文章编号:1000-5862(2021)02-0111-07

认知诊断计算机化自适应测量技术在心理障碍诊断与评估中的应用

汪大勋,涂冬波*
(江西师范大学心理学院,江西 南昌 330022)

摘要:以现代测量理论为基础,该文尝试将认知诊断与计算机化自适应测验 2 项新技术应用于心理障碍(抑郁症)的诊断与测评,一方面探讨新技术在抑郁症诊断中的科学性与合理性,另一方面开发基于认知诊断计算机化自适应测验技术的抑郁症测评工具(简记为 CD-CAT-D). 研究共调查被试 2 492 人,经大样本数据标定及测量学指标筛选,题库最终保留 136 题;研究结果表明:在认知诊断理论框架下,CD-CAT-D 具有较高的诊断分类一致性信度;若以 PHQ-9 量表作为效标,CD-CAT-D 具有较理想的收敛效度和效标关联效度;同时,CD-CAT-D 的灵敏度与特异度平均在 0.850 左右,以及 AUC 指标在 0.80 ~ 0.90 之间,这些都表明 CD-CAT-D 具有较理想的预测效果. 这为心理障碍的诊断与评估提供了一种全新的方法和技术支持.

关键词:认知诊断;计算机化自适应测验;认知诊断计算机化自适应测验;项目反应理论;抑郁症
中图分类号:B 841.7 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2021.02.01

0 引言

作为新一代测量理论核心的认知诊断理论突破了传统测量理论(如经典测量理论(CTT)和项目反应理论(IRT)),注重对被试宏观特质水平的测量与评估,而且能为被试提供更为精细的微观认知属性的诊断与评估,从而达到了以测验促进被试发展的功能. 目前,认知诊断理论虽然主要应用于能力领域,但由于其数学模型的可扩展性强,已被学者们[1-4]应用于心理障碍的诊断与评估,并体现了比传统测量理论更多的优势(如能为患者提供详细的症状谱等).

计算机化自适应测验(CAT)是近 20 年来在项目反应理论(IRT)基础上发展起来的新测量技术,它强调对被试因人施测,不同被试由计算机根据相关的 IRT 算法智能化地实现对被试的测量. 与传统的纸笔测验(P&P)相比,CAT 最大的优点是可以在不损失测量精度的条件下大幅减少测试的题量并实现智能化的因人施测,从而减轻被试的测试负担. 随着测量理论的不断发展,学者们为充分发挥认知诊断理论(CDT)和计算机化自适应测验(CAT)的优势,将它们相结合,从而产生了认知诊断计算机化自适应测验(CD-CAT). CD-CAT 因能兼顾 CD 和 CAT 的优势,已被广大测量学研究者和实际应用者推崇,并成为当前国际上心理测量领域最前沿的技术之一.

目前,抑郁症是国际上公众关注度较高的一种心理障碍疾病,它具有高患病率、高复发率、高疾病负担、高致残率和高自杀率的特点. WHO 以平均伤残调整生命年(DALYS)作为评价指标,指出抑郁症在全球疾病总负担中位居第 4 位,并预测到 2030 年将成为仅次于艾滋病的第 2 大疾病负担[5]. 由此可见,抑郁症给人类社会带来的危害是十分巨大的.

与医学其他领域不同,在精神病学中的心理障碍不仅存在躯体症状,还存在精神症状. 躯体症状的表现可以通过患者客观的躯体表现出来,但精神症状却是主观的. 正是由于精神症状的主观性,心理学或精神病学的诊断评估缺乏客观的诊断工具(如仪器),医生往往依赖于患者对自身症状和严重程度的报告做出诊断评估. 因此,如何有效、准确而又客

收稿日期:2020-05-18
基金项目:国家自然科学基金(31660278)资助项目.
通信作者:涂冬波(1978—),男,江西南昌人,教授,博士,博士生导师,主要从事心理统计与测量研究. E-mail:tudongbo@aliyun.com

观地获取患者的自我报告成为一个关键问题. 经过大量心理学家与精神病学家的努力与探索,研制了用于测量心理障碍症状的心理量表,并成为目前国内外一种相对有效的评估筛选工具.

