北京大学学报(自然科学版) 第 52 卷 第 2 期 2016 年 3 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 2 (Mar. 2016)
doi: 10.13209/j.0479-8023.2015.123

用项目反应理论编制中文版远程联想测验(RAT)

肖微 姚翔† 邱永桃

北京大学心理学系行为与心理健康北京市重点实验室, 北京 100871; †通信作者, E-mail: xiangyao@pku.edu.cn

摘要 遵循远程联想理论, 筛选适用于中国中学生群体的远程联想测验题本; 运用项目反应理论, 构建双参数模型, 使用来自广东、甘肃和湖北 3 所中学 2659 名初中学生的结果进行项目分析检验。遵照标准参照测验的编制原则, 选取难度为−3.00~3.00, 区分度为 0.30~2.50 以及 $\theta(\pi_0)$处信息量大于 0.20 的项目, 编制了中文版远程联想测验(RAT)(计 91 个项目)。该中文版 RAT 与托伦斯创造力测验、伦科创造力测验和瑞文智力测验的成绩正相关, 该测验可以预测中学生的学业成绩和教师评价的创造力表现。此外, 还讨论了该测验在实践中的应用价值。

关键词 创造力; 远程联想测验; 项目反应理论; 信息函数; 学业成绩

中图分类号 B84

Constructing Chinese Remote Associates Test (RAT) with Application of Item Response Theory

XIAO Wei, YAO Xiang†, QIU Yongtao

Beijing Key Laboratory of Behavior and Mental Health, Department of Psychology, Peking University, Beijing 100871;
† Corresponding author, E-mail: xiangyao@pku.edu.cn

Abstract The aim of the current study is to construct the Remote Associates Test (RAT) in Chinese version, which is based on the theory of associative creativity. Item response theory (IRT) with Binary Logistic Models was used for item selection. Participants were 2659 middle-school students from Guangdong, Gansu, and Hubei Province. Based on the principle of Criterion-Referenced Test, the item difficulty was limited between −3.00 and 3.00, the item discrimination was limited between 0.30 and 2.50, and the item information was above 0.20. In addition, the RAT scores were positively correlated with scores on the Torrance Test of Creativity, Runco Test, and Raven's Standard Progressive Matrices. Results of RAT were also positively correlated with teachers' evaluations of creativity and could significantly predict middle-school academic performance. Potential applications of RAT are also discussed.

Key words creativity; Remote Associates Test; item response theory; information function; academic performance

创造力(creativity)对个体的生存和发展具有举足轻重的作用[1]。创造力是一个特别受人青睐却很复杂、很难界定的概念[2], 长期以来, 心理学家基于不同的创造力理论开发了不同的创造力测评工具, 这些工具在研究和实践中得到广泛应用[3]。概括而言, 心理学家主要从发散式思维、创造性作品、创造力潜能以及创造性人格特征等角度来测量创造力。在诸多测量方法中, 测量发散式思维的托伦斯创造力测验(Torrance Test of Creativity Test, TTCT)广泛应用于教育领域和商业领域[4]。研究表明, TTCT 测量的创造力与智力和年龄正相关[5]。此外, 基于同感理论的同感评估技术(Consensus Assessment Technique, CAT)也是心理学领域经常使用的创造力测量方法, 这种方法强调通过评估创造

国家自然科学基金(31200792, 91224008, 91324201)资助
收稿日期: 2014−10−28; 修回日期: 2015−05−11; 网络出版日期: 2015−09−25

产品来测量创造力[6]。与其他测量创造力的方法相比，同感评估技术具有较好的生态效度[7]，广泛用于特殊领域中创造力的测量[8]，也被国内学者用于儿童语言创造力的研究[9]。对目前常用创造力测量方法的主要形式的简要比较见表 1。

如何快速简便地评估创造力仍然值得心理学家继续探索[10]。首先，创造力测量领域中的大部分方法有赖于专家评估，结果容易受到评估者的主观影响[11]。以 TTCT 为例，该测验词汇册的记分主要从流畅性、灵活性和独创性 3 个维度开展；图画册的记分主要参照流畅性、独创性、精致性、抗过早封闭性和标题抽象性等 5 个指标。上述分数的获得有赖于受过专门训练的评分者逐个评价测验成果[12]。其次，现有的某些创造力测验(如物品的非常规用途题目)往往在被试思维过程中引入太多的变量，很难清楚地区分，解决问题的时间过长，难以在标准化的测验中加以掌控。上述两个原因导致现有的创造力测验难以在实践中大规模推广，尤其是针对中学生群体的大规模标准化施测。此外，由于记分手段的限制，现有的测量工具很难运用在创造力神经机制的研究中。

本文尝试从发散式思维的角度出发，采用远程联想理论(The Theory of Associative Creativity)[13]解决上述问题。因为发散式思维被认为是创造力最主要的表现形式，有关发散式思维的测验广泛应用于教育和研究领域中创造力的测量[3]。远程联想理论认为创造性思考是将联想得来的元素重新整合的过程，因此，新结合的元素相互之间联想的距离越远，思维过程或者问题解决就越具有创造力。研究结果表明，创造力高的人更倾向于，并且更有能力进行远程联想[14]。

