文章编号:1000-5862(2020)05-0454-08

计算机化自适应测验技术在情绪智力智能测评中的初步应用
——基于项目反应理论

张龙飞,刘 凯,宋 鸽,涂冬波*
(江西师范大学心理学院,江西 南昌 330022)

摘要:在项目反应理论(IRT)框架下,采用计算机化自适应测验技术实现对情绪智力的智能测评. 基于IRT系列分析(含单维性检验、模型拟合检验、局部独立性检验以及项目质量分析),构建了符合IRT测量学要求的情绪智力测评的题库,并以此为基础探讨了计算机化自适应测验技术在情绪智力智能测评(CAT-EI)中的应用. 实验结果表明:(i)CAT-EI相关算法具有较高的参数估计精度,同时具有较理想的测量信度和效度;(ii)CAT-EI可使用较少的题量(M_{ean}=9.88题)达到使用整个题库(67题)的测量精度,它一方面能做到减轻被试的测试负担,另一方面实现了对情绪智力高效、快速、准确的智能测评. 总之,该研究为实现对情绪智力智能测评提供了一种新的测量技术支持.

关键词:计算机化自适应测验;项目反应理论;情绪智力
中图分类号:B 841.7 **文献标志码**:A **DOI**:10.16357/j.cnki.issn1000-5862.2020.05.02

0 引言

情绪智力是当前心理学研究中比较热门的专题,这个概念可用于解释以下现象:在生活中,一些人的智力水平可能比不上另外一些智力较高的人,但是这些智力较低者的成就却可能高于后者. 传统意义上的智力对于个体成就水平的预测是不够精准的,个体是否能够取得成功的关键往往在于一些非智力因素(nonintellective factors)[1]. 基于这种现象,情绪智力的概念便逐渐发展了起来. P. Salvery等[2]定义情绪智力为个体监控自己及他人的情绪和情感,并识别、利用这些信息指导自己的思想和行为的能力. D. Goleman[3]在《情绪智力》中指出,情绪智力的好坏是决定一个人成为社会栋梁还是庸碌之辈的关键因素. 因此,情绪智力的重要性可见一斑. 进入新世纪后,中国学者对情绪智力进行了一系列研究,考察情绪智力对于各类心理特征的影响并取得了较为丰富的成果,如发现了情绪智力能够影响大学生的人际适应能力[4],情绪智力可以显著预测工作绩效[5],情绪智力与人们心理健康相关显著[6]. 由于情绪智力对于人们学习、工作和生活等各方面的成就均有较高的预测作用,因此开发并完善有效的情绪智力测验具有重要的实践意义.

到目前为止,对于情绪智力的测量主要通过纸笔施测(Paper & Pencil,P&P),它是基于经典测量理论(classical test theory,CTT)开发而成. 经典测量理论虽然有易于理解、操作且便于实施等优点,但该理论存在一系列的问题,如观察分数与真分数之间呈线性关系、对被试能力的估计依赖于题目的难度、测验参数对被试样本具有较强的依赖性、误差与真分数的假设在实际测试中很难满足等. 鉴于CTT的种种不足,新的测量理论——项目反应理论(item response theory,IRT)诞生了,IRT克服了CTT许多方面的缺陷. 项目反应理论具有项目参数不依赖于被

收稿日期:2019-10-29
基金项目:国家自然科学基金(31960186,31760288,31660278)和江西师范大学研究生创新基金(YJS2019089)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的研究. E-mail:tudongbo@aliyun.com

试样本的特点,与 CTT 只能在整个测验上计算出总信度不同,IRT 能计算出每一个人在每一道题上的信度,称为“信息量”(Information,IF). 计算机自适应测验(computerized adaptive testing,CAT)是基于项目反应理论而开发的,被学界誉为是一种非常适合各类心理评估的测量方法[7-8]. 它具有“因人施测,量体裁衣”的特点,针对不同能力的被试,计算机能够给出符合其能力水平的试题,通过使试题难度匹配被试能力,能够提供最大的信息量. 因此,CAT 可以实现真正意义上的“千人千卷”,并且测试结束,被试可立即获得测试分数. 一般的纸笔测验存在题量大、测验负担重、测验精度低等不足之处,而计算机自适应测验只需要较少的项目就可以实现精确的测量,从而减少被试作答时间[9]. 而且,由于 CAT 可实现“千人千卷”,即不同被试作答的测验题目不一样,因此 CAT 也可以有效地防止作弊行为.

