心理学探新 2023,Vol.43,No.3,278 - 288
PSYCHOLOGICAL EXPLORATION

基于项目作答反应时间的作弊甄别研究进展

杨志明 徐庆树*
(湖南师范大学外国语学院,长沙 410081)

摘 要:随着基于计算机的测试逐渐普及,搜集、记录、分析考生的项目作答反应时间数据成为可能,越来越多的研究者开始基于这一数据开展考试的作弊甄别研究。然而,此类研究国外较多,国内则非常之少。提出“两种三类”的作弊行为分类标准,从参数建模法、非参数建模法两个维度,对基于项目作答反应时间的作弊研究进行梳理,评述其在甄别各类作弊行为中的应用实践和甄别效果,并对未来的研究方向做出展望。

关键词:项目作答反应时间;作弊甄别;个人拟合
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003 - 5184(2023)03 - 0278 - 11

1 引言

作弊甄别研究一直是考试研究的重点课题,考试中的作弊现象不仅干扰了考试秩序(胡佳琪 等,2020;骆方 等,2020),影响了考试的信效度,影响评分标准或合格线划定(Sinharay,2021),而且违反国家有关考试的法律法规(Crittenden,Hanna,& Peterson,2009;彭恒利,2015)。一直以来,大量的研究者采用心理计量学、统计学的研究方法甄别各种作弊现象(Cizek & Wollack,2017;Sinharay,2017;van der Linden & Lewis,2015)。近年来,随着基于计算机的测试(computer based test,CBT)的快速发展,考生在考试中的项目作答反应时间可以被搜集、记录、使用(Zopluoglu,2019)。正如 van der Linden 所说,项目作答反应时间一直以来都被看作是识别个人行为的重要信息源,但只在基于计算机的测试普及后,这一信息源才能得到解码使用(van der Linden,2006)。

最近 10 年间,越来越多的研究者开始尝试使用考生的项目作答反应时间甄别考生的异常作答情况(Man & Harring,2021;Nagy & Ulitzsch,2021)。然而,本研究发现国外研究者应用此类方法较多,而国内研究者应用此类方法较少。为更好的介绍这种作弊甄别方法,本文首先对考试作弊现象进行界定与分类,接着介绍基于项目作答反应时间建模的研究假设、建模方法和个人数据拟合方法,然后进一步阐释各类建模方法、拟合方法在不同类型作弊行为甄别中的应用和实践,最后对这种作弊甄别方法中存在的相关问题开展总结和讨论。

2 考试作弊现象的界定与分类

考试作弊呈多发态势(彭恒利,2015),作弊形式趋于多样化,作弊工具趋于高科技化、作弊行为向团伙化、集团化、专业化方向发展,考试作弊逐渐形成产业链,向商业化运作模式转变。面对当前纷繁复杂的作弊现象,没有任何一种研究方法或者统计模型可以全部适用。因此,开展考试作弊现象研究,首先要对考试作弊行为进行界定、分类(Sinharay,2020)。

基于作弊主体的分类较多。有研究者基于作弊主体的规模将作弊行为分为个人作弊行为和团体作弊行为(胡佳琪 等,2020;骆方 等,2020)。我国政府对作弊行为进行的分类主要也是基于作弊主体开展的。教育部颁布的《国家教育考试违规处理办法》界定了 3 类作弊行为,分别是:考生的作弊行为、考试工作人员的作弊行为、组织作弊的行为。《刑法修正案(九)》按作弊主体分别做出作弊犯罪的处罚规定,其中对组织作弊者的处罚规定为:“在法律规定的国家考试中,组织作弊的,处三年以下有期徒刑或者拘役,并处或者单处罚金;情节严重的,处三年以上七年以下有期徒刑,并处罚金。”

当然,也有研究者基于作弊行为的分布特点、动机和发生场景进行了分类。Cizek 提出 3 大类 59 种作弊行为,其中最常见的是抄袭(Cizek,1999)。从作弊动机的角度,Cizek 进一步将作弊行为分为 2

* 通讯作者:徐庆树,E - mail:xuqingshu@hunnu.edu.cn。

类:一类是考试欺诈(test cheating),主要指需要参加考试并获得成绩的个人或者群体进行的作弊行为;另一类是考试盗窃(test theft),主要指作弊者无需参加考试、从考试作弊中牟利的行为(Cizek & Wollack,2017)。Wollack 等按照作弊发生的场景将作弊行为分为3类:第一类是答案抄袭与共谋作弊(answer - copying and collusion),指的是作弊考生单独或与其他考生协作,进行答案抄袭的现象;第二类是泄题(item preknowledge),指的是考生通过各种手段在交卷前获取考试的项目信息现象;第三类是篡改答案(test tampering),指的是考生、教师、考试机构工作人员等通过修改考生答案的方式进行的作弊行为(Wollack & Fremer,2013)。

本研究在前述作弊分类方案的基础上进行总结,提出"两种三类"的作弊行为分类方法:"两种"指的是作弊主体,也就是:个人作弊、团体作弊,两种作弊行为;"三类"则分别指的是:泄题、抄袭和篡改答案。

3　基于项目作答反应时间的作弊甄别方法

基于项目作答反应时间的作弊甄别方法基本可以分为2类:一类是参数法,一类是非参数法。参数法的基本假设为:假定项目作答反应时间是一个连续变量,该变量的分布规律符合某种特定分布,如对数正态分布(lognormal distribution)(van der Linden,2006;van der Linden,2008;van der Linden,2009)、伽马分布(gamma distribution)(Verhelst, Verstralen, & Jansen,1997)、指数分布(exponential distribution)(Scheiblechner,1979,1985)、偏正态分布(孟祥斌,2016)等,研究者基于项目作答反应时间的这种分布特点构建模型。在个人数据拟合中,研究者再将模型的预测值与实际观测值进行比较,如果二者相差过大,就怀疑考生或有作弊、试题或被泄露(Qian, Staniewska, Reckase, & Woo,2016;van der Linden & Guo,2008)。非参数法则另做假定:考生的项目作答反应时间是一个离散变量,单个题项的项目作答反应时间的分布可以和考生总体的项目作答反应时间进行对比。基于这种假设,研究者就可省去参数建模的步骤,直接采用如KL离散度等方法比较考生个体和考生群体的作答反应时间模式,进行作弊甄别,也取得了不错的效果(Man, Harring, Ouyang, & Thomas,2018)。