英文语境下的远程联想测验(RAT)主要基于 Mednick[15]的远程联想理论。在具体实施中，RAT 的基本方法是：向被试呈现 3 个词汇，然后请他们想出第 4 个词。例如：same-tennis-head，答案词是“match”。这种联系方式既可以是同义词(same-match)，又可以是语义联系(tennis-match)。另外，还有一个专门针对组合复合词组的复合远程联想测验(Compound Remote Associate Problems, CRAP)。复合远程联想测验将答案词与题目的这种关联固定到一种联系方式，例如：land-hand-house，答案词是“farm”，3 个给定词与答案词之间均是一种“复合词”关联[16]。曾有中国学者尝试按照复合远程联想测验的思路编制中文版本①。

RAT 测验体现了创造力的 3 个重要属性：在测验过程中被试的思维高度发散，测验题目不指向答案，甚至可能误导答案；解决问题的具体思考过程无法报告；测验过程伴随着顿悟的发生[16]。研究表明，RAT 确实能够测量个体在不同领域中的创造力。RAT 能够区分研究人员创造力的高低，RAT 得分较高的工程师能提出更多的改善方案[15]。另一项研究结果也表明，与 RAT 得分较低的科学家相比，得分较高的科学家能够提出更好的研究成果[17]。RAT 的分数与智商测验分数的相关值是 0.40，与教师创造能力的评定结果、建筑设计新颖性测试的结果正相关[14]。更重要的是，与其他现有

表 1　创造能力测量的主要形式简要比较

Table 1　Comparison of the most important creativity rating tests

测验方式	理论基础	测量内容	测验形式	主要变体	优点	缺点
TTCT	Guilford 的 SOI 智力结构	发散式思维	言语/图形	非常规用途测验(UUT)	全面反映发散式思维	分数解读，原创性/图形部分的计分依赖专家经验
远程联想测验	Mednick 的连接理论	发散式思维	言语		计分和解读操作简便	个体的语言思维能力/词汇量会影响最终得分依赖专家经验
CAT	Amabile 的同感理论	对作品创造性的评估	创造性作品		生态效度好	缺乏常模，打分依赖专家经验，易被低估
创造力检核表	多维智力理论	创造性潜能	评估	天才评估量表(GRS)	评估全面系统	专家打分，受评估者效应影响较大
创造人格测量 CPA	“大五”人格理论	人格特征	自陈式报告	开放性/ KAI	简便	自陈式报告，受自我偏差的影响

① 王烨，周晓林．编制中文复合远程联想测验(CRAP)．北京大学本科生科研项目结题报告，2006

的创造力测验相比，RAT 具有现场施测方面的优势。因为 RAT 存在明确的参考答案，题目形式简单固定，计分标准化，便于未经训练的人员使用，因此广泛应用于创造力的各种研究中[18]。

值得注意的是，现有的英文版 RAT 不能通过翻译或者修订直接适用于中国被试。首先，让呈现词和选出词组成复合词的形式很难在中文情景中实现；其次，研究者们认为 RAT 过分依赖语言能力和词汇量[12]。基于上述原因，对RAT不能通过简单翻译原版量表的形式加以修订，必须在中文语境中按照远程联想理论重新建立测验，题目内容的设计也需要更多地从形象思维的角度出发，降低语言能力或者词汇量的影响。因此本研究尝试在中文语境中编制中文版的 RAT，并进一步验证该测验与智力测验和其他主要的创造力测量工具的关系，以及该测验对于学生的学业成绩与创造力表现的预测作用。

研究编制中文版 RAT 是基于项目反应理论(Item Response Theory, IRT)。IRT 作为新兴的心理与教育测验理论，是在分析与克服经典测验理论局限性的基础上发展起来的[19]。IRT 能精确估计被试的潜在特质，依据更为精确的区分度、难度和信息函数峰值等挑选出更具有代表性的项目，提高量表的质量[20]。IRT 广泛应用于能力和成就测验的编制中，如项目反应理论在大规模选拔性考试试题质量评价中的应用[19]等。现有创造力测验多根据经典测验理论(CTT)编制，而 CTT 存在样本依赖性和信度估计欠精确等不足，IRT 有样本自由性和结果准确性等优点，因此根据 IRT 编制的创造力测验具有更好的信效度、更强的科学性。

1 测验材料的编制

1.1 材料收集

本研究遵循英文 RAT 的编制理论，采用同样规则，通过查找《现代汉语实词搭配词典》、《现代汉语词典》、《近义词反义词词典》等工具书以及网络搜索等形式收集测验项目。中文版 RAT 的题目同样由 3 个呈现词和一个答案词组成；呈现词之间无关联，即这 3 个词不同时出现在同一个时空关系或者语言情境中；呈现词与答案词的关系来自两个或者两个以上的关系类别，这些关系类别包括类别范畴、语义相同、组合词组等。例如呈现词“远行–红军–火箭”，答案词为“长征”。“长征”与“远行”语义相同，与“红军”、“火箭”是复合词组。通过初步收集和编写，共得到 355 个项目的远程联想原题检核表。

