基于认知诊断理论的网络安全自适应测试技术

齐 斌 王 宇 邹红霞 李冀兴
(航天工程大学航天信息学院 北京 101416)

摘 要 为进一步研究人员的网络安全素养,准确诊断人员网络安全知识和技能的水平,结合心理测量学和计算机测试技术,开发了基于认知诊断的多级属性评分的自适应测试技术。首先,为更好适应多元化复杂的网络安全知识结构且便于测试模型,在网络安全领域设计了复杂的层级网络安全知识库模型;然后,在多级评分认知诊断模型的基础上引入了属性层级的概念进行综合改进,并提出了准确、高效的参数估计方法和同模型相适应的选题策略。实验结果表明,多级属性评分的网络安全自适应测试技术较传统的多级评分模型提高了10.5%的效率,为计算机自适应测试领域的研究提供了参考。

关键词 自适应测试,认知诊断,网络安全,PH-DINA,素养测评
中图法分类号 TP309.2,TN915.08 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.016

Adaptive Testing Technology Based on Cognitive Diagnostic in Cybersecurity

QI Bin WANG Yu ZOU Hong-xia LI Ji-xing
(Department of Information, Space Engineering University, Beijing 101416, China)

Abstract To further effectively study the cybersecurity literacy of personnel and accurately diagnose the specific level of personnel knowledge and skills, the paper developed adaptive cybersecurity testing technology based on multi-level attributes scoring of cognitive diagnosis by combining psychometrics and computer testing technology. Firstly, a hierarchical cybersecurity knowledge model is designed for better adapting to the complex knowledge structure and verifying the research. Then, the hierarchy of attribute is input based on polytomous scoring cognitive diagnosis model to implement comprehensive improvements. Accurate and efficient parameter estimation method and suitable selection strategy are proposed to improve performance. The experimental results show that the adaptive cybersecurity testing technology of multi-level attributes scoring improves the efficiency by 10.5% compared with the traditional multi-level scoring model, which provides a reference to the research of computerized adaptive testing.

Keywords Adaptive testing, Cognitive diagnostic, Cybersecurity, PH-DINA, Evaluation of literacy

1 引言

网络安全是当今焦点领域,层出不穷的数据泄露、计算机病毒、恶意软件感染等事件暴露出数据安全防护工作中的重大弊端,即人员安全[1]。现有的大多数APT攻击事件往往是因为人为失误或人为诱导等进行自我破坏,究其本质则是相关工作岗位的人员严重缺乏网络安全意识或是人们掌握的安全知识和技能同工作岗位并不匹配[2]。

有效、可控地量化人员安全风险是网络安全管控中的必要一环。通过对人员安全素养的评测,准确而有效地度量个人网络安全知识、技能水平,将有助于及时预警并减少网络威胁,从而降低或避免更大的损失[3]。根据测评情况,还可针对性地编制教育方案,从而高效地提高个人网络安全素养。

网络安全素养测评实际上是对人员网络安全知识和技能的考核与评价[4-5],目前采用较多的是基于认知诊断的计算机自适应测试技术(Cognitive Diagnosis Computerized Adaptive Testing,CDCAT)[6],该技术可根据被试者的作答反应,自适应地调整和选择主、客观测试项目(包括技能测试题型),提高了测试效率并提供了准确的诊断信息,便于被试者进一步提高知识和技能短板。

认知诊断模型是CDCAT的核心,DINA(Deterministic Inputs, Noisy "and" gate model)模型因为具有易解释性和计算便捷等优势[7],在工程实践中得到了广泛的应用,国内外学者也针对不同的应用方向进行了相应的改进,产生了多级评分DINA、高阶DINA模型等[8]。但现有的认知诊断模型在处理高维知识数据时仍面临很大的挑战,尤其是处理类似于

到稿日期:2018-06-03 返修日期:2018-08-27 本文受国家863计划项目(2015AAxxx2078),省部级科技创新工程(ZYX14030011)资助。
齐 斌(1994-),男,硕士,主要研究方向为网络空间安全;王 宇(1971-),男,博士,教授,CCF会员,主要研究方向为保密技术,E-mail:1364742701@qq.com(通信作者);邹红霞(1968-),女,硕士,副教授,主要研究方向为计算机应用技术;李冀兴(1993-),男,硕士,主要研究方向为网络空间安全。

网络安全知识体系这类层次复杂的知识属性结构时,其在处理精度和测试曝光率等指标的量化中效果并不理想[8-9]。

为提高和改善自适应测试在多维知识数据处理中的准确性和安全性,本文建立了符合实际的复杂层次的知识库模型,改进了基于多级属性评分的认知诊断模型,并设计了同模型相适应的选题策略。其在网络安全领域得到了很好的验证,为自适应测试和人员安全风险评估等工作提供了参考。