3.1　参数建模法

参数建模法主要有3类参数模型:一是基于作答反应时间的模型,这类模型以对数正态模型应用最为广泛;二是基于项目作答反应时间和作答正误情况联合建模的层次框架模型;三是在第二类模型基础上增加眼动等其他过程数据的联合建模。

3.1.1　对数正态模型

van der Linden(2006)采用对数正态模型对考试中的项目作答反应时间进行建模,这种建模方法拟合优度较高,获得了广泛认可。此方法假定考试中的项目作答反应时间具有随机性,且呈对数正态分布(lognormal distribution),因此借鉴项目反应理论的两参数模型的建模方法,提出项目作答反应时间的对数正态分布模型(Lognormal Model),构建考生做题速度参数τ、题项参数β_i、α_i,用以拟合考生作答反应时间的分布情况,公式如下:

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

其中,t_i代表考生t在进行题项i作答时所用的时间,τ则代表了考生的做题速度,τ取值越大,则考生在该题项上花费的时间越少。β_i则代表了试题的时间消耗度(time intensity),β_i的取值越大,考生在该试题上花费的时间就越多。而α_i如同在项目反应理论中一样,是一个作答时间区分度指标,取值大于0,且取值越大,考生在第i个题的作答时间的对数分布越集中;取值越小,则越分散。因此,α_i这一指标可以实现以题项为单位区分做题速度不同的考生。

模型背后的原理可以对照对数正态模型的公式(公式2)进行解读:

$$f(x,\mu,\sigma) = \begin{cases}\dfrac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\dfrac{1}{2\sigma^2}(\ln x-\mu)^2\right], x>0\\ 0, x\leqslant 0\end{cases},\quad(2)$$

对照可知,项目作答反应时间对数正态分布模型的均值和标准差分别为:

$$\mu_i = \beta_i - \tau,\quad(3)$$

$$\sigma = \frac{1}{\alpha_i}。\quad(4)$$

公式1中的t_i(考生t在第i题的做题时间)的对数$\ln t_i$服从正态分布:其均值由题项的时间消耗度β_i和考生的做题速度τ决定(公式3);标准差为试题的作答时间区分度α_i的倒数(公式4)。

为了提高模型的可识别性,van der

Linden(2006) 对公式 1 中提到的变量进行了约束。速度参数 τ,应该符合公式 5:

$$\sum_{j=1}^{N}\tau_j = 0, \quad (5)$$

也就是说,第 j 个人的做题速度可以取正值或者负值,所有人的做题速度之和等于 0。

结合公式 3 可知,第 j 个考生在第 i 个题项上的作答反应时间的对数分布模型的均值等与第 i 个题项的时间消耗度 β_i 减去第 j 个考生的做题速度 τ_j,公式如下:

$$\beta_i - \tau_j = \mu_{ij}, \quad (6)$$

由此可知:

$$n^{-1}\sum_{i=1}^{n}\beta_i - N^{-1}\sum_{j=1}^{N}\tau_j = \left((nN)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N}\mu_{ij}\right), \quad (7)$$

套入公式 5 的约束,可以得到:

$$n^{-1}\sum_{i=1}^{n}\beta_i = \left((nN)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{N}\mu_{ij}\right), \quad (8)$$

这就是说试卷中所有试题的时间消耗度的均值等于所有考生作答所有题目所需的项目反应时间的对数正态分布的均值。

van der Linden(2006) 认为可以采用 MCMC 方法(吉布斯采样) 对这个模型进行参数估计。但是,同项目反应理论一样,在参数估计前需要先进行局部独立性假设:一是,考生个体之间具有局部独立性;二是,给定考生作答所有试题的项目作答反应时间之间具有局部独立性。在此基础上,考生的参数 τ_j 符合正态分布:

$$\tau_j \sim N(\mu_\tau, \sigma_\tau^2)。 \quad (9)$$

同时,他指定试题参数作答时间区分度 α_i 服伽马分布(公式 10),题项时间消耗度 β_i 服从正态分布(公式 11):

$$\alpha_i \sim G\left(\frac{v}{2}, \frac{v}{2\lambda}\right), \quad (10)$$

$$\beta_i \mid \alpha_i \sim N[\mu_\beta, (\alpha_i^2\kappa)^{-1}]。 \quad (11)$$

如公式 11 所示,β_i 的正态分布标准差中含有参数 α_i,且 α_i 服从 gamma 分布,因此,(β_i, α_i) 服从 normal - gamma 分布。

在 Gibbs 采样中,交替给定考生参数 τ 和试题参数 (α, β) 按照公式(12) 进行参数估计:

$$f(\tau, \alpha, \beta \mid t) \propto \prod_{j=1}^{N}\prod_{j=1}^{n} f(t_{ij}; \tau_j, \alpha_i, \beta_i) f(\tau_j; \mu_\tau, \sigma_\tau) f(\beta_i \mid \alpha_i; \mu_\beta, \kappa) f(\alpha_i; v, \lambda), \quad (12)$$

其中,$\prod_{j=1}^{N}\prod_{j=1}^{n} f(t_{ij}; \tau_j, \alpha_i, \beta_i)$ 为似然函数,$f(\tau_j; \mu_\tau, \sigma_\tau) f(\beta_i \mid \alpha_i; \mu_\beta, \kappa) f(\alpha_i; v, \lambda)$ 为先验分布。这样经过多次迭代循环,在收敛后即可完成参数估计。

这个模型的拟合优度,则可以通过检验个人和题项两类参数的贝叶斯残差进行估计。当吉布斯采样收敛后,可以对模型预测的第 j 个考生在第 i 个题项中所用的作答时间 $\widetilde{\ln t_{ij}}$,和实际观测的 $\ln t_{ij}$ 进行对比,进而确定模型的拟合优度。

3.1.2 层次框架模型

van der Linden(2009) 提出考生在作答速度和作答正误情况之间会做出权衡(speed - accuracy tradeoff),并提供示意图:

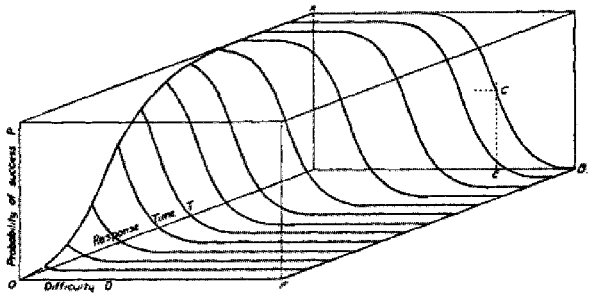

图 1 瑟斯多恩作答反应面示意图

(Thurstone's response surface)

图中横轴为试题难度、纵轴为作答时间、竖轴为作答正确率,图示作答时间越长,正确率越高。很多研究者基于此开展项目作答反应时间和项目作答正误情况的联合建模分析。De Boeck 和 Jeon(2019) 认为这种联合建模一般可以分为四类:第一类是将项目作答反应时间作为因变量,将做题正误情况作为自变量的建模(Thissen,1983);第二类则是将项目作答正误情况作为因变量而将项目作答反应时间作为自变量的建模(Roskam,1987;Verhelst et al.,1997);第三类则是将项目作答反应时间和项目作答正误情况同时作为因变量且二者之间无因果关系的建模分析 van der Linden(2007);第四类则是将项目作答反应时间和项目作答正误情况同时作为因变量且将二者之间的因果关系也考虑到数据模型中的建模方法(dependency model,本文译作“依存模型”)(Bolsinova & Maris,2016)。在所有这些模型中,van der Linden(2007) 提出的层次框架模型是使用最多且最为广泛的模型,本文重点介绍这一模型。

作答反应时间经常和考生的准确率一起联合建

模。这种建模方式的基本假设是：考生可以选择高准确率低速度答题，也可以选择低准确率高速度答题，其中速度是自变量，而准确率是因变量。根据项目反应理论，我们可以估计出能力值θ。在此基础上，van der Linden(2007) 基于作答反应时间和考生的准确率进行联合建模，提出了层次框架模型。这种建模方式主要基于以下5个关键假设：一是个体考生的做题速度相对固定；二是个体考生在作答单个题项时，其做题速度和作答对错情况均属于随机变量；三是题项的参数(如时间消耗度、时间区分度、难度、区分度) 和考生参数(做题速度、能力值) 等可以分开计算；四是在给定做题速度和做题能力的前提下，作答对错和项目反应时间之间条件独立；五是可以用样本考生的数据对考生总体的作答速度和准确率分别进行估计。

在此基础上，van der Linden(2007) 提出2层模型。第一层模型选用三参数IRT模型(正态肩形模型Normal Ogive Model或logistic模型) 对作答的对错情况进行建模，同时选用对数正态模型对项目反应时间进行建模，公式如下：

$$f(u_j,t_j;\xi_j,\psi) = \prod_{i=1}^{I} f(u_{ij};\theta_j,a_i,b_i,c_i) f(t_{ij};\tau_j,\alpha_i,\beta_i), \quad (13)$$

其中，考生j的考生参数ξ_j有2个，分别为做题速度(τ_j) 和能力值(θ_j)；试题参数ψ_i有5个，分别为区分度(a_i)、难度(b_i)、猜测指数(c_i)、时间消耗度(β_i)、时间区分度(α_i)。

第二层模型同样包含2个模型，公式如下：

$$f(u,t;\xi,\psi) = \prod_{j=1}^{J}\prod_{i=1}^{I} f(u_j,t_j;\xi_j,\psi_i) f(\xi_j;\mu_P,\Sigma_P) f(\psi_i;\mu_i,\Sigma_i), \quad (14)$$

其中，考生参数ξ_j的取自于考生总体P，其参数符合多元正态分布，公式如下：

$$\xi_j \sim f(\xi_j;\mu_P,\Sigma_P)。 \quad (15)$$

试题参数ψ_i同样取值于多元正态分布：

$$\psi_i \sim f(\psi_i;\mu_i,\Sigma_i)。 \quad (16)$$

第二层模型在第一层模型的基础上对考生总体的参数和题项间的关系进行了估计。层次模型框架(hierarchical framework) 同样采用了吉布斯采样的方式进行参数估计。

自发布以来，层次框架模型广受好评，多位研究者认为层次框架模型是一种插件模型(plugin model)(Bolsinova，Tijmstra，& Molenaar，2017；Molenaar，Bolsinova，& Vermunt，2018；Molenaar & de Boeck，2018)，研究者可以将表示作答准确率的单维项目反应模型换成多维项目反应模型或认知诊断模型(Zhan，2022；Zhan，Man，Wind，& Malone，2022；詹沛达，2019)，又或者多级计分模型(汪大勋，郭莹莹，2022)，也可以将项目作答反应时间的对数正态模型换成多维反应时间模型(Zhan，Jiao，Man，Wang，& He，2021；Zhan，Jiao，Wang，& Man，2018；詹沛达，Jiao，Man，2022)，还可以在模型中增加协变量(Qiao & Jiao，2022)，郭小军等(2022) 还进一步探讨了多维潜在特质速度之间可能存在层阶关系，并提出了高阶对数正态作答时间模型与双因子对数正态作答时间模型。Ranger(2013) 认为van der Linden的层次框架模型是标准化测验中有关考生作答和考生作答反应的标准建模操作流程，Wang(2018) 等研究者更是指出这种层次框架模型在考生作答和考生作答反应的多种统计建模中是最为流行的一种。严娟等(2022) 将这种建模方法应用到了多维人格测验中。

当然，我们也应注意到在层次框架模型以外，也有不少研究者尝试提出其他类型的模型，其中比较重要的如双向异常值检测模型(two - way outliers detection model)(Chen，Lu，& Moustaki，2019) 和线性模型(Molenaar & Bolsinova，2017；Molenaar，Tuerlinckx，& van der Maas，2015a，2015b)，尤其线性模型是在层次框架模型基础上发展而来。