1.2 题目筛查

来自于北京大学的 9 名心理学研究者对 355 个检核表的项目按照以下 3 个规则进行独立筛选，并对筛除结果共同讨论。每轮筛选讨论，保证有超过 3 名研究者参与。具体规则如下。

1) 呈现词与答案词有关联。避免同时使用近义词和反义词，因为同时使用这两个类别的词汇，会引发截然相反的联想方向，导致答案词有多个。此外，在检核过程中还进行关联难度的检查。3 名研究者分别独立思考答案词，超出 20 秒未解答的项目为“过难”，需要修改或予以剔除。

2) 呈现词之间彼此不相关。一方面，呈现词两两之间不能同时语义相关，如果呈现词属于相同语义范畴称为语义相关，如“春季–秋日–冬天”；如果呈现词来自不同的语义范畴，则称为语义不相关，如“苹果–月亮–黄昏”[21]。另一方面，呈现词不能创设同一情境模型，如“警察–抓捕–小偷”创造追捕逃亡的情境模型，“冰雪–寒冷–腊月”创造冬天的情境模型。

3) 呈现词和答案词的表面内容应该符合要求。为了避免提供答案线索，呈现词与答案词没有重复，相邻项目的连接方式不相同；避免使用人名；所使用词汇的意义属于中学生的知识范围内(如某些中学生没见过扁担，想不出“扁担”与搬运的联系，故放弃)；避免使用生僻字、生僻词、自造词等。

经过多次讨论修订或删除不符合规则的项目后，最终保留 300 道题。中文版 RAT 的计分方法是：答对 1 道题计 1 分，被试答对题目的总和为最后得分。除给定答案词外，符合中文版 RAT 关系规则的答案也计分并进行记录，供备选答案库使用。

1.3 确定测试复本

对深圳某中学 36 名学生(男性和女性各 18 名，初中一、二年级学生各 18 名)的初测表明，完成 100 道测试题目的平均耗时为 34.61 分钟，与中文复合远程联想测验 15 分钟完成 50 道题的时间基本上一致。根据初中学生注意力保持的特点，在中学生群体中施测的时间以控制在 1 个小时内为宜；为了便于实施其他效标测验，最终施测选定中文版 RAT 的题目以 100 道左右为宜。所以研究者将 300

道题分成 3 个复本对 3 个独立样本分别施测。

2　正式测量

2.1　被试

参考已有研究对创造力测量学指标选取的思路，中学生创造力水平总的趋势是向前发展的，具有足够的个体差异性和测量学上的区分性[6]。因此，本研究选取广东、甘肃和湖北 3 所初中的学生进行测量学指标的检验，被试均是接受九年制义务教育的初中一年级和二年级学生，问卷的收发均获得校方的允许和授权。共发放问卷 2659 份，回收后对问卷进行差错补缺等工作(具体方法见测试流程)，最终按照字迹潦草程度(难以辨认)和测验的完成情况(缺失值达 30%以上)，进行无效问卷判别。由此获得的最终有效问卷为 2572 份(有效问卷回收率 96.7%)。被试中年龄最低 10 岁，最高 16 岁，平均年龄 13.80 岁(标准差 0.88 岁)；其中男性 1342 人，女性 1230 人，男女平均年龄无显著差异。所有被试均无此类测验经验。

2.2　测试流程

第一步，施测。测验由中学各班班主任监督实施。实施现场有研究者负责主持，测验的每个部分均有明确的时间限制，现场施测的成功关键是避免跨区作答。时间安排上，中文版 RAT 计 33 分钟(100 道题)，瑞文测验计 10 分钟(12 道题)，TTCT 计 10 分钟(2 道题)，伦科测验(RCAB)计 5 分钟(3 道题)，上述时间均包含解释说明的时间。正式测验时，使用学校的闭路广播播放标准话束(包括针对测验的总体说明、各个部分的具体说明和进程控制)。学生必须在规定时间内独立完成测验，不得跨区答题。问卷的现场发放和回收由班主任监督完成。

第二步，评分。23 名心理学硕士研究生对问卷独立评分。确保每次评分时至少有 7 名评分者。正式评分前，向评分者介绍测验的规则和要求，确保评分者掌握规则。数据分析前，研究者对评分和录入工作进行查错、补缺以及逻辑检查，处理缺失值。缺失值的定义是：1) 中文版 RAT 中没有作答的项目视为缺失；2) 在瑞文测验中，如果某个项目同时选定两个或者以上的选项、某个项目选定了题目中没有给出的选项以及没有作答的选项等情况视为缺失。使用序列均值代替缺失值。

第三步，数据检核和录入。对问卷录入整理后，导入 SPSS 备用。

2.3　统计方法

数据处理采用 SPSS19.0 和 CONQUEST1.0 软件进行。需要分析的内容包括数据与模型拟合检验、试题的参数、信息函数曲线、被试能力估计等。同时收集受测者的期末考试成绩和其他创造力测验的结果进行相关分析。

