

基于速度与准确率权衡的心理测量学模型及应用

郭小军 罗照盛

(江西师范大学心理学院, 南昌 330022)

摘要:在大规模被试评价中,任务完成的准确率一直是评价的主要指标。然而,在各种任务情景中,反映执行者素质的指标除了结果的正确性之外,决策过程的时效性同样是极其重要的。因此,开发一个适合大规模评价情景,同时考虑速度与准确率权衡的模型,探索被试作答准确率与速度间的关系将具有重要的价值。基于认知心理学实验中的速度与准确率研究,构建一个基于速度与准确率权衡的心理测量学模型。新模型参数能非常稳定而又精确地被估计,同时模型中的变量及其关系也能够很好地得到实测数据的支持。

关键词:速度与准确率权衡;大规模评价;认知实验;心理测量

中图分类号:B841.2

文献标识码:A

文章编号:1003-5184(2019)05-0451-10

1 引言

人们在解决现实问题做出决策时,总是希望得到一个正确的结果。然而,在任何情景下,时间几乎是一个不可避免的因素,在决策的过程中,总是需要考虑时间的有限性,不可能对任务进行无限制地思考和讨论,比如智力测验中既要保证结果的正确性,同时又要保证快速作答;而在人格测验中,如果评价结果涉及到高利害关系时(比如工作应聘),被试就可能进行更长时间的思考,以选择有利的结果,当然,作答时间过长可能会被视为虚假作答的反映;而在解决现实问题时,处理方式的合理性和做出决定的快速性有时可能会同等重要,被试只能在权衡结果和速度的基础上,做出一个最优的决策。因此,在评价被试完成任务的质量时,应该同时考虑结果的准确性和反应过程的流畅性水平。

然而,在实际的大规模评价项目中,准确率一直是评价被试任务完成质量的主要指标。研究者和实践者往往忽视被试在任务完成过程中的时间使用效率,而反应时其实能揭露准确率所不能反映的一些关键信息与心理活动,能够反映被试在不同维度上的素质水平。在被试评价指标中,准确率反映的主要是被试的信息完备性或知识丰富性水平,而反应时则反映了被试的信息加工流畅性水平。信息完备性指的是被试是否拥有所考察项目规定的知识结点及正确的联结模式;而流畅性水平反映了被试所有相应知识结点的强度以及知识结点之间的联结强度(熟悉度)。在完成一个测验任务时,被试要正确作答测验项目,前提是被试头脑中已存储了相应的知识结点及正确的联结模式,而且知识结点的强度与知识结点之间的联结强度能够支持被试流畅地提取

相应的信息。被试作答测验项目时,不仅仅反映了其能力水平,同时也反映了被试的信息提取的流畅性水平。所以,要更全面,更加科学地评价被试的水平,应该将准确率指标和反应时指标有机地结合在一起,在实际任务解决中,既考虑结果的准确性,又考虑决策过程的时效性。

在认知心理学中,已经进行了大量的基于速度与准确率指标及其权衡关系(speed-accuracy tradeoff, SAT)的研究,如概念加工(McElree, Jia, & Litvak, 2000), 句子理解(McElree, 2000; McElree, Foraker, & Dyer, 2003), 记忆(McElree, 1998), 注意(McElree & Carrasco, 1999; McElree & Carrasco, 2001; Carrasco, McElree, Denisova, & Giordano, 2003; Carrasco, Giordano, & McElree, 2005; Giordano, McElree, & Carrasco, 2009)等领域。在不同实验条件下,通过对 SAT 模型参数组合进行变化,计算模型与数据拟合的指数(Reed, 1976),最后确定最佳的参数组合形式,从而评价这些条件对心理现象的影响。在认知心理学实验研究中,较为深入地研究了认知加工任务中速度与准确率指标,研究了不同刺激条件下的速度与准确率的权衡变化关系。但是,认知心理学实验关于速度与准确率权衡的研究仅仅局限于实验室内,并且进行的都是非个人层面的小样本研究;在被试信息处理上,对于错误作答的项目时间,则该项目的反应时会用其他被试均值等方法进行替换,当被试作答错误项目过多时,则被试的作答结果一般认为是无效的;另外,认知心理学实验任务往往都是非常简单的重复任务,较难应用于复杂的任务加工中,这就严重限制了速度与准确率及其权衡范式的研究范围。

在心理测量学研究中,总是假设完成任务的时间足够充分,被试能力得到完全测量。然而,实际测试中总是在强调准确率的同时,会限定任务完成的时间。因此测试结果会在一定程度上反映被试速度与准确率的结合。前期研究中对反应时的忽视或许是由于客观条件的限制,但是随着电脑使用的普及,搜集被试的反应时变得越来越简单与便利。心理测量学研究者也逐渐认识到单纯依靠准确率信息来评价被试的局限性,于是,反应时研究开始受到关注。van der Linden(2006)构建了一个对数正态模型分析被试反应时数据;考虑到对数正态分布对反应时分布存在不适用情况,Klein Entink, van der Linden 和 Fox(2009)提出具有一般性的 Box - Cox 正态转换的反应时模型;之后,Meng, Tao 和 Shi(2014)将对数正态模型扩展到多级计分模型;并且孟祥斌(2016)发现对数偏正态比对数正态拟合反应时分布更佳。为了减小对反应时分布的依赖性,部分研究者借鉴生存分析,将反应时分布作为一个半参数部分构建反应时分析模型。Ranger 和 Ortner(2012, 2013)以及 Ranger 和 Kuhn(2014, 2015)提出潜在特质的比例风险模型以及相关反应时模型;Wang, Fan, Chang 和 Douglas(2013)将潜在特质的比例风险模型与准确率模型进行联合估计,发现能较好拟合实测数据。Wang, Chang 和 Douglas(2013)还基于半参数基础上,提出一个更有一般性的线性转换模型。在大规模评价项目中,如 PISA(Program for International Student Assessment)、TIMSS(The Third International Mathematics and Science Study)、NAEP(The National Assessment of Educational Progress)等,测量的项目任务较为复杂,信息加工量比较大,信息加工过程更加复杂;既可以对群体水平进行评价,也可以对个体加工水平进行评价;在作答信息利用上,错误作答信息与正确作答信息具有同样重要的价值,共同为准确评价被试水平或项目质量提供有用信息。但是,在已提出的大部分反应时测量学模型中,都无法反映出被试作答过程中速度与准确率权衡的现象;对于在相同时间内作答错误与作答正确的不同被试,现有的反应时模型无法区分两种被试的流畅性水平;特别是当被试作答过程中存在权衡时,注重准确率或者注重反应时的两种被试,现有模型也无法充分有效估计被试的流畅性水平;另外,速度与准确率权衡作为作答过程中的基本现象,现有的模型无法进行有效的描述与刻画。考虑到“被试流畅性水平”似乎更能表示一种内隐的、潜在的特质,下文将以流畅性水平表示加工速度。

