人工智能生成内容技术在教育考试中应用探析

王 蕾
（教育部教育考试院，北京 100084）

摘要：初步探讨了人工智能生成内容（AIGC）技术在教育考试领域的应用，特别是在试题自动生成与技术增强型试题方面的应用潜力。通过AIGC技术辅助试题自动生成，可以自动、快速地生成大量高质量内容，从而降低命题成本；技术增强型试题能够丰富题目类型，加强核心素养考查，使试题更为接近真实的生活和学习场景。两者有机结合不仅可以解决题目资源不足的现实问题，还能够实现考试试卷的个性化，更加突出能力和素养考查。建议在高中学业水平考试合格考、高等教育自学考试、高校的拔尖创新人才选拔考试等项目中开展相关试点，并加快开发适用于我国教育考试的大语言模型。

关键词：人工智能生成内容；试题自动生成；技术增强型试题；大语言模型

【中图分类号】G405 **【文献标识码】**A **【文章编号】**1005-8427（2023）08-0019-9

DOI: 10.19360/j.cnki.11-3303/g4.2023.08.003

近期，以ChatGPT为代表的人工智能生成内容（artificial intelligence generated content，AIGC）技术引发全球的高度关注。AIGC技术是通过人工智能算法自动地生成内容，这里的内容包括各种类型传递信息的数据，如文本、音频、图像、视频等。AIGC技术的优点是可以根据要求自动、快速地生成海量高质量的内容，从而节省时间和人力，提高效率和精确度。AIGC技术在网络营销、客服、翻译、医疗、教育等多个领域都有广泛的应用。除聊天功能外，文本创作、文字转图像、自动摘要等都属于AIGC技术的应用范围。大语言模型（large language models，LLMs）是AIGC的核心技术之一，其在ChatGPT产品的成功应用，揭示了传统人工智能算法在海量数据和参数支持下能够“智慧涌现”。为此，几乎所有有影响的国内外高科技公司都在竞相开发大语言模型，或探索在不同行业的垂直应用。

2022年初，教育部启动教育数字化战略行动，推动国家智慧教育公共服务平台建设。信息化长期以来是我国教育和考试改革的重要主线之一，从信息化到数字化，不仅意味着对信息技术深入、全面的运用，更是让业务与技术真正产生交互，改变传统以物理资源为核心的业务运作模式。这种变革形成了以收集数据、分析数据、预测数据为核心的思维模式和业务模式，从而催生业务创新，解决许多传统业务模式无法解决的问题。为此，我国考试工作者提出“打造智慧考试，服务智慧教育”[1]的基本理念，并围绕其内涵和外延，组织高水平大学、科研机构和考试机构进行了一系列研究探索。ChatGPT的横空问世，从一定程度上推动了智慧教育和智慧考试研发的紧迫性。对于AIGC技术的迅猛发展及由此带

收稿日期：2023-05-24

作者简介：王 蕾（1969— ），女，教育部教育考试院研究员。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

来的考试题型、考试形式、数据收集模式、结果分析和应用等变化，各级考试机构和广大研究人员应给予高度关注并作出快速反应。

1 试题自动生成是AIGC技术在教育考试中颇具潜力的应用

试题自动生成（automatic item generation，AIG）起源于20世纪60年代美国等西方发达国家的大学和考试机构，基于认知心理学和心理计量学的基本理论和模型，借助编程等技术，自动生成满足特定教育或心理测量目标的试题[2]。试题自动生成既可以应用于纸笔考试，又可以应用于计算机化考试，其主要目标是扩充现有题型下的题目数量。

试题自动生成的优势在于能够有效降低命题成本，同时大幅增加可用的题目数量，以解决传统命题方式因成本过高、缺少命题教师等难题。由于国外尤其是发达国家人工成本昂贵，因此，试题自动生成技术一直受到考试机构的高度重视，并在一些国际著名的考试项目中得到应用。

