·课程理论与教学改革·

为高阶学习而评价

——表现性评价及其在高等教育学习成果评估中的应用

彭 湃①
(华中科技大学 教育科学研究院,湖北 武汉 430074)

摘 要:我国大学教育在促进高阶学习方面与先进国家差距很大,在学习成果评估中对其重视也不足,这已影响创新型人才的培养。表现性评价是一种评价高阶学习的方法。支撑它的认知理论是建构主义,其指向的教育目标是高阶的学习成果,如分析、综合、评价能力以及批判性的思维能力,与之相适应的课程教学是以学生为中心的。对大学学习评价(CLA)的样题分析显示,表现性评价具备认知高阶性、情境真实性、问题复杂性等诸多特色。严格的标准化程序保障了大规模测评中表现性评价的质量。我国高等教育需要表现性评价。

关键词:表现性评价;学习成果;高阶学习;建构主义;大学学习评价
中图分类号:G420 **文献标志码**:A **文章编号**:1000-4203(2015)11-0055-09

Assessment for Higher-order Learning: Performance Assessment and Its Application in Large-scale Assessment of Student Learning Outcomes in Higher Education

PENG Pai
(*School of Education, Huazhong University of Science & Technology, Wuhan* 430074, *China*)

Abstract: There was a large gap between universities in advanced countries and in China in terms of higher-order learning, which was not emphasized in student learning outcome assessment. It had a negative impact on producing creative talents. Performance tasks (PT) could be used to assess higher-order learning, of which the underlying cognitive theory was constructivism. PT assessed higher-order learning outcomes, such as abilities of analysis, synthesis, evaluation, and critical thinking skills. It aimed student-centered curriculum and teaching. An analysis of CLA sample task indicated that PTs in large-scale assessment had the following features: higher-order cognition, real contexts, and complex problems. Strict standardized procedures ensured the quality in large scale assessment. Higher education in China needed PT.

Key words: performance assessment; learning outcomes; higher-order learning; constructivism; Collegiate Learning Assessment

① **收稿日期**:2015-09-25
基金项目:国家社科基金项目(14CGL047)
作者简介:彭 湃(1981—),男,安徽六安人,华中科技大学教育科学研究院讲师,教育学博士,从事教育评价研究。

一、引 言

创新是我国国家发展战略的追求，培养创新型人才是我国高等教育的重要目标。反映在大学教学领域中，培养学生的高阶思维能力应该是一个极为重要的教学目标。然而，当前我国大学教学对高阶学习的重视不足。有研究表明，在课程教学方面，我国本科课程在高阶学习方面与美国差距非常大；在学习成果评估方面，“加强短时记忆的背诵迎考是本科学习的难忘经历”[1]，“考试目的忽视培养思考能力和创造能力”[2]。可见，我国高等教育学习成果评估中缺少对于高阶学习的关照。

表现性评价是对高阶学习进行评价的重要方法。它要求学生“在某种特定的真实或模拟情境中，运用先前所获得的知识完成某项任务或解决某个问题，以考察学生知识与技能的掌握程度、问题解决、交流合作和批判性思考等多种复杂能力的发展状况”。[3]表现性评价因为与传统的“选择—反应”测试（即所谓客观性试题）的差异显著，经常被称为“替代性评价”（alternative assessment），也有人将其称为“真实性评价”（authentic assessment），其意在强调评价任务中问题或情境的真实性。表现性评价已在国际范围的高等教育学习成果评估中得到了应有的重视。我国研究者正在密切关注这一领域，但对具体评价方法的深度解读较为少见。由于评价学生的学习成果一向被视为大学教师个人的“自留地”，而他们很少接受专业指导，对于如何科学有效地开展评价了解不多。无独有偶，国外学者也发现“很少有教师在给学生的表现和成就打分方面受过专业训练”。[4]“在高等教育领域，我们仍然似乎不擅长评价。”[5]可以说，对评价方法研究的不足导致评价实践的相对滞后。

