《心理学报》审稿意见与作者回应

题目：开放式情境判断测验的自动化评分

作者：徐静；骆方；马彦珍；胡路明；田雪涛

第一轮

审稿人 1 意见：

文章开发了开放式情境判断测验，探索了自动化评分的应用。研究问题较为前沿，方法选择得当。针对开放式情境测验自动化评分的分析，目前研究尚处于起步阶段。文章所采用的研究方法也能恰当地自动化评分的问题，具有一定的创新。建议文章聚焦研究问题，做出以下修改和调整，以便更好指导其他研究者。

意见 1：就整体研究问题而言，研究者所关注的重点为“开放式情境判断测验的自动化评分”，涉及三个研究问题，分别为：（1）评分标准的问题；（2）自动化评分的问题；（3）自动化评分的解释性和效度验证问题。建议作者聚焦文章的研究目的，对文章结构作出适当调整。在引言部分，建议对开放式情境判断测验“设置评分标准的方法”、“自动化评分的方法”、“评分的解释性和效度验证”等领域的文献做出阐述和补充，在此基础上提出本文的研究问题。在研究 1 和研究 2 部分，建议将研究 1 中测验的编制部分作为文章的研究材料，且需详细说明测验编制的过程和信效度。并将测验评分规则的制定作为研究结果的第一点。不再拆分研究 1 和研究 2。讨论部分需要继续加强，引用适当的文献，围绕文章聚焦的研究问题进行展开说明。结论部分需要更加聚焦文章研究问题所得到的研究结论。

回应：非常感谢审稿专家细致认真的审阅，谢谢您的宝贵意见和建议！您的建议对我们提升文章的逻辑性有非常大的帮助，我们参照您给出的意见，通读全文，对文章结构做了调整。

意见 2：摘要中的关键词提到“自动短答案评分”，但是在文章中并没有出现这个专业术语。建议文章统一专业术语，在相关概念出现时做出解释，在后期避免出现新的概念，这样会降低读者的阅读连贯性。

回应：非常感谢审稿专家细心地指出这个问题。“自动短答案评分”确实存在术语不统一问题，其在原文中的阐释如下(原文 p.4)：“根据文本长度，目前主观题自动评分主要可分为 2 大类型，长文本类型如自动论文评估(Automated Essay Scoring, AES)，短文本类型如短答案自动评分(Automatic Short Answer Grading, ASAG)，书面回答式的开放式测验评分介于前二者之间”。关于这一类型问题，学者叫法不一，有“短答案自动评分”、“自动短答案评分”、

“简答题自动评分”等，为了方便读者阅读，我们将这一术语在全文中统一表述为“简答题自动评分”。

由于这一概念对厘清研究问题比较重要，而在原文 p.4 中引言部分的辨析确实不甚清晰，因此调整了这部分的文字阐释，使其更加清晰易读，分为三部分来解释：作文自动评分、简答题自动评分、开放式 SJT 的自动评分，这三部分评分方式不尽相同。

此外，我们再次检查了全文其余部分，将评分规则中的反应项统一表述为“行为反应项”，代表某情境下的某一行为，方便读者阅读和理解。

参考文献：

Burrows, S., Gurevych, I. & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *Int J Artif Intell Educ, 25*, 60–117.

意见 3：第 7 页的测验编制部分，如果问卷不能公布，建议对编制题目的问题情境和选项等作出举例说明。第 3 段话中，为什么仅对最优和最差的反应进行评定？题目难度均值为 0.66。为什么提供题目难度，目的是什么？

回应：十分感谢审稿专家的意见。

（1）根据您的建议，我们在文章附录部分增加了题目修订阶段中专家评定的示例，以及正式施测版本中一道题目的示例，并且展示了这道题的评分规则。

（2）对于您提出的第二个问题（“为什么仅对最优和最差的反应进行评定？”），在原文 p.7，在测验初步编制完成后，一轮修改以及一线教学专家评定的主要目的有三：①情境是否真实、贴合教学实际；②是否有表述不清或歧义；③采用带选项的传统 SJT，被试是否存在优势作答倾向。

“由 8 名心理测量学领域研究者对……以及最优和最差的反应项进行评定。”主要是为了防止被试猜出明显更“正确”的选项，如果研究团队的组员一致认为某选项是最优反应项，则可能代表此题还有很强的猜测性，开发者会对选项的文字表述和做法进行修改。

完成这样的一轮修改之后，由一线教学专家评定最优做法和最差做法，是想验证被试存在优势作答倾向。题干采用的是“如果是你，你实际最有可能做什么？”这样的行为倾向指导语，结果发现被试实际选项和其认为的最优做法选项一致性程度高，一定程度上也说明了被试的优势作答倾向，也说明了采用开放式情境判断测验的必要性。

考虑到上述内容在我们之前的行文中没有进行充分的说明，这可能对读者理解这部分内容造成困难。因此，我们在文中对该内容作了补充。

（3）对于第三个问题（“为什么提供题目难度，目的是什么？”）对于 SJT 题型来说，难度这一指标的意义不大，在原文中将难度删去。

意见 4：第 8 页 2.2.2 研究工具中，课堂评价量表从四个维度进行评价，这里有相关依据吗？

需要提供参考文献。2.2.3 中，第二段话，“从行为的有效性、逻辑性、全面性、合理性维度确定分值(0~3)”，这里的“有效性、逻辑性、全面性、合理性”是如何体现的？

回应：感谢审稿专家指出问题，这些部分在原文中确实表述不够清晰。

（1）在正文 p.8 中补充表述为：“依据课堂观察量表(凌晨，2020)，从课堂管理、教学内容、思维培养、情感关注 4 个维度对教学视频进行评价，该量表内部 α 系数为 0.83，验证性因素分析结果为 CFI=0.897，TLI=0.868”。

（2）对于您提出的第二个问题(“有效性、逻辑性、全面性、合理性”是如何体现的？”)，我们在文中 p.9 做了如下补充：“为行为反应项赋予分值。依据作答结果与胜任力模型即测验维度的吻合程度来评分，为更加贴近胜任力特征的回答赋予更高分值，两名评分员结合本题维度，同时关注行为的丰富度、具体性、全面性、逻辑顺序等所体现的思维水平和能力的差异，对各个反应项赋分，采用 3 分制(0 ~ 3 分)，1=差，3=优秀，0=偏题或无效作答。赋分环节中最重要的一点是分数公平性与合理性，权重分数需能够有效地体现行为反应项的差异。每道题的赋分皆在两名评分员的讨论后确定，直至达成一致性评价。”

参考文献：

凌晨. (2020). *课堂观察量表的开发——促进初任教师专业发展* (硕士学位论文). 北京师范大学.

意见 5：第 9 页中，胜任力总分与教学设计、课堂评价和学生作业呈现显著相关。相关系数最高为 0.263，显著性可能是样本量导致的。此处需要做出解释。

回应：十分感谢审稿专家的意见。完整提交有效教学材料的样本量为 181，胜任力总分虽与三个效标皆呈显著相关，但相关系数皆不大(教学设计 r=0.263，课堂评价 r=0.203，学生作业 r=0.223)。但在我们的研究中，我们认为大体上是可以接受的。McDaniel 等(2001)进行了第一个关于情境判断测验效标关联效度的元分析，检索了 1887-2000 年 95 个实证研究，情境判断测验分数与工作绩效之间的相关是 0.3(校正后)。McDaniel 等(2007)在 118 个研究的基础上再次对情境判断测验进行了元分析，在不考虑指导语类型的前提下，效标关联效度系数为 0.26。O’Connell 等(2007)研究了情境判断测验与绩效的相关，结果表明，与任务绩效的相关为 0.14，与关系绩效的相关为 0.10。或许是由于测验维度本身很难做到单维，情境判断测验这种测验类型在效标关联效度上无法达到过高的相关。

然而，您的提问也给我们以启发，或许在效标选取上我们还可以做深一步地探索，我们选取的教学表现的效标与教师胜任力对应的不是很贴切，其中可能还是有一些变异存在，而由专家对教学表现评价，其中也难免存在偏误，在这样的情况下，还能够得到显著的相关系数，也是能够接受的结果。我们将有关效标的思考也补充在了讨论部分之中。

参考文献：

刘晓梅，卞冉，车宏生，王丽娜，邵燕萍. (2011). 情境判断测验的效度研究述评. *心理科学进展, 19*(5), 740–748.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M.A., & Braverman, E.P. (2001).Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730–740.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63–91.

O'Connell, M. S., McDaniel, M. A., Grubb, W. L., Hartman, N. S., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*, 19–29.

意见 6：第 9 页研究 2 中，需要对有监督的文本分类流程做出解释。

回应：感谢审稿专家对此做出提示。在正文 p.11 中就“有监督的文本分类流程”补充了以下说明文字：有监督的文本分类流程包括两个阶段，第一个阶段是在有标签的训练数据上进行模型训练，第二个阶段是应用训练好的模型对无标签的测试数据进行预测并进行性能评估。在两个阶段中，文本数据需要进行相同的预处理和特征提取操作，例如去停用词、统计词频等，从而获取计算机可直接计算的数值型文本表征。所训练的分类模型可以看作从文本表征到分类标签的映射函数，通过指定的机器学习算法进行训练得到，并实现对待分类文本所关联的标签做出预测。

意见 7：第 10 页中，为什么会采用文档层面的多标签文本分类，文章前后内容中均没有涉及。为什么只用了第一题进行尝试？

回应：感谢审稿专家的提问。根据具体任务的输入输出形式，自动化评分建模可以有多种建模思路，在实践中一般需进行多种尝试并选取更加简单有效的建模方法。本研究中输入为被试的作答文档，输出为该文档涉及的多个反应项或多级评分，这种输入输出形式可以直接对应机器学习领域的多标签分类任务，因此研究首先尝试了文档层面的多标签文本分类。这种建模方法没有引入句子级别的标注信息，如果能够达到可用的性能可优先使用。然而，在实践中，多标签分类结果欠佳，文章仅以第一题为例说明了这个过程，在后续的研究中认为句子级的自动化评分能够取得更有效的结果，采用了这种思路，并在文章中进行了详细的描述。

意见 8：第 11 页中，建议句子层面文本多分类的结果能够结合具体案例进行解释。

回应：感谢审稿专家提出的建议。我们在文档的附录部分增加了一道题目的示例，并且展示了这道题的评分规则，对句子层面文本多分类的逻辑作了说明，以方便读者理解和参考。

意见 9：第 15 页讨论中，提到“将带有权重的关键行为反应作为评分标准”，前面内容中没有地方提到是如何进行加权的。

回应：感谢审稿专家的提问。原文 p.8 评分规则中“其次，赋分。每个反应项由两名评分员从行为的有效性、逻辑性、全面性、合理性维度确定分值(0～3 分)，采用 3 分制，1=差，3=优秀，0=偏题或无效作答。除了反应项评分，还设定了综合评级，每道题的作答文本往往包含多个反应项，加和合成总分后，依据百分位数换算成等级分数，前 27%等级分数为 3 分，后 27%为 1 分。”

原文 p.17“评分的核心主要包含以下几个步骤：首先，专家对反应项编码聚类……其次，为反应项赋予权重分值。依据作答结果与胜任力模型即测验维度的吻合程度来评分，为更加贴近胜任力特征的回答赋予更高分值，以得分高低来体现作答者的胜任水平差异。赋分环节中最重要的一点是分数公平性与合理性，权重分数需能够有效地体现反应项的差异。”

因此，“将带有权重的关键行为反应作为评分标准”可以理解成：评分标准分为 2 部分，一，确定行为反应项；二，为这些反应项赋予 0-3 分作为权重。

您的问题使我们意识到，评分规则这一部分应当是我们重点向读者阐述的部分，而原文在研究方法和讨论中都有涉及，却表述不甚清晰，因此，我们修改了关于评分规则的表述，详见正文 p.9、p.13。

