个水平)，另一因素为测验项目数(分别为 20 题和 30 题两个水平)，所有项目为 0，1，2，3 多级评分。

参数估计采用 MCMC 方法，利用 JAGS 和 R(Version 3.3.1 64－bit)中的 R2jags 包((Version 0.5－7；Su & Yajima，2015)对参数进行估计。生成 2 条独立的马尔科夫链，每条链长 5000，取后 2500 平均。

4.2　评价指标

(1)平均离差指标(Bias)

Bias 反映了参数估计值与真值之间的偏离程度，Bias 越小，表明参数估计的越准确。

$$Bias(\hat{\nu}) = \sum_{n=1}^{N} \frac{\hat{\nu}_n - \nu_n}{N} \tag{6}$$

其中 $\hat{\nu}_n$ 为参数估计值，ν_n 为参数真值。

(2)均方根差指标(RMSE)

RMSE 反映了参数估计值与真值之间偏差的平均大小，RMSE 越小，表明参数估计的越准确。

$$RMSE(\hat{\nu}) = \sqrt{\sum_{n=1}^{N} \frac{(\hat{\nu}_n - \nu_n)^2}{N}} \tag{7}$$

4.3　Monte Carlo 模拟过程

(1)给定 JRT－GPCM 模型参数分布，并从相应分布随机生成参数真值。被试参数服从二元正态分布，且设定 $\rho_{\sigma_\theta\sigma_\tau}$ 为 0.5，σ^2_θ 与 σ^2_τ 均为 1，被试的均值向量 μ_θ 和 μ_τ 均为 0，被试参数的方差协方差矩阵如下。

$$\sum person = \begin{bmatrix} \sigma_\theta^2 & \rho_{\sigma_\theta\sigma_\tau} \\ \rho_{\sigma_\theta\sigma_\tau} & \sigma_\tau^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

项目的区分度参数 $a \sim \log N(0, 0.27)$，项目各个等级参数与时间强度参数也服从多元正态分布，且设定 $\sum item$ 的方差协方差矩阵上项目等级参数的方差为 1，时间强度参数的方差为 2，所有的项目参数均假设为不相关，且项目均值向量 μ_{b_1}、μ_{b_2}、μ_{b_3} 均为 0，μ_ξ 为 3，项目参数的方差协方差矩阵为：

$$\sum item = \begin{bmatrix} \sigma_{b_1}^2 & \sigma_{b_1b_2} & \sigma_{b_1b_3} & \sigma_{b_1b_\xi} \\ \sigma_{b_2b_1} & \sigma_{b_2}^2 & \sigma_{b_2b_3} & \sigma_{b_2b_\xi} \\ \sigma_{b_3b_1} & \sigma_{b_3b_2} & \sigma_{b_3}^2 & \sigma_{b_3b_\xi} \\ \sigma_{b_\xi b_1} & \sigma_{b_\xi b_1} & \sigma_{b_\xi b_3} & \sigma_{b_\xi}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

(2)根据参数真值及 GPCM 模型(公式 2)和对数正态模型(公式 3)生成被试得分矩阵和反应时矩阵。

(3)根据生成的得分矩阵和反应时矩阵，采用 JRT - GPCM 模型的自编 MCMC 算法利用 JAGS 和 R 软件再估计所有参数。

(4)将自编 MCMC 算法软件估计出的参数与第(1)步的参数真值进行比较，以计算参数估计的精度指标 Bias 和 RMSE。

(5)重复(1)至(4)步 30 次，即重复实验 30 次，以减少随机误差。

4.4 研究 1 结果

表 1 和表 2 分别是项目参数和被试参数的返真性情况。总体而言，在各个实验条件下，模型的所有参数的返真性都比较理想。对于被试参数来说，在四个实验条件下，速度参数的 Bias 和 RMSE 均要小于潜在特质参数，说明速度参数的返真性要优于潜在特质参数。对于项目参数，时间强度的返真性是最佳的，然后是项目区分度参数，而最差的是项目难度参数。表 3 为均值向量、$\sum person$ 和 $\sum item$ 的返真性情况，总的来说，它们的返真性都很好。除了被试潜在特质参数的方差的 Bias 和 RMSE 大于 0.2，其他各个参数的返真性都较好，同时项目参数和被试参数 30 次实验结果的离散程度(标准差)均很小，且在四个实验条件下，参数的返真性均没有太大的差异，说明程序估计的稳健性较强。由表 1、表 2、表 3 可以看出，在被试相等的情况下，题目数量越大，估计精度越高，在题量相等的情况下，被试数量的多少对估计精度的影响较小。