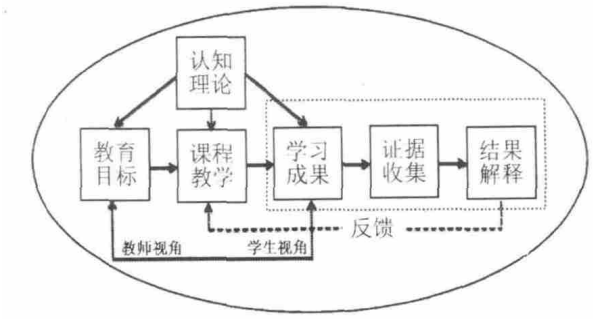

图 1　教育评价的基本框架

教育评价是在特定认知理论的基础上，根据具体的教育目标，在课程教学的某个阶段，系统地收集学生的学习成果并解释结果的过程，这一过程还可能包括对评价结果的反馈（见图 1）。将评价方法置于这一分析框架中，有助于加深我们对具体方法的理解。本文将借助这一框架，深度分析表现性评价方法，并以美国的大学学习评价项目（CLA）为例，着重介绍其在大规模学习成果评估中的应用。

二、认知理论、课程教学与表现性评价

认知理论指的是涉及待评价的知识和能力的理论和信念，或者说评价背后的知识观、能力观与学习观。“评价的性质必须与被评价的学习的性质相一致”[6]，因此，认知理论决定教育评价的内容，回答的是“应该评价什么”的问题。课程教学是在认知理论指导和教育目标指引下帮助学生获得学习成果的基本途径，也是教育评价的服务对象，它回答了“为什么而评价”的问题。

在过去的一个多世纪中，心理学研究使人们对学习的认知产生了巨大的变化。从研究动物开始，巴甫洛夫、斯金纳、班杜拉等先驱逐渐发展出经典条件反射、操作条件反射和观察学习等基于联想和模仿的学习理论。这种学习理论因为强调心理学研究要基于可观测的行为变化而被称为“行为主义”。简单来说，行为主义者认为知识是独立于学习者的客观存在，知识和技能的学习具有刺激—反应性、机械性、累积性和渐进性的特点，是一个传授、记忆、吸收和掌握的过程。行为主义所对应的课程观和教学观强调知识和技能的传授效率，以其为基础的教学被归纳为“以教师为中心的教学”（见图 2）。斯金纳甚至认为“必须告诉学生结果而不管他做得对或错，如果做得对，再引导其进入到下一步”。[7]他将学生视为被动的、完全依靠外部激励的学习机器。行为主义所关注的学习成果一般为事实性的、分散性的和程序性的知识以及较低层次的技能，因此其对应的或者说最适合的教育评价方法是传统的选择题及其变体（如判断题、配对题等），一般被统称为“选择—反应”测试或者说“客观”测试。这类评价方法规定了学生的反应类型，结构性强，因此具有简单易行（尽管其设计需要科学考量）、信度高、评分客观等优点。教育评价学家谢帕德（Shepard）在讨论评价对于学习的意义时，曾指出 20 世纪的主流范式是社会效率导向的课程理论、行为主义的学习理论以及科学测量导向的评价理论的结合。其中，科学测量的

主要形式是“客观”测试，主张“一题考一个技能”。[8]