意见 10：在 17 页中，4.2 部分，什么是“细粒度的评分模型”？建议文章中的专业术语要统一。4.2 中提到文章中使用了自上而下和自下而上相结合的方法。这种方法在前面内容中是如何体现的。

回应：感谢审稿专家的提问。

（1）对于第一个问题（什么是“细粒度的评分模型”），传统的机器通过对训练集的学习，完成对每道题的打分，却并不能回答为什么这样的回答比那样的回答分数更高。细粒度的评分模型突出了本研究的特色，我们关注作答者的不同做法，建立起更加具体的行为层面的评分规则，能够更细致刻画被试的行为差异，是更细粒度的评分模型，也更具有可解释性。根据您的建议，我们检查了文中的专业术语，对于同一概念统一了术语表达。

（2）对于第二个问题（自上而下和自下而上相结合的方法在前面内容中是如何体现的），做出回应如下：

自上而下：“一方面，依赖专家确定评分规则，人工编码评分作为机评的有监督数据，机评以人工评分为标准做验证”这一部分为前文自动评分总体流程中的内容；

自下而上：“另一方面，也从数据出发，机评结果能够为题目设计、评分规则提供修正思路，从而不断完善题目以及评分模型。”运用同样的分类模型，各题上的评分准确率不同，准确率若不理想，我们可以去发现一些问题：

①评分规则是否需要修正？比如，是否类别过多或过少、或是存在某些类别赋分不合理。

②题目本身是否存在问题？比如，第 1、3、10 题是否题目质量不高，存在歧义等。

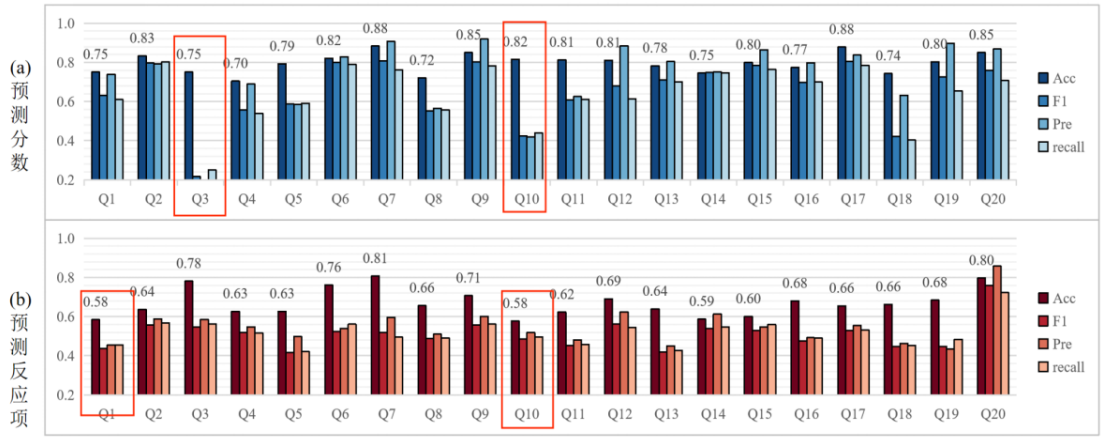

图 1 CNN 在 20 道题上的结果图

③题目维度是否设定不合理？比如某一维度人机评分的相关过低，也可考虑是否是维度设置的不合理，把不属于本维度的题包含进来了。

本研究想通过这样的尝试给其他研究者带来启发，题目开发、评分规则、评分性能这些环节紧密相关，互受影响，此问题属于心理测量和计算机的交叉领域，测验和机器评分不可割裂，需要同时对心理测量和文本分类有深入的理解。

意见 11：部分文字表达不准确。例如，第 3 页引言第 1 段话中“适合测量非认知能力如实践智力、内隐知识(Sternberg et al., 1993)”，放在现在这个时期已经不合理。目前，对于学生高级思维能力的测量，也更多采用真实情境问题的测验。第 4 页第 3 段中提到，“PEG 主要用于评估文本的写作风格评价，未考虑语义信息”，而后面又提到“不仅能如 PEG 那样评判语言质量”。“文本的写作风格评价”和“语言质量”等价吗？

回应：非常感谢您在文字表述方面提出的建议，我们做了如下更改：

（1）已在正文中删去“适合测量非认知能力如实践智力、内隐知识(Sternberg et al., 1993)”这一表述；

（2）“PEG 评判语言质量”确实表述不明，与“文本写作风格”表达含义不同，因此将“语言质量”更正为“语言风格”。

意见 12：文章还存在一些语句的小错误，请再次通读全文，并修正。例如，文章第 3 页，最后一行中“情境判断测验能在一定程度上能降低作假的影响”多了“能”。

回应：十分感谢审稿专家对本文全面、精准而细致的审阅，感谢您提出的意见和建议。我们再次通读了全文，检查语法、错别字等错误。再次感谢您对本文的辛苦审阅！

审稿人 2 意见：

研究以教师胜任力测评为例，探索了开放式情境测验的开发和自动化评分的应用，结果证明从句子层面应用卷积神经网络的效果较好，具有较高的评分准确率和人机评分一致性。研究为自动评分方法的应用做出了有益探索，具有一定实践价值，但是显得创新性不足。具体意见如下：

意见 1：研究关注的重点是自动化评分，为什么又要做测验开发的研究而不用已有的测验？在问题提出部分没有很好的阐述做测验开发（研究一）的原因。即，前面并未提到 SJT 在开发测验上也有局限性，为什么把这个也作为研究的一部分？

回应：非常感谢审稿专家的审阅！

（1）对于您提出的第一个问题（“为什么要做测验开发的研究而不用已有的测验？”），这个问题想必也能代表着读者的疑问。诚如您所说，如果本研究是评分问题，那么文章的重点在于评分的准确性究竟有多高、是否达到甚至超过人类评分员的一致性、是否能够代替部分人力工作。我们看到计算机学科的自动评分问题重点也在于此，即通过算法的改进来提升评分的准确性。

但不同于批改一道数学题、不同于评价一篇文章是否优秀这样的问题，前者重点在于作答文本与参考答案之间的相似程度有多高、后者在于评价文本的结构、风格、语言习惯，而本研究并不是一个单纯的计算机评分问题，而是基于心理测量的准则来设计题目和计分的，前后连根共树，不可分割。

作为心理测验中人格类测验问题，在评分准确性达到可用的标准之上，我们关注的是文本中体现出的人格特质是什么，而人格特质本身是无好坏之分的，我们评价的是文本中体现出的某些特质是否属于我们所测评的特质。具体而言，本研究针对教师胜任力问题，采用情境判断测验这种形式，评价作答文本与教师胜任力模型之间的匹配度，分数越高，则作答者的胜任水平越高，可能更加适合从事教师行业。

那么作答文本与胜任模型是如何建立联系的呢？本研究给出的答案是通过“行为反应”这样一个桥梁（为了方便行文，文中表述为“行为反应项”），每道题单独设定一个评分规则，这个规则列举了当前情境下所有可能的行为反应，符合本道题测评维度的行为反应项被赋予更高权重。评分大体逻辑为：作答文本——这段文本包含了哪几种行为表现——对应评分规则——得到本道题包含的若干行为表现对应的分数——单道题分数合成，转成标准分——加和、合成整套测验分数。可见，需在对测验维度有深入理解后才能更好建立评分模型，并不单纯是计算机简单的文本处理问题。

另一方面，我们也试图从心理测量者的研究视角，探索一种情境判断测验的开放式形式，并通过自动评分的方式来解决评分工作量大的问题，鼓励实践领域更多的研究者借助智能化的文本挖掘方式达成自主测评目标。

（2）对于您提到的第二点（“在问题提出部分没有很好的阐述做测验开发的原因”），在引言部分补充了 SJT 在开发测验上的局限性：“另一方面，从测验开发技术的角度来看，情境判断测验的选项编制有诸多标准：选项需能激活特定构想；对不同特质的人群具有区分度；尽可能涵盖所有被试反应；选项之间无较大差别以防止社会称许性和猜测；计分键科学，各选项分值之差能够代表能力水平的差别等(正文 p.4)。”

参考文献：

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 730–740.

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: a warning concerning warning statements. *Human Performance*, *21*(1), 89–106.

Lievens, F., De Corte, W. & Westerveld, L. (2015). Understanding the building blocks of selection procedures: effects of response fidelity on performance and validity. *Journal of Management*, *41*(6), 1604–1627.

Marentette, B. J., Meyers, L. S., Hurtz, G. M., & Kuang, D. C. (2012). Order effects on situational judgment test items: A case of construct-irrelevant difficulty. *International Journal of Selection and Assessment*, *20*(3), 319–332.

意见 2：问题提出部分不够聚焦。例如，p.3 最后一段，应当讲封闭式作答形式的缺陷，中间又阐述了讲情景判断测验在降低作假影响上的优势“尽管相比量表题，情境判断测验能在一定程度上能降低作假的影响，但在选拔的情况下仍然存在作假行为(刘晓梅 等, 2011)。”再例如，文中提到的评分问题阻碍了开放式情境判断测验的发展，“主要困难包括：（1）评分标准的制定。（2）自动评分问题。（3）自动评分的解释与效度验证问题。”。（3）也属于自动评分问题（或者（2）应当总结为自动评分的实现？）。但其实本研究只解决了后两个问题。对第一个问题的考虑是什么？

回应：感谢审稿专家的提问，您的意见帮助我们重新梳理了问题提出部分的文字表述和语言逻辑。

（1）对于您提到的“p.3 最后一段中讲封闭式作答形式的缺陷却又提到了其优势”，本段确实重在讲封闭式作答形式的缺陷，即使 SJT 一定程度上能降低作假，但是在选拔的情况之下 SJT 的作假问题仍然是需要解决的问题。但是诚如您所说，上述内容在我们之前的行文中不够聚焦，可能对读者梳理和理解论文内容造成影响。为了解决这一问题，我们修改了引言中这部分的表述，使其更加清晰：“封闭式方便标准化处理和快速计分，是目前主流的测验形式。然而，这类选择题易受个体作答态度、猜测和应试策略的影响，测验结果有时不能反映被试真实的能力水平，研究者证明了在选拔中作答者基于相关动机有意识地做出偏好或期望反应，选择最优项而非实际做法(McDaniel et al., 2001; Robson et al., 2007)(正文 p.4)。

（2）第二点“自动评分问题”，更确切地说是机器自动评分过程的实现，是一个输入文本输出分数的过程。根据您的建议，我们在正文中将其修改为“自动评分的实现”。

（3）关于评分标准问题，体现在原文 p.8 评分规则、原文 p.17“评分的核心主要包含以下几个步骤：首先，专家对反应项编码聚类……其次，为反应项赋予权重分值……”为了更加清晰地表述评分标准的制定(评分规则)，我们将此部分进行更清晰的阐释，详见正文 p9、p.13。

参考文献：

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*(4), 730–740.

Robson, S. M., Jones, A., & Abraham, J. (2007). Personality, faking, and convergent validity: a warning concerning warning statements. *Human Performance*, *21*(1), 89–106.