综上所述,研究拟探索的问题是,借鉴已有的认知心理学实验研究成果和心理测量学研究成果,开发一个考虑速度与准确率及其权衡关系的心理测量学模型,以期实现认知心理学实验成果从实验室走向大规模评价应用中。

2 基于认知心理学实验 SAT 函数模型的构建

2.1 认知心理学实验 SAT 函数模型

Reed(1973)提出了一个基于速度与准确率权衡模式下的函数模型,如下函数式(1):

$$d'(t) = \lambda * (1 - \exp(-\beta * (t - \delta))) \quad t > \delta \text{ 且 } t \neq 0 \quad (1)$$

式(1)中, d' 表示的是被试辨别力水平,在信号检测实验中,用击中概率与虚报概率各自对应的正态化标准分数之差来表示; t 表示的是控制加工时间,指的是以刺激呈现的时间为起点的被试认知加工时间,在 SAT 实验范式中作为操作变量由实验设计者来预先设定; λ 是渐近线水平参数,反映的是在加工时间足够时,被试群体能达到的最大辨别力水平; δ 是截距参数,是辨别力处于随机水平时($d' = 0$)的加工时间; β 为加工速率,是辨别力水平随着时间变化的变化速率,反映函数曲线的陡峭程度。该函数描述了被试信息加工的动态过程。在认知实验过程中,主试通过设置不同的控制加工时间(函数(1)中的时间 t),并获取被试在对应控制加工时间条件下的辨别力水平,从而可以了解被试群体在实验条件中的加工速度与准确率之间的权衡关系(如图 1 所示, $\delta = 0.2, \beta = 2, \lambda = 3.1$)。

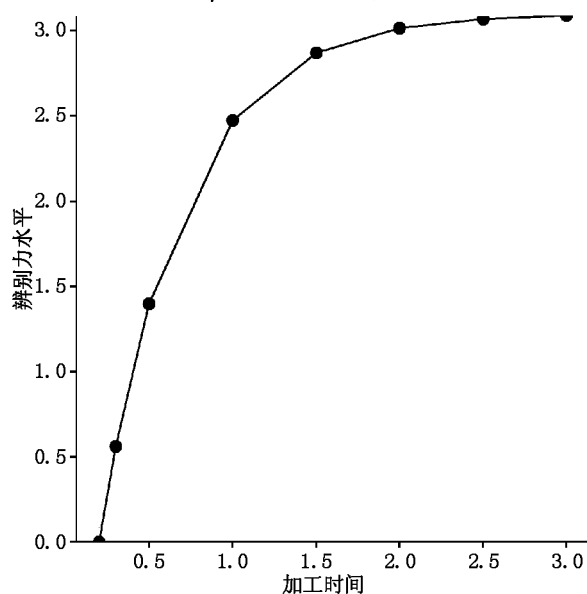


图 1 加工时间与辨别力关系

大量认知心理学研究(Dosher, 1976, 1984; Reed, 1973, 1976; Wickelgren, 1977; Carrasco et al.,

2003;Giordano et al.,2009)表明,函数(1)能较好地拟合认知任务加工中速度与准确率权衡的动态数据。

2.2 认知心理学实验 SAT 函数模型变量的分析与再建

在函数式(1)中, β 为加工速率,反映曲线的陡峭程度,与项目反应理论模型的区分度内涵一致,因此, β 值在0.1到2.5之间(罗照盛,2012)。

在信号检测论中, d' 表示的是被试在噪音背景中识别信号的辨别力水平;上渐近线参数 λ 表示的是在充足反应时间条件下,被试群体理论上能达到的最高辨别力水平。以二值记分项目为例,信号指正确选项,噪音指错误选项,击中为选择正确选项,即为 p ,虚报为选择错误选项即为 $Q=1-p$,则被试辨别力水平为 $d'=Z_p-Z_{1-p}$,根据正态分布性质, $Z_p=-Z_{1-p}$,则 $d'=2Z_p$,则有 $2Z_{p=1}=2Z_{max}=d'_{max}=\lambda$,并且项目的渐近线参数 λ 对所有被试都是相同的,因此, $\frac{d'}{\lambda}$ 取值范围为(0,1)之间,即 $\frac{d'}{\lambda}=p$ 。函数式(1)可以转换为:

$$p(t)=1-\exp(-\beta*(t-\delta)) \quad t>\delta \text{ 且 } t \neq 0$$

关于反应时间变量,需要明确三类概念。第一类是预先设定的控制加工时间,对每个被试都是固定不变的,即函数(1)中的时间 t 。第二类是被试的期望加工时间 t^* (vander Linden,2009),即被试在项目上的理论作答时间。第三类是被试的实际观测加工时间 \hat{t} ,可以任意变化,不同被试完成同一项目的加工时间可能不同。在认知实验 SAT 范式中,通过设置不同控制加工时间点来研究和了解被试的速度与准确率权衡关系。但是在大规模评价中,通过重复测量,为测验项目设置不同的控制加工时间是不现实的。于是,通过期望加工时间 t^* 与实际观测加工时间 \hat{t} 的组合关系 $\hat{t}-t^*$ 来表示认知实验中不同控制加工时间点的情况,同时又能够应用于大规模评价项目中。在函数(1)中, δ 指随机加工时间, $t-\delta$ 则反映了被试的精细加工时间, $t^*-\delta$ 和 $\hat{t}-\delta$ 亦为同理。根据 SAT 实验范式,当控制加工时间 t 低于被试完成项目的期望加工时间时,则对应时间差 $\hat{t}-\delta-(t^*-\delta)=\hat{t}-t^*<0$,即实际观测加工时间小于期望加工时间,被试作答时间不足,或者被试追求更快的解题速度;当控制加工时间 t 等于被试完成项目的期望加工时间时,则对应 $\hat{t}-t^*=0$,即实际观测加工时间与期望加工时间相等,被试在准确率与速度方面进行了很好的权衡;当控制加工时间 t 高