2　知识库模型

ACM SIGCSE2018 国际会议正式发布了网络空间安全学科知识体系(CSEC2017),其将人员安全的重要性提升至最高层次,指明在软件安全、数据安全、组件安全等基础领域之上考虑人员安全,这既符合现实意义,也体现了人员网络安全重要性的理论支撑[10]。人员安全知识主要包括个人数据保护、个人隐私保护和安全威胁的化解,也涉及用户的行为、知识和隐私对网络空间安全的影响。

人员安全知识库不仅需要独有的社会工程学等安全知识,还应该包括数据安全、系统安全等。但在客观条件下,人员处在不同行业领域,不同岗位对个人安全知识的掌握(内容、层次和程度等)要求不同。因此,本文综合分析国内外现有的网络安全知识体系的分类标准[11],结合具体行业领域规范,构建了基于“行业、岗位、人员”的知识库模型,如图 1 所示。

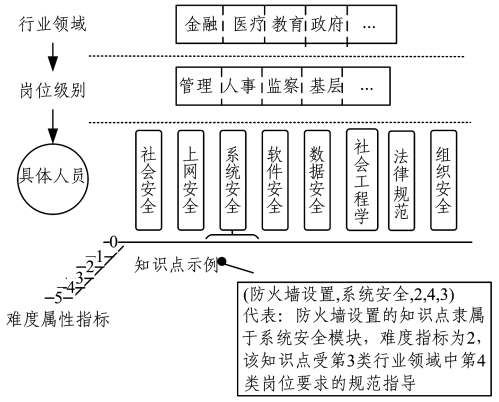

图 1　人员安全知识库模型

Fig. 1　Personnel safety knowledge base model

该模型的核心是具体人员的知识分类和属性指标,参考权威的网络安全知识体系和相关标准,人员需要掌握的安全知识内容可大致分为社会安全、上网安全、系统安全、软件安全、数据安全、社会工程学、法律规范、组织安全等 8 类顶层知识域[10],在不同的知识领域下仍然可根据内容再次划分具体的知识点,并根据行业领域的客观要求赋予不同的权重。但因人员岗位的要求不同,即使是对于同一知识点,也仍可根据诸如难度等指标继续分类,因此有必要将每一知识点都额外设置具体的属性指标,以便于知识的抽取和试题的分类补充。

测试题库是指根据测试需求以知识库模型为模板先建立个人的网络安全知识图谱,即基于具体人员角色建立网络安全知识库,并根据相应知识点补充所属的不同类型的测试题。但客观上存在同一道测试题可考核多个知识点的试题类型,即每一道试题至少包含一个知识属性,则试题 Q 可表示为 $Q_j=(q_{j1},q_{j2},\cdots,q_{jk})$。其中,$j$ 为测试题目的编号,k 为题目 j 待考核知识点的最大数目,$q_{jk}=\{0,1,2,\cdots,n\}$ 表示第 k 个测试属性具有 n 个级别。若 $q_{jk}=n\geqslant 1$ 则代表考查难度为 n 的第 k 个知识属性;反之,$q_{jk}=0$ 则代表不考查该知识属性。

3　基于 PH-DINA 的自适应测试设计

3.1　认知诊断模型

认知诊断模型是认知心理学与心理计量学的产物,它不仅可以宏观评价个体心理的特质水平,还可以诊断个体的认知加工特点,因而在教育测量学领域得到了广泛的拓展应用,其中因 DINA 模型只涉及失误和猜测两个参数,比其他模型更加简洁、灵活和易于解释,因此得到了广泛的理论研究。

DINA 模型[7]是典型的非补偿模型,要求必须掌握待测的全部技能或知识属性 α_i 才可被认定正确作答,项目所考查的技能或属性则全部被包含在待测项目 q_j 中。

$$\text{DINA}:P(Y_{ij}=1|\alpha_i)=(1-s)^{\eta_{ij}}\cdot g_j^{1-\eta_{ij}} \tag{1}$$

