

《心理学报》审稿意见与作者回应

题目：一种基于 Q 矩阵理论朴素的认知诊断方法

作者：罗照盛；李喻骏；喻晓锋；高椿雷；彭亚凤

第一轮

审稿人 1 意见：

该文借鉴广义距离判别法的思路，利用海明距离定义观测反应模式与理想反应模式之间的距离，从而给出了一种新的认知诊断方法。该方法的优点是计算简单，便于应用，但该方法的关键在于当某个观测反应模式与多个理想反应模式距离相等且最小时的判别方法，作者给出了两种方法 B 方法和 R 方法。

意见 1：建议作者说明与多个理想反应模式距离相等且最小的观测反应模式占全部观测的比例。如果这类观测反应模式比较少，则判别效果一定比较好。

回应：已经采纳审稿专家的建议，对该比例进行说明（详见“5 研究结果”的表 1 和图 1 及对应的文字说明）。由表 1 可以看出，五种属性层级关系下该比例由低到高分别为直线型、收敛型、独立型、发散型、无结构型，印证了审稿专家的判断，若该比例较小，则对应的判别效果更好。

意见 2：模拟研究表明 B 方法优于 R 方法，为了便于他人应用文章中给出的 B 方法，文章中应该给出先验概率的选取原则，总体分布是什么及后验概率的计算公式。

回应：已经采纳审稿专家的建议，对 HDD 的 B 方法的操作步骤进行了详细论述。给出了后验概率的计算公式，先验概率及条件概率的选取方法（详见“3.3 应用海明距离作判别”部分，并将所参考的文章加入到“参考文献”中去）。总体分布本研究采用的是标准正态分布，在正文中已有说明，参见“4.2 被试作答反应矩阵的模拟”中第一段的第三行。

意见 3：文章中 R 方法“若某个被试的 ORP 同时与两种或两种以上的 IRP 的海明距离相等且最小时，则随机选取一种作为被试的 IRP，该 IRP 所对应的 IMP 则为被试的 IMP”。建议作者解释为什么随机选取一种 IRP 作为被试的 IRP，其判别效果也不错。

回应：已经采纳审稿专家的建议，增加了对 R 方法的合理性的论述。坦言，在提出 R 方法之时，我们认为这种方法从逻辑上来说合理的，因而能够预计其可以提高对被试的判别分类的准确率，但究竟能提高多少需要结合具体的测验情境，在不进行模拟实验的情况下无从得知。所以，对于“R 方法的判别效果为什么也不错”的问题，我们采用如下方式进行解释：先阐述 R 方法的合理性，后用实验数据证明该方法的合理性，同时通过实验数据直观地说明不同实验条件下 R 方法的判别效果。

审稿人 2 意见：

《一种基于 Q 矩阵理论的朴素的认知诊断方法》审稿意见：

本文提出一种基于海明距离的认知诊断判别分类方法，该方法与其他方法相比最显著的特点是容易理解和计算简便性，文稿表述清晰，逻辑合理。通读全文，我认为可以从以下几个方面改进本文：

意见 1：摘要中文献的表述方式是否恰当，需要检查；第五行“即海明距离判别法”是不是可以改为“称为海明距离判别法”；第二次提到“海明距离判别法”时是否改用 HDD？

回应：感谢审稿专家的细致评阅，已修改。

意见 2: 引言中的英文表述不全一致, 特别是有些的首字母大写, 有些的有没有大写;

回应: 感谢审稿专家的细致评阅, 已修改。

意见 3: 引言第四段第三行讲“GDD 引入 IRT 的正确作答概率 P 以及对应的项目参数 $a, b...$ ”与“2.2 广义距离判别法”中“GDD 通过被试的 ORP 和 IRP 的组合估计认知诊断项目参数 $a, b...$ ”, 这两处的意思是不是不一致?

回应: 感谢审稿专家的细致评阅。正如审稿专家所说, 这两处说明的对象是一样的, 表达的意思理应一致。而原始文稿中的表达容易使读者产生误解, 有失妥当, 现已修改。

意见 4: “2.2 广义距离判别法”中第五行“再根据其定义的计算公式求取.....”是否应该把公式补充在下面;