于被试完成项目的期望加工时间时,则对应的 $\hat{t}-t^*>0$,即实际观测加工时间大于期望加工时间,被试有剩余时间思考,从而追求更高的准确率。所以,时间差 $\hat{t}-t^*$ 能有效实现控制加工时间 t 的功能。在认知实验中,加工时间可以作为操作变量由主试控制,以研究速度与准确率的权衡关系;而在实际的大规模评价中,加工时间只能由被试自身控制,但可以通过与期望加工时间比较,同时结合作答准确率来评价被试准确率与作答速度之间的权衡关系。根据函数(1)的定义,要求 $t-\delta>0$,选择指数转换,即 $\exp(\hat{t}-t^*)$ 。

2.3 心理测量学 SAT 模型构建

通过对函数(1)中各变量内涵的分析与表达形式的再建,并结合 van der Linden(2009)的分析,构建出一个新的反映了加工速度与作答准确率权衡关系的模型,如下式(2):

$$p(U_{ij}=1 | \tau_i, d_j, \beta_j) = 1 - \exp(-\beta_j * \exp(t_{ij} - \frac{d_j}{\tau_i})) \quad (2)$$

其中, p 指被试在项目上的准确率; U_{ij} 表示被试 i 在项目上 j 的作答情况,1为正确作答,0为错误作答; t_{ij} 指被试 i 在项目 j 上的实际观测作答时间; τ_i 为被试 i 的流畅性水平参数, τ 越大,则被试流畅性水平越高, τ 越小,流畅性水平越低; d_j 是项目 j 的时间压力参数,也就是作答项目所需的时间量或工作量; β_j 为项目 j 的区分度参数,也叫速率参数,是反映项目曲线的陡峭程度。

vander Linden(2009)分析了被试的实际观测作答时间、流畅性水平参数和项目的压力参数三者之间的关系,并通过式(3)进行表示,

$$t_{ij} = \frac{d_j}{\tau_i} + \varepsilon_{ij} \quad (3)$$

式(3)中, $\frac{d_j}{\tau_i}$ 表示了被试的期望作答时间 t^* ; ε_{ij}
 $= t_{ij} - \frac{d_j}{\tau_i}$ 表示了实际观测作答时间与期望作答时间之间的差异大小,与前文中 $\hat{t}-t^*$ 表示了相同的意义,反映了被试作答过程中受到其他因素的影响。观测时间和期望时间差与准确率结合可以反映被试作答中的权衡关系。当被试作答倾向又快又好时,则时间差偏向负数且准确率高;当被试作答倾向好却慢时,则时间差偏向正值且准确率高;当被试作答倾向差却快时,则时间差偏向负值且准确率低;当被试作答又差又慢时,则时间差偏向正值且准确率低。

被试的权衡不同,则作答结果也会不同,但是又快又好是所有被试以及决策者都青睐的目标。

2.4 心理测量学 SAT 模型参数估计

2.4.1 参数估计的 MCMC 算法

在估计心理测量学 SAT 模型的参数时,首先假设被试与项目各自之间的作答时间与准确率是相互独立的,即项目之间的作答时间与准确率以及被试之间作答时间与准确率分别独立。研究使用 R 语言自编程序进行参数估计,其 M-H 的 Gibbs 抽样过程如下所述。

(1) 被试流畅性水平参数

$\hat{\tau}_i^{v+1}$ 从对数正态分布 $\text{LN}(\hat{\tau}_i^v, \sigma_1^2)$ 中随机抽取,同时设定 $\sigma_1^2 = 1$, $\hat{\tau}_i^v$ 向 $\hat{\tau}_i^{v+1}$ 转移的函数式为:

$$p(\hat{\tau}_i^v, \hat{\tau}_i^{v+1}) = \min \left\{ \frac{L(\hat{\beta}^v, \hat{d}^v; \hat{\tau}_i^{v+1})}{L(\hat{\beta}^v, \hat{d}^v; \hat{\tau}_i^v)}, 1 \right\}$$

(2) 项目参数 $\{\hat{\beta}, \hat{d}\}$

$\{\hat{\beta}^{v+1}, \hat{d}^{v+1}\}$ 分别从对数正态分布 $\text{LN}(\hat{\beta}^v, \sigma_2^2)$ 和对数正态分布 $\text{LN}(\hat{d}^v, \sigma_3^2)$ 中随机抽取,且设定 $\sigma_2^2 = \sigma_3^2 = 1$ 。则 $\{\hat{\beta}^v, \hat{d}^v\}$ 向 $\{\hat{\beta}^{v+1}, \hat{d}^{v+1}\}$ 转移的概率式为:

$$p(\{\hat{\beta}^v, \hat{d}^v\}, \{\hat{\beta}^{v+1}, \hat{d}^{v+1}\}) = \min \left\{ \frac{L(\hat{\beta}^{v+1}, \hat{d}^{v+1}; \hat{\tau}_i^v)}{L(\hat{\beta}^v, \hat{d}^v; \hat{\tau}_i^v)}, 1 \right\}$$

表 1 参数返真性(平均 MSE 和 BIAS)

参数	平均指标	J = 20			J = 50		
		N = 200	N = 500	N = 1000	N = 200	N = 500	N = 1000
τ	MSE	0.132	0.139	0.137	0.060	0.056	0.054
	BIAS	-0.007	0.019	-0.001	-0.004	0.008	-0.006
d	MSE	0.097	0.048	0.014	0.097	0.040	0.021
	BIAS	-0.063	-0.018	-0.017	-0.084	-0.036	-0.028
β	MSE	0.031	0.020	0.020	0.031	0.019	0.014
	BIAS	0.052	0.039	0.039	0.055	0.036	0.027

表 1 中结果反映了被试量与测验长度对心理测量学 SAT 模型的参数估计精度的影响情况。项目参数和随着被试量的增加, MSE 和 BIAS 值都明显降低;被试参数随着测验长度增加, MSE 和 BIAS 都明显降低。结果说明 MCMC 估计方法对模型参数的估计是非常有效的。

