AI助力 心理测评与人才管理论坛

北京師範大學心理学部

发散思维测验人工评分与机器评分探索

段海燕

2016 级心理测量与人力资源管理专业硕士

现工作单位：中国工商银行远程银行中心

指导老师：徐建平 副教授

1 项目简介

本研究采用发散思维测验中常见的“续写故事结尾”题目，探讨和研究发散思维测验人工评分的改进和机器评分的开发。通过机器学习，通过哈工大的语言技术 (LIP) 平台进行分词，采用隐含狄利克雷分配模型 (LDA，Latent Dirichlet Allocation)、K-means 聚类分析以及四分模型等对训练集数据学习，形成评分模型，实现用机器评分对发散思维进行评分。

2 研究目的

对发散思维的计分方法进行创新，探索机器评分应用于发散思维评分的可行性。

3 研究过程

研究一：

对发散思维测验中被试答案的内容以及出现的频次进行编码与统计，开发新的计分方式，以对发散思维进行人工评分。

结果：

确定了答案编码所对应的新颖性分数以及流畅性和变通性的计分方式。

研究二：

对研究一的效果进行检验。

结果：

对于发散思维的三个维度，两次评分之间均显著相关。

研究三：

探索机器评分在发散思维测验的可能性。

结果：

实现了发散思维测验的机器评分程序的编写。

研究四：

对研究三形成的机器评分方式的评分效果进行检验。。

结果：

发散思维机器评分各维度得分及总分均与同感评估得出的创造力之间显著相关；人工评分与机器评分在发散思维的三个维度及总分也存在显著相关，聚合效度较好。仅机器评分新颖性维度预期显著正相关。本研究也根据大五人格的区分情况检验机器评分的区分效度。

4 研究结论

（1）对前人评分方式进行改进，开发出了针对“故事结尾续写”题目的新人工评分方式。

（2）通过对训练集数据进行机器学习，探索开发出发散思维测验中“故事结尾续写”题目的机器评分模型与系统，实现发散思维机器评分的初步尝试与探索。

（3）开发的发散思维测验新的人工评分方式，聚合效度、效标关联效度以及区分效度均较为良好，但是仍然存在较大的提升空间，而对于机器评分而言，其除了聚合效度较好之外，其他的效标关联效度以及区分效度相对较差，这与训练集数据量的大小有很大关系。但是其在评分速度和评分稳定性上的独特优势仍然非常显著，其在发散思维评分方面具有很大的可行性。

5 创新点

（1）聚焦于发散性思维评分方式上。

（2）主要将焦点放在“续写故事结尾”这种能够发挥更大的想象空间的题型。

（3）基于传统的发散思维评分理论与方法，对人工评分的方式进行了优化，开发了可以直接比对评分的评分方式；又探讨了机器评分对发散思维评分的可行性。

（4）结果迁移性较大，可以迁移到发散思维其他题型的评分中。

AI助力心理测评与人才管理论坛

北京師範大學心理学部

复杂问题解决能力的游戏化测量

樊真如
2017 级心理测量与人力资源管理专业硕士
现工作单位：北京世纪好未来教育科技有限公司
指导老师：黎坚 副教授

1 项目简介

本研究结合已有的问题解决模型，将复杂问题解决能力划分为规则分析、制定计划与执行以及监测反思这三个维度。同时结合复杂问题解决能力的特性，采用游戏测量的形式，依照循证设计范式中的概念评估框架，以迷你地铁游戏为原型开发出一款新型的测量工具。实验结果证明该测验具有足够的内部一致性信度（ α=0.676 ）。同时在汇聚效度方面，受测者在迷你地铁 CPS 版测验中的部分行为表现得分与其在基因实验室测验以及瑞文推理测验中的得分具有中度相关关系（ 0.457 ⩽ r ⩽ 0.624，p < 0.01 ）。本研究证明了使用其他游戏任务测量复杂问题解决能力的可能性，是一次面向未来的有益尝试。

2 研究目的

开发改编迷你地铁游戏作为复杂问题解决的测量工具以弥补现有工具的不足，探索复杂问题解决能力测验的新方向。

3 研究过程

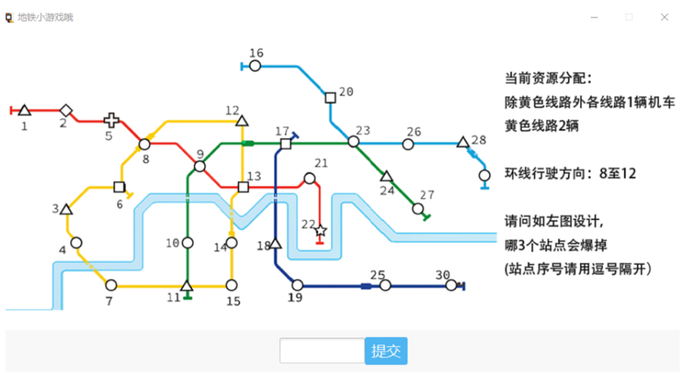
工具开发

（1） 模型确立

本研究通过文献综述提出了复杂问题解决能力的三维模型。规则分析维度要求人们通过多种方式收集并组织信息形成对问题空间的初步了解；制定计划与执行维度要求人们依据掌握的规则制定可操作的计划并使用各种工具与手段执行计划；监测反思维度要求人们不断地处理外部的反馈信息，评估各阶段的成果，探查分析超出预测的情况，并在适当时调整行动优化策略。

（2） 任务设计

本研究使用 JAVA 语言及 JAVAFX 框架，结合循证设计的设计范式，使用随机生成站点与乘客的线路建造类游戏为原型设计了共 4 阶段 11 回合的游戏测验。



研究设计

（1） 被试

本研究选取大学生被试 39 名，剔除指标数据有缺失的 6 名被试以及在第二阶段未覆 盖所有车站的不认真作答的 5 名被试后剩余有效被试 28 名

（2） 工具

迷你地铁(Mini metro): 用于测量复杂问题解决能力。游戏共 4 阶段 11 回合，测验时长 50 分钟左右，由计算机自动记录并生成玩家行为日志，累计记录 14 个指标。

瑞文标准推理测试（ Standard Progressive Matrices,SPM;Raven,1989 ）：用于测量一般智力中的抽象推理能力，共 60 题。

基因实验室（ Genetics Lab;Sonnelitner et al.,2012 ），以最小微观世界与线性方程模型为 框架的测量工具，用于测量复杂问题解决能力。该工具测验时长约为 35 分钟左右，共包含 15 道题目，其中 3 道为练习题。

（3） 实验流程

所有被试在主试监督下在实验室中完成三项测试并收集基本的人口学信息，全程 2 小时

（4） 特征提取

本测验实际记录指标 46 个，最终提取 14 个合成指标。其中规则分析维度 6 个，制定计划与执行维度 4 个，监测反思维度 4 个。

研究结果

表 1 列出了迷你地铁测验各维度的内部一致性信度，由此可知三个维度都具有较好的内部一致性信度。

表 1 内部一致性信度

	规则分析	制定计划与执行	监测反思	总分
科隆巴赫 α	.692	.696	.650	.676

表 2 列出了迷你地铁测验的效度检验结果。由表可知，迷你地铁测验中规则分析维度以及制定计划与执行维度与基因实验室测验、瑞文推理测验均有显著的相关关系，并且制定计划与执行维度与被试的年级排名也存 在显著相关关系。这一结果与前人研究相符（ Greiff,2012; Stadler, 2018 ）。监测反思维度与 所有外部效标之间不存在显著相关关系，表现较差。

表 2 迷你地铁各维度与效标的相关关系表

	规则分析（MM）	制定计划与执行	监测反思	规则分析（GL）	规则应用（GL）	瑞文推理
制定计划与执行	.500**					
监测反思	-.126	.036				
规则分析（GL）	.624**	.457*	.109			
规则应用（GL）	.479**	.560**	-.030	.399*		
瑞文推理	.589**	.398*	-.150	.538*	.688**	
年级排名	-.189	-.456*	.007	-.307	-.472**	-.332

4 研究结论

本研究旨在开发出一款可以有效测量个体复杂问题解决能力的游戏测验。通过对 28 名被试的实验数据进行分析，我们发现迷你地铁测验可以有效测量被试的规则分析能力以及制定计划与执行的能力。这一结果证明了本研究所使用的循证设计框架的有效性，有助于之后的研究者以更加开放的心态去探索有效测量复杂问题解决能力的方法。