这些结果均表明 MCMC 估计算法可以很好地估计 JRT - GPCM 模型参数，且估计的精度较高。

表 1 被试参数的 Bias 与 RMSE 返真性

实验条件	指标	θ		τ	
		Bias	RMSE	Bias	RMSE
$N=1000, I=20$	Mean	-0.003	0.277	-0.003	0.118
	SD	0.029	0.004	0.039	0.007
$N=1000, I=40$	Mean	-0.004	0.190	-0.003	0.088
	SD	0.032	0.038	0.036	0.006
$N=2000, I=20$	Mean	0.005	0.276	0.006	0.113
	SD	0.021	0.006	0.022	0.003
$N=2000, I=40$	Mean	0.001	0.204	-0.001	0.081
	SD	0.027	0.004	0.021	0.003

表 2 项目参数的 Bias 与 RMSE 返真性

实验条件	指标	a		b		ξ	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$N=1000, I=20$	Mean	0.019	0.084	-0.006	0.168	-0.002	0.037
	SD	0.029	0.017	0.035	0.027	0.040	0.022
$N=1000, I=40$	Mean	0.017	0.084	-0.002	0.167	-0.002	0.037
	SD	0.022	0.011	0.042	0.018	0.038	0.018
$N=2000, I=20$	Mean	0.047	0.083	-0.002	0.144	0.005	0.023
	SD	0.020	0.022	0.025	0.028	0.023	0.011
$N=2000, I=40$	Mean	0.021	0.068	-0.002	0.133	-0.001	0.021
	SD	0.019	0.010	0.032	0.013	0.021	0.010

表 3　方差与协方差矩阵与均值向量估计的返真性

参数		$N=1000, I=20$		$N=1000, I=40$		$N=2000, I=20$		$N=2000, I=40$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\sum person$	$\sum \theta\tau$	−0.018	0.042	−0.020	0.041	−0.018	0.030	−0.012	0.031
	$\sum \theta\theta$	−0.268	0.269	−0.270	0.271	−0.269	0.269	−0.261	0.262
	$\sum \tau\tau$	−0.010	0.047	−0.008	0.044	−0.003	0.030	−0.007	0.028
$\sum item$	$\sum b_1\xi$	0.041	0.230	0.036	0.218	0.050	0.220	0.048	0.146
	$\sum b_2\xi$	−0.006	0.361	0.030	0.208	−0.127	0.370	0.089	0.208
	$\sum b_3\xi$	−0.001	0.345	−0.038	0.292	0.042	0.446	−0.057	0.178
μ_{item}	μ_{b_1}	0.061	0.230	−0.005	0.128	−0.071	0.211	0.056	0.172
	μ_{b_2}	0.008	0.195	0.001	0.164	−0.044	0.200	0.061	0.124
	μ_{b_3}	−0.007	0.229	0.026	0.160	−0.037	0.204	−0.042	0.146
	μ_{ξ}	−0.019	0.233	−0.004	0.189	−0.045	0.214	−0.025	0.142

注：$\sum person$ 为被试参数的方差协方差矩阵；$\sum item$ 为项目参数的方差协方差矩阵；μ_{item} 为项目均值向量。

5　研究 2：JRT－GPCM 模型在人格测量中的应用

5.1　大五人格量表及其神经质分量表

研究 2 采用量表是由 Costa 和 McCrae(1985)编制的大五人格量表(NEO)，它由聂衍刚(2008)根据中国的文化背景进行了翻译和修订的大五人格量表。该问卷共 60 道题，包括 5 个分量表(神经质量表、外向型量表、开放性量表、宜人性量表和尽责性量表)，每个分量表均有 12 题，问卷采取五级评分(1－5 为完全不同意到完全同意)。为了便于说明问题，本文选取其中的神经质分量表进行研究。本研究中该分量表的 alpha 系数为 0.842，分半系数为 0.867，表明该分量表具有很好的内部一致性信度。

5.2　被试

本文的施测群体为大学生，采用电脑作答方式进行，共收集到 1030 份数据(包括在每题的作答用时数据即反应时数据)，通过剔除无效数据(如缺失数据过多/在测谎题上做答异常)，最终有效数据为 845 份。详细的人口统计信息见表 4。

表 4　人口学信息描述统计

人口学特征	类别	频数	百分比(%)
性别	男	405	47.9
	女	440	52.1
是否独生	是	241	28.5
	否	604	71.5
年级	大一	28	3.3
	大二	120	14.2
	大三	174	20.6
	大四	523	61.9
专业	文科	324	38.3
	理科	214	25.3
	工科	284	33.6
	艺术类	23	2.7
户籍类型	农村	498	58.9
	城镇	347	41.1

5.3　数据分析

在运用 NEO－神经质分量表数据分析中，采用 JAGS 包实现 JRT－GPCM 模型参数估计，其 MCMC 设置为：使用两个随机初始值的马尔科夫链，每条链进行 20000 次迭代，每条链的前 10000 次迭代作为燃值，剩余的 10000 次迭代用于模型参数估计。