DINA 模型的反应函数表示被试 i 在掌握属性的前提下正确回答项目 j 的概率,$Y_{ij}=1$ 代表被试可以正确作答项目 j,$\eta_{ij}\in\{0,1\}$ 表示被试在理想情况下(不考虑猜测和失误的情况)作答的结果,其计算公式可表示为:

$$\eta_{ij}=\prod_{K=1}^{K}\alpha_{ik}^{q_{jk}} \tag{2}$$

其中,$\alpha_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{ik}\,|\,\alpha_{ik}=\{0,1\})$ 表示被试 i 对各个属性的掌握情况,$\alpha_{ik}=1$ 说明被试 i 掌握了 k 属性,为 0 则表示未掌握;$q_j=(q_{j1},q_{j2},\cdots,q_{jk}\,|\,q_{jk}=\{0,1\})$ 表示项目 j 对各个属性的考查情况,$q_{jk}=1$ 说明项目 j 考查了属性 k,为 0 则表示未考查;K 表示测试属性的数量。项目 j 的失误参数 $s_j=P(Y_{ij}=1|\eta_{ij}=1)$,指被试在掌握项目 j 考核的属性下仍答错的概率,$g_j=P(Y_{ij}=1|\eta_{ij}=0)$ 为项目 j 的猜测参数,指被试在尚未完全掌握项目 j 考核的属性下答对的概率。

为拓展认知诊断在实际中的应用与发展,涂冬波等[12-13]在传统 DINA 模型的基础上开发了多种评分模型 P-DINA(Polytomous-DINA),增加了测试项目的应用范围,模型反馈的信息也更加丰富,总体提高了测试效率。P-DINA 模型的项目反应函数如下:

$$P(Y_{ij}=t|\alpha_i)=P^*(Y_{ij}=t|\alpha_i)-P^*(Y_{ij}=t+1|\alpha_i) \tag{3}$$

$$P^*(Y_{ij}=t|\alpha_i)=(1-s_{jt})^{\eta_{ij}}\cdot g_{jt}^{1-\eta_{ij}} \tag{4}$$

其中,$P(Y_{ij}=t|\alpha_i)$ 表示被试 i 在项目 j 上得 t 分的概率,$P^*(Y_{ij}=t|\alpha_i)$ 表示被试 i 在项目 j 上得 t 分及以上的概率,P-DINA 模型中理想反应得分 η_{ij} 仍采用传统 DINA 模型进行计算,但式(4)的猜测参数和失误参数需满足 $s_{jt}\leqslant s_{jt+1}$,即对于需要掌握项目 j 考核属性的被试而言,其得 t 分的失误概率要小于 $t+1$ 分的失误概率;$g_{jt}\geqslant g_{jt+1}$,即对未全部掌握项目 j 考核属性的被试而言,猜对 t 分的概率要大于猜对 $t+1$ 分的概率,从而保证了被试答对的概率恒不为负。

根据项目反应理论的局部独立型假设,P-DINA 的似然函数为:

$$L(s,g;\alpha)=\prod_{i=1}^{N}\prod_{j=1}^{m}\prod_{t=0}^{mf_j}p_{ijt}^{u_{ijt}} \tag{5}$$

其中，p_{ijt} 指被试 i 在项目 j 上得 t 分的概率，$u_{ijt}=\{0,1\}$ 指被试 i 在项目 j 上得 t 分的事实判断，N 是参加测试的人员总数，m 是对被试 i 进行测试的试题总数，mf_j 指第 j 个题目的满分值。

3.2 PH-DINA 认知诊断模型的开发思路

对于 DINA 和 P-DINA 模型而言，项目属性 $\boldsymbol{Q}$ 矩阵由 0-1 元素构成，因而只能单纯进行布尔运算，测试反馈的信息较少，很多测试数据并不能够较好地拟合 DINA 及其改进模型（G-DINA 模型、P-DINA 模型以及 MS-DINA 模型等）[14]。因此，为了适应网络安全等领域的复杂知识结构和匹配的网络安全知识库模型，本文引入属性层级（Hierarchical）的概念，在 P-DINA 模型的基础上提出了基于属性多级评分的认知诊断模型 PH-DINA（Polytomous-Hierarchical-DINA）。属性多级的 $\boldsymbol{Q}$ 矩阵可以进行任意整数赋值，如 $q_{jk}=3$ 代表项目 j 考查指标为 3 的 k 属性，$\alpha_{ik}=2$ 则代表被试 i 掌握了 k 属性的第 2 层次。如果被试要正确作答项目则需要掌握考核属性指标水平及其以上的层次，如项目 j 测量的属性 $p_j=(1,3,2)$，属性 $A1,A2,A3$ 分别具有 2，4，3 种层次，若要答对项目 j，则要求被试的属性掌握模式为 $\alpha_i=\{(1,3,2)|(2,3,2)|(1,3,3)|(2,3,3)|(3,3,2)|(3,3,3)\}$。

