人工智能教育评估应用的潜力和局限

袁 莉[1] 曹梦莹[2] 约翰·加德纳[3] 迈克尔·奥利里[4]

(1. 北京师范大学 未来教育学院,广东珠海 519085;
2. 上海开放大学 上海开放远程教育工程技术研究中心,上海 200433;
3. 英国斯特林大学 教育学院, 英国; 4. 爱尔兰都柏林城市大学 教育学院,爱尔兰)

[摘要] 随着机器学习和大数据的发展,如何利用人工智能技术优化教学过程和改进教学评估,已成为教育主管部门、科研人员、教育科技公司和教育工作者共同关注的话题。近年来,数十亿学习者在各种学习平台随时随地进行正式和非正式学习,形成了特定的活动轨迹和大量学习数据,应用人工智能技术对数字化学习环境中海量学习数据进行分析,给学生提供自动反馈和评估得到了广泛认可。因此,运用智能技术和大数据分析提高教育评估的效率和有效性也引起研究者越来越多的关注。但人工智能在高风险考试中应用的合理性和有效性备受质疑,其在形成性评估中应用的潜力和局限仍有待探讨。本文通过文献研究,从计算机测评领域相对成熟的两个自动测评系统:作文自动评分系统(AES)和计算机化自适应测验(CAT)的应用以及学术界对其存在问题的争论入手,对人工智能应用于教育评估的前景进行分析,并对人工智能和机器学习在形成性评估中的应用潜力和局限开展讨论。本研究认为,尽管人工智能的算法和大数据分析提高了自动测评系统的反馈速度和准确性,但其对学生深度学习和能力发展评价的应用价值仍然有限,教育评估中应用人工智能要掌握和了解计算机在总结性评估(如 AES 和 CAT 等)中的特征和局限,充分利用学习分析在形成性评估中的潜力,促进学生在数字化学习环境下创造力和自主学习能力的发展和培养。

[关键词] 人工智能;机器学习;作文自动评分;计算机自适应测验;学习分析;形成性评估
[中图分类号] G40-057 [文献标识码] A [文章编号] 1007-2179(2021)05-0004-11

近年来,随着人工智能技术的迅速发展,如何有效地利用这些技术推进教育评价改革,帮助教师全面了解和掌握学习者状态,并根据学习者产生的数据对学习状况进行准确评估,促进学习者综合能力和素质的提高,已成为教育界关注的问题。教育评估既要评估学生成绩,也要通过评估提高其能力;评估标准既要有筛选价值,也要涵盖个体的增值性评价(张生等,2021)。然而,许多面向教学的人工智能技术聚焦减轻教师负担,如自动完成作业批改、单元测试和考试评估等,随之也带来一些问题,如教师需要通过检查作业了解学生对知识的理解和掌握程度,为备课提供依据;人工智能技术能否对学生学习

[收稿日期]2021-06-05 [修回日期]2021-08-16 [DOI 编码]10.13966/j.cnki.kfjyyj.2021.05.001
[作者简介] 袁莉,教授,北京师范大学未来教育学院,研究方向:人工智能与未来教育(l.yuan@bnu.edu.cn);曹梦莹,上海开放大学上海开放远程教育工程技术研究中心,研究方向:学习分析、人工智能技术教育应用等(caomy@sou.edu.cn);约翰·加德纳(John Gardner),教授,英国斯特林大学教育学院,研究方向:教育评估政策与实践(john.gardner@stir.ac.uk);迈克尔·奥利里(Michael O'Leary),教授,爱尔兰都柏林城市大学教育学院教育评估研究、政策与实践中心,研究方向:国内外教育评估研究(michael.oleary@dcu.ie)。
[引用信息] 袁莉,曹梦莹,约翰·加德纳,迈克尔·奥利里(2021).人工智能教育评估应用的潜力和局限[J].开放教育研究,27(5):4-14.

作出准确判断,并避免数据偏见和算法黑箱。迄今为止,很少有证据表明,基于人工智能和大数据的学习分析确实改善了学习结果(Ferguson & Cllow,2017)。因此,运用人工智能对学生进行评估和考核应该格外慎重,否则不但不能促进学生学习,反而会带来很多负面影响。

受疫情影响,2020 年英国高考 A-level 和中考 GCSE 被取消。英国考试监管机构(Ofqual)公布了评定成绩的"标准化模型"——一种旨在避免分数膨胀的神秘算法,结果发现该算法对所有 A-level 学科成绩预测的准确性只有 60%,即近 40% 的学生成绩低于教师给出的预估分数(Ofqual,2020)。鉴于准确率过低,英国政府不得不取消人工智能提供的成绩,以教师评估为标准。英国考试监管机构希望通过使用人工智能算法提高学生成绩预测的准确性,但没有足够的数据,无法获得精准的预测模型。因此,承认历史数据的局限性对于考核评估算法应用尤为重要。

计算机应用于教育教学测评由来已久,特别是在作文批改、语言教学及数学等标准化学科考试中的应用尤为广泛。代表性的两个应用系统是"作文自动评分"(Automated Essay Scoring, AES)和"计算机自适应测验"(Computerised Adaptive Testing, CAT)。作文自动评分系统被广泛应用于大型考试的智能作文评分,如美国、英国、澳大利亚等国家研究生管理专业入学考试(GMAT)写作部分和一些作文批改评分平台(批改网、Grammarly 等),计算机自适应测验系统主要被应用于美国研究生入学考试 GRE 和 GMAT。本研究基于对这两大核心系统应用和相关研究的分析,阐述人工智能应用于教育评估面临的问题和挑战,以及运用学习分析对学生进行形成性评估的潜力和意义。