目前计算机化自适应测验已用在许多国内外的大规模考试中,如中国汉语水平考试(HSK)、美国研究生入学考试(GRE)和托福考试(TOFEL)等. 作为一种智能化的测评模式,CAT 具有许多优良特性,可作为未来测验发展的一种趋势. 纵观国内外,当前对于情绪智力的 CAT 化研究还处于初始阶段,对于情绪智力评估的测试方式还使用传统的纸笔测验模式,这既跟不上新测验模式的发展潮流,也不利于完善情绪智力的测量. 本研究旨在解决这一问题,首次开发出情绪智力的计算机化自适应测验. 通过情绪智力测验的 CAT 化,可减轻被试的作答负担,使得原本需要动辄几十分钟才能完成的纸笔测试,提升为可因人施测、测量精度较高且测验耗时较短的智能化测评系统. 本研究拟在项目反应理论的基础上,探讨计算机化自适应测验技术在情绪智力测评中的应用及其优势,一方面为情绪智力的测量提供新的技术手段和测评工具,另一方面为实现情绪智力的 CAT 化提供理论和技术支持.

1　研究条件

1.1　情绪智力量表

本研究的项目选自国内外使用比较广泛、学术界较为认可的情绪智力量表,包括情绪智力量表(EIS)、自陈式情绪智力量表(WLEIS)和鹿特丹情绪智力量表(REIS).

1.1.1　*中文版情绪智力量表(EIS-C)*　情绪智力量表英文版(EIS)由 N. S. Schutte 等[10]于 1998 年编制而成,王才康等[11-12]于 2002 年对其进行了汉化. 中文版情绪智力问卷(EIS-C)共包含 33 个项目,测量了 4 个维度:情绪感知、自我情绪调控、调控他人情绪和运用情绪. 采用 5 级计分,由低到高,“1”表示“很不符合”,“5”表示“很符合”,EIS-C 具有良好的信度(克龙巴赫 α 系数为 0.84)和效度.

1.1.2　*中文版自陈式情绪智力量表(WLEIS-C)*　自陈式情绪智力量表英文版(WLEIS)由 C. S. Wong 等[13]于 2002 年编制. 中文版由王叶飞[14]在 2002 年汉化,共包括 16 个项目,测量 4 个维度:自我情感评估、他人情绪评估、运用情绪和调节情绪. 采用 7 级计分,“0”表示“非常不赞同”,“6”表示“非常赞同”. 王叶飞在对大学生、公务员、企业员工样本的施测中,发现 WLEIS-C 的信度较高,克龙巴赫 α 系数分别为 0.83、0.83、0.91,并且具有较高的效度.

1.1.3　*中文版鹿特丹情绪智力量表(REIS-C)*　鹿特丹情绪智力量表英文版(REIS)由 K. A. Pekkar 等[15]于 2017 年编制. 本研究将其引入国内,并通过以下步骤形成中文版量表(REIS-C):首先,经得原量表开发者的同意,由本文几位作者将其译为中文,形成初稿;然后,根据心理测量学理论和技术对量表进行修订,具体分析包含基于经典测量理论(CTT)下的测验信效度分析、基于验证性因素分析(CFA)的结构效度分析以及基于项目反应理论(IRT)的项目质量分析和测验质量分析;最后,形成中文版鹿特丹情绪智力量表(REIS-C),其克龙巴赫 α 系数为 0.91,分半信度为 0.88;原量表的 4 维结构(即关注自我情绪并评估、关注他人情绪并评估、对自我情绪的调节和对他人情绪的调节)仍然成立. REIS-C 共包括 28 题,它保留了原量表的所有项目,计分方式与原量表完全一致,即采用 5 级计分,从低到高,“1”表示“完全不同意”,“5”表示“完全同意”.

1.2　被试

将上述 3 个情绪智力量表同时对同一批被试进行纸笔施测,被试主要来自某省 4 所高校的大学生,有效被试数为 865 人,被试年龄分布为 16 ~ 25 岁,平均年龄为 19.28 岁(S_D = 1.22). 其中男生 347 人,女生 518 人;城镇户籍 332 人,农村户籍 533 人;文

科 344 人,理工科 521 人.

2 研究方法

2.1 基于 IRT 的情绪智力计算机化自适应测评系统(CAT-EI)的题库开发

为获得符合 CAT 标准的高质量题库,本研究运用项目反应理论对被试在初始题库上的作答进行分析以及参数估计. 最终题库须满足如下要求:测验的单维性(即所有题目整体上测量情绪智力这一公共维度)、局部独立性(即被试与被试之间、项目与项目之间作答是局部独立的)、项目符合测量学要求(即项目具有高区分度,与 IRT 模型拟合,以及无项目功能差异). 因此,通过以下步骤检验原始题库各项目在 IRT 框架下对于各项假设的拟合程度,筛选出合适的项目组成 CAT 题库,并获取各项目的参数.

