修改说明

尊敬的评审专家：

根据您的意见下面进行了回复并在文中进行了修改，主要修改了模拟研究、实测数据分析相关内容，文中修改部分用红色标注。非常感谢您的宝贵意见和建议！

第二轮评审专家意见及回复

本文提出的模型，本质是上一个由单维 IRT 中的 theta（连续）与若干认知技能（二分）组成的多元混合潜变量分布。因此，本文的模拟研究没有探讨一个非常重要的变量：这些潜变量之间的相关程度。请作者结合多维 IRT 模拟研究设计要求，探讨不同相关程度对“如何充分利用能力和知识状态之间的关系提高诊断准确性”这个关键性问题。同时，在分析实际数据事，也请给出这些潜变量的相关估计，验证模拟研究的结果。 另外，根据这个模型构造思路，除了文中提到的 DINA，HO-DINA 以及 2PLM 模型，还有一个非常重要的竞争模型：把 theta 也看作一个认知技能的 DINA 模型。如果原来的模型是 IRT+K 个认知技能的模型，那么请检验一下 K+1 维的 DINA 表现如何。 最后，作者在多处把“模型”写成了“摸型”。

回复：谢谢您提出的十分重要的意见和建议。根据您所提的评审意见，主要对模拟研究和实测数据分析部分进行了修改，并对错字进行了修改。

（1）修改了模拟研究的实验目的、实验条件和实验结果

HO-DINA 模型中属性斜率类似于 2PLM 区分度，能力与属性之间的关系紧密程度由属性斜率大小决定，且属性之间的关系也由属性斜率大小决定(Wang, Chang, & Douglas, 2012)。属性斜率越大，说明能力与属性关系以及属性之间的关系越强。因此，在研究一中通过变化不同属性斜率参数，探讨新模型能否充分利用能力和属性之间的关系提高诊断准确性。

新增加了对比实验条件，即考虑了两个水平的属性斜率，原属性斜率和降低 0.5 后的属性斜率。当新模型中属性斜率参数各降低 0.5 时，其能力与知识状态的返真性如表 7 和 8（新增加的表）。对比表 3 和表 4，可见属性斜率越大，能力返真性越高，绝大多数情况下属性判准率也越高。

在研究二中考虑了不同题长条件下属性斜率变化对新模型的影响，结果见表 14 和 15（新增加的表）。对比表 10 和 14，以及表 11 和 15，在各种相同题长条件下，属性斜率越大，能力和知识状态返真性越高。

表 7 各属性斜率降低 0.5 时新模型的能力返真性

参数	条件	*BIAS*	*ABS*	*RMSE*
θ	1	0.028	0.323	0.405
	2	0.019	0.318	0.407
	3	0.010	0.316	0.398
	4	-0.003	0.307	0.392

表 8 各属性斜率降低 0.5 时新模型的知识状态返真性

条件	*AMR(1)*	*AMR(2)*	*AMR(3)*	*AMR(4)*	*AMR(5)*	*MMR*	*PMR*
1	0.754	0.812	0.672	0.794	0.772	0.7608	0.284
2	0.766	0.818	0.72	0.798	0.736	0.7676	0.306

3	0.98	0.974	0.948	0.954	0.958	0.9628	0.842
4	0.98	0.978	0.944	0.956	0.946	0.9608	0.828

表 14 各属性斜率降低 0.5 时新模型的能力返真性

参数	2PLM 题长	DINA 模型题长	*BIAS*	*ABS*	*RMSE*
θ	5	25	0.029	0.541	0.672
	10	20	0.007	0.397	0.499
	15	15	0.003	0.381	0.476
	20	10	0.007	0.317	0.395
	25	5	-0.002	0.279	0.346

表 15 各属性斜率降低 0.5 时新模型的知识状态返真性

2PLM 题长	DINA 模型题长	*AMR(1)*	*AMR(2)*	*AMR(3)*	*AMR(4)*	*AMR(5)*	*MMR*	*PMR*
5	25	0.968	0.976	0.980	0.980	0.986	0.978	0.898
10	20	0.946	0.978	0.960	0.982	0.974	0.968	0.858
15	15	0.966	0.968	0.952	0.950	0.97	0.961	0.836
20	10	0.908	0.884	0.878	0.928	0.944	0.908	0.612
25	5	0.878	0.732	0.694	0.882	0.83	0.803	0.338

（2）实测数据部分增加了四属性 DINA 模型并对结果进行了修改

在实测数据分析中，使用新模型、HO-DINA 模型、DINA 模型、2PLM、四属性 DINA 模型(将能力看作一个属性，在 Q 阵加一列全 1)分析了 ECPE 数据。从表 17 可以看出，新模型优于除 2PLM 外的其他模型。