一、人工智能与教育评估

人工智能一词,最早是由计算机专家约翰·麦卡锡(John McCarthy)等学者提出来的,指运用计算机软硬件模拟人类某些智能行为的基本理论、方法和技术(黄欣欣,2017)。机器学习作为人工智能的分支,"是对计算机算法的研究,允许计算机程序通过经验自动改进"(Mitchell, 1997)。人工智能本质是机器"学习",即让计算机具有"学习"能力,通过对数据分析,"训练"出一个模型对新数据进行预测。因此,大量的数据和机器学习是人工智能的基础。麦肯锡公司(McKinsey Co.,2011)和高德纳公司(Gartner Glossary,2019)强调,"大数据"是需要新的处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。大数据概念起源于工程、量子物理和天文学等科学研究的大规模计算环境,数十亿份实验数据都要经过大规模算法分析以辨别模型、发现因果关系和预测可能的结果。大数据已应用到现代社会医疗诊断、消费趋势分析、天气预报等领域。以机器学习为核心的这些应用程序的"智能"特性体现在两个层面:无监督学习和监督学习。无监督学习指计算机从大量数据集中学习并更新识别模型。监督学习指计算机利用人工标注的数据训练预设好的模型,从而学习海量数据集中的相关性,并对已有模型完善和优化。

在教育评估中,机器学习概念同样适用。如果计算机可以检查学生"学"的效果,那么它就可以"教"学生知识,并对学生的知识掌握情况进行评估。简言之,如果训练计算机"学习"评估标准,评判学生对知识的理解和掌握情况(无论是书面的还是口头的),并按照既定标准对学生答案进行匹配,那么它就有应用于教育评估的潜力。但是,对教育工作者来说,掌握知识和能够理解并灵活应用知识存在差异。因此,教育评估中人的判断和机器的判断是不同的,这一差异在以计算机为基础的学生写作评估中尤为明显。

二、计算机在教育评估中的应用

(一)作文自动评分系统的应用

1. 作文自动评分系统的发展及其使用特征

基于计算机技术的作文自动评分系统近年被越来越多的高等教育机构用于评估学生写作。四个领先的商业作文自动评分系统包括项目作文评分(Project Essay Grading,PEG)、智能测评(Intellimetric)、智能作文评估软件(Intelligent Essay Assessor,IEA)和电子评分器(e-Rater)。项目作文评分主要应用于教师执照考试(Praxis)、GRE 考试和英语分级考试;智能测评用于 K-12 标准参照测验(K-12 norm-referenced test);智能作文评估软件主要用于

GMAT 考试;电子评分器用于 GMAT 考试和书面英语考试(Test of Written English,TWE)(Ben-Simon & Bennett,2007)。使用作文自动评分工具进行大规模评估优势明显,包括及时反馈、低成本和评分一致性。此外,作文自动评分工具应用于课堂评估,可以减少写作教师的工作量,并为每个学生提供即时反馈(Bull, 1999)。

早在 1966 年,项目论文评分的发明者、作文自动评分系统的先驱埃利斯·佩奇(Ellis Page)发表了题为《计算机作文评分的必要性》的文章,描述了使用计算机技术批改论文的想法,期望将英语教师从批改作文的负担中解脱出来,并预测随着自然语言处理技术的不断成熟,机器能够像人一样不断学习、理解和评估写作的不同指标(Wrech, 1993)。四十多年后,到 2007 年,本-西蒙和贝内特(Ben-Simon & Bennett,2007)分析了四个先进的商业作文自动评分系统发现,作文评估的基础(如所使用的指标)几乎没有什么变化,但是,这些系统使用的人工智能引擎更加精准,其容量和效率有了很大提高。例如,美国教育考试服务中心的电子评分器使用相对容易检测的四个评估指标(Deane,2013):语法(如主谓语不一致、代词误用、所有格错误等)、用词(如定语和介词错误、词的形态错误等)、技巧(如字母大小写、标点符号、拼写错误等)、风格(如单词重复使用等)。同时,该系统还使用较复杂的指标,如文章立意(如观点陈述、要点、讨论深度)、词汇的复杂性(如使用不寻常或复杂的单词等)、句子变化、资料引用、观点一致性(如是否偏离文章主题等)。

作文自动评分系统必须了解这些指标与学生作文质量之间的关系,才能给出相应的分数。因此,人们要先"训练"计算机,即计算机需要从大量的学生作文中识别这些指标并与人工分数进行比较。人工评分和范文越多,计算机给出的成绩与人工给出的分数越接近。因此,通常情况下,机器评分和人工评分在一定范围内能够达成一致。但对特别优秀或非常糟糕的作文,人工评分与机器评分会出现较大差异。例如,机器在检测语法、词汇和技巧方面比人工更准确迅速,但对写作的高级表现形式,如创造力、独特的论据、分析与综合能力等,因人工智能处在低级智能阶段,数据驱动的评价标准无法全面、深刻地反映学生写作的真实水平,相比之下,人工测评往往会在全面、深刻理解的基础上,给出较为准确的评价。

最初使用作文自动评分系统评估写作的研究,希望通过大规模、高效和准确的机器作文自动评分降低评估成本。但是,对于从事英语母语和第二外语教学的教师来说,与更为复杂和高级的写作构思相比,写作的技术指标是次要的。美国中学成功写作框架(CWPA et al., 2011)表明了写作过程建构的复杂性,强调教师要通过写作训练,发展学生的能力,为高等教育的学习做准备。它主要集中在以下方面:

- 好奇心:渴望更多地了解世界的愿望;
- 开放性:愿意接受新的生活方式和思维方式;
- 敬业度:全心投入和参与学习的意识;
- 创造力:用新的方法解决问题、研究和表达新的想法的能力;
- 持久性:对短期和长期项目保持兴趣和注意力的能力;
- 责任感:具有较强行动能力并能对行为后果进行判断;
- 灵活性:适应环境,能达到相应的期望或要求的能力;
- 多元认知:能够不断反思并提高自己的认知和文化认同能力。

可见,对于作文自动评分系统研发人员来说,面临的挑战是如何将高效、准确和低成本的写作评估转化为更加复杂和高级的指标。"项目作文评分"的修订版使用语法检查器和词性标记器等自然语言处理工具(Page, 1994; Page & Petersen, 1995)。2003 年发布的电子评分器第二版(Attali & Burstein, 2005;Burstein et al., 2004)确定了 12 条优秀作文的评价标准,涉及五个维度:语法、用词、技巧和风格,组织和发展,主题分析(即特定提示词汇),词的复杂性,论文长度(Attali & Burstein, 2005),有助于程序开发人员理解测评内容。

