《心理学报》审稿意见与作者回应

题目：人工智能方法在探究小学生作业作弊行为及其关键预测因子中的应用
作者：赵立; 郑怡; 赵均榜; 张芮; 方方; 傅根跃; 李康

第一轮

审稿人 1 意见：

意见 1：本研究运用了机器学习的方法分析了小学生的问卷数据来预测其作弊行为。研究的亮点在于机器学习的集成算法详尽考察了可以预测作业作弊行为的变量，包括了个人的态度，感知到的同伴行为（即社会规范印象）等。文章写作流畅，结构合理，结果呈现恰当，关于机器学习的数据分析逻辑流程也很清楚，值得赞扬。
回应：非常感谢审稿人对本研究的认可。

意见 2：研究问题：为何要考察作业的作弊行为。从作者的前言综述来看，前人对小学生学业作弊的研究极少，引用的文献只有来自 100 年前的研究(Hartshorne & May, 1928)，而作业作弊的研究可以说几乎没有。作者提出研究极少的原因是西方小学的作业较少，而且我国在新的双减政策下作业也在逐渐减少。这样的背景下，考察作业作弊是否有现实意义？为何作者不研究考试作弊这样后果更加严重的作弊行为，或者学业、学校环境中的其他不道德行为（例如对老师、家长说谎的频率）？
回应：感谢您指出了这个问题。我们在正文的文献综述部分对于为什么要考察作业作弊行为进行了进一步说明（详见第 113-121 行）。

一方面，正如审稿人所言，“双减”政策包含的内容主要涉及减轻学生过重的作业负担。但即使“减轻”，作业依然占据着小学生大部分的学业时间（远高于考试的时间和频率）。换句话说也就是:“作业依然是我国义务教育阶段小学生最主要的学业任务（见第 113-114 行）”。

另一方面，为了响应“双减”政策，已有不少中小学开始禁止学校公布学生的考试成绩和排名，杜绝“唯分数论”。强调学校需要“提高作业管理水平”。说明国家的新政策在减少作业量的同时，也提出了对作业质量的要求。未来比起考试成绩，可能反而更强调重视小学生的作业等其他日常表现（见第 118-121 行）。考试反而变得不如作业重要。因此，对这片空白领域的探索研究其实是非常迫切的。

而对于小学生说谎行为的研究，其实已经有了一定的国内外研究基础(如 Bussey, 2010; Popliger et al., 2011; Lee, 2013; Talwar & Lee, 2002)。因此，开展目前几乎为空白的小学生作业作弊的研究，相对于研究说谎行为而言，也是非常迫切和重要。

参考文献：

Bussey, K. (2010). Lying and truthfulness: children's definitions, standards, and evaluative reactions. *Child Development, 63*(1).

Lee, K. (2013). Little liars: Development of verbal deception in children. *Child development perspectives, 7*(2), 91-96.

Popliger, M., Talwar, V., & Crossman, A. (2011). Predictors of children's prosocial lie-telling: motivation, socialization variables, and moral understanding. *Journal of Experimental Child Psychology, 110*(3), 373-392.

Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development, 26*(5), 436-444.

意见 3：研究使用了机器学习进行预测，也详尽说明了机器学习的方法学优势。但不清楚的是，如果使用较为传统的线性回归或二元回归，是否也有极高的准确率？和传统的线性回归方法比，机器学习的统计优势在哪里？是有更高的预测准确率，还是发现了传统的方法没有发现的预测变量？

回应：为了回答审稿人的这一问题，我们在此对机器学习及其较之线性回归、二元回归等传统分析方法的优势说明如下（由于文章篇幅有限，我们在正文中的介绍是相对精简的，如果审稿人觉得我们应该在正文中也进行与下面一样详细的说明，我们很乐意进行补充）。

