# 《心理学报》审稿意见与作者回应

题目：基于信号检测论的认知诊断评估：构建与应用

作者：郭磊，秦海江

---

## 第一轮

首先非常感谢三位审稿人提出的宝贵意见，使本文质量得到进一步提升。以下是我们分别对三位审稿人的审稿意见做出的回答。

**审稿人 1 意见：**

本研究在 DeCarlo(2021)研究的基础上，将信号检测论引入到认知诊断领域，并提出了基于信号检测论的认知诊断模型 SDT-CDM；与以往模型相比，新模型具有如下优势：无需对干扰项选项进行属性层面的编码，能获得传统诊断模型无法提供的题目区分度和难度参数，以及可以直接表达每个选项之间的合理性差异，对题目性能刻画更加细微全面，便于其用于题目质量诊断及修订。同时文章还进一步推导出了新模型的 EM 算法，并实现了其参数估计。随后文章通过模拟研究与实证研究来检验新模型的性能。总之，本审稿人认为本文研究视角新颖，具有一定的创新性。为了进一步提升文章的质量及可读性，建议作者参考以下意见进行修改：

**意见 1**：2022 年 JEBS 期刊(见下)发表了一篇关于 MC 认知诊断建模的研究，该研究并不需要本文作者提到的 “…干扰项的编码需要是正确选项编码的子集、不同干扰项之间也要有包含关系…”这个条件。所以作者在文献综述时，一方面需要引入该最新研究成果，另一方面需要对文章中以往研究中存在的不足的论述进行修改。

Wang, Y., Chiu, C., & Kohn, H. F. (2022). Nonparametric Classification Method for Multiple-Choice Items in Cognitive Diagnosis. Journal of Educational and Behavioral Statistics. https://doi.org/10.3102/10769986221133088

**回应**：感谢审稿人提出的宝贵建议。我们在阅读了您提到的该研究后已将不恰当的描述进行

了修改，修改如下：“虽然前期的研究要求干扰项的编码需要是正确选项编码的子集、不同干扰项之间也要有包含关系(郭磊 等, 2013)，但最近的研究已突破了该限制，即干扰项的编码无需嵌套于正确选项编码中(Wang et al., 2022)，进一步推动了该领域研究。”再次感谢您提供的宝贵学习经验和文献。

**意见 2**：对于 MC 的认知诊断，目前国际上主流研究(如 Wang et al., 2022; Ozaki, 2015; DiBello et al., 2015; de la Torre, 2009)基本上都充分利用 MC 干扰选项的信息，不仅是利于了在干扰选项上的作答信息，还会充分利用干扰选项上的测量属性信息(q 向量信息)，从而真正将干扰选项与认知诊断充分结合。本研究新开发的模型并没有将选项的测量属性信息考虑进来，但建议作者在讨论中还是需要对该问题进行充分说明。尤其是结合最新研究(Wang et al., 2022)已突破了作者在文章中说的前提条件。

**回应**：感谢审稿人提出的宝贵建议。我们已将该部分内容补充在 6.1.1 部分。

**意见 3**：建议作者对附录中的公式推导等进行核查，以进一步确认公式推导的可读性、连贯性及准确性。

**回应**：感谢审稿人的建议。我们已对全文的所有公式及其推导过程进行了检查。

**意见 4**：讨论部分作者重点论述讲了未来的研究方向，但并没有将本研究的主要发现与以往的研究进行深入比较讨论，请补充。

**回应**：感谢审稿人的建议。我们已对本研究的主要发现及其与传统的模型 NRDM 进行了比较讨论，请参见 6.1 部分。

**……**

**审稿人 2 意见：**

文章提出了基于信号检测论的认知诊断模型 SDT-CDM。通读全文，文章需要解决以下问题：

**意见 1**：文章提出的 SDT 认知诊断模型仅是基于 DeCarlo(2021)将信号检测论在项目反应理论(IRT)应用的拓展。对比文中 IRT 的 SDT 模型(公式 1)和认知诊断的 SDT(公式 2)，两者的公式差异似乎只存在于 theta 到 alpha(知识状态)的改变。