5 创新点

本研究证明了使用其他游戏任务测量复杂问题解决能力的可能性，是一次面向未来的有益尝试。

AI助力 心理测评与人才管理论坛

北京師範大學心理学部

基于过程数据的合作问题解决能力测量

袁建林

2018 届博士生

现工作单位：湖南大学教育科学研究院助理教授

指导老师：刘红云 教授

1 项目简介

本研究以过程性测评理念为指导，采取两人交互的形式，设计开发交互性测验任务，获取学生在测验过程中的操作行为与对话并形成过程流数据，基于过程流数据抽取指标项目，开发自动化评分程序完成对学生过程流数据的评分，最后采用多维度 IRT 模型对两位合作者的能力进行估计。

2 研究目的

探讨基于计算机的过程性测评形式测量合作问题解决能力的可行性，为复杂能力的测评提供方法层面的参考。

3 研究过程

测验任务设计与开发:

采取“两人交互”的合作形式，利用“信息不对称”和“达成共同目标”的二元机制诱发学生合作完成共同任务，共开发 5 个任务，其中任务 1 如图 1：任务目标是给水杯装满 4 升的水，两学生利用对话框分享各自拥有的信息，讨论解决方案，协同执行操作完成任务。 过程中系统会以“关键事件”的形式记录学生所有操作行为和对话信息。

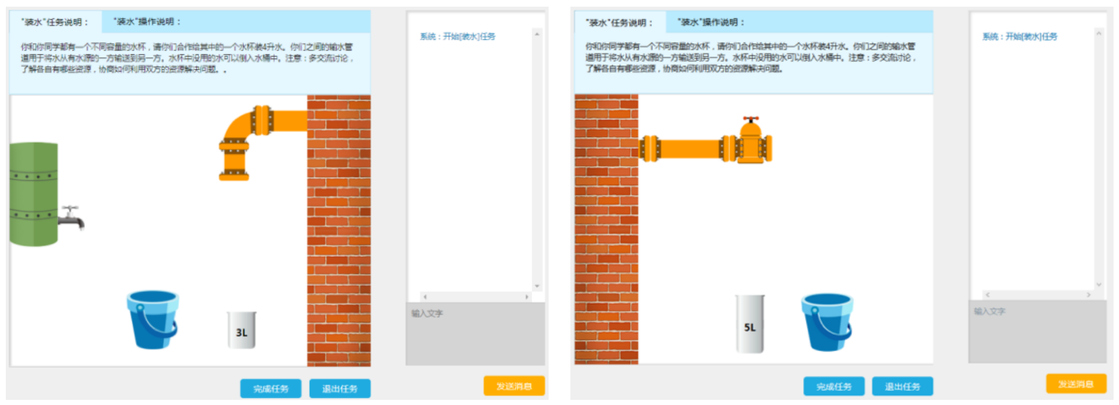


图 1

数据收集:

434 位学生参与测试，随机分配成 217 个小组，学生在测验过程中的操作行为和对话记录成行为过程流数据如图 2。此外，学生完成任务之后要求填写“同伴互评”问卷，要求对合作搭档的测验过程进行评价。

student_name	task_name	log_type	log_content	log_time	ROLE_TYPE
李张要	平衡木	task_start	task_start	2017-01-16 13:46:19 239	B
徐珏萌	平衡木	task_start	task_start	2017-01-16 13:46:19 491	A
李张要	平衡木	chat	我没砝码，你有不	2017-01-16 13:46:54 890	B
徐珏萌	平衡木	chat	我有4个	2017-01-16 13:47:15 811	A
李张要	平衡木	chat	都是啥么样的？	2017-01-16 13:47:40 863	B
徐珏萌	平衡木	chat	50 100 300 500g	2017-01-16 13:48:06 818	A
徐珏萌	平衡木	start_drag	mass:g100	2017-01-16 13:48:33 787	A
徐珏萌	平衡木	send	mass:g100	2017-01-16 13:48:37 115	A
李张要	平衡木	start_drag	mass:g100	2017-01-16 13:49:03 906	B
李张要	平衡木	hang	mass:g100postion:p3	2017-01-16 13:49:05 100	B
徐珏萌	平衡木	start_drag	mass:g300	2017-01-16 13:49:12 044	A
徐珏萌	平衡木	hang	mass:g300postion:p1	2017-01-16 13:49:14 103	A
徐珏萌	平衡木	balance	left:300 on p1,;right:100 on p3,;	2017-01-16 13:49:14 218	A
李张要	平衡木	start_drag	mass:g100	2017-01-16 13:49:32 460	B
李张要	平衡木	hang	mass:g100postion:p3	2017-01-16 13:49:35 537	B
李张要	平衡木	balance	left:300 on p1,;right:100 on p3,;	2017-01-16 13:49:35 731	B
徐珏萌	平衡木	chat	这样可以了吧	2017-01-16 13:49:38 177	A
李张要	平衡木	chat	可以	2017-01-16 13:49:44 714	B
李张要	平衡木	next	next:page2	2017-01-16 13:50:13 029	B
徐珏萌	平衡木	next	next:page2	2017-01-16 13:50:13 340	A
李张要	平衡木	left_mass_fill	g300	2017-01-16 13:50:20 127	B
李张要	平衡木	left_position_fill	p1	2017-01-16 13:50:21 698	B
李张要	平衡木	right_mass_fill	g100	2017-01-16 13:50:23 389	B
徐珏萌	平衡木	left_mass_fill	g100	2017-01-16 13:50:23 448	A
李张要	平衡木	right_position_fill	p3	2017-01-16 13:50:24 369	B
徐珏萌	平衡木	left_mass_empty	g100	2017-01-16 13:50:26 502	A
徐珏萌	平衡木	right_mass_fill	g100	2017-01-16 13:50:32 905	A
徐珏萌	平衡木	right_position_fill	p3	2017-01-16 13:50:34 266	A

图 2

指标抽取:

本研究以 ATC21S 界定的合作问题解决能力框架为基础，基于行为过程流数据，定义映射构念元素的具体行为模式或行为序列，以此作为学生能力的观测项目。针对“两人交互”的测验模式，提出“个人项目”和“小组项目”两种类型共存的项目设计思路，邀请评审专家对评审设计项目，算法编写程序对行为过程流的自动化评分，最后包含个人层面项目 44 个，小组层面项目 8 个。

建模分析:

采用项目内两维度 Rasch 模型对抽取的测验项目进行分析，其建模思路如图 3。学生 A 和学生 B 被视为模型的两个维度，对于学生 A，CA1、C A2…代表为学生 A 定义的共同项目，DA1、DA2…代表为学生 A 定义的独立项目；对于学生 B，同样会有其自身的共同项目 CB1、CB2…以及独立项目 DB1、DB2…；G1、G2…是小组项目，是两位学生共同作用的结果，所以同时附着于学生 A 和学生 B。同时，由于 CPS 包含合作和问题解决两个维度，我们应用项目内四维度 Rasch 模型对每个学生的合作能力和问题解决能力进行了估计。

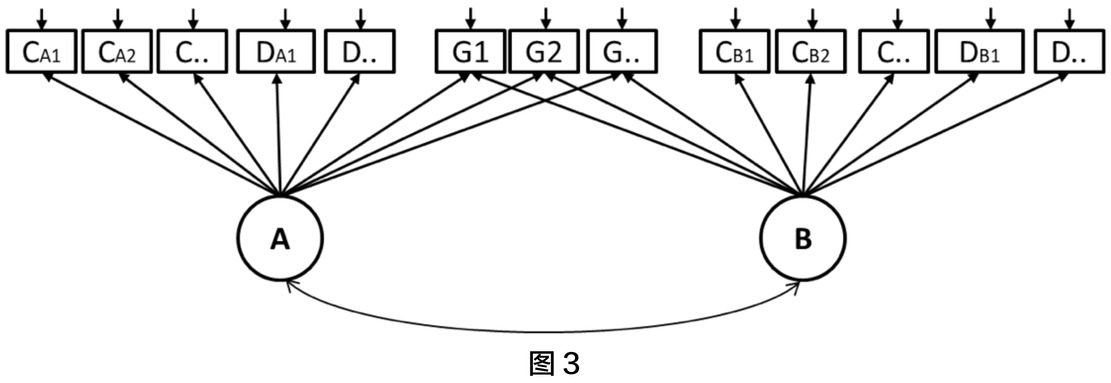


图 3

4 研究结论

总体项目分离度为 0.981，被试分离度分别为 0.886 和 0.891，两维度之间相关 0.561。所有项目的 MNSQ 的值处于 0.87~1.34 之间，平均值 1.0。基于 Conquest 输出的“怀特图“，符合教育测验中学生能力正态分布的一般假设。利用结构方程模型对测验估计的学生合作问题解决能力预测发现，同伴互评的合作能力对测验所得到的合作能力有显著预测作用，预测系数 0.209；同伴互评的问题解决能力对测验所得到的问题解决能力有显著预测作用，预测系数为 0.314。研究表明：基于过程数据测量复杂技能是一种可行的途径，“两人交互”的合作问题解决能力测验具有较高测量学质量。