首先，机器学习是一个总称，事实上它可以使用一系列不同的统计算法和方法，通过开发数学方程或模型来描述以及归纳手头的数据。这意味着我们可以同时使用很多机器学习算法（详见第 157-164 行），包括线性回归、逻辑回归、随机森林、XGboost、支持向量机、多层感知器和卷积神经网络等。其实，线性回归分析方法是研究人员开发并使用的第一个机器学习算法，之后又引入了逻辑回归算法。随后，新的机器学习算法层出不穷，这也为我们解决问题在算法/统计方法上带来了更多选择。

传统统计方法和现代机器学习方法的一个主要区别在于是否进行了交叉验证。传统方法(主要是一般线性模型中的线性回归和逻辑回归，或者更高级的广义线性模型)通常会一次性将手头所有数据投入分析，因此没有交叉验证的可能性（见 line148—156）。这种统计方法被广泛应用于心理学和许多其他学科中，但经常会出现过度拟合的问题，即模型对手头数据的预测效果良好，但却很难在一个新的数据集中推广(即效度低)。

为了克服这个问题，研究人员最初开发了 bootstrapping 算法，该算法倾向于使用 n-1 个数据点来构建一个模型，然后用之前闲置的一个数据点来测试这个模型，并且通过重复若干次上述过程 (重复次数等于或小于 n-1)得到最终模型。这种算法在机器学习语言中被称为“留一法”(leave-one-out approach)。现今，依靠 SPSS 软件就能够进行传统方法和留一法的分析。

留一法一度被认为远优于传统方法，因为它提供了一个交叉验证的可能性：不仅可以验证模型的性能是否依赖于某些特定的数据点，而且还可以估算模型在新数据中的泛化性。这是因为使用 n-1 模型，人们可以计算出模型预测的均值、标准差和总体均值 95%的置信区间(本质上是对模型的泛化性进行显著性检验)。然而，留一法有一个局限：只有 n-1 个测试可以用来测试模型的泛化性。

为了克服这个问题，研究人员开发了一种训练-测试方法。他们将数据按一定的比例随机划分成两个子集，一个子集用于训练模型(称为训练集)，一个子集用于测试模型(称为测试集)，测试集 n > 1。数据通过划分-重组-划分所能得到的数据集数量是指数级的，只要数据集足够大，研究人员就可以开发大量的模型，并且能够使用不同的测试数据点集来测试它们，一定程度上克服了留一法的局限性。例如，一个数据集有 80%被用于训练，20%被用于测试，得到一个模型后，再将训练集和测试集重新组合起来，并再次随机分割为 80%和 20%进行下一轮模型训练……这个过程被重复多次后，生成了多个模型，而后可针对多个不相同的测试集对这些模型进行测试，从而对模型的可泛化性进行更稳健的估计。然而，该方法也有一个局限：由于相同的数据被划分为不同的训练集被反复使用，训练的模型往往很相似，我们依然不知道训练的模型是否可以推广到一个全新的数据集中去(即外部效度)。

为了进一步解决这个问题，研究人员又开发了一套训练-测试-验证的方法。具体来说，首先将数据随机按一定比例分为三个子集：训练集、测试集和留出集。模型在训练集上训练，在测试集中测试，然后将训练集和测试集重新组合起来，并随机分成另一个训练集和另一个

测试集……在重复多次这个过程后，用事先预留的留出集数据对这些模型的预测力做进一步评估，即验证这些模型的预测力能否概化到新的数据中（见第 152-156 行）。通过使用留出集对模型进行验证，我们可以进一步确定模型的外部效度。本研究中采用的就是最后这种方法，这是目前机器学习研究中最常见的方法。

此外，通过机器学习，我们可以获得 Shapley 值，量化不同影响因素在整个模型中的相对重要性，这也是相较于传统的回归分析，机器学习非常关键的优势之一(详见 line165-174)。