**回应：**感谢审稿人提出的宝贵建议。该项研究看起来似乎“仅是基于 DeCarlo(2021)将信号检测论在项目反应理论(IRT)应用的拓展”，但在拓展过程中要解决的问题，所构建模型的参数调整，以及参数估计的 EM 算法推导等环节均有较大差异。我们认为一个研究的创新性是要看对该领域研究的现实/实际贡献，在 SDT-CDM 提出之前，仅有 NRDM 能够在无需对干扰项进行编码时，对选项层面数据进行处理。但正如文中所述，即使是简化的 NRcRUM 也需要极大的样本量(至少 5000 人)才能保证其参数估计的精度(Templin et al, 2008)，何况饱和的 NRDM 模型。然而，现实情境中的诊断测验很少能够同时收集如此多被试的作答数据，导致 NRDM 及其简化的 NRcRUM 在现实情境中难以适用，并且 NRDM 也无法提供题目的难度和区分度参数，使其在刻画题目质量上信息有限。因此，需要开发一种更加高效的模型来帮助实际使用者分析数据，不仅能获得选项层面的参数信息，还能够提供题目区分度和难度指标以全面评估题目质量。事实上，能够从一批数据中充分挖掘信息也是推动模型开发的一个重要方向，很多模型的开发思路是将辅助数据包含进来以获得模型估计能力的提升，如 JRT-DINA 模型就将反应时数据包含在高阶 DINA 模型中(Zhan et al., 2017)，而本研究无需额外辅助数据便可获得比 NRDM 及传统模型如 GDINA 等更丰富的分析结果，而且 SDT-CDM 也无需对选项编码，比 MC-DINA 等模型更实用高效，这正是我们研究的重要贡献。

此外，在 CDM 的模型开发中，大部分的研究也是采用“拓展”的思想。如 JRT-DINA 模型就是借鉴了 van der Linden(2007)的层级框架模型的理念，将三参数 IRT 模型拓展至 DINA 中。再如 Ma 等(2021)将原始的 GDINA 模型拓展至多组情境，仅在原始 GDINA 模型的题目参数下角标增加表示组别 $g$ 的符号。类似的研究还有将 Wilson 等(2006)提出的解释性 IRT 模型拓展至 CDM 中，提出了解释性 CDM(Park et al., 2014)；Huang(2017)将多水平 IRT 模型(Huang, 2015)拓展至多水平 CDM 中。

综上理由，我们认为将 SDT-IRT 推广至 SDT-CDM 是一种创新的研究，解决了目前 CDM 所不能解决的问题。

**相关文献：**

Huang, H.-Y. (2017). Multilevel Cognitive Diagnosis Models for Assessing Changes in Latent Attributes. *Journal of Educational Measurement, 54,* 440-480.

Huang, H.-Y. (2015). A multilevel higher-order item response theory model for measuring latent growth in longitudinal data. Applied Psychological Measurement, 39, 362–372.

Ma, W. Terzi, R., & de la Torre, J. (2021). Detecting Differential Item Functioning Using Multiple-Group Cognitive Diagnosis Models. *Applied Psychological Measurement, 45*, 37-53.

Park, Y. S., & Lee, Y. S. (2014). An Extension of the DINA Model Using Covariates: Examining Factors Affecting Response Probability and Latent Classification. *Applied Psychological Measurement, 38*, 376-390.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.

Wilson, M., Boeck, P. D., & Carstensen, C. H. (2006). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects*. Göttingen: Hogrefe & Huber.

Zhan, P. D., Jiao, H., & Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*, 262-286.