3　结果

测量学指标检验遵循如下的思路：分别使用 3 个独立的样本对 3 个中文版 RAT 测验的复本进行项目检验和分析；随后，按照检验结果筛选的最终题本进行信度检验。因为现场施测的复杂性和不可控性，效标测验的数据收集工作伴随着整个测验的施测过程进行。

3.1　测验的单维性检验

项目反应理论要求编制和修订的量表具有单维性，即只测量被试的某一种潜在特质，忽略其他潜在特质对测验结果的影响。被试对测验中任一项目的反应是该单一特质 θ 的函数。先前的研究推荐使用因素分析法进行单维性假设检验。单维性的检验标准是主成分分析结果的第一因子的特征根大于第二个因子特征根的 3 倍或以上，称为 Hambleton 标准[22]。分析结果表明，初筛题目的第一个因子特征根与第二个因子特征根的比值在不同题本中分别为 5.61/1.83，7.27/2.23 和 7.50/2.49，符合单维性假设的 Hambleton 标准。

3.2　参数估计与项目初选

为了解决经典测量理论对样本依赖性大的问题，项目反应理论采用局部独立性假设与样本独立项目校准的方法[23]。由此获得的项目参数具有不变性，各被测者或群体所得的项目参数具有可比性，所以我们对初筛题目的不同题本汇总，进行比较分析和筛选。

应用项目反应理论，采用双参数模型分析量表项目，使用边际极大似然估计法(MML)估计各项目的区分度参数、难度参数和各个项目与整个测验的拟合程度，从而挑选适合的项目。根据以往研究的经验[20]，设定难度和能力参数的取值范围均为[–3.00，+3.00]，区分度的取值范围为[0.30，2.50]，共有 113 个项目符合难度和区分度的要求。113 个项目的难度实际范围是[–2.44，2.45]，平均难度为

0.48; 区分度的实际范围是[0.30, 0.47], 拟合度的实际范围是[0.91, 1.03], 符合标准参照测验编制的标准。

3.3 划界分数的确定

为了选取信息量较大的项目, 首先需要选定测验的划界分数。项目反应理论具有同时估计被试能力参数和题目难度等参数的功能。正如可以按照被试的能力水平给被试排序一样, 也可以按照在某一特定概率水平上答对每一道题目所需要的被试能力水平来给题目排序。在实践操作中, 研究者们习惯使用书记法(bookmark)来确定测验的划界分数[24]。按照书记法的标准, 运用项目反应理论模型分析正确应答或获得相应分数的概率为 2/3 时所要求的能力值 θ, 在双参数模型中, $\theta = b + 0.693/1.7a$[25]。

项目分析完成后, 由专家按照试题难度排序的试题册和书签记录表来设置划界分数。评判专家从试题册第一试题页开始, 逐页判断合格水平最低能力的考生对该试题做出正确应答的概率或者获得相应等级分数的概率是否落在 2/3 以下, 如果回答是否定的, 就进入下一试题页判断; 否则, 就将书签安放在本页, 并在书签记录表中填写相应的试题页码[25]。本研究初步选定中文版 RAT 的划界分数 $\pi_0 = 0.40$, 经 IRT 测验信息函数验证, 当 $\pi_0 = 0.40$ 时, $\theta(\pi_0)$处的信息量处于测验的峰值范围内, 具有一定的合理性。

3.4 项目信息量分析

完成上述工作后, 挑选在划界分数水平附近具有最大信息量的项目。由项目反应理论分析得到的信息函数, 是潜在能力 θ 的连续函数, 当用极大似然法估计 θ 时, 估计量随样本量的增大而渐近正态分布, 则测验信息函数可以定义为能力估计值的方差的倒数, $I(\theta)=1/\mathrm{SE}(\theta)^2$。测验信息与测量误差是一一对应的。信息量越大, 测量精度越高。测验信息函数由每个项目信息函数累加, 每个条目可以单独对量表总信息做贡献, 贡献量大小不受量表其他条目的影响, 因此可以为增加或者删除条目提供依据[26]。

为了使测验在划界分数 π_0 处误差较小, 当测验长度较大时(如 $m>90$), 若 0 或 1 记分的题目 $\theta(\pi_0)$ 处信息量小于 0.10, 则此题应删除不用; 若 $\theta(\pi_0)$ 处信息量大于等于 0.10 而小于 0.20, 则此题可以修改后再用; 若 $\theta(\pi_0)$处信息量大于等于 0.20, 则此题较适合使用[26]。根据项目信息量的要求, 本文删除 $\theta(\pi_0=0.40)$处的信息量小于 0.20 的项目。图 1(a)为项目 32 的信息函数曲线, 可以看出该项目的作用不大, 不仅在能力区间[–3.00, 3.00]的信息量小, 而且在划界分数点上的信息量更小, 对整个测验在划界分数点上的信息量的贡献甚微, 不能很好地区分合格和不合格的被试, 质量较差, 予以删除。