意见 4：作者提到小学生的道德水平发展比较低，从结果来看，随着年龄增长（从 2 到 3 年级），小学生的道德水平应该是增长，但作弊行为也在增长，请作者根据研究的数据和结果讨论为何道德认知水平和作弊行为同时增长这一现象。
回应：感谢您的宝贵建议。事实上，我们最初认为，随着小学生道德水平的发展，其作弊率可能会越来越低。但结果却发现，在 3 年级之后，小学生的作业作弊率似乎趋于平缓。根据审稿人的建议，我们在文中对这一结果进行了补充讨论(详见第 472-473 行；第 476-478 行)。

意见 5：2 年级相比 3 年级等作弊水平较低，是否考虑了不同年级的作业频率和作业量？
回应：我们非常赞同您的这一观点，并据此在讨论部分进行了补充完善（见第 475-476 行）。

意见 6：作者提到，研究的一大意义是“即可输出该学生作业作弊的可能性（即作弊倾向）。依据这一结果，教师或家长便可对高作弊倾向的学生进行针对性的教育和干预。”
虽然研究本身很好回答了研究问题，但该研究也带来了一系列问题，对于作弊倾向，如何定义“高”？多高才是高？本研究所采用测试的信度（一致性）如何？学生回答问卷的分数，多少程度反映了稳定的个人特质（作弊、不受规矩、不诚信），多少反映了不稳定的场景影响（当时的心情、最近的成绩、同伴冲突）？施测一次，和施测多次，是否会有影响？
请作者回答：根据一次问卷的得分对学生进行“作弊倾向”分类，会不会给学生带来刻板印象和压力？家长会如何看待自己的孩子被归为“高作弊倾向”？教师会如何看待学生被分类为“高作弊倾向”“低作弊倾向”？
回应：我们非常赞同您的考虑，并对原文中的措辞进行了修改（见第 174-175 行；第 566 行；第 569-572 行）。具体来说，模型能够输出的是小学生参与作业作弊行为的可能性或概率（从 0%到 100%）。正如审稿人所言，我们并不能依据一次测试的结果就给学生“贴标签”。关于如何使用预测模型的结果，还需结合研究伦理和学校教育方式加以综合考虑。根据审稿人的建议，我们对文章讨论部分进行了相应的修改，具体详见第 569-572 行。

意见 7：与此相关的是，给社会个体分类和贴标签需要谨慎谨慎再谨慎，特别是此类有很强道德含义的和作弊、诚信相关的标签。请作者在讨论中对这些问题进行详尽的讨论，避免今后发表后可能带来的社会争议。
回应：谢谢审稿人的建议。针对可能存在的“贴标签”问题，我们已在讨论部分进行了详细说明，以避免这一问题（详见第 569-572 行）。

意见 8：对于说谎/诚实这一经典问题，来自哲学、心理学和神经科学的研究一直在争论，说谎是一时兴起，还是蓄谋已久？不知本研究的结果是否可以为这一经典问题提供一些观点？
回应：感谢您提出这个问题。我们深受启发，已在引言的理论创新部分增加了相应的内容（见第 554-558 行）。

事实上，说谎和作弊是两种不同的不诚信行为。前者是指个体有意图地通过一定的方式来操纵他人的错误信念的行为(Ding et al., 2015)，后者则是指个体通过破坏既定、强制性的社会规则来获取利益或赢得优势的行为(Green, 2004）。在过去的发展心理学研究中，研究者

更多的研究关注的是说谎行为。审稿人提到的“说谎是一时兴起，还是蓄谋已久”也是说谎领域的一个常见的争议性问题。

而对于作弊而言，同样存在一个类似的争议：作弊是情境性的还是特质性的。过去的研究似乎更加认可作弊是情境性而非特质性的(Hartshorne & May, 1928)。但在本研究中，我们发现影响小学生作业作弊最重要的因素分别有：小学生对作业作弊行为的可接受性、同伴的作业作弊行为，以及小学生自身的成绩水平。可见，不仅仅是情境因素，小学生对情境的认知对作业作弊行为也具有较大的影响。

参考文献：

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science, 26*(11), 1812-1821.