长期以来，试题自动生成在我国并未得到充分的重视，这是出于我国独特的国情和考情原因。我国历来重视考试，考试结果在各类升学、就业、选拔和提级决策中发挥着至关重要的作用，社会对考试的公平性异常敏感。考生和教辅机构为备考往往要投入大量的时间、精力和财力、物力，对试题的研究深入到细枝末节。笔者注意到，我国考试中存在能力降级现象，即一种题型在其出现早期往往具有较高的统计难度，而随着考生对此题型的逐渐熟悉，即使同样题型的题目更换新的场景和参数，难度也会显著降低。考生对题型的熟悉程度是决定试题难度的重要因素。因此，我国大规模社会化考试试题需要在考试内容和方式上保持一定的变化，如此才能突出能力和素养的考查，不能像西方国家那样从题库中随机抽取一部分试题反复使用。那样的话，很容易导致考试分数的膨胀和考查能力的降级，进而导致考试信度和效度的降低。

进入21世纪以来，随着互联网，尤其是移动互联网的迅猛发展，人们已经深刻体会到信息技术的重要性。从教育模式的演变来看，农业社会的特征是分散化，教师个人在这个阶段起到决定性的作用。进入工业社会，教育特征转变为标准化，统一的课程标准和教材通常为全国同龄学生所使用。在数字化时代，教育的特征转变为个性化，每个人都可以依据自身的能力、兴趣以及专业成长目标选择适合自己的学习内容和进度。理想的考试编制和组织实施模式应该让每个考生都能面对最适合其知识能力水平的试卷，这样才能最大程度地发挥其能力，确定自己在群体中所处的位置，找到合适的学习材料。有些大规模社会化考试允许考生多次参加，同时一些考试形式的改革引进新的试卷生成方式。这些多元化的因素使得考试机构对题目数量的需求呈现出快速乃至指数级增长趋势。例如，在我国高考中，以前只需要一份全国统一的试卷，但现在每年都需要命制几十套试卷。

我国作为全球经济最富活力的国家之一，拥有最发达的互联网应用，政府和人民都高度重视教育，我国无疑应该站在数字化创新发展的前列。因此，有必要重新探讨试题自动生成在我国教育考试中的应用。试题自动生成在广义上属于AIGC技术的一种应用。如果说在ChatGPT出现之前，由于其固有的技术局限难以满足我国大规模考试的需求，那么ChatGPT的出现无疑让人们重新认识到试题自动生成技术的潜在价值。试题自动生成有可能成为满足考试改革和发展的一条新途径，以ChatGPT为代表的AIGC技术在智慧考试命题中展现出广阔的应用前景，为我国从考试大国迈向考试强国提供了新机遇。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 以试题自动生成满足大规模社会化考试的迫切需求

无论是从纸笔考试向机考、网考的形式转变,还是从常模参照考试到标准参照考试的理论模型演变,都呈现出相同的变化趋势:即每次考试所需要的题目总数量在增加,而每道题目在总体中所占的权重逐渐减小。因此,如果某一道题目出现质量问题甚或泄题,其造成的影响也会相应地减少。

从考试形式角度看,纸笔考试的最大特征是“千人一卷”,因此对题目数量的需求相对有限,但对于因题目泄露等因素引起的考试安全问题十分敏感。在大规模考试中,一旦发生泄题,其后果往往是灾难性的。因此,考试机构不遗余力地确保命题和制卷过程的绝对保密性。像教育部负责命题的国家考试试题,就必须进行严格的查重和避重检验,与公开流通的大量题目进行比对,以防止出现重复题目引发公众舆情。这意味着尽管题目数量有限,但命题成本依然高昂。

计算机化考试(computerized test 或 computer assisted test)将纸笔考试搬到计算机上实施,早期阶段的研究重点在于比较其与纸笔考试在功能、公平性和成本等方面的差异,以证明考试计算机化的必要性和可行性。后续进一步发展的自适应性考试则采用动态的试卷生成策略,即根据考生的答题情况动态调整试题难度。如果算法判断考生的能力水平较高,就会提供一个难度更大的题目,反之亦然。从理想程度看,每个考生得到的试卷版本是不同的,这意味着即使通过舞弊在考前得到部分试题,也未必能直接应用于自己的考试,因此实行自适应性考试后,题目泄露的安全风险大大降低,但这需要试题库储备更多的试题。考试工作者为能够在提高考试信度、效度与降低安全风险和控制成本之间找到平衡,理论上可行的自适应考试的形式在实施中研发了很多变型,如多阶段考试等。