2. 作文自动评分系统的应用及学界对其有效性的质疑

近年来,人工智能技术在作文自动评分系统的应用,促使其不断完善,并在为学生和老师提供写作技能的快速反馈方面取得了重要进展。中国 2011 年上线的批改网是一款基于语料库和云计算技术的

机改作文系统,此系统以大学英语四级作文考试要求为模板,可以在1.2 秒内自动批改学生的英语作文,并给出分数、总评、按句纠错的批改反馈(张芳等,2021)。批改网提供的多维分析报告可以帮助教师及时了解学生的写作水平,指导课堂英语教学;教师可以基于批改网积累的语料库素材,分析学生作文数据,了解学生学习轨迹,从而辅助其科研。批改网也能激发学生英语写作兴趣,帮助他们提高英语写作能力(张芳等,2021)。但研究人员也发现一些问题,如"无法准确评估作文内在质量"(何旭良,2013),智能评价系统对"篇章结构和逻辑思维"无法做出准确判断(张芳等,2021)。作文自动评分系统对作文质量的测量仍以语法和语义内容为主,对思想、立意、创新性等深层属性的测量还远远不够(杨丽萍等,2021)。此外,浙江大学与杭州增慧网络科技有限公司联合开发了冰果英语智能作文评阅系统,但有研究人员(张仲德等,2013)通过实践发现该系统评阅得分有时与人工评阅出入较大,且程式化写作往往得分较高。这些与作文自动评分系统应用出现的问题一致,即系统只能对语言的表层现象加以评定,对深层次的语言现象评判不足,忽视了写作过程中的修辞、认知、思维发展过程等因素(张荔等,2016)。

长期以来,不少机构和研究人员反对用作文自动评分系统评估写作质量。例如,美国全国英语教师理事会参考大量与自动评估相关的文献,对作文自动评分提出质疑,如计算机无法识别或判断那些与高水平写作相关的元素(如逻辑性、清晰度、准确性、创新风格、更强的吸引力、不同的组织形式、说服类型、证据质量、幽默或讽刺,以及重复的有效使用等)。使用计算机评估学生写作剥夺了学生在写作中获得除特定要求外的任何发挥的机会;迫使教师忽略写作教学中最重要的因素,而去教学生一些毫无意义的东西。计算机按照编程特定的提示给作文打分,减少了教师通过评估改进写作教学的兴趣和创新的机会 (NCTE, 2013)。佩雷尔曼(Perelman,2012a)是主要的批评者之一,明确指出"作文自动评分简直就是荒谬的"。佩雷尔曼设计了被称为机器评分克星的自动语言生成器"Babel",以揭示自动评分的弱点和缺陷。机器评分克星通过运用计算机的疯狂填词游戏(Mad Libs)创作出毫无意义的作文,却能在机器评分中获得高分。佩雷尔曼(Perelman,2018)认为计算机只能计算,不能真正理解意义,往往只是按照设定的算法执行。学生一旦掌握了计算机批改的技巧就可以在考试中通过使用大量复杂词汇、复杂句子和关键短语来愚弄算法。他分析了许多作文自动评分系统,并以此作为批判作文自动评分的研究证据,认为作文自动评分鼓励"使用难懂的、晦涩的和矫饰的语言"(Perelman,2012b,p126),并严厉地驳斥了所有作文自动评分"不理解意义,也没有感知能力"(Perelman,2012b,p125),指责它们在评估中过度强调作文长度等(Perelman,2014)。其他语言专家,如康登(Condon,2013)支持利用"作弊"的方式查找作文自动评分的弱点,但认为这种做法不能解决核心问题,只是将注意力转移到不相关的争论上,即"把焦点放在作文自动评分提供的分数是否与人工评分一致上,实际上是将两个不相关的度量认为是相关的"。迪恩(Deane,2013)指出作文自动评分系统专注于"衡量作品的最终质量",人工评分更关注学生的写作技能,即阅卷人在阅读学生作文时侧重于理解写作者的思想,而作文自动评分系统强调识别文本中的标识。即使人工和计算机评分结果一致,它们的内在含义差异也很大:"没有作文自动评分系统可以达到在理解作者意思的基础上进行评估"(Deane,2013)。迪恩(Deane,2013) 认为,写作的目的是为了与人交流。如果学生的第一次写作经历是对着一台机器,这可能意味着写作不被视为人际交流,反过来可能降低评估的有效性。此外,由于计算机写作评分的算法是根据过去的经验和知识设定的,我们也无法知道评分中是否包含了特定类型的偏见。因此,他强烈反对在入学考试、分班或期末测验等高风险考试中使用作文自动评分系统进行写作评估。

从短期看,作文自动评分系统可以帮助老师减轻批改负担,及时给予学生反馈(张荔等,2016)。但从长远角度看,学生容易在自动批改评分中形成写作的固化思维,影响真正的写作,而且面对计算机评估的写作本身违背了写作的社会属性(Deane,2013)。如果大学的作文评估都是机器评分,很可能导致高中的评分系统和写作训练都基于自动评分,从长远来看,这不利于通过写作培养学生的交流能力。作文自动评分系统的发展,除了扩展评估指

标的范围,还需要加强对写作结构的理解。另外,如何将作文自动评分系统用于写作教学支持教学创新,帮助学生提高认知能力并与社会实践相结合,从而提高学生的写作水平也非常重要(Deane,2013)。汉姆普-莱昂斯和康登(Hamp-Lyons & Condon,2000)研究证明了将写作评估视为一个涉及迭代、学习和多方利益相关者互动过程的重要性。edX、麻省理工学院和哈佛大学等相继使用基于机器的作文自动评分系统评估慕课书面作业。Coursera 在此基础上,增加以人为基础的"校准的同行评审",来对学生写作进行评分并提供反馈(Balfour,2013)。这种模式将作文自动评分系统用于慕课写作教学,给学生提出反馈和修改意见,然后使用校准的同行评审进行最终评估(Sandeen,2013)。这使得一些简单问题能尽早得到纠正,从而改善作文质量,且比单一的人工评估或机器评估更为准确和高效(Balfour,2013)。