目前,比较常用的自评量表有抑郁自评量表(SDS)、流调用抑郁自评量表(CES-D)和贝克抑郁自评量表(BDI)等. 这些知名抑郁自评量表已被广泛接受并应用于抑郁症的临床诊断与评估中. 然而,由于这些量表编制的时间较早,而且都是基于传统的经典测量理论(CTT)编制而成,有许多不足: 首先,大部分抑郁量表的编制不是依据国际上通用的关于抑郁症诊断标准开发的. 如在美国精神医学学会颁布的 DSM-5 和由世界卫生组织颁布的 ICD-10 中,对抑郁症的诊断有明确的诊断标准及症状标准,然而 SDS、CES-D 和 BDI 等知名量表并没有测量 DSM-5 或 ICD-10 中的所有症状标准,因而不符合国际上通用的 DSM-5 或 ICD-10 的诊断标准. 其次,已开发的抑郁量表需要被试完成整个量表的所有题目,无法做到因人施测或因人诊断,从而会降低测试的效率. 最后,当前抑郁自评量表只能从宏观水平对患者进行综合诊断评估,无法提供患者在抑郁症各症状标准上的详细信息(即每个被试的症状谱,Symptom Spectrum). 然而,实际的情况是在传统抑郁量表中得分相同且均被评为抑郁患者的 2 个或多个被试的症状表现却不同,即症状谱不尽相同(有相同的测验分数但却有不同的症状表现),从而不利于对患者有针对性地预防、干预及治疗. 因此,开发一个智能化的因人施测、高测量精度而且能提供被试/患者抑郁症症状谱的测评工具显得十分必要. 而现代测量理论中涌现出的一些新的测量理论和新技术,尤其是认知诊断(CD)、计算机化自适应测验(CAT)和认知诊断计算机化自适应测验(CD-CAT)的出现,为这一智能化测评工具的开发、实现及应用提供了重要的理论、方法和技术支持.

鉴于此,为了弥补在当前国内外抑郁测评中存在的不足,本文以 ICD-10 关于抑郁症诊断标准为理论指导,以现代测量理论中的认知诊断理论、计算机化自适应测验及 CD-CAT 为方法及技术支持,开发一个智能化的因人施测/诊断、减轻测试负、提高测量精度而且能提供被试/患者抑郁症症状谱的测评工具(为描述方便,本文把该测评工具简称为 CD-CAT-D),以实现对患者高效、快速、准确的测评,并为患者提供症状谱等更为详细诊断信息. 本文一方面为抑郁症的诊断评估提供了一种全新的方法支持,另一方面为抑郁症的诊断评估提供了一种新的智能化的测评工具. 该测评工具有望实现对被试抑郁障碍进行高效、快速、准确的筛查与诊断,并为患者/被试提供更为详细的抑郁症症状谱信息,从而有利于对被试/患者开展有针对性的预防、干预及治疗.

1 研究方法与过程

1.1 抑郁症的诊断标准

目前国际上有 2 个知名的抑郁症诊断标准,分别是由美国精神医学学会颁布的 DSM-5 和由世界卫生组织颁布的 ICD-10. DSM-5 和 ICD-10 对抑郁症的诊断给出了详细的诊断及症状标准,具体症状标准如表 1 所示,目前 DSM-5 和 ICD-10 均被广泛应用于抑郁症的临床诊断中.

表 1 在 DSM-5 和 ICD-10 中关于抑郁症的症状标准

DSM-5
(i) Depressed mood
(ii) Markedly diminished interest or pleasure
(iii) Significant weight loss
(iv) Insomnia or phypersomnia
(v) Psychomotor agitation or retardation
(vi) Fatigue or loss of energy
(vii) Feelings of worthlessness or excessive or inappropriate guilt
(viii) Diminished ability to think or concentrate, or indecisiveness
(ix) Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan or a suicide attempt or a specific plan for committing suicide
ICD-10
Typical symptom criteria
(i) Depressed mood
(ii) Loss of interest and enjoyment
(iii) Increased fatigability
Common symptom criteria
(i) Reduced concentration and attention
(ii) Reduced self-esteem and self-confidence
(iii) Ideas of guilt and unworthiness (even in a mild type of episode)
(iv) Bleak and pessimistic views of the future
(v) Ideas or acts of self-harm or suicide
(vi) Disturbed sleep
(vii) Diminished appetite

DSM-5 和 ICD-10 的症状标准基本相似,但与 DSM-5 不同的是,ICD-10 将抑郁症的症状划区分为典型症状(typical symptom)和普通症状(common symptom). 在 DSM-5 中指出患者具备在 9 个症状中的 5 个或 5 个以上可考虑诊断为重度抑郁(Major

Depressive Episode); 而 ICD-10 指出患者具备 2 条典型症状以及 2 条普通症状可考虑轻度抑郁(Mild Depression) ,具备 2 条典型症状以及 3 ~ 4 条普通症状可考虑中度抑郁(Moderate Depression) ,具备 3 条典型症状以及 4 条普通症状可考虑重度抑郁(Severe Depression) . 考虑到 DSM-5 侧重于重度抑郁症患者的诊断,而 ICD-10 可诊断出轻度、中度和重度抑郁症 3 类患者. 相比较而言,ICD-10 提供的诊断分类信息更为丰富,因此本文采用 ICD-10 作为抑郁症症状标准.