回应: 感谢审稿专家的细心指出。对于是否应该给出 GDD 的公式这一问题, 我们的理解如下: 第一, “2.2 广义距离判别法”隶属于“2 规则空间方法、广义距离判别法及海明距离概述”, 其目的是使用较为精简的语言对 GDD 进行介绍, 当读者需要对 GDD 进一步深入了解时, 我们鼓励读者查阅原作者所著文章, 从中可以得到更加详细而准确的信息; 第二, 若要将 GDD 的计算公式给出, 则前文所提到的 RSM 的计算公式同样也应当给出, 且在给出公式的同时, 有必要对公式进行一定的解释, 这样一来就需要较多的文字进行论述, 而本文意在提出一种新的方法, 因此, 将论述重点放在对新方法的论述上应当是合乎情理的。

基于以上考虑, 我们认为, 不给出计算公式不失为一种合理的处理方式。

意见 5: “3.3 应用海明距离做判别”此部分内容和“3.2 海明距离判别法的合理性”第二段有重复之处, 请修改以使表述简洁, 另外这部分应该加上“后验概率”的计算方法;

回应: 感谢审稿专家的细致评阅。这两段确实有重复之处, 分别进行如下修改: “3.2 海明距离判别法的合理性”的第二段对于 B 方法和 R 方法分别只用一句话概括其方法的实质, 即 B 方法和 R 方法“是什么”, 把重点放在对“为什么合理”这一问题的论述; “3.3 应用海明距离做判别”部分则重点论述 B 方法和 R 方法“怎么做”, 增加了对 B 方法的操作步骤的详细论述, 包括如何获得先验分布、条件分布, “后验概率”的计算方法也在此部分给出。

意见 6: 我很想知道这种方法中情况: “被试的 ORP 同时与两种及以上 IRP 的海明距离相等且最小”的比例大概有多少? 这就决定了是 B 方法还是 R 方法在其主要作用。

回应: 已在“5 研究结果”中添加表 1, 报告了本研究中每种实验条件下该比例的大小。请参见对审稿专家 1 的意见 1 的回应。

意见 7: 模拟研究部分还应该加上一部分内容“模拟研究的判别分类法”, 因为从结果可以看出你用了 HDD 和 GDD, 但在此部分没有加以说明。

回应: 已经采纳审稿专家的建议, 在“4 模拟研究”中添加了“4.4 模拟研究使用的判别分类法”部分, 报告了本研究中所使用的判别分类法。

意见 8: 你采用的海明距离判别法显得太简单了, 以至于我不太敢相信其判别的结果是否如你所说, 另外为什么不采用欧氏距离代替海明距离呢?

回应: 同审稿专家一样, 我们的第一反应也是不敢相信。在进行这项研究之前我们的团队进行了一系列的专题讨论, 其主题是如何将认知诊断评价理论和方法应用到实践当中, 以便让

更多的非测量学专业的学科专家也能够应用认知诊断评价方法。基于讨论结果，团队成员开始进行系列研究，其中就包括对属性模式判别分类方法的研究。我们的初衷就是探索一种足够简单的方法，然后在此基础上进行改进，在既保证判别分类准确性的情况下，又做到容易理解和应用。在这个想法的驱动下，我们首先想到的就是本研究所提出的海明距离判别法(HDD)的雏形——将被试的ORP与IRP进行直接比较(简称直接比较)。出乎我们意料的是，在初步的试验性研究后就发现直接比较已经具有一定程度的分类准确性。因此，我们尝试针对直接比较这种方法进行改进，提出了HDD，并设计了完整的实验方案，与另外一种刚提出的较简单且效果较好的方法——广义距离判别法(GDD)进行比较。最后的研究结果显示，HDD的表现大大超出了我们的预期。对这个结果我们团队始终都半信半疑，原因正如审稿专家所说，HDD太简单了。我们认为很有可能是研究中的某个步骤出了问题，才会导致这样“异常的”结果。因此，我们组织团队所有成员对研究的设计、研究的过程、程序代码的编写和结果的处理进行了反复地核查，才最终确定了研究结果的确是准确无误的。

审稿专家提到用欧氏距离代替海明距离，给我们提供了另外一种判别方法的研究思路。我们讨论后认为可以作为另外一项独立的对比研究。感谢专家给我们提供的新想法。

意见 9: 我不知道你的方法受测验长度的影响大不大？请请你增加一个实验：例如采用随机生成项目考核模式（如共 20 题），每种 IMP 的被试 100 人，参加这 20 题的作答，采用你的方法判别分类，看看结果如何？

回应: 根据审稿专家的提议，我们选用独立型，考察属性个数为 6 的实验条件。该种条件下的 IMP 共有 64 种，每种 IMP 的被试 100 人，一共 6400 人。在作答反应失误概率(slip)分别为 2%、5%、10%和 15%的情况下的判别结果（50 次模拟的平均值）如下表所示：