2.1.1 *单维性检验* 为保证 CAT-EI 题库所有项目整体测量情绪智力这一公共维度,需要进行单维性检验. 本研究采取 D. Andrich 等[16-18]提出的经典方法及标准,在探索性因素分析(exploratory factor analysis,EFA)中,第 1 因子特征根与第 2 因子特征根的比值大于 4,且当第 1 因子的方差解释率大于 20%时,可认为测验基本符合单维性;在探索性因素分析过程中,对于第 1 因子负荷小于 0.3 的项目[18],需要进行删除,并再进行探索性因素分析. 本研究采用以上方法及标准进行单维性检验.

2.1.2 *IRT 模型选择* 选用合适的 IRT 模型是保证后续数据分析和参数估计准确的前提. 本研究考察当前使用较广泛的 2 种多级评分 IRT 模型对数据的拟合程度,这 2 个模型分别是等级反应模型(Graded Response Model,GRM)[19]和拓广分部评分模型(Generalized Partial Credit Model,GPCM)[20]. 根据模型与数据的拟合程度选择 IRT 模型. 模型拟合的指标选用经典的 $-2LL$(-2^* Log-likelihood),AIC(Akaike′s information criterion)[21]和 BIC(Bayesian information criterion)[22],这 3 种指标越小说明模型拟合越好.

2.1.3 *局部独立性检验* 只有保证了测验的局部独立性,才能在后续 IRT 分析中准确估计各类 IRT 模型参数. 本研究采用 Q_3 统计量指标[23]检验测验的局部独立性,若这 2 个项目之间 Q_3 统计量大于 0.36,则这 2 项目违背局部独立性假设[24]. 此时必须从这 2 个中删去与其他项目 Q_3 统计量累加较大的一个项目.

2.1.4 *项目质量分析* 项目质量分析针对单个项目的测量学指标进行分析,删除不符合相应标准的项目,以保证进入 CAT-EI 题库的项目均具有较高的质量. 项目质量分析包括 3 种:项目区分度、项目拟合度(item-fit)检验和项目功能差异(differential item functioning,DIF)检验.

项目区分度是指该项目能在多大程度上能区分不同能力的被试. 本研究删除区分度小于 0.7 的题目[25].

项目拟合度检验是指选用的 IRT 分析模型对于该项目上实际作答资料的拟合程度,本研究使用 $S-X^2$ 统计量[26]进行项目拟合检验,删除 $S-X^2$ 对应的 p 值小于 0.01 的项目[27],即删除与使用的 IRT 模型不拟合的项目.

项目功能差异(DIF)检验是指该项目上的作答是否会受到人口学变量的影响,如性别是否会影响某道题的作答. 本研究采用 Logistic 回归方法(LR)进行 DIF 检验,具体使用 McFadden′s pseudo R^2 的变化量进行评估[28],当 R^2 的改变量大于 0.02 时,项目存在 DIF[29],应予以删除. 本研究分别对被试性别、户籍所在地做了 DIF 检验.

2.2 CAT-EI 的算法性能及效度验证

CAT 的算法主要涉及以下几方面:初始题选取、选题策略、能力估计方法及终止策略[30]. 由于初始阶段对被试心理特质情况一无所知,因此初始题选取采用随机法,即测验的第 1 题是从 CAT 题库中随机抽取题目给被试作答;选题策略采用常用的最大 Fisher 信息量法(Maximum Fisher Information,MFI)[31],即从已作答的剩余题库中选取能使当前能力估计值信息量最大化的项目;能力参数估计方法采用期望后验(Expected a Posterior,EAP)方法;终止策略为不定长,即不对被试的作答项目数量(测验长度)进行规定,而规定达到一定测量误差(standard error,SE)或信息量(information,IF)便终止测验. IRT 测验信息量与测量误差存在关系 $I_F = 1/S_E^2$,即测量信息量越大,测量的误差越小. 在本研究中,通过规定测量误差(S_E)来设置 5 种终止规

则:$S_E<0.30$,$S_E<0.35$,$S_E<0.40$,$S_E<0.45$,$S_E<0.50$. CAT-EI 的算法和性能研究主要包括下述 2 个模拟实验.

2.2.1　实验 1:基于模拟被试的 CAT 模拟研究　实验 1 为基于模拟被试的实验,题库使用本研究开发的真实的 CAT-EI 题库及参数;被试为模拟被试,能力值区间为 -3.5 ~ 3.5,每隔 0.25 生成 100 名被试,共计 2 900 名模拟被试,这种模拟被试方法的最大好处是可以充分考察 CAT-EI 对每种类型被试的测量性能. 通过 CAT 算法获得 2 900 名模拟被试的模拟作答数据,并获得其能力的最终估计值. 通过分析真实值与估计值之间的相关指标以评估 CAT-EI 的性能及有效性.