图 2　认知理论及其所对应的课程教学与教育评价方法的演变

如果将有关学习的认知理论看成一个存在两极的连续体，那么行为主义在一端，建构主义则在另一端，并随着时间推移，后者正处于越来越受认可的地位（见图 2）。从 20 世纪末开始，建构主义的学习理论在高等教育领域逐渐兴起。建构主义主要源自于杜威的进步主义教育哲学以及心理学家皮亚杰和维果斯基有关认知发展阶段和“最近发展区”的研究。建构主义者认为知识不是经由他人传授获得的，甚至都不是客观存在的，而是学生在一定情境中借助一定的学习资料在他人的帮助下通过意义建构出来的。建构主义对应的课程和教学强调真实情境的创设、挑战性的任务、小组合作和自主探究。教师的角色不再是知识的传授者，而是学生学习的辅助者，因此基于建构主义的教学经常被称为“以学生为中心的教学”。基于问题的学习（problem-based learning）、翻转课堂（flipped classroom）等时下流行的教学模式均有建构主义的理论基础。建构主义强调学习者自己主动去构建知识。这非常符合成人学习的特征，比如“越成熟，越自我导向；累积的经验是学习的丰富资源；只有在体验到有需要学习某项知识或技能的时候才会准备去学习；比儿童更多以问题为中心；最有效的动机是内在的”[9]。针对这种学习特征的教育评价任务也需要学生通过建构，而不是通过记忆和再现来完成。因此，与其相适应的教育评价方法也被称为“建构—反应”测试，具体包括结构化和开放性程度不等的表演、实验、创造等实践任务以及简答、论述、作文等书面任务。其中结构化程度最低、开放性最强的评价被称为表现性评价。它需要学生能够运用批判性思维解决复杂问题并将知识运用于真实情境中。这些能力的评价是传统的“客观”测试无法胜任的，比如评价论文写作能力的选择题只能测试学生能否记住写作的基本步骤和要素，但事实上只有真实的写作任务才能对学生的写作能力进行评价。需要提及的是，教育评价方法同样被视为存在两极的光谱，其一端为“客观”测试，另一端则为复杂的表现性评价。在这个光谱中还存在填空、解释性练习（interpretive exercise）等结构性和自由度居于中间位置的评价方法。表 1 显示了表现性评价与“客观”测试之间的区别。

表 1　表现性评价与“客观”测试的区别

区分维度	“客观”测试	表现性评价
认知理论	行为主义	建构主义
任务形式	选择题及其变体	论述、作文、创作、表演
考察的知识	事实性（陈述性）知识	过程性、原理性与战略性知识
考察的能力	低阶思维；单一维度能力	高阶思维；多种维度能力
问题结构	结构性强、自由度弱	结构性弱、自由度强
问题情境	无情境	真实情境
结果解释	客观判断＋标准答案	主观判断＋预定质量标准
评分信度	信度较高	信度较低

三、教育目标、学习成果与表现性评价

教育目标（objectives）是对宽泛的、一般化的教育目的（goals）的操作化。它的行为主体是学生，通常表征为具体的、可测量的、可观察的学生行为。换言之，教育目标是教师视角中预期的学习成果；而学习成果则是学生视角中的教育目标，它在教育评价的基本框架处于中心地位。教育目标受认知理论指导，因为它“必须与一种学习心理学相联系”[10]。它回答了“具体评价什么”的问题。清晰表述的教育目标不仅有助于确定课程标准，也有助于度量学习成果的质量，更是制定评分准则的直接依据。

教育目标和学习成果可以粗略地分为认知性和非认知性两类。表现性的实践任务可以用来评价非认知性的学习成果。比如可以通过表演来评价情感表现力，通过实验来评价动手操作能力，通过演讲来评价口头表达能力等。而表现性的书面任务一般只能用来评价认知性的学习成果。在认知性的学习成果中，表现性评价对其中建构性较强的部分更有针对性和内容关联效度。例如，布鲁姆等将认知性的学习成果分为知识、领会、运用、分析、综合和评价。在“知识”这一类别中，无论是具体的知识还是普遍原理与抽象概念的知识，学生“所涉及的主要心理过

程是记忆”。[11]因此，对于这一类教育目标，传统的客观测试就能发挥很好的评价作用。在“领会”这一类别中，学生需要能转化、解释和推断交流内容。对于这一类教育目标，传统的客观测试也能发挥较好的评价作用，但评价效度就比知识类别要弱。对于后面四个类别的教育目标，需要学生能进行复杂的联想和判断，选择题效度更弱，能发挥的评价作用更小，最多只能处于辅助地位，而表现性评价则效度更高、更有用武之地。

单纯从知识的角度来看，也能发现表现性评价更加适应哪些教育目标。例如，谢韦尔森（Shavelson）将知识分为四种类型：陈述性知识（知道事实）、过程性知识（知道怎么去做）、原理性知识（知道为什么）和战略性知识（知道何时适用）。[12]陈述性知识的获取依赖学习者主动构建的成分较少，记忆和接受教师讲授传递是学习的主要形式，因此其评价通过传统的选择题即可，但后三种知识的掌握更加需要学习者的主动参与和构建，需要更高阶的思维能力，因此表现性评价更加合适。