5 创新点

1. 提出基于过程数据测量复杂技能的基本内涵和方法过程。
2. 提出复杂测验任务设计与开发的基本方法和过程。
3. 阐释了复杂过程数据中抽取测量证据的测量学原理。
4. 基于新的思路设计了合作问题解决能力测验项目和测量模型。

AI助力 心理测评与人才管理论坛

北京師範大學心理学部

基于游戏的大学生复杂问题解决能力测验开发

王帅

2017 级心理测量与人力资源管理专业硕士

现工作单位：科大讯飞

指导教师：骆方 副教授

1 项目简介

本研究开发了一套基于游戏的测评大学生复杂问题解决能力的系统，并尝试分析过程性数据。经检验其具有较好的测量学属性，结合过程性指标使测评更加标准化，同时在被试群体中有较高的接受度，能发挥出基于游戏的测评的优势，有效调动起学生解决问题的兴趣，帮助其发挥出最佳水平，并在一定程度上避免了学生与任务情境相关的先验知识、游戏经验和推理能力的干扰。

2 研究目的

本研究希望能够在已有系统的基础上，从复杂系统的设计、测评维度和问题情境等角度适当创新，开发出一套能有效测量大学生复杂问题解决能力的复杂系统，并利用学生在复杂情境中操作的过程性数据，挖掘学生的行为模式和问题解决策略等信息，以更全面客观地评价其复杂问题解决能力，帮助家长和老师完善对学生的教育和教学。

3 研究过程

被试： 37 名北京地区大学生

研究材料：

1）基因实验室：一套由卢森堡大学开发、已经过测量学指标验证的基于计算机测评学生复杂问题解决能力的复杂系统。

2）小怪兽诊所：本研究开发的基于游戏的测评大学生复杂问题解决能力的系统。

3）标准版瑞文推理测验

4）目标导向与学习策略问卷

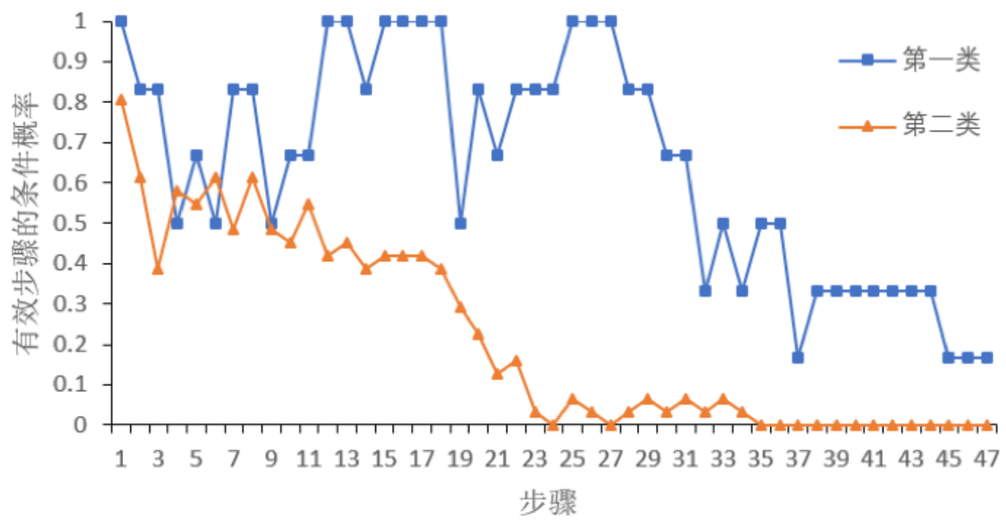
5）个人信息及任务体验问卷

实测程序：

采取平衡设计的方法确定测试顺序

研究结果：

1）本研究开发的测评系统具有较好的接受度，不受无关因素影响，同时具备较好的效度和区分度，可以用于测量大学生复杂问题解决能力。



2）经对过程性数据的分析发现，学生经过前期的摸索，能否尽快适应系统的变化模式并找到合适的策略，将对整个探索过程起着至关重要的作用。

3）经对过程性数据的分析发现，在学生与复杂系统不断互动的过程中，得到的信息越来越多，使用有效策略的难度也越来越大。



4）经对过程性数据的分析发现，整个探索过程中，步骤对预测学生复杂问题解决结果的重要程度呈下降趋势。

4 研究结论

1）本测验具备较好的测量学属性，可以有效评价大学生的复杂问题解决能力。

2）本研究尝试以学生操作过程中的最优性指标和关键步骤等作为评价指标，评价学生在具体维度中的表现，在一定程度上证明了过程性数据对测评学生复杂问题解决能力的意义。

3）学生在探索过程中前二分之一的操作对预测学生能否有效解决复杂问题至关重要，而后二分之一的操作对识别出高能力学生有重要意义。

5 创新点

从测评维度、问题情境和复杂系统构建的角度创新

AI助力心理测评与人才管理论坛

北京師範大學心理学部

利用大数据预测社会化问答平台中回答采纳行为

刘宇宁

2016 级心理测量与人力资源管理专业硕士

现工作单位：北京正略钧策咨询股份有限公司

指导老师：黎坚 副教授

1 项目简介

互联网及搜索引擎的诞生使我们获得了获取丰富信息的途径，但同时伴随着信息的泛滥化和碎片化的问题，使人们产生了高效获取有针对性和个性化信息的需求，因此，互联网问答服务平台应运而生，随着自媒体和网络媒体的发展，问答平台还逐渐融合了社交属性从而形成了社会化问答平台。然而在内容体量巨大，鱼龙混杂的回答信息中，回答被用户所接受和采纳的程度也千差万别。探究哪些因素会影响用户对回答的采纳行为，对回答者有针对性地提升影响力有着重要影响，同时也能够优化回答质量的评估模型。本研究运用大数据方法，以国内应用最广泛的社会化问答网站——知乎为研究对象，结合双路径理论和信息采纳模型，获取并形成 2011 年 -2018 年的回答内容对应回答文本特征、回答形式特征、回答背景特征以及用户特征等，采用机器学习的方法构建了问答社区回答采纳行为预测模型，并进一步分析了各类特征的重要性。

2 研究目的

对知乎中的高赞回答，通过自下而上的数据分析结合自上而下的理论，从数据中提取可量化的影响采纳行为的特征，通过实证研究，建立不同话题下采纳行为的预测模型。

3 研究过程

研究样本

（1）研究话题

本研究选择了知乎上受到较为广泛关注的三个话题：“儿童教育”；“医疗”话题；“互联网金融”。

（2）研究问题

选择的问题为三个所选知乎话题页面下精华回答栏目列表中所涉及到的题目。使用 python 进行相应话题下精华回答中所包含的问题信息的抓取。

数据预处理和特征获取

本研究机器学习建模的预测指标用回答点赞数量进行量化，所获取的样本中包含了大约三分之一点赞数量为 0 的样本。为减少无关样本，提升预测的准确性，本研究在分析中去除了预测指标（回答点赞数量）为 0 的样本，此外还去除了包含缺失值（由于匿名回答缺乏回答者特征）的样本，最终分析样本量为 13029 个。

清理数据后，使用 python 及 TextMind 软件获取各项特征指标。使用的特征包括：（1）回答文本特征（2）回答形式特征（3）回答背景特征（4）其他特征，如回答评论数量；（5）回答者特征

模型训练

本研究使用了 Linear Regression、Lasso Regression、Ridge Regression 等多种算法，建立对回答采纳行为的预测模型。模型训练过程中使用了交叉验证的方法，随机划出 20% 的样本作为评估集，80% 的样本用于训练模型。

研究结果

最佳分类模型和模型评价指标见表 1。图 1 中较为直观的体现了各个特征在总体以及各个分话题模型间的稳定性，在各个模型中某特征出现的次数越多，表示其在各个话题上对采纳行为具有越为普遍的预测能力，也就是越为稳定。

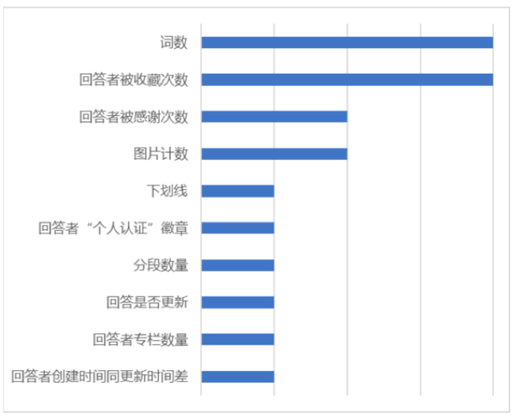


图 1 各类特征在话题间的预测稳定性

	最佳模型	样本量	r	R2	RMSE	特征数量
总体	RidgeCV Regression	13029	0.39	0.15	0.64	116
儿童教育话题	RidgeCV Regression	6294	0.32	0.10	0.76	116
医疗话题	Linear Regression	3839	0.40	0.16	0.70	122
互联网金融	RidgeCV Regression	2896	0.30	0.09	0.24	116