1.2　认知诊断模型

在认知诊断理论框架下,认知诊断模型(CDMs) 将在 ICD-10 中的 10 项抑郁症症状标准划分为一个二分潜变量(具备或不具备) . K 个症状标准将会有 2^K 种症状模式(symptom profiles) ,每一种症状模式就是一个潜在类别. 将第 c 种症状模式记为 $\boldsymbol{\alpha}_c = (\alpha_{c1},\alpha_{c2},\cdots,\alpha_{ck},\cdots,\alpha_{cK})$,若患者具备症状 k 则 $\alpha_{ck}=1$,否则 $\alpha_{ck}=0$. 如 $\boldsymbol{\alpha}_c=(0,0,\cdots,0)=(\mathbf{0})_K$ 表示被试不具备 K 个症状的任何一个. 认知诊断模型将被试在量表上的作答反应与他们的症状模式进行数学建模,从而实现对患者症状模式的分类与诊断. 本文涉及的认知诊断模型主要有 G-DINA、DINA、A-CDM、LLTM、rRUM 等模型.

1.3　CD-CAT-D 开发与信度和效度验证

1.3.1　**项目来源**　在 ICD-10 框架下,题库最初的题目从 15 个国内外知名的抑郁自评量表中精心选取而来,选取的原则是项目须至少测量在 ICD-10 中关于抑郁症症状的 1 条症状; 同时少量自编项目以及少量 R. D. Gibbons 等[8] 使用的项目,共获取195 道试题. 需要说明的是: PHQ-9 仅作为效标量表,不参与本文的题库开发; 由于本文涉及的是自评量表,因此国际上知名的一些他评量表(如汉密尔顿抑郁量表) 不适合本文的测验.

1.3.2　**测验 Q 矩阵标定**　对初步选取的 195 题,邀请 5 位具有 5 年以上临床经验的抑郁症治疗师和 4 位有丰富心理测量经验的测量学者一起构建了测验 Q 矩阵,以 K_{appa} 系数作为评价多位专家评定结果一致性程度,并删除在专家标定 Q 矩阵中 K_{appa} 系数低于 0.4 的项目[7].

1.3.3　**分析过程与步骤**　**Step 1**　认知诊断模型选用. 选择恰当的认知诊断模型进行数据分析是在认知诊断中非常核心的环节. 尽管在认知诊断领域中开发了一系列模型,但人们在实际中往往不清楚哪个模型更适合哪个项目. 本文使用 Wald 统计量[8] 为每个项目选择一个最优认知诊断模型,本文采用认知诊断模型主要有 G-DINA、DINA、DINO、rRUM、A-CDM 和 LLM 等模型,包含了当前认知诊断领域主要的 CDMs.

Step 2　项目分析. 使用 Step1 选用的模型分析 CD-CAT-D 所有项目的测量学特征,包括项目区分度分析、项目水平的模型-资料拟合检验和项目功能差异(DIF) 诊查. 项目区分度指标采用 J. de la Torre[9] 提出的基于认知诊断的区分度指标,其计算公式为

$$D_{isc_j}=P(X_j=1\mid\alpha_{cj}^*=1)-P(X_j=1\mid\alpha_{cj}^*=0),$$

$P(X_j=1\mid\alpha_{cj}^*=1)$ 为具备项目 j 测量的所有症状的被试在项目 j 上的反应概率,$P(X_j=1\mid\alpha_{cj}^*=0)$ 为不具备项目 j 测量的任何一个症状的被试在项目 j 上的反应概率. 项目 j 区分度 D_{isc_j} 等于这 2 个概率之差,D_{isc_j} 值越大说明项目 j 越能将具备项目 j 所有症状的被试和不具备项目 j 任何症状被试区分开来,即项目的区分度越大. 同时,采用 R_{MSEA} 统计量检验每个项目与模型之间的拟合度[10] ,其计算公式为

$$R_{MSEA_j}=\Big(\sum_{l=1}^{2^K}p(\boldsymbol{\alpha}_c)\big(P_{\text{expected}}(X_j=1\mid\boldsymbol{\alpha}_c)-P_{\text{observed}}(X_j=1\mid\boldsymbol{\alpha}_c)\big)^2\Big)^{1/2},$$

$p(\boldsymbol{\alpha}_c)$ 为被试属于 $\boldsymbol{\alpha}_c$ 的边际比例(marginal proportion) ,K 为测验属性数. 根据 O. Kunina 等[10] 提出的标准,$R_{MSEA}>0.1$ 表示拟合较差,$R_{MSEA}<0.1$ 表示基本拟合. 最后使用 Wald 检验法[11] 诊查项目功能差异 DIF(如性别 DIF 等) .