Green, S. P. (2004). Cheating. *Law and Philosophy, 23*(2), 137-185.

审稿人 2 意见：

意见 1：研究采用大样本问卷和机器学习相结合的方法考察了影响小学生作业作弊的因素及各因素的相对重要性。研究主要亮点在于研究方法上的创新。

回应：非常感谢审稿人的认可。

意见 2：但对于研究问题本身的创新与价值，以及支撑研究问题的理论框架仍需进一步思考和整理。首先，教育学和心理学围绕学业诚信（academic integrity/academic dishonesty）问题的国内外研究是很多的。许多学者围绕着人格特质因素、态度因素、动机因素对学业不端的潜在影响开展了一系列研究工作（可参见 Lee, Kuncel, & Gau, 2020 的元分析）。但作者没能很好地从这些文献中梳理出一个可靠的理论模型来进行验证性地分析，而是采用机器学习的方法对“可能的”模型进行了探索性的数据拟合，虽然拟合效果看起来不错，但整体研究看上去理论支撑不足。具体表现为，在进行文献综述时，作者只是泛泛地例举了一些现有研究成果，因此变量间的逻辑关系不清。例如，作者提出了三个主要方面的预测因素：作弊后果的严重性；行为可接受性和同伴行为；以及对预防措施有效性的评价。在提出第一个变量的时候提到了“理性行动理论”，但在提出第二和第三套变量的时候并没有足够的理论支撑。很明显，“同伴行为”应属社会规范里的描述性规范，但“作弊行为的可接受性”为什么要和“同伴行为”放在一起作为第二个方面呢？而第三个方面的预测性因素“预防措施有效性评价”又是从什么理论衍生出来的，和其他两方面的变量之间的关系是怎样的呢？对于这些问题的回答，都需要作者从已有文献中梳理出理论框架，以作为变量间逻辑关系的支撑。

回应：谢谢审稿人的宝贵建议。我们已经对文章引言部分的理论框架进行了加强（详见第 176-225 行）。具体来说，本研究主要以 Murdock 和 Anderman(2006)的作弊动机理论模型为框架。该模型框架主要针对大中学生群体提出，是现今为止最具影响力的作弊动机模型之一。在该模型框架下，我们结合以往相关实证研究结果和有关研究实践意义的考虑，对本研究所考察的有关个体作弊的预测变量进行了进一步筛选。这些预测变量的设置，可为进一步检验 Murdock 和 Anderman(2006)的作弊动机模型是否适用于小学生群体及其作业作弊行为提供了新的科学依据。

本研究结果发现，Murdock 和 Anderman(2006)提出的一些影响大中学生作弊行为的关键要素能够预测小学生的作业作弊——个体自身对作业作弊行为的接受性、同伴作业作弊的普遍性和发生频率，及其自身的成绩水平是预测小学生作业作弊的关键变量。

然而，本研究也发现，另一个被 Murdock 和 Anderman (2006)认为能够预测大学生考试作弊

的至关重要的作弊动机——后果严重性，对小学生作业作弊的影响力却远不如预期之大。可见小学生群体具有其发展的特殊性，影响大学生考试作弊的主要因素与影响小学生作业作弊的主要因素并不完全相同（详见第 546-549 行）。

意见 3：在文献综述时，作者指出“针对小学生作业作弊的实证研究尚属空白”（line129-131），但后面作者提出的三个主要变量也是基于针对中学生和大学生样本的大量研究结果而提出的，并没有针对小学生特点的变量，这减弱了研究的创新性和价值。

回应：为了解决审稿人提出这一问题，我们在讨论部分进一步明确了本研究的理论创新(见第 541-542 行及第 546-549 行)，以及有关未来在理论上可以进一步研究和开发的展望(见第 591-596 行)。