3.1.3　与其他过程数据的联合模型

随着基于计算机的考试(CBT) 进一步普及，考生越来越多的生物信息(如眼动、脑电、心率) 开始被采集、记录和分析，Man和Harring(2021) 在综合对比多种生物信息和传统考试信息的基础上进行分析，提出了一种基于项目反应、作答反应时间、注视点个数的联合建模，并用这种模型分析了团体作弊行为，他们通过对335名大学生的眼动实验，验证了模型具有较好的数据拟合性，也能量化呈现不同学生群体的作答准确率、作答效率和视觉参与度，为作弊甄别提供量化依据。在疫情的影响下，这一研究也为如何开展线上考试的作弊甄别提供了一种新的解决方案。还有研究者将作答反应时间数据、作答正误数据以及作答中鼠标的拖拽和点击等数据进行联合建模(Liang，Tu，& Cai，2023)，也为作弊甄别提供了

一种新的研究方法。

3.1.4 个人拟合分析

本文前述模型需要通过个人拟合分析的方法判断考生是否作弊。在考试甄别中，基于对数正态模型和层次框架模型的个人拟合方法分别为：以标准作答为基础的索引法（Marianti，Fox，Avetisyan，Veldkamp，& Tijmstra，2014）和以层次框架模型为基础的贝叶斯残差分析法（van der Linden & Guo，2008）。

标准作答索引法（Marianti et al.，2014）是指在作答反应时间的对数正态模型的基础上，基于 Iz 统计法（Drasgow，Levine，& Williams，1985）进行个人拟合分析，查看观测到的考生项目作答反应时间数据在模型预估到的项目作答反应时间分布中的概率，并将该概率与门槛值 C 进行对比，判断偏离情况，确定是否需要标注、告警。并基于公式 1 建立公式如下：

$$l^t(T_j^*) = \sum_{i=1}^{I} Z_{ij}^2。 \tag{17}$$

其中，$Z_{ij} = \frac{(t_{ij}^* - \mu_{ij})}{\sigma_i}$（参见公式 6：$\beta_i - \tau_j = \mu_{ij}$）是一个服从正态分布的变量，$t_{ij}^* = \ln(t_{ij})$ 则代表了第 j 个考生在第 i 个题项中所用的作答反应时间的对数，$\mu_{ij} = \beta_i - \tau_j$，$\sigma_i = \frac{1}{\alpha_i}$。$l^t$ 则符合 I 度（I 代表试题的数量）卡方分布，其他指标如 α_i，β_i，τ_j 等参数的值均通过贝叶斯估计获得。l^t 的值越大，越容易被怀疑为作弊。当作弊行为绝对数较少的时候，这种方法的检出率较高，并且检出率随着样本量的增加而增加。

贝叶斯残差分析法是在层次框架的基础上提出 2 个公式，用以计算考生 j 在任意一个题项 i 作答中所用的项目反应时间的观测值与预测值的差异，并计算在贝叶斯后验分布中该观测值出现的概率。如果考生在某个题项的作答反应时间过于短，则怀疑该考生有可能提前获知题项，存在泄题（preknowledge）的问题；如果考生在某个题项的作答反应时间过于长，则怀疑该考生有可能是在背诵试题，用以对外售卖或者传播。

概率计算公式采用了反常积分（Improper Integral）的计算方法，考生在某个题项答题时间短于预测值的概率的公式如下：

$$\pi_{ij}^l = \int^{t_{ij}^*} f(\tilde{t}_{ij}^* \mid t_{j\backslash i}^*) d\tilde{t}_{ij}^* \text{ 或者} \int^{t_{ij}^*} f(\tilde{t}_{ij}^* \mid t_{j\backslash i}^*, u_{j\backslash i}) d\tilde{t}_{ij}^*， \tag{18}$$

考生在某个题项答题时间长于预测值的概率的公式如下：

$$\pi_{ij}^u = \int_{t_{ij}^*} f(\tilde{t}_{ij}^* \mid t_{j\backslash i}^*) d\tilde{t}_{ij}^* \text{ 或者} \int_{t_{ij}^*} f(\tilde{t}_{ij}^* \mid t_{j\backslash i}^*, u_{j\backslash i}) d\tilde{t}_{ij}^* \tag{19}$$

其中，t_{ij}^* 为观测到的考生 j 在第 i 个题项上的作答时间的对数的向量，$\tilde{t}_{ij}^*$ 则是代表的预测到的值。$t_{j\backslash i}^*$ 则是观测到的考生 j 在除了第 i 个题项之外的题项上的作答时间的对数的向量，$u_{j\backslash i}$ 指的是考试 j 在除了第 i 个题项外其他所有题项上的作答反应。

这种方法不仅可以进行作弊考生检测，还可以进行泄露题项检测（Wang et al.，2018）。在检测作弊考生时，研究者通过计算 π_{ij}^l 和 π_{ij}^u 确定 p 值，将 p 值与名义的 α 值比较，如果 p 值过小，则代表该考生在该题项上的作答情况存在问题。在检测泄露题项时，研究者将所有考生在题项 i 上的作答时间进聚合，得出题项 i 的指标 $\hat{\pi}_i$，同样为 $\hat{\pi}_i$ 设置门槛，如果超过门槛值，则代表了该题项已经被泄露。

当然，也有研究者提出了其他的个人拟合分析方法，如在三参数的项目反应模型和项目作答反应时间的对数正态模型的基础上提出混合模型法（mixed hierarchical model，MHM），构建异常作答指标 Δ_{ij}（第 i 个题项，第 j 个考生的异常作答情况），并对其进行建模运算（Wang et al.，2018），这种拟合方法层次框架模型的一种延伸和修正。

当数据污染情况较重时，残差分析法对题目参数估计准确性会大幅降低，刘玥等（2022）利用混合模型法（MHM）对原有残差计算方式进行优化，提出了固定参数标准化残差法（conditional estimate standard residual，CSR），该方法先通过混合模型法（MHM）筛选正常作答的考生，进而获得较为准确参数估计结果，研究显示在数据污染较为严重时，该方法效果优于其他方法。

也有研究者在个人拟合统计量的基础上尝试使用变点分析法（change point analysis，CPA）进行异常作答甄别。张龙飞等（2020）对这种方法进行了系统介绍，变点分析的原理在于使用个人拟合分析统计量（person－fit statistics，PFS）判断作答序列中是