意见 3：文章第二段介绍了开放式情境判断测验的几种类型，后面讲到自动评分可以书面回答式，其他类型无法应用自动评分吗？

回应：感谢审稿专家的提问。

对于书面回答式(written-constructed)的自动评分：被试的作答是书面文本。

对于视听构建式(audio-visual constructed)与情景面试(Oostrom et al., 2010, 2011, 2012)的自动评分：被试的作答可以是口头表达也可以表演出来，只是后者增加了主考官与被试的互动(面对面或网络视频连线)。

事实上，在查阅国内外相关文献时，即使是书面回答式的题型，也并未有一种较好的、适合情境判断测验这类题型的自动评分的方式。原文中提到：“尽管随着技术的进步，越来越多的研究者开始探索开放式的作答形式，但目前的研究仍处于起步阶段(Cucina et al., 2015)。如 Lievens 等(2019)尝试了书面回答式和视听构建式的情境判断测验，但此研究仍采用人工评分的方式。”

本研究探索了对文本自动评分的可行性，而对于语音、肢体动作也可以实现自动评分，如今，一些研究者探索了 AI 面试，其技术实现方式可以为情境判断测验的语音、肢体等数据的自动评分提供借鉴。

参考文献：

Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T.(2010). Webcam testing: Validation of an innovative open-ended multimedia test. *European Journal of Work and Organizational Psychology*, *19*, 532–550.

Oostrom, J. K., Born, M. Ph., Serlie, A. W., & Van der Molen, H. T. (2011). A multimedia situational test with a constructed-response format: Its relationship with personality, cognitive ability, job experience, and academic performance. *Journal of Personnel Psychology*, *10*, 78–88.

Oostrom, J. K., Born, M. P., Serlie, A. W., & van der Molen, H. T. (2012). Implicit trait policies in multimedia situational judgment tests for leadership skills: Can they predict leadership behavior? *Human Performance*, *25*(4), 335–353.

Lievens, F., De Corte, W. & Westerveld, L. (2015). Understanding the building blocks of selection procedures: effects of response fidelity on performance and validity. *Journal of Management*, *41*(6), 1604–1627.

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority-majority differences and validity.*Journal of Applied Psychology*, *104*(5), 715–726.

意见 4：“然而，在选拔这种高利害场景中，情境判断测验的作假问题尤为严重，使得测评和 选拔效果大打折扣。因此，选定教师群体作为研究对象，开发开放式教师胜任力情境判断测验具有实践意义。”这段话不构成因果关系。

回应：感谢评审专家指出此问题。我们在原文中将此段话删去，直接展开讲述测验编制的过程。

意见 5：研究中测验的效标是：

（1）教学设计，教师依据统一的规范和要求提供一节课时的教学设计；

（2）教学视频，一个完整的 30 分钟以上的课堂教学视频；

（3）学生作业，按照优良差的标准各提交 3 份，共 9 份具有代表性的学生作业；

（4）课堂评价量表。

其中前三项是如何赋 分的？作者应当呈现这些项目作为效标的依据，即，有研究证明教师胜任力和这些因素有关。

回应：感谢审稿专家的建议，此部分叙述不够充分。

（1）关于赋分问题。我们修改了部分文字，增加了赋分的说明，以更详细描述研究中使用到的效标。具体如下：

“627 名教师中，共有 148 名教师提交了完整的教学监测材料，包括：(1)教学设计，教师依据统一的规范和要求提供一节课时的教学设计；(2)教学视频，一个完整的 30 分钟以上的课堂教学视频；(3)学生作业：布置作业，按照优良差各 3 份共 9 份具有代表性的学生作业。教学监测材料由 6 名教学专家进行评分，每个维度满分 3 分。教学设计的评价标准包括教学依据、目标、重点、难点、教学方法、教学过程 6 个方面；依据课堂观察量表(凌晨，2020)，从课堂管理、教学内容、思维培养、情感关注 4 个维度对教学视频进行评价，该量表内部 α 系数为 0.83，验证性因素分析结果为 CFI=0.897，TLI=0.868；对于作业部分，从作业设计、作业评价标准以及教师对学生作业的分析 3 个部分评分。”

（2）关于效标选取问题。研究者发现教师胜任力与实际工作表现相关，教师胜任力是预测工作绩效的有效工具(Arifin, 2015; Wahyuddin, 2016)。而教学设计、课堂教学、设计作业是一个完整的教学过程，是中小学教师教学能力的直观体现，是其工作绩效的代表性表现。

参考文献：

Arifin, H. M. (2015). The influence of competence, motivation, and organisational culture to high school teacher job satisfaction and performance. *International Education Studies*, *8*(1), 38–45.

Wahyuddin, W. (2016). The relationship between of teacher competence, emotional intelligence and teacher performance madrasah tsanawiyah at district of serang banten. *Higher Education Studies*, *6*(1), 128–135.

意见 6：“剔除作答时间少于 1000 秒与明显不认真作答的受测者 10 人”，明显不认真作答是如何判断的？

回应：感谢审稿专家的提问。每道题回答限定 100 字以上，而部分受测者，回答中实质内容仅有寥寥数字，通过重复的复制粘贴达到字数要求，以及部分研究者的回答与提问毫无关系，这些的数据判定为不认真作答。在文中已做补充说明。

意见 7：结果部分是否应当呈现人工评分的一致性，在一致性良好的基础上，才能对比人机评分的结果？

回应：非常感谢审稿专家的建议！诚如您所言，在一致性良好的基础上，对比人机评分的结果更加合理。

本研究中，为了保证人工编码评分的质量，选取 4 名心理学专业学生作为编码员，编码前统一接受半天培训，培训内容包括测评维度、编码标准、软件操作、遇争议项的处理原则(p.9)，题目被随机分配，使用 Nvivo 11 软件进行编码。

接着，测试其中两名编码员的默契度，在第 1 题上，两名编码员彼此独立编码，并依据编码结果计算各个作答者的得分，在此基础上计算人工评分一致性(r = 0.783, p＜0.01)，受评分主观因素影响，此一致性可以接受。紧接着，我们计算了人机评分的相关系数(r = 0.877, p＜0.01)，高于人工评分。

受研究条件所限，并未对其他题目进行独立双评。本研究对 300 人的 6000 道题目共 647322 字的作答文本进行人工编码，4 名编码员花费 3 个月时间，总计标注 19368 个句子，每道题的句子标注数量在 724~1453 句之间，如果其余 19 道题，每道题都由两人编码，将会是十分繁重的工作量。

但在研究过程中，我们尽量采取方法降低编码的主观性，编码和评分过程是分开的。

1) 首先，每道题由两名编码员先通读作答文本，分别独立确定行为反应项，再一起修改合并反应项，为反应项聚类，确立典型行为反应项(10-30 类以内，多为十几类)，对反应项的理解达成一致认同。

2) 接下来，进行编码：一名编码员在 Nvivo 软件中逐句编码标注，另一名编码人员对编码结果进行核查，提出不同意见，编码过程中还可对编码规则继续合并完善。

3)为行为反应项赋分也是经两名编码员讨论后确定，依据本题维度，同时关注行为的丰富度、具体性、全面性、逻辑顺序等所体现的思维水平和能力的差异，对各个反应项赋分(0~3)，直至达成共识(正文 p.9)。

意见 8：研究二中，“在心理测量理论的框架下对开放式测验进行自动评分的这类研究问题大体上分为四个环节：测验开发、设定评分规则、文本分类、验证自动评分性能 ，且互为基础、上下相承。”这部分阐述是否应当放在前文中，这个研究只考察了评分环节？

回应：感谢您的建议，在新一版的正文中，我们将其放在引言的最后(p.6-7)，作为整体研究思路的介绍。

意见 9：文中使用了很多评价指标，如准确率(accuracy)、精确率(precision)、召回率(recall)、F1 值等。应当详细介绍这些指标的目的、公式、标准。例如，作者提到“如图 4 所示，计算机在 20 道题上预测分数(0 ~ 3 分)的准确率在 70%~ 88%之间，结果较好。预测具体反应项的准确率在 57%~ 79%之间，考虑到……故此准确率仍属较为不错的结果。”有没有同类研究预测分数和准确率的结果可以参照？多少可以接受？

回应：感谢审稿专家提出的意见。

（1）针对第 1 个，评价指标问题，在正文中增加说明如下：在机器学习领域，对分类任务的评价一般采取准确率(accuracy)、精确率(precision)、召回率(recall)、F1 值四个指标。以二分类为例，对四个指标的计算过程进行说明。假设二分类包括正类和负类，下表所示为二分类情况下的混淆矩阵，矩阵中的元素定义为：1)TP(True Positive)：实际为正类且预测为正类的样本个数，2)TN(True Negative)：实际为负类且预测为负类的样本个数，3)FP(False Positive)：实际为负类且预测为正类的样本个数，4)FN(False Negative)：实际为正类且预测为负类的样本个数。

表 1 二分类的混淆矩阵表

		预测	
实际		TP 真正例	FN 假负例
		FP 假正例	TN 真负例

准确率反映模型在所有样本上的预测性能，等于分类正确的样本数除以总体样本数，即混淆矩阵中的对角线元素之和除以矩阵中所有元素之和，即准确率 Acc=TP+TN/TP+FN+FP+TN。精确率、召回率和 F1 值三个指标在每个类别上需单独计算。以二分类中的正类为例，精确率等于将正类样本预测为正类的数量除以所有预测为正类的样本数量，即 P=TP/(TP+FP)；召回率等于将正类样本预测为正类的数量除以真实的正类样本数量，即 R=TP/(TP+FN)；F1 值为精确率和召回率的调和平均值，即 F1=2PR/(P+R)。本文中的文本分类主要为多分类任务，在计算评价指标时先分别在每个类别上计算 P、R、F1，然后根据每个类别的样本数量计算加权平均值得到最终的精确率、召回率和 F1 值的评估结果。

（2）针对自动化评分性能的评估标准问题：反应项或者子分数的预测性能并没有绝对参考标准，原则上自然越高越好，一般要通过个体层面的信效度来评估模型性能是否达到可用。在机器学习领域，可根据预测结果和随机预测结果的差值来评价模型预测结果的有效性。根据本研究中的反应项和子分数数量，随机预测准确率在 5%-25%之间，可见本文应用的分类模型在反应项和子分数的预测上具有较好的效果，其应用性也在个体层面上实现了验证。

意见 10：文中涉及到的分类方法，如 CNN、RNN、R-CNN、RNN+Attention，没有英文全称及相应描述。

回应：感谢审稿专家指出问题。原文 p.11 描述“使用 CNN、RNN、R-CNN、RNN + Attention 多种分类模型对比实验性能”比较简单。

在更新的版本中，相应的描述修改为以下内容：使用多种成熟的深度学习方法进行文本分类建模，从而实现从作答文本到标签体系的自动化映射。在具体操作中，输入文本通过 Jiaba 分词和 Word2vec 预训练词向量转化为数字矩阵形式，再连接具有可训练参数的神经网络层、全连接层和 SoftMax 层，最终输出文本所属各个标签的概率。所应用的深度学习方法主要不同体现在模型的神经网络层，包括卷积神经网络(Convolution Neural Network, CNN)(Kim, 2014)、循环神经网络(Recurrent Neural Network, RNN)(Zhao et al., 2019)、循环神经网络串联卷积神经网络(Recurrent Convolution Neural Network, R-CNN)(Lai et al., 2015)和循环神经网络串联注意力网络(Recurrent Neural Network + Attention, RNN + Attention)(Pang et al., (2021)。其中，CNN 主要通过卷积核参数来捕捉各类标签的文本局部深度特征；RNN 通过循环单元结构来捕捉各类标签的文本全局深度特征；R-CNN 同时发挥两者的优势将 RNN 和 CNN 进行串联使用；Attention 则通过神经网络计算文本中每个词的权重来优化文本深度表征，通常与 RNN 进行串联使用。

参考文献：

Kim,Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing*, 1746–1751.

Zhao, Y., Shen, Y., & Yao, J. (2019). Recurrent neural network for text classification with hierarchical multiscale dense connections. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5450–5456.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2267–2273.

Pang, N., Zhao, X., Wang, W., Xiao, W., & Guo, D. (2021). Few-shot text classification by leveraging bi-directional attention and cross-class knowledge. *Science China Information Sciences. 64*(3).