Step 3　选择高质量试题构建 CD-CAT-D 的最终题库. 根据 Step2 对项目的分析,选择区分度大于 0.3($D_{isc_j}>0.3$) 、项目基本拟合($R_{MSEA}<0.08$) 并且无 DIF($p<0.05$) 的项目组成最终题库,即删除凡是区分低或者不拟合或者有 DIF 的项目.

Step 4　CD-CAT-D 算法设置. Step1 ~ Step3 主要是构建 CD-CAT-D 的题库,接下来需对 CD-CAT-D 的相关算法进行设置. CD-CAT-D 的算法涉及初始题选取规则、选题策略、参数估计以及终止策略. 认知诊断模型采用 Step1 为每个项目选择的模型,初始题选取采用随机法,选题策略采用当前 CD-CAT 领域应用较成熟的 PWKL 方法[12] ,其计算公式为

$$P_{WKL_j}(\hat{\boldsymbol{\alpha}}^t)=\sum_{c=1}^{2^K}\sum_{x=0}^{1}\Big(\big(P(x_j=x\mid\hat{\boldsymbol{\alpha}}^t)\log\big(P(x_j=x\mid\hat{\boldsymbol{\alpha}}^t)/P(x_j=x\mid\boldsymbol{\alpha}_l)\big)P(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)\Big)\Big),$$

其中 $P(\boldsymbol{\alpha}_c\mid\boldsymbol{X}_i)=P(\boldsymbol{\alpha}_c)L(\boldsymbol{X}_i\mid\boldsymbol{\alpha}_c)/\Big(\sum_{c=1}^{2^K}P(\boldsymbol{\alpha}_c)L(\boldsymbol{X}_i\mid$

$\boldsymbol{\alpha}_c$)),$P(\boldsymbol{\alpha}_c)$ 是知识状态 $\boldsymbol{\alpha}_c$ 的先验概率,$P(\boldsymbol{\alpha}_c \mid X_i)$ 是知识状态为 $\boldsymbol{\alpha}_c$ 的后验概率. 被试参数估计采用最大似然估计(MLE),其计算公式为

$$\hat{\boldsymbol{\alpha}}_i = \operatorname{argmax} \prod_{c=1}^{2^K} P(\boldsymbol{\alpha}_{il})^{X_{ij}} (1 - P(\boldsymbol{\alpha}_{il})^{1-X_{ij}}).$$

采用不定长的终止策略,当测验精度达到一定水平时,测验停止. 即若某被试在 CD-CAT-D 中达到某一预先设定的测量精度,则停止测试. 不定长终止策略的特点是被试的测量精度基本一致,但被试所有的题量不尽相同. C. Hsu 等[13] 以及蔡艳等[14] 在其研究中用后验概率 $P(\boldsymbol{\alpha}_c \mid X_i)$ 作为不定长 CD-CAT 的测量精度指标,即当被试的某个知识状态的后验概率达到预先设定值时,终止测试.

在本文中,为探讨 $P(\boldsymbol{\alpha}_c \mid X_i)$ 的何种标准对 CD-CAT-D 更为合理,本文设置了 4 种水平,分别比较了在 $P(\boldsymbol{\alpha}_c \mid X_i) > 0.75$、$P(\boldsymbol{\alpha}_c \mid X_i) > 0.80$、$P(\boldsymbol{\alpha}_c \mid X_i) > 0.85$ 和 $P(\boldsymbol{\alpha}_c \mid X_i) > 0.95$ 这 4 种终止规则下 CD-CAT-D 的效果,即在每种终止规则下所有被试使用题目数量的平均数和标准差,以及 CD-CAT-D 诊断的结果与使用题库所有题目诊断结果的一致率(C_R).

Step 5 CD-CAT-D 的信度与效度验证. 主要考察在 CTT 下 Cronbach 的 α 信度和分半信度,以及在认知诊断下症状分类一致性信度[15]. 效度主要考察效标关联效度和收敛效度,效标为PHQ-9量表和自我报告抑郁程度. 同时,分析在 CD-CAT-D 的 ROC 曲线下的面积(AUC),以及灵敏度(sensitivity)与特异度(specificity)指标,以充分考察 CD-AT-D 的预测效用(predictive utility).