否存在可将该序列划分为具有不同统计学属性的两个部分的点，常用的统计检验方法有基于似然比检验的 L_{max} 法，基于 Wald 检验的 W_{max} 法，基于得分检验的 S_{max} 法和基于加权残差的 R_{max} 法。钟小缘等(2022)将变点分析拓展性的应用到了作答时间数据的分析中，发现此方法在加速作答检测中效果较好，I 型错误水平较低。

目前，有大量的研究者使用参数法进行建模和个人拟合，在检出率和误检率方面都取得了不错效果(Qian et al.,2016;Zopluoglu,Kasli,& Toton,2021)。但是，也有研究者对参数建模方法提出了批评(Man et al.,2018;Meijer & Sijtsma,2001)，他们认为参数法存在着比较多的问题：一是算法过于复杂，如使用 MCMC 进行参数估计的时会因为迭代数、起始点、马氏链长度等参数的设置造成不同；二是如果异常作答的数据较多，用单一模型进行数据拟合的难度将会变得非常大。

3.1.5　参数法在作弊甄别中的应用

参数法在近些年间获得了广泛应用，从本研究界定的“三类”作弊行为而言：这种作弊甄别方法在泄题和抄袭类的作弊行为的甄别中应用更多且更有效，但是在篡改答案型的作弊甄别中应用较少。这主要是因前两类作弊行为会更加系统化、规模化的影响作答反应时间(Sinharay,2021)。从本文界定的“两种”作弊行为而言：这种作弊甄别方法在团体作弊行为和个人作弊行为的甄别中都取得了不错的效果，并且在团体作弊行为甄别中的效果更优。

在抄袭类作弊行为甄别中的应用。van der Linden 基于作答反应时间开展二元对数正态模型(bivariate lognormal)建模，进行抄袭类作弊行为的甄别研究(van der Linden,2009)。二元对数正态模型主要是在对数正态模型的基础上引入了 2 个考生在个别考试题项上作答反应的一致性参数 ρ_{jk}，假定 $\rho_{jk}=0$(即 2 个考生的作答没有一致性)。但在观测中，如果发现 $\rho_{jk}>0$，并且大于门槛值 C，则这 2 个考生之间可能存在着答案抄袭或者共谋作弊的情况。在实际的估算和检验中，van der Linden(2009)采用了 LM 检验的方法(Lagrange multiplier test，拉格朗日乘数检验)进行抄袭甄别确认。研究结果显示，基于二元对数正态模型的抄袭甄别检验在实际数据中获得了不错的效果，尤其 LM 检验比普通的相关性检验能更好的反应 2 个考生之间的相似度。

在泄题类作弊行为甄别中的应用。考试泄题、考生提前获取试题已经成为当前考试实践中面临的一个重要问题(Zopluoglu et al.,2021)。Eckerly(2017)对泄题类作弊甄别做了分类：一是针对泄题类作弊考生开展的甄别研究；二是针对泄题题项的甄别研究；三是同时针对作弊考生和泄题题项的甄别研究；四是针对团体作弊考生甄别研究。Eckerly(2017)指出，基于反应时间的泄题类作弊行为研究主要应用在第三类。前文介绍的 2 种参数建模方法在泄题类作弊甄别中均有应用：单纯基于项目反应时间数据库开展的泄题类作弊甄别研究有很多(Qian et al.,2016;Van der Linden & Van Krimpen - Stoop,2003;van der Linden & Guo,2008)，基于项目反应时间数据和其他数据联合建模进行作弊甄别的研究也有不少(Meijer & Sotaridona,2006;Wang et al.,2018;Zopluoglu,2019)。如，在一项基于真实数据的泄题甄别研究中(Qian et al.,2016)，研究者从项目作答反应时间角度对两个行业从业资格考试进行泄题甄别。该研究选定分属于金融、卫健行业的 2 个行业准入考试，基于层次框架模型(hierarchical framework)(van der Linden & Guo,2008)进行作弊甄别。结果显示，这种方法在检出率和误报率等指标中都获得了比较好的结果，检测出了 111 个题项中 2 个有可能被泄露的题项和 1 172 个考生中有 2 个有可能掌握泄题资料的考生。

在答案篡改类作弊行为甄别中的应用。目前尚未有研究者采用项目作答反应时间的方法针对考试篡改答案情况进行研究。这主要是因为考试篡改答案行为，如更改考生答案等，污染了原有的项目作答反应数据，不易识别。

在团体类作弊行为甄别中的应用。根据本文的分类，团体类型的作弊也会出现如答案抄袭和共谋作弊(answer copying and collusion)、泄题(preknowledge)、考试篡改答案(test tampering)等类型。有研究者提出了一种多维数据联合建模的方式开展团体作弊行为中的泄题行为甄别，并取得了较好的作弊甄别效果(Man & Harring,2021)。也有研究者更新了对数正态模型(Cengiz Zopluoglu et al.,2021)，采用了增加门控制机制(Gating Mechanism)对数正态模型的方法对团体泄题类型的作弊进行了甄别研究。该结果虽然显示甄别效果较好，但是研究者也坦陈了效应量的问题。因为不同

数据集的效应量可能来自于多种因素，如考生特点、泄题的具体情况、和试题的特点等，这种研究方法不具有普适性，研究者需要针对数据特点选择建模方法和拟合分析方法。

3.2 非参数建模法及应用

非参数建模法则主要采用 K－L 散度（相对熵）的方法进行建模，当然，研究者也会在建模中引入其他作答数据，如作答准确率、考生的其他生物信息等。非参数法与参数法的本质不同在于，非参数法把考生的作答时间看成了一种离散变量，而参数法则把考生的作答时间看作了一种连续变量，且其对数服从正态分布。非参数的这种检测方法在模拟数据和真实数据中都取得了较好的效果（Man，Harring，Ouyang，& Thomas，2018），尤其是在同样的数据分析中，取得了比标准作答索引法（Marianti et al.，2014）相对更优的效果。Man 等研究者（2018）针对参数建模的缺点，采用 K－L 散度（Kullback－Leibler Divergence，也称相对熵）进行了基于项目作答反应时间的作弊甄别研究。K－L 散度可以度量两种分布之间的差异。在研究中，研究者使用相对熵的方法对考生个人的作答反应时间分布情况和考生总体的作答反应时间分布情况进行了对比，公式如下：

$$D_{KL}(f \parallel g) = \sum_i f(i) \ln\left(\frac{f(i)}{g(i)}\right), \qquad (20)$$