(二)计算机自适应测验的应用和面临的挑战

1. 计算机自适应测验

计算机自适应测验具有设计标准化和操作简单等特性,但题目选择与评估的算法和技术复杂。它与作文自动评分系统的主要区别是:作文自动评分系统的机器学习试图模仿人工评分标准进行判断,计算机自适应测验使用一系列测试题目决定考生能力,标准是预先设定的,即根据题目难度判断考生的知识掌握水平。在计算机自适应测验评估中,计算机根据考生对测试题的反应有目的地选择下一题,直到可以评估考生是否达到被测能力的极限。诺一琼斯(Noijons,1994)将自适应测试定义为"在计算机帮助下诱发和评估语言表现的综合程序,包括生成测试、与参与者互动、反馈评价"。计算机自适应测验与纸笔测试、一般计算机测验的区别是具备估算被试者的能力、不依赖于测试题目的特性;可以根据题目的信息量,选择与被试者能力相匹配的题目;测试管理灵活,测试结果可以立即显示,减少考生考试焦虑;与传统测验相比,计算机自适应测验可节省成本(赵茜等,2020;Mulkern, 1998;Weiss,1990;Straetmans & Eggen, 1998)。通过计算机自适应测验弹性水平策略,考生通常被给予适合其特定水平的测试,不需要回答对他们来说太难或太易的问题(Larson & Madsen,1985)。

计算机自适应测验评估是个迭代过程(Rudner,1998),算法通常包括:1)根据预估的考生能力和水平,对题库的所有考题进行评估,确定适合考生水平的题目,选定合适的考题,由考生回答;2)根据考生答案重新计算其能力和水平;3)重复 1 到 2 步骤,直到确定考生最终能够达到的标准。简言之,当计算机选择第一道题时,其预期是考生可以给予正确答案。计算机自适应测验为了确定适当的难易水平,测试前问考生一些代表其能力水平的指标问题,然后根据这些数据决定第一道考题。在缺乏了解考生能力的情况下,计算机自适应测验通常会在尝试过程中提供难度较低的题目(如 70% 的考生可以回答正确),然后逐步加大难度,直到能够判断考生可以达到的能力和水平的上限,即"测评的终结点"。在相对简单的及格测评设计中,"测评的终结点"是考生超过及格分数线或没有达到及格分数线。因此,计算机自适应测验常用于总结性评估,如高风险的考试测评(大学入学和就业等)。

2. 计算机自适应测验的应用及存在的问题

教育评估中两种著名且成功的基于计算机自适应测验的考试是美国研究生管理入学委员会的 GMAT 和美国教育考试服务中心的 GRE 考试。全球商学院都使用 GMAT 成绩挑选 MBA 申请者。当学生开始 GMAT 考试时,计算机假设学生有一个平均分数,给出一道中等难度的题目。如果学生回答正确,电脑会给出更难的考题,并增加难度。反之,如果回答错误,计算机会降低难度。学生分数由预设的算法决定,该算法不仅根据学生答案的对与错,还根据他们回答问题的难度计算学生能力。此外,GMAT 写作评估采用六分制,由人和计算机共同完成,学生考试结束后可立即收到非官方的 GMAT 成绩,并选择保留或取消他们的 GMAT 成绩(KAPLAN,2020)。GRE 科目测试旨在测量学生特定学科领域(如数学、历史或英语文学)的成绩(Stocking et al. , 2000)。

美国 K-12 基础教育考试服务执行主任马特森认为,计算机自适应测验的最大优势是学生作答试题的难度与学生能力匹配。这一优势能带给学生更有益的考试体验,能力低的学生不会遇到无法回答的题目,因而不会备受打击;能力高的学生不会因回答过于简单的试题而失去兴趣,这能够提高学生的

测验参与度和动机(王超,2017)。

由于计算机在语言测试中的重要作用,很多学者开始比较计算机测试与笔试的差异。侯赛尼等(Hosseini et al.,2014)在慕尼黑大学随机抽取了106名伊朗英语学习者计算机测试与笔试的成绩。结果表明,参与者的笔试成绩优于计算机测试成绩。此外,其他学者也发现,受试者书写测试比计算机测试表现更好(Coniam, 2006; Cumming et al.,2006; Salimi et al., 2011;Mazzeo et al.,1991)。计算机自适应测验基于写作反应理论模型,此模型不能用于所有写作,因为它不适用于开放式问题和不容易校准的写作(Rudner, 1998)。计算机自适应测验的另一个缺点是,考生无法在测试结束前退回去更改答案,因为下一道题目是根据前一道题的结果给出的(Rudner, 1998)。安全性是计算机自适应测验的另一个重要问题。如果题库被用来测试考生的知识,在测试过程中,有些题目可能比其他题目更频繁地出现,这些题目可能被记住并传递给其他考生(Wainer & Eignor,2000)。

另外,莫内塔-克勒等(Moneta-Koehler et al.,2017)反对美国教育考试服务中心将GRE考试分数作为研究生入学的唯一标准。他们以范德堡大学国际研究生项目为例发现,GRE成绩不能预测学生能否顺利博士毕业或发表更多论文。因此,他呼吁生物医学科学招生委员会应考虑最大限度地少用GRE分数预测学生学习成绩和创作力。在过去十年,计算机自适应测验应用结果的有效性几乎没有变化。有研究者担心,其设计可能会限制妇女和少数民族人员进入科学等关键领域,例如,米勒和斯坦森(Miller & Stassun,2014)指出,女性候选人的GRE成绩平均比男性少80分,非裔美国人比白人少200分。