图 1(b)为项目 45 的信息函数曲线, 可以看出, 该项目虽然在能力区间[–3.00, 3.00]有一定的信息量, 但是当能力达到 3.30 左右时, 项目的信息量最大。在划界分数处的信息量较小, 不符合要求, 予以删除。

图1(c)是项目21的信息曲线函数, 具有较好的信息量, 在能力区间[–3.00, 3.00](包括划界分数点)上都有高信息量, 对整个测验信息量贡献大, 测量误差小, 质量良好, 予以保留。

按上述方法共删除 22 个项目, 最终保留 91 个项目。最终版本的难度、区分度和拟合度指标见表 2。整个测验的信息函数见图 2。可以看出, $\theta(\pi_0=$

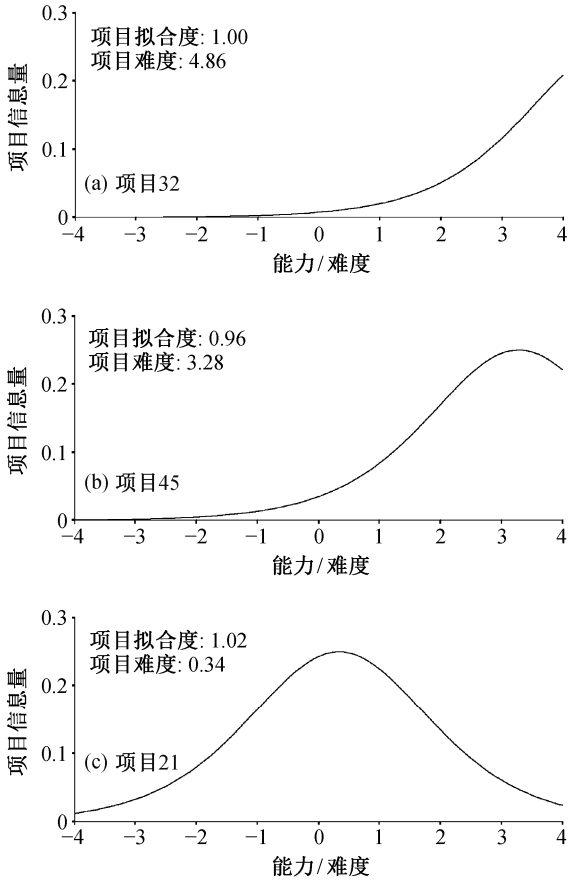

图 1 项目 32, 45 和 21 的信息函数
Fig. 1 Information function of the three items

0.40~0.80)之间的信息量普遍较大(大于20)，在一定程度上能区分不同水平的被试，符合标准参照测验的思想(整个测验在划界分数点附近有高信息量)[26]。经典测验方法的中文版 RAT 的内部一致性信度达到 0.92。

3.5　效度检验

为了检验中文版 RAT 的预测效度和区分效度，我们选取甘肃某中学初中一、二年级的学生(初中一年级学生 n_1= 484，初中二年级学生 n_2= 452)为研究对象。选取 n_2 样本进行与学业成绩的相关分析，中文版 RAT 与学生期末考试成绩的相关性见表 3。可以看出，中文版 RAT 与各科测验成绩之间均显著相关，相关值在 0.20~0.40 之间波动，其中语文和英语测验成绩与中文版 RAT 的相关程度相当于甚至高于数学和物理测验成绩与中文版 RAT 的相关程度。

表 2　中文版 RAT 测验的难度、区分度和拟合度(按难度降序排列)

Table 2　Difficulty, discrimination and weighted MNSQ of RAT (descending order of difficulty)