意见 11：研究发现文档层面分类效果不好，即采用句子层面文本分类，但并未呈现分类效果的结果(如准确性)，说明这种方法更好。

回应：谢谢审稿人的严谨细致的审阅。原文对文档、句子两个层面的评分思路进行了尝试，由于篇幅限制，句子层面仅给出了准确率(accuracy)的结果，在原文 p.14 中描述如下：“实验结果表明，不管是在分数预测任务上，还是行为反应项分类任务上，CNN 的综合表现最为优异”。

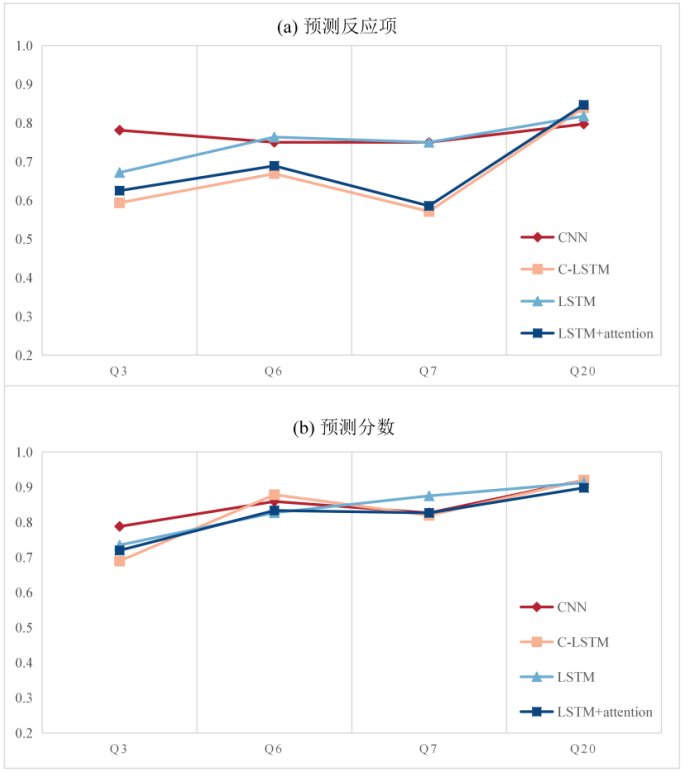

图 2 四种模型在四道题目上的准确率

由于过于简略，在文章中增加完整结果如下：

（1）反应项预测任务

四种算法在准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值这四项指标上的效果大体为：C-LSTM＜LSTM+attention＜LSTM＜CNN。

表 2 四种模型在四道题目上预测反应项任务的结果对比

题目	模型	Acc	Pre	Rec	F1
Q7	CNN	0.750	0.554	0.480	0.488
	C-LSTM	0.571	0.244	0.272	0.250
	LSTM	0.750	0.409	0.374	0.368
	LSTM+attention	0.586	0.280	0.273	0.261
Q3	CNN	0.781	0.585	0.563	0.546
	C-LSTM	0.594	0.297	0.296	0.289
	LSTM	0.672	0.254	0.291	0.268
	LSTM+attention	0.625	0.283	0.295	0.281
Q20	CNN	0.797	0.858	0.722	0.759
	C-LSTM	0.839	0.827	0.758	0.784
	LSTM	0.818	0.763	0.788	0.769
	LSTM+attention	0.847	0.799	0.769	0.778
Q6	CNN	0.750	0.723	0.577	0.585
	C-LSTM	0.669	0.552	0.574	0.552
	LSTM	0.764	0.626	0.615	0.611
	LSTM+attention	0.689	0.592	0.599	0.583

（2）分数档预测任务

四种算法在不同指标上效果大体如下表所示。

表 3 四种模型在四道题目上预测分数任务的结果对比

题目	模型	Acc	Pre	Rec	F1
Q7	CNN	0.826	0.641	0.551	0.583
	C-LSTM	0.819	0.670	0.505	0.547
	LSTM	0.875	0.614	0.585	0.598
	LSTM+attention	0.826	0.619	0.576	0.591
Q3	CNN	0.788	0.636	0.549	0.563
	C-LSTM	0.689	0.245	0.274	0.249
	LSTM	0.735	0.519	0.345	0.362
	LSTM+attention	0.720	0.367	0.297	0.287
Q20	CNN	0.920	0.916	0.739	0.797
	C-LSTM	0.920	0.779	0.877	0.808
	LSTM	0.912	0.910	0.782	0.834
	LSTM+attention	0.898	0.910	0.812	0.845
Q6	CNN	0.859	0.858	0.863	0.859
	C-LSTM	0.878	0.895	0.867	0.879
	LSTM	0.827	0.829	0.843	0.830
	LSTM+attention	0.833	0.816	0.816	0.815

实验结果表明，不管是在反应项分类任务上，还是分数预测任务上，CNN 性能表现都最为优异。

意见12：图 4 和图 6，为什么只标 Acc？图 4 中 Q3 的 Pre 是 0？结合图 6 看出 Q3 始终不好，可能的原因是什么？

回应：感谢审稿专家的提问。以下作出回应：

（1）为了图显示更加清晰，仅标注了 Acc 准确率的具体数值，准确率也是本研究中重点关注的指标。

（2）图 4 中 Q3 的 pre 精确率为 0.190909。

（3）对于您提出的问题（“结合图 6 看出 Q3 始终不好，可能的原因是什么？”）我们查阅了第 3 题的详细数据：

表 4 第三题的模型预测效果

ID	任务	题号	*Acc*	*Pre*	*Recall*	*F1*
201-300	预测反应项	Q3	0.781	0.585	0.563	0.546
201-300	预测分数	Q3	0.750	0.191	0.250	0.216
301-627	预测分数	Q3	0.707	0.294	0.675	0.287

发现在预测反应项上，模型结果尚可，但在预测分数上，F1、Pre、Recall 的结果都不理想，于是我们找到了本题的题目、评分规则和人工编码的情况。

第 3 题题目为：在“角的分类”这节课上，你按照预设的教学程序，引导学生相互交流、共同归纳，得出结论：“小于 90 度的是锐角；等于 90 度的是直角；大于 90 度小于 180 度的是钝角；等于 180 度的是平角；等于 360 度的是周角。”当你准备进入下一个教学环节时，有学生突然举手发问：“老师，大于 180 度而小于 360 度的角叫什么角？”听到这个问题，你感到有些吃惊，因为你根本没想过这个问题。此时，你会怎么做？

表 5 第三题评分规则

分档	编号	反应项	分值	句子标注数量
0	0301	偏题	0	34
1	0302	给予解答	1	37
	0303	引导其他学生解答	1	28
	0304	告诉是优角并讲解	1	19
	0305	加上简单分析	1	18
	0306	注意保持教学进度	1	15
	0307	让他找其他老师/数学老师	1	4
2	0308	表扬学生善于思考	2	**235**
	0309	与学生一起查资料	2	**200**
	0310	承认自己无法回答	2	69
	0311	课下与同学交流	2	26
	0312	鼓励其他同学也要善于思考	2	21
	0313	课后与其他老师讨论	2	5
	0314	留悬念，下节课再讨论	2	**4**
3	0315	作为专题，全班一起讨论	3	65
	0316	引导思考	3	17
	0317	备课、教学反思	3	16
	0318	鼓励在此话题下继续探索	3	**7**

我们猜测，可能是由于本题作答者回答比较趋同，出现了严重的标签标注数量不均衡的情况。预测分数是一个四分类任务(0、1、2、3)，大部分反应项是 2 分档，但其中的语言表达却同质性不高，比如“表扬学生善于思考”和“与学生一起查资料”在语义向量空间上距离应当是比较远的，但是机器学习的时候，习得的他们的输出是一样的，都是 2。在预测反应项上不存在这个问题，因为此时机器习得的输出是不同的)。因此，在未来的探索中，我们将对评分规则继续修订，同时增加数据集中个别反应项的标注数量。

意见 13：作者用图 5 比较了人工评分与机器评分分数频率分布，能否给出具体的分布指标进行对比？

回应：感谢审稿专家的建议。人工评分与机器评分的分布较相似，以下是各题的情况：

表 6 人机评分结果在各题上的对比表(n=94)

题号	均值 M		标准差 SD		区分度 D		相关系数 r
	人评	机评	人评	机评	人评	机评	
Q1	1.86	1.87	0.770	0.722	0.565**	0.646**	0.877**
Q2	1.95	1.87	0.739	0.691	0.468**	0.432**	0.639**
Q3	1.94	1.77	0.787	0.710	0.462**	0.538**	0.801**
Q4	1.83	1.86	0.838	0.649	0.613**	0.537**	0.707**
Q5	1.81	1.94	0.752	0.716	0.491**	0.477**	0.777**
Q6	1.88	1.88	0.716	0.716	0.563**	0.670**	0.602**
Q7	1.81	1.83	0.692	0.650	0.551**	0.557**	0.883**
Q8	1.78	1.83	0.642	0.728	0.569**	0.499**	0.631**
Q9	1.94	1.98	0.730	0.733	0.555**	0.409**	0.479**
Q10	1.87	1.99	0.691	0.755	0.477**	0.422**	0.821**
Q11	1.89	1.83	0.725	0.728	0.574**	0.584**	0.841**
Q12	1.67	1.94	0.612	0.787	0.452**	0.445**	0.537**
Q13	1.81	1.87	0.752	0.765	0.592**	0.597**	0.836**
Q14	1.70	1.69	0.801	0.748	0.506**	0.506**	0.814**
Q15	1.82	1.88	0.655	0.653	0.477**	0.463**	0.854**
Q16	1.77	1.83	0.809	0.812	0.644**	0.620**	0.741**
Q17	1.82	1.79	0.789	0.788	0.614**	0.552**	0.681**
Q18	1.65	1.94	0.714	0.759	0.582**	0.583**	0.752**
Q19	1.82	1.83	0.855	0.757	0.471**	0.573**	0.650**
Q20	1.76	1.82	0.714	0.816	0.561**	0.542**	0.902**
整体均值	1.82	1.86	0.74	0.73	—	—	

注：* p <0.05，** p<0.01。

意见 14：“人机评分的相关系数(r = 0.877, p＜0.01)高于两名人工编码员(r = 0.783, p＜0.01)。”人工编码员的相关是什么？

回应：感谢审稿专家的细致审阅。同上述意见 7 中的回应，人工编码员的相关指两名编码员分别独立对第一题评分所求得的 *pearson* 相关系数。见表述：“接着，测试其中两名编码员的默契度，在第 1 题上，两名编码员彼此独立编码，并依据编码结果计算各个作答者的得分，在此基础上计算人工评分一致性(r = 0.783, p＜0.01)，受评分主观因素影响，此一致性可以接受。紧接着，我们计算了人机评分的相关系数(r = 0.877, p＜0.01)，高于人工评分。”

意见 15：模型泛化评价目的是“以评估模型在新文本上的泛化能力。”但是作者又做了一遍测验质量分析，这样就能说明泛化能力吗？

回应：非常感谢审稿专家提出的问题，这使我们对模型泛化能力评分部分作了新的思考。当模型在面对新的数据集时也有很好的预测，我们才说这个模型就是表现好的，也就是泛化能

力强的。本研究中，共收集了 627 人的数据，对于前 300 人的数据，训练集和测试集按照70%、30%的比例划分，模型报告了在测试集上的优异表现。为了测试在新文本上的分类能力，数据集中使用了后 327 人的数据。除了报告模型性能，我们还依据机评结果，计算了测验的信效度，原意是假设模型信效度是特定的，在不同人群上施测，应该不会有太大的差异，从这一角度来说明模型的效果。但是，经过谨慎思考，我们认为信度分析有些偏离主题，因此在新一版的正文中将其删去。保留效度部分，以说明其在新的文本上机器评分的有效性。