三、学习分析与教育评估

(一)大数据与学习分析

大规模数据分析是人工智能用于教育评估的基础,例如,作文自动评分系统基于众多人工测评对大量学生写作进行评判,计算机自适应测验基于许多学生多项选择题测试结果。大数据在这些评估中的共性,也就是机器学习在科学、医学和技术发展中的应用都是过程数据,即使用应用程序可以有目的或偶然地在线捕获数据。分析这些学习和评估数据的技术通常被称为学习分析(有时称为教育数据挖掘)。学习分析是对与学习活动相关的学习者数据采集、分析和干预的过程(Gaševic et al.,2015),最常使用的定义是第一届国际学习分析和知识会议(Long & Siemens,2011)提出的:"……关于学习者及其学习数据和情境的测量、收集、分析和报告,目的是理解和优化学习及其发生的环境"。这些目标的实现很有价值。埃利斯等(Ellis et al.,2013)认为,大部分学习分析活动通常专注于预测,如识别大数据中与特定结果相关的模型,以提高学生的考试成绩。然而,越来越多的人认识到课堂或培训环境中形成性评估的重要性,因此,如何使用教学大数据进行智能分析,帮助学习者形成性地自我调节从而改进学习显得尤为重要。

柯普等(Cope & Kalantzis,2016)将学习过程中机器评估收集的各种数据分为结构化数据(即由计算机专门预测和捕获的数据)和非结构化的偶然数据,如迪赛尔博和贝伦斯(DiCerbo & Behrens,2014)提出的"数据废气"(data exhaust)。后者包括点击数、日志文件等时间、击键和编辑历史记录或"轨迹"。通过分析,我们可以了解学生是如何解决问题、所犯的错误和所做的修改、对概念的误解,以及面对学习进展缓慢或没有进展时的反应和应对能力等。此外,他们通过摄像机、录音机、智能手表和手环等对学生进行眼动追踪、面部表情、身体姿势、手势和课堂发言等方面的数据收集,以此作为教学活动过程的评价参数,如分析同伴互动甚至情感状态,包括困惑、沮丧、无聊和参与度等。对智能导师系统(Intelligent Tutoring Systems)捕获的数据进行深入分析可以帮助教师更好地了解学生并改进教学策略。莫莱纳尔等(Molenaar et al.,2021)将计算机自适应测验的绩效评估数据运用到自适应学习平台,帮助教师选择合适的学习资源(教学材料)和确定适合于不同学生的问题。教师如果能够及时分析这些不同类型的数据,了解学生如何对待学习任务以及他们在哪些方面掌握了知识,哪些方面面临困难,可以作为形成性反馈及时提供给学生。例如,Embrace系统使用动态跟踪数据,为学生可视化在线阅读理解任务的表现提供即时形成性反馈(Walker et al.,2017)。阿尔乔哈尼等(Aljohani &

Davis,2013）使用手机数字仪表板让学生查看测验结果,为学生提供有关学科整体成绩的及时反馈,并按照布鲁姆分类学对学生认知水平进行评估。上述例子虽然都是按照学校要求对学生的学习进行评估,是一种描述性的可视化分析,但也展示了人工智能既有用于总结性评估实时反馈的特性,也有用于形成性评估的潜力。

（二）学习分析与形成性评估

蒂勒等（Thille et al.,2014）认为可以从三方面对大规模评估数据进行多样化评价:1）连续性（始终自动收集数据）;2）反馈性（为老师和学生提供实时的数据分析、解释和报告）;3）多样性（可以收集点击数、日志文件,以及自动记录的多种数据）。随着新的评估测试和测试群体的不断增长,大规模评估数据分析,可以提供全面的学习“轨迹”建模（“专家系统”）,从而将个体学生的学习表现与以大多数学生群体为基础所建立的典型表现模型进行比较。其中非常重要的一点是“专家系统”对学生的评价是由系统自动给出的“提示”,并以脚手架教学或形成性评估干预的形式,在学生解决问题的过程中适当地提供实时反馈。另外,蒂勒等（Thille et al.,2014）还指出,在某些情况下,“专家系统”提出的一个步骤到另一个步骤的学习策略与学生的实际轨迹和决策不符,通过更好地了解学生不同的学习方法,能够更好地改进专家系统,提供更加精准的学习路径推荐。如果要在实施过程中向学生提供有意义的反馈,就要求学习分析及时准确,这就是柯普等（Cope & Kalantzis,2016）提出的学习分析结束了“教学和评估分离的历史”,并具有“随时提供反馈”的潜力。然而,将这些形成性评估技术从蒂勒等的小规模在线环境（智能导学系统、编码实践和应用慕课）迁移到更多样化的学习环境,可能因数据捕获的挑战性,前景有限。

毫无疑问,形成性评估在教育中的作用越来越重要,人们也越来越对运用大数据和智能分析帮助学生在在线学习环境发展自主学习能力感兴趣。自主学习（self-regulated learning,SRL）是一种复杂的现象,受每个人的个性特征、学习习惯和学习环境的影响。例如,西奇内利等（Cicchinelli et al.,2018）确定了与学生自我计划和监督相关的指标,这些指标与学生的学习和考试分数直接相关。另外,贾维拉等（Jarvela et al.,2020）认为,最近掀起的学习分析热潮,通过对学生的情绪、社交和认知等进行跟踪,使以前完全不透明的自主学习过程变得清晰可见,特别是在协作学习环境中更是如此。基于这些发展,美国高考（ACT）测试研究小组采用移动应用程序 Companion 进行实时测试,对学生学习进度和成果进行及时分析和反馈（ACTNext,2020）。该系统使用“动态认知诊断模型和机器学习算法”分析测试结果和学习资源的使用情况,并承诺可以通过亚马逊的 Alexa 和苹果的 Siri 等工具融入学生日常生活。无疑,学习分析反馈对任何学习环境的个性化形成性评估都非常有用,特别是在慕课或其他大规模在线学习环境中,成千上万的学习者共同学习一门课程。在这种情况下,自主学习变得尤为重要,因为外界的形成性评估和及时性反馈受到教师当面答疑甚至异步互动成本的限制。学伴评估,如果准确的话,可以帮助解决慕课形成性评估问题（Garcia-Martinez et al.,2018）,但简森等（Jansen et al.,2020）提出,使用学习分析和内置干预,即按照自主学习设计的课程资源视频,可以提高慕课课程完成率,越来越多的研究证实了这一结论（例如,Jarvela et al.,2020;Martin & Ndoye,2016;Tempelaar et al.,2013;Gutierrez Rojas & Crespo Garcia,2012）。正因为如此,对于在线学习开发人员来说,能够在慕课中提供自动化且具有成本效益的个性化形成性评估和反馈是首要目标。