高等学校旨在培养高级专门人才，因此教育目标突出了专业知识和技能，学习成果评价更是围绕专业知识技能展开。面对日新月异的社会和已见端倪的知识经济，过程、方法、科研思维、批判性思维、跨学科的一般能力等教育目标同样值得重视。有调查显示，80％左右的美国雇主认为大学毕业生需要加强批判性思维以及分析性推理能力、分析和解决复杂问题的能力、口头和书面表达能力、以及在真实世界中对知识和技能的应用能力。[13]这些能力显然不是某个学科专业所特有的，但对于学生的学习和未来的就业和职业成功都有重要意义。这需要高校在教学和教育评价中予以回应。很早就有人提出，选择题使更有智慧和创造力的人处于不利的地位[14]，现实中重要的问题往往具有“非结构性”，因此基于行为主义的“选择—反应”任务在应对这些方面的学习成果时是力不从心的，表现性评价则大有用武之地。

四、表现性评价的证据收集与结果解释

证据指的是要求学生表现（选择出、说出、写出、表演出或创造出答案等）以展示出他们所具有的知识与能力的任务或情形。证据收集专指评价方法，回答的是“如何评价”的问题。在学习成果的证据收集方面，表现性评价具有如下特征：第一，问题具有真实情境：表现性任务中待解决的问题不会被孤立地表征，而是一定出现在某一具体的情境中。构成情境的信息是符合逻辑的，有可信的来源，具有极高的表面效度。任务的情境可分为个人情境、社会情境、职业（工作）情境、教育（科学）情境等。其中，高等教育学习成果评估涉及后三种情境居多。第二，问题与提供的信息是复杂的：问题有时没有被清晰陈述，需要学生自己去识别和提炼；信息经常是杂乱无章的、模糊的，有时甚至相互矛盾，需要学生从中提取若干关键信息来辅助解决问题。学生还需要有洞察力地意识到可能缺失的信息。第三，答案是开放的：表现性任务一般没有唯一的“正确”答案，问题解决具有开放性。学生需要提供可能的解决方案，决定如何行动，识别出不同方案的预计效果和负面影响，以及不同方案所面临的约束，并进行权衡和取舍。可以说，没有标准答案是表现性任务的基本“标准”。第四，角色扮演以融入情境：表现性任务经常为学生设置一个角色，这个角色一般是学生所不熟悉的。学生在角色扮演中，需要不停地问自己“在这种情形中我该怎么做？”。学生需要面对其他角色所持的不同视角或相反意见，需要秉持包容和开放的态度。表现性任务具有“百老汇的戏剧教学特点”。[15]

结果解释指的是如何从证据中得出评价结果并据此对学生的学习成果进行推论的过程，回答的是“评价结果是什么含义”的问题。表现性评价与结果解释的关系要看其运用的情境。在班级教学评价中，教师负责解释评价的结果，他们一般对收集到的证据进行直觉或定性的解读，这种解读是非正式的，随意性较大。在我国大学本科教学实践中，许多教师在有意识或无意识地使用表现性评价，比如让学生撰写论文、承担设计任务等。不可否认，有的评价任务质量很高，效度很好，但对评价结果的解释则完全依赖教师的主观判断。主观判断的特点是没有正式的评分工具和准则，容易带有教师个人的偏见和误差。比如评分中容易存在“光环效应”，即教师容易给自己偏爱的、在课堂上表现积极的学生高分。在大规模评价中，有着严格的评分标准，并从多个方面来保证评价结果的客观性。结果的解读主要依赖统计模型，来对学生能力以及熟练程度进行汇总及分层，这种解读是正式的、复杂的、高度依赖标准的，科学性很强。表现标准和计分准则在大规模表现性评价中起到非常关键的作用。

五、表现性评价的应用：CLA 案例解读

大学学习评价（Collegiate Learning Assessment，CLA）是一项大规模的高等教育学习成果评估项目。大规模评价不仅能报告学生的学习成果，还能反映学生能力或素养结构中社会认为重要、值得认可和奖励的方面，能为教师、学生、学校乃至整个教育体系提供值得追求的目标，能为全面的、系统性的教育教学改革发挥刺激作用。相对于基础教育阶段数量众多的大规模甚至国际性的学习成果评价项目（比如 PISA，TIMSS），高等教育界内大规模学习成果评价的实践为数不多。这主要有两个原因：第一，高等教育的最大特点是多样性，它不具备基础教育阶段相对统一的教育目标。不同类型、层次高校的愿景和培养目标相差甚远。第二，高等教育在绝大多数国家是一种专业教育，不同专业之间难以有统一的学习成果。正如认知科学所认为的那样，“知识和技能是按一定结构组织起来的，具有领域特殊性”。[16]也正是因为这两个原因，高等教育领域中的大规模学习成果评价项目所聚焦的均不是专业学习成果，而是跨学科的通用技能，主要指那些能够在多种职业中应用的技能。CLA 亦是如此。