**意见 2**：文中提到：“SDT 用于认知诊断评估有以下优势：①无需对 MC 题目的选项进行编码，节省大量人力物力。”审稿人认为一般的认知诊断模型似乎也不需要对 MC 题目的选项进行编码，如作者在文中提到的 NRDM 诊断模型，是否需要对 MC 选项编码信息呢？如果传统 CDM 无需编码选项信息，那 SDT 的这点优势是否还成立？

**回应**：感谢审稿人提出的宝贵建议。目前大部分传统的 MC-CDM 均需要对选项的 $q$ 向量进行编码，如 MC-DINA (de la Torre, 2009)，MC-S-DINA (Ozaki, 2015)等，以及基于选项的非参方法(郭磊 等, 2021; Wang et al., 2022)。但也存在无需编码的处理方法，如 NRDM 与本研究提出的 SDT-CDM，这两种模型都将 MC 数据视为称名数据进行分析，传统 CDM 是视作 0-1 计分来分析，而称名数据比 0-1 数据提供的信息更加丰富。但 NRDM 需要在极大样本量情境下才能有较好表现，这也是开发 SDT-CDM 的一个重要原因。在现实测验情景中，我们建议：若能够对较多的选项进行编码，且被编码选项的 $q$ 向量彼此间不雷同，可使用传统的 MC-CDM 进行数据分析。但据我们所了解，目前能够对选项层面进行编码的诊断测验非常少，绝大多数测验仅有题目层面的 $q$ 向量编码，因此，SDT-CDM 的现实实用性更强。而且根据本研究的模拟和实证研究均验证了新模型具有较好的表现，值得推广使用。

**意见 3**：此外，文中还提到：“③由于模型更加简洁，模型和数据的拟合可能会进一步提升”，如果审稿人没理解错，模型更加简单，作者应该指的是 SDT 和 NRDM 相比，SDT 的项目参数更少。将文中公式(3)和 NRDM 相比，似乎 SDT 只是少了一个截距参数；此外，审稿人认为，该点优势成立是由于作者采用的是 NRDM(Templin et al., 2008)的复杂模式，Templin 等人(2008)提出 NRDM 模型的研究中，也提到为了简化模型以更好地应用于实际测验，曾

将题目参数加以限制，初始复杂的 NRDM 模型可以简化为 NRcRUM (Nominal Response cRUM)，此时，NRDM 模型的参数会大幅度减少，如果将 SDT 和 NRcRUM 相比，SDT 的第(3)点优势好像也难以满足。

**回应：**感谢审稿人提出的宝贵建议。SDT-CDM 没有截距项是因为需要符合现实情境：当被试未掌握题目考察的任何属性时，混合参数$\lambda_{1j}$的值需要等于 0，即表示不会作答。此外，饱和的 NRDM 共有$2^{K_j^*} * (H-1)$个参数需要估计，$K_j^*$为题目 $j$ 所考察的题目数量，$H$ 为选项个数。以一道 4 个选项的题目考察了 3 个属性为例，饱和的 NRDM 需要估计$2^3 * (4-1) = 24$个参数：3 个截距，9 个主效应，9 个二阶交互效应，3 个三阶交互效应，共 24 个参数。但饱和的 SDT-CDM 仅需要估计 3 个主效应，3 个二阶交互效应，1 个三阶效应和 3 个合理性参数 $b$，共 10 个参数。即便是简化模型 NRcRUM，也有 3 个截距和 9 个主效应，共 12 个参数需要估计，而简化的 SDT-CDM 只需要估计 6 个参数：3 个主效应和 3 个合理性参数 $b$。因此，SDT 的第(3)点优势仍然可以满足。

**意见 4：**实验设计方面，在研究一中，作者只是验证了 SDT-CDM 的参数估计精度，并没有与其他类似的 CDM 进行比较，因此，研究一的结果只可以说明 SDT-CDM 的参数估计算法是科学、可行的。审稿人认为如果要体现 SDT-CDM 的优势，在模拟实验阶段，仍需要将 SDT-CDM 和类似的几种 CDM 加以比较，从参数估计精度等方面，检验 SDT-CDM 的价值。

**回应：**感谢审稿人提出的宝贵建议。由于 SDT-CDM 分析数据时无需对选项的 $q$ 向量进行编码，为了公平比较起见，需要选择一个也是仅根据题目 $q$ 向量就能分析称名数据的模型，因此如 MC-DINA，MC-S-DINA，以及基于选项层面的非参方法等不予考虑，最终选择了 NRDM 进行全面比较。从新补充的模拟研究 2 结果可以看出：SDT-CDM 从各方面都要优于 NRDM，通过详尽的模型比较研究，进一步证明了新模型的优势。