题号	难度	区分度	拟合度	题号	难度	区分度	拟合度	题号	难度	区分度	拟合度
1	2.32	0.30	0.96	32	0.91	0.40	0.95	63	0.35	0.37	0.98
2	2.30	0.32	0.96	33	0.90	0.33	0.97	64	0.34	0.32	1.02
3	2.30	0.34	0.94	34	0.88	0.32	1.00	65	0.29	0.30	1.01
4	1.96	0.31	0.97	35	0.88	0.32	0.99	66	0.25	0.37	0.97
5	1.80	0.31	0.98	36	0.87	0.31	0.99	67	0.23	0.32	0.99
6	1.64	0.32	0.98	37	0.83	0.37	0.95	68	0.19	0.30	1.01
7	1.62	0.30	0.99	38	0.81	0.35	0.99	69	0.16	0.34	1.00
8	1.57	0.30	0.97	39	0.80	0.32	0.98	70	0.14	0.31	1.00
9	1.55	0.33	0.97	40	0.77	0.32	0.98	71	0.12	0.33	0.99
10	1.48	0.32	0.97	41	0.73	0.35	0.97	72	0.09	0.36	0.99
11	1.45	0.34	0.97	42	0.71	0.30	1.02	73	0.09	0.46	0.91
12	1.44	0.30	0.97	43	0.69	0.32	0.99	74	0.05	0.40	0.95
13	1.41	0.30	1.00	44	0.65	0.33	0.99	75	0.02	0.34	0.98
14	1.34	0.31	0.97	45	0.64	0.39	0.95	76	0.01	0.32	1.00
15	1.32	0.36	0.95	46	0.64	0.31	0.99	77	–0.16	0.36	0.97
16	1.30	0.31	0.99	47	0.62	0.37	0.97	78	–0.20	0.30	1.01
17	1.27	0.33	0.98	48	0.61	0.32	0.98	79	–0.24	0.31	0.99
18	1.23	0.31	0.97	49	0.61	0.31	1.00	80	–0.24	0.36	0.98
19	1.22	0.31	0.98	50	0.59	0.34	1.01	81	–0.34	0.37	0.97
20	1.20	0.31	1.00	51	0.59	0.30	1.00	82	–0.40	0.30	1.01
21	1.20	0.34	0.97	52	0.58	0.35	0.97	83	–0.41	0.36	0.97
22	1.19	0.33	0.98	53	0.57	0.42	0.95	84	–0.45	0.47	0.91
23	1.10	0.34	0.96	54	0.48	0.32	0.99	85	–0.47	0.37	0.97
24	1.09	0.32	0.98	55	0.45	0.32	1.01	86	–0.55	0.33	0.99
25	1.08	0.36	0.95	56	0.40	0.30	1.01	87	–0.55	0.31	1.00
26	1.04	0.32	0.98	57	0.38	0.36	0.99	88	–0.59	0.35	0.98
27	0.98	0.33	1.00	58	0.38	0.44	0.92	89	–0.61	0.33	0.98
28	0.96	0.35	0.96	59	0.38	0.37	0.96	90	–0.76	0.40	0.94
29	0.95	0.34	0.96	60	0.38	0.35	0.98	91	–0.96	0.36	0.95
30	0.92	0.33	1.00	61	0.37	0.31	1.01				
31	0.92	0.30	1.00	62	0.37	0.32	0.99				

表 3 初中二年级创造力测试结果与学业成绩的相关

Table 3 Mean, SD and correlation of RAT results and student academic performance

测验	*M*	SD	中文版 RAT	数学	语文	英语	物理
中文版 RAT	32.12	11.18					
数学	103.66	10.90	0.29**				
语文	85.40	8.40	0.29**	0.56**			
英语	94.06	17.34	0.37**	0.69**	0.63**		
物理	73.35	13.79	0.34**	0.68**	0.51**	0.66**	
四科总分	356.30	43.20	0.38**	0.85**	0.75**	0.90**	0.85**

注: *表示相关性在 0.05 水平上显著; **表示相关性在 0.01 水平上显著(双尾检验)。

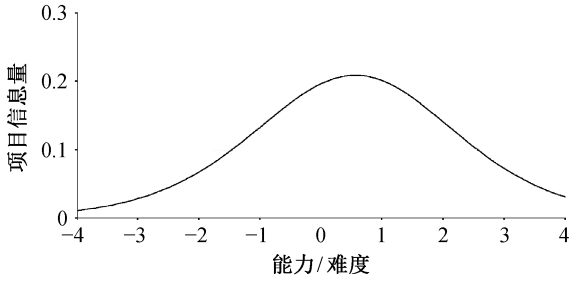

图 2 测验的信息函数

Fig. 2 Test information function

因为该样本包含初中两个年级的学生，所以采用这两个样本进行区分效度的检验。独立样本 T 检验的结果表明，初中一年级学生(M_1= 28.46)和初中二年级学生(M_2= 32.06)的中文版 RAT 结果差异显著(t=–5.30, p<0.01, η^2=0.03)，说明该测验具有较好的区分效度，与张景焕等[6]的研究结果一致。

选取湖北某中学的初中一、二年级学生进行中文版 RAT 的效标关联效度检验(n_3= 716)。关联效标选用瑞文测验、TTCT 测验和 RCAB 测验。其中，中文版 RAT 随机抽取最终题本中的 40 个题目。中文版 RAT、瑞文测验、TTCT 和 RCAB 相关性分析见表 4。结果表明，中文版 RAT 与瑞文测验、TTCT 以及 RCAB 的得分均显著正相关，与瑞文测验的相关系数达 0.34 (p<0.01)，与 TTCT 和 RCAB 的相关达 0.18 (p<0.01)和 0.28 (p<0.01)。

选取深圳某中学的两个班级(n_4= 93)进行预测效度的检验，由 3 名参与班级教学、总教学年限不少于 6 年的教师评价每名学生的创造力。使用学生行为特质评定量表的创造力维度(该量表采用 6 点计分，1= 从不，6= 总是)。分析得到，教师评价结果与中文版 RAT 正相关(r= 0.38, p<0.01)。文献[27]的研究表明创造能力测验与教师评估的相关在 0.20~0.70 之间，因此本测验的校标效度可以接受。