由于 “针对小学生作业作弊的实证研究尚属空白”，我们在教育等实践工作中只能借鉴那些针对大学生和中学生的研究成果。因此，本研究的理论价值在于，验证这些对于大中学生来说非常重要的变量，是否也对小学生有影响（见第 541-544 行）。比如，结果发现，“年级”这一变量对小学生作业作弊具有较强预测作用，但这个变量对大学生和中学生学业作弊行为中却不存在显著效应（见第 552-553 行）。

意见 4：在采集被试时，作者表示本研究通过了所在高校学术伦理委员会的伦理审查，但未对研究对象的伦理问题进行说明，例如是否获得小学生及小学生父母的知情同意书等等。尤其是作为一个高社会赞许性的研究问题，在询问小学生是否有作弊（同伴作弊）行为后，可能会让他们产生消极不安情绪，作者应对这方面的处理手段进行一定的说明。

回应：谢谢审稿人的提醒。关于知情同意，我们在正文中进行了补充说明（见第 233-234 行）。对于被试的不安情绪，我们在正文中也进行了说明：本研究问卷调查采用匿名形式，并且在发放、填写和回收问卷的整个过程中，学生所熟悉的班主任或任课老师均全程不在场、不参与（见第 268-269 行）。

意见 5：最后，我虽然对于机器学习的方法不了解，但通过作者的描述大概理解为对机器学习的四种算法结果进行集成后，最终模型的预测率可达 80%。这虽然看起来是一个较好的结果，但我仍存在两个主要疑问。首先，较好的预测率看起来是数据导向而不是理论导向的，即计算机通过算法将数据进行了很好的拟合，但如果理论支撑不足的话，拟合效果很好的数据是否能够排除其他变量（例如指令性规范、作弊动机等）作为更有效预测变量的可能？其次，这种算法与一般的线性回归方程、结构方程的本质区别是什么？或者说通过简单线性回归或结构方程是否也能得到同样的结论？

回应：根据您的宝贵建议，我们已重新对理论部分进行了梳理（详见第 176-225 行）。基于这些理论基础，我们选出了本研究所考察的预测变量，并通过机器学习获得了预测率达到了80%的最终模型。虽然 80%的预测准确性已经很高，但与 100%相比仍有 20%的差距。这表明，还有其他可能显著预测变量是本研究尚未涉及到并值得未来进一步研究的。对此我们在文章讨论部分进行了补充说明（见第 590-596 行）。

此外，关于“机器学习算法与一般线性回归方程等的区别在哪里”这一问题，由于我们已经在回答审稿人 1 的问题（2）中进行了详细的说明，故不在此一一赘述了。概括说来，线性回归是研究人员开发并使用的第一个机器学习算法，逻辑回归算法等也都是机器学习的算法之一。与传统的回归相比，在机器学习中的这些算法采用的是交叉验证的逻辑：将整体数据按比例进行子集划分（训练集、测试集、留出集），先用训练集数据训练模型，再用测试集验证模型，最后用留出集验证模型的外部效度，通过重复一百遍上述过程获得最终模型。这种交叉验证的方法，不但能够解决过度拟合的问题，也能够提高模型的外部效度(详见 line148-156 行)。

意见 6：综上，作为一项心理学研究，理论支撑对于问题的提出而言是很重要的，作者应着重对这一部分进行凝练和补充。
回应：再次感谢您的宝贵建议，对此我们已在引言有关理论创新部分进行了完善（见第 176-225 行）。

……

审稿人 3 意见：

意见 1：本研究聚焦小学生作业作弊行为，设计了评分量表，进行了较为广泛的调查问卷。此外，本研究采用了机器学习的方法，建立了分类模型，对不同小学生作业作弊的可能性进行了较为准确的预测；并基于这种方法提取了对于预测作业作弊可能性具有较大贡献的特征。总的来说，本研究的思路较为新颖，写作较为规范，使用的数据分析分析方法合理，得到的发现有望为儿童诚信行为发展的理论构建以及学业作弊的早期干预提供一定的科学依据。但是还有一些问题需要进一步解释清楚。
回应：非常感谢审稿人对本文的肯定。