从理论模型角度看,考试逐渐从常模参照性考试向标准参照性考试演变,这不仅体现在社会上众多新兴的证书考试,很多传统选拔性考试也逐渐改革为标准参照考试。例如,我国香港地区的高中文凭考试2012年改革为“采用水平参照模式汇报成绩”[6]。标准参照性考试的目的和功能是判断考生是否达到某种预设标准,这类试题的特点是根据人为设定的外部标准(如课程标准、成就标准等)命题,主要考查标准所规定的基本知识、能力和素养,设计上避免难题、偏题、怪题等。

综合以上各方面因素,考试形式和考试模型理论的演变给试题自动生成提供了新的生存和发展的空间。Kurdi 等将试题自动生成方法归类为3种:模板法、规则法和统计法[2]。早期的试题生成主要采用的是模版法和规则法,通过人工设定相应的模板和规则,从而创建问题。例如,当需要设计一道考查个位数加法的题目时,在传统命题方式中,命题教师只需要在纸上写出“5+7=”。而基于模板法的试题生成通过设定原始模板为“{参数1}+{参数2}=”,编写一段计算机程序,随机产生两个参数并通过验证确保参数是10以内的整数,就能够得到无数道考查相同知识点、具有难度相近的新题目。这是试题生成中最简单的场景之一,而在实际应用中设定的模板会不断复杂化。例如,考查的知识点扩展到“个位数的加减法(不涉及负数)”,此时模板需要增加第3个参数,即“{参数1}{参数3}{参数2}=”,其中参数3的取值范围为“+”或“-”,同时需要增加一条验证规则:“参数1大于或等于参数2”。近年,随着机器学习和深度神经网络的兴起,试题生成多采用基于统计的方法,通过训练复杂的语言模型,从而创建问题。中国科学技术大学、西安交通大学等单位进行了相关研究[4-5]。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

新高考改革实施以来，高考科目分为两部分：一部分是以确保达到课程标准为主要目标的高中学业水平考试合格考；另一部分是为满足高校选拔目标而设定的高中学业水平等级考试以及语文、数学和外语3门国家统一考试。其中，高中学业水平考试已从原来的国家统一命题转变为省级教育考试机构负责命题。然而，许多省级考试机构反映，这一改变使得其在命题和组织考试方面的负担显著增加。此外，《关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见》明确要求“坚持以学定考，进一步提升中考命题质量，防止偏题、怪题、超过课程标准的难题”[7]。对于主要功能设定为确保高中教育质量、减轻学生不合理课业负担的高中学业水平考试合格考，试题自动生成可能提供了一种非常有前景的解决方式，可以从根本上减轻各级考试机构的命题负担。高考、研究生入学考试等高利害选拔考试，以选拔国家发展所需要的拔尖创新人才为主要目的，承担着维护教育公平的重担，试题自动生成技术在这些考试中在可以预见的将来应用可能性很小。

我国学生的应试能力普遍较高，因此以模板和规则为主要特征的早期试题自动生成技术所产生的题目，由于题型固定、缺少变化，容易导致能力降级现象。ChatGPT发布后，国家考试机构组织队伍，成立智慧考试课题组，研究试题自动生成技术对我国教育考试的潜在影响和可能应用。在文献梳理过程中，有两项国外的相关研究引人关注：一是美国医学考试协会案例。美国医学考试协会研究案例发表于2019年。该研究使用基于Transformer技术的GPT-2模型生成医疗认证测试材料。具体来说，该模型首先使用一份由互联网通用语料库训练得到的345M参数预训练模型作为基础。其次，为适应医学文本语境，使用开源数据库PubMed的80万篇医学文章对模型进行微调。经过6天训练，使用调优模型生成电子病历描述和医疗多项选择题的干扰项。实验结果表明：GPT-2模型具备生成可被人类创作、加工的文本草案的潜力[8]。未来，使用更新的Transformer模型结合现有的项目数据进行实验，有望进一步提高实验性能，并促进测试材料的开发。二是Duolingo案例。Duolingo语言教学与考试机构进行的研究案例发表于2022年[9]。该研究应用GPT-3模型，产生交互式阅读理解题目，并自动对考生作答进行打分。该研究使用超过14 000段语篇，由GPT从中挑选789段，经由人工审查保留454段。GPT基于这些语篇出题和设计评分细则。随后，通过Duolingo APP招募考生试测题目，试测持续了21天，共5 246道题目，平均每道题目收集到425名考生的作答数据。研究者对收集到的数据进行了统计分析。