意见 16：文章讨论部分总结“标签数量影响分类效果，本研究中二十道题目的标签数量皆 在二十项左右，标签数量多导致分类效果不佳。”能否对已有标签进行合并呢？已有研究是怎么解决这个问题的？

回应：感谢您对此部分提出的建议。我们确实有过这样的思考，本研究在确定评分规则这一步中，需要将所有可能的反应梳理成类别，最初一般一道题能梳理出 50 类以上，但是让机器学习一个 50 分类的任务，显然是困难的，而且每一类的样本标注也不够，因此大致梳理出类别之后，进行聚类、总结、再次提炼，至 10~20 类比较适当，变为十几分类任务，当然也可以继续概括至 5 类以内，这样的分类任务应当更简单了，可能会有更高的准确率，但我们认为 5 类以内过于“粗”了，对区分被试差异的意义不大。因此采用 10~20 类这样的一个分类概括程度。

意见 17：文章还有一些需要修订的小错误。

（1）英文缩写第一次出现时应当写出全称并解释，例如摘要中的 QWK，正文中的 NLP。

（2）文中有一些名词并不常用，需要作出解释。例如“增益效度”，“行为倾向指导语”。

（3）表下方应当有所有英文指标缩写的全称。

（4）p.11，“四种深度学习模型”是指什么应交代清楚。

回应：非常感谢审稿专家的建议。

（1）已检查文中所有的英文缩写，首次出现时均写出了全称。

（2）增加了对名词的解释：“增益效度(情境判断测验在同一类效标上的增量程度和变异贡献量更大(Sechrest，1963)。”。“行为倾向指导语”的解释如下：情境判断测验有行为倾向和知识导向的两类指导语，行为倾向型指导语要求作答者在备选项中选出自己最倾向于怎么做(would do)，知识导向型指导语要求作答者选择最好的一项(should do)。不同类型的指导语会影响到个体反应的理解、提取和判断的过程。在被试诚实作答的情况下，使用行为倾向型指导语，作答者的这三个过程主要受到人格因素的影响；使用知识型指导语，这三个过程主要受到认知因素的影响(McDaniel et al., 2001; 2007)。因篇幅原因，不在正文中详细解释，删去“行为倾向型”这一名词，改为此句：采用指导语“在这样的情况下，你会怎么做？”。

（3）已补充表下方所有英文指标缩写的全称。

（4）已在原文中补充了具体的深度学习模型及简介。

参考文献：

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M.A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

Mcdaniel, M. A. , Hartman, N. S. , Whetzel, D. L. , & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: a meta-analysis. *Personnel Psychology, 60*(1), 63–91.

漆书青，戴海琦. (2003). 情景判断测验的性质、功能与开发编制. *心理学探新, 23*(4), 42–46.

凌斌，顾金良，孙丽君. (2016). 情境判断测验的研究述评. *心理技术与应用, 4*(9), 538–548.

意见 18：作为一项指导应用的文章，本文应当附上测验和相关的程序语句。

回应：非常感谢审稿专家的审阅！感谢您细致地指出文章书写上的问题，耐心地指出不足，您的意见对本文的完善提供了巨大的帮助。如我们在自检报告中提到的那样，目前此测验正用于某市的新任教师选拔项目中，暂时还不具备公开的条件。研究突出的是开放式测验的评分思路和自动评分技术，是对这种范式的探索，其他研究者可在此思路下自主开发测验。我们在附录中附上了 1 道题的示例，后期也会将代码上传，与其他研究者共享。

第二轮

审稿人 1 意见：

作者对提出的问题进行一一的回复和修改，但是文章的研究问题还需要继续聚焦。具体意见如下。

意见 1：研究题目是“开放式情境判断测验的自动化评分”，重点是自动化评分。研究问题中也提到：(1)评分标准的问题；(2)自动化评分的问题；(3)自动化评分的解释性和效度验证问题。那么我们应该围绕开放式情境判断测验“设置评分标准的方法”“自动化评分的方法”“评分的解释性和效度验证”等领域的文献做出阐述和补充，在此基础上提出本文的研究问题。目前文献综述依然不聚焦，对于开放式情境判断测验评分标准的设定、自动化评分解释和效度方面还没有做出相关的解释。另外，从文章整体结构来看，前后是不一致的。目前研究结果方面主要集中在评分标准、自动化评分和效度方面。但是文章的讨论部分和结论部分的框架却是测验的开发、评分规则的设定、人工评分的问题。因此，文章需要继续调整结构，聚焦研究问题进行阐述和说明。

回应：感谢审稿专家对我们的文章提出的建设性意见，本轮我们对全文做出了大范围的修改，修改部分以蓝色字体标注，以下我们将对审稿意见做出一一回应。

在新一版文章中，将“设置评分标准的方法”“自动化评分的方法”“评分的解释性和效度验证”三项内容作为文献综述、研究结果、讨论、结论部分的主体。具体修改为：

一，在文献综述部分，围绕这三点作了文献补充；

二，在研究过程中，针对“设置评分标准的方法”“自动化评分的方法”做详细阐述。

三，在结果、讨论和结论部分，皆围绕这三点展开。

意见 2：文章中胜任力的维度包括学生导向、问题解决、情绪智力和成就动机四部分。为什么会作者选择用教学设计、教学视频和学生作业作为效标？

回应：谢谢审稿人提出的意见，效标问题确实是我们研究中很重视的一部分。我们在确立教师胜任力特征时选择从个人特质层面构建，这些是更深层的、更稳定的特征，更能影响教师的长期发展、绩效和职业适应，我们认为专业知识与技能一般会在教师资格考试中涉及，并未纳入到我们的胜任力特征中。因此，在选取效标时，较难找到非常贴合的效标。

而新教师一般都会有指导教师帮扶（有的学校称为“师傅”），我们在数据采集时也编制了一套胜任力量表，收集了指导教师对被测教师的评价，但由于指导教师的评价大多接近 5 分满分，SJT 的得分与指导教师评价胜任力的得分并不显著相关，所以未在正文中报告。

另一方面，教师胜任力是预测工作绩效的有效工具(Arifin, 2015; Wahyuddin, 2016)，因此我们选择一些能够代表教师工作能力的指标作为效标，在数据采集时纳入教学能力综合评估，让教师设计并提交一节课完整的前中后准备工作，包括教学设计、课堂录像、设计课后作业三部分，这三项是一个完整的教学过程，是中小学教师教学能力的直观体现，是其工作绩效的代表性表现，我们邀请了 4 位教学专家对提交的内容进行打分。

但仅有教学能力评估可能还不够具有说服力，我们还在新一版正文中补充了其他效标(p.7-p.8)，“为了对情境判断测验的效度做验证，还让被试填写了教师工作满意度问卷、自编公用教学理念问卷、自编学科教学理念问卷”，为效度提供更多方面的证据。

意见 3：目前 3.1 评分规则的内容不清晰，具体说明制定了哪些评分规则，可以在正文中举例说明。另外，评分规则的效度怎么样？

回应：感谢审稿人提出的意见。

首先，我们对 3.1 部分做了修改，现在正文中 3.1 包含“人工编码数据集生成”和“二十道题的评分规则”两部分，并展示了一道题的评分规则和评分切片示例。对于您提到的第二个问题——评分规则的效度如何，评分规则的效度可以从以下两个方面说明：

（1）编码一致性：选取第一道题做编码一致性检验，两个评分者在 627 份作答数据上的人工编码一致性 $r = 0.78$，二次加权 Kappa 系数为 0.82，一致性较高。

（2）通过测验结果证明：使用评分规则进行评估，分析评分结果(见 p.14 中 3.2 测验质量分析)，如果信效度指标良好，则可以说明评分规则效度高。本研究在当前采用的评分

规则下，测验的信效度指标是较好的。另外，采取 *pearson* 相关来计算题目区分度，以项目得分与测验总分的相关作为区分度指标，计算出各题的区分度在 0.450 ~ 0.625 之间，区分度较好。

意见 4：3.3.2 中作者呈现了人工评分与机器评分数据的分布，如何说明两种分布形态接近？请用相关数据来表达说明。

回应：感谢您的建议。我们在 p.17 的 3.4.1 中增加了人机评分总分的峰度和偏度值，“人工评分总分(36.36±7.99)的峰度为-0.592，偏度为 0.175，机器评分总分(37.23±7.83)的峰度为-0.345，偏度为 0.151”，另附上各题的均值和标准差。

表 7 人机评分结果在各题上的对比（*n*=94）

题号	均值 *M*		标准差 *SD*	
	人评	机评	人评	机评
Q1	1.86	1.87	0.770	0.722
Q2	1.95	1.87	0.739	0.691
Q3	1.94	1.77	0.787	0.710
Q4	1.83	1.86	0.838	0.649
Q5	1.81	1.94	0.752	0.716
Q6	1.88	1.88	0.716	0.716
Q7	1.81	1.83	0.692	0.650
Q8	1.78	1.83	0.642	0.728
Q9	1.94	1.98	0.730	0.733
Q10	1.87	1.99	0.691	0.755
Q11	1.89	1.83	0.725	0.728
Q12	1.67	1.94	0.612	0.787
Q13	1.81	1.87	0.752	0.765
Q14	1.70	1.69	0.801	0.748
Q15	1.82	1.88	0.655	0.653
Q16	1.77	1.83	0.809	0.812
Q17	1.82	1.79	0.789	0.788
Q18	1.65	1.94	0.714	0.759
Q19	1.82	1.83	0.855	0.757
Q20	1.76	1.82	0.714	0.816
整体均值	1.82	1.86	0.74	0.73

意见 5：3.3.2 中，作者提到采用 325 名教师样本，通过优秀班主任作为区分标准。首先，这部分的数据的作用，在研究方法那部分没有提前介绍。其次，这部分效标选取方法和前面教学设计、教学视频和学生作业的效标选取不一样，请做出相关解释。

回应：感谢您的提问和建议。以下将对这两点做出回应：

（1）在 2.1 介绍被试的部分增加描述：“还收集了被试的学历、职称职务、执教时间、

荣誉与获奖情况等个人信息”。

（2）325 名教师的这部分文本数据，主要用来测试模型在新文本上的泛化能力，一般在计算机领域，只报告模型性能指标即可，如 Accuracy、Precision 等，但由于本研究的评分任务应用在心理测验上，我们进一步使用机器评分的分数做了信效度检验，由于这部分被试没有提交教学评估的材料（教学设计、录制的课堂视频和学生作业），因此无法做效标关联分析，仅能从教师的个人信息中选用一些指标分析，其中就发现了“是否为优秀班主任”可以作为预测效度的检验方式。但确实，增加这一部分信息可能会造成前后不一致，也容易给读者造成误解，因此在新一版中删去“优秀班主任”的效度分析。

意见 6：对于研究局限，作者从样本量、测验公平性、测验本身、测验效度等方面进行阐述，最后一点才提到自动化评分的方法问题。文章的主要内容是自动化评分，应该围绕研究问题阐述文章的局限。从这点也可以看出，文章目前的研究问题还是不聚焦，需要继续做调整。
回应：感谢审稿专家的建议，研究局限部分我们修改了描述，见正文 p.22 讨论部分。除此之外，如意见 1 中的回应，我们对文章整体都做了调整，围绕核心的三点内容来展开叙述。再次向您的审阅致以衷心感谢！

……

审稿人 2 意见：

感谢作者根据审稿意见对文章进行的仔细修改，使得文章质量有了一定提高。但是，在二审过程中仍然发现一些问题，在此提出请作者酌情修改。

意见 1：图 4 和图 6，建议修改纵坐标从 0 开始，作者解释图 4 中 Q3 的 pre 精确率为 0.190909，比纵坐标的起点还低，读者是无法读取这个信息的。
回应：非常感谢您细心指出这个问题，我们在新一版正文 p.17 和 p.19 中对图 4 和图 6 重新制图，已解决此问题。