5.4　结果

5.4.1　MCMC 收敛性分析

本文采用 Brooks 和 Gelman(1998)改进的 Gelman－rubin 收敛统计量 $\hat{R}$ 作为标准，来评估各参数的收敛性。$\hat{R} < 1.1$ 表示达到收敛标准(Brooks & Gelman, 1998)。该量表各项目参数估计的 $\hat{R}$ 指标参见图 1 至图 4。从这些图可以看出所有项目的参数估计值的 $\hat{R}$ 指标基本上都小于 1.1，说明 MCMC 算法的参数估计收敛。

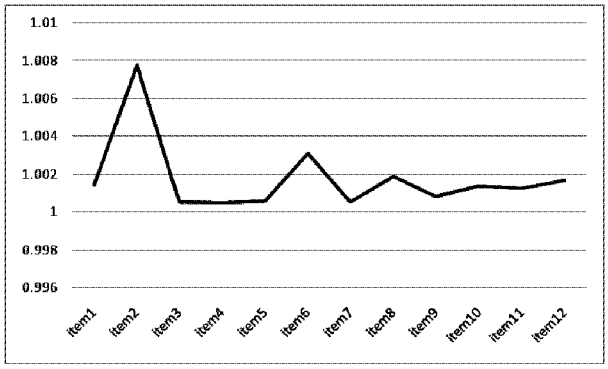

图 1 项目区分度参数

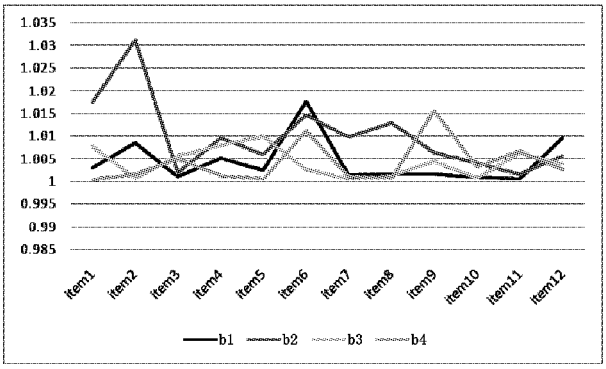

图 2 项目难度参数

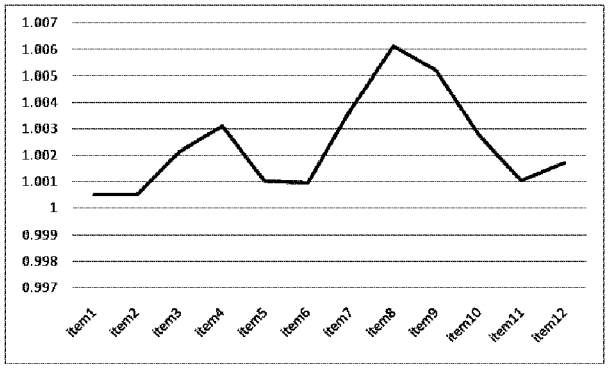

图 3 时间强度参数

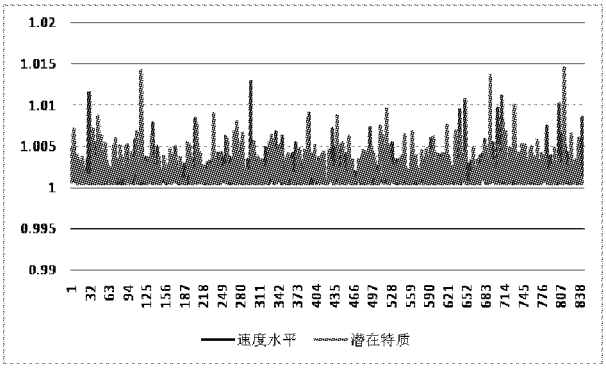

图 4 被试参数

5.4.2 研究 2 结果

表 5 给出了估计的项目参数和被试参数的方差和协方差矩阵,以及估计的项目参数均值向量。各个参数的方差和协方差的标准差均较小,说明模型在实证研究中具有很好的稳健性。