上述两个案例反映了考试领域国际先进技术的发展趋势，说明国外考试机构已经开始研究AIGC技术在试题自动生成领域的应用，这无疑给考试领域带来一场革命，使考试命题从传统的以专家经验为主的模式，演变为通过AI产生素材和半成品、再经专家最终审查成题更具有质量和效率的新模式。这场新的革命有可能从根本上改变考试命题模式，彻底解决题目资源不足的难题。

智慧考试课题组在2023年6月尝试通过ChatGPT生成英语阅读和理解试题，并组织国内3所大学近千名学生试测。考生反馈题目自然流畅，感觉不到与人工命题有什么差别。智慧考试课题组由此得出3个方面的认识：1）与ChatGPT聊天，正确地提问非常重要。在人与人的交流中，可以消除不会对交流造成显著影响的因素，如语言的歧义、人脑的潜意识、常识类的假设等，但这些因素对ChatGPT的影响则大相径庭。例如，用高考规范格式要求ChatGPT“作文长度不超过800字”时，它仅写出495个字，这肯定会被严

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

重扣分,这是由于ChatGPT缺少对“常识性假设”的理解造成的。在写作要求改为“作文长度介于700~800字”时,以人的意识来看其实什么都没有改变,ChatGPT却写出799个字。因此,有必要开发专门用于考试命题的ChatGPT交流模板,或者帮助提问的规范套路,作为命题人员与ChatGPT的沟通中介。2)不要简单要求ChatGPT提供成型的题目甚或一套试卷,要致力于让ChatGPT发挥其知识面广的重要优势,从而辅助命题人员出好每一道题。ChatGPT对任何前沿或敏感的问题回答都是四平八稳,滴水不漏,很多回答看似正确,实则内容空泛,甚至在不知道正确答案的情况下,也会无所顾忌地编造答案。因此,用户要具备高度的批判性思维能力和技巧,深挖ChatGPT的逻辑规律,调动它最大的知识储备,这样才能充分发挥ChatGPT的作用。3)用于训练已有模型的数据集是有偏的,不同国家开发的模型必然存在不同意识形态的立场或偏见,不可能存在国际通用的模型。因此,开发适用我国教育考试领域的专用模型就显得十分必要和紧迫。

3 以技术增强型题型满足能力和素养考查的需求

加强能力和素养的考查始终是我国考试内容改革的主要目标。在我国考试改革实践中,将试题自动生成与技术增强型试题(technology-enhanced items)配套使用,能够减少试题自动生成模式化带来的副作用。技术增强型试题是指通过应用先进的计算机网络和多媒体技术,实现考生和考试环境之间的交互,相比传统纸笔考试或早期计算机化考试,具有更科学、高效、个性化交互等特点。这种形式的试题更贴近真实的学习环境,可以更有效地收集考生对试题的反馈,从而更好地考查考生的知识、能力和素养[3]。技术增强型试题通常伴随即时结果反馈、不同后续发展等动态试题、试卷生成技术,一般常见于基于网络或计算机的考试,是对现有考试题型种类的扩充。