意见 2：摘要部分过于简单。如果没有字数限制的话，作者应该重新组织摘要，要将研究背景、目的、方法以及结果与结论简明扼要地介绍清楚。
回应：200 字是心理学报对论文摘要的字数要求。我们很同意审稿人的意见。若编辑部允许，我们很乐意对摘要进行扩充，以最大程度地并用最简洁的文字对文章的背景、目的、方法和结果与结论进行介绍。

意见 3：建议将被试的人口学特征也已表的方式介绍。
回应：感谢审稿人的宝贵建议，我们已将人口统计学特征整理成了表 1（见第 242 行）。

意见 4：AUC 主要用于描述分类器的综合分类性能，而敏感性和特异性则用来评估在某一特定风险阈值设置下的具体分类性能，因此“用于评估模型的敏感性(即其在多大程度上能够准确预测“存在作业作弊行为”这一情况)和特异性(即其在多大程度上能够准确预测“不存在作业作弊行为”这一情况)。”以及“模型总体具有较高的敏感性和特异性(1－假阳性率)”等语句描述不准确。
回应：针对审稿人提出的这一问题，我们在正文中对 ROC 和 AUC 进行了补充说明（见第 359-367 行）。正如审稿人所言，AUC 是帮助我们将机器学习分类器的性能进行可视化的重要指标，它实际上是 ROC 曲线所覆盖的区域面积。

ROC 全称为受试者工作特征曲线，它表现了分类模型的特征，可以衡量机器学习分类模型的好坏。具体来说，ROC 是通过真阳率（True Positive Rate，正确预测出的阳性的数量/所有阳性的数量）和假阳率（False Positive Rate，将阴性误判为阳性的数量/所有阴性的数量）绘制出的曲线（见图 1，第 368-371 行）。真阳率越高，假阳率越低，ROC 曲线越优，曲线下面积（即 AUC）越大，分类器分类效果越好。

其中，敏感性即真阳率，特异性为 1-假阳率（即正确预测出的阴性的数量/所有阴性的数量）。敏感性和特异性之间存在着一种函数关系，即在不同敏感性条件下，特异性会朝相反的方向变化。当 ROC 曲线在对角线位置时，两者达到最优组合。需要说明的是，在实际应用过程中，因具体应用情境不同，在 ROC 曲线上以多少的敏感性和特异性作为标准，需要视实际情况来决定。

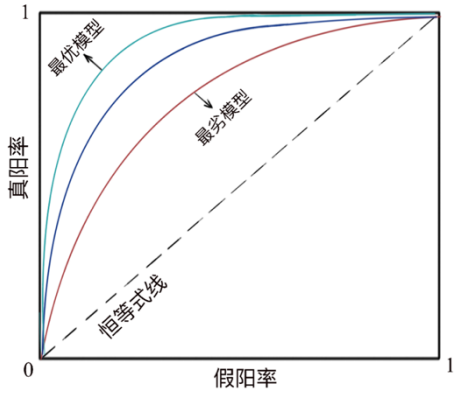

图 1 计算机模型的受试者工作特征曲线(ROC)描绘了模型敏感性(真阳率)随着特异性(假阳率)变化的规律。曲线下的面积(AUC)表示模型的整体性能，曲线距离恒等式线(虚线对角线)越远，代表模型的分类能力越好，反之则越差。

意见 5：图 2 所示，由于但 3、4、5、6 四个年级间的作弊率两两差异不显著(ps > 0.05)，而且五年级作弊了还略有下降，因此不能笼统地说“作弊行为的发生率大致呈现随年级增长而上升的趋势”。

回应：非常同意审稿人提出的这个问题，我们已对文中的相应表述进行了修改（见第 390 行及 第 472-473 行）。

意见 6：机器学习模型建立的描述中一些细节没有详细介绍，比如：如何进行参数调优、总共有多少特征输入等。

回应：针对审稿人提出的这一问题，我们已在正文中补充了有关特征输入、参数设置的说明。其中，关于特征输入见第 338-342 行。关于参数设置见第 324-334 行（补充见第 351-354 行）。