如果教师对学生学习的反馈主要用于改善教学质量而非与其他学生比较,并关注学生个体进步,那么学生可以通过形成性评估反馈知道下一步该做什么,从而产生控制自己学习的满足感,激发学习动机和有效完成学习目标（Brookhart,2008;Crooks,1988）。学习分析与形成性评估都具有及时反馈的特征。形成性评估帮助学生了解学习过程中的表现,而学习分析为学生提供了有关个人表现的信息,并可提高教师对不同学生表现的认识深度（Aljohani & Davis,2013）。布莱克等（Black & William,1998）总结了 250 多项形成性评估研究后发现,形成性评估是课堂工作的重要组成部分,可以提高学生学习成绩。基于学习分析的形成性评估对于教学过程的评价更为科学、系统和智能,进而可实现数据驱动下的“以学定教和因学定导”（毕鹏晖,2021）。

斯佩克特等(Spector et al.,2016)强调了个性化形成性评估的重要性,但发现在某种程度上,智能导学系统可能会“一刀切”。这些系统虽然能够判断学生存在的问题并帮助学生弥补不足,但对存在同样问题的学生给出的建议都一样。他们认为,学习分析系统可以通过对学生进行更深入的分析,并结合各种绩效分析技术提供反馈,以满足不同学生的需求。个性化学习分析可以通过捕获学生学习过程的“隐性”数据,如监测那些连续的、内在的和不明显的行为,识别学生的学习习惯,并结合“显性”的学生画像,包括他们的爱好、兴趣和态度等外在数据,对学生的学习提供及时、有效的干预。但是,与营销和其他个人在线活动分析一样,个性化学习分析使用不当可能会引发伦理问题。这种反馈机制“尚未能够大规模和持续运用”(Spector et al.,2016)。

四、结 语

尽管计算机算法和大数据分析技术有了突飞猛进的发展,但目前教育评估中人工智能运用的基本原理和功能几乎没有明显变化,即以总结性评估为主的计算机自动测评而非以学习过程数据为基础的形成性评价为主。不可否认,人工智能评估的效率、速度和精准都有了显著提高,能够达到与人工评估非常相近的结果,甚至在某种程度上,特别是在标准化测试和自适应教学系统中发挥了计算机评估和反馈更为迅速、客观、准确的优势。通过分析作文自动评分系统和计算机自适应测验的发展状况,本研究发现这些教育评估的核心应用已从机器学习的技术进步中受益,人工智能技术仍会不断迭代。但是,寄希望于运用人工智能系统取代人工评估仍然是不现实的。将基于学生能力特征分析的计算机自适应测试、模仿人工判断的作文自动评分系统,与复杂学习过程的各种数据整合,可以提高教育评估的效率和有效性。精准的学习分析可以通过手机等移动设备为学生在慕课和智能导学系统的学习提供合适和有目的的形成性评估反馈,支持学生的自主学习。随着计算机硬件技术和软件系统的不断进步,人工智能在教育评估中的应用,特别是在形成性评估中的应用潜力值得期待,但数据采集和算法方面的局限仍有待探讨。

[参考文献]

[1] ActNext (2020). Educational Companion[OL]. Retrieved from https://actnext.org/research-and-projects/holistic-learning-mobile-app/.

[2] Aljohani, N. R., & Davis, H. C. (2013). Learning analytics and formative assessment to provide immediate detailed feedback using a student centred mobile dashboard[C]//Proceedings of the Seventh International Conference on Next Generation Mobile Apps, Services and Technologies, IEEE:262-267.

[3] Attali, Y., & Burstein, J. (2005). Automated essay scoring with e-rater v.2.0.[R]. ETS Research Report Series:ETS RR-04-45.

[4] Balfour, S.P. (2013) Assessing writing in MOOCs: Automated essay scoring and calibrated peer review[J]. Research & Practice in Assessment, 2013, 8(1):40-48.

[5] Ben-Simon, A., & Bennett, R. E. (2007). Toward a more substantively meaningful automated essay scoring [J]. Journal of Technology, Learning and Assessment, 6(1):47.

[6] 毕鹏晖(2021).基于学习分析的形成性评估融入在线混合式教学的行动研究[J].中国多媒体与网络教学学报(上旬刊),(3):17-19.

[7] Black, P.,& Wiliam, D. (1998). Assessment and classroom learning[J]. Assessment in Education: Principles, Policy & Practice, 5(1): 7-74.

[8] Brookhart, S. M. (2008). How to give effective feedback to your students. Alexandria[M]. VA: Association for Supervision and Curriculum Development:121.

[9] Bull, J. (1999). Computer-assisted assessment: Impact on higher education institutions[J]. Educational Technology & Society, 2(3):123. Retrieved from http://www.ifets.info/journals/2_3/joanna_bull.pdf.

[10] Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service [J]. AI Magazine, 25(3): 27-36.

[11] Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., et al.. (2018). Finding traces of self-regulated learning in activity streams[C]//Proceedings of the 8th International Conference on Learning Analytics and Knowledge. New York, USA: ACM: 191-200.

[12] Condon, W. (2013). Large-scale assessment, locally developed measures, and automated scoring of essays: Fishing for red herrings? [J]. Assessing Writing,18(1): 100-108.

[13] Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test[J]. ReCALL,(18):193-211.

[14] Cope, B., & Kalantzis, M. (2016). Big Data comes to school: implications for learning[J]. Assessment and Research. AERA Open, 2(2): 1-19.

[15] Crooks, T. J. (1988). The impact of classroom evaluation practices on students[J]. Review of Educational Research, 58(4):438-481.

[16] Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated prototype writing tasks for new TOEFL (TOEFL Monograph No. MS-30)[R]. Princeton, NJ: Educational Testing Service:1-87.