其中，$f(i)$ 代表了考生总体作答反应时间的概率质量函数，而 $g(i)$ 则代表了个体考生作答反应时间的概率质量函数。当考生个体的作答反应时间特点与考生总体的作答反应时间特点趋于一致的时候，$\frac{f(i)}{g(i)}$ 趋近于 1，其对数趋近于 0。反之，如果其对数的绝对值与 0 相差巨大，则说明该考生的作答时间存在问题，需要进一步查验。

有研究者在非参数建模中引入了作答准确率的数据，对项目作答反应时间进行了细分，提出了“有效反应时间”（effective response time）的概念，用以描述个体考生答对某一题项所花费的时间（Meijer & Sotaridona，2006）。研究者假定，获得泄题（pre-knowledge）数据的考生作答时间与普通考生有较大差异，并采用堪萨斯大学 528 位大学一年级学生参与的摘要推理考试（abstract reasoning test，ART）数据进行假设检验。研究发现，基于“有效反应时间”模型的误检率（type Ⅰ error）较低。随后，他们又在真实数据的基础上生成了模拟数据，在真实考试数据中抽样选取部分考生，然后将其在原始考试题项中比例为 50% 或者 75% 的题项上的作答反应时间改为原始数据的 1/2 或者 1/4，发现检出率较以往的方法有了大幅提升。但是也应该注意到，这项研究中的模拟数据情况较为极端（Meijer & Sotaridona，2006）。

4 总结与讨论

4.1 项目作答反应时间数据不能被污染

有关项目作答反应时间的考试作弊甄别研究成立的一个基本假设为：考生没有刻意操纵自己的作答时间。但是，如果作弊考生了解到，现有作弊甄别技术是通过监控其作答时间进行作弊甄别的，考生有可能会刻意控制自己的项目作答时间，这将为甄别效果带来巨大的挑战。不过，也应该看到，考生在刻意控制作答时间的时候，眼动等其他生物信息、敲击键盘的信息等会与积极作答的考生有所不同（Nagy & Ulitzsch，2021；Zopluoglu，2019），所以将考生作答反应时间与其他考生信息进行联合建模将有力的促进研究者优化建模方法、提升模型拟合度、提高作弊的甄别效率。

另外，还需要注意到，也有研究者（郭小军，罗照盛，2019；Domingue et al.，2022）对速度与准确率之间的权衡进行了分析与讨论，他们认为作答反应时间与作答准确性之间可能不是线性关系，随着反应时间的增加，准确率提高到某种程度之后会停滞或者降低。这些研究对有关项目作答反应时间数据的假设也提出了一定挑战。

4.2 多类型过程数据联合建模成为趋势

随着基于计算机的考试（Computer Based Tests）进一步普及，随着各类信息追踪设备和软件的轻量化、普及化发展，考生在考试中的各类信息可以被实时搜集（Man & Harring，2021），这些信息既包含机械信息：如作答反应时间、如击键记录（keystroke logging）；也包含生物信息如眼动追踪、注视点个数、眨眼频率（blinking rates）、瞳孔直径（pupil diameters）、血氧度（blood oxygen level）等（Liang et al.，2023；Man，Harring，& Zhan，2022）。这些信息可以和项目作答反应时间、考试作答数据等其他数据进行联合建模，进而量化考生的总体情况，从更全面的角度、更为精确得开展作弊甄别。

在这种多类型数据的联合应用过程中，不仅要

扩大数据量、拓展数据种类，也要进一步提升数据模型的拟合优度。有研究者尝试将机器学习、深度学习的研究方法应用到了作弊甄别中，取得了不错的效果。在机器学习方面，Man 等(2019)研究者对多种机器学习方法进行了对比，包括无监督学习方法 K 均值算法(K - means)，有监督学习方法支持向量机(SVM)、K 近邻(K - Nearest Neighbor)、随机森林(Random Forests)等，建议在作弊甄别实践中将多种机器学习方法合并使用，可以获得较好的检测效果。Pan 等采用机器学习方法进行了泄题题项和获取泄题信息的考生的作弊甄别(Pan & Wollack，2021，2023)。也有研究者采用机器学习中集成学习的方法开展作弊甄别，Zhou 和 Jiao(2022)使用集成学习的 stacking 算法开展了大规模考试的作弊甄别，Zopluoglu(2019)采用了集成学习中的 boosting 的方法进行了作弊甄别，两个模型越都取得了较好的作弊甄别效果。Meng 和 Ma(2023)也选定了 11 种特征，并使用随机森林(Random Forests)、逻辑回归(Logistic Regression)、支持向量机(SVM)等方法训练模型，发现支持向量机(SVM)和随机森林(Random Forests)在作弊甄别中的效果更好。在深度学习方面，Kamalov 等研究者采用循环神经网络的方法(RNN)进行了作弊甄别(Kamalov，Sulieman，& Santandreu Calonge，2021)。Zhen 和 Zhu(2023)采用了深层神经网络 TabNet 进行了作弊甄别，研究发现相较于其他机器学习模型而言，TabNet 具有较强的优势效果。

同时，如基于卷积神经网络(CNN)和 YOLO 算法的机器视觉研究也开始逐渐被使用到考场监控视频的识别研究中(窦刚，刘荣华，范诚，2021)，研究者也可以尝试将考场抓取的机器视觉信息与作答反应模型、项目作答反应时间模型等数据联合使用，提升作弊甄别的效率。

4.3　模型应用范围仍需进一步拓展

从研究数据角度看，大部分研究者都拿不到真实数据，或同一数据集被反复使用(Zopluoglu et al.，2021)，模拟数据研究的支撑作用较大(van der Linden & Guo，2008)，采样软件有 JAGS 或 R 语言的 LNIRT 等。换言之，虽然科研领域对这类作弊甄别模型研究较多，但是在实际考试中的模型的应用仍然较少。随着国内基于计算机的考试(Computer Based Tests)大规模普及，考试组织方和考试研究单位可以积极开展联合协作，推动作答反应时间类数据的采集和作答反应时间类作弊甄别模型的广泛应用，提升考试的安全性。

参考文献

窦刚，刘荣华，范诚.(2021).基于卷积神经网络的考场不当行为识别.*中国考试*，(2)，56 - 62 + 94. doi：10. 19360/j. cnki. 11 - 3303/g4. 2021. 02. 006

郭小军，罗照盛.(2019).速度与准确率权衡：被试反应状态评价与建模.*心理与行为研究*，(5)，589 - 595.