2.2.2　实验 2:基于真实被试的 CAT 模拟研究　实验 2 为基于真实被试及真实的 CAT-EI 题库的实验,所有被试均为在开发 CAI-EI 题库时的真实被试,实验 2 需要调用这批真实被试的作答数据到 CAT-EI 实验过程中,即被试在纸笔测验中相应项目的作答数据视为在 CAT-EI 过程中的作答. 因此,实验 2 中被试为真实被试,其得分数据都是真实的,实验 2 需要做的是模拟这批真实被试在 CAT 中的自适应答题过程. 最后,通过分析基于全题库估计的能力值与基于 CAT-EI 估计的能力值之间的相关指标以评估 CAT-EI 的性能、有效性及其效度.

2.2.3　评价指标　实验 1 采用 CAT-EI 估计的能力与模拟的真实能力值之间的偏差(Bias)、平均绝对误差(mean absolute deviation, MAD)、均方根误差(root mean square error, RMSE)及相关系数(correlation coefficient)共 4 项指标来评价 CAT-EI 算法性能.

B_{ias} 是偏差,可反映测验能力估计是否有偏差,B_{ias} 越接近 0 表明能力估计值与真值之间的偏差越小,其计算方法为

$$B_{ias}=\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i), \tag{1}$$

其中 N 为被试人数,θ_i 为被试 i 的能力真值,$\hat{\theta}$ 为其能力估计值.

M_{AE} 是平均绝对误差,可反映能力估计的返真性,在 B_{ias} 的基础上它表明能力估计的偏差大小,其计算方法为

$$M_{AE}=\frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}_i-\theta_i|. \tag{2}$$

R_{MSE} 是均方根误差,可反映能力估计的稳定性,其计算方法为

$$R_{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)^2}. \tag{3}$$

实验 2 评价指标包括考察被试在 CAT-EI 中的平均测验长度、信息量、测量误差(standard error, SE)和边际信度(marginal reliability, MR)($M_R=1-S_E^2$)以及使用整个题库估计的能力值与使用 CAT-EI 估计能力值之间的相关系数. 并且,本研究以 3 个情绪智力量表(REIS-C、EIS-C 和 WLEIS-C)作为效标,分别计算被试在 3 个量表中的原始得分与其在 CAT-EI 估计中的潜在特质 $\hat{\theta}$ 之间的相关系数,以考察 CAT-EI 的效度.

2.3　分析软件

本研究中单维性检验使用 SPSS 22.0 软件进行 EFA,其余过程均使用 R 语言软件. 其中,模型拟合检验、局部独立性检验、区分度估计、项目拟合检验采用 R 语言 mirt 包[32],项目功能差异检验采用 R 语言 lordif 包[33],CAT 算法均采用 CATR 包[34].

3　结果

3.1　基于 IRT 的 CAT-EI 题库开发

3.1.1　单维性检验　先对 3 个量表共 77 题进行 EFA 分析,删除在第 1 因子负荷上小于 0.3 的项目,共删除 5 题,剩余 72 题. 对余下 72 题再次进行 EFA 分析,此时因子分析的第 1 特征根为 19.363,第 2 特征根为 3.775,第 1 特征根与第 2 特征根之比为 5.13,且第 1 特征根的方差解释率为 25.15%,满足第 1 特征根与第 2 特征根之比大于 4 且第 1 特征根方差解释率大于 20% 的条件. 因此,剩余 72 题符合 D. Andrich 等[16-18]关于单维性的标准,满足 IRT 的单维性假设.

3.1.2　模型选择　令 GRM 和 GPCM 这 2 个 IRT 模型作为备选模型,分别对作答数据进行拟合,根据这 2 类模型的拟合统计量选择最优模型. 在表 1 中,GRM 模型在 -2LL、AIC 和 BIC 这 3 项拟合指标上均小于 GPCM 模型,这说明 GRM 模型的拟合效果更佳,因此本研究采用 GRM 模型作为后续分析中的 IRT 模型.

表 1 2 模型拟合指标

模型	-2LL	AIC	BIC
GRM	74 063.73	148 917.5	150 800.1
GPCM	74 491.12	149 772.2	151 654.9

3.1.3 局部独立性检验 将选定的 GRM 模型用于局部独立性检验,发现第 70 题与其他项目的 Q_3 统计量大于 0.36,因此删去该题,此时题库剩余 71 题.

3.1.4 项目质量分析 区分度检验结果显示第 7、37 题区分度小于 0.7,删去这 2 题,剩余 69 题. 模型拟合检验显示第 25、47 题的 $S-X^2$ 对应的 p 值小于临界值 0.01,删除这 2 题,剩余 67 题.