表 1 模式判准率

属性层级关系 类型	反应失误概率 (slip)	模式判准率 (PMR)		
		HDD 的 B 方 法	HDD 的 R 方 法	GDD
独立型	2%	0.4420	0.5587	0.4307
	5%	0.3988	0.5018	0.3737
	10%	0.3519	0.4235	0.3144
	15%	0.2820	0.3322	0.2463

表 2 属性平均判准率

属性层级关系 类型	反应失误概率 (slip)	属性平均判准率 (AAMR)		
		HDD 的 B 方 法	HDD 的 R 方 法	GDD
独立型	2%	0.4761	0.8645	0.4552
	5%	0.4801	0.8415	0.4296
	10%	0.5165	0.8091	0.4159
	15%	0.5190	0.7572	0.4014

由表 1 可以看出，在当前实验条件下，PMR 由高至低分别为 HDD 的 R 方法，B 方法，

GDD, 且三种方法的 PMR 均较低。我们认为三种方法的判别效果整体水平较低的原因如下: 第一, 当属性个数为 6 个, 属性层级关系为独立型时, 其典型项目考核模式有 63 种, 而当前实验条件下是随机生成 20 个项目考核模式, 意味着大多数考核模式并未出现, 分类准确率必然大打折扣; 第二, 随机生成的 Q 矩阵中很有可能未包含 R 矩阵, 这也降低了分类准确率 (丁树良, 汪文义, 杨淑群, 2011)。

由表 2 可以看出, 在当前实验条件下, AAMR 由高至低分别为 HDD 的 R 方法, B 方法, GDD。较为特别的是, HDD 的 R 方法的 AAMR 显著高于 B 方法及 GDD。经分析, 我们认为原因如下: 在当前的实验条件下, 项目考核模式的严重不足影响了对被试的判别分类, 使得在计算某种 IRP 的条件概率时, 会出现概率为 0 的情况 (在已经判别分类的被试中该 IRP 对应的 IMP 一次都未出现, 条件概率的计算公式中的分子为 0), 进一步导致计算出来的后验概率也为 0; 当需要对 2 个 IRP 进行 Bayes 判别时, 它们的后验概率会出现均为 0 的情况; 对此, HDD 的 B 方法并没有相应的处理方式, 故默认这种情况下是无法进行判别的, 则该部分被试的属性掌握模式处于未知的状态, 也就无法计算其 AAMR。而 R 方法不存在上述问题, 它能够对所有被试进行判别。同理, GDD 也存在类似的问题, 只不过是换成了被试的 ORP 与多个 IRP 的广义距离相同且均为最小值, 在其方法中无法对这类型的被试进行判别分类, 他们的属性掌握模式也同样是未知状态, 使得 GDD 的 AAMR 也处于较低水平。

意见 10: 自检报告“第一条”中答的最后一行, “但是”是不是应该改为“而且”?

回应: 感谢审稿专家的细致评阅, 此处使用“而且”确实比“但是”更恰当, 已修改。

第二轮

审稿人 1 意见:

修改后稿件, 体现了修改建议。但下面一个小问题还需要再修改:

意见 1: 第 3.3.1 部分中公式 $P(IRP_t|Y_i) = \frac{P(Y_i|IRP_t)g(IRP_t)}{P(Y_i)}$ 的下面的 $P(Y_i|IRP_t)$ 解释有问

题, 建议将“ $P(Y_i|IRP_t)$ 为被试的真实的属性掌握模式对应的理想作答恰好是第 t 种 IRP 的条件概率”修改为“ $P(Y_i|IRP_t)$ 为被试 i 的理想反应模式为第 t 种 IRP, 但观测反应模式为 Y_i 的条件概率”。

本文有一定的创新, 方法简单便于实际应用, 建议发表。

回应: 非常感谢审稿专家的肯定。二位审稿专家均指出修改稿中对于条件概率的描述存在问题。已经采纳二位审稿专家的建议, 并做出相应修改。

审稿人 2 意见:

《一种基于 Q 矩阵理论的朴素的认知诊断方法》第二次审稿意见:

论文在初稿上做了一定补充, 经过仔细阅读, 我认为还存在下述问题:

意见 1: 2 部分对规则空间、GDD 和 HD 的概述的 GDD 部分增加广义距离公式 (这个好像

不多吧); 海明距离的概念可以提到第 3 部分; 第 3 部分 HDD 中左后的“下面, 以第 j 个下面为例...”的内容可以简化为几句话加入在距离定义合理性部分, 太简单的内容用如此多的话语显得多此一举。