与传统题型相比,技术增强型试题的优势是能够显著提升考试效度,可以测量更为复杂、多元的知识、能力或素养,使试卷和题目更接近真实的生活和学习场景。同时,随着自动评分、即时反馈等技术的应用和信息技术的快速发展,它还可以通过减少评分误差显著提高考试的信度,快捷、详细地报告考试结果,从而改进考生的考试体验,降低考试实施和管理的成本。

为了最大限度地避免命题模板化产生的副作用,并进一步提升智慧考试在考查能力和素养方面的水平,智慧考试课题组并行开展了试题自动生成和技术增强型题目的研究。笔者认为,将这两种方法并列讨论和研究,能够更好地发挥考试的正面功能。下面通过案例介绍几种当前国际流行的技术增强型题型,包括改进型选择题、多阶段试题、模拟与仿真试题、多媒体试题以及信息技术应用试题。

3.1 改进型选择题

以“四选一”为主要特征的多项选择题,是教育考试标准化、工业化的典型代表。这种形式的题目不仅可以有效控制考试评分的误差,还能大大降低考试成本,对推动考试公平性乃至实现教育的普及化起到了极其重要的作用。然而,选择题因远离正常学习和应用场景、限制了考生的思维,长期以来备受诟病。最简单的技术增强型试题是对简单选择题的改进。例如,通过下拉式菜单提供足够的备选项,从而使猜测效率极小化,或根据考生选择的备选项呈现“第二小题”等。

图1展示了一道统计题,要求学生在给出的数据中选取出最大值与最小值,进而判断在移除最大值和最小值后,剩余数据的何种属性会与原始数据保持一致。这主要考查学生对平均数、中

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

位数以及极差方差的理解和掌握程度。如果在传统纸笔考试中,作为一道“四选一”类型的选择题,按照命题经验最多给出10个数,这可能导致学生的猜测答案的概率很大,从而影响考试的信度和效度。相比之下,改进后的技术增强型题给出了真实场景中的100个数,并提供了一键排序工具,在确保考试信度和效度的同时,还可以考查学生是否会使用工具。

3.2 多阶段试题

在真实的学习环境中,完成学习任务有困难的考生可以随时向教师或同学求教,这是真实学习过程的一部分,但在纸笔考试中却无法实现。事实上,这会降低考试的信度,因为考生并非完全不懂难题,在获得提示或帮助后,可以完成任务并获得这道题目的部分分数。图2是一道有关生物实验的多阶段试题,要求学生知道合成同位素标记的多肽链所需的材料和设备,并通过拖拽相关组件完成实验。这主要考查学生生物学基因表达和蛋白质合成的相关知识。多阶段试题通过提供提示、根据考生的答题或选择提供下一问等手段,使考试题目更贴近真实的学习情境,增强个性化和定制化,从而适应不同能力水平的考生,从而提高考试信度。

3.3 模拟与仿真试题

对纸笔考试的另一大批评是其无法考查科学课程中必不可少的实验环节。理想的解决方案是让学生进入实验室亲自动手,但在现代许多行业中,如飞行员训练,绝大部分目标可以通过模拟和仿真技术实现,而无需亲临操作现场;此外,很多儿童和成年人都沉迷于通过模拟和仿真技术构建的数字世界,表明这些技术已经相当成熟。图3展示了一道结合模拟和仿真的技术增强型试题,改编自以前高考纸笔卷的一道生物选择题,通过仿真技术模拟真实的生物实验场景,目的是让考生通过核糖体上的图示和碱基序列,推断对应的tRNA序列,并正确地将tRNA放置在图中的正确位置。这主要考查学生生物学蛋白质合成过程的理解和掌握程度。考生可以反复进行交互实验达到实验目标。因此,模拟和仿真技术可以大大扩展考试题型的功能,甚至要求考生展现极高的思维和判断能力。