3 心理测量学 SAT 模型合理性分析

3.1 心理测量学 SAT 模型理论逻辑分析

根据模型(2)及设定相应条件后的参数估计结果,项目正确作答概率与被试流畅性水平和项目时间压力的理论关系分别如图 2 与图 3 所示。

图 2 表示,在时间压力 $\ln(d) = 0$,两项目区分度分别为 $\beta = 0.5$ 和 $\beta = 1.5$ 时,不同流畅性水平被

其中 $\hat{\tau}_i^v$ 是被试流畅性水平 $\hat{\tau}$ 的第 v 次取样状态, \hat{d}^v 指项目时间压力参数 \hat{d} 的第 v 次取样状态, $\hat{\beta}^v$ 则是项目速率参数 $\hat{\beta}$ 的第 v 次取样状态。

2.4.2 模拟研究及结果

为了验证模型参数估计的可行性与精度,采用 Monte Carlo 方法进行模拟研究。首先,被试参数分布为 $\ln(\tau) \sim N(0, 1)$, 取值范围为 -3 到 3 之间,项目时间压力参数 d 分布为 $\ln(d) \sim N(0, 1)$, 取值范围为 -3 到 3 之间,项目速率参数 β 分布为 $\beta \sim \text{LN}(0, 1)$, 取值为 0.1 到 2.5 之间;其次,通过公式(3)计算理论加工时间,由于被试在作答过程中受到权衡的影响,观测时间以 $t_{ij} = t^* + U(-t^*, t^*)$ 生成,其中 U 指均匀分布,当均匀分布生成的值在 $(-t^*, 0)$ 之间时,意味被试在提前作答,追求速度,当均匀分布生成的值在 $(0, t^*)$ 之间时,则被试愿意花费更多的时间追求准确率,由此生成作答时间矩阵;然后,通过公式(2)计算 p 与随机数比较,生成作答矩阵。

模拟研究涉及两个变量,分别为被试量与测验长度,其中被试量分为 $N = 200, 500, 1000$, 而测验长度分别为 $J = 20, 50$, 每种条件重复 50 次。参数返真性指标采用均方误差 MSE 和相对偏差 BIAS。

试在两道项目上的正确作答概率。从图中可以看出,流畅性水平 τ 越高,项目正确作答概率越高;同时,项目区分度 β 越高,项目特征曲线越陡峭,表现出越强的区分作用或对流畅性水平的变化越敏感。

图 3 表示,当固定被试的作答时间,设定被试流畅性水平 $\ln\tau = 0$ 和 $\ln\tau = 2$, 且项目区分度 $\beta = 1$ 时,随着项目时间压力参数 d 的增大,被试的项目正确作答概率逐渐降低;同时可以看出,对相同的时间压力,流畅性水平越高,正确作答概率越高,这是由于时间压力越大,被试作答该项目所需时间越多,而对于时间压力越低的项目,流畅性水平越高则时间就越充分,所以正确作答概率就越大,也就是说,项目的时间压力参数 d 与被试正确作答概率 p 成反比

关系。

为了解释被试作答时的权衡关系,参照认知实验中的控制时间设置,在估计项目参数、被试流畅性水平后,模拟了被试在7个不同的观测时间点上(分别用A、B、C、D、E、F、G代表)的项目作答结果,然后绘制时间差与准确率关系权衡曲线,最后,依据被试在权衡曲线上的实际位置,作为被试作答时间与准确率权衡发展趋势的评价依据,示例如图4。图中,项目时间压力参数 $\ln(d) = 0$, 速率参数 $\beta = 1$, 选择了三个流畅性水平分别为 $\text{Intau} = -1$ 、 $\text{Intau} = 0$ 和 $\text{Intau} = 1$ 的被试。为便于解释,将关系图分割成四个象限。在百分制中,通常80分被定义为优秀的起始值,于是将0.8定义为高准确率,研究者也可以根据实际情况定义不同的高准确率标准,观测作答与期望作答时间差为0被当作被试作答时间权衡的

一个分割点。根据这两个标准将图形分成四个象限。在四个象限中,第一象限上的作答结果被定义为慢而好,第二象限作答结果被定义为又快又好,第三象限作答结果为快而差,而第四象限作答结果则是又慢又差。图4表示了被试在不同的模拟观测时间点上的作答准确率,当然,在实际作答中的观测作答时间点只有一个,但还是可以根据模型参数的估计结果模拟被试在多个观测作答时间点上的作答结果,并以此作为评价每个被试作答每个项目过程中的权衡状态。如图4中右端 $\text{Intau} = 1$ 的被试,如果该被试的实际观测作答时间点位于A点,可以认为他过分追求作答速度,而如果他的实际观测作答时间点位于E、F或G点,则可以认为他作答太谨慎,过分追求准确率。