表 1 机器学习模型预测结果

4 研究结论

（1）问答平台上的采纳行为受到文本内容特征、形式特征、回答者特征以及背景特征的影响；

（2）区分话题能够提高采纳行为预测的准确性，不同话题采纳行为预测模型有共性也有差异；

（3）不同话题预测难度有差异；不同特征对于各个有些特征类型个话题采纳行为的预测稳定性有差别；

（4）词数是最有力的采纳行为的指示特征，其次是回答者的收藏数量，而有些特征只对于特定话题的采纳行为有预测能力。

5 创新点

本研究使用了大数据的研究方法，通过获取社会化问答平台中的大量行为数据，对采纳行为进行分析，在选题和研究方法上就有一定的时代特色和创新性，补充了相关领域的研究并且较为系统地梳理了前人的相关研究，建立了较为完整的研究框架。并且得出的总体及各个分话题的采纳行为预测模型的预测指标较好，能够以较为精简的特征进行预测，同时模型具有较低的标准误。

AI助力心理测评与人才管理论坛

北京師範大學心理学部

让自适应测验更知人善选
——基于推荐系统的选题策略

王璞珏
2017 级学术硕士
指导老师：刘红云 教授

1 项目简介

研究提出全新基于协同过滤推荐的 CAT 选题策略，利用已有答题者数据，提出两种可利用已有答题者数据的 CAT 选题策略：直接基于答题者推荐（DEBR）和间接基于答题者推荐（IEBR）。通过蒙特卡洛模拟研究，考察 DEBR 在选题精度和对题目曝光率控制方面的表现。

2 研究目的

改进 CAT 选题策略，提出全新选题策略，构建更智慧的选题推荐系统。

3 研究过程

研究 1:

研究问题：

探究传统选题策略和不同长度测验的答题者信息对两种推荐选题策略表现的影响。

研究材料：

400 道 0-1 计分的题目，基于 3 参数 Logistic 模型（3PLM）。

模拟流程：使用传统选题策略、推荐选题策略分别对 1000 名答题者进行 CAT 模拟。贝叶斯后验期望法进行能力估计。BAS 策略中，题库为 4 层。

评价指标：

均方误差，平均绝对误差，真实能力值与最终能力估计值的相关，题目实际曝光率与理想分布的卡方值，测验重叠率，曝光不足，曝光过度，答题者调用率。

研究结果：

不同特点的已有答题者数据会影响推荐选题策略表现的趋势。测验长度不影响新策略特定数据下的表现，但会影响各指标上的绝对大小，包括更高的精度和答题者调用率等。

研究 2:

研究问题：

研究现实情境下推荐选题策略的表现。

研究材料：

科学测验 276 题，125 道基于 2PLM，151 道基于 3PLM

模拟流程：

使用传统选题策略、推荐选题策略分别对 1000 名答题者进行 CAT 模拟，将共 2000 名答题者数据合并为已有答题者数据，再次使用推荐选题策略对第三批 1000 名答题者 CAT 模拟。使用 BAS 策略时，题库 4 层。

研究结果：

真实题库不影响两种推荐选题策略的选题特点和性质，且 DEBR 和 IEBR 的优势表现得更明显，特点更鲜明。

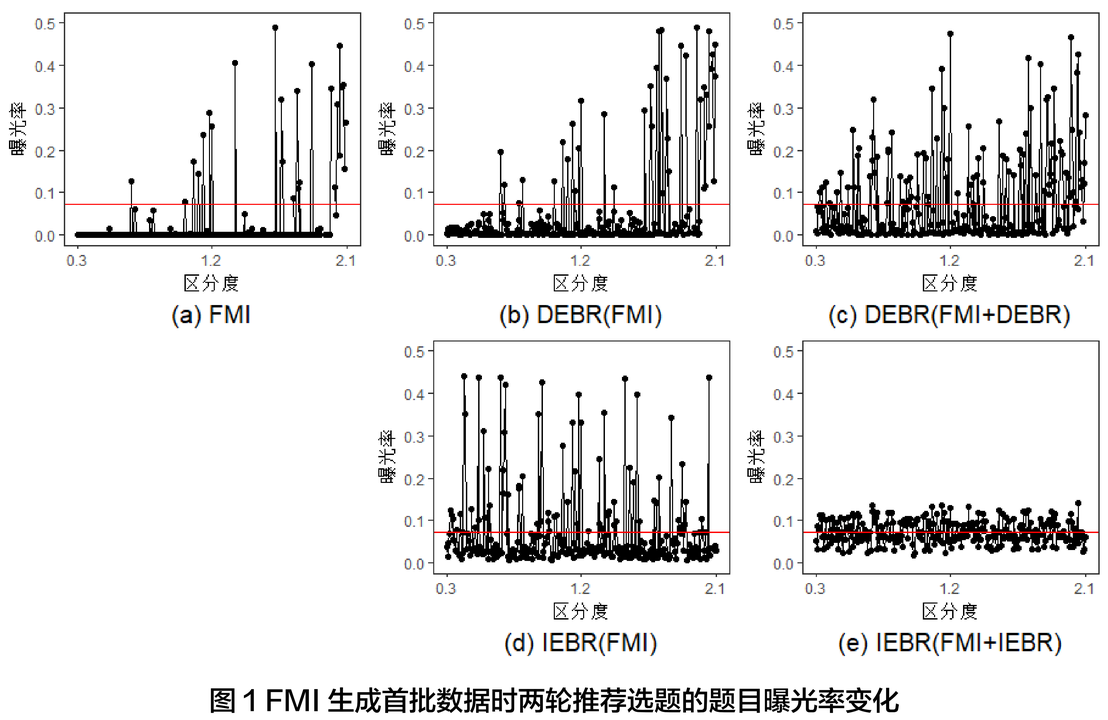


图 1 FMI 生成首批数据时两轮推荐选题的题目曝光率变化

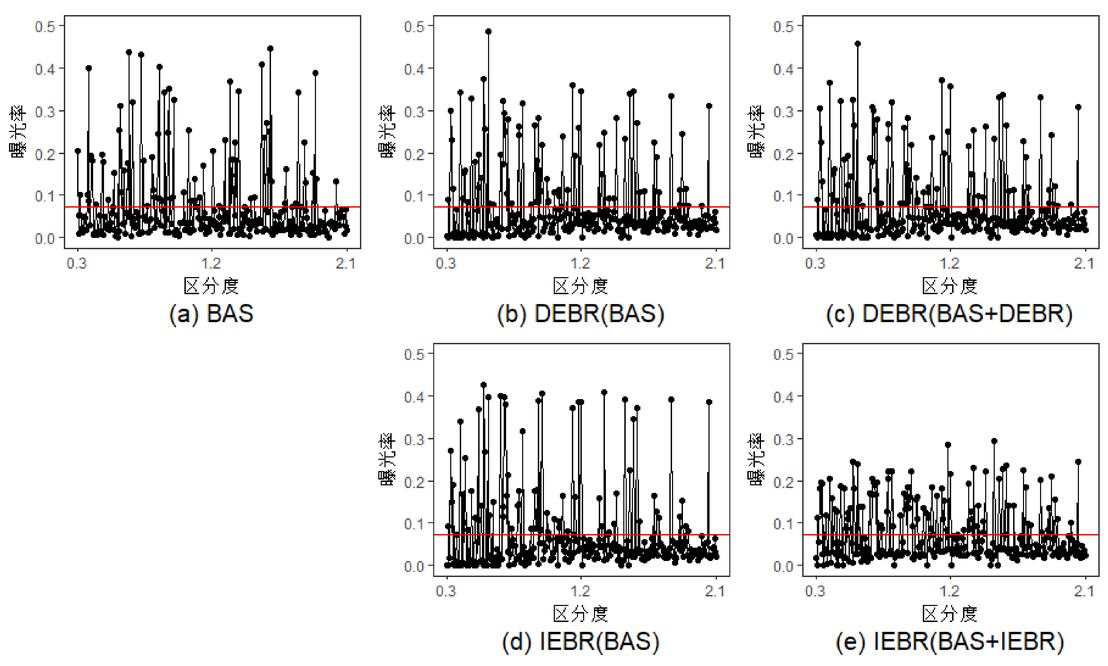


图 2 BAS 生成首批数据时两轮推荐选题的题目曝光率变化

4 研究结论

（1）推荐系统中的协同过滤推荐可移植于 CAT 选题，设计出的推荐选题策略保证测量精度且对题目曝光率的控制更好；

（2）已有答题者数据特点、质量是影响推荐选题策略表现的主要因素。

5 创新点

1. 发现已有答题者数据对选题的价值。

2. 发现推荐系统和 CAT 选题的共通性，借鉴协同过滤推荐技术建立选题规则，证明协同过滤推荐的底层假设同样适用 CAT 选题场景。

AI助力 心理测评与人才管理论坛

北京師範大學心理学部

使用改编的游戏评估个体的元认知调节能力

林琪

2017 级心理测量与人力资源管理专业硕士

现工作单位：小米通讯技术有限公司

指导老师：黎坚 副教授

1 项目简介

自元认知的概念首次被提出至今，研究者们对于元认知的概念与成分已经形成了基本共识。同时，有关元认知研究的重点也从对概念理论的梳理逐渐转变为对应用领域的探索。在这些研究中，如何测量个体的元认知能力水平是非常重要的一环。对此，国内外的研究者们进行了诸多探索，目前在测量个体元认知能力时使用的方法主要为问卷法、访谈法、出声思维法和认知操作法。但传统的测量方法存在的部分问题影响了评估结果的准确性，而基于游戏的评估可以在一定程度上改善此现状。本研究基于改编的魔塔游戏，挖掘游戏行为数据，探索了魔塔游戏在评估个体元认知调节能力上的可行性。