回应: 感谢审稿专家的耐心指正。已针对本条意见中的三个意见进行逐条修改: 增加了广义距离公式; 将海明距离的概念提到第 3 部分, 并对相应的二级标题进行修改; 将修改稿的第 3 部分中“下面, 以第 j 个项目为例...”的内容精简并合并至“距离定义合理性”部分。

意见 2: 3.3.1 我认为后验概率的表述不准确。例如 $P(Y_i | IRP_i)$ 的含义是理想反应模式为 IRP_i

的被试中实际反应为 Y_i 的条件概率; 另外, 此段中有些用语不恰当, 比如“及计算条件概率即可”, 完全可以去掉“计算”二字, 此段末尾“论述”二字, 感觉不太恰当。

回应: 感谢审稿专家的细致评阅, 修改稿中条件概率的表述确实有误, 已根据意见做出修改。对此段中用词不当之处修改如下: 去掉了“计算”二字, 将“下面对先验概率和条件概率的获取方法进行分别论述”一句改为“先验概率和条件概率的获取方法如下”。

意见 3: 3.3.1 先验概率部分末尾“本研究采用的是标准正态分布。”那么利用标准正态分布公式计算 $g(IRP_i)$ 的自变量 x 等于什么?

回应: 在本研究中, 采用标准正态分布时 $g(IRP_i)$ 的自变量 x 为 IRP 对应的测验总分种类。

这是由研究中所采用的被试知识状态的模拟方法决定的。若根据被试属性掌握个数来模拟服从标准正态分布的被试的知识状态(蔡艳, 涂冬波, 丁树良, 2013), 则自变量 x 为属性个数。

意见 4: 条件概率的计算中 $P(Y_i | IRP_i) = \frac{P(Y_i \cap IRP_i)}{P(IRP_i)}$, 意即应该等于被试中理想属性掌握模式为 IRP_i 且反应为 Y_i 的被试人数除以理想属性掌握模式为 IRP_i 的被试人数。看了你的

表述, 好像不是这么算的, 为什么?

回应: 感谢审稿专家的细致评阅。已在文稿中“条件概率”部分增加采取当前的计算方法的理由。理由如下: 为了降低判别分类方法本身的复杂程度, HDD 采用了非参数的方式, 这使得 HDD 的 B 方法无法同其他认知诊断方法一样先通过参数估计获得所需的项目参数和能力参数(如 GDD 中的项目参数 a 和 b , 以及被试的能力参数 θ), 再结合被试在所有项目上的作答计算出似然函数, 即条件概率。因此, 在 B 方法中, 我们对条件概率的计算方式做出了相应的调整, 改为求取 IRP_i 在当前测验情境中出现的概率。该概率同样也是对当前测验

情境所提供的关于 IRP_i 的信息的描述。采用该概率所依据的逻辑是, 当某个被试的 ORP 与多个 IRP 的海明距离相等且均为最小值时, 这些 IRP 中在当前测验情境中出现的概率较大者对应的 IMP 恰好是该被试的理想掌握模式的概率更大。

意见 5: 文中公式缺乏编号。

回应: 感谢审稿专家的细致评阅, 已对修改稿中的公式进行统一编号。

意见 6: 4.1 末尾一句完全可以放到 4.2 部分。4.2 第一段末尾一句表述不完整。

回应: 感谢审稿专家的细致评阅, 已将该句改为 4.2 第一段首句; 4.2 第一段末尾, “服从”一词后面补上“标准正态分布”。

意见 7: 文中表格不规范, 可以按窗口自动调整。

回应: 感谢审稿专家的细致评阅, 已按窗口自动调整所有表格。

意见 8: 表 2 和表 3 完全可以合并为一张表格。

回应: 已采纳审稿专家的建议, 将表 2 和表 3 整合为一张表格。

意见 9: 图 1 我认为可以删掉。

回应: 已采纳审稿专家的建议, 将图 1 删去。

意见 10: 5.1 中缺乏结果原因的解释。

回应: 感谢审稿专家的建议, 已增加对结果原因解释。经考虑, 我们认为“研究结果”部分, 重点应放在对实验结果的描述上, 实验结果的原因的解释应置于“讨论”部分较为合适。故对文章的结构进行微调, 把修改稿中“5.1 HDD 的 B 方法及 R 方法与 GDD 的分类准确率比较”删除(此部分都是对实验结果的描述, 不需要单独的二级标题); 将修改稿中“6 讨论及展望”改为“6 小结与讨论”, 在此部分增加“6.1 研究结论”对实验所得结果进行总结和概括, 并对实验结果可能的原因进行分析和解释; 将原二级标题“6.3 展望”改为一级标题“7 展望”。