对于属性多级模型，α_{ik} 和 q_{jk} 的取值共有 L_k 种，即属性 k 的层级有 $L\geqslant 2$ 种，因此如果属性 k 的数值为非 0-1 元素，则理想反应得分 η_{ij} 和项目反应函数不再适用，且增加了参数估计的难度和计算量。为了保持认知诊断模型的简洁性和易解释性，需要通过 Discriminant 函数将多级 α 和 q 转换为 0-1 元素。Discriminant 函数如下所示：

$$\alpha'_{ik}=\begin{cases}1, & \alpha_{ik}\geqslant q_{jk}\\ 0, & \alpha_{ik}<q_{jk}\end{cases} \tag{6}$$

$$q'_{jk}=\begin{cases}1, & q_{jk}\geqslant 1\\ 0, & q_{jk}=0\end{cases} \tag{7}$$

$$\eta'_{ij}=\prod_{K=1}^{K}(\alpha'_{ik})^{q'_{jk}} \tag{8}$$

模型此时虽然实现了属性多级化的计算处理，满足了多级属性的客观考查和量化要求，但理想得分式(8)仍为 0-1 评分，无法准确表达被试在项目 j 上获取合适的得分，因此利用 Weight 函数将 0-1 评分拓展为多级评分，进而使观察得分、理想得分的表现均为多级评分。将式(8)改进为基于权重的理想得分函数：

$$\eta^{*}_{ij}=round\sum_{k=1}^{K}\rho_{ijk}\cdot\omega_{jk}\cdot mf_j \tag{9}$$

$$\rho_{ijk}=\begin{cases}\dfrac{\alpha'_{ik}}{q'_{jk}}, & q'_{jk}=1\\ 0, & q'_{jk}=0\end{cases} \tag{10}$$

其中，ω_{jk} 是项目 j 考查属性中 k 属性所占的权重，$\dfrac{\alpha'_{ik}}{q'_{jk}}$ 为被试 i 在项目 j 上掌握属性的比例。为便于模型的理解和参数估计运算，本文对理想得分进行了取整操作，但实际上由于式(11)是一个门函数，使得 η^{*}_{ij} 的取整与否并不影响模型的表达。

为保持认知诊断模型的优势，并使其适用 P-DINA 项目反应函数式(3)，本文对式(4)进行了多级化拓展，则该转化公式为：

$$P^{*}(Y_{ij}=t|\alpha_i)=(1-s_{jt})^{\varphi_{ijt}}\cdot g_{jt}^{1-\varphi_{ijt}} \tag{11}$$

$$\varphi_{ijt}=\begin{cases}1, & \eta^{*}_{ij}\geqslant t\\ 0, & \text{otherwise}\end{cases} \tag{12}$$

式(3)、式(9)、式(11)即为 PH-DINA 模型的项目反应概率函数。与 P-DINA 模型相比，PH-DINA 模型不仅增加了被试和项目多级属性指标的运算，还拓展了理想反应得分 η_{ij} 的计算方法，增加了项目的实际考查范围和反馈的信息量，现举例说明和验证改进模型的优势。假设某项目 j 的测试属性为 $q_j=(1,1,1,1)$，满分 $mf_j=3$，则得分概率的对比如表 1 所列。

表 1 P-DINA 模型和 PH-DINA 模型的得分概率对比

Table 1 Comparison of probabilities between P-DINA model and PH-DINA model

t	P-DINA		PH-DINA			
	$\eta_{ij}=0$	$\eta_{ij}=1$	$\eta_{ij}=0$	$\eta_{ij}=1$	$\eta_{ij}=2$	$\eta_{ij}=3$
0	$1-g_{j1}$	s_{j1}	$1-g_{j1}$	s_{j1}	s_{j1}	s_{j1}
1	$g_{j1}-g_{j2}$	$s_{j2}-s_{j1}$	$g_{j1}-g_{j2}$	$1-s_{j1}-g_{j2}$	$s_{j2}-s_{j1}$	$s_{j2}-s_{j1}$
2	$g_{j2}-g_{j3}$	$s_{j3}-s_{j2}$	$g_{j2}-g_{j3}$	$g_{j2}-g_{j3}$	$1-s_{j2}-g_{j3}$	$s_{j3}-s_{j2}$
3	g_{j3}	$1-s_{j3}$	g_{j3}	g_{j3}	g_{j3}	$1-s_{j3}$

在同等条件下，P-DINA 模型只能将项目反应的概率分为 2 类，而改进后的 PH-DINA 模型则可将其分为 mf_j+1 类，测试反馈的信息也更丰富。假设猜测参数 g 及失误参数 s 的最小值为 0.1，且式(3)的模型参数要求满足 $s_{jt}\leqslant s_{jt+1}$，$g_{jt}\geqslant g_{jt+1}$，为便于直观表示，以概率差值 0.1 进行仿真验证，表 1 的模型得分概率可在多次参数赋值中呈现出相应的离散点图，具体如图 2 所示。可以明显看出，P-DINA 模型的得分概率两极分化严重，白色点主要集中在 0 分和满分 mf_j 两个极端，但其避免了 0-1 的极化缺陷，黑色点近乎平均分布于不同得分的附近，$\eta_{ij}=t$ 时被试得 t 分的概率倾向于最大，实现了多级观察得分与理想得分的对应关系[13]。