1.3.4 研究被试 研究被试分 2 个部分:第 1 部分被试用于 CD-CAT-D 题库构建,这部分被试主要用于项目质量分析与项目筛选分析,更为重要的是用于估计 CD-CAT-D 题库的项目参数,记为样本 1;第 2 部分被试主要用于 CD-CAT-D 系统的信度与效度验证,记为样本 2. 这 2 部分有效被试总人数为2 492 人,其中抑郁症患者 816 人,正常被试 1 676 人. 考虑到 CD-CAT-D 题库建设及参数估计需要大样本容量,所以第 1 部分被试为 2 316 人(抑郁症患者 786 人),第 2 部分被试为 176 人(抑郁症患者 30 人).

2 研究结果

2.1 测验 Q 矩阵的标定

经删除 K_{appa} 系数低于 0.4 的项目,保留 172 题,这些试题的平均 K_{appa} 系数为 0.70($p < 0.05$). 表 2 是专家最终界定的项目 Q 矩阵示例,以第 2 题为例,它既反映出被试的兴趣和愉快感缺失(S_2)的情况,也可以说明其劳累、活动减少、精力减退(S_3),还伴随着消极的自我评价,反映了症状自罪和无价值感(S_6),即第 2 题同时测量了 S_2、S_3 和 S_6 共 3 个症状.

表 2 部分项目 Q 矩阵示例

项目内容	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
我觉得闷闷不乐,情绪低沉.	1	0	0	0	0	0	0	0	0	0
我几乎什么事都干不了.	0	1	1	0	0	1	0	0	0	0
我觉得我能把所有事做好.	0	0	0	1	1	1	0	0	0	0

注:$S_1 \sim S_{10}$ 表示在 ICD-10 中关于抑郁症的 10 项症状标准.

2.2 项目分析

根据项目区分度、拟合度和 DIF 这 3 项指标,删除凡是区分度低于 0.3 或者不拟合($R_{MESE} > 0.08$)或者存在 DIF($p < 0.05$)的项目,共删除 36 题,保留高质量的 136 题,组成 CD-CAT-D 的最终题库,它测量了在 ICD-10 中关于抑郁症的所有 10 项症状,平均每个症状被 21.8 题测量,测量症状最少的题数为 8(S_{10}),最多为 45(S_1);测量 1 个、2 个、3 个和 4 个症状的题数分别有 73、45、16 和 2. 整个题库 Q 矩阵包含了 4 个完整的 R 阵[16-17] 的题目,这些均符合认知诊断测验的要求.

表 3 呈现了 CD-CAT-D 题库部分项目的测量学特征,整个题库项目最小区分度为 0.310,最大区分度为 0.900,平均区分度达 0.601($S_D = 0.160$),这说明题库在整体上有较高的区分能力,题目的质量较理想;同时所有题目不存性别、年龄和地区的 DIF,也不存在所选用认知诊断模型不拟合的情况. 从总体来看,CD-CAT-D 的最终题库符合测量学要求,具有较高的质量.

2.3 CD-CAT-D 的测量性能与终止规则

表 4 为在 4 种终止规则下 CD-CAT-D 的测试效果. 由表 4 可知,不论在哪种终止规则下,CD-CAT-D 的诊断结果与使用所有题库诊断结果的一致率(C_R)均达 95% 以上,这说明 CD-CAT-D 可以达到理想的测量结果. 对 4 种终止规则比较发现,C_R 指标相当且均在 96% ~97% 之间,无实质性差异,但在终止规则为 $P(\boldsymbol{\alpha}_c \mid X_i) > 0.75$ 时,使用的题量最少. 因此,综合考虑题量使用量与 C_R 指标,终止规则

$P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i) > 0.75$ 更具优势,把该终止规则作为 CD-CAT-D的最终终止规则.

表 3　CD-CAT-D 题库的项目测量学特征(前 5 题)

题号	选用模型	区分度	R_{MSEA}	DIF(p 值)		
				性别 DIF	地区 DIF	年龄 DIF
1	RRUM	0.66	0.02	0.69	0.24	1.00
2	LLM	0.48	0.02	0.53	0.63	1.00
3	RRUM	0.58	0.02	0.83	0.26	1.00
4	GDINA	0.83	0.02	0.07	0.78	0.99
5	GDINA	0.34	0.04	0.77	0.17	0.94

表 4　不同终止规则的表现

终止规则	使用题量		C_R
	平均	S_D	
$P(\boldsymbol{\alpha}_c\|\boldsymbol{X}_i) > 0.75$	29.39	8.96	0.96
$P(\boldsymbol{\alpha}_c\|\boldsymbol{X}_i) > 0.80$	31.34	8.83	0.96
$P(\boldsymbol{\alpha}_c\|\boldsymbol{X}_i) > 0.85$	33.77	7.89	0.96
$P(\boldsymbol{\alpha}_c\|\boldsymbol{X}_i) > 0.95$	41.65	10.96	0.97