表 5 基于 NEO－神经质分量表数据的方差与协方差矩阵与均值向量

参数			Mean(S. D)	Correlation
$\sum person$	被试潜在特质和作答速度的协方差	$\sum\theta\tau$	0.036(0.015)	0.121**
	被试作答速度的方差	$\sum\tau\tau$	0.144(0.008)	
$\sum item$	第 1 个难度参数的方差	$\sum b_1b_1$	0.364(0.211)	
	第 1 个与第 2 个难度参数的协方差	$\sum b_1b_2$	0.058(0.089)	0.594*
	第 1 个与第 3 个难度参数的协方差	$\sum b_1b_3$	0.064(0.110)	0.404
	第 1 个与第 4 个难度参数的协方差	$\sum b_1b_4$	0.030(0.112)	0.244
	第 1 个难度参数与时间密度参数的协方差	$\sum b_1\xi$	－0.037(0.088)	－0.310
	第 2 个难度参数的方差	$\sum b_2b_2$	0.158(0.079)	
	第 2 个与第 3 个难度参数的协方差	$\sum b_2b_3$	0.048(0.065)	0.676
	第 2 个与第 3 个难度参数的协方差	$\sum b_2b_4$	0.016(0.063)	0.254*
	第 2 个难度参数与时间密度参数的协方差	$\sum b_2\xi$	0.011(0.056)	0.162
	第 3 个难度参数的方差	$\sum b_3b_3$	0.224(0.118)	
	第 3 个与第 4 个难度参数的协方差	$\sum b_3b_4$	0.066(0.082)	0.672*
	第 3 个难度参数与时间密度参数的协方差	$\sum b_3\xi$	－0.036(0.066)	－0.399
	第 4 个难度参数的方差	$\sum b_4b_4$	0.212(0.114)	
	第 4 个难度参数与时间密度参数的协方差	$\sum b_4\xi$	－0.051(0.065)	－0.703*
	时间密度参数的方差	$\sum\xi\xi$	0.164(0.084)	

续表 5

参数			Mean(S. D)	Correlation
μ_{item}	第 1 个难度参数的均值	μ_{b_1}	-2.139(0.187)	
	第 2 个难度参数的均值	μ_{b_2}	0.106(0.124)	
	第 3 个难度参数的均值	μ_{b_3}	0.185(0.143)	
	第 4 个难度参数的均值	μ_{b_4}	2.593(0.150)	
	时间密度的均值	μ_{ξ}	1.675(0.119)	

注:Mean 为后验均值;$\sum person$ 为被试参数的方差协方差矩阵;$\sum item$ 为项目参数的方差协方差矩阵;* $=p<0.05$;** $=p<0.01$。

表 6 为神经质分量表各个题目的项目参数估计值及其估计的标准误。从表 6 可知,神经质分量表的 12 个题目的区分度范围在 0.895 ~ 1.209,均大于 0.7(Fliege,2015),说明 12 道题目的质量都较好,而且大部分项目参数估计的标准误基本介于 0.01 ~ 0.2 之间,基本上在可接受的范围之内。

表 6　神经质分量表数据的项目参数估计值

Items	a	b_1	b_2	b_3	b_4	ξ	σ_{ε}
item1	0.952(0.096)	-2.502(0.221)	0.189(0.099)	-0.202(0.101)	2.183(0.149)	2.319(0.024)	0.542(0.015)
item2	0.977(0.101)	-1.86(0.171)	0.378(0.099)	0.227(0.102)	2.61(0.171)	1.859(0.02)	0.493(0.013)
item3	0.958(0.1)	-2.481(0.221)	-0.066(0.101)	-0.163(0.093)	2.242(0.14)	1.835(0.021)	0.466(0.012)
item4	1.065(0.113)	-2.588(0.223)	-0.166(0.094)	0.049(0.093)	2.477(0.147)	1.607(0.022)	0.496(0.013)
item5	1.133(0.117)	-1.948(0.16)	0.254(0.09)	0.462(0.105)	2.966(0.204)	1.544(0.022)	0.499(0.013)
item6	0.953(0.101)	-1.223(0.136)	0.158(0.106)	0.249(0.106)	2.663(0.178)	1.469(0.022)	0.477(0.012)
item7	1.114(0.122)	-2.428(0.203)	-0.168(0.094)	0.044(0.098)	2.963(0.178)	1.493(0.023)	0.538(0.015)
item8	0.953(0.092)	-2.003(0.177)	0.506(0.095)	1.011(0.115)	3.001(0.226)	1.582(0.021)	0.46(0.011)
item9	1.003(0.1)	-2.559(0.215)	0.008(0.09)	0.379(0.096)	2.554(0.157)	1.686(0.022)	0.48(0.012)
item10	1.209(0.127)	-2.454(0.203)	0.055(0.091)	0.105(0.095)	2.847(0.181)	1.5(0.024)	0.542(0.012)
item11	0.949(0.096)	-2.602(0.219)	-0.096(0.098)	-0.023(0.096)	2.635(0.162)	1.56(0.022)	0.48(0.012)
item12	0.895(0.094)	-1.652(0.157)	0.188(0.098)	0.19(0.098)	2.526(0.167)	1.494(0.021)	0.429(0.01)