**意见 5：**实证数据取自 PISA2000 的英语阅读，审稿人认为实证数据不一定重复引用已有研究使用过的数据，况且该数据年代间隔过于久远，在数据的时效性方面可能存在不足。

**回应：**感谢审稿人提出的建议。为了保证数据的时效性，同时也需要满足选项答案数据公开、题目已存在 Q 矩阵等条件，我们选择了 Ma 和 de la Torre 在 2020 年发表文章中使用过的 TIMSS 2011 数据，该数据共包含 748 名来自美国的被试在 23 道数学测验题目上的作答原始数据，本研究选择其中的 14 道选择题进行分析。

……

**审稿人 3 意见：**

基于“选择题作答可被视为从噪音中提取信号的过程”的思想，文章对 DeCarlo(2021)提出的 SDT 选择模型进行拓展，得到适用于认知诊断评估的 SDT 选择模型(SDT-CDM)。总体而言，论文的整体框架完整，选题具有较好的实践价值，但创新性有待提高，文字表达还需进一步完善。具体意见如下：

**意见 1**：DeCarlo(2021)将 SDT 与项目反应理论结合用于选择题的题目分析，并对模型作了详细推导。而本研究只是将 DeCarlo 的模型用于认知诊断，并将原本反映被试总体水平的 拓展为被试会作答题目 *j* 的概率 。因此从创新性的角度来看，本文还有提升空间。比如，从第 21 页的结果来看，SDT-CDM 得出的易度估计结果与 2PL/3PL 模型的“易度”结果的相关并不算高。所以，要使“新模型可以提供与 IRT 模型近似的难度参数表达，用以反映题目的难度水平”，可能在模型构建方面还可以做些改进。

**回应**：感谢审稿人认真细致的审阅及提出的宝贵建议。关于研究的创新性说明，可以参见对审稿人 2 的意见 1 的回复。我们在修改稿中，增加了实证研究中关于 SDT-CDM 和 2PL/3PL 模型的难度和区分度参数相关系数的说明。具体为：根据 Cohen(1988；P82)提出的标准，相关系数$r \geq 0.5$即为大效应量(Large Effect Size)；此外根据张厚粲和徐建平(2015；P150)提出，相关系数在 0.6 至 0.8 之间即为强相关，0.8 以上即为非常强相关，并且以上 4 个相关系数均显著，因此表明新模型可以提供与 IRT 模型近似的难度参数表达，用以反映题目的难度水平。

需要注意的是，根据信号检测论的思想，SDT-CDM 中题目难度包括了两类：会作答的被试答题时的“易度”（$e_K$）和不会作答的被试答题时的“易度”（$e_{DK}$），这比 IRT 中单一刻画题目难度的参数有更丰富的解释性。即便如此，本研究中$e_K$与$e_{DK}$所表征的难度都与 IRT 模型中的难度表征高度相似。本研究将 SDT-CDM 模型的易度参数与 IRT 模型中的难度参数进行相关分析，最重要的目的不是为了以 IRT 的难度参数为标准来评价模型优劣，因为它们彼此基于不同的理论基础、概念在本质上也有所不同。这样做的目的，仅是为了将 SDT-CDM 中的“易度”与以往理论中能够表达“难度”的相似方法相互比较与印证，以此证明 SDT-CDM 中易度表征的合理性，该做法与 DeCarlo(2021)研究中的做法是完全一致的。

SDT-CDM 是根据信号检测论角度提出的，因此，遵循 SDT 框架内对于“难度”或称作“易度”的表达。作为将 SDT 初次引进 CDA 领域的研究，我们需要首先检验模型的性能及其表现，但我们非常感谢和支持您提出的可以尝试从新的视角来构建模型的想法，未来可以进一步思考在现有 SDT-CDM 基础上，找到更加适合表达题目“难度”或“易度”的参数。

**相关文献：**

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. New York, NY: Erlbaum.

张厚粲, 徐建平. (2015). *现代心理与教育统计学(第 4 版)*. 北京：北京师范大学出版社.