需要说明的是，我们采用的是默认参数，并未对参数进一步调优。这是因为本研究旨在检测机器学习法在领域内的可行性。当然，若想在实践中应用最终模型，未来仍然可以对参数进行进一步调优（目前 AUC 准确率为 80%左右，除了对可能的预测因素进行进一步探索外，还可通过参数调优来进一步提高准确率）。在调优的过程中，需要进一步收集数据，再增加一个外部验证的数据集，进而使模型的 AUC 值进一步提高。对此，我们也在文中的讨论部分进行了相应补充（见第 596-599 行）。

意见 7：在集成模型的建立中，为什么将测试集的结果也作为输入？在训练集成模型时，训练集、测试集和留出集的样本比例是多少？

回应：针对审稿人提出的这一问题，我们在文中进行了一些补充说明（详见第 351-354 行）。集成模型的基本步骤与其他四种算法完全一致，但集成学习法的内部逻辑是：先将训练集的数据通过四种方法进行训练，然后对四种模型训练的结果按照 Stacking 法进行整合，而后再进行测试集验证和留出集验证（整个流程由软件后台直接完成，并不会中途输出可视化数据）。所以，集成算法的三个子集的比例分布也与其他几种算法一致：训练集占总被试量的 64%，测验集占 16%，留出集占 20%。

意见 8：图 3，建议对 4 种不同的机器学习模型分别画出对应的 ROC

回应：谢谢审稿人的建议，我们已在文中增加了其他几种算法的 ROC 曲线（见第 426-430 行，图 3）。

第二轮

审稿人 1 意见：

仔细读了修改稿，作者的修改和回复很好地回答了我之前的问题。我没有其他问题了，推荐发表。

回应：非常感谢您的宝贵建议。

审稿人 2 意见：

意见 1：作者较好的回答了我提出的问题，但随之而来产生了一个新的问题那就是：研究在方法（确切的说是算法）上确实有较强的创新性和应用价值，但是在研究的理论价值上只能说是部分验证了传统的理论模型，因而显得理论价值不足。总体而言，我同意作者的文章发表，但希望作者能够更深入的思考并完善其研究的理论价值。毕竟，这并不是一篇单纯的方法论的文章。

回应：非常感谢您宝贵建议。针对您所提出的以上问题，我们在讨论部分对本研究的理论价值进行了补充说明(详见 line 538-559)。概括来说，本研究的理论创新主要体现在：**初步建立了关于小学生作业作弊影响因素的综合模型，并发现小学生作业作弊的影响因素具有特殊性，与大中学生学业作弊行为的影响因素构成并不完全相同；此外，本研究为解决作弊是情境驱动还是个体特质驱动这一长期以来存在的学术争论提供了新的依据。**

与此同时，我们还在研究局限部分对未来开展有关理论方面的考察进行了展望(详见 line 593-598)：**以本研究为起点，未来研究可以通过更充分地挖掘影响小学生作弊的因素，来最终建立一个科学而完备的、专门适用于小学生作业作弊的理论模型**

意见 2：此外，作者需要修改英文摘要（包括一些语法问题），使中英文摘要对应。

回应：感谢您的建议。我们已根据学报的要求（中文摘要在 300 字以内；英文摘要总字数不少于 500 字，按 4 个标题分 4 个自然段写出，其中背景部分不得超过 200 字，方法部分不少于 100 字，结果部分不少于 150 字，结论部分不得超过 150 字），对英文摘要进行了修改。

审稿人 3 意见：

作者已经正确地回答了我上次提出的问题，建议发表。

回应：非常感谢您的宝贵建议。

编委意见：

比较完善了。建议修改后同意发表。

主编意见：

文章经过多轮修改，已经达到学报发表要求。同意发表。谢谢审稿专家和作者的辛勤付出！