[17] CWPA, NCTE, & NWP (2011). National Framework for success in postsecondary writing[R/OL]. Council of Writing Program Administrators, the National Council of Teachers of English, and the National Writing Project. Retrieved from http://wpacouncil. org/files/framework-for-success-postsecondary-writing. pdf.

[18] Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct [J]. Assessing Writing, 18(1): 7-24.

[19] DiCerbo, K. E., & Behrens, J. T. (2014). Impacts of the digital ocean on education[R]. London: Pearson:1-44.

[20] Ellis, C. (2013). Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics[J]. British Journal of Educational Technology, 44 (4): 662-664.

[21] Ferguson, R., & Clow, D. (2017). Where is the evidence? [C]//Proceedings of the 7th International Conference on Learning Analytics and Knowledge (LAK '17). 13-17 March 2017, Vancouver, BC, Canada. New York: ACM:56-65.

[22] Gaševic, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning[J], Tech Trends, 59 (1): 64-71.

[23] Gartner Glossary (2019). Big Data[EB/OL]. Retrieved from https://www. gartner. com/en/ information-technology/glossary/big-data.

[24] Gutierrez, I., & Crespo Garcia, R. M. (2012). Towards efficient provision of feedback supported by learning analytics[C]//12th IEEE Conference on Advanced Learning Technologies (ICALT), IEEE: 599-603.

[25] Hamp-Lyons, L., & Condon, W. (2000). Assessing the portfolio: Principles for practice, theory, and research [M]. Creskill, NJ: Hampton Press:1-216.

[26] 何旭良(2013). 句酷批改网英语作文评分的信度和效度研究[J]. 现代教育技术,23(5):64-67.

[27] Hosseini, M., Zainol Abidin, M., & Baghdarnia, M. (2014). Computer-based tests (CBT) and paper and pencil tests (PPT) among English Language Learners in Iran[J]. Procedia-Social and Behavioral Sciences, 98: 659-667.

[28]黄欣欣(2017).探析人工智能在人类生活中的应用及未来发展态势[J].中国新通信,19(21):87.

[29] Jansen, R. S., Leeuwen, A. V., Janssen, J., Conijn, R., & Kester, L. (2020). Supporting learners´ self-regulated learning in Massive Open Online Courses [J]. Computers and Education, 146, 103771.

[30] Jarvela, S., Gasevic, D., Seppanen, T., Pechinizky, M., & Kirschner, P. (2020) Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning [J]. British Journal of Educational Technology, 51(6): 2391-2406.

[31] KAPLAN (2020). What is the GMAT? [EB/OL]. https://www. kaptest. com/gmat/what-is-the-gmat.

[32] Larson, J. W., & Madsen. H. S. (1985). Computer-adaptive language testing: Moving beyond computer-assisted testing [J]. CALICO Journal, 2(3): 32-6.

[33] Long, P., & Siemens, G. (2011). Penetrating the fog[J]. Educause Review, 46(5):31-40.

[34] Martin, F., & Ndoye, A. (2016) Using learning analytics to assess student learning in online courses [J]. Journal of University Teaching & Learning Practice, 13 (3):7.

[35] Mazzeo, J. Druesne, B., Raffeld, P., Checketts, K., &Muhlstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations (Report No. 91-5)[R]. Princeton, NJ: ETS:24.

[36] McKinsey Co. (2011). Big data: The next frontier for innovation, competition and productivity[EB/OL]. https://www. mckinsey. com/ ~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our% 20Insights/Big% 20data% 20The% 20next% 20frontier% 20for% 20innovation/MGI_big_data_exec_summary. ashx.

[37] Mitchell, T. M. (1997). Machine Learning [M]. McGraw-Hill Education, New York:1-414.

[38] Miller, C. & Stassun, K. (2014). A test that fails[J]. Nature ,510(7504): 303-304.

[39] Molenaar, I., Horvers, A., & Baker, R. S. (2021) What can moment-by-moment learning curves tell about students' self-regulated learning? [J]. Learning and Instruction. 72(1):101206.

[40] Moneta-Koehler, L., Brown, A. M., Petrie, K. A., Evans, B. J., & Chalkley, R. (2017) The Limitations of the GRE in Predicting Success in Biomedical Graduate School[J]. PLoS ONE 12(1): e0166742. Retrieved from https://journals. plos. org/plosone/article?id = 10. 1371/journal. pone. 0166742.

[41] Mulkern, A. E. (1998). Frequently Asked Questions about Computer-Adaptive Testing (CAT)[EB/OL]. [2005-0411]. http://carla. acad. umn. edu/CATFAQ. html.

[42] NCTE (2013). Position Statement on Machine Scoring. National Council of Teachers of English[EB/OL]. [2013-04-20]. Retrieved from http://www2. ncte. org/statement/machine_scoring/.

[43] Noijons, J. (1994). Testing computer assisted language tests: Towards a checklist for CALT [J]. CALICO Journal, 12(1): 37-58.

[44] Ofqual (2020). Ofqual's guide to the 2020 AS and A level results in England[EB/OL]. [2020-08-13]. https://www. gov. uk/government/news/guide-to-as-and-a-level-results-for-england-2020.

[45] Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test [J]. Phi Delta Kappan, 76(7): 561-565.

[46] Page, E. B. (1994). Computer grading of student prose, u-

sing modern concepts and software [J]. Journal of Experimental Education, 62(2): 127-142.

[47] Perelman, L. (2012a). Mass marketing assessment of writing is bullshit[M]//Elliott, N. & Perelman, L. (Eds). Writing Assessment in the 21st Century: Essays in Honor of Edward M. White. New York:Hampton Press:425-437.

[48] Perelman, L. (2012b). Construct validity, length, score and time in holistically graded writing assessments: The case against automated essay scoring (AES)[M]//Bazerman, C., Dean, C., Early, J. S., Lunsford, K. J., Null, S., Rogers, P. & Stansell, A. (Eds). International Advances in Writing Research: Cultures, Places, Measures. Parlor Press:121-131 .