**意见 2**：模拟研究部分只考察新模型的表现，并未将新模型与已有模型进行比较。审稿人认为，如果要推行一种新模型，是不是应该将新模型和以往表现较好的模型进行比较。关于这一点，作者是如何考虑的？

**回应**：感谢审稿人提出的宝贵建议。我们在修改稿中已补充了“模拟研究 2”，用以全面检验 SDT-CDM 和 NRDM 之间的表现和差异。

**意见 3**：相比于已有模型，新模型的优势之一在于能够得到题目的区分度和难度参数，从而指导测验编制。但模拟研究结果表明：新模型的参数估计精度受样本量影响最大。实证研究也是对大样本数据进行分析。是否意味着新模型用于指导测验编制的条件有限？在最后的讨论和展望部分，作者只对新模型的拓展进行了展望，而没有对新模型的不足和适用条件等进行讨论和展望。

**回应**：感谢审稿人提出的建议。从模拟研究 1 的结果不难看出，当样本量在 1000 人时，参数估计精度其实已经可以接受了，只是在继续增大样本量的情况下，参数估计精度会进一步提升。该现象也普遍存在于大部分关于模型的研究中，即样本量的增加会提升参数估计的精度，因此本研究出现该结果并不会存在矛盾。此外，即使是简化的 NRcRUM 也需要 5000 人才能达到较好的参数估计精度，从这点来说，SDT-CDM 的样本量是完全可以接受的。在修改后的新的实证研究中，仅有 748 人。如果要将 SDT-CDM 用于小样本情境中，可以采用 NRDM 类似的做法，也将 SDT-CDM 进行约束，将其变为补偿的 SDT-CDM 模型，但本研究并未尝试，这值得在未来研究中进行探索。对于新模型的不足和适用条件等内容我们补充在了讨论部分。

**意见 4**：稿件中的公式符号问题：

(1)第 9 页第第 7 行，j 为下标，公式(3)和(4)中的高阶交互效应项也存在这个问题。

(2)第 10 页公式(7)和(8)中，表示估计值的符号上方一般用 hat，而不是 bar。

(3)公式(11)缺少下标 k。

**回应：**感谢审稿人提出的建议。我们已针对审稿人发现的问题修改，并对全文的其他公式进行了检查与修改。

**意见 5：**稿件中还存在一些表述不规范/语句不通顺/格式有误的地方。比如：

(1)文中部分地方使用英文括号，部分地方使用中文括号，请统一。

**回应：**已统一为英文括号。

(2)第 5 页倒数第 4 行至第 3 行，实证研究由谁开展？是指 DeCarlo(2021)吗？

**回应：**已修改。

(3)第 7 页第 7 行至第 8 行的“… 因此被试将做出正确选择答对题目”不通顺。

**回应：**已修改为合适的表达：“因此被试将做出“选择正确答案 B”的反应”。

(4)第 7 页第 10 行，“等价于 IRT 中区分度 a 参数的概念”值得商榷。

**回应：**感谢审稿人提出的建议。IRT 中的区分度特指 IRT 模型中的 $a$ 参数，因此原文的表达在文义上似乎将 SDT-CDM 中的区分度参数 $d$ 与 IRT 中的 $a$ 参数作了完全等价，确有不妥。现已更改为更合适的表达：“与 IRT 中区分度$a$参数类似”。

(5)第 8 页公式(2)下面一行，什么是“alpha_l 的被试”？

**回应：**应为“知识状态为$\boldsymbol{\alpha}_l$的被试”，已修改。

(6)第 9 页第 8 行，应为“…部分或全部属性的效应之和”。

**回应：**感谢审稿人提出的建议，已根据您的建议修改了表述。

(7)第 9 页倒数第 1 行以及第 23 页 6.1.2 节第 7 行的文献引用格式有误。

**回应：**已修改。

(8)第 10 页第 2 行，准确的表述形式应为“**参数从**分布中抽取”。公式(5)上 2 行，不是属性分布采取**分布生成，而是知识状态或属性掌握模式。公式(5)下第 2 至 3 行，表述也不规范。