意见 11: 我很疑惑你对我之前提到的意见 9 中的答复和你文中独立结构的判准率的差异为何这么大? 其实, 你并没有明白我之前的意思。我之前的想让你多一个实验, 里面分别包含 5 个属性结构的测验, 每个测验长 20, 项目均匀地取自典型项目考查模式, 每个知识状态的被试 100 人, 进行模拟实验, 检验测验长度对不同方法的影响。

回应: 感谢审稿专家的提醒。关于判准率差异大的问题, 在第一轮对您的意见 9 的回复中, 我们进行了简单分析(详见意见 9 的回复中加粗部分), 但没有详细解释, 在此作补充说明。受意见 9 的回复中所做实验里项目考核模式生成方式(随机生成 20 个)所限, Q 矩阵只包含有少量的典型项目考核模式(共有 64 种), 且无法保证 Q 矩阵中包含 R 矩阵, 此时的 Q 矩阵不是充分必要 Q 阵(丁树良, 杨淑群, 汪文义, 2010; 丁树良, 汪文义, 杨淑群, 2011)。这使得 IMP 与 IRP 无法建立起一一对应关系, 降低了判准率。以 3 个属性、独立型为例, R

矩阵为 $\begin{pmatrix} 100 \\ 010 \\ 001 \end{pmatrix}$ 。若 Q 矩阵中未包含 R 矩阵中的(100)行, 如 Q 矩阵为 $\begin{pmatrix} 010 \\ 001 \\ 110 \\ 101 \\ 011 \\ 111 \end{pmatrix}$ 时, 则 IMP(100)

和 IMP(000)所对应的 IRP 完全一致, 均为(000000)。当判定被试的理想作答模式为(000000)时, 无法进一步判断被试的 IMP 究竟是该 IRP 对应的(100)还是(000), 这必然导致判准率下

降。

关于您建议多做一个实验，以检验测验长度对不同方法的影响的问题，同时也为了考察 Q 矩阵中是否包含 R 矩阵对本文所采用的认知诊断方法的判准率的影响，我们针对您所要求的“5 个属性结构的测验，每个测验长 20，项目均匀地取自典型项目考查模式，每个知识状态的被试 100 人”的实验进行了扩展，属性个数与正文保持一致，设定为 6 个属性，将每种实验条件都分为 Q 矩阵包含 R 矩阵和不包含 R 矩阵两种情况。其中，包含 R 矩阵的 Q 矩阵的模拟方法为 Q 矩阵中固定包含 1 个 R 矩阵，其余项目均匀地取自典型项目考核模式。实验结果见下表（为了表格的简洁，将 Q 矩阵中包含 R 矩阵和不包含 R 矩阵两种情况分别记为“R”和“无 R”）。

从表中数据可以看出：

(1). 相对于正文中的实验结果，当前五种属性层级关系中除无结构型外，判准率都有所下降。属性层级关系越紧密，判准率下降的幅度越小。

(2). Q 矩阵中未包含 R 矩阵会在一定程度上降低判准率。属性层级关系越紧密，降低的幅度越小。

(3). 属性层级关系对当前实验条件下三种方法的判准率的影响与正文中的实验结果一致，即属性层级关系越紧密，判准率越高。

(4). 三种方法的判准率由高至低分别为 HDD 的 R 方法，HDD 的 B 方法，GDD，但三者间的差距不大。

因此，从总体上看，项目数量较少的情况下，三种方法的判准率都有所下降。但是，这种下降在当前实验条件下并不单纯是由测验长度的变化所引起的，还受到其他因素的影响，如 Q 矩阵的生成方式（是否为充分必要 Q 矩阵）、被试知识状态的模拟方式（当前采取的是每个知识状态的被试 100 人的方式，使得不同属性层级关系的被试数量均不相同；而且，这种方式实质上是生成了服从均匀分布的被试群体，进一步会影响 HDD 的 B 方法的判准率）等。因此，若要系统考察测验长度对于认知诊断方法的判准率的影响，还需要更加精细和完整的实验设计。此外，影响认知诊断方法的判准率的因素，不仅有测验长度，还包括属性个数、属性正确界定与否、属性层级关系正确界定与否、被试分布、被试数量等等，需要综合考虑。已有学者进行了相关研究(蔡艳, 涂冬波, 丁树良, 2013; 涂冬波, 蔡艳, 戴海琦, 2013)。