在人格测量领域有一个非常重要的理论被称为“距离 - 困难假设”(distance - difficulty hypothesis;Ferrando & Lorenzo - Seva,2007),它的含义可以这样表述:当被试的潜在特质水平越接近项目的阈值时,被试对该项目做出反应的难度也会随之增加。例如:艾森克人格问卷中的一个项目,你是否活跃?选项为“是”或着“否”,一个非常活跃的人会毫不犹豫地回答“是”,相反,一个很不活跃的人则会很快做出“否”的回答,而一个活跃程度一般的人,会由于不确定而很难做出反应,因此犹豫不定,反应时间则会随之延长。这一假设是类比一个众所周知的心理物理结果而得到的,即当刺激接近个体的心理物理阈值时,响应时间会增加(Vickers,1980)。

根据“距离 - 困难假设”,Ferrando 和 Lorenzo - Seva(2007)首次提出了一个用于描述人格测验中反应时间的数学模型,并将该模型成功用于一个人格测验数据集,证明了反应时间包含了关于个体潜在特质的信息。在 IRT 框架下,Ferrando 和 Lorenzo - Seva(2007)使用被试的潜在特质与项目阈值之间的绝对差异来衡量“距离”:

$$\delta_{ij} = \sqrt{a_j^2(\theta_i - b_j)^2}$$

这里同时用项目的区分度进行加权,表明被试反应倾向发生转变的突然程度,项目的区分度越大,被试的反应倾向转变也就越突然。不难发现,该模型仅适用于 0 ~ 1 计分的数据,因为在多级计分的模型(如 GRM 模型,GPCM 模型等)中,由于存在多个阈值参数,类似的“距离”较难定义。本研究依据经典测量理论来划分“距离”(Kuncel,1977),图 5 中本研究超过一半的被试得分范围在 2 ~ 3 区间,对应项目难度参数(location parameter)为 b_2,4 ~ 5 分的被试人数最少。各个难度参数与时间强度参数的相关,具体相关情况可见表 5,b_2 与时间强度参数呈正相关外,其余各个难度均与时间强度参数呈现负相关,而 $r_{b_4\xi}$ 的相关达到 -0.703,其次分别为 $r_{b_3\xi}$、$r_{b_1\xi}$。

被试的整体水平接近 b_2，而项目的第四个难度参数被试的整体距离相距最远，因此 b_4 与时间强度参数存在较大的负相关，题目的难度参数越高那么时间强度就越小，表明被试距离神经质倾向越远，那么做出反应的时间越长，这与 Jochen(2013) 的研究结果一致。

早期的相关研究发现，在人格构念的特质维度两端的被试对相应项目做出反应更快，另一方面，代表极端特质的项目比中等特质范围的项目的反应时间更短(Amelang，Eisenhut，& Rindermann，1991)，这表明，项目的阈值与该项目的时间强度存在一定程度的相关。

基于此，在 JRT－GPCM 的框架下，我们可以认为项目的难度参数(location parameter)与其时间强度参数存在相关，即项目的难度参数的绝对值越大(偏离中心值，则说明代表了偏向极端的特质水平)，那么被试在作答该项目时所需要的时间越少，即时间强度较小。

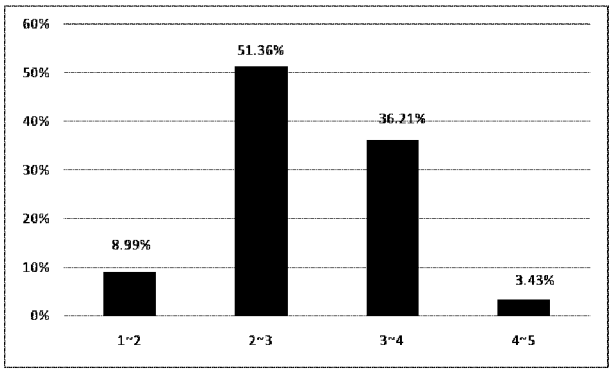

图 5　被试在神经质量表得分频率表

表 6 中，JRT－GPCM 估计的潜在特质 θ 与神经质总分的相关系数为 0.993，存在非常明显的正相关，作答速度水平与神经质总分的相关系数为 0.121($p<0.01$)，但也达到了显著性水平，说明潜在特质、作答速度水平均与总分存在正相关，即被试的神经质倾向越高潜在特质越高，并且作答速度也越快，见图 6。

表 7　作答速度水平、潜在特质与总分的相关

	作答速度水平 τ	潜在特质 θ	神经质分量表总分
作答速度水平 τ	1		
潜在特质 θ	0.121**	1	
神经质分量表总分	0.113**	0.993**	1