意见 2：p.4，第三自然段，“困难主要有两点，一是评分标准不易制定，二是评分成本高。”，后面又讲到“开放题主要由人工评分，往往不够稳定，存在评分者效应(rater effects)，且评分周期长，不能及时反馈分数，因此对效率要求高的大规模测验往往会谨慎使用开放题。”评分结果不稳定和评分周期长，又是第三和第四个困难。希望这一段能综合、全面总结困难。
回应： 感谢您的建议。在新一版的正文 p.2 中，已经整体调整了这一段的描述。

意见 3：p.4，第四自然段最后一句，“主观题自动评分目前主要分为两大类问题，长文本类型如作文自动化评分(Automated Essay Scoring, AES)，短文本类型如简答题自动评分

(Automatic Short-answer Grading, ASAG)。”与前文衔接不强。加在这里很突兀。

回应：谢谢您指出这个问题，在新一版的正文 p.3 中，注重文章写作的流畅性，已经修改如下：“第二，关于开放式 SJT 自动化评分的算法。文本自动化评分的研究最早可以追溯到Page(1966)的早期工作，目前按照文本长短大致可以将文本自动化评分的研究分为两大类，长文本类型如作文自动化评分(Automated Essay Scoring, AES)，短文本类型如简答题自动评分(Automatic Short-answer Grading, ASAG)。开放式 SJT 的问题类型介于这两者之间，尚无明确有效的自动化评分算法。”。

意见 4：p.5，第二自然段，“常见关键词匹配法”，句子不通，应为“常见的方法有关键词匹配法，即……”。

回应：非常感谢您的建议，已经修改为：“常见的方法有关键词匹配法，即作答文本的关键词越多分数越高”。

意见 5：p.6，第二自然段，文本分类之后是如何找到类别与所测心理特质的关系的？

回应：感谢审稿人的问题，这个问题涉及设置评分模型的核心，是我们研究的重点。文本分类的目的是将文本分配到一个或多个预定义的类别中，通常使用分类器来完成这项工作。对于将文本分类与所测心理特质之间建立相关，涉及整个研究过程的设计，我们做了以下工作：

（1）数据收集与标注：由人工对作答文本编码，梳理每个作答者在某问题情境下具体的反应行为是什么。

（2）设计评分规则：给行为反应项赋予分值作为权重。通过“行为反应项”这样一个桥梁建立文本与所测心理特质的关系，每道题单独设定一个评分规则，这个规则列举了当前情境下所有可能的行为反应，符合本道题测评维度（心理特质）的行为反应项被赋予更高权重。

（3）文本分类：特征提取，采用 Word2Vec 进行词向量编码，提取文本的特征；分类模型训练，将已标注好的数据集分为训练集和测试集，并用训练集来训练分类模型，比如 CNN、RNN、LSTM 等，然后使用测试集来评估模型的性能；分类结果分析，发现高分文本与低分文本在特征上的差异。

意见 6：p.6，第三自然段，“(1)评分标准的制定”没有相应的解释（可参照第二点）？

回应：感谢审稿专家指出这个问题，现已修改为：“(1)评分标准的制定。目前的评分多依赖于专家经验”。

意见 7：图 1 上面文字总结“自动评分总体分为三个环节：设定评分规则、自动文本分类、验证自动评分性能”。但在图中三个大环节只是流程中的小节点。能否在图中更好地总结小

的流程节点和大环节？后面的研究过程能否与这个流程对应？

回应：按照您的建议，我们对图 2 自动评分的总体流程重新制图（正文 p.6），使图中内容与研究过程部分内容与顺序一致。

意见 8：p.7，研究预期部分，没有关于评分效度的预期。

回应：感谢审稿专家指出这个问题，现已修改为：“（3）机器评分具有较好的信效度，机器评分稳定性高，人机评分存在正向的强相关。”。

意见 9：p.8，测验编制中的题目修订部分，为什么没有进行测验信效度的验证？

回应：感谢审稿专家的提问，以下将做出解释。

在专家评定与题目修订这一环节，主要的目的是减少题目表述歧义、由一线教师评定情境是否真实、测试在有选项的情况下被试是否有优势作答倾向，来确认开放式 SJT 的必要性，因此，重点对题目本身做了修订。另一方面，在试测环节进行信效度检验主要是为了保证测验质量，而试测版本是有选项的 SJT，其信效度与正式施测中的开放式 SJT 的信效度不一定相近，因而未作信效度检验。不过，如果试测中加入信效度分析，确实有利于保证测验质量，这也是我们研究设计中欠缺考虑的部分。

意见 10：2.1.2 教学监测数据”，这个标题描述不够准确，需要说明该数据的用途。

回应：“非常感谢您的建议，目前这一标题已经做了修改，也在 p.7 中 2.2.2 的效标部分，增加了数据用途的描述：“为了对情境判断测验的效度做验证，同时让被试作答教师工作满意度问卷、自编公用教学理念问卷、自编学科教学理念问卷，并提交教学能力综合评估材料（一节课的教学设计、课堂录像、学生作业）。”。

意见 11：研究过程特别是数据处理等部分，应当说明每一步的目的，和下一步的关系。

回应：感谢审稿人的建议，在新的版本中，已经注意在每个章节开始的部分增加一部分综述，来说明每个步骤设计的用意，也着重修改了具体的语言表述，使其过渡更加自然。

意见 12：“2.3.1 人工编码”，为什么只算第一道题的人工评分一致性？

回应：感谢审稿人的问题。以下是我们对这个问题的思考：

（1）采取方式降低人工评分的主观性。通过多名评分员评分，能够发现某些评分员的极端分值，能在一定程度上减少评分的主观性，但这只是保证评分信度的方式之一。我们在研究的最初阶段，也是采取了双评的方式，结果发现人工评分的一致性只能达到 0.5 左右，因此，我们不再对回答直接评分，而是将“直接评分”转变为先编码，再给此编码一定的分值。在本研究中，编码更类似于对文本的观点进行概括总结，被归为有限的十几类中的一种，

因此编码工作相较于直接评分，主观性的空间相对小一些。“选取第一道题做编码一致性检验，两个评分者在627份作答数据上的人工编码一致性 $r = 0.78$，二次加权 Kappa 系数为0.82。”这也说明，先编码再评分的方式能够一定程度上提高评分信度。

（2）受研究条件所限。我们想要实现自动评分就是想要解放人力工作，而编码是一项非常繁重的任务。目前的版本中，对 64 万字的文本，标注了一万九千余次，仅由一名编码员编码就耗费了 4 名编码员近 4 个月的时间，如果其余 19 道题每道题都由两人编码，时间和人力成本将非常高，与我们的研究目的相悖。

经过综合考虑，我们认为只在第一道题上双人编码是可以接受的，我们也在研究结果中发现，“在 20 道题目上人机评分的相关系数 r 依次为 0.88、0.64、0.80、0.71、0.78、0.60、0.88、0.63、0.48、0.82、0.84、0.54、0.84、0.81、0.85、0.74、0.68、0.75、0.65、0.90”，总体上可以接受。但我们的考虑可能不够全面，也欢迎审稿人对此问题再次作出指导。

意见 13：整个编码流程写的不够清晰，例如，什么是“行为锚”，建议作者在描述时举例说明，让读者能够看明白。

回应：谢谢您的提问。根据审稿专家的意见，我们重写了编码流程这一部分详见 p.9，2.4.2 人工编码，使其更加清晰。另外，关于“行为锚定”的评分思路，在引言中增加了介绍，“人工评分参照了行为锚定评分表(behavioral anchored rating，BAR)进行评分，评分标准更加具体化和客观化，该表是 Smith 和 Kendall 在 1963 年提出的，它是一种行为测量工具，用于员工的工作绩效评级。”这种评分思路更关注回答中的具体行为是什么，以此来给予分数。

意见 14：“2.4 测验质量分析”中，对结构效度的验证为什么不做验证性因素分析？

回应：谢谢审稿人指出问题，在原一版本的正文中在这一部分有做验证性因子分析，但可能在段中，不够清晰，“使用 Mplus7.0 对数据进行验证性因素分析。$\chi^2/df = 1.34$，RMSEA = 0.034，SRMR = 0.042，CFI = 0.955，TLI = 0.948，拟合情况较好，各项目在各因子上的因子载荷在 0.420 ~ 0.623 之间”。在新一版的正文中，p.14 的 3.2“测量质量分析”中，增加部分内容：“为检验自编测验的结构效度，设定并比较了四种验证性因子分析模型”，使其更清晰展示。

意见 15：文章讨论部分，“4.1 应在规范的心理测量框架下开发开放式情境判断测验”，建议换为其他内容。文章主要关注自动化评分，测验开发并不是重点，因此不应放在讨论的第一点。

回应：非常感谢您的建议，目前讨论部分的内容已做了大幅修改，分为三点：“评分标准的设计”、“自动化评分过程的影响因素”、“自动化评分的效度和可解释性”，详见 p.20。

意见 16：文中有很多笔误，比如“在人工编码的数据集中中”，“3.2.1 文档层面层面文本多分类”。建议作者自查修改。

回应：谢谢审稿人指出问题，我们为自己的粗心感到非常抱歉。在本轮的修改稿中，我们已经以出声朗读全文的方式来检查并修改书写上的错误，感谢您的指正。

第三轮

审稿人 1 意见：

感谢作者的回复和修改。文章的结构相比上一版也有了很大的进步，文章主体内容也变得清晰。但还有几个问题需要继续修改和完善。具体意见如下：

意见 1：目前文章篇幅太长，希望作者进行精简，更加突出文章主体内容。

回应：我们在本轮修改中，进行了大幅删减，在不改变结构的情况下尽可能保证精简，原文正文字数 18182 字，删减后字数为 13961 字。

意见 2：2.2 提到还收集了“被试的学历、职称职务、执教时间、荣誉与获奖”，但是后文并未将这些指标用于分析中。

回应：谢谢审稿人提出的问题，新的一版中删去此描述，以使文章更简练。

意见 3：在 3.1.1 中，前面提到“选取 300 份进行编码”，后边又提到 627 份作答数据，前后样本量不一致。

回应：感谢审稿人细心指出这一问题。本研究在初始阶段，为了测试不同的分类模型，把第一题数据(627 份)做了全部编码，大致选定实验的几种模型之后，为提高效率，后面的十九道题目，只编码 300 份。但这一背景信息并未在正文交代，因此在描述中，确实会给读者造成误解。考虑后，我们决定第一题仅计算前 300 人的数据，删除 627 份的描述，不给读者带来阅读上的困惑。本部分更新为：“选取第一道题做编码一致性检验，两个评分者的人工编码一致性 $r = 0.84$，二次加权 Kappa 系数为 0.78”。

意见 4：为什么人工评分的结果与工作满意度、公用教学理念、学科教学理念以及教学设计、课堂评价、学生作业等显著相关。而机器评分中显示仅与公用教学里面显著相关。请做出相关解释。

回应：谢谢审稿人的提问。人机评分与效标的相关性和显著性表现不同，可能的原因有：

一，人机评分之间的偏差。人机评分的结果本身就是有差异的，因此当它们与效标做相关系数的计算时，结果亦有不同。

二，样本量也是很大的影响因素。人工评分的效标相关性是在 300 人份的数据上来计算的(去除无效后余 290 人)，机器评分的效标相关性是在 100 人份的数据上来计算的(去除无效后仅有 94 人)，样本量不同的原因是：人工编码 300 份，在总数据集中取后 100 人作为测试集预测文本的标签和分数，因此机评只有这 100 人的结果。样本量相差悬殊，样本量较小时，受取样偶然因素的影响较大，因此相关性的表现出现了较大的差别。