对于实验结果中 HDD 的 B 方法与 GDD 的 AAMR 显著低于 HDD 的 R 方法的原因，我们在第一轮对您的意见 9 的回复中也进行了分析（详见意见 9 的回复中“由表 2 可以看出…”部分）。

HDD 的 B 方法、R 方法和 GDD 的模式判准率(PMR)和属性平均判准率(AAMR)(Q 矩阵中包含 R 矩阵和未包含 R 矩阵)

属性层级关系类型	反应失误概率(slip)	HDD 的 B 方法				HDD 的 R 方法				GDD			
		PMR		AAMR		PMR		AAMR		PMR		AAMR	
		R	无 R	R	无 R	R	无 R	R	无 R	R	无 R	R	无 R
直线型	2%	0.9853	0.9303	0.9964	0.9407	0.9887	0.9598	0.9981	0.9932	0.9861	0.9310	0.9976	0.9407
	5%	0.9629	0.9055	0.9907	0.9334	0.9676	0.9375	0.9944	0.9892	0.9645	0.9070	0.9939	0.9362
	10%	0.9161	0.8710	0.9823	0.9353	0.9217	0.8990	0.9857	0.9814	0.9140	0.8688	0.9845	0.9340
	15%	0.8462	0.8176	0.9634	0.9314	0.8533	0.8409	0.9712	0.9688	0.8460	0.8111	0.9705	0.9307
收敛型	2%	0.9854	0.9108	0.9967	0.9214	0.9873	0.9487	0.9978	0.9913	0.9850	0.9103	0.9975	0.9220
	5%	0.9561	0.8869	0.9896	0.9196	0.9596	0.9268	0.9929	0.9870	0.9549	0.8837	0.9921	0.9171
	10%	0.9048	0.8549	0.9784	0.9286	0.9112	0.8813	0.9836	0.9783	0.9007	0.8465	0.9821	0.9262
	15%	0.8326	0.7511	0.9630	0.8735	0.8372	0.8005	0.9674	0.9602	0.8248	0.7435	0.9661	0.8707
发散型	2%	0.9736	0.7533	0.9933	0.7721	0.9773	0.8538	0.9960	0.9715	0.9679	0.7457	0.9942	0.7652
	5%	0.9275	0.7492	0.9839	0.8004	0.9308	0.8325	0.9871	0.9654	0.9116	0.7283	0.9832	0.7811
	10%	0.8423	0.6952	0.9641	0.8036	0.8469	0.7638	0.9691	0.9483	0.8225	0.6659	0.9644	0.7751
	15%	0.7352	0.5958	0.9352	0.7675	0.7402	0.6594	0.9422	0.9160	0.7200	0.5705	0.9392	0.7355
无结构型	2%	0.9472	0.5854	0.9876	0.6144	0.9515	0.7000	0.9911	0.9251	0.9421	0.5721	0.9893	0.5959
	5%	0.8712	0.5353	0.9709	0.6120	0.8772	0.6360	0.9760	0.9028	0.8575	0.5053	0.9724	0.5639
	10%	0.7356	0.4935	0.9371	0.6495	0.7412	0.5670	0.9442	0.8781	0.7168	0.4513	0.9399	0.5727
	15%	0.6019	0.4220	0.8989	0.6468	0.6064	0.4779	0.9069	0.8444	0.5888	0.3800	0.9045	0.5483
独立型	2%	0.9246	0.4112	0.9827	0.4412	0.9291	0.5370	0.9868	0.8586	0.9196	0.3989	0.9849	0.4205
	5%	0.8188	0.3845	0.9592	0.4649	0.8245	0.4865	0.9650	0.8352	0.8072	0.3576	0.9614	0.4100
	10%	0.6523	0.3395	0.9168	0.4980	0.6577	0.4154	0.9236	0.8013	0.6367	0.3013	0.9192	0.4036
	15%	0.5052	0.2882	0.8689	0.5239	0.5109	0.3400	0.8778	0.7616	0.4950	0.2474	0.8752	0.3986

第三轮

审稿人 2 意见:

论文《一种基于 Q 矩阵理论的朴素的认知诊断方法》审稿意见:

论文在上次修改建议的基础上做了较大的修改和调整,但是通读全文认为还存在以下不足之处:

意见 1: 论文不够简洁和流畅。

回应: 感谢审稿专家的细致审阅。通读全文后,结合专家在正文中所作批注(即意见 5),我们修改了前言中各段间的过渡句和其余段落的部分文字描述,删减了语意重复或不够简洁的语句。