注：** $=p<0.01$

通过整个测验可以获得每个被试的测验的总分，结合被试的作答速度水平，可以绘制被试的潜在特质、作答速度水平与神经质总分的关系分布，如图 7。从图中可以看出，相同作答速度水平的被试，得分不同，可能的原因是，相同作答速度水平的被试，其潜在特质可能存在差异，从而影响被试的得分；同时，相同分数的被试，其作答速度水平可能不同，这可能是由于潜在特质相同被试的信息提取效率有差异，因为有的被试对文字的提取能力较快，有些被试对文字信息的提取能力较慢，有效提取需要消耗更多的时间。因此，作答速度水平与潜在特质可能描述了被试两个不同层面的特质。

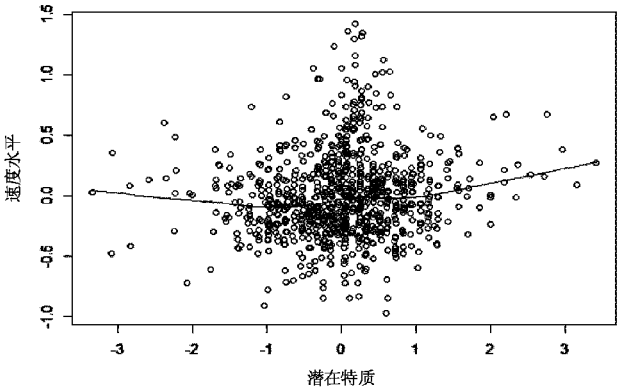

图 6　被试潜在特质与反应速度的关系

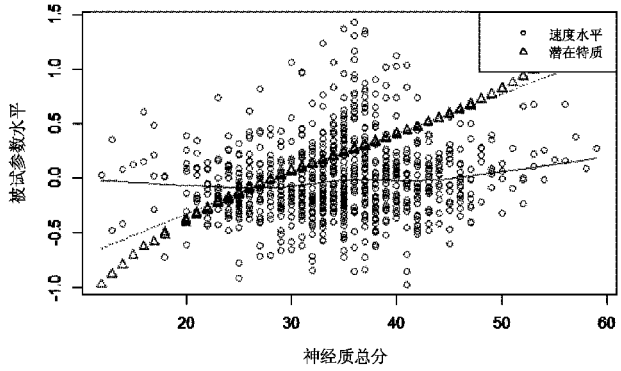

图 7　潜在特质、反应速度与神经质总分的关系

6　讨论及结论

随着计算机测试技术的发展，采集反应时已经成为许多大型测试的常规工作，但当前有关反应时的模型只适用与 0－1 评分数据的试题。然而在教育考试或心理测验中，选择题、填空题等一类的 0－1 评分试题所提供的信息十分有限，而诸如综合分析题、应用题、作文题等多级评分题型考核题目提供的信息更为丰富。但是当前国内外，均没有融合多级评分的反应时有关的研究，为填补这一空白，本研究在 GPCM 模型基础上开发了一种新的融合反应

时的多级评分 IRT 模型。

研究使用全息贝叶斯 MCMC 方法进行模型的参数估计，并探讨了 JRT - GPCM 模型在四种不同的实验条件下的性能，研究结果表明：MCMC 算法对 JRT - GPCM 模型的参数估计精度较理想；且在不同题长和被试人数下，JRT - GPCM 模型均有较好的适用性，说明 JRT - GPCM 模型具有较好的稳健性。此外，本研究探讨了 JRT - GPCM 模型在大五人格测量中的应用，数据分析结果表明，被试的潜在特质参数和被试的速度参数之间存在正相关，同时证实了。Ferrando 和 Lorenzo - Seva（2007）提出的“距离 - 困难度假设”，即当被试的潜在特质与项目的难度阈限距离越远，那么被试会花费更多的时间对项目进项作答。

尽管在本研究中仅使用 GPCM 模型进行了说明，但联合多级评分建模方法可以扩展到其他的多级评分模型，以进行进一步的研究。本文在 JRT - GPCM 模型中采用了经典的对数正态 RT 模型（van der Linden，2006）。因此，JRT - GPCM 模型存在一些限制。例如，假设在整个测试中受访者的速度特征是恒定的（Fox & Marianti，2016）；对于给定的被试而言，项目作答和反应时间是独立的（Bolsinova & Maris，2016；Meng et. al.，2015）；对数转换后，对数反应时遵循正态分布（Klein Entink，van der Linden，et al.，2009）。其他反应时模型可以在将来的研究中进行探索和比较。此外，这项研究假设对被试潜在特质与项目难度距离的划分依据与被试的作答情况，在今后的研究中可进一步对这种“距离”提供更为科学的依据。

在 IRT 框架中，多项研究表明，记录的反应时可用于改良测试设计，优化计算机自适应测试中的项目选择和项目校准，以及检测检测异常反应行为（例如 Lee & Chen，2011；Meyer，2010；van der Linden，2008；van der Linden，Breithaupt，Chuah，& Zhang，2007；van der Linden & Guo，2008；Wang & Xu，2015；Wise & DeMars，2006；Wise & Kong，2005），未来研究还需进一步探讨在多级计分反应时 IRT 模型下，如何使用反应时信息来优化 CAT 选题以及个体异常反应行为等研究。

参考文献

聂衍刚，林崇德，郑雪，丁莉，彭以松.（2008）. 青少年社会适应行为与大五人格的关系. *心理科学*，*31*（4），774 - 779.