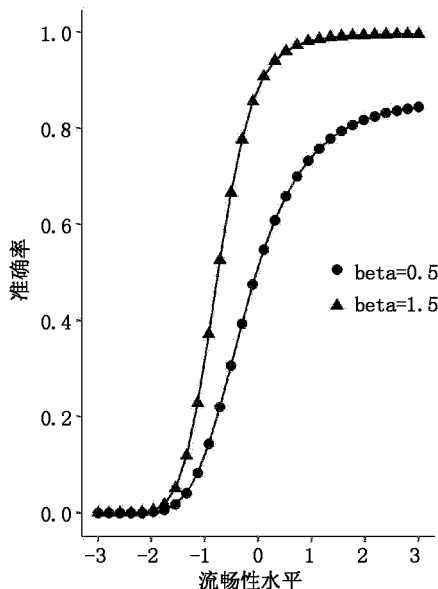


图2 流畅性水平与准确率关系

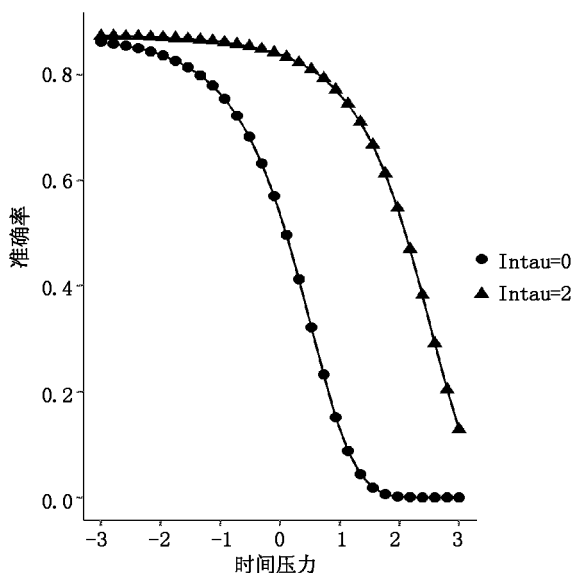


图3 时间压力与准确率关系

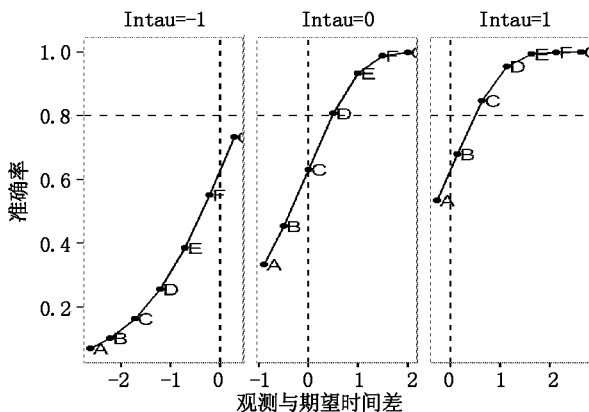


图4 时间与准确率关系

3.2 模型与实测数据拟合情况分析

理论模型与实测数据之间的拟合检验是验证模

型合理性的关键。如果理论模型与实测数据之间能够很好拟合,则说明实测数据能够通过理论模型进行描述,或者说,理论模型能够很好地解释实测数据结果。

为了检验模型与实测数据的拟合情况,采集了瑞文标准推理测验(张厚燾,王晓平,1985)实测数据进行分析。在测验设计时,期望测试中被试作答存在明显的速度与准确率权衡行为,而不是纯能力测验或者纯速度测验,测验指导语要求被试又快又准确地作答,同时对作答又快又准确的被试承诺给予一定奖励。被试选自两所县城高中的学生,数据收集程序采用 E-prime1.1 编写,开始时由主试在机房主机上统一操作演示与讲解,之后再由学生进行作答,每次测试安排三个主试机动。测验长度为

60 题,控制测验最长作答时间为 45 分钟。共施测 340 名被试,排除无效被试 20 名。参数估计采用两条链,每条链长为 20000, burn-in 设置为 10000, 参数拟合收敛采用潜在量尺缩减因子 (potential scale reduction factor, PSRF) (Brooks & Gelman, 1998), 通常 $PSRF < 1.1$ 或 1.2, 研究选择参数的 PSRF 均要求小于 1.1 来表示参数估计已经拟合。

在参数估计后,首先,将被试按流畅性水平大小分成九个组,获取每个组的被试在每个项目上的平均作答时间与平均正确作答比例(实际观测值),以及每组被试在每个项目上的理论正确作答概率(通过模型参数估计得到);然后,描述实际观测值与理论正确作答概率之间的关系,以此为基础,评价理论

模型与实测数据之间的拟合情况。

模型与数据拟合情况主要通过项目水平上,作答准确率与被试流畅性水平理论与观测曲线来评价,如图 5 所示。图形的两条直线表示实际观测值与理论值的散点趋势线,两条直线越一致,则表明实际观测结果与理论值越一致。

在准确率与流畅性水平的关系曲线上,任意选择了测验中的四个项目,分别绘制了四个项目(第 2、12、36、53 题)的观测与期望曲线图,0 指的是观测曲线,E 则是期望曲线。从第 2 题到第 53 题,项目的时间压力参数是逐渐增大的,但是四个项目的理论与观测曲线都非常趋近,而且图形也呈现出随着流畅性水平增大,准确率提高的趋势。

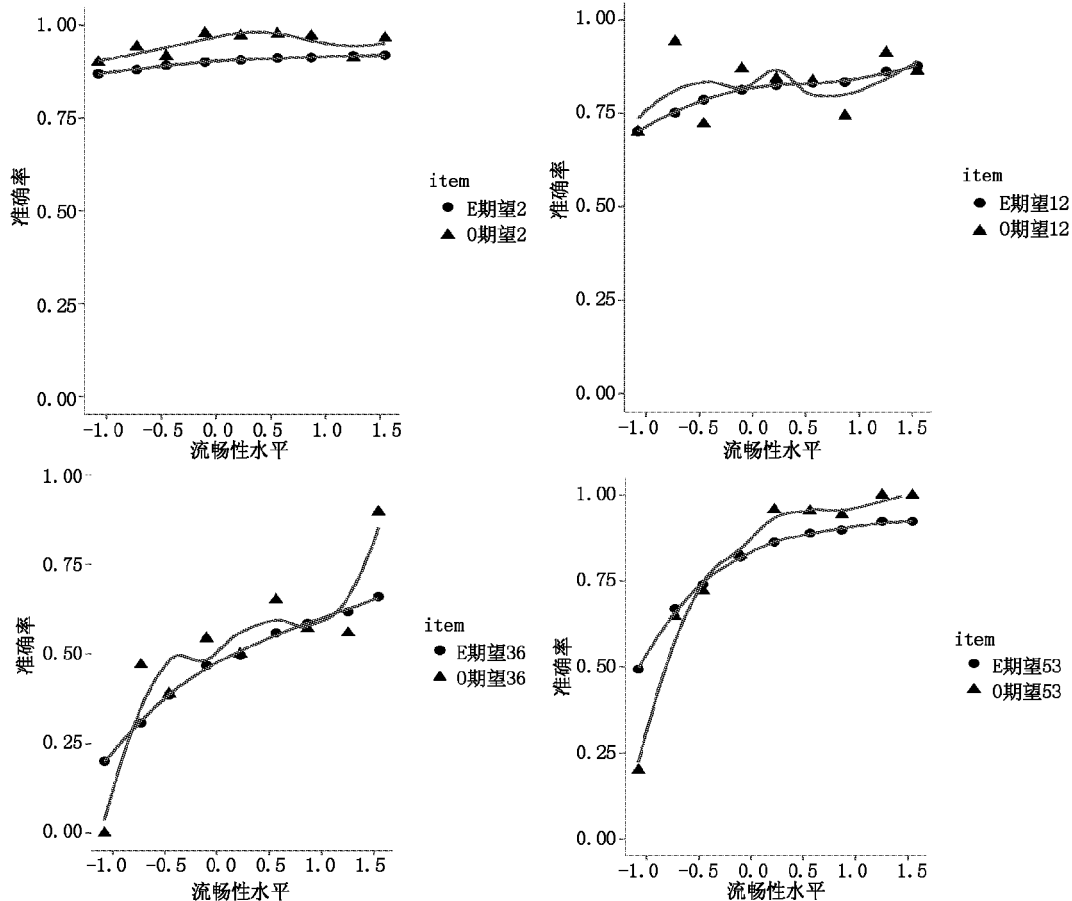


图 5 不同项目流畅性水平与准确率关系

另外,采用 Yen 统计量对 SAT 模型与实测数据拟合情况进行评价,如式(4)所示。Yen 统计量服从自由度为 $m - k$ 的 χ^2 分布,其中 m 为组数, k 为项目参数个数, O_{ij} 是类 j 对项目 i 的正确作答比例, E_{ij} 是类 j 对项目 i 正确作答的理论比例, r_j 是类 j 的被试数。