2 研究目的

改编魔塔游戏作为测量工具，结合游戏行为数据，探索魔塔游戏对评估个体元认知调节能力的有效性。

3 研究过程

研究参与者：

国内本科生及研究生，一共 49 人。剔除无效作答后，有效被试一共 37 人。其中筛选被试的方法为：（1）在某一维度上数据缺失，删除受测者在该维度上的数据；（2）缺失游戏数据超过 70% 以上，删除受测者的全部数据；（3）测试过程中出现明显不认真行为，删除受测者的全部数据；（4）不能正确回答所设定的质量监控题，删除受测者的全部数据。

研究材料：

1. 魔塔游戏

原版的《魔塔》是一个固定数值的益智类 RPG 游戏，最早出现于日本，后由网友将其汉化后引入中国。本文部分借鉴了《21 层魔塔》游戏的主体运行脚本、图片素材和原型，使用 C# 语言改编出了适用于本研究的魔塔游戏。游戏界面如图 1 所示。

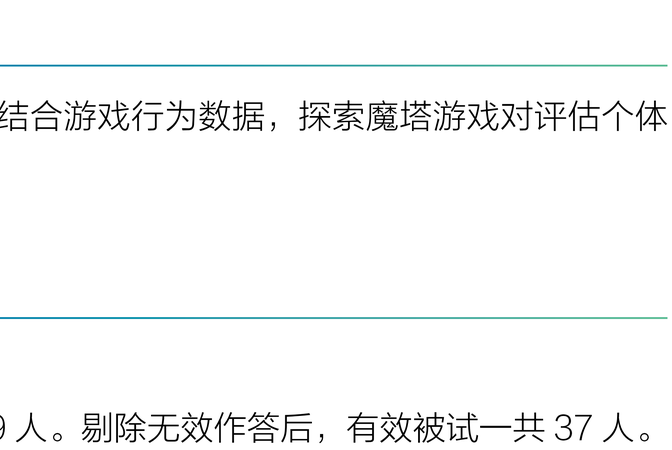


图 1 魔塔游戏界面

2. 元认知量表

元认知意识量表（Metacognitive Assessment Inventory, MAI）和状态元认知量表（State Metacognition Inventory）用于评估个体的元认知调节能力。

研究过程：

所有受测者均在实验室中参加测试，受测者需先完成游戏（耗时约 120 分钟），再填写元认知问卷（耗时约 15 分钟）。整个过程由主试监督，以保证被试可以在安静的环境中独立作答。

研究结果：

魔塔与 MAI 量表、状态元认知量表的相关系数见表 1。

表 1 魔塔与 MAI 量表、状态元认知量表的相关系数表

	预测	计划	策略选择	监测	修正	评价
陈述性知识	.234	.377*	.421*	.351*	.118	.398*
程序性知识	.128	.326	.302	.239	.348	.241
条件性知识	.103	.063	.261	.216	.246	.043
计划	.574**	.359*	.428*	.399*	.061	.282
信息管理策略	.164	.172	.297	.332	.227	.037
理解监测	.340*	.256	.187	.289	.253	.082
修正策略	.248	.352*	.480**	.478	.191	.172
评价	-.075	.200	.244	.151	-.058	.177
MAI 总分	.341*	.389*	.464**	.448**	.222	.279
元认知意识	.065	.456**	.542**	.210	.311	.452*
认知策略	.041	.417*	.577**	.402*	.411	.418*
计划	.144	.305	.483**	.178	.172	.384*
自我检查	.026	.346*	.392*	.261	.392	.177
状态元认知总分	.077	.430**	.564**	.299	.373	.407*

4 研究结论

（1）利用魔塔游戏的游戏行为数据在区分不同个体在计划和策略选择过程上的差异时具有较好的结果；

（2）魔塔游戏在区分不同个体在预测、监测和评价过程上的差异时结果一般；

（3）魔塔对个体修正过程的差异的区分效果较差；

（4）魔塔在计划和策略选择两个过程上的区分度较差。

5 创新点

本研究采用魔塔游戏作为能力测量工具，实现了更加隐蔽、客观、准确的测量，有助于解决传统心理测验存在的不足。本研究的指标建构较前人更加丰富和复杂，有助于深入挖掘游戏过程性数据，提高预测的准确性。

AI助力 心理测评与人才管理论坛

北京師範大學心理学部

使用基于游戏的评估和计算机日志数据预测学生的能力组别

李昌锦
2017 级学术硕士
指导老师：黎坚 副教授

1 项目简介

本研究基于推箱子游戏，充分挖掘计算机日志数据，探索了游戏化测量在评估学生的整体认知能力上的有效性，并使用机器学习算法建立了分类模型。结果表明在区分学生能力组别的任务上，机器学习模型的准确率在 0.762 到 0.900 之间，达到了较好的分类效果。

2 研究目的

利用推箱子游戏作为测量工具，并深入挖掘计算机日志信息，建立更加完善的指标体系，探索推箱子游戏作为能力的测量工具的有效性。

3 研究过程

研究被试:

超常儿童来自北京市某重点中学的超常儿童实验班，共 174 人；
普通儿童和数学学习障碍儿童来自北京市某普通中学，根据一定标准选取数学学习障碍儿童 56 人，普通儿童 171 人。

研究材料:

1. 推箱子游戏

推箱子游戏 (Sokoban) 是一款经典益智游戏，玩家需要操纵小人上下左右移动，将箱子推入目标位置。经过研究者对程序的改编，计算机后台可以自动生成计算机日志文件，记录受测者每一步的时间与动作。游戏界面如图 1 所示。

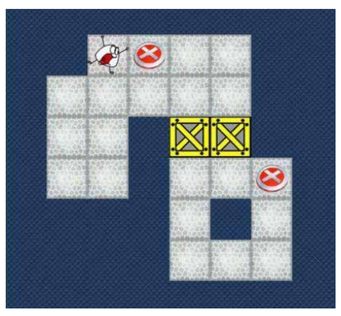


图 1 推箱子游戏界面

2. 其他能力测试

瑞文标准推理测验 (Standard Progressive Matrices, SPM) 用于测量一般智力中的抽象推理能力。

3. 研究者获得了学生的三次数学和语文测验的成绩 (期中、月考、期末)，将三次数学和语文成绩分别取平均值作为学生的数学和语文学业表现的指标。

研究流程:

数据预处理和测量指标选取

从游戏日志数据中收集到的原始过程数据包括玩家每一步进行的操作 (目标、起点和终点) 以及每一步操作结束到下一步开始操作之间的时间，计算机记录的受测者顺利通过某一关的典型数据形式如图 2 所示。

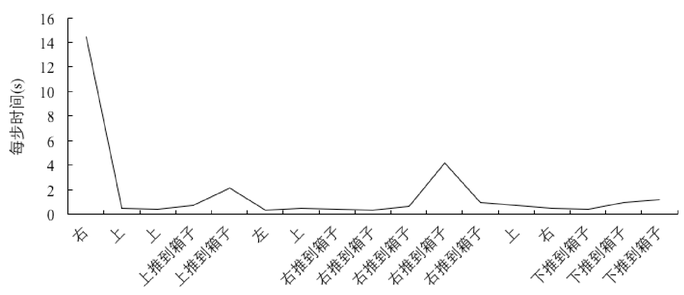


图 2

模型训练

为了区分超常儿童 (Sample A)、数学学习障碍儿童 (Sample B) 与普通儿童 (Sample C)，分别在 A 与 B、B 与 C、A 与 C 三组样本上，使用随机森林 (Random Forests, RF)、梯度提升决策树 (Gradient Boosting Decision Tree, GBDT)、支持向量机 (Support Vector Machine, SVM) 和逻辑回归 (Logistic Regression, LR) 四种机器学习算法建立分类模型。