表 4 中文版 RAT、瑞文测验、TTCT 和 RCAB 的描述统计与相关系数

Table 4 Mean, SD and correlation of RAT, Raven's SPM, TTCT, RCAB

测验	*M*	SD	中文版 RAT	瑞文测验	TTCT
中文版 RAT	15.55	6.185			
瑞文测验	5.11	2.719	0.34**		
TTCT	17.71	7.014	0.18**	0.16**	
RCAB	9.40	4.924	0.28**	0.13**	0.47**

注: *表示相关性在 0.05 水平上显著; **表示相关性在 0.01 水平上显著(双尾检验)。

4 讨论

中文版 RAT 的编制首次将项目反应理论运用到创造力测验的项目分析中。项目反应理论通常用于教育考试的研究中，关于创造力测验的研究较少，国外关于 RAT 的项目检验还停留在经典测量的领域[16]。模拟研究的结果表明，IRT 方法得到的测验结果与 CTT 的结果相比，能更准确地估计被试的特质水平[28]。本研究选用双参数模型对测验的每一个项目进行分析，并选取划界分数下信息量大于 0.20 的项目，保证整个测验的信息量，提高了测验的精确度。

4.1 中文版 RAT 与校标测验关系的分析

RAT 测验与诸多效标测验之间的关系有 3 个方面的问题值得注意。

首先，虽然中文版 RAT 与 TTCT 和 RCAB 的相关性显著，但相关系数低于 0.30。这可能是因为与智力的概念相类似，创造力的操作性定义存在很大差异，导致基于不同创造力理论的测验测量了创

造力的不同成分，所以各个测验间的相关性不高。研究表明，Abedi-Schumacher Creativity Test (CT)和TTCT 各分量表的相关性虽然达到 0.01 水平的显著，但平均相关系数只有 0.11[29]。

其次，创造力(中文版 RAT)与智力(瑞文测验)的测验结果呈中等程度的相关，与前人研究结果[5]相符。智力的一些成分是创造力的重要基础，创造力思维测验中需要多种能力以及智力因素的参与，如定义和重新定义问题的能力以及洞察力[30]。但是二者之间又不完全相同，瑞文测验主要测查被试的推理能力，而中文版 RAT 主要测查被试的创造性思维及远程联想能力。因此，中文版 RAT 与瑞文测验呈中等程度的相关。

最后，在学科成绩方面，学生中文版 RAT 与语文、数学等学业测验成绩显著相关。这与大多数研究结果相符合，已有研究表明，用于培养分析思维、创造性思维、记忆力等的学科的训练(如阅读、社会学习、科学、数学等)均能提升学业成就表现[31]。值得注意的是创造力(中文版 RAT)与语文成绩的相关性高于数学成绩。这可能是由于 RAT 是运用联想词语来测量创造力，所以 RAT 与被试的语言能力以及词汇量相关性显著，语文成绩好的学生在作文等文字表达中能运用更丰富的语言表达形式，所以两者呈现更显著的相关性。文献[5]表明，中文版 TTCT 与中学生语文成绩在 7 个得分上显著相关(共 11 个得分)，与数学成绩在 9 个得分上显著相关。这也在一定程度上表明，RAT 所测量的创造力与 TTCT 所测量的创造力存在一定的差异。

4.2　中文版 RAT 的实践价值

中文版 RAT 的编制在应用和推广中有特有的优势。首先，中文版 RAT 的 100 个题目在 33 分钟左右即可完成，均有供参照的标准答案，可以做到标准化评分且计分简单。基于以上特性，中文版 RAT 可以以团体测验的形式大规模评估中学生的创造力，可以应用于口语报告范式以及研究创造力的脑机制等研究中。其次，本测验在编制过程中，尽量选择通俗常见的词汇，无生僻字、生僻词，不受专业词汇的影响，保证了测验的普适性，因而可以进一步尝试推广到其他人群，为今后在企业中测量员工的创造力提供了可能性。

4.3　测验的不足与展望

本研究存在以下不足。首先，中文版 RAT 侧重于发散式思维以及远程联想组织能力，对创造性人格以及其他社会文化环境因素的体现相对较少，对创造过程的研究更多依赖理论或者假设，使测验本身在解释创造力方面有一定局限。其次，本次研究的对象为中学生，样本群体相对单一。在后续研究实践中，可以尝试扩大样本容量和层次，研究该测验对不同群体的适用性以及对现实生活中创造力表现的预测能力，将其与实践中的创新表现联系起来进行检验和修订，使测验进一步优化。此外，由于中学生的注意力特点以及施测的总体题目数量的限制，研究的效标证据是随着 3 个研究复本的项目筛选展开的，这并不是检验效标的标准方法。在后续的研究中，可以尝试克服题目数量和测验时长的限制，统一收集最终题本和效标的证据，进一步给出效标效度的结果。

未来的研究中，拟将中文版 RAT 的最终版本在企业中施测，分析在企业员工测量中的测量学指标。在完善和标准化中文版 RAT 的基础上，进一步提供认知神经方面的证据，根据 ERP 和 fMRI 的呈现结果对个体进行创造力训练，再检验创造力培训的结果，在科学测量创造力的基础上达到有效培养创造力的最终目的。