意见 2: 针对修改意见的修改有些地方没有经过仔细思考,直接把相应内容补充到文中,没有注意表达的连贯性,请详见修改稿中的几段修改样例。

回应: 感谢审稿专家的批评。我们逐一阅读了正文中所有的修改部分,将其中我们认为不通顺、用词不当或重复累赘之处进行了修改。

意见 3: 针对文 3.4 部分介绍的条件概率 $P(Y_i | IRP_t) = \frac{n}{N}$ 的求法,我还是存在很大的疑惑。

因为, $P(Y_i | IRP_t)$ 表示理想反应模式为 IRP_t 的被试中实际反应模式为 Y_i 的条件概率,从数

学推理上讲 $P(Y_i | IRP_t) = \frac{P(Y_i \cap IRP_t)}{P(IRP_t)}$,即可用被试总体中实际反应和理想反应一致的人

数 m ,除以第 t 种理想反应的被试人数 n 。作者根据实际反应是否对应一种最小理想反应的情况将被试分为两类,并依据第一类(即实际反应对应一种最小海明距离的理想反应)的情况求取条件概率,那么一定会遇到有些实际反应 Y_i 不在第一类,即不能够求出

$P(Y_i \cap IRP_t)$ 。换句话说,这种方法并没有得到所有的条件概率。另外,作者采用

“ $P(Y_i | IRP_t) = \frac{n}{N}$,其中, N 为在当前被试总体中,有且仅有一个 IRP 与其 ORP 的海明距离最小的被试(即第一类被试)总人数, n 为在这 N 个被试中 AMP 为第 t 种 IRP 对应的 IMP 的被试人数。”计算条件概率也是我久经思考却不能理解的做法,我觉得你的做法更像是求

$P(IRP_t)$ 的概率。正如你在解释中也提到“因此,在 B 方法中,我们对条件概率的计算方式

做出了相应的调整,改为求取 IRP_t 在当前测验情境中出现的概率。”。这样就有悖于方法的初衷,那么你就需要对方法的描述进行修改了。

回应: 感谢审稿专家的提醒。原先我们对于原稿中 $P(Y_i | IRP_t) = \frac{n}{N}$ 的解释有误。其中分母

N 是被试总人数, n 是某种属性掌握模式所分类的人数, n/N 是该种属性掌握模式的后验概率。正文没有表述清楚,实际上 B 方法的判别过程和 RSM 中的 Bayes 判别是一样的。RSM 是从待选的纯规则点中取后验概率最大(人数最多)的一个判给被试,其使用的是马氏距离。HDD 的 B 方法只是使用了不同的距离(海明距离),而 Bayes 判别方法并未改变,即还是选取与被试的 ORP 海明距离相同且最小的几个 IRP 中后验概率最大(人数最多)的判给被试。

例如，在当前测验中，APM 被判为 IRP_{i1} 的被试人数多于 IRP_{i2} ，当某一被试的 ORP 与 IRP_{i1} 和 IRP_{i2} 的海明距离均为最小值时，该被试的 AMP 是 IRP_{i1} 所对应的 IMP 的概率更高，故将 IRP_{i1} 判给该被试。

正文中引用后验概率的标准化公式只是为了阐述后验概率的定义。而 B 方法中并未直接使用该公式。但原稿中紧接该公式的 B 方法操作步骤的文字叙述又是按照该公式的各个部分进行展开。这使得专家和读者容易产生误会，认为我们使用的是该公式，却没有严格采用公式的计算方法。故将该公式及相应文字描述删除，改为直接描述 B 方法的操作步骤。

此外，经审稿专家的提醒，我们发现 B 方法中后验概率的计算有可能会出现 n 为 0 的情况。对于这种情况，原先并没有考虑到，故在原文中进行补充说明：当某一被试的 ORP 与多个 IRP 的海明距离相同且最小时，若其中部分 IRP 的 n 为 0，则依旧采用原方法进行判别分类；当多个 IRP 的 n 均为 0 时，我们建议对于此类被试，采用 R 方法进行判别。

意见 4：引言第一段中“DINA 模型(Deterministic Inputs, Noisy “and” Gate Model)”应该是“决定性输入，噪音’与’门模型 (Deterministic Inputs, Noisy “and” Gate Model, DINA) ”