**回应：**已修改。

---

## 第二轮

非常感谢两人位审稿人提出的宝贵意见，使本文质量得到进一步提升。以下是我们分别对两人位审稿人的审稿意见做出的回答。

**审稿人 1 意见：**

经过作者修改，文章质量得到了很大提升，已达到发表水平，目前文稿的篇幅有点多，建议作者做些调整与缩减。建议小修改后发表。

**回应：**感谢审稿人提出的宝贵建议，已对文稿进行精简提炼与压缩，将部分表格内容移至附录部分呈现，大幅减少了正文篇幅。

**审稿人 3 意见：**

针对审稿人的意见，作者进行了认真回复，并在稿件中做了详细修改。还有一个小问题，描述如下，供作者参考：

作者在修改稿的实证研究部分添加了“关于 SDT-CDM 和 2PL/3PL 模型的难度和区分度参数相关系数”的描述，并通过呈现 4 个相关系数“相关较强且显著”，来说明“新模型可以提供与 IRT 模型近似的难度参数表达，用以反映题目的难度水平”。我关注的点是在“近似”，请注意两个变量相关较高且显著，并不代表它们的值“接近”或“近似”。

建议修后发表。

**回应：**感谢审稿人认真细致的审阅及提出的宝贵建议。我们经过讨论，认为原稿中“近似”一词在使用上可能会让人误解，因此做了如下两个思考和修改：

① 经审稿人提醒，我们发现原稿中“近似”一词容易让读者将该词理解为“数值相近”的含义。而我们想要表达的意思是：SDT-CDM 能够提供题目的难度参数，就像 IRT 模型也能表达题目难度参数一样，是为了说明 SDT-CDM 和 IRT 模型一样都能够对题目进行难度表征。现已将文稿中相应部分的描述调整为：“因此表明新模型与 IRT 模型一样，都可以对题目进行难度表征，以此来反映题目的难度水平。”

② 文稿中关于新模型与 IRT 模型在难度参数表达上“近似”的陈述，也可能让读者混淆 SDT-CDM 与 IRT 中两者难度参数的含义。文稿中“2. SDT 模型简介”部分呈现了详细

描述，SDT-CDM 的难度由两个易度参数表达，即$e_{DK}$与$e_K$，“两者含义均为被试感知到的正确选项的合理性与剩余最高的合理性之间的差值”，而 IRT 模型中的难度则由模型中的难度参数进行表达。现已删除文稿中的“近似”一词，并调整该句表达，见①。

以上两点思考与回应中，我们认为您更有可能是想表达第一个意思，因此在修改稿中，我们增加了对①的修改。若您仍有其他建议或意见，我们会做进一步修改。

再次感谢两位审稿人的细心审阅！

---

**编委意见：**

**意见 1**：引言最后一段编委建议另起一段，加上一两句说明本文的研究目的，或者直接说本文要研究什么，然后才是文章结构。

**回应**：感谢您的宝贵建议。我们按照您的意见进行了修改，具体内容为：“综上所述，信号检测论视角的 MC 题型认知诊断评估将具备诸多优势，因此本文拟探讨基于信号检测论的 MC 题型认知诊断评估方法与技术，构建 SDT-CDM 模型并推导其参数估计方法，并在模拟和实证测验中检验新模型的性能和有效性。”

**意见 2**：一方面，文章已经很长；另一方面，对算法推导有兴趣的读者很少，不建议将附录放在期刊纸质版，但可以放在网络版。

**回应**：已将 EM 算法推导过程标记为参见网络版附录。

**意见 3**：结果中的几个图，下面还有表列数据，如果两者表示的信息相同，只保留其中一种便可，要么都用图呈现，要么都用表呈现。没有必要两种同时呈现。

**回应**：感谢您的建议，我们已进行相应修改，只保留了图的部分，删除了表格部分。

**主编意见**：本论文将信号检测论引入到认知诊断领域，提出了基于信号检测论的认知诊断模型，并通过模拟和实证研究证明了模型的性能。本论文的研究视角新颖，研究框架清晰，所获得研究结论较为可信且具有一定理论和实际应用价值。