周明元，甘登文，丁树良.（2005）. GPCM 模型项目参数估计程序的开发与研究. *心理学探新*，*25*（1），57 - 60.

Amelang，M.，Eisenhut，K.，& Rindermann，H.（1991）. Responding to adjective check list items：A reaction time analysis. *Personality and Individual Differences*，*12*，523 - 533.

Andrich，D.（1978）. A rating formulation for ordered response categories. *Psychometrika*，*43*，561 - 573.

Brooks，S. P.，& Gelman，A.（1998）. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*，*7*，434 - 455.

Bolsinova，M.，& Maris，G.（2016）. A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*，*69*，62 - 67.

Costa，P. T. Jr.，& McCrae，R. R.（1985）. *The NEO personality inventory manual.* Odessa，FL：Psychological Assessment Resources.

Ferrando，P. J.，& Lorenzo - Seva，U.（2007）. An item response theory model for incorporating response time data in binary personality items. *Applied Psychology*，*31*，525 - 543.

Fliege，H.，Becker，J.，Walter，O. B.，Bjorner，J. B.，Klapp，B. F.，& Rose，M.（2005）. Development of a computer - adaptive test for depression（D - CAT）. *Quality of Life Research*，*14*（10），2277 - 2291.

Fox，J. - P.（2010）. *Bayesian item response modeling：Theory and applications.* New York，NY：Springer.

Fox，J. - P.，& Marianti，S.（2016）. Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*，*51*，540 - 553.

Gelfand，A. E.，& Smith，A. F. M.（1990）. Sampling - based approaches to calculating marginal densities. *Journal of the American Statistical Association*，*85*，398 - 409.

Jackson，D. N.（1986）. The process of responding in personality assessment. In A. Angleitner & J. S. Wiggins（Eds.），*Personality assessment via questionnaires*（pp. 123 - 143）. Berlin，Germany：Springer - Verlag.

Jochen，R.（2013）. Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*，*55*（4），361 - 382.

Klein Entink，R. H.，Fox，J. - P.，& van der Linden，W. J.（2009）. A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*，*74*，21 - 48.

Klein Entink，R. H.，van der Linden，W. J.，& Fox，J. - P.（2009）. A box - cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*，*62*，621 - 640.

Lee，Y. - H.，& Chen，H.（2011）. A review of recent response - time analyses in educational testing. *Psychological Test and Assessment Modeling*，*53*，359 - 379.

Locke, E. A. (1965). Interaction of ability and motivation in performance. *Perceptual and Motor Skills*, *21*, 719 – 725.

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. *Learning and Individual Differences*, *21*, 124 – 128.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York, NY: Oxford University Press.

Masters, G. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149 – 174.

Meng, X. – B., Tao, J., & Chang, H. – H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*, 1 – 27.

Meyer, J. P. (2010). A mixture rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521 – 538.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*, 197 – 219.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological*, *16*(2), 159 – 176.

Plummer, M. (2015). *JAGS Version* 4. 0. 0 *user manual.* Retrieved from http://sourceforge. net/ projects/mcmc – jags/.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer – based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38 – 47.

Su, Y. – S., & Yajima, M. (2015). *R2jags: Using R to run 'JAGS'. R package version* 0. 5 – 7. Retrieved from http://CRAN. R – project. org/package = R2jags.

Suh, H. (2010). *A study of bayesian estimation and comparison of response time models in item response theory.* Doctoral dissertation, University of Kansas, Lawrence, KS.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181 – 204.

van der Linden, W. J. (2007a). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287 – 308.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007b). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117 – 130.

van der Linden, W. J. (2007c). Conceptual issues in response – time modeling. *Journal of Educational Measurement*, *46*, 247 – 272.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5 – 20.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response – time patterns in adaptive testing. *Psychometrika*, *73*, 365 – 384.

van der Linden, W. J. (2009). Conceptual issues in response – time modeling. *Journal of Educational Measurement*, *46*, 247 – 272.

van der Linden, W. J., & Fox, J. – P. (2015). Joint hierarchical modeling of responses and response times. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol.* 1. *Models* (pp. 481 – 500). Boca Raton, FL: Chapman & Hall/CRC.