表 17 五个模型对 ECPE 全部数据的拟合统计量

模型	-2LL	AIC	BIC	DIC	PPP
新模型	79882	80020	80433	85316	0.51
DINA	81228	81354	81731	86543	0.43
2PL	80552	80664	80999	83930	0.36
HO-DINA	80834	80972	81385	85838	0.48
四属性 DINA	81243	81385	81810	86717	0.46

从表 20 可以看出，新模型的属性分类准确性高于部分拟合 HO-DINA、DINA 模型，可见新模型能够利用能力信息提高所有属性的分类准确性。新模型中三个属性的斜率分别是 2.93、2.15 和 2.94，截距分别是 0.061、-0.466 和-0.574。相比部分拟合模型，新模型在对题数较少的属性 2 上分类准确性提高幅度较大。在部分拟合 Q 矩阵中，属性 1、属性 2 和属性 3 分别被 7 题、3 题和 10 题所考查。

表 20 各模型的属性分类准确率

模型	属性 1	属性 2	属性 3
新模型	0.900	0.801	0.902
全部拟合 DINA	0.910	0.857	0.910
全部拟合 HO-DINA	0.880	0.830	0.906
部分拟合 DINA	0.864	0.760	0.880
部分拟合 HO-DINA	0.837	0.774	0.893
四属性 DINA	0.792	0.758	0.819

图 4 给出了被试能力与属性掌握概率之间的散点图，可以看出：能力越高，属性掌握概率越高；属性 1 和 3 斜率较属性 2 的斜率大，考察属性 1 和 3 的题目的项目参数 (s+g)/2 的均值(0.3215、0.3373)小于属性 2（0.4055），同时考查属性 2 的题数较少，导致属性 1 和 3 的分类准确率高于属性 2 的。

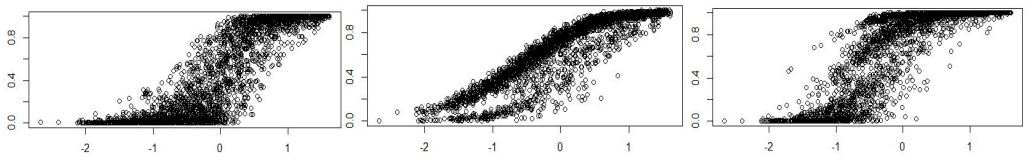

图 4 能力与属性掌握概率之间关系（从左往右，依次属性 1、2 和 3）

第一轮评审专家意见及回复

评审专家 1 意见及回复

本文提出了能力测量与知识状态诊断兼顾的计量模型。本文模型适用的场景是一张试卷中分别包含两种试题：一类专门用于能力测评，另一类用于知识诊断测评；两类试题之间不具备对方的功能。基于对本文的详细阅读,本人建议作者进行重大修改，具体理由如下：1. 使用场景的非典型性。一般来说，能力测评用于大型考试或者高利害考试，而认知诊断用于教学过程中的小测评，因而与“嵌入式测评”、“过程性评价”有很多的重叠。把这两种不同的考试情景混合起来在现实中并不多见。双重功能测试主要来自与张华华等人利用现有能力测评数据来回溯挖掘诊断信息的研究，这样的研发具有很大的意义，因为可以对已有数据进行再发现与再利用。但是本人作者提出的情景并不符合教育的实际情况，特别是目前认知诊断研究发展迅猛，我们的目标应该是开发更科学的嵌入式测评，而不是这种混合测评。2. 参数估计技术没有新贡献。由于试卷的组成，本文的参数估计相当于把认知诊断与 IRT 中的 MCMC 整合起来，不能构成重大的技术贡献。请作者详细回答本人以上两个问题，本人需要作者详细回复本人的新贡献与价值，才能进一步做出决定。