三，机评的效标并未选用教学能力评估部分的教学设计、课堂评价、学生作业这三个效标，这部分需要教师精心备课和录制，因此提交成本很大，符合要求的材料仅有 181 份。机器评分的后 100 人中，仅有 30 人提交，数量过少，即使与机评分数的相关显著地高，由于样本量原因也不能算作充分的效度证据，因此在此未纳入比较。

……

审稿人 2 意见：

感谢作者的修改和回复，目前文章质量有了进一步提高。目前较大的问题一是希望文章更聚焦，脉络清晰；二是文章创新性不足，需要突出其独特贡献。仍有一些问题，与作者商榷。

意见 1：p. 8，效标介绍部分，学科教学理念问卷和公用教学理念问卷不同吗？是否有学科特异性的题目设计？为什么没给出效度指标？

回应：非常感谢审稿人细心和耐心的审阅。学科/公用教学理念问卷是不同的问卷，《公用教学理念》要求全体教师作答，《学科教学理念》针对语文、数学、英语三科分别设计题目。语数英教师共 269 人，语文学科为 99 人，数学 86 人，英语 84 人，需要报告 3 部分各自的验证性因子分析结果，因此在前一版的描述中，篇幅限制未做全部报告，但其验证性因子分析结果也是比较好的，CFI、TLI 值都在 0.9 以上。

意见 2：结果部分，“选取 300 份进行编码（ID 为 1~300），已编码文本数据共 647322 字，每道题的句子标注数量在 724~1453 句之间，总计标注 19368 个句子。选取第一道题做编码一致性检验，两个评分者在 627 份作答数据上的人工编码一致性 $r = 0.78$，二次加权 Kappa 系数为 0.82。”其中，后面检验评分一致性的是否应为 300 份，因为只有 300 份进行了编码而不是所有的 627 份？

回应：感谢审稿人指出，第一位审稿专家在本轮中也提出了这个问题，确实是上一版本论文中表述不清。本研究在初始阶段，为了测试不同的分类模型，把第一题数据(627 份)做了全部编码，大致选定实验的几种模型之后，为提高效率，后面的十九道题目，只编码 300 份。

但这一背景信息并未在正文交代，因此在描述中，确实会给读者造成误解。考虑后，我们决定第一题仅计算前 300 人的数据，删除 627 份的描述，不给读者带来阅读上的困惑。本部分更新为：“选取第一道题做编码一致性检验，两个评分者的人工编码一致性 r = 0.84，二次加权 Kappa 系数为 0.78”。

意见 3：部分标题显得不够简练，如“3.1.2 二十道题的评分规则”。
回应：谢谢您，此标题已改为“评分规则”。

意见 4：“在 20 道题目上人机评分的相关系数 r 依次为 0.88、0.64、0.80、0.71、0.78、0.60、 0.88、0.63、0.48 、0.82、0.84、0.54、0.84、0.81、0.85、0.74、0.68、0.75、0.65、 0.90，p＜0.001。”其中有些题目的相关较低，如 0.48，可能造成的原因是什么？能否进行补救？
回应：感谢审稿人的提问，人机评分的各题与总分的相关系数不同，其背后的原因是复杂的，包括：

第一，题目本身的问题，题目本身是否与教师胜任力的特质联系不强，导致人评和机评结果出现了较大偏差，但通过计算人评的题总相关系数（见下表），发现其在正常范围内，因此排除是题目本身的问题。

第二，评分规则。各题的评分规则是不同的，类别数不等，因此机器的分类效果有差异。本题评分规则中共有 22 类，类别较多，确实可能导致机器分类效果不佳，但我们发现，本题预测反应项的准确率为 0.71，预测分数为 0.85，在全部题目中属于较好的水平，因此可能也并不是由于类别过多。

第三，标注的数量与质量。本题标注 1162 个句子，标注数量较多，那么可能在标注质量上可以进一步探究。综合来看，可能的原因是，虽然机器分类的准确性还不错，但在未正确分类的句子中，或许是对 1 分反应项和 3 分反应项的区分力不够强，导致该题的机评结果的区分度不够，具体的表现就是该题机评的题总相关在所有题目中最低(如下表 1 所示)，如果以人评为参照，低了 0.15。

补救的思路是，（1）检查一遍数据标注集，看是否存在标注错误，但预期这一步骤提升的空间较小。（2）改进分类模型，提升模型识别语义的准确率，进一步使机评向人工评分靠拢，提升人机评分的一致性。模型改进和算法升级是本研究团队近期正在努力的事情，自动化评分系统将不断升级，引入更加准确和高效的评分手段。

表 8 各题的题总相关、人机评分相关表

维度	题目编号	人评-题总相关	机评-题总相关	人评-机评相关
学生导向	Q1	0.57	0.65	0.88
	Q8	0.57	0.50	0.63
	Q9	0.56	0.41	0.48
	Q10	0.48	0.42	0.82
	Q12	0.45	0.45	0.54
	Q16	0.64	0.62	0.74
	Q20	0.56	0.54	0.90
问题解决	Q3	0.46	0.54	0.80
	Q4	0.61	0.54	0.71
	Q6	0.56	0.67	0.60
	Q7	0.55	0.56	0.88
	Q17	0.61	0.55	0.68
	Q18	0.58	0.58	0.75
情绪智力	Q2	0.47	0.43	0.64
	Q5	0.49	0.48	0.78
	Q11	0.57	0.58	0.84
	Q19	0.47	0.57	0.65
成就动机	Q13	0.59	0.60	0.84
	Q14	0.51	0.51	0.81
	Q15	0.48	0.46	0.85

意见 5：模型泛化能力评估部分，采用了未文本标注的样本，那是如何得到预测准确率的？

回应：感谢审稿人的提问，没有标签的数据无法计算准确率指标。此部分对模型的泛化能力做评估是一种外部验证，在训练集和测试集之外，再单独获取一批新的数据进行预测。假设一套测验的信效度指标是特定的、稳定的，那么在前 300 份和后 327 份上得到的信效度指标应该是接近的，借以比较前后两份文本上的信效度的具体数值，来观察机评的结果在后一份新文本上有无较大的偏差，我们在这样的逻辑上来做分析，以此说明模型在新文本上的泛化能力。而实际操作中，我们对模型的训练是取了已经标注的句子来训练的，然后将新文本输入进这个训练好的模型中，因此这里的预测准确率其实也使用到了已标注的数据集。在本轮修改中，对这个问题进行了再三考虑，认为泛化能力评估这一部分，验证不够充分，可能会带来读者的疑问，且此部分并不是一个自动评分系统中必不可少的，另外，我们的文稿字数已经严重超过了建议字数一万字，需要在本轮修改中进行大幅删减，因此，将模型泛化能力的验证这一部分做了整体删除。

意见 6：讨论部分，自动化评分过程的影响因素考虑了文本分类的层面，以及一些人工因素，影响因素是否还应包括采用的深度学习模型？这些模型的特点是什么？

回应：感谢审稿专家的建议。诚然，不同类型的深度学习模型在处理文本分类任务时，具有

独特的优势和限制，这些特点会在自动化评分性能方面产生不同的影响。

例如，卷积神经网络(CNN)在文本中主要捕获局部特征，如词组、短语等，对于需要考虑长程依赖关系的任务，CNN 可能表现较差，因为它无法有效地处理长文本序列中的全局信息；循环神经网络(RNN)及其变体，如长短时记忆网络(LSTM)和门控循环单元(GRU)，在处理序列数据时能够捕捉上下文信息，适用于对文本中长期依赖关系较强的任务。然而，传统的 RNN 有梯度消失和爆炸问题，导致难以处理长序列，虽然 LSTM 和 GRU 在一定程度上缓解了这些问题，但仍受到文本长度的限制而在一些文本分析任务上表现较差；注意力机制(Attention)使模型能够在处理文本时聚焦于关键部分，有助于更好地捕捉重要信息，但早期的注意力机制通常与 RNN 绑定使用，容易受到 RNN 模型的限制。

本研究中，由于句子层面的反应项分类任务通常与特定的词组、短语等相关联，因此 CNN 在本研究上的性能最优是可以理解的。在广泛的研究任务中，不同类型的深度学习模型在自动化评分中具有各自的特点，它们的选择将对评分性能产生显著影响。根据任务的特殊要求，结合模型的优势和限制，选择合适的模型有助于提高自动化评分的准确性。此外，在预训练语言模型(Pre-trained language model)、大语言模型(如 ChatGPT)等被提出后，自动化评分模型也有了更丰富的选择，但考虑到场景的特异性，仍然需要经过严谨的性能评估、信效度检验才能确定评分模型的可用性。

我们也将这些思考加入到了新一版正文中的讨论部分。

意见 7：目前文章篇幅太长，很多地方举的说明例子可以放到附录中，以免干扰整个文章介绍方法过程的脉络。此外，文章有一些重复的地方，例如“自动化评分的效度和可解释性”中对自动化评分系统性能良好的说明，前文中已经阐述过。希望作者进一步精简和聚焦。

回应：非常感谢审稿专家的建议，我们在新一版中已将举例说明的部分放进了附录中，对于讨论部分我们也做了较大调整，删去赘述，新增了部分新的思考。

意见 8：总的来说，研究的创新性不足。应用深度模型进行文本分类，实现机器评分，已经在很多测验上得到了应用，该研究只是将其应用于开放式情境判断测验中。应当进一步突出该研究的独特贡献。

回应：感谢审稿专家的建议，以下是我们对这个问题的思考。

一个研究的创新性可以从理论、技术和方法等多方面来评价，或是与同领域研究相比有独特的观点，或是解决领域内无法解决的问题，或是对领域内的应用和实践产生影响，应当都算作创新之列。虽然深度模型已经被广泛应用于文本分类等领域，但是应用于开放式情境判断测验的研究非常少，这一个新的问题类型，需要新的解决方案。

本文的价值在于对开放式测验开发-编码-设定评分规则-自动评分-效果验证这样一个完整的范式的探讨。在心理测量学上，本研究更关注测验的质量和自动评分的实现，核心在于

“如何评分”，以及“如何自动化地评分”。

对于“如何评分”，本文提出根据关键行为设置题目评分反应项的思路，摒弃了以往类似研究中设定评分要点或者设定参考答案之类的做法，这是研究思路上的特别之处。

对于“如何自动化地评分”，可以笼统地把这类问题归为主观题评分问题，按照文本长短，长文本有作文自动评分问题，短文本有简答题自动评分问题，这两类都有丰富的研究成果。但是开放式情境判断测验属于无标准答案的短文本类型，目前这一块的研究成果很少，已在引言部分介绍。尤其是在算法方面，没法直接参照其他研究中聚类或者相似度计算的算法。因此本文提出的编码评分方法，以及对这一大类问题的研究范式的探索，是具有一定学术意义的。

深度学习模型并不是研究的主体，在正文中没有详细、大篇幅地介绍深度模型，本文不认为这一部分是本研究中最有创新意义的部分，因为总是可以采用更先进的算法和模型来替代这一部分，本研究团队目前也正在改进这部分的算法、以使模型训练效果更佳。

新版的论文将研究意义的描述做了调整，作为讨论的第 4 部分，详见 p.17。

……

审稿人 3 意见：

本文使用深度学习算法探究了开放式情境判断测验的自动化评分问题。研究具有创新性，研究整体完善，有一些细节性问题有待进一步解决：

意见 1：P2 “行为选项编制和计分方式的不合理设置会使得测验易受猜测，通常信度表现较低、效度难以保障”，“使得测验易受猜测” 句法问题。

回应：谢谢审稿人指出，因篇幅限制，已经将此两句删去。

意见 2：P2“开放式作答一定程度上可以解决这些问题，这种形式不局限于固定答案，能够给予受测者更多自由表达的空间(Finch et al., 2018)，促进对主题材料的深入理解”，促进谁？