研究结果:

最佳分类模型和模型评价指标见表 1，基于树模型计算的特征重要性见图 3。

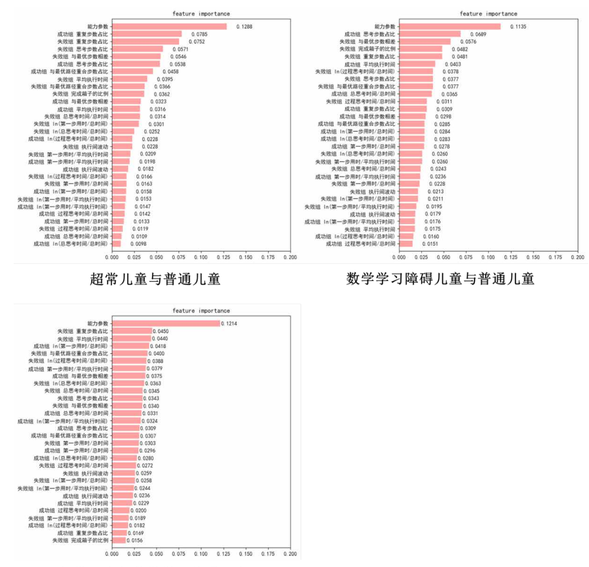


图 3 基于树模型的特征重要性

表 1 机器学习模型预测结果

分类任务	最佳模型	精确率	召回率	F1
超常儿童 – 普通儿童	GBDT	0.772	0.762	0.757
数学学习障碍儿童 – 普通儿童	SVM	0.732	0.783	0.752
超常儿童 – 数学学习障碍儿童	RF 和 LR	0.913	0.900	0.893
状态元认知总分	.077	.430**	.564**	.299

4 研究结论

（1）利用推箱子游戏的日志数据和机器学习分类模型能较好地预测个体的能力组别；
（2）三个分类任务中，IRT 能力参数都是最重要的特征，这表明结果性指标对于预测学生的能力非常重要；
（3）对于不同样本的分类任务，过程性指标的重要性排序不尽相同，说明不同能力组别学生之间可能存在不同方面的差异。

5 创新点

本研究采用推箱子作为能力测量工具，实现了更加隐蔽、客观、准确的测量，有助于解决传统心理测验存在的不足。本研究的指标建构较前人更加丰富和复杂，有助于深入挖掘游戏过程性数据，提高预测的准确性。

AI助力 心理测评与人才管理论坛

北京師範大學心理学部

天气对中国足球超级联赛运动员绩效表现的影响

杨上琦

2016 级心理测量与人力资源管理专业硕士

现工作单位：招商银行股份有限公司北京分行

指导老师：黎坚 副教授

1 项目简介

本研究利用网络爬虫技术去获取中国足球超级联赛近 3 个赛季的所有上场球员的进攻、防守等详细技术统计和这 3 年来比赛时的天气数据；从关系绩效和任务绩效的角度出发，将收集到的数据进行划分并利用机器学习的方法建立天气与球员工作绩效之间的预测模型。

2 研究目的

本研究主要的目的不仅仅为了研究极端天气和空气质量对暴露在空气中的足球运动员工作绩效的影响，还想通过大数据进行机器学习建模，从而生成一个预测模型。这样不仅能够帮助球队根据天气进行技战术调整，还能够帮助球员更科学地了解空气污染和极端天气的危害，并做出防范措施。

3 研究过程

研究问题：

探究恶劣气候和空气污染对球员的比赛的影响，即足球运动员的任务绩效和关系绩效的影响，并利用机器学习的方法去建立天气与球员工作绩效之间的预测模型。

研究材料：

中国足球超级联赛 2015 赛季至 2017 赛季 3 年内所有上场的 438 名运动员，合计 10153 个数据点。其中有两场比赛的数据大面积缺失，最终剩余球员 438 名，合计数据点 10122 个。

特征提取：

研究变量：足球运动员相关变量、气候变量、空气质量变量以及其他变量。

气候数据：包括日均温度、日均湿度、日均气压、日均风力、8-20 时降水量、蒸发量、日照时间、地表气温。空气质量主要利用空气质量指数（AQI）和污染物 CO、NO2、O3、PM10、PM2.5 以及 SO2 的浓度。

其他：球员的年龄、身高、体重以及球赛开始的时间点，足球比赛是星期一至七的哪一天。

模型建立：

自变量和因变量的选择以前锋的任务绩效（进攻变量）为例，如表 1 中所示：

表 1 前锋任务绩效模型的因变量与自变量

因变量	自变量	
以进攻为目的的传球次数、射门次数、助攻次数、关键传球次数、进球数之和。	气候指标	日均温度、日均气压、日均风力、8-20 时降水量、日照时间、地表气温
	空气质量指标	空气质量指数（AQI）、空气质量评级、PM2.5 浓度、PM10 浓度、SO_2 浓度、CO 浓度、NO_2 浓度、O_3 浓度
	球员特质变量	球员身高、球员体重、上赛季平均表现

在建立模型时，使用岭回归、Lasso 回归以及径向基函数核函数（Radial Basis Function, SVR-RBF）和神经网络核函数（Sigmoid kernel, SVR-Sigmoid）的支持向量回归（SupportVector Regression, SVR）的机器学习算法分别进行建模。

研究结果：

1. 前锋球员任务绩效模型中，天气对其任务绩效表现的解释率较高（R2 = 0.425，λ = 0.01），且球员特质的解释率偏低（R2 = 0.0009）；前锋球员关系绩效模型中，天气对其关系绩效的解释率较高（R2 = 0.241，λ = 0.01），且球员特质的解释率偏低（R2 = -0.005）。这可以说明天气能够比较好地预测前锋球员的任务绩效和关系绩效。

2. 后卫球员任务绩效模型中，天气对其任务绩效表现的解释率较高（R2 = 0.207，λ = 0.001），并且球员特质的解释率偏低（R2 = -0.001）；后卫球员关系绩效模型中，天气对其关系绩效的解释率较低（R2 = 0.054，λ = 0.01），球员特质的解释率也偏低（R2 = 0.022）。这可以说明天气能够预测后卫球员的任务绩效，一定程度上可以预测关系绩效，但是解释率不高。

表 2 前锋球员任务绩效模型拟合结果

模型		Lasso			岭回归			SVR 径向基核函数			SVR 神经网络核函数		
		MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2
特质		1.001	1.001	0.0009	1.0001	1.0007	0.0009	0.871	0.933	0.130	1.003	1.002	-0.0028
天气		0.576	0.759	0.425	0.579	0.761	0.422	0.574	0.758	0.426	0.579	0.761	0.421
两者		0.575	0.758	0.425	0.580	0.761	0.421	0.577	0.759	0.424	0.579	0.761	0.422
		特质	天气	两者	特质	天气	两者	特质	天气	两者	特质	天气	两者
超参数	λ	0.0001	0.01	0.01	0.1	10	10						
	C							1000	1000	100	1	1000	1000
	Gama							10	0.0001	0.001	0.1	0.0001	0.0001

其中 N=3831；λ ∈ $[10^{-10}, 10^4]$；C ∈ $[10^{-5}, 10^3]$；gama ∈ $[10^{-5}, 10^3]$。

为了提高搜索效率，超参数取值按照 10 的对数取值，因为算法本身的复杂性，本次建模只选择了 7 个数字进行搜索。

表 3 前锋球员关系绩效模型拟合结果

模型		Lasso			岭回归			SVR 径向基核函数			SVR 神经网络核函数		
		MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2
特质		1.038	1.019	-0.005	1.038	1.019	-0.005	1.067	1.033	-0.033	1.076	1.037	-0.043
天气		0.783	0.885	0.242	0.786	0.887	0.238	0.809	0.899	0.216	0.810	0.900	0.215
两者		0.783	0.885	0.241	0.787	0.887	0.237	0.806	0.898	0.219	0.810	0.900	0.216
		特质	天气	两者	特质	天气	两者	特质	天气	两者	特质	天气	两者
超参数	λ	0.0001	0.01	0.01	0.1	100	100						
	C							1000	100	1	100	10	10
	Gama							10	0.0001	0.01	0.01	0.001	0.001