2.4　CD-CAT-D 信度与效度

2.4.1　**基于 CTT 和 CD 的信度分析**　CD-CAT-D 的 α 信度和分半信度分别为 0.984 和 0.977,具有较高的内部一致性信度. 在认知诊断理论(CDT)框架下,CD-CAT-D 的症状分类一致性信度从 0.896 到 0.996(见图 1),这说明 CD-CAT-D 对 ICD-10 中 10 项抑郁症症状均具有较理想的分类一致性信度. 但相比较而言,CD-CAT-D 对症状 S_{10} 的信度相对要低,其主要原因在于题库中测量症状 S_{10} 的题最少(仅有 8 题). 综上所述,不论是从 CTT 还是从 CDT 来看,CD-CAT-D 整体上均具有较理想的信度.

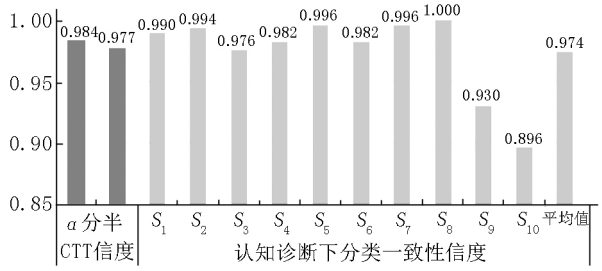

图 1　CD-CAT-D 信度

2.4.2　**CD-CAT-D 效度分析**　由表 5 可知,CD-CAT-D 不论是对被试整体诊断结果还是对每个症状的诊断结果均与 PHQ-9 量表以及被试自我报告的抑郁程度有着显著的相关性($p < 0.001$),且相关系数从 0.258 到 0.769,这说明 CD-CAT-D 具有较好的收敛效度和效标关联效度. 研究还发现,CD-CAT-D 的 AUC 介于 0.800 ~ 0.900 之间,灵敏度和特异度平均在 0.850 左右,这说明 CD-CAT-D 具有较好的预测准确性.

图 2 是抑郁组和正常组的被试(根据 PHQ-9 划分)被 CD-CAT-D 诊断为抑郁概率的 95% 置信区间图. 在图 2 中,抑郁组被试被 CD-CAT-D 诊断为抑郁概率的平均值为 0.814($S_D = 0.354$),而在正常组中被 CD-CAT-D 诊断为抑郁概率的平均值仅为 0.110($S_D = 0.290$). 经统计检验,$t = 14.481$,$d_f = 174$,$p < 0.000\ 1$,效果量 Cohen's $d = 2.178$,这说明 2 组被试在 CD-CAT-D 上的表现有非常大的差异,也说明 CD-CAT-D 能较好地区分出正常被试与抑郁症患者被试,具有较好的区分效应.

表 5　CD-CAT-D 效度

CD-CAT-D	PHQ-9 量表		自我报告
	测验总分	等级划分	抑郁程度
诊断结果	0.734	0.713	0.381
S_1	0.695	0.666	0.438
S_2	0.576	0.561	0.327
S_3	0.657	0.619	0.331
S_4	0.586	0.541	0.341
S_5	0.631	0.585	0.548
S_6	0.679	0.652	0.305
S_7	0.721	0.733	0.364
S_8	0.661	0.637	0.390
S_9	0.505	0.450	0.261
S_{10}	0.447	0.404	0.258