$$\text{Yen 统计量} = \sum_{j=1}^m \frac{r_j (O_{ij} - E_{ij})^2}{E_{ij} (1 - E_{ij})} \quad (4)$$

选择上文中的四个测验项目(2, 12, 36, 53),通过计算四个项目的 Yen 统计量与 $\chi^2_{0.01(7)} = 18.48$ 临界值进行比较,结果发现四个项目的 Yen 统计量的值分别为 10.68、8.61、14.20、14.56,都明显要小于临界值 18.48。在所有 60 个项目上,小于临界值的项目为 52 个,超过临界值项目为 8 个。在 8 个非拟合项目中, χ^2 值低于 18.8 的项目占 3 个; χ^2 值在 20 左右有 4 个项目,只有项目 59 明显超过临界值。

综合而言,通过观测曲线与期望曲线的一致性以及 Yen 统计量检验的结果,同时考虑模型稳健性的特点,可以认为模型与实际数据拟合结果良好。

对于被试的权衡状态可以采用两个方法进行评价。方法一,绘制被试在所有项目上观测时间和期望时间差与准确率的散点图,根据各散点位于 $X=0$ 与 $Y=0.8$ 的四个区间的聚集趋势进行评价;方法二,首先,基于理论模型和参数估计结果,模拟多个观测时间点,绘制模拟观测时间点和期望时间差与准确率的关系曲线,然后,通过实际观测时间与准确率的点在曲线的位置,推断被试在项目上的权衡倾向。

对于第一种评价方法,选择了两个流畅性水平非常接近的被试,绘制他们在全部及部分试题上时间差与准确率关系散点图,分别如图6左图与右图所示。在图中,两个被试流畅性水平皆为 $\text{Intau} = -$

0.07,两虚线分别对应 $Y=0.8, X=0$ 。在左图中 Intau1 被试,第一象限到第四象限项目数分别为15、28、12、5,从阴影部分也反映出第二象限项目数占明显优势,也就是又快又好,所以该被试整体上在准确率与速度上进行了较好的权衡;对被试 Intau2 ,其第一象限到第四象限的项目数分别为38、9、10、3,从阴影部分也可以看出第一象限项目数占明显优势,也就是好却慢,也就是该被试整体上更多追求的是正确率。为了进一步分析两个被试的权衡特征,选择了项目参数非常接近的14个项目,绘制了图9中的右图。在右图上,通过两条参考线,明显可以看出,被试 Intau2 的各点都在第一象限,即倾向准确率,观测时间基本都高于期望时间,正确率也较高;而被试 Intau1 的各点倾向第二象限,观测时间基本低于期望时间,同时也具有较高的正确率。

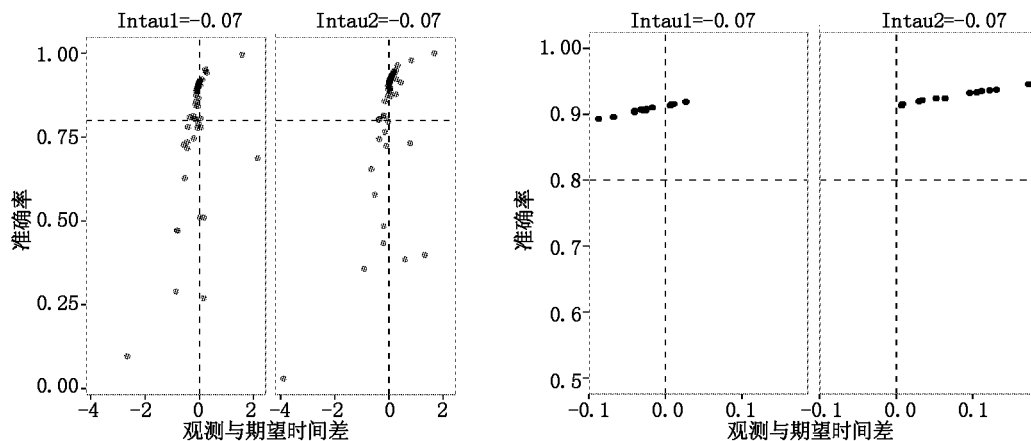


图6 相同流畅性水平时间与准确率权衡

对于第二种评价方法,选择了流畅性水平分别为 -0.385 和 0.555 的两个被试,同时选择了四个有代表性的项目(β 参数分别为 1.10 、 2.36 、 1.45 、 0.134 , 时间压力参数 d 分别为 0.393 、 -0.741 、 1.48 、 -1.80), 每个项目均模拟了6个观测时间点,加上实际观测时间,共7个时间点,绘制观测时间和期望时间差与准确率关系散点和曲线,如图7所示。在图中, E 对应的是模拟观测时间与准确率的散点, O 为实际观测时间与准确率的散点, 数字1-4 分别代表四个不同项目, 两条虚线分别为 $X=0, Y=0.8$ 。

对于流畅性水平为 -0.385 的被试, 项目1 的观测时间点 O1 处的准确率接近达到最大值, 同时观测时间明显高于期望加工时间, 说明被试在项目1 上作答比较谨慎, 更倾向于提高准确率, 属于好却慢; 而在项目2 上, 观测时间点 O2 处的准确率仍然