其中 N=3831；λ ∈ $[10^{-10}, 10^4]$；C ∈ $[10^{-5}, 10^3]$；gama ∈ $[10^{-5}, 10^3]$。

为了提高搜索效率，超参数取值按照 10 的对数取值，因为算法本身的复杂性，本次建模只选择了 7 个数字进行搜索。

表 4 后卫球员任务绩效模型拟合结果

模型		Lasso			岭回归			SVR 径向基核函数			SVR 神经网络核函数		
		MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2
特质		0.990	0.995	-0.001	0.989	0.995	-0.0008	0.966	0.983	0.023	0.998	0.999	-0.009
天气		0.785	0.886	0.207	0.786	0.887	0.205	0.796	0.892	0.194	0.797	0.893	0.194
两者		0.784	0.885	0.207	0.785	0.886	0.206	0.795	0.892	0.196	0.796	0.893	0.194
		特质	天气	两者	特质	天气	两者	特质	天气	两者	特质	天气	两者
超参数	λ	0.001	0.001	0.001	1000	100	100						
	C							1000	10	10	10	100	100
	Gama							10	0.001	0.001	1	0.001	0.001

其中 N=6305；λ ∈ $[10^{-10}, 10^4]$；C ∈ $[10^{-5}, 10^3]$；gama ∈ $[10^{-5}, 10^3]$。

为了提高搜索效率，超参数取值按照 10 的对数取值，因为算法本身的复杂性，本次建模只选择了 7 个数字进行搜索。

表 5 后卫球员关系绩效模型拟合结果

模型		Lasso			岭回归			SVR 径向基核函数			SVR 神经网络核函数		
		MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2	MSE	RMSE	R^2
特质		0.954	0.977	0.022	0.954	0.977	0.022	0.858	0.926	0.121	1.049	1.024	-0.074
天气		0.947	0.973	0.033	0.951	0.975	0.026	1.040	1.019	-0.066	1.045	1.022	-0.071
两者		0.923	0.961	0.054	0.932	0.965	0.045	1.022	1.011	-0.046	1.031	1.016	-0.056
		特质	天气	两者	特质	天气	两者	特质	天气	两者	特质	天气	两者
超参数	λ	1e-05	0.01	0.01	0.1	100	10						
	C							1000	0.1	1	100	100	1000
	Gama							10	0.01	0.01	0.01	1	0.001

其中 N=6305；λ ∈ $[10^{-10}, 10^4]$；C ∈ $[10^{-5}, 10^3]$；gama ∈ $[10^{-5}, 10^3]$。

为了提高搜索效率，超参数取值按照 10 的对数取值，因为算法本身的复杂性，本次建模只选择了 7 个数字进行搜索。

4 研究结论

（1）气候指标（温度、降雨量、风力等）处于极端情况时，天气能够预测球员的关系绩效和任务绩效。

（2）空气质量指标（PM2.5、PM10、CO 等），能够较好地预测两类球员的任务绩效和前锋的关系绩效，却不能预测后卫的关系绩效。

5 创新点

本研究丰富了天气对工作绩效领域的研究，并且为这些领域提供了一个新型的研究逻辑、研究对象和研究方法。

AI助力心理测评与人才管理论坛

北京師範大學心理学部

现场和微信无领导小组讨论中评价顾忌、社会临场感对表现真实性的影响研究

吴兆霖

2015 级心理测量与人力资源管理专业硕士生

现工作单位：大悦城控股集团股份有限公司

指导老师：徐建平 副教授

1 项目简介

本研究提出借助社交软件 – 微信组织无领导小组讨论的人才选拔方法，并且通过微信、现场和考官在场与否的实验设计，探究了这种方法对于应聘者的评价顾忌、社会临场感和表现真实性的影响。

2 研究目的

探究在微信视频组织的无领导小组讨论中，应聘者的评价顾忌、社会临场感和表现真实性的情况。

3 研究过程

研究问题：

探究应聘者在微信视频讨论且考官可见、微信视频讨论且考官不可见、现场讨论且考官可见和现场讨论且考官不可见四种情境下社会临场感、评价顾忌和表现真实性的情况。

研究材料：

特殊处理的座签 24 张。考官手册 3 份，评分表 12 张，记录表 12 张，装有题目的信封 6 份，空白 A4 纸 36 张，黑色签字笔 9 支。装有屏幕视频录像软件 Bandicamp 的笔记本电脑一台。能够与上述笔记本电脑同时登陆同一个微信账号的手机一个。电子版考场规则 1 份。分发给 6 位参与讨论的应聘者的讨论题目的电子版 6 份。

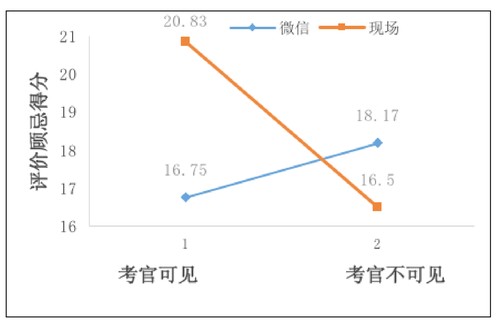
无领导小组讨论的题目由本人自编完成。无领导小组中资源争夺式的题目往往能够使得参与讨论的被试展开激烈的讨论，从而激发更多与测评维度相关的行为，因此本研究设计了资源争夺式的题目。

实验流程：

第一步，培训考官，筛选达到测试标准的考官；第二步，为参与现场面对面讨论的被试安排统一时间和统一讨论场地进行现场讨论，为参与微信视频讨论的被试安排统一时间进行连线讨论；第三步，组织正式讨论；第四步，要求被试填写评价顾忌问卷和社会临场感问卷；第五步，要求考官、被试的舍友和被试的班长对被试进行打分。第六步，分析收集的数据。

评价指标：

非参数方差，打分差值的绝对值，社会临场感得分。



研究结果：

结果一：在考官可见时，应聘者在现场情境的评价顾忌水平显著高于在微信情境的评价顾忌水平；在现场时，应聘者在考官可见情境的评价顾忌水平显著高于应聘者在考官不可见情境的评价顾忌水平；

结果二：应聘者意识临场感维度得分在现场情境和微信情境之间的差异达到显著；

表 1 应聘者在不同情境下的无领导小组讨论中的社会临场感的得分情况 (N=24)

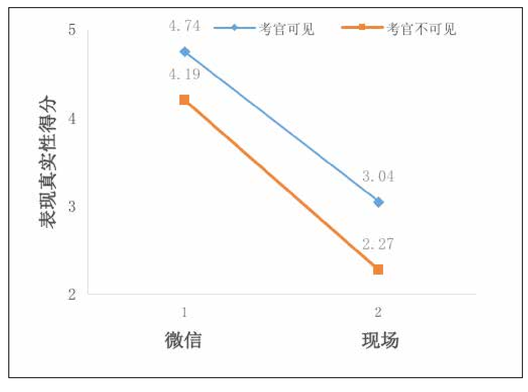
	讨论媒介	平均数	标准差
社会临场感总分	微信	35.67	6.12
	现场	36.50	4.44

表 2 应聘者在不同情境下的无领导小组讨论中的意识临场感的得分情况 (N=24)

	微信		现场	
	平均数	标准差	平均数	标准差
1. 我能够感知到其他个体的存在。	4.33	0.917	4.67	0.565
2. 我会密切关注其他个体的存在。	4.04	0.999	4.58	0.584

对应聘者在两种情境下的意识临场感、认知临场感和情绪临场感分别进行独立样本 t 检验，所得结果表明，只有在意识临场感上的差异显著 ($p \leqslant 0.05$)，在情感临场感和认知临场感上的差异均不显著。

结果三：应聘者在现场情境的表现真实性显著高于应聘者在微信情境的表现真实性。



4 研究结论

（1）考官可见与否会影响应聘者在无领导小组讨论时的评价顾忌。在现场情境时，应聘者在考官可见情境下的评价顾忌水平要高于在考官不可见情境下的评价顾忌水平；在微信情境时，应聘者在考官不可见情境下的评价顾忌水平要高于在考官可见情境下的评价顾忌水平；（2）应聘者在微信视频讨论与现场讨论中的的意识社会临场感维度有显著性差异；

（3）应聘者在微信视频情境下的表现真实性显著低于在现场情境下的表现真实性。

5 创新点

第一，发现了在小组工作中，旁观者可见与否会影响个体的评价顾忌水平；

第二，发现了在不同媒介中进行无领导小组讨论，应聘者的表现真实性会受影响。

AI助力 心理测评与人才管理论坛

北京師範大學心理学部

虚拟现实情境中面试的环境对应聘者压力知觉的影响

魏毅

2015 级心理测量与人力资源管理专业硕士

现工作单位：清华大学工业工程系人因研究所

指导老师：徐建平 副教授

1 项目简介

研究搭建了虚拟现实面试官面试场景，采用虚拟现实技术探究面试环境与应聘者间的关系。发现面试环境对应聘者行为存在影响，在领域性较高的环境面试应聘者表现出的行为更加积极。虚拟现实技术可以控制面试过程中面试官产生的无关因素，可用于心理学研究与面试。