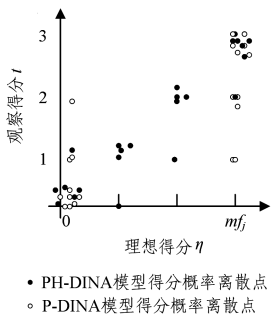

图 2 模型的得分概率离散点对比

Fig. 2 Comparison of probability discrete point between two models

经蒙特卡洛模拟参数估计的研究，PH-DINA 模型在项目参数和被试知识属性参数上的估计精度和判准率均比 P-DINA 模型高。当独立属性的个数为 6、题量为 120 时，P-DINA 的判准率为 81.3%，PH-DINA 的判准率为 96.0%。并且测试项目类型越复杂，PH-DINA 模型的估计精度和判准率表现得越好（限于篇幅，未列出全部实验数据）。

3.3 PH-DINA 模型的参数估计

计算机自适应测试的参数估计一般包括被试知识属性参数条件估计和项目参数条件估计。被试参数估计通常是指在项目参数已知的情况下,采用极大似然估计、期望后验估计等计算规模小、效率较高的算法估计被试的知识属性。因此本文结合 PH-DINA 模型参数改进了极大似然估计算法以便于适应计算。

假设 $L(Y_i|\alpha)$ 是被试 i 在多级属性评分下的似然函数,则有

$$L(Y_i|\alpha)=\prod_{j=1}^{J}\prod_{t=0}^{mf_j}P(Y_{ij}=t|\alpha)^{u_{ijt}} \tag{13}$$

$$u_{ijt}=\begin{cases}1, & Y_{ij}=t\\ 0, & \text{otherwise}\end{cases} \tag{14}$$

因此,PH-DINA 模型的似然函数为:

$$L(Y_i|\alpha)=\prod_{j=1}^{J}\prod_{t=0}^{mf_j}\left[(1-s_{jt})^{\varphi_{ijt}}\cdot g_{jt}^{1-\varphi_{ijt}}-(1-s_{jt+1})^{\varphi_{ijt+1}}\cdot g_{jt+1}^{1-\varphi_{ijt+1}}\right]^{u_{ijt}} \tag{15}$$

则被试 i 的知识属性的极大似然估计的计算式为:

$$\hat{\alpha}_i=\arg\max\{L(Y_i|\alpha_c)\} \tag{16}$$

对于项目参数估计中的 s 和 g 而言,精确度是首要考虑的要素,且项目参数的精确与否将会直接影响被试知识属性的判准率,因此选用 MCMC 算法对其进行估计。项目属性参数引入了层级属性结构,设置模型参数的先验概率分布为 $g_j\sim 4-\text{Beta}(0,0.6,1,2)$,$1-s_j\sim 4-\text{Beta}(0.4,1,2,1)$,$\alpha_{ik}\sim U(0,L_k)$。

根据 Bayes 定理,待估参数的近似满条件分布为:

$$P(s,g|Y,\alpha)\propto L(s,g|\alpha)P(s)P(g)$$

因此,$\{s_j^{n+1}\}$ 从均匀分布 $U(s_j^n-\delta_s,s_j^n+\delta_s)$ 中随机抽取,$\{g_j^{n+1}\}$ 从均匀分布 $U(g_j^n-\delta_g,g_j^n+\delta_g)$ 中随机抽取。现假定 $\delta_s=\delta_g=0.1$,则参数转移概率公式为:

$$p(\{s_j^n,g_j^n\},\{s_j^{n+1},g_j^{n+1}\})=\min\left\{\frac{L(s_j^{n+1},g_j^{n+1}|\alpha_{ik}^{n+1})P(s_j^{n+1})P(g_j^{n+1})}{L(s_j^n,g_j^n|\alpha_{ik}^{n+1})P(s_j^n)P(g_j^n)},1\right\} \tag{17}$$

由于仅需要估计 PH-DINA 模型的项目 j 参数,因此有效似然函数为:

$$L(s,g|\alpha)=\prod_{i=1}^{N}\left[\{(1-s_j)^{\eta'_{ij}}g_j^{1-\eta'_{ij}}\}^{Y_{ij}}\cdot\{1-(1-s_j)^{\eta'_{ij}}g_j^{1-\eta'_{ij}}\}^{1-Y_{ij}}\right] \tag{18}$$