有较大的增长空间, 被试此时做出了决策, 不过被试的准确率仍然达到了 0.8 以上, 同时观测时间低于期望时间, 说明被试在项目2 上是又快又好。在项目3 和项目4 上, 两个项目的时间压力都偏大, 准确率偏低, 但是两个项目权衡曲线的陡缓程度存在差异, 这是由项目的速率参数决定。在项目3 上, 速率参数较小, 权衡曲线增长缓慢, 被试很难在合理的时间内完成作答, 被试提前做出了决策, 属于快而差; 而项目4 的时间压力大且速率参数也偏大, 权衡曲线陡峭, 虽然有较好的上升空间, 但观测时间已经高于期望时间, 被试的努力尝试仍未达到较高的准确率, 权衡收益相对成本不占优势, 在追求一定准确率后做出了决策, 属于又慢又差。

对于流畅性水平为 0.555 的被试, 情形与上一被试非常相近, 只是在项目1 上速度与准确率权衡较好。

速度与准确率权衡是被试作答项目过程中的一个基本现象,每个被试对每个项目的权衡倾向可能

都存在差异。

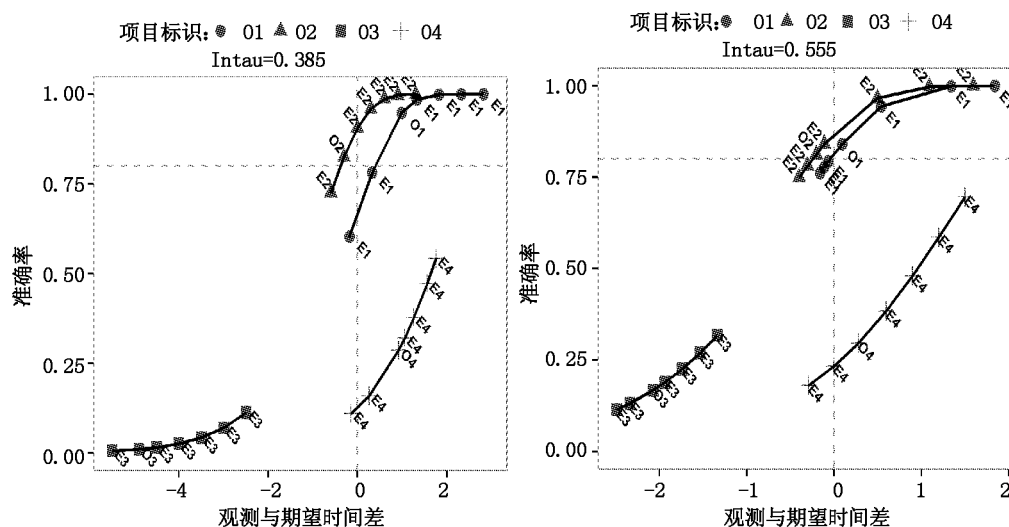


图7 不同流畅性水平时间与准确率权衡

4 讨论

目前,关于反应时的研究正在逐步受到关注。但是已有的反应时模型都未涉及被试作答过程中速度与准确率权衡现象的处理,更多的是单纯针对反应时信息而言,而速度与准确率权衡是被试决策过程中的一个基本现象,直接影响到被试的作答时间与作答准确率。因此,开发一个可以应用于大规模评价中的速度与准确率权衡的心理测量学模型将具有重要的价值。

反应时能反映出被试作答项目时的重要信息,能从不同角度评价被试完成任务时的决策过程。研究以认知心理学实验的速度与准确率权衡模型为基础,通过理论逻辑分析和数据分析,提出了一个新的心理测量学的速度与准确率权衡模型。

新模型借鉴已有的认知心理学实验研究成果和心理测量学研究成果,首先,它能够使速度与准确率权衡模型应用于更复杂的认知任务情景中,实现认知心理学实验成果从实验室走向大规模评价应用中;另外,相对已有的测量学模型而言,新模型考虑了速度与准确率及其权衡关系。而且,新模型并不依赖于反应时的分布,它是基于被试作答过程的权衡机制构建,从而能够针对被试作答项目的不同权衡与反应进行区别对待,而不是像现有的反应时模型将不同作答结果等效处理。在正确的项目作答时间上,现有的模型(van der Linden, 2006; Ranger, et al, 2012, 2013; 等等)和文中模型都是作为有效作答信息进行处理,但是速度与准确率权衡模型同时兼顾考虑了被试在作答过程中的权衡状态;而对于错

误的项目作答时间,现有的反应时模型仍然将其作为完全有效数据进行处理,也就意味着在同一个项目上,作答正确的时间与作答错误的时间代表同样的意义,而文中速度与准确率权衡模型则将错误作答时间识别为部分有效时间,也就是说,被试实际所花费的作答时间与正确作答该项目所需的时间是有差距的,从认知实验角度而言,项目正确作答的时间是必须经历正确提取项目考察知识结点和相关联结以及应用的时间过程,而错误作答项目只经历其中部分过程,同时兼顾速度与准确率权衡,将被试识别为不同的权衡状态。为了验证速度与准确率权衡模型,通过模拟分析和实测数据分析发现,新的基于速度与准确率权衡的心理测量学模型参数能非常稳定而又精确地被估计,同时能很好地应用于大规模心理测验任务解决中。

人们在解决现实问题并做出决策时,反映了被试三个方面的心理特征与品质,分别为标志知识技能完备性的能力水平、标志知识技能熟练度的流畅性水平、标志速度与准确性权衡标准的倾向性(Kagan, Rosman, Day, Albert, & Phillips, 1964; Grigorenko & Sternberg, 1995)。能力与流畅性水平都是被试潜在特质的不同方面,只是能力反映被试作答准确率高低,而流畅性水平反映被试作答时间长短。有些被试可能三者均有很好的表现,而有的被试可能表现更加冲动,反应快,但是精确性差,而有些被试喜欢反复验证,反应慢,但精确性水平更高。心理测量模型可以应用于有时间限制条件下的问题处理情景,比如在大规模被试能力水平评价项目中,被试作