1. CLA 简介

CLA 的研发机构为美国教育援助委员会（Council for Aid to Education，CAE），该机构由一群公司高管在通用汽车领导人斯隆的带领下成立的，旨在加强校企合作，以及企业对美国高等教育的支持力度。CAE 迄今已有 60 多年的历史，目前专门提供教育评价服务。也许正是研发机构的历史特征决定了 CLA 的特点，即它所评价的焦点是那些社会和雇主十分珍视的，以及期望大学毕业生需要掌握的认知技能。雇主得不到这些信息，因为它们在高等教育领域内没有被充分评价到。由于高等教育机构中普遍存在的分数膨胀现象，雇主仅通过应聘者毕业学校的声望以及学生的平均绩点（GPA）来了解未来雇员已经变得不那么可靠①。正如 CAE 主席本杰明（Roger Benjamin）所引用的，“高盛集团并非一定要哈佛的毕业生，他们只要所需职位的最佳人选。精英大学学位传统上被视为（人才质量的）最佳代理指标，只是因为目前没有精确的衡量标准”。[17]而 CLA 提供了这样一种衡量的可能性。具体来说，CLA 并不评价具体的学科知识，而是聚焦于大学本科生的批判性思维、分析性推理、问题解决以及书面表达能力。这些能力对于每个学科专业的学生来说都是适用的。比尔·盖茨在阅读了基于 CLA 数据分析的著作《随波逐流的学术：大学校园中的有限学习》②之后写道，“有人批评本书没有关注学科知识的学习，但我认为大部分人都会同意 CLA 所评价的批判性思维、复杂性推理和写作能力非常重要”[18]。有趣的是，那些在 CLA 所测量的领域中非常优秀的人在大学后职业发展的初期有着更为成功的体验。[19]

CLA 于 2002 年开始试点。虽然问世不过十余年，属于非常年轻的学习成果评价项目，但它已经发展成为高等教育领域内最为知名的大规模评价项目。CLA 每年举办一次，目前已经有超过 700 所高校参加过该项目，其中绝大多数为美国高校，还有少数高校来自加拿大、日本、西班牙等国家和香港地区。尽管 CLA 来自美国，但它正在获得国际认同。经合组织（OECD）即将正式实施的大规模、国际性的“高等教育学习成果评价”项目（AHELO），其通用技能评价部分在可行性研究中很大程度上参考了 CLA 的表现性任务，可以被视为 CLA 的一个修订版本。[20]

CLA 主要利用“建构—反应”测试来评价如前所述的若干认知技能。学生可以选择完成一个表现性任务或者一个分析性写作任务。表现性任务要求学生根据特定情境和给定的文档信息来完成一个写作任务（90 分钟）；分析性写作任务则要求学生首先就一个话题完成一篇立论文（45 分钟），然后根据一段短文完成一篇驳论文，要求识别并描述该段短文中的逻辑缺陷（30 分钟）。CLA 在与时俱进，2013 年推出了新版本 CLA＋。新版本在内容上增加了对学生科学和定量推理能力、批判性阅读和评估能力以及观点驳斥能力的评价；在方法上缩短了表现性任务的完成时间（60 分钟），删除了分析性写作部分，增加了基于文档阅读的 25 个选择题（30 分钟），以适应新的评价内容、增加评价的信度；在成绩报告中增加了标准参照（即熟练水平或掌握水平）的结果解读等等。无论是 CLA 还是 CLA＋，表现性任务均处于评价的中心地位。