3.4 多媒体试题

动图、音频和视频等多媒体材料可以拓展考试的考查功能,无须赘述。图4展示了一道有关

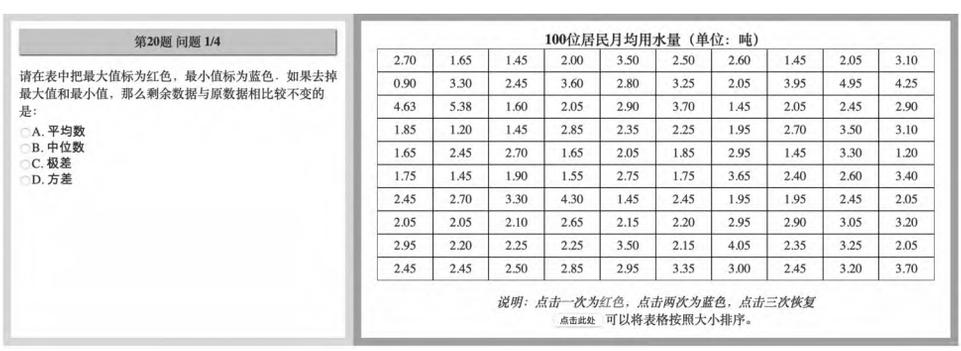

图1 改进型选择题

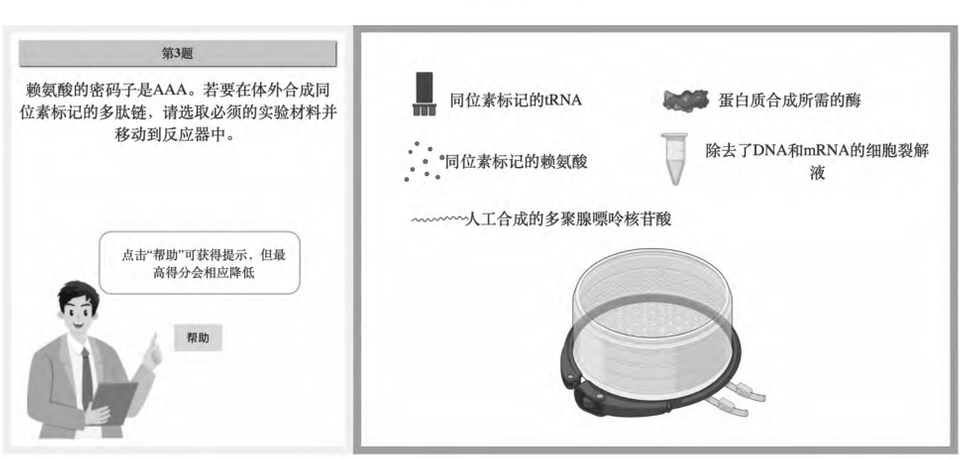

图2 多阶段试题

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

环境保护方面的多媒体试题,改编自以前高考纸笔卷的一道地理主观解答题,要求学生理解和评估人类活动对环境的影响。这涉及学生对生态环境影响的基本理论和概念的理解,如荒漠化等概念的理解。原题通过4张静态图展示了环境变化,而技术增强型的多媒体试题则使用一段8秒视频展示。这不仅增强了试题的表现力,同时也从防偷拍、防作弊角度探索了更多的防范手段。教育部提出2023年高考要让手机“带不进、用不了、传不出”的目标[9],如果未来考试题目以多媒体形式呈现,就会杜绝使用手机拍照外传的作弊行为。