2 研究目的

探究虚拟现实面试的可行性、面试环境与应聘者压力知觉及起面试得分之间的关系。

3 研究过程

研究被试： 北京地区高校在校本科生或研究生 34 名。

研究设计：

实验采用 2（环境领域性：主要领域、次要领域）×2（环境私密性：高私密性、低私密性)混合实验设计。环境领域性为被试间因素，环境私密性为被试内因素。

实验材料：

（1） 面试内容：以校园面试为主题，根据半结构化面试流程及面试考察侧重点，完成情景面试题目编制。共设计开场白 1 段，主干问题 4 题，追问 8 题，结束语 1 段。相应面试评分标准 1 份。

（2）面试情景制作：制作 2（环境领域性）*2（环境私密性）*13（问题 + 追问 + 等待）共 52 段面试视频，使用 Kolor Autopano Video Pro 制作全景视频。

（3）杨廷忠教授（2003）年翻译的压力感知量表（CPSS）

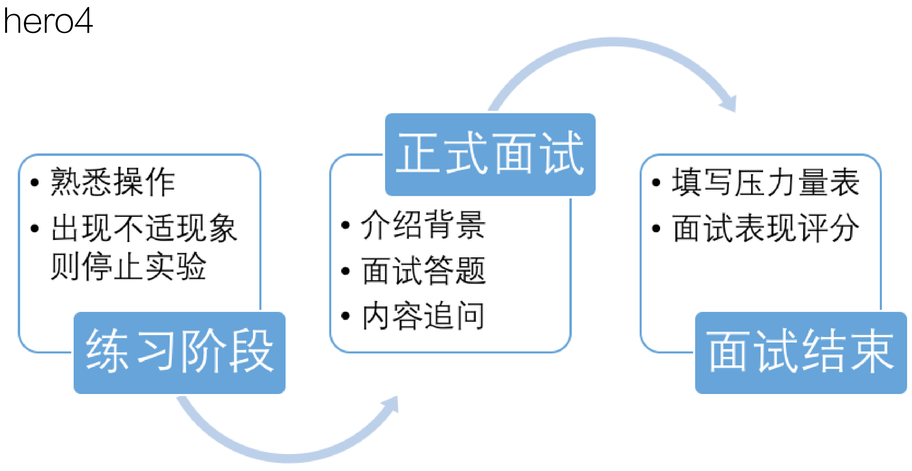
实验仪器：

（4）Oculus Dk2（5.7 英寸 AMOLED 显示屏、1080*960 分辨率，100 度可视角）

（5）ASUS 游戏级台式机（i7-6790k，18G 内存，GTX-960 显卡）

（6）GoPro hero4

实验流程：



实验结果：

文科被试比理科被试更能融入虚拟现实技术构建的面试场景之中，同时日常生活压力越大的被试体验虚拟现实面试的真实感与沉浸感越差。

主要领域与次要领域对被试在试验中感受到的面试真实度、场景真实度以及面试动机存在影响，虚拟现实场景内容越丰富，被试对情景的保真度体验越高。环境的领域性会通过影响被试的日常压力水平从而提升对被试的压力感知。

4 研究结论

面试环境的领域性会对应聘者的面试动机、面试压力造成影响，主权显示越明显的面试环境给应聘者带来的面试压力越高，同时也会带来应聘者更多的积极行为反应。

在虚拟现实下的场景模拟具有较高的保真度，可以采用虚拟现实技术进行模拟面试研究；应聘者表示中感受到的面试压力程度不足。在面试后期沉浸感最为强烈，面试行为表现最为明显，在实验后期发现模拟面试官行为死板，应聘者感受到的压力程度有所降低。

5 创新点

（1）客服应聘者的印象管理行为、晕轮效应、面试作假等行为，使用模拟面试王城重复性劳作，降低时间与人力资源成本。

（2）精确控制虚拟面试官的表情、言语、肢体动作等无关变量，做到反复实验不产生系统误差。探索与验证前人提升面试效度假设的道路。

（3）探究心理学与虚拟现实的结合和应用，开阔心理学实验新的方向与思路。



1. 主要领域高私密性面试场景设计



2. 次要领域高私密性面试设计

AI助力心理测评与人才管理论坛

北京師範大學心理学部

游戏形式的创造力测验的开发

袁玉琢

2016 级学术硕士

现工作单位：中国基础教育质量监测协同创新中心博士

指导教师：骆方 副教授

1 项目简介

创造力测量是创造力研究的基础部分，虽然创造力测验数量庞大、种类繁多，然而质量却良莠不齐，饱受质疑。随着信息技术在心理与教育测量领域的发展与应用，同时为满足当前复杂高阶认知能力的测评需求，人机交互心理测评技术——基于游戏的测评（Game-based Assessment, GBA）得到了研究者的广泛关注。

2 研究目的

本研究以“证据为中心的设计”（Evidence-centered Design）为测验设计指导框架，旨在开发基于问题情境的、游戏形式的创造力测验。创造力游戏测验以假想的“史前人类生活”为问题情境，要求被试通过组合卡片的方式探索出尽量多的创造性方案。

3 研究过程

被试： 515 名山西省大学一年级学生

测量工具：

- 背景信息问卷
- 新编创造思考图形测验 (吴静吉 等 ., 1998)
- 新编创造思考语文测验 (吴静吉 等 ., 1998)
- 顿悟性问题测验 (邱发忠 , 2005)
- 中文远距离联想测验
- 瑞文标准推理测验
- 威廉斯创造倾向量表 (王木荣 & 林幸台 , 1986)

实测过程：

编号	施测内容	施测方式	时长
PART 0	总指导语，填写并回收知情同意书		5 min
PART 1	信息调查和自陈问卷	计算机施测	15 min
	背景信息填写	在线链接	
	威廉斯创造倾向量表 50 题		
PART 2	创造力游戏测验 *	计算机施测	70 min
	指导语，完成游戏	单机软件	
	创造力游戏体验问卷	在线链接	
休息			
PART 3	创造力效标测验 *	纸笔施测	40 min
	新编创造思考语文测验 *	纸质测验	
	新编创造思考图形测验 *	纸质测验	
	顿悟性问题测验 *	纸质测验	
PART 4	瑞文标准推理测验 60 题	在线链接	40 min

注：* 表示在本研究中需平衡设计的测验

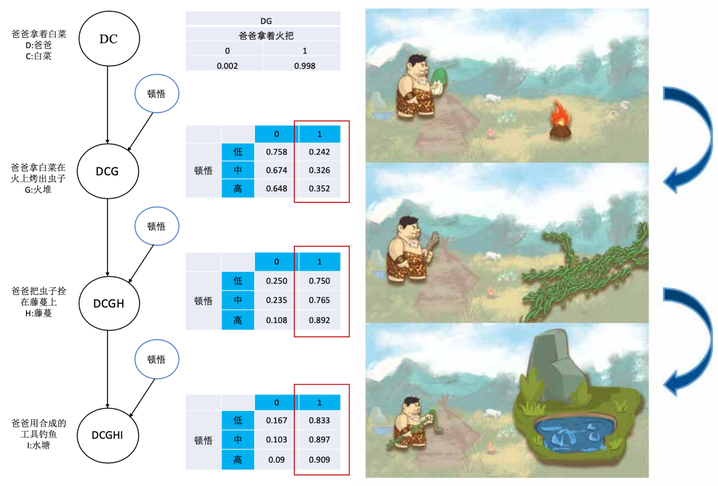
研究结果：

（1）研究一结果：

- 创造力游戏测验具备良好的心理测量学属性；
- 创造力游戏测验能够测量创造力的多种认知思维能力，并能被测量指标所记录；
- 创造力游戏测验能够体现游戏测评的优势，具有较好的作答体验性和趣味性。

（2）研究二结果：

不同顿悟水平的被试在研究预设的正确路径步骤上表现出差异。具体表现为，在已知同样合成上一步组合卡的条件下，高、低、中三组顿悟水平个体合成当前组合卡的条件概率表现出差异。



4 研究结论

（1）创造力游戏测验具有良好的心理测量学属性，能够有效的测量到创造力的多种认知成分。

（2）创造力游戏测验的任务设计调动了个体的顿悟认知过程，并发现某些关键步骤与顿悟能力存在直接关联，证明了测验的结构效度。

5 创新点

（1）创造力测量方法的创新：

- 融合发散性测验和顿悟类测验的优势，考虑创造力评价的适宜性；
- 构建不良定义的问题和半封闭性的答案集合，避免作答开放性过强；
- 组合卡片的形式产生创造性解决问题的方案；
- 计分简便，避免评分者主观因素；
- 较为良好的测量学指标。

（2）创造力测量的问题情境创新

- 采用人机交互的基于游戏的测评形式；
- “原始人生活”的问题情境；
- 诱发创造性问题解决的认知过程；
- 创造力任务的作答情境轻松且有趣。