2. CLA 表现性任务分析

表 2 展示了 CLA 表现性任务的一个例子。学生需要了解任务背景，扮演假设性的角色，并从文档库中选择适当的信息进行分析和推理来完成写作任务。这些文档类型繁多，包括论文摘要、图表、备忘录、博客文章、新闻报道、地图等，以尽可能贴近现实

环境。根据本文前述的教育评价框架，CLA 的表现性任务具有如下特点：第一，它依据的认知理论不再是行为主义，而是建构主义。CLA 所测量的能力必须靠学生在平日的建构中获取，而不是靠各种版本的“刺激—反应”式的灌输学习。正如 CAE 网站所介绍的那样，“CLA＋独特和有意为之的特性是没有什么速成课能让学生在评价中表现优秀，因为它测度的是批判性思维技能而不是信息的积累，测试的准备需要学生花费数年时间去发展和磨炼批判性思维”。[21]第二，它所指向的教育目标与低阶的知识记忆和再现相去甚远，分析、评价以及批判性思维等高阶学习成果才是它真正关注的对象。第三，大学课程教学与 CLA 所评价的学习成果有紧密关系。如果学校不创设好的环境，不提供机会；教师不以学生为中心来组织课程教学（包括实践教学）；学生不自主学习，不主动实践，不利用和创造机会去锻炼问题解决能力，CLA 所关注的批判性思维能力是绝对不会自动获取的。

表 2　CLA 表现性评价样题

背景：你是一家机构的职员，该机构主要分析政治候选人提出的政策要求，并为拥护某个候选人提供建议。S 正在竞选连任 G 州 J 市市长一职。S 市长的竞选对手是 E 博士。E 是 J 市议员，他在一次电视采访中提出了反驳 S 的三个观点：第一，E 认为，S 市长提出利用增加警察数量来降低犯罪的办法不可取。E 说“这样做只会导致更多的犯罪”。E 用一个图表来支持他的观点，该图表显示相对于人均市民所需警察数较少的城镇来说，人均市民所需警察数较多的城镇犯罪率更高。第二，E 说“我们可以将那些原本要用来雇佣警察的钱用在某戒毒计划中”。E 用华盛顿社会研究所发布的一个简报来支持自己的观点，该简报描述了某戒毒计划的效果。E 说，还有另外一些科学研究显示某计划是十分有效的。第三，E 说，J 市的吸毒和犯罪关系紧密，因此减少吸毒者的数量就能降低城市的犯罪率。为了支持自己的观点，E 展示了一个图表，该图表将 5 个邮政区域的吸毒者比重同犯罪率进行了比较，数据来源于 J 市警察部门提供的信息。	任务：在辩论开始前，你必须写一篇评论 E 观点的文章，并就应该支持 S 还是 E 提出建议。完成时间为 60 分钟。 文档库（具体内容略） (1)私家侦探给 S 市长的备忘录（E 与某戒毒计划没有经济联系，但有私人联系） (2)J 市日报的文章（与吸毒有关的犯罪在上升） (3)J 市警察局提供的犯罪与吸毒数据图表 (4)华盛顿社会研究所的研究简报（C 镇某戒毒计划很成功） (5)G 州公共安全部数据（2000 年各县警察数量与犯罪率之间有正相关关系） (6)E 自己所绘的图表（J 市盗窃和抢劫数量与成人吸毒率之间有正相关关系） (7)有关某戒毒计划的 3 个研究摘要

注：文档库部分括号的内容为笔者在阅读后所加，以反映文档的主要观点。

资料来源：CLA＋Sample Performance Task. http://cae.org/images/uploads/pdf/CLA_Plus_Practice_PT.pdf，并经笔者编辑简化。