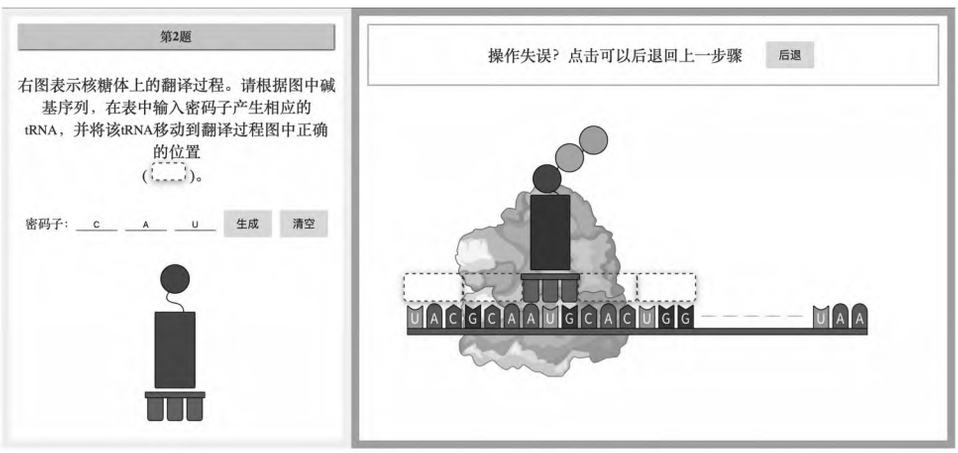

图3 模拟与仿真试题

图4 多媒体试题

3.5 信息技术应用试题

随着信息技术的发展,人机互动形式越来越多样化。考生可以在屏幕上设定热点、移动对象、绘制图形等,这些技术的应用使考试更贴近现实生活和工作场景。近年来,考试内容改革提出情境化命题的方向,而技术增强型试题无疑更有利于这一目标的实现。图5展示了一道有关明代卫所的信息技术应用试题,改编自以前高考纸笔卷的一道历史主观解答题,要求学生在图片区域标示明代

图5 信息技术应用试题

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

卫所集中分布的一个区域，并在文字区域解释分布的原因。该题主要考查了学生对明代军事制度、地理环境以及历史背景的知识。考生可以通过与该题进行交互，在试题图片区域上自由选择并标示卫所位置，然后在文字作答区域详细阐述理由。新型人机交互试题不仅形式更加丰富多样，而且更贴近实际生活场景。

在实际应用中，智慧考试目前主要采用两种技术路线：一是智慧考试课题组开发了一个名为“技术增强型试题开发平台”的工具，学科秘书和学科专家可以利用这个平台，通过“所见即所得”的工作方式，编制技术增强型试题。二是预计在未来，相关领域的技术服务公司将通过商品化的方式，开发一款类似于Office的命题办公软件，供命题人员使用。然而，对于需要用到诸如建模、模拟、仿真等技术的复杂题目，可行的技术路线是在命题组配备前端开发工程师，通过在命题过程中与学科专家间的反复交流，共同完成题目的制作。

4 智慧考试在我国教育考试中的应用路径

以试题自动生成和技术增强题型为代表的智慧考试代表大规模社会化教育考试未来的发展方向，打造智慧考试的最终目的是服务智慧教育和更好地服务考生。智慧考试的研究面向3个主要目标：汇聚考试资源、强化考试服务、创新考试生态。

多年来，国家考试机构在计算机等级考试等项目上积累了丰富的机考、网考经验，最近又全面启动了高等教育自学考试专网机考的探索。我国的国情和考情具有其独特性，一些考试如研究生入学考试、高考、中考等，涉及的利益相关方众多，事关国家治理和社会稳定，在这些考试中，任何事关体制机制和考试内容的改变都必须慎重、稳妥。因此，智慧考试可以首先在低利害、风险小的项目中先行试点，取得经验后再进行推广。在国家考试层面，高等教育自学考试是较为适合的试点项目；在省级考试层面，高中学业水平考试合格考是相对适合的试点项目。此外，高水平大学开发的一些针对拔尖高中生的选拔考试也适合进行智慧考试试点。这是因为高水平大学具有独特的人才优势和智力优势，这类选拔是分散的和小规模的，能够规避对整个教育系统产生负面影响。因此，国家应该鼓励高水平大学探索选拔培养拔尖创新人才的新途径，研发智慧考试新技术。