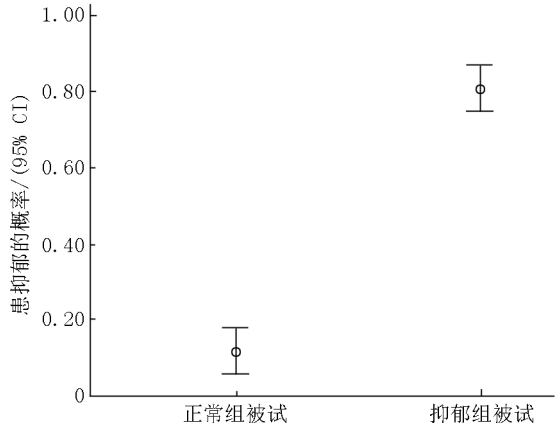

图 2　2 类被试被 CD-CAT-D 诊断为抑郁概率的 95% 置信区间

2.5　CD-CAT-D 个体诊断结果报告

图 3 是 2 位抑郁症患者症状谱,这 2 位患者在 SDS 量表上的原始分相同(均为 51 分),转换后抑

郁严重指数得分为 51/80 = 0.637 5,介于 0.600 ~ 0.690 之间,根据 SDS 量表诊断为中度抑郁. 同时,根据患者在 CD-CAT-D 上测试结果,被试 A 和被试 B 也均诊断为中度抑郁,与传统的知名抑郁症量 SDS 表诊断的结果一致. 但 CD-CAT-D 还可以为被试 A 和被试 B 提供更为详细的抑郁症症状断谱(见图 3),而这是传统抑郁量表(如 SDS、PHQ-9、CES-D)所无法提供的信息. 仔细对照这 2 位患者的抑郁症状谱,可以发现这 2 位患者尽管 SDS 量表总分相同,但所表现出的抑郁症状却相差较大:这 2 人都出现了 S_1(心境低落)、S_2(兴趣和愉快感丧失)、S_4(注意力降低)、S_7(无望感)、S_{10}(食欲紊乱)共 5 项症状,此外,被试 A 还表现出了 S_5(自我评价低);被试 B 身上还表现出了 S_6(自罪和无价值感)、S_8(自伤、自杀)、S_9(睡眠障碍)3 种不同的症状. 若主治医生只根据传统量表测验结果(相同的测量分数)采用同样的治疗方案,则显然是不合适的;但通过症状谱图,患者具体的情况一目了然,医生可做到对症下药. 因此,与传统的抑郁量表相比,本文开发的 CD-CAT-D 不仅可以提供对被试抑郁严重程度的评估与诊断,还可以为每位患者提供更具价值的抑郁症状谱,从而为患者及治疗师有针对性地进行干预及治疗提供重要支持;症状谱的另一个优势是还可用于评估康复患者的治疗效果,同时医生也可以清晰地了解每种治疗方案具体对于哪种抑郁症状是最有效的.

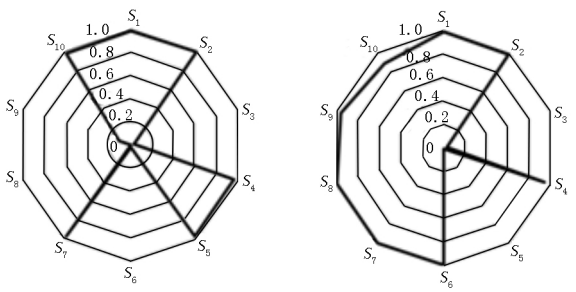

图 3 2 位中度抑郁症患者症状谱

3 研究结论与讨论

本文以 ICD-10 抑郁症诊断标准为理论基础,尝试将认知诊断与计算机化自适应测验 2 项新技术应用于抑郁症的诊断与测评,一方面探讨新技术在抑郁症诊断中的科学性与合理性,另一方面开发基于认知诊断计算机化自适应测验技术的抑郁症测评工具(简记为 CD-CAT-D). 与传统抑郁症测评工具相比,CD-CAT-D 不仅可以为每位被试/患者提供宏观的抑郁症诊断,还可以为每位被试/患者提供微观的抑郁症症状谱,为随后的抑郁症的预防、干预甚至治疗提供重要的指导,弥补了传统抑郁测量的不足. 研究结果表明:CD-CAT-D 具有较高的测量信度,在经典测量理论下 CD-CAT-D 的 α 信度指数和分半信度达 0.984 和 0.977;而在认知诊断理论框架下,CD-CAT-D 对 10 项症状的平均分类一致性信度也达到 0.974;若以 PHQ-9 量表作为效标,则 CD-CAT-D 具有较理想的收敛效度和效标关联效度;同时,CD-CAT-D 的灵敏度与特异度平均在 0.850 左右以及 AUC 指标在 0.80 ~ 0.90 之间,这意味着 CD-CAT-D 有较理想的预测效果.

限于时间及精力,本文在以下几方面还有待进一步探讨:(i)选用的是 ICD-10,未来研究可以进一步考察 DSM-5 的适合性及科学性,同时可以比较 DSM-5 和 ICD-10 在实际应用中的效果;(ii)选用了多个认知诊断模型,涉及到 G-DINA、A-CDM、RRUM、LLM 等模型. 而在认知诊断领域中,测量学者们开发了更多的认知诊断模型,因此未来也可以进一步考察其他认知诊断模型(如 LCDM 模型、GDM 模型等)在抑郁症测评中的可行性与合理性,从而为郁症的测评实践者和研究者提供更多可供选用的模型;在研究中仅有症状 S_{10} 的分类一致性信度低于 0.9,因此未来研究中还需进一步增加高质量的测量症状 S_{10} 的项目,以同时提升整个测评工具的信度;(iii)在研究中用于 CD-CAT-D 的信度、效度及测试效果的验证的样本量(即样本 2)偏少,因此为了进一步稳定 CD-CAT-D 信度、效度以及灵敏度和特异度指标的估计值,未来还需使用更多的被试做进一步的验证.