回复：谢谢您提出的十分重要的意见和建议。根据您所提的评审意见，下面详细叙述新模型的应用场景和主要贡献。

1、应用场景

新时代教育评价改革提出严格控制教育评价活动数量和频次，减少多头评价、重复评价，切实减轻基层和学校负担。开发新的认知诊断模型，充分利用测验上不同试题上作答反应中能力与知识状态信息，对于发挥评价诊断与改进功能十分重要。新模型主要应用于形成性评价中认知诊断。如您所说，本模型适应的场景是一张试卷包含两类试题：一类无需标 Q 矩阵，用于能力测评；另一类需标 Q 矩阵，用于诊断测评；两类试题虽不含相同试题而属于同一章节测试范围。新模型主要服务于面向学习的诊断测评，新模型中的能力类似于部分认知诊断模型中能力，并非局限于大规模测评的能力。许多认知诊断模型也对能力进行了建模。例如，高阶确定性输入噪音与门模型(HO-DINA 模型;de la Torre, 2004)，借助高阶能力构建知识状态条件分布。还有目前认知诊断研究发展迅猛的纵向认知诊断模型，如基于 HO-DINA 模型构建的纵向认知诊断模型（Lee, 2017；Wang，Yang，Culpepper，& Douglas，2018; Zhan, Jiao, Liao, & Li，2019；詹沛达，潘艳方，李菲茗, 2021）和多水平认知诊断模型(Lee, 2017)，都借助于不同时间点能力变化或相关刻画属性状态分布的变化。基于以上考虑，考虑到标题“兼顾能力测量与知识状态诊断的计量模型”不是太适合，原标题将能力测量与知识状态诊断视为同等地位，并且“计量模型”过于一般化。这容易让读者将能力测量和知识状态诊断两种不同的考试目的或情景混合起来。因此，我们将标题修改为“融入能力信息的认知诊断模型开发与应用”，并对文中相关描述进行了修改，将本文重点聚焦于认知诊断。

2、主要贡献

（1）新模型可提高知识状态分类准确性。新模型利用能力与属性状态之间的关系或相互蕴含的信息，提高知识状态诊断的准确性。就如同将能力视为知识状态的协变量，利用协变量与知识状态之间的桥梁关系，从而可间接利用能力测评试题上的作答反应中的信息，用于提高知识状态诊断的准确性。Wang，Yang，Culpepper 和 Douglas（2018）将能力作为协变量，提出结合认知诊断模型、高阶模型和隐马尔可夫模型的学习模型，用于追踪学生技能掌握，并用于评价不同学习干预措施的效果。

（2）新模型可提高能力估计精度。新模型充分利用无需标 Q 矩阵的试题上提供的能力信息，并附加被试属性状态中蕴含的能力信息，可以减小被试能力估计的误差小，这很好地克服了 HO-DINA 模型的高阶能力误差过大的问题（Hsu & Wang, 2015）。值得注意的是，新

模型用于嵌入式测评数据分析时，因为测验长度相对较短且含两类试题仅能覆盖较小范围内容领域，因此，新模型所得能力不具备大规模测评能力的泛化能力。

（3）新模型对于认知诊断走进课堂具有重要实际意义。尽管认知诊断模型研究发展十分迅速，但是标定测验 Q 矩阵是认知诊断评估真正走向实际应用的基本前提。新模型中仅需标定已有数据测验中部分试题的 Q 矩阵，或只需要开发少数认知诊断试题而其他试题可以使用常规测验试题，这样可减少 Q 矩阵标定难度及测验开发代价。

（4）新模型有助于发现与再利用已有数据的信息。张华华等人利用现有能力测评数据来回溯挖掘诊断信息的研究，因为可以对已有数据进行再发现与再利用，这样的模型框架具有重要意义。对于能力测评数据，要求项目反应理论模型，尤其是认知诊断模型都很好地拟合各个测验题目，存在一定的难度。如果部分试题只可以较好地拟合单个模型，在张华华等人提出的模型框架下，这部分题目便会被删除而信息被浪费，而新模型是可以适合此应用情景。

（5）新模型有可能为认知诊断等值研究开辟全新思路。因为认知诊断模型通常是对离散潜在知识状态进行建模，这给认知诊断等值研究带来了挑战。而本文提出的新模型，结合了能力建模，项目反应理论框架下成熟的等值设计、等值方法的相关成果，是否可为认知诊断等值研究开辟全新思路，值得进一步研究。

评审专家 2 意见及回复

摘要中“模拟研究结果显示，新模型对能力和知识状态估计质量高，对项目参数估计的返真性较好；新模型在英语水平证书考试的实测数据上表现也相当不错”要具体些，要有数据说明。论文主要从统计学角度来验证模型，还没验证能力和属性的诊断结果是否符合学科特征和被试特征。

回复：谢谢您提出的十分重要的意见和建议。主要对摘要和实测数据分析部分进行了较大修改。

（1）对摘要进行了仔细修改。摘要中“模拟研究结果显示，新模型对能力和知识状态估计质量高，对项目参数估计的返真性较好；新模型在英语水平证书考试的实测数据上表现也相当不错”，已经修改为“模拟研究考查了新模型在四种参数分布和五种题量下的表现，并与 DINA 模型与 2PLM 进行了比较。结果显示：在相同题量(如 15)下，新模型能力参数的均方误差（0.449）低于 2PLM 的均方误差（0.490），同时新模型的模式判准率（0.868）高于 DINA 模型的模式判准率（0.786）。在英语水平证书考试的实测数据上，新模型相对拟合指标优于 DINA 和 HO-DINO 模型；新模型虽稍逊于 2PLM，但两者所得能力与测验总分相关较高；新模型在属性 2 上分类准确性高于其他模型，分析发现新模型可利用能力信息提高 Q 阵中考查次数较少属性的分类准确性。”