试题自动生成和技术增强型试题是智慧考试的重要组成部分。自古以来，我国教育一直追求的理想是孔子提倡的有教无类和因材施教。今天，有教无类在我国已经基本实现，因材施教在信息化和数字化的推动下显露曙光。考试数字化转型是考试强国建设的新赛道，也是考试现代化建设的必由之路。在某种意义上，考试是考生和试卷的一种交互，有交互就会产生数据，交互越充分、越深入，得到的交互数据就越丰富，对被试者的评价也才越全面、越深入。目前统一命题和千人一卷的考试模型相对传统，考生的作答模式与其在生活中收集、分析、反馈信息的模式相去甚远，这种模式已跟不上时代的发展。智慧考试不仅能实现考试的个性化，还能增强考生和试卷之间更加充分的交互，获得更丰富的数据，包括答题时间、答题思路、答题路径等信息，从而从更多维度上对考生、命题、教育进行评价和反馈，使以学定考、以考促学、教考相长，实现全体学生全面而有个性的发展。我国高水平大学和各级教育考试机构应携起手来，发挥制度优势，尽快建设适用于我国教育和考试需求的大语言模型。基于我国的经济社会的实际情况，开发人工智能在教育和考试领域的垂直应用，从根本上改变我国长期以来的追赶者地位，使我国成为世界教育考试发展的引领者。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献:

[1] 于涵.打造智慧考试 服务智慧教育[J].中国考试,2023(5):1-10.

[2] KURDI G, LEO J, PARSIA B, et al. A systematic review of automatic question generation for educational purposes[J]. International journal of artificial intelligence in education, 2020(30): 121-204.

[3] What is a technology-enhanced item? [EB/OL].(2023-05-17)[2023-05-18]. https://support.goguardian.com/s/article/What-is-a-Technology-Enhanced-Item-1629330276146.

[4] HUANG Z, LIU Q, GAO W, et al. Neural mathematical solver with enhanced formula structure[C]. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, July, 25-30, 2020, Xian, China,

[5] ZENG H W, ZHI Z, LIU J, et al. Improving paragraph-level question generation with extended answer network and uncertainty-aware beam search[EB/OL].(2021-04-20)[2023-05-18]. https://sci-hub.wf/10.1016/j.ins.2021.04.026.

[6] 香港中学文凭简介[EB/OL].(2021-03-31)[2023-05-18]. https://www.hkeaa.edu.hk/tc/hkdse/introduction/.

[7] 中共中央办公厅 国务院办公厅印发《关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见》[EB/OL].(2021-07-24)[2023-05-18]. http://www.gov.cn/zhengce/2021-07/24/content_5627132.htm.

[8] VON DAVIER M. Training optimus prime, md: generating medical certification items by fine-tuning OpenAI's gpt2 transformer mode[EB/OL].(2019-08-23)[2023-05-18]. https://arxiv.org/abs/1908.08594.

[9] ATTALI Y, RUNGE A, LAFLAIR G T, et al. The interactive reading task: Transformer-based automatic item generation[EB/OL].(2022-07-22)[2023-05-18]. https://www.frontiersin.org/articles/10.3389/frai.2022.903077/full.

[10] 教育部:把防范手机作弊作为今年高考安全的重中之重_新闻频道_央视网[EB/OL].(2023-05-10)[2023-05-18]. https://news.cctv.com/2023/05/10/ARTIXoNro7WErcm8d3tiRolC230510.shtml.

Widening the Paths of Assessment Development via Artificial Intelligence-Generated Content: Potential Applications of Automatic Item Generation and Technology-Enhanced Items in China

WANG Lei

(National Education Examinations Authority, Beijing 100084, China)

Abstract: This article delves into the application of Artificial Intelligence Generated Content (AIGC) technology in educational assessments, especially its potential for automatic item generation and technology-enhanced items. Assisted by AIGC technology, a large amount of high-quality content can be automatically and quickly generated, reducing the cost and increasing the number of items. At the same time, technology-enhanced items can enrich item types, strengthen the assessment of abilities and literacy, and make the test papers and items closer to real-life and learning scenarios. The combination of both not only solves the real problem of insufficient item resources but also realizes the personalization of examination papers, highlighting the assessment of abilities and literacy more than the traditional mode. Consequently, it is recommended that relevant pilots be carried out to target areas such as the threshold level of the high school academic examinations, higher education self-study examination, and gifted and talented students' selection examinations organized by top colleges and universities and to accelerate the development of Large Language Models suitable for China's educational examinations.

Keywords: artificial intelligence-generated content; automatic item generation; technology-enhanced items; large language model

(责任编辑:周黎明)

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net