4 参考文献

[1] Templin J L, Henson R A. Measurement of psychological disorders using cognitive diagnosis models [J]. Psychological Methods, 2006, 11(3): 287-305.

[2] Jaeger J, Tatsuoka C, Berns S M, et al. Distinguishing neurocognitive functions using partially ordered classification models [J]. Schizophrenia Bulletin, 2006, 32(4): 679-691.

[3] de la Torre J, van der Ark L A, Rossi G. Analysis of clinical data from a cognitive diagnosis modeling framework [J]. Measurement and Evaluation in Counseling and Development, 2015, 51(4): 281-296.

[4] Tu Dongbo, Gao Xuliang, Wang Daxun, et al. A new measurement of internet addiction using diagnostic classifica-

tion models [J]. Frontiers in Psychology,2017,8:1-9.
[5] Mathers C D,Loncar D. Projections of global mortality and burden of disease from 2002 to 2030 [J]. PLoS Medicine, 2006,3(11):e442.
[6] Gibbons R D,Weiss D J,Pilkonis P A,et al. Development of a computerized adaptive test for depression [J]. Archives of General Psychiatry,2012,69(11):1104-1112.
[7] Viera A J,Garrett J M. Understanding inter-observer agreement: the kappastatistic [J]. Family Medicine, 2005, 37(5):360-363.
[8] de la Torre J. The generalized DINA model framework [J]. Psychometrika,2011,76(3):179-199.
[9] de la Torre J. An empirically based method of *Q*-matrix validation for the DINA model: development and applications [J]. Journal of Educational Measurement, 2008, 45(4):343-362.
[10] Kunina O,Rupp A,Wilhelm O. The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models [J]. Journal of Educational Measurement,2012,49(1):59-81.
[11] Hou L,De la Torre J,Nandakumar R. Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA model [J]. Journal of Educational Measurement, 2014, 51(1):98-125.
[12] Cheng Ying. When cognitive diagnosis meets computerized adaptive testing: CD-CAT [J]. Psychometrika, 2009, 74(4):619-632.
[13] Hsu C L,Wang Wenchung,Chen Shuying. Variable-length computerized adaptive testing based on cognitive diagnosis models [J]. Applied Psychological Measurement,2013, 37(7):563-582.
[14] 蔡艳,苗莹,涂冬波. 多级评分的认知诊断计算机化自适应测验 [J]. 心理学报,2016,48(10):1338-1346.
[15] Cui Ying,Gierl M J,Chang Huahua. Estimating classification consistency and accuracy for cognitive diagnostic assessment [J]. Journal of Educational Measurement,2012, 49(1):19-38.
[16] Templin J,Bradshaw L. Measuring the reliability of diagnostic classification model examinee estimates [J]. Journal of Classification,2013,30(2):251-275.
[17] Tatsuoka K K. Rule space: an approach for dealing with misconceptions based on item response theory [J]. Journal of Educational Measurement,1983,20(4):345-354.

The Application of Cognitive Diagnostic Computerized Adaptive Testing on Diagnosis and Assessment of Psychological Disorder

WANG Daxun,TU Dongbo*
(College of Psychology,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Based on modern measurement theory,it is attempted in this paper to apply two new technologies,cognitive diagnosis and computerized adaptive testing to the diagnosis and evaluation of mental disorders (depression). On the one hand,the science and rationality of new technologies are explored in the diagnosis of depression. On the other hand,a depression assessment tool based on cognitive diagnostic computerized adaptive testing (called as CD-CAT-D) is constructed. After model estimation of the data from 2 492 subjects,the items are screened by several Psychometric indicators,and 136 questions are finally retained in the question bank. The research results also show that under the framework of cognitive diagnosis theory,CD-CAT-D has high diagnostic classification consistency reliability. If the PHQ-9 scale is used as the criterion,CD-CAT-D has ideal convergent validity and criterion-associated validity. Meanwhile,the sensitivity and specificity of CD-CAT-D are around 0.850 on average and the AUC index is between 0.80 and 0.90 indicating that CD-CAT-D has good predictive validity. In short,a new method and technical support are provided for the diagnosis and evaluation of mental disorders.

Key words: cognitive diagnosis; computerized adaptive testing(CAT); cognitive diagnostic computerized adaptive testing(CD-CAT); item response theory; depression

(责任编辑: 冉小晓)