在项目功能差异(DIF)检验中对于户籍和性别变量进行了检验,考察项目在这 2 个变量上是否存在 DIF. 结果显示所有项目在这 2 个变量上的 McFadden's R^2 变化量均高于临界值 0.02,因此无项目存在 DIF.

到目前为止,在题库中仍剩余 67 题,在题库建设流程中共有 13 题由于以上原因被剔除,对这剩余的 67 题再次进行单维性检验、IRT 模型选择、局部独立性检验、区分度检验、模型拟合检验和项目功能差异检验,结果表明题库剩余各项目均较好地符合各项标准. 并且对区分度指标的进一步分析发现,剩余 67 题的最大区分度为 1.72,最小区分度为 0.80,平均区分度为 1.22($S_D=0.23$),这说明最终 CAT-EI 题库在整体上质量较高.

图 1 表示整个 CAT-EI 题库提供给在整个能力区间上所有被试的测验信息量(I_F)和边际信度(M_R)(2 者之间关系为 $M_R=1-1/I_F$),由 67 题构成的题库的边际信度和测验信息量当能力值在 -3.5~2.0 之间时比较高,因此随后的 CAT-EI 可以对此能力区间的被试情绪智力评估产生较好的效果,最终的 67 题可以构建高质量的 CAT 题库.

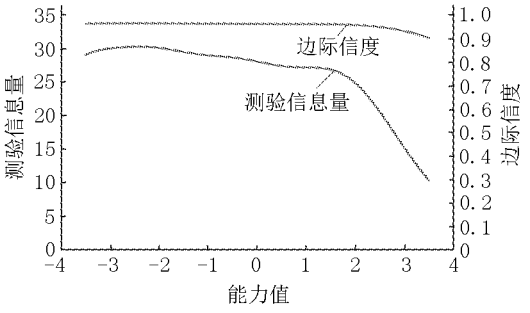

图 1 CAT-EI 题库测验信息量与测验边际信度

3.2 CAT-EI 的算法性能及效度验证

3.2.1 实验 1:基于模拟被试的 CAT 模拟研究 表 2 为基于模拟被试的 CAT-EI 各项性能指标. 在各终止规则下偏差(B_{ias})均接近 0,这表明 CAT-EI 参数估计具有无偏性;在各终止规则下平均绝对误差(M_{AE})均小于 0.500,这表明结果具有较强的返真性(Recovery);在 $S_E<0.30$、$S_E<0.35$ 和 $S_E<0.40$ 下均方根误差(R_{MSE})指标均小于 0.50,这表明估计值与真实值在这 3 种精度条件下十分接近,且估计值与真值的相关系数(r)在 5 种终止规则下均大于 0.950. 上述结果表明 CAT-EI 的算法参数估计精度较高,CAT-EI 的算法基本合理.

表 2 实验 1:模拟被试下不同终止规则 CAT-EI 性能指标

终止标准	B_{ias}	M_{AE}	R_{MSE}	估计值与真值相关系数
$S_E<0.30$	0.001	0.252	0.320	0.990
$S_E<0.35$	0.007	0.319	0.400	0.986
$S_E<0.40$	0.006	0.357	0.446	0.983
$S_E<0.45$	0.008	0.418	0.518	0.979
$S_E<0.50$	0.010	0.469	0.579	0.973

3.2.2 实验 2:基于真实被试的 CAT 模拟研究 表 3 为基于真实被试的 CAT 各项测量学特征指标,包括在不同终止规则下所用题量平均数(Mean)和标准差(standard error, SD)、平均测量误差(Mean SE)、边际信度以及基于 CAT 作答的估计值与基于全题库作答的估计值之间的相关系数(r).

表 3 实验 2:真实被试下不同终止规则 CAT-EI 性能指标

终止规则	所用题量		均值 S_E	边际信度	r
	均值	S_D			
使用整个题库	67.00	0.00	0.19	0.96	1.00
$S_E<0.30$	18.72	2.25	0.30	0.91	0.96
$S_E<0.35$	13.25	1.49	0.34	0.88	0.94
$S_E<0.40$	9.88	0.93	0.39	0.85	0.92
$S_E<0.45$	7.67	0.86	0.44	0.81	0.90
$S_E<0.50$	6.21	0.63	0.48	0.77	0.89

从表 3 中可以看出,当终止规则设定为 $S_E<0.30$、$S_E<0.35$ 和 $S_E<0.40$ 时,边际信度均在 0.85 以上,因此这 3 种条件下测验具有较高的可靠性. 而且相关系数在 $S_E<0.30$、$S_E<0.35$、$S_E<0.40$ 和 $S_E<0.45$ 这 4 种终止规则下均大于 0.90,所以在这 4 种终止规则下 CAT 估计的能力值十分接近使用整个题库估计出来的能力值. 而在所用题量上,在

$S_E<0.30$的终止规则下平均每名被试用了 18.72 题,$S_E<0.35$ 为 13.25 题,$S_E<0.40$ 为 9.88 题.