3.4 选题策略

CD-CAT 可根据被试的作答情况抽取最适合的题目,但知识属性往往是非连续性的,同时考虑到 PH-DINA 模型的参数的多维性,HKL(H-KullbackLeibler)信息量[15]选题策略难以直接运用到基于 PH-DINA 的自适应测试中,因此本文在参数计算上有针对性地拓展了 HKL,记为 PH-HKL 选题策略,以实现在 PH-DINA 模型下计算选题项目的信息量。

PH-KHL 信息量不仅考虑了后验概率加权,而且进一步考虑了被试之间知识属性的相似性,其计算公式为:

$$\text{PH-HKL}_j(\hat{\alpha})=\sum_{c=1}^{2^k}\sum_{t=0}^{mf_j}\frac{1}{d(\alpha_c,\hat{\alpha})}\cdot\{[\log(\frac{P(Y_j=t|\hat{\alpha})}{P(Y_j=t|\alpha_c)})P(Y_j=t|\hat{\alpha})]\pi(\alpha_c|Y_j)\} \tag{19}$$

其中,$P(Y_j=t|\hat{\alpha})$,$P(Y_j=t|\alpha_c)$,指不同属性状态的被试在项目上得分的反应概率;$\pi(\alpha_c|Y_j)$ 是指知识属性为 $\alpha_c(c=1,2,\cdots,2^K)$ 的后验概率,记 $p(\alpha_c)$ 为知识状态 α_c 的先验概率,则后验概率的计算公式为:

$$\pi(\alpha_c|Y_j)=\frac{p(\alpha_c)L(Y_j|\alpha_c)}{\sum_{c=1}^{2^K}p(\alpha_c)L(Y_j|\alpha_c)} \tag{20}$$

$d(\alpha_c,\hat{\alpha})$ 指不同被试的知识状态间的相似性,具体公式为:

$$d(\alpha_c,\hat{\alpha})=\sqrt{\sum_{k=1}^{K}(\alpha_{ck},\hat{\alpha}_k)^2} \tag{21}$$

4 实验分析与论证

4.1 评价指标

评价指标一般分为属性判准率和非约束指标。判准率[16]通常采用平均属性边际判准率(Average Attribute Match Ratio,AAMR)、模式判准率(Pattern Match Ration,PMR)两类评价指标;非约束类指标主要包括题库的安全性和测试效率(Testing Efficiency,TE),其中安全性包括题目曝光率(Exposure Rate,ER)、测验重叠率(Testing Overlap Ration,TOR)。假设在 CD-CAT 中对 N 个被试的 K 个属性进行认知诊断测验,α_i 是第 i 个被试的真实知识状态,$\hat{\alpha}_i$ 是估计的被试属性掌握模式,具体计算公式及含义的阐述如下。

边际判准率(Marginal Match Rate,MMR)是单个属性在实验中的判准率,若被试 i 的第 k 个属性评估准确,则记为 $g_{ik}=1$,不准确则记为 $g_{ik}=0$,有:

$$MMR(k)=\sum_{i=1}^{N}g_{ik}/N \tag{22}$$

AAMR 为全部属性的平均判准率:

$$AAMR=\sum_{k=1}^{K}MMR(k)/K \tag{23}$$

PMR 中,若 $\alpha_i=\hat{\alpha}_i$,则 $h_i=1$,否则为 0,即判断被试作为个体的知识属性掌握情况是否满足要求,其计算表达式为:

$$PMR=\sum_{i=1}^{N}h_i/N \tag{24}$$

ER 反映了试题库的曝光程度,一般采用卡方指标:

$$\chi^2=\sum_{j=1}^{M}\frac{[ER_j-E(ER_j)]^2}{E(ER_j)} \tag{25}$$

其中,$ER_j=f_j/N$ 是第 j 题的曝光率,f_j 是第 j 题被抽取的次数,$E(ER_j)$ 是试题 j 期望的曝光率,ER_j 越小,则曝光率越低,安全性越高。测试中的理想情况是所有试题都可被均匀抽取,即 $E(ER_j)=L/M$,L 是平均测验的长度,M 是题库的试题总数。因此,χ^2 用来统计观察曝光率与期望曝光率间的距离,χ^2 的值越小,说明题库安全性越高。

TOR 用于反映不同被试抽取相同试题的重叠情况,因此计算式与题目曝光率、测验长度和被试人数有直接关系,重叠率越高,题目越不安全,其公式如下:

$$\overline{\hat{T}}=\frac{N\times\sum_{j=1}^{M}ER_j^2}{(N-1)\times L}-\frac{1}{N-1} \tag{26}$$