（2）对实测数据分析结果进行了修改。除了从统计上报告模型拟合指标外，我们增加了能力与测验总分关系、属性分类准确性指标估计的结果和相关描述。由图 3 可见，新模型与 2PLM 所得的能力与测验总分相关较高。除预烧外链长中每隔 100 取样计算属性掌握概率，然后得出属性分类准确性(Wang et. al., 2015)。从表 18 可以看出，新模型的属性分类准确性均值与 HO-DINA、DINA 模型基本相当，并且新模型在属性 2 上分类准确性高于其他模型。对测验 Q 矩阵分析发现，测验中有 13 题考查属性 1，有 18 题考查属性 3，而仅有 6 题考查属性 2。这说明，新模型可利用能力信息提高 Q 阵中考查次数较少属性的分类准确性。

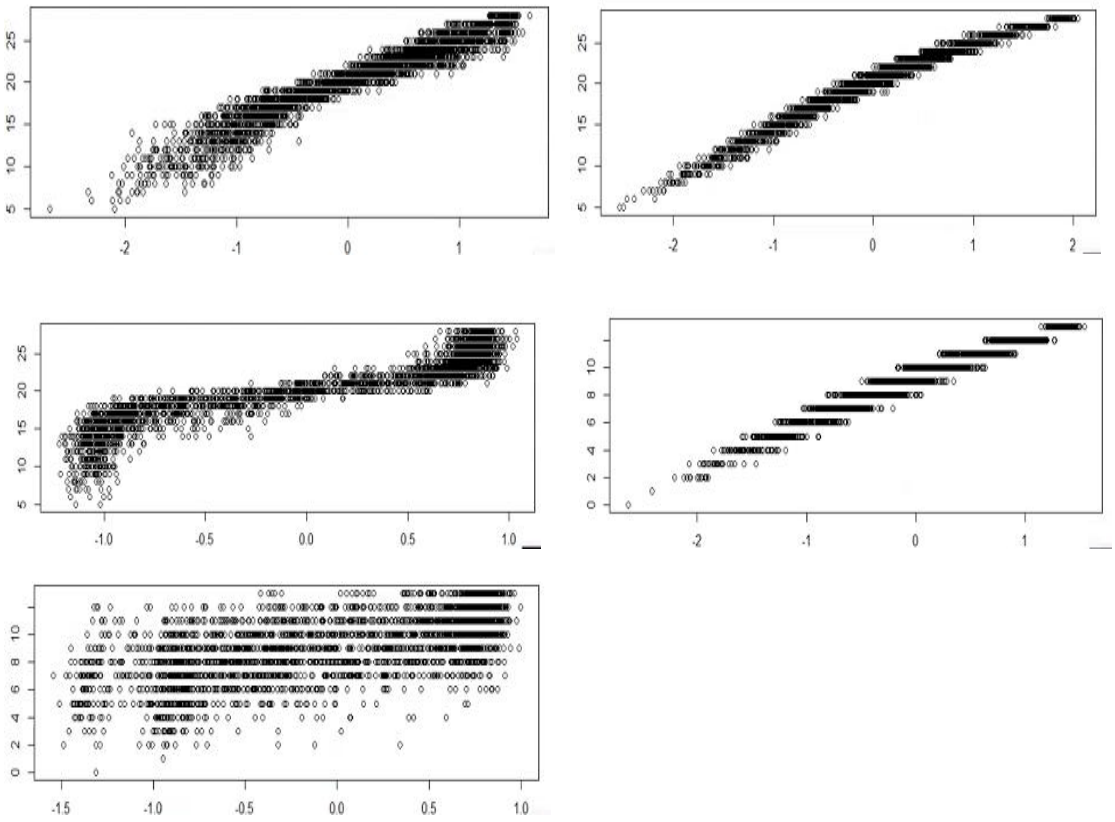

图 3 不同模型下能力与测验总分的关系

（上从往下，从左往右依次为新模型、全部 IRT、全部 HO-DINA、部分 IRT、部分 HO-DINA）

表 18 各分析模型的属性分类准确率

模型	属性 1	属性 2	属性 3
新模型	0.8948	0.8890	0.8838
全部拟合 DINA	0.9092	0.8588	0.9107
全部拟合 HO-DINA	0.9132	0.8586	0.9107
部分拟合 DINA	0.8632	0.7970	0.8800
部分拟合 HO-DINA	0.8656	0.8501	0.8809