在 $S_E<0.30$、$S_E<0.35$ 和 $S_E<0.40$ 这 4 种终止规则下,分析被试 CAT-EI 能力估计值与所用题量和信息量的关系(见图 2 ~ 图 4),可以认定 $S_E<0.40$ 为最优终止规则. 该终止规则下能力值为 -2.5 ~ 2.5 之间的被试所用题量在平均值(9.88)附近密集分布,信息量均可以达到 6.25 以上,相当于边际信度 0.84 以上,能够取得较好的估计效果. 因此,在终止规则 $S_E<0.40$ 下,被试的能力估计可以较为精确,同时能够极大地减少测验负担. 以上结果表明:若使用 CAT-EI 系统进行情绪智力评估,则每名被试平均只需要完成 9.88 道题目便可达到使用全题库 67 题才能达到的精度.

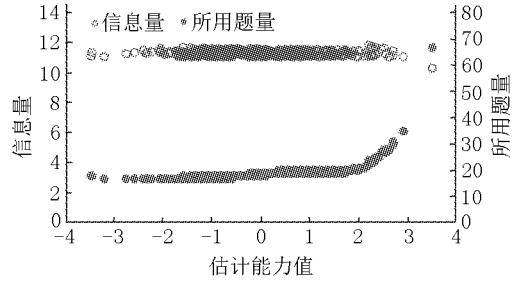

图 2　在 $S_E<0.30$ 下,被试估计能力值与测验信息量和所用题量的关系

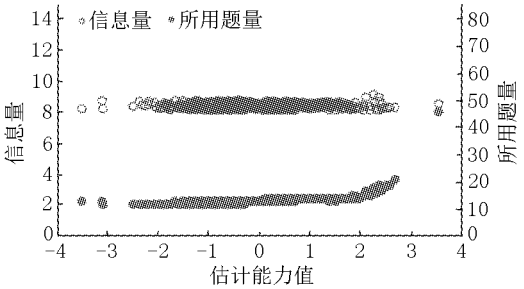

图 3　在 $S_E<0.35$ 下,被试估计能力值与测验信息量和所用题量的关系

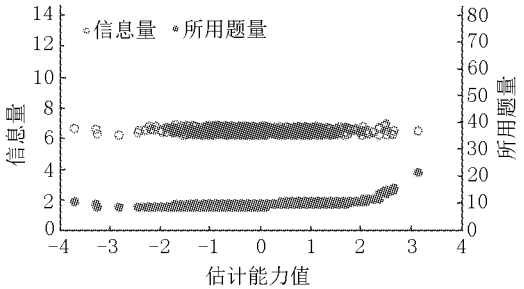

图 4　在 $S_E<0.40$ 下,被试估计能力值与测验信息量和所用题量的关系

3.2.3　CAT-EI 的效度验证　本研究以 REIS-C、EIS-C 和 WLEIS-C 作为 CAT-EI 的效标量表,分别计算被试的 CAT-EI 能力估计值与其在这 3 个量表上得分的相关系数获得效标关联效度,以考察 CAT-EI 的有效性. 结果显示:CAT-EI 对被试心理特质的估计值与 3 个量表得分都存在显著的相关($p<0.01$),且相关系数均大于 0.75,其中,CAT-EI 估计值与 REIS-C、EIS-C 和 WLEIS-C 的相关分别为 0.79、0.76 和 0.87,这说明 CAT-EI 具有较理想的效标关联效度.

4　结论与讨论

本研究目的在于构建一个基于 IRT 框架的高质量 CAT 题库,最终开发出情绪智力的 CAT 算法 CAT-EI,并通过模拟实验探究算法的性能和效度. 研究发现:

1)经过在 IRT 框架下各项题库建设流程后,最终的 CAT-EI 题库包含 67 个严格遵循单维性、局部独立性假设且较好地拟合等级反应模型 GRM、各类测量学指标良好的项目. 题库中项目的区分度指标均在 0.80 ~ 1.72 之间,平均值为 1.22,因此各项目对于被试均具备较高的区分能力,题库的边际信度在 0.90 以上;

2)实验 1 为基于模拟被试的 CAT 研究,其结果表明,CAT-EI 相关算法具有较高的参数估计精度,B_{ias} 均接近 0,M_{AE} 均小于 0.50,R_{MSE} 基本小于 0.50,且在各终止规则下 CAT 估计值与真实值之间的相关系数都在 0.97 以上;

3)实验 2 为基于真实被试的 CAT 研究,其结果表明,在研究所选用的 $S_E<0.40$ 终止规则下,被试使用 CAT-EI 进行测验,题量仅需 9.88 个,占总题量的 13.43%,节约了 86.57% 的题目,并且能力估计值与全题库估计值之间的相关系数高达0.92. 因此,CAT-EI 可以使用较少的题量($M_{ean}=9.88$ 题)达到使用整个题库 67 题的测量精度,它不但可以减轻被试的测试负担,而且可实现对情绪智力的高效、快速准确的智能测评.