回应：已修改。

意见 5：通读引言部分，感觉写的不是很流畅。另外，我认为你第一段应增加对认知诊断理论的概念和开展认知诊断意义的介绍，然后说明常用的认知诊断方法。

回应：已采纳审稿专家的建议，增加了相关论述，并在参考文献中补充所引用的文献。

意见 6：建议分小点的方式介绍 RSM 的基本步骤，感觉是倒清不楚的。

回应：已采纳审稿专家的建议，分点介绍 RSM 的基本步骤。

意见 7：正文 2.2 部分中，是不是按下述方式修改更好呢？改为“孙佳楠等人(2011)提出的 GDD 方法首先通过被试的 ORP 及 IRP 估计项目反应理论下的项目的参数 a 和 b 、被试的潜在特

质水平 θ 并据此计算被试 i 在项目 j 上的正确作答概率 P_{ij} ，进而得到项目 j 上被试 i 的观察

反应 Y_{ij} 与项目 j 上第 t 种理想反应 $I_j^{(t)}$ 的广义距离

$d(Y_{ij}, I_j^{(t)}) = |Y_{ij} - I_j^{(t)}| P_j(\theta_i)^{Y_{ij}} (1 - P_j(\theta_i))^{1 - Y_{ij}}$ 。然后对所有 J 个项目求取被

试的 ORP 与每种 IRP 之间的广义距离 $d(Y_i, I_t) = \sum_{j=1}^J d(Y_{ij}, I_j^{(t)})$ ，并按距离最短的准则对被试进行分类。研

究表明 GDD 具有良好的分类效果(孙佳楠等, 2011; 涂冬波, 蔡艳, 戴海琦, 2013; 蔡艳, 涂冬波, 丁树良, 2013)。”

回应：已接受审稿专家的建议，采用上述更简洁的表达。

意见 8：公式 (4) 的最后一个等式表示不准确，可以直接去掉。

回应：已采纳审稿专家的意见。原公式 (4) 的目的在于介绍海明距离，考虑到 3.1 部分已有相关文字描述，且公式 (4) 最后一个等式所要表达的意思在当前正文的公式 (2) 相同，故将原公式 (4) 及其对应的文字描述一并删除。

意见 9: 表述错误, 应为: “ J 、 Y_i 和 I_t 的含义与 GDD 中一致”

回应: 已改正错误。

意见 10: 建议 3.3 部分交换 B 和 R 方法的次序, 然后在此基础上修改 3.4 部分。

回应: 已采纳审稿专家的意见, 交换了次序, 并修改了 3.4 部分, 将 3.4 中 R 方法的步骤描述删除 (3.3 部分已经做了论述)。

意见 11: 3.4 部分可以不独立, 删掉标题和 3.4.2 部分的内容; “其中 $P(Y_i|IRP_t)$ 为理想反应

模式为 IRP_t 的被试中实际反应模式为 Y_i 的条件概率, 即 $P(IRP_t|Y_i) = \frac{P(Y_i|IRP_t)g(IRP_t)}{P(Y_i)}$ 。

$g(IRP_t)$ 是第 t 种 IRP 的先验概率, 特别地本文假设其服从标准正态分布; $P(Y_i)$ 是被试 i 的 ORP 的边际概率, 即 ???。由于在对被试 i 的判别过程中, 边际概率值是固定不变的。”补充上述??指代的内容, 第二, 我还是觉得你的条件概率求法和表达式的含义不一致。

回应: 感谢审稿专家的建议。修改如下:

(1)3.1-3.4 部分的内容是按照“是什么、为什么、怎么做”的整体思路来行文, 且 3.4 与 3.3 并不是完全一致的内容, 不宜进行合并处理。因此, 我们保留了 3.4 标题, 并改为“应用海明距离法(HDD)作判别”, 便于读者了解该方法的操作步骤;

(2)删除了 3.4 中与 3.3 部分重复的 R 方法相关内容;

(3)对部分用词前后文不一致的情况进行修改, 如“实际反应模式”改为“观察反应模式”;

(4)先验概率部分的论述已经删除, 参见意见 3 的回应;

(5)关于“条件概率求法和表达式的含义不一致”这一点的回应, 参见意见 3 的回应。

意见 12:

回应: 感谢审稿专家的细致评阅。该句后为公式 6, 其中使用的是“IRP”, 而该句使用的是“IMP”, 虽然二者是一一对应关系, 但是这样前后不一致确实容易引起读者的不解, 已统一为“IRP”。

意见 13: 5 研究结果中最后一段的 slip 是什么?

回应: 已将“slip”改为“作答反应失误概率(slip)”, 便于阅读。

审稿人 3 意见:

没有修改意见, 建议接受发表。