TE 是综合评定测试效能比的指标,指在相同测量精度

下平均耗用的试题数量，TE 值越低代表效率越高，其中 L_i 是指测试中被试 i 平均耗用的题目数量。

$$TE=\sum_{i=1}^{N}L_i/N \tag{27}$$

4.2 测试结果与分析

为了进一步研究和验证基于 PH-DINA 的自适应网络安全测试技术的可行性，本文采用蒙特卡洛法进行模拟实验。因网络安全知识库模型有 8 项主要知识内容，所以实验设置了 8 个相对独立的认知属性，每一个属性都包含低、中、高 3 类级别，即 $k=\{0,1,2,3\}$，因此理论上知识属性的全部掌握模式有 $4^8=65536$ 种。但因为网络安全知识结构中因为网络安全知识结构中指标层级的限制，被试个人的知识属性的掌握模式远达不到全集数量，所以本文在实际测试实验中选取某领域、某岗位、某类人员的知识属性结构，如表 2 所列，计有 $3\times2\times4\times3\times2=144$ 种。从这些属性掌握模式中随机抽取 $N=\{1000,2000\}$ 名被试的模式测试情况。项目参数从均匀分布中随机产生，$s_{jt}\sim U(0,0.3)$，$g_{jt}\sim U(0,0.3)$，严格控制 $s_{jt}\leqslant s_{jt+1}$，$g_{jt}\geqslant g_{jt+1}$，且所有试题均采用 $mf_j=3$ 的评分方式。

表 2 知识属性指标

Table 2 Knowledge property indicators

属性	$\alpha1$	$\alpha2$	$\alpha3$	$\alpha4$	$\alpha5$	$\alpha6$	$\alpha7$	$\alpha8$
层级	1	3	2	1	4	3	2	1
值域	0	0,1,2	0,1	0	0,1,2,3	0,1,2	0,1	0

为更契合实际测试环境和数据情况，本文基于非定长测试考虑了两种因素进行对比实验设计。第一类因素为被试数量，有 1000，2000 两种情况；第二类因素为测量精度指标，即后验概率 p，共设置了 0.75，0.80，0.85 这 3 个水平。实验中分别对比 P-DINA 模型和 PH-DINA 模型在不同指标下的测试结果。选题策略分别为 HKL 和 PH-HKL，测试结果如表 3 所列。

表 3 基于 P-DINA 和 PH-DINA 模型的测试结果指标对比

Table 3 Comparison of testing results based on P-DINA model and PH-DINA model

样本大小	测量精度（后验概率）	P-DINA					PH-DINA				
		AAMR	PMR	ER	TOR	TE	AAMR	PMR	ER	TOR	TE
1000	$p=0.75$	0.935	0.844	66.32	0.24	9.46	0.954	0.862	62.11	0.21	8.30
	$p=0.80$	0.950	0.890	67.12	0.25	9.52	0.965	0.914	64.02	0.22	8.48
	$p=0.85$	0.956	0.902	68.02	0.26	9.68	0.971	0.928	65.12	0.22	8.89
2000	$p=0.75$	0.935	0.853	66.32	0.24	9.46	0.954	0.888	62.11	0.21	8.30
	$p=0.80$	0.951	0.894	67.12	0.25	9.52	0.965	0.916	64.02	0.22	8.48
	$p=0.85$	0.956	0.903	68.02	0.26	9.68	0.971	0.928	65.12	0.23	8.89

上述结果表明，在固定精度的条件下，基于 PH-DINA 模型的自适应测试的属性边际判准率 $AAMR$ 均超过 95%，模式判准率 PMR 均高于 86%，因此可以证明 PH-DINA 模型具有较高的准确率。相比于传统 P-DINA 模型而言，PH-DINA 不仅在测试准确性上平均提高了 2%，在安全性和测试效率上也有一定的进步。PH-DINA 模型在曝光率和重叠率上相比于原模型分别降低了 6%和 3%，使得测试题库的安全性得到了更好的保障。PH-DINA 测试平均使用了 $(8.30+8.48+8.89)/3=8.56$ 道题目，相比原模型的 $(9.46+9.52+9.68)/3=9.55$ 道题目，提高了 10.5%的效能。两次不同样本容量的测试实验结果表明，基于多级属性评分的 PH-DINA 模型具有较好的稳定性和合理性，诊断正确率高且安全性也较理想，综合测试效率能够满足现有自适应测试的要求，同时也适用于网络安全及其他领域学科的自适应测试。

结束语 CD-CAT 在心理测量领域和计算机技术领域已经开展了较为广泛的研究，涉及认知诊断模型、参数估计、选题策略等多个研究方向。但国内外目前的研究成果仍难以满足实际测试环境，往往要求测试题目先满足模型要求，使得测试反馈的信息较少，题目数量较大，测试效率较低。在实际应用和测试环境中，多级属性、多级评分的数据大量存在，传统模型显然无法满足要求，也不利于自适应测试的推广，因此本文所述的基于多级属性评分的计算机自适应测试不仅有助于更多学者的进一步研究，也助力于网络安全人员测试的进一步发展。