郭小军，罗照盛，严娟.(2022).项目间多维测验作答时间数据分析：潜在特质速度间效应建模.*心理科学*，(5)，1222 - 1229. doi：10. 16719/j. cnki. 1671 - 6981. 20220525.

胡佳琪，黄美薇，骆方.(2020).考试作弊甄别技术的研究进展：个体作弊的甄别.*中国考试*，(11)，32 - 36. doi：10. 19360/j. cnki. 11 - 3303/g4. 2020. 11. 006

刘玥，刘红云，游晓锋，杨建芹.(2022).用于处理不努力作答的标准化残差系列方法和混合多层模型法的比较.*心理学报*，(4)，411 - 425.

骆方，王欣夷，徐永泽，封慰.(2020).考试作弊甄别技术的研究进展：团体作弊的甄别.*中国考试*，(11)，37 - 41. doi：10. 19360/j. cnki. 11 - 3303/g4. 2020. 11. 007

孟祥斌.(2016).项目反应时间的对数偏正态模型.*心理科学*，(3)，727 - 734. doi：10. 16719/j. cnki. 1671 - 6981. 20160332.

彭恒利.(2015).惩治考试作弊的困境与出路之我见.*中国考试*，(2)，13 - 19. doi：10. 19360/j. cnki. 11 - 3303/g4. 2015. 02. 003

汪大勋，郭莹莹.(2022).融合反应时的多级评分 IRT 模型开发及其应用研究.*心理学探新*，*22*(3)，269 - 278 + 288.

严娟，郭小军，罗照盛.(2022).多维人格测验的反应与反应时联合建模分析.*江西师范大学学报*(自然科学版)，(5)，453 - 459. doi：10. 16357/j. cnki. issn1000 - 5862. 2022. 05. 03.

詹沛达.(2019).计算机化多维测验中作答时间和作答精度数据的联合分析.*心理科学*，(1)，170 - 178. doi：10. 16719/j. cnki. 1671 - 6981. 20190126.

詹沛达，Hong Jiao，Kaiwen Man.(2020).多维对数正态作答时间模型：对潜在加工速度多维性的探究.*心理学报*，(9)，1132 - 1142.

张龙飞，王晓雯，蔡艳，涂冬波.(2020).心理与教育测验中异常反应侦查新技术：变点分析法.*心理科学进展*，(9)，1462 - 1477.

钟小缘，喻晓锋，苗莹，秦春影，彭亚风，童昊.(2022).基于作答时间数据的改变点分析在检测加速作答中的探索——已知和未知项目参数.*心理学报*，(10)，1277 -

1292.

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 62 – 79. doi: 10.1111/bmsp.12059

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257 – 279. doi: 10.1111/bmsp.12076

Chen, Y., Lu, Y., & Moustaki, I. (2019). Statistical Analysis of Item Preknowledge in Educational Tests: Latent Variable Modelling and Statistical Decision Theory. *arXiv e – prints*. arXiv: 1911.09408.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. New York, NY: Routledge.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. New York: Routledge.

Crittenden, V. L., Hanna, R. C., & Peterson, R. A. (2009). The cheating culture: A global societal phenomenon. *Business Horizons*, *52*(4), 337 – 346. doi: 10.1016/j.bushor.2009.02.004

Domingue, B. W., Kanopka, K., Stenhaug, B., Sulik, M. J., Beverly, T., Brinkhuis, M., ... Yeatman, J. (2022). Speed – accuracy trade – off? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real – world settings. *Journal of Educational and Behavioral Statistics*, *47*(5), 576 – 602.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.00102

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67 – 86.

Eckerly, C. A. (2017). Detecting preknowledge and item compromise: Understanding the status quo. *Handbook of Quantitative Methods for Detecting Cheating on Tests*, 101 – 123.

Kamalov, F., Sulieman, H., & Santandreu Calonge, D. (2021). Machine learning based approach to exam cheating detection. *PLoS One*, *16*(8), e0254340. doi: 10.1371/journal.pone.0254340

Liang, K., Tu, D., & Cai, Y. (2023). Using Process Data to Improve Classification Accuracy of Cognitive Diagnosis Model. *Multivariate Behav Res*, 1 – 19. doi: 10.1080/00273171.2022.2157788

Man, K., & Harring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiple – group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, *81*(3), 441 – 465. doi: 10.1177/0013164420968630

Man, K., Harring, J. R., Ouyang, Y., & Thomas, S. L. (2018). Response time based nonparametric kullback – leibler divergence measure for detecting aberrant test – taking behavior. *International Journal of Testing*, *18*(2), 155 – 177. doi: 10.1080/15305058.2018.1429446

Man, K., Harring, J. R., & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, *56*(2), 251 – 279.

Man, K., Harring, J. R., & Zhan, P. (2022). Bridging models of biometric and psychometric assessment: A three – way joint modeling approach of item responses, response times, and gaze fixation counts. *Applied Psychological Measurement*, *46*(5), 361 – 381. doi: 10.1177/01466216221089344

Marianti, S., Fox, J. – P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426 – 451. doi: 10.3102/1076998614559412

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107 – 135. doi: 10.1177/01466210122031957

Meijer, R. R., & Sotaridona, L. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (LSAC research report series; No. CT 03 – 03). Law School Admission Council.