当然本研究也存在一些不足之处,限于文章篇幅列出以下几点:首先,本研究只初步实现了 CAT-EI 的核心算法部分,尚未形成完整的 CAT 系统,仅通过纸笔测验收集的作答数据作为被试的作答反应,并非真正意义上的实证研究,模拟与实际之间势必存在出入,因此下一步研究工作应随机抽取一批或多批真实被试并用 CAT-EI 系统对其进行现实施测,以进一步检验测验系统的外部效度;其次,由于 CAT-EI 算法中选题策略采用的是最大 Fisher 信息

量法,Fisher 信息量又称为局部信息量(Local Information)[35],该算法是一种"贪婪"算法,会优先选用那些区分度较高的项目,这使得那些区分度较低的项目较少被使用[36]. 因此,可以在后续研究中采用曝光率控制(exposure control)技术平衡各题的出现次数,或可以尝试在 CAT-EI 中采用其他选题策略,如 KL 信息量选题方法[36]和最大优先级指标选题方法[37]等,以比较不同选题策略对测量精度的影响;再者,整个题库的测验信息量和边际信度在能力值为 2 以后下降得较快,这说明 CAT-EI 对于能力在 2 以上的被试测量误差更大,应当在后续研究中增加一些适合测量高情绪智力被试的项目到题库中去;最后,本研究题库中的项目共 67 个,一个优秀的 CAT 题库必须具备足够数量且高质量的项目,而且难度的覆盖范围也要足够大[38],因此未来研究还需继续充实和完善题库,可采用在线标定(On-line Calibration)的技术引入更多测量情绪智力的项目.

5 参考文献

[1] Alexander W P. Intelligence, concrete and abstract: note [J]. British Journal of Psychology, 1938, 29(1): 74.

[2] Salovey P, Mayer J D. Emotional intelligence [J]. Imagination, Cognition and Personality, 1990, 9(3): 185-211.

[3] Goleman D. Emotional intelligence [M]. New York: Bantam Books, 1995.

[4] 屠嘉俊,万娟,熊红星,等. 父母支持对大学生人际适应性的影响:情绪智力的中介作用 [J]. 心理科学,2016,39(4):964-969.

[5] 张辉华,王辉. 个体情绪智力与工作场所绩效关系的元分析 [J]. 心理学报,2011,43(2):188-202.

[6] 罗榛,金灿灿. 中国背景下情绪智力与心理健康关系的元分析 [J]. 心理发展与教育,2016,32(5):623-630.

[7] Embretson S E. The new rules of measurement [J]. Psychological Assessment, 1996, 8(4): 341-349.

[8] Meijer R R, Nering M L. Computerized adaptive testing: overview and introduction [J]. Applied Psychological Measurement, 1999, 23(3): 187-194.

[9] 涂冬波,蔡艳,戴海琦. 认知诊断 CAT 选题策略及初始题选取方法 [J]. 心理科学,2013,36(2):469-474.

[10] Schutte N S, Malouff J M, Hall L E, et al. Development and validation of a measure of emotional intelligence [J]. Personality and Individual Differences, 1998, 25(2): 167-177.

[11] 王才康. 情绪智力与大学生焦虑、抑郁和心境的关系研究 [J]. 中国临床心理学杂志,2002,10(4):298-299.

[12] 王才康,何智雯. 父母养育方式和中学生自我效能感、情绪智力的关系研究 [J]. 中国心理卫生杂志,2002,16(11):781-782,785.

[13] Wong C S, Law K S. The effects of leader and follower emotional intelligence on performance and attitude: an exploratory study [J]. The Leadership Quarterly, 2002, 13(3): 243-274.

[14] 王叶飞. 情绪智力量表中文版的信效度研究 [D]. 长沙:中南大学,2010.

[15] Pekaar K A, Bakker A B, van der Linden D, et al. Self-and other-focused emotional intelligence: development and validation of the rotterdam emotional intelligence scale (REIS) [J]. Personality and Individual Differences, 2018, 120(1): 222-233.