限于时间和篇幅，上述研究存在很多有待进一步探讨及完善的方面。本研究主要拓展了非补偿型的认知诊断模型，但针对补偿型的多级属性、多级评分认知诊断模型仍需进一步研究。由于引入了多级属性，虽然参数估计的精确度和测试效率在一定程度上得到了提高，但却是以计算规模为代价，实际应用中的可靠性也有待进一步探讨；多级属性的属性标定、修正、检验甚至估计都是未来研究有待进一步解决的问题。

在传统 CD-CAT 研究的基础上，本文开发设计了可以处理多级评分、多级属性指标的 CD-CAT，创新点在于提出了较为科学的网络安全知识库模型，开发设计了多级属性评分的认知诊断模型，并验证处理了多级属性评分的自适应测试，弥补了现有的 CD-CAT 处理多级属性评分数据时的不足，PH-HKL 选题策略同样具有较理想的被试属性判准率、题库安全性和高测验效率，对于进一步拓展 CD-CAT 在实践中的应用提供了重要的理论设计，为人员网络安全素养的测评提供了有力的技术支撑，进一步为人员安全预警和定制化教育提供有效参考。

参 考 文 献

[1] CONTEH N Y, SCHMICK P J. Cybersecurity: risks, vulnerabilities and countermeasures to prevent social engineering attacks [J]. International Journal of Advanced Research in Computer Science, 2016, 6(23): 31-38.

[2] GRATIAN M,BANDI S,CUKIER M,et al. Correlating Human Traits and Cybersecurity Behavior Intentions[J]. Computers & Security,2018,73(3):345-358.

[3] BASSETT G. System and method for cyber security analysis and human behavior prediction:US 20160205122. A1[P]. 2016-3-22.

[4] YOUNG H,VLIET T V,VEN J V D,et al. Understanding Human Factors in Cyber Security as a Dynamic System[C]// AHFE 2017:8 th International Conference on Applied Human Factors and Ergonomics. Los Angeles,Springer,2017:244-254.

[5] ZHANG H L,YU H N,FANG B X,et al. Research on China's cyberspace security practice qualification system [J]. Chinese Engineering Science,2016,18(6):44-48. (in Chinese)
张宏莉,于海宁,方滨兴,等. 我国网络空间安全执业资格认证体系研究[J]. 中国工程科学,2016,18(6):44-48.

[6] SMITS N,PAAP M C S,BÖHNKE J R. Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes[J]. Quality of Life Research,2018,27(4):1055-1063.

[7] GU Y,XU G. The Sufficient and Necessary Condition for the Identifiability and Estimability of the DINA Model[J]. Psychometrika,2018(2):1-16.

[8] TORRE J D L,MINCHEN N. Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework [J]. Psicología Educativa,2014,20(2):89-97

[9] MCS P,KROEZE K A,TERWEE C B,et al. Item usage in a multidimensional computerized adaptive test (MCAT) measuring health-related quality of life[J]. Quality Life of Research,2017,26(11):2909-2918.

[10] RAJ R K,PARRISH A. Toward Standards in Undergraduate Cybersecurity Education in 2018[J]. Computer,2018,51(2):72-75.

[11] QI B,WANG Y,ZOU H X,et al. The Analysis of Measurement Method in the Knowledge System of Network Security Based on Information Entropy[C]// ICCT 2017:17th IEEE International Conference on Communication Technology. Sichuan, China:IEEE Press,2017:1328-1333.

[12] TU D B,CAI Y,DAI H Q,et al. A Polytomous Cognitive Diagnosis Model:P-DINA Model[J]. Acta PsychologicaSinica,2010,42(10):1011-1020. (in Chinese)
涂冬波,蔡艳,戴海琦,等. 一种多级评分的认知诊断模型:P-DINA 模型的开发[J]. 心理学报,2010,42(10):1011-1020.

[13] CAI Y,MIAO Y,TU D B. The polytomously scored cognitive diagnosis computerized adaptive testing[J]. Acta PsychologicaSinica,2016,48(10):1338-1346. (in Chinese)
蔡艳,苗莹,涂冬波. 多级评分的认知诊断计算机化适应测验[J]. 心理学报,2016,48(10):1338-1346.

[14] XU G. Identifiability of restricted latent class models with binary responses[J]. The Annals of Statistics, 2017, 45(2): 675-707.

[15] HSU C L,WANG W C,CHEN S Y. Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models. [J]. Applied Psychological Measurement,2013,37(7):563-582.

[16] KAPLAN M,TORRE J D L,BARRADA J R. New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing[J]. Applied Psychological Measurement,2015,39(3):167-188.