[49] Perelman, L. (2014). When 'state of the art' is counting words[J]. Assessing Writing, 21: 104-111.

[50] Perelman, L. (2018). Interview on his Babel Generator Tovia Smith podcast: More states opting to 'robo-grade' student essays by computer [N]. [2018-06-30]. https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer?t=1564600373190.

[51] Rudner, L. M. (1998). An online, interactive, computer adaptive testing tutorial [EB/OL]. [2015-04-16]. http://EdRes.org/scripts/cat.

[52] Salimi, H., Rashidy, A., Salimi, A. H., & AminiFarsani, M. (2011). Digitized and non-digitized language assessment: A comparative study of Iranian EFL language learners[C]//International Conference on Languages, Literature and Linguistics. Singapore: IACSIT Press: 607.

[53] Sandeen, C. (2013). Assessment's place in the new MOOC world [J]. Research & Practice in Assessment, 8(1): 5-12.

[54] Spector, J. M., Ifenthaler, D., Sampson, D., Yang, L., Mukama, E., & 12 additional authors (2016). Technology enhanced formative assessment for 21st century learning[J]. Educational Technology and Society, 19(3): 58-71.

[55] Straetmans, G. J. M., & Eggen T. J. H. M. (1998). Computerized adaptive testing: What it is and how it works [J]. Educational Technology, 38(1):45-52.

[56] Stocking, M. L., Smith, R., & Swanson, L. (2000). An investigation of approaches to computerising the GRE subject tests[M]. NJ: Education and Testing Services. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2000.tb01827.x.

[57] Tempelaar, D. T., Cuypers, H., van de Vrie, E., Heck, A., & van der Kooij (2013) Formative assessment and learning analytics [C]//Proceedings of LAK'13, Leuven, Belgium:205-209.

[58] Thille, C., Kizilcec, R., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data-enriched assessment [J]. Research and Practice in Assessment 9: 5-16.

[59] Wainer, H., & Eignor, D. (2000). Caveats, pitfalls and unexpected consequences of implementing large-scale computerized testing[M]//H. Wainer et al. (Ed.), Computer adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates:271-299.

[60] Walker, E., Wong, A., Fialko, S., Restrepo, M. A., & Glenberg, A. M. (2017) EMBRACE: Applying Cognitive Tutor Principles to Reading Comprehension[M]//André E., Baker R., Hu X., Rodrigo M., du Boulay B. (eds). Artificial Intelligence in Education. AIED 2017:578-581.

[61] 王超,陆宏(2017).计算机自适应测验的研究策略与应用实践[J].现代教育技术,27(12):44-49.

[62] Weiss, D. J. (1990). Adaptive testing[M]//H. J. Walberg & G. D. Haertel (Eds.). The International Encyclopedia of Educational Evaluation. Oxford: Pergamon Press: 454-458.

[63] Wresch, W. (1993). The imminence of grading essays by computers:25 years later [J]. Computers and Composition 10(2): 45-58.

[64]杨丽萍,辛涛(2021). 人工智能辅助能力测量:写作自动化评分研究的核心问题[J]. 现代远程教育研究,33(4):51-62.

[65] 赵茜,马力,温红博(2020). PISA2018 阅读素养计算机化自适应测试的技术与方法探析[J].中国考试,(11):74-78.

[66] 张芳,高佳佳(2021). 批改网对英语专业学生英语写作结果的影响[J]. 教学研究,44(1):59-65.

[67] 张荔, Mark Warschauer,盛越(2016). 自动写作评阅反馈系统研究述评与展望[J].当代外语研究,(6): 54-61.

[68] 张生,王雪(2021).齐媛.人工智能赋能教育评价:"学评融合"新理念及核心要素[J].中国远程教育,(2):1-8+16+76.

[69] 张仲德,李雅萍(2013). 基于文本基础上冰果智能英语作文的分析与研究[J].长春大学学报,23(8):1047-1050.

(编辑:魏志慧)

The Potentials and Limitations of Artificial Intelligent in Education Assessment

YUAN Li[1], CAO Mengying[2], GARDNER John[3] & O'LEARY Michael[4]

(1. *College of Education for the Future, Beijing Normal University, Zhuhai*519085, *China*;
2. *Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, Shanghai* 200433, *China*;3. *Faculty of Social Sciences, University of Stirling, Stirling, UK*;
4. *Centre for Assessment Research, Policy and Practice in Education, Institute of Education, Dublin City University, Dublin, Ireland*)

Abstract: *With the development of machine learning and big data, how to use Artificial Intelligence (AI) to optimize teaching and learning processes and improve the quality of educational assessment has become a hot topic among educational policymakers, educational researchers, technology developers, and educators. The latest emergence of MOOCs platforms has provided opportunities for learners worldwide to learn anytime and anywhere, which has generated a large amount of learning data to help identify patterns of various learning activities conducted by learners. It is widely recognized that AI can capture data in the learning process in a digital learning environment. It can be analyzed to provide students with instant feedback and evaluate their learning. Therefore, the use of AI and big data to improve the efficiency and effectiveness of educational assessment have attracted great attention. However, in practice, the rationality and effectiveness of the application of artificial intelligence in high-stakes tests are challenged, and the potential and limitations of AI in formative assessment need to be further explored. In evaluating the state of play of Artificial Intelligence in formative and summative educational assessment, this paper offers a critical perspective on the two core applications: Automated Essay Scoring systems and Computerized Adaptive Tests. It also, along with the Big Data analysis approaches to machine learning that underpin them. In this regard, this paper showed that AI had improved the efficiency, speed, and sophistication of summative assessment, especially in analyzing large-scale assessment process data. However, their application value for deep learning of life and evaluation of capacity development is still limited. Therefore, when applying AI in educational assessment, it is important to understand the characteristics and limitations of computerized summative assessment applications (e.g., AES and CAT) and explore the potential of appropriate, and purposeful learning analytics for formative assessment to support learners in developing their ability and skills on creativity and self-regulated learning in a digital world.*

Key words: *artificial intelligence; machine learning; automated essay scoring; computerized adaptive tests; learning analytics; formative assessment*