Meng, H., & Ma, Y. (2023). Machine learning – based profiling in test cheating detection. *Educational Measurement: Issues and Practice*, *42*(1), 59 – 75. doi: 10.1111/emip.12541

Molenaar, D., & Bolsinova, M. (2017). *A heteroscedastic generalized linear model with a non – normal speed factor for responses and response times*. Paper presented at the British Journal of Mathematical and Statistical Psychology.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi – parametric within – subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 205 – 228. doi: 10.1111/bmsp.12117

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279 – 297. doi: 10.1007/s11336 – 017 – 9602 – 9

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015a). A bivariate generalized linear item response theory

modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56 – 74. doi: 10.1080/00273171.2014.962684

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015b). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197 – 219. doi: 10.1111/bmsp.12042

Nagy, G., & Ulitzsch, E. (2021). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement*, 00131644211045351. doi: 10.1177/00131644211045351

Pan, Y., & Wollack, J. A. (2021). An unsupervised – learning – based approach to compromised items detection. *Journal of Educational Measurement*, *58*(3), 413 – 433. doi: 10.1111/jedm.12299

Pan, Y., & Wollack, J. A. (2023). A machine learning approach for the simultaneous detection of preknowledge in examinees and items when both are unknown. *Educational Measurement: Issues and Practice*, *42*(1), 76 – 98. doi: 10.1111/emip.12543

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. doi: 10.1111/emip.12102

Qiao, X., & Jiao, H. (2022). Explanatory cognitive diagnostic modeling incorporating response times. *Journal of Educational Measurement*, *58*(4), 564 – 585. doi: 10.1111/jedm.12306

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, *78*(3), 538 – 544. doi: 10.1007/s11336 – 013 – 9324 – 6

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. *Progress in Mathematical Psychology*, 151 – 171.

Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*(1), 18 – 38. doi: 10.1016/0022 – 2496(79)90003 – 8

Scheiblechner, H. (1985). Psychometric models for speed – test construction: The linear exponential model. *Test Design, Developments in Psychology and Psychometrics*, 219 – 244.

Sinharay, S. (2017). Detecting fraudulent erasures at an aggregate level. *Journal of Educational and Behavioral Statistics*, *43*(3), 286 – 315. doi: 10.3102/1076998617739626

Sinharay, S. (2020). Detection of item preknowledge using response times. *Appl Psychol Meas*, *44*(5), 376 – 392. doi: 10.1177/0146621620909893

Sinharay, S. (2021). Latent – variable approaches utilizing both item scores and response times to detect test fraud. *Open Education Studies*, *3*(1), 1 – 16. doi: 10.1515/edu – 2020 – 0137

Thissen, D. (1983). Timed testing: An approach using item response theory. *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, 179 – 203.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181 – 204. doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287 – 308. doi: 10.1007/s11336 – 006 – 1478 – z

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5 – 20. doi: 10.3102/1076998607302626

van der Linden, W. J. (2009). A bivariate lognormal response – time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*(3), 378 – 394. doi: 10.3102/1076998609332107

van der Linden, W. J. (2009). Conceptual issues in response – time modeling. *Journal of Educational Measurement*, *46*(3), 247 – 272. doi: 10.1111/j.1745 – 3984.2009.00080.x

van der Linden, W. J. (2009). Conceptual issues in response – time modeling. *Journal of Educational Measurement*, *46*(3), 247 – 272.

van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, *80*(3), 689 – 706. doi: 10.1007/s11336 – 014 – 9409 – x

van der Linden, W. J., & Van Krimpen – Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251 – 265. doi: 10.1007/BF02294800

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response – time patterns in adaptive testing. *Psychometrika*, *73*(3), 365 – 384. doi: 10.1007/s11336 – 007 – 9046 – 8

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time – limit tests. *Handbook of Modern Item Response Theory*, 169 – 185.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469 – 501. doi:

10.3102/1076998618767123

Wollack, J. A., & Fremer, J. J. (Eds.). (2013). *Handbook of test security*(1st Ed.). Routledge. New York, NY: Routledge.

Zhan, P. (2022). Joint - cross - loading multimodal cognitive diagnostic modeling incorporating visual fixation counts. *Acta Psychologica Sinica*, *54*(11), 1416 - 1432. doi:10.3724/SP.J.1041.2022.01416

Zhan, P., Jiao, H., Man, K., Wang, W. C., & He, K. (2021). Variable speed across dimensions of ability in the joint model for responses and response times. *Front Psychol*, *12*, 469196. doi:10.3389/fpsyg.2021.469196

Zhan, P., Jiao, H., Wang, W. - C., & Man, K. (2018). A multidimensional hierarchical framework for modeling speed and ability in computer - based multidimensional tests. *arXiv preprint. arXiv*:1807.04003.

Zhan, P., Man, K., Wind, S. A., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts: Providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, *47*(6), 736 - 776. doi:10.3102/10769986221111085

Zhen, Y., & Zhu, X. (2023). An ensemble learning approach based on tabNet and machine learning models for cheating detection in educational tests. *Educational and Psychological Measurement*. doi:10.1177/00131644231191298

Zhou, T., & Jiao, H. (2023). Exploration of the Stacking Ensemble Machine Learning Algorithm for Cheating Detection in Large - Scale Assessment. *Educational and Psychological Measurement*, *83*(4), 831 - 854. doi: 10.1177/00131644221117193

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large - scale testing using extreme gradient boosting (XGBoost). *Educ Psychol Meas*, *79*(5), 931 - 961. doi:10.1177/0013164419839439

Zopluoglu, C., Kasli, M., & Toton, S. L. (2021). The effect of item preknowledge on response time: Analysis of two datasets using the multiple - group lognormal response time model with a gating mechanism. *Educational Measurement: Issues and Practice*, *40*(3), 42 - 51. doi:10.1111/emip.12428

Advances in Cheating Detection Research Based on Item Response Time: A Review

Yang Zhiming Xu Qingshu

(Foreign Studies College, Hunan Normal University, Changsha 410081)

Abstract: As Computer Based Tests (CBT) become more and more popular, it is possible to collect, record, and analyze test takers' item response time. More and more researchers have started to conduct cheating screening studies based on this data. In this paper, we sort out the cheating detection researches based on item response time from two dimensions: parametric modelling method and non - parametric modelling method. At the same time, this paper proposes "two kinds and three types" of cheating behaviour classification criteria, and introduces the application practices and detection results of parametric and non - parametric modelling methods in the detection of various types of cheating behaviours, which provides a reference for related researchers. Furthermore, this paper also points out the direction of future studies in this field.

Key words: item response time; cheating detection; person fit analysis