回应：感谢审稿人的提问，已更改为：“促进受测者对主题材料的深入理解”。

意见 3：P3“简答题自动评分侧重于评估语义内容而非一般语法风格，考察特定知识点”，第一，是简答题考察特定知识点，不是简答题自动评分考察特定知识点，主语混乱。第二，将其放在“而非”后面，容易引起歧义。

回应：感谢审稿人严谨的审阅，已更改为：“简答题考察特定知识点，其自动评分侧重于评估语义内容”。

意见 4：P4“文本分类的流程主要包括：文本预处理、向量化表示、特征抽取、模型训练、

模型预测与评估、模型部署等部分。” 这里的文本分类流程是否特指本文研究中所用到的文本分类流程，如何是的话，那么从后面的内容可以知道，该研究采用了 word2vec 这种词嵌入方式来实现文本的向量化表示，这种方式本身就是一种自动的特征提取方法，因此这里文本分类流程中向量化表示和特征抽取本质上指代同一个过程。如果这里的文本分类方法是泛指的话，它与图一的内容不符，图一中没有向量化表示这一步骤。希望作者能够统一逻辑。

回应：感谢审稿人的提问，如您所指的后一种，引言中对文本分类方法的介绍是泛指。目前已经把这一段文字改为：“文本分类的流程主要包括：文本预处理、特征提取、模型训练、模型评估、模型优化与应用等部分”，并且再次检查了图示内容，使得几处表述统一。

意见 5：P4“第二个阶段是应用训练好的模型对无标签的测试数据进行预测并进行性能评估”，应该是应用测试好的模型对测试数据进行预测，将预测标签与真实标签进行比较以评估模型性能。如果你的测试数据没有标签，我不能理解你是如何进行性能评估的？

回应：感谢审稿人严谨细致的审阅，如您所言，测试集也是有人工标注的标签的，此为表述错误，在新一版的正文中删去“无标签的”。

意见 6：研究过程：无论是预测反应项还是预测分数，从作者描述和表 2 可以看出训练数据极有可能存在样本的类别不平衡问题。作者是否统计了训练数据中每种类别的比例，如果不存在样本不平衡问题，请提供证据，如果存在，作者在训练模型过程中是否通过欠采样或者调整损失函数中不平衡类别的权重值来解决这一问题，如果样本不平衡问题没有得到解决，作者在之后的模型泛化性能评估突出准确率作为参考显然是不够客观的。

回应：感谢审稿专家的提问。如您所述，对于反应项的人工编码结果，确实存在类别间样本不均衡的问题，比如下表 9 展示的第 15 题编码示例，可以看出，在 18 个类上人工编码标定的句子在一些反应项上数量多，在一些反应项上数量少，从 15 句到 90 句不等。本题总计标注 862 句。

首先，针对此存在类别间样本不均衡的数据，在实际的模型性能验证过程中，我们划分训练集和测试集时，基本保证了训练集和测试集在各个类别上的比例基本一致，因此实验过程并没有引起额外的样本不均衡问题。

其次，对于您提出的“准确率作为参考不够客观”的问题，在机器学习领域中，只有在二分类任务上或者多分类任务中某一类样本数量占据绝对优势时，准确率才不足以反映模型的整体性能的。在本研究的数据集上，在每一道题目上均有多个反应项的样本数量较多，数据集能够反映大多数人的作答模式，且能够支撑模型训练在多个反应项上的预测能力，因此通过准确率来反映模型的整体性能是合理的。此外，研究中报告的多分类 F1 值采用的是 Macro-F1，也是适合于评估不均衡分类的常用指标。

最后，为了进一步澄清此问题，图 1 展示了该测试集上预测结果的混淆矩阵，可以看出

样本数量极少的类别上的性能确实容易低于样本数量多的类别但影响不大。对于各类别上训练数据只有较少样本的情况，很难通过解决不均衡问题来处理，而应该引入机器学习领域的小样本学习进行研究，这也是本研究团队下一步的研究方向。由于这不是本研究的重点，本文没有在正文中对此作出特别处理和说明，仅在讨论部分有提及。谢谢审稿人的提问，使我们再次梳理了这个问题。

表 9 第 15 题的评分规则表

编号	行为反应项	分值	人工标注句子数量
1500	偏题	0	15
1501	寻找突破点	1	24
1502	个人分析	1	55
1503	家校合作	1	62
1504	向学校寻求帮助	1	81
1505	严明纪律	1	28
1506	班干部管理	2	45
1507	处理个别同学	2	56
1508	共同规划	2	66
1509	建立感情	2	65
1510	奖罚分明	2	49
1511	了解情况/找出原因	2	90
1512	树立榜样	2	48
1513	与学生谈话沟通	2	21
1514	自我鼓励	2	27
1515	创新方法	3	16
1516	反思与提升	3	52
1517	强化集体意识	3	43
1518	学生自我管理	3	34

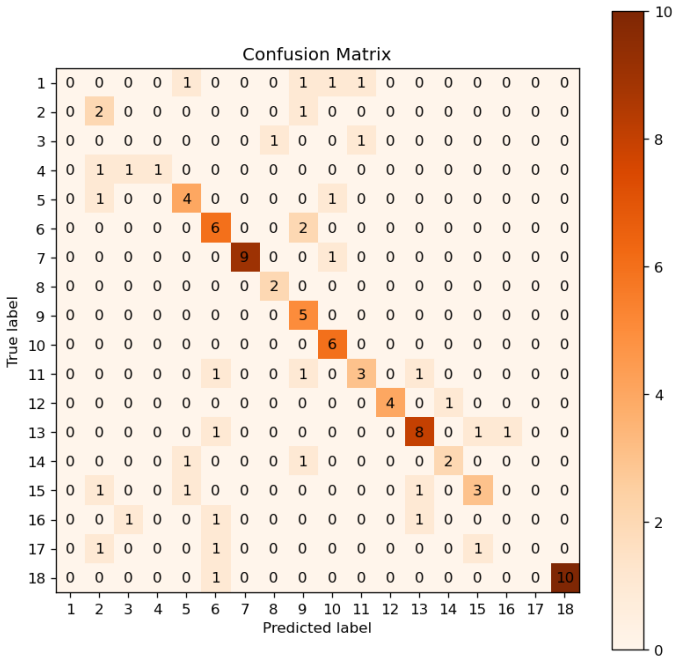

图 3 混淆矩阵图

注：1-18 分别为：“寻找突破点、个人分析、家校合作、向学校寻求帮助、严明纪律、班干部管理、处理个别同学、共同规划、建立感情、奖罚分明、了解情况/找出原因、树立榜样、与学生谈话沟通、自我鼓励、创新方法、反思与提升、强化集体意识、学生自我管理”。

意见 7：结果：3.4.3“对 627 份数据集中的剩余 327 份未标注的文本作机器自动化评分”，如果这 327 份文本没有标注，那么是如何计算准确率的？你的真值是什么？

回应：感谢审稿专家的提问，第 2 位审稿专家也在意见 5 中提出了这个疑问。我们的回复如下：没有标签的数据无法计算准确率指标。此部分对模型的泛化能力做评估是一种外部验证，在训练集和测试集之外，再单独获取一批新的数据进行预测。假设一套测验的信效度指标是特定的、稳定的，那么在前 300 份和后 327 份上得到的信效度指标应该是接近的，借以比较前后两份文本上的信效度的具体数值，来观察机评的结果在后一份新文本上有无较大的偏差，我们在这样的逻辑上来做分析，以此说明模型在新文本上的泛化能力。而实际操作中，我们对模型的训练是取了已经标注的句子来训练的，然后将新文本输入进这个训练好的模型中，因此这里的预测准确率其实也使用到了已标注的数据集。在本轮修改中，对这个问题进行了再三考虑，认为泛化能力评估这一部分，验证不够充分，可能会带来读者的疑问，且此部分并不是一个自动评分系统中必不可少的，另外，本文的文稿字数已经超过了建议字数一万字，需要在本轮修改中进行大幅删减，因此，将模型泛化能力的验证这一部分做了整体删除。

意见 8：“新文本上的模型性能如图 6 所示，模型在测试集上各题目的分数预测准确率在 68% ~ 95%之间，准确率较高。”这里的测试集与 2.5.1 中“以 300 名教师的测验数据作为数据集，20 道题目共 6000 道回答，总计标注 19368 个句子。训练集和测试集按照 70%、30%的比例划分。”的测试集是否指代同一个？如果是同一个，我没有看出来 70%和 30%是如何划分的。

回应：感谢审稿人的提问。是指同一个测试集，“句子按照 70%和 30%的划分”，是将数据集的人群按 2:1 来分，前 200 人的句子标注集合是训练集，后 100 人的句子标注集合是测试集。新版正文中也修改了描述：“每道题的标注文本中，按照人群的 2:1 划分训练集和测试集”。

第四轮

审稿人 1 意见：

本文经过作者的再次修改，对审稿人的意见有详尽的回复，文章的质量有了明显提高。目前文章中还有一些小错误，需要作者再仔细检查。第 5 页 2.2.1 中，采取行为事件访谈法，这里落了“采”。讨论部分，4.1 评分标准的设计中，合理赋分中的(1)和(2)的序号表达方式需要与(2)有所区别。

回应：非常感谢审稿专家的审阅，对于您提到的两点问题，

第一点，已修改为“采取”；

第二点，原文描述中，评分规则构建的目标与难点，分开进行论述，皆采用了(1)、(2)的序号表示方式，确实不够清晰。因此，本轮修改了“4.1 评分标准的设计”这一段落的表述结构，按照“合理分类”和“合理赋分”两点分开论述，不再出现过多序号(详见 p.15)。

审稿人 2 意见：

作者已经逐一对意见进行了回复和修改。我没有更多的建议。“4.4 研究意义”放在讨论部分似乎不太合适，或者改成“实践启示”？供参考。

回应：非常感谢审稿专家对本文的所作的详细指导，4.4 部分的标题名已改为“实践启示”。

审稿人 3 意见：

感谢作者根据审稿意见对文章进行的仔细修改，文章中还有一些通读起来别扭的句子，例如引言部分的“有相关研究者对书面回答式(Lievens et al., 2019)和视听构建式(Oostrom et al., 2010, 2011)进行了探索，是富有创新性的尝试，然而仍采用人工评分的方式。”；4.2 中“在广泛的研究任务中，不同类型的深度学习模型在自动化评分中具有各自的特点，“它们的选择”将对评分性能产生显著影响。”希望作者能够仔细的通读全文并作出相应的修改，使文章更符合中文的语法习惯。除此之外，文章没有其他问题，符合录用标准。

回应：非常感谢审稿专家的耐心审阅和指导！对于您提到的语言表达问题：

一，引言部分，此句已修改为：“有研究者对书面回答式(Lievens et al., 2019)和视听构建式(Oostrom et al., 2010, 2011)的测验形式进行了探索，是富有创新性的尝试，然而评分环节仍采用人工评分的方式。”。

二，4.2 部分，“它们的选择”已修改为：“模型的选择”。

此外，在本轮修改中，我们再次检查了语言表达的流畅性和精准性，修改了部分语句，在正文中以蓝色字体表示。

第五轮

编委意见：同意发表。

主编意见：本研究以教师胜任力测评为例，针对教学中典型问题场景开发了开放式情境判断测验，并借助机器学习方法，探索了开放式情境判断测验中自动化评分的应用。本论文的选题具有一定新颖性，研究方法选用恰当，获得的研究结论真实可信。