参考文献

[1] Yao X, Yang Q, Dong N, et al. The moderating effect of Zhong Yong on the relationship between creativity and innovation. Asian Journal of Social Psychology, 2010, 13 (1): 53–57

[2] 施建农, 陈宁, 杜翔云, 等. 创造力心理学与杰出人才培养. 中国科学院院刊, 2012, 27(增刊): 164–173

[3] Dietrich A, Kanso R. A review of EEG, ERP, and neuroimaging studies of creativity and insight. Psychological Bulletin, 2010, 136(5): 822–848

[4] Kim K H. Can we trust creativity tests? a review of the torrance tests of creative thinking (TTCT). Creativity Research Journal, 2006, 18(1): 3–14

[5] 张景焕, 张广斌. 中学生创造性思维发展特点研究. 当代教育科学, 2004(5): 52–54

[6] Horn D, Salvendy G. Consumer-based assessment of product creativity: a review and reappraisal. Human Factors and Ergonomics in Manufacturing & Service Industries, 2006, 16(2): 155–175

[7] 宋晓辉, 施建农. 创造力测量手段: 同感评估技术(CAT)简介. 心理科学进展, 2005, 13(6): 739–744

[8] Han K. Domain-specificity of creativity in young

children: how quantitative and qualitative data support it. The Journal of Creative Behavior, 2003, 37(2): 117–142

[9] 曲小军, 施建农. 评价和奖赏对场依存, 场独立儿童语言创造力的影响. 中国心理卫生杂志, 2005, 19(6): 408–412

[10] Kaufman J C, Plucker J A, Russell C M. Identifying and assessing creativity as a component of giftedness. Journal of Psychoeducational Assessment, 2012, 30(1): 60–73

[11] Hempel P S, Sue-Chan C. Culture and the assessment of creativity. Management and Organization Review, 2010, 6(3): 415–435

[12] 王烨, 余荣军, 周晓林. 创造性研究的有效工具. 心理科学进展, 2005, 13(6): 734–738

[13] Mednick S. The associative basis of the creative process. Psychological Review, 1962, 69(3): 220–232

[14] Ansburg P I, Hill K. Creative and analytic thinkers differ in their use of attentional resources. Personality and Individual Differences, 2003, 34(7): 1141–1152

[15] Mednick S A. The remote associates test. The Journal of Creative Behavior, 1968, 2(3): 213–214

[16] Bowden E M, Jung-Beeman M. Normative data for 144 compound remote associate problems. Behavior Research Methods, Instruments, & Computers, 2003, 35(4): 634–639

[17] Andrews F M. Social and psychological factors which influence the creative process. Perspectives in Creativity, 1975, 117: 145

[18] Eysenck H J. The measurement of creativity // Boden M A. Dimensions of Creativity. Cambridge, MA: MIT Press, 1994: 199–242

[19] 戴海琦, 丁树良. 项目反应理论基础. 北京:北京师范大学出版社. 2012: 88–112

[20] 赵守盈, 石艳梅, 朱丹. 项目反应理论在大规模选拔性考试试题质量评价中的应用. 教育学报, 2013, 9(1): 71–77

[21] 方燕红, 张积家. 图词干扰范式下的语义效应. 心理科学进展, 2007, 15(5): 781–787

[22] Muñiz J. Review of “Handbook of modern item response theory”. European Journal of Psychological Assessment, 1998, 14(1): 91–93

[23] 钟轶, 季晓辉. 两种教育测量理论在试卷质量控制和评价中的应用及其展望. 南京医科大学学报: 社会科学版, 2013, 13(1): 66–71

[24] Chad W, Russell W, James C, et al. A comparison of angoff and bookmark standard setting methods. Journal of Educational Measurement, 2002, 39(2): 253–263

[25] 王晓华. Bookmark 法在基于标准的教育考试中设置划界分数的应用. 中国考试, 2014(7): 10–18

[26] 熊建华, 丁树良, 漆书青. 用测验信息量分析试卷质量. 江西师范大学学报: 自然科学版, 2002, 26(3): 225–228

[27] Cropley A J . Defining and measuring creativity: are creativity tests worth using?. Roeper Review, 2000, 23(2): 72–79

[28] 朱宁宁, 张厚粲. CTT 与 IRT 方法对人格测验结果处理的比较研究. 心理学探新, 2003, 23(3): 48–51

[29] Auzmendi E, Villa A, Abedi J. Reliability and validity of a newly constructed multiple-choice creativity instrument. Creativity Research Journal, 1996, 9(1): 89–95

[30] Sternberg R J. Implicit theories of intelligence, creativity, and wisdom. Journal of Personality and Social Psychology, 1985, 49(3): 607–627

[31] Grigorenko E L, Jarvin L, Sternberg R J. School–based tests of the triarchic theory of intelligence: three settings, three samples, three syllabi. Contemporary Educational Psychology, 2002, 27(2): 167–208

[32] Sternberg R J. What does it mean to be smart?. Educational Leadership, 1997, 54(6): 20–24