Vickers, D. (1980). Discrimination. In A. T. Welford (Ed.), *Reaction times* (pp. 25 – 72). New York: Academic Press.

Wang, C., Chang, H., & Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *Journal of Mathematical and Statistical Psychology*, *66*, 144 – 168.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456 – 477.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323 – 339.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort – moderated IRT model. *Journal of Educational Measurement*, *43*, 19 – 38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer – based tests. *Applied Measurement in Education*, *18*, 163 – 183.

Zhan, P., Jiao, H., & Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 262 – 286.

(下转第 288 页)

Development, Reliability and Validity Test of the Multidimensional Self – Control Scale for Adolescents

Liu Jinglun Dong Shenghong Ye Baojuan Yu Zhiwen
(School of Psychology, Key Lab of Psychology and Cognitive Learning, Jiangxi Normal University, Nanchang 330022)

Abstract: To adapt to the theory of multidimensional self – control, a scale for measuring multidimensional self – control of Chinese adolescents was developed and its reliability and validity were tested. The initial questionnaire was formed through literature review and expert evaluation. Taking junior and senior high school students as subjects, 1424 subjects were selected for the preliminary test, and 681 subjects were selected for the formal test. The initiation self – control scale includes three dimensions of emotion regulation, goal maintenance and self – control strategies, and the inhibition self – control scale includes three dimensions of attention control, impulse control and emotional control. Both scales have good construct validity, criterion – related validity and reliability. In conclusion, the initiation and inhibition self – control scales in the Multidimensional Self – Control Scale for Adolescents meet the psychometric standards, could be used to evaluate the self – control ability of Chinese adolescents.

Key words: self – control scale; self – control; adolescents; reliability; validity

(上接第 278 页)

Research on Development and Application of Polytomous IRT Model Incorporating Response Times

Wang Daxun[1], Guo Yingying[2]
(1. Jiangxi Normal University, Nanchang 330022; 2. Hailiang Education Group Inc., Hangzhou 310052)

Abstract: With the development of computer testing technology, collecting reaction time has become a routine work of many large – scale tests. However, most current IRT models for fusion reaction time are only applicable to 0 – 1 score data, which greatly limits the application of IRT model in practice. Based on the traditional two – level scoring response time IRT model, this paper intends to develop a multilevel scoring response time model. Under the framework of hierarchical modeling, the extended partial scoring model (GPCM) and the log – normal model (jrt – gpcm) were used to construct the multi – stage scoring IRT model (jrt – gpcm) for fusion reaction, and the parameter estimation of the new model was realized by the holographic bayesian MCMC algorithm. In order to verify the feasibility of the newly developed jrt – gpcm model and its application in practice, this paper carried out two studies: Study 1 for simulation experiment research, the use of 2 x 2 double factor experiment design, one factor for the number of participants (1000 and 2000 respectively, the two level), another factor for the test number (20 and 30 respectively two levels), all items of 0, 1, 2, 3 multistage grading, using holographic bay leaf, MCMC algorithm for parameter estimation, and validates the feasibility of MCMC algorithm and JRT – GPCM model to estimate accuracy; Study 2 for JRT – GPCM model in the application of the big five personality – neurotic subscales, testing group for college students, this paper USES the computer answer way, collected a total of 1030 data (including the answer in each available data that reaction time), by eliminating the invalid data (such as too many missing data/answer exception) on lie detection problem, the final valid data is 845. Study 1 results show that under the JRT – GPCM model, the estimated method of MCMC algorithm by fairly robustness, and the precision of the item and the person the parameters was preferably great, model has good robustness, and the topic, the more the higher estimation precision, It indicated that the number of subjects indicated that the rrt – gpcm model was reasonable and feasible. Study 2 shows that the parameter estimation indexes of all items are basically less than 1.1, indicating the convergence of parameter estimation of MCMC algorithm. The variance of each parameter and the standard deviation of the covariance are small, which indicates that the model has good robustness in empirical research. The 12 questions on the neurotic subscale ranged from 0.895 to 1.209, all of which were greater than 0.7 (Fliege, 2015), indicating that the 12 questions were of good quality. There was a positive correlation between the potential traits and the response speed of the subjects. The higher the neurotic tendency of the subjects, the higher the potential traits and the faster the response speed. Project step parameters (the location parameter) and its parameters is related to the intensity of time, the greater the absolute value of that project step parameters (off center value, then represents to the characteristics of extreme levels), so the participants answers in the less time needed for the project, namely the time intensity is small, the results support Ferrando and Lorenzo – Seva (2007) proposed "distance – difficult holiday". In conclusion, this study provides a new method to expand the application of response time information in psychological measurement and education.

Key words: item response theory; GPCM model; JRT – GPCM model; MCMC algorithm