答同时需要考虑时间与结果的准确性问题。而对于具有充足作答时间的能力测验(纯能力测验)或者只考察速度的速度测验(纯速度测验),速度与准确率权衡模型也能适用,但是由于被试在纯能力测验上,不需要太在意时间的影响,而纯速度测验,测验项目普遍偏向简单,则被试在作答过程中,就不需要特别在意权衡速度与准确率的倾向,难以激发被试在作答过程中的权衡行为,也就无法发挥速度与准确率权衡模型的优势。在测验项目上,项目所考察知识技能点的复杂度一方面表现为项目难度,另一方面是项目的时间压力水平。对于项目时间压力水平,在项目时间不变前提下,项目越复杂,项目时间压力水平越高;当项目复杂度相同,项目设定时间越短,项目时间压力水平越高。项目时间压力是项目复杂度与项目设定时间的结合。在整个测验上,测验的时间设置,标准的把握都是基于被试评价的实际需要。针对相同的测验,当目标是评价被试知识技能完备性,也就是能力时,测验的时间就需要尽力充足,以避免因为时间不足或者知识运用不熟练,影响作答结果;当目标是评价被试知识技能熟练度,也就是流畅性水平时,需要压缩测试时间,给予被试足够的时间压力,以免因为时间设置不当,时间压力过低;最后当评价目标为被试综合素质,也就是能力与流畅性水平兼顾评价时,则需要设置时间压力不是特别大,同时时间又不充裕时,由被试在作答过程中的权衡倾向,综合考察被试的素质。因此,在模型数据收集与测验项目选择上,针对不同的被试评价目标,测验的时间设置需要慎重考虑。

当然,还有许多问题需要今后进一步探讨。比如,权衡指标的分析,虽然通过被试的观测时间和期望时间差与准确率结合分析,可以将被试的权衡倾向进行识别,但是将权衡倾向整合成一个统一的评价指标,或者以参数的形式在模型中反映出来,将更便于了解与分析被试的权衡状态。另外,如何在模型中整合被试能力水平是今后需要进一步研究的问题。

参考文献

- 罗照盛. (2012). 项目反应理论基础. 北京: 北京师范大学出版社.
- 孟祥斌. (2016). 项目反应时间的对数偏正态模型. 心理科学, 39(3), 727 - 734.
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434 - 455.
- Carrasco, M., McElree, B., Denisova, K., & Giordano, A. M. (2003). Speed of visual processing increases with eccentricity. *Nature Neuroscience*, 6(7), 669 - 700.
- Carrasco, M., Giordano, A. M., & McElree, B. (2005). Attention speeds processing across eccentricity: Feature and conjunction searches. *Vision Research*, 46(13), 2028 - 2040.
- Doshier, B. A. (1976). The retrieval of sentences from memory: A speed - accuracy study. *Cognitive Psychology*, 8, 291 - 310.
- Doshier, B. A. (1984). Degree of learning and retrieval speed: Study time and multiple exposures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 541 - 574.
- Giordano, A. M., McElree, B., & Carrasco, M. (2009). On the automaticity and flexibility of covert attention: A speed - accuracy trade - off analysis. *Journal of Vision*, 9(3), 1 - 10.
- Grigorenko, E. L., & Sternberg, R. J. (1995). Thinking styles. In D. H. Saklofske & M. Zeidner (Eds.), *Perspectives on individual differences*. International Handbook of Personality and Intelligence.
- Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs: General and Applied*, 78(1), 1 - 37.
- 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.
- McElree, B., Jia, G. X., & Litvak, A. (2000). The time - course of conceptual processing in three bilingual populations. *Journal of Memory & Language*, 42, 229 - 254.
- McElree, B. (1998). Attended and non - attended states in working memory, accessing categorized structures. *Journal of Memory & Language*, 225 - 252.
- McElree, B. (2000). Sentence comprehension is mediated by content - addressable memory structures. *Journal of Psycholinguistic Research*, 29, 111 - 123.
- McElree, B., & Carrasco, M. (1999). The Temporal Dynamics of Visual Search, Evidence for Parallel Processing in Feature and Conjunction Searches. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1517 - 1539.
- McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48, 67 - 91.
- McElree, B., & Carrasco, M. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of The National Academy of Sciences of The United States of America*, 98(9), 5363 - 5367.
- Meng, X. B., Tao, J., & Shi, N. Z. (2014). An item response model for Likert - type data that incorporates response time in personality measurements. *Journal of Statistical Computation and Simulation*, 84, 1 - 21.
- Ranger, J., & Ortner, T. (2012a). A latent trait model for re-

- sponse times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, 65, 334 – 349.
- Ranger, J. , & Ortner, T. (2013). Response Time Modeling Based on the Proportional Hazards Model. *Multivariate Behavioral Research*, 48, 503 – 533.
- Ranger, J. , & Kuhn, J. T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, 67, 388 – 407.
- Ranger, J. , & Kuhn, J. T. (2015). Modeling Information Accumulation in Psychological Tests Using Item Response Times. *Journal of Educational and Behavioral Statistics*, 40(3), 274 – 306.
- Reed, A. V. (1973). Speed – accuracy trade – off in recognition memory. *Science*, 181, 574 – 576.
- Reed, A. V. (1976). List length and the time course of recognition in human memory. *Memory & Cognition*, 4, 16 – 30.
- 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. (2008). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287 – 308.
- van der Linden, W. J. (2009). Conceptual issues in response – time modeling. *Journal of Educational Measurement*, 46, 247 – 272.
- Wang, C. , Chang, H. H. , & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, 66, 144 – 168.
- Wang, C. , Fan, Z. W. , Chang, H. H. , & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, 38, 381 – 417.
- Wickelgren, W. (1977). Speed – accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41, 67 – 85.

A Psychometric Model for Speed – accuracy Tradeoff and Application

Guo Xiaojun Luo Zhaosheng

(Psychology College, Jiangxi Normal University, Nanchang 330022)

Abstract: The accuracy of completing a task has always been the main evaluation index in the large – scale assessment. However, during a variety of task situations, all the indexes indicating the quality of the doers are extremely important, including the correctness of the result as well as the timeliness of the decision – making process. Therefore, the precision of results and fluency level of reaction should be regarded as the two indispensable indexes that evaluate the quality of each individual in finishing tasks. Grounded on the research of speed and accuracy in cognitive experiment, the manuscript will make the cognitive experiment of the speed – accuracy tradeoff model out of the lab and make it a large – scale assessment of model that can be applied to more complex situations of cognitive tasks by building a psychometric model based on the speed – accuracy tradeoff. The new psychometric model of parameters based on speed – accuracy tradeoff can be estimated very stably and accurately. At the same time, the variables of model and their relations can be well supported by the real data. Finally, the quality of the subjects can be evaluated from different methods by the speed – accuracy tradeoff model.

Key words: speed – accuracy tradeoff; large – scale assessment; cognitive experiment; psychometric model