[16] Andrich D. A general hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Llikert methodologies [J]. British Journal of Mathematical and Statistical Psychology, 1996, 49(2): 347-365.

[17] Reckase M D. Unifactor latent trait models applied to multifactor tests: results and implications [J]. Journal of Educational and Behaviornal Statistics, 1979, 4(3): 207-230.

[18] Reeve B B, Hays R D, Bjorner J B, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS) [J]. Medical Care, 2007, 45(5): S22-S31.

[19] Samejima F. Estimation of latent ability using a response pattern of graded scores [J]. Psychometrika, 1969, 34(1): 1-97.

[20] Muraki E. A generalized partial credit model: application of an EM algorithm [J]. Applied Psychological Measurement, 1992, 16(2): 159-176.

[21] Akaike H. Stochastic theory of minimal realization [J]. IEEE Transactions on Automatic Control, 1975, 19(6): 667-674.

[22] Schwarz A S. The partition function of degenerate quadratic functional and Ray-Singer invariants [J]. Letters in Mathematical Physics, 1978, 2(3): 247-252.

[23] Yen Wendy M. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model [J]. Applied Psychological Measurement, 1984, 8(2): 125-145.

[24] Cohen J. Statistical power analysis for the behavioral science [J]. Technometrics, 1988, 31(4): 499-500.

[25] Fliege H, Becker J, Walter O B, et al. Development of a computer-adaptive test for depression (D-CAT) [J]. Quality of Life Research, 2005, 14(10): 2277-2291.

[26] Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models [J]. Applied Psychological Measurement, 2000, 24(1): 50-64.

[27] Flens G, Smits N, Terwee C B, et al. Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank [J]. Evaluation and the Health Professions, 2017, 40(1): 79-105.

[28] McFadden D. The measurement of urban travel demand [J]. Journal of Public Economics, 1974, 3(4): 303-328.

[29] Choi S W, Gibbons L E, Crane P K. Lordif: an R package for detecting differential item functioning using iterative Hybrid ordinal logistic regression/item response theory and Monte Carlo simulations [J]. Journal of Statistical Software, 2011, 39(8): 1-30.

[30] Chen S K, Cook K F. SIMPOLYCAT: an SAS program for conducting CAT simulation based on polytomous IRT models [J]. Behavior Research Methods, 2009, 41(2): 499-506.

[31] Baker F B. Item response theory: parameter estimation techniques [M]. New York: Marcel Dekker, 1992.

[32] Chalmers R P. Mirt: a multidimensional item response theory package for the R environment [J]. Journal of Statistical Software, 2012, 48(6): 1-25.

[33] Choi S W. Lordif: logistic ordinal regression differential item functioning using IRT [EB/OL]. [2019-08-14]. http://cran.utstat.utoronto.ca/web/packages/lordif/.

[34] Magis D, Barrada J R. Computerized adaptive testing with R: recent updates of the package CATR [J]. Journal of Statistical Software, 2017, 76(1): 1-19.

[35] Chang Huahua, Ying Zhiliang. A global information approach to computerized adaptive testing [J]. Applied Psychological Measurement, 1996, 20(3): 213-229.

[36] Chang Huahua, Ying Zhiliang. *a*-stratified multistage computerized adaptive testing [J]. Applied Psychological Measurement, 1999, 23(3): 211-222.

[37] Cheng Ying, Chang Huahua. The maximum priority index method for severely constrained item selection in computerized adaptive testing [J]. British Journal of Mathematical and Statistical Psychology, 2009, 62(2): 369-383.

[38] Wainer H, Dorans N J, Eignor D, et al. Computerized adaptive testing: a primer [J]. Quality of Life Research, 2001, 10(8): 733-734.

The Application of CAT on Emotional Intelligence with Item Response Theory

ZHANG Longfei, LIU Kai, SONG Ge, TU Dongbo*

(School of Psychology, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: The computerized adaptive testing is developed for emotional intelligence (CAT-EI) assessment based on item response theory (IRT). A high-quality item bank is built after a series of IRT analyses contain one-dimensional test, local independence test, model selection and item quality analysis, after that, the CAT-EI is developed and its performance is investigated in the simulation studies. The results show that CAT-EI makes a good estimation of characteristics of emotional intelligence and has reasonable marginal reliability and criterion-related validity. CAT-EI reaches the measurement precision of the whole CAT-EI bank (67 items) with the average of 9.88 items under the stop rule of $S_E < 0.40$. In brief, the CAT-EI not only can reduce respondent burden for the examinees that means saving considerable time for them, but also can make an efficient, quick and accurate assessment of emotional intelligence. In a word, a new psychometric technology for emotional intelligence is provided by this study.

Key words: computerized adaptive testing; item response theory; emotional intelligence

(责任编辑:冉小晓)