对样题的进一步分析显示，CLA 的表现性任务至少可以评价学生批判性思维能力的如下方面：(1)识别并区分相关与不相关信息的能力。例如，7 号文档中有 3 个论文摘要，但其中的信息只有部分支持 E 的观点，E 在论证时是有选择性的，学生需要能够识别。(2)识别出结论背后所隐含假设的能力。E 的结论是“投资于某戒毒计划能够降低 J 市的犯罪率”，其背后隐含的假设是“某戒毒计划是有效的＋毒品是引发犯罪的决定性因素→投资某戒毒计划能够降低犯罪率”。如果学生能够识别出这种假设和逻辑，那么在论证中只要找到相关信息来支持或攻击其证据链即可。(3)识别出给定证据中的局限或错误的能力。例如，6 号文档（E 所提供的图表）所绘制的犯罪数量与成人吸毒率之间有正相关关系，但此图表是根据 3 号文档绘制出来的。事实上这两个变量之间的性质不同，一为数量，一为比率，绘图并不合适；此外，E 还选择性忽视了 3 号文档中的其他重要信息。(4)识别出错误推理的能力。例如，5 号文档中警察数量与犯罪率之间的相关关系被 E 直接认为是因果关系，学生需要能识别出这种错误。(5)从所给资料中得出有效结论的能力。如果学生支持 E，那么 1 号和 4 号文档中的信息就需要被重点使用，并做论证；如果学生支持 S，那么更多要识别出 E 的论证以及所使用证据的不足。除这 5 种批判性思维能力外，CLA 还涉及解释关系的能力、辨认并陈述问题的能力、解释图表和数据的能力、对观点进行评价的能力等一系列复杂性推理和高阶思维的能力。

除批判性思维能力之外，由于学生必须以书面形式对问题做出反应，因此学生的写作能力也是 CLA 直接评价的内容。它主要包括在行文中合理组织论点和论据的能力、逻辑一致连贯的论证能力、流畅表达的能力、举例类比能力、细节阐述能力以及突出重点的能力等。当然，写作任务对于语法、遣词造句、标点符号、拼写等基本技能也有要求。此外，由于文档库存在大量的书面材料，所以学生的阅读能力也是 CLA 间接评价的内容。学生需要识别材料中的细节、辨别其中变量的关系并对内容进行推论等。

3. CLA 表现性任务的评分及报告

如前所述，表现性任务的评价相对于客观测试来说具有更大的主观性，因此评分准则对保证评价

的质量极为重要。CLA 的表现性评价采用的是分项评分法，对“分析与问题解决”、“写作效能”和“写作基本技能”三个评价标准分别评分后进行汇总。表 3 展示了“分析与问题解决”评价标准的质量描述和评分准则。可见，CLA 的评分规则用一系列的程度副词、形容词以及量词非常具体、详细地描述了学生反应的质量区别以及赋分规定，这在很大程度上缓解了评分主观性的问题。

表 3　CLA＋表现性评价中对“分析与问题解决”标准的评分规则

分值	质量描述：做出合乎逻辑的决定/得出合理的结论/持有恰当的立场；从文档库中使用合适的信息（事实/观点/计算的数值/显著的特征）来支撑决定/结论/立场	
1	表达或暗示了决定/结论/立场	仅提供了极少的分析（比如仅涉及一份文档中的一个观点）；分析完全不准确、缺乏逻辑、不可信或者与决定/结论/立场缺乏联系
2		仅对一些观点进行分析；一些观点不准确、缺乏逻辑、不可信或者与决定/结论/立场缺乏联系
3		提供了一些有效的支撑，但遗漏了或错误陈述了关键信息；只有肤浅的分析和对文档的部分理解；没有能够说明对立的信息（如适用）
4	清楚地表达了决定/结论/立场	提供了有效的支撑，涉及多个重要的、可信的信息；展现出足够的分析能力和对文档的理解能力；遗漏了一些信息；能够试图处理对立的信息或者决定/结论/立场（如适用）
5		提供了很强的支撑，运用了许多重要的、可信的信息；展现出很好的分析能力和对文档的理解能力；反驳了对立的信息或者决定/结论/立场（如适用）
6		提供了全面的支撑，运用了几乎所有重要和可信的信息，展现出突出的分析能力和对文档的理解能力；全面地反驳了对立的证据或者决定/结论/立场（如适用）

资料来源：CLA＋Scoring Rubric. http://cae.org/images/uploads/pdf/CLA_Plus_Scoring_Rubric.pdf.

CLA 是标准参照式的教育评价项目，因此需要对学生的表现预先设置“绝对”的质量标准，而不是设置相对标准。CLA 采用书签法（Bookmark method），通过来自不同领域内的专家对学生的掌握水平进行标准设定，即确定获得多少分才可以被认定达到某种掌握水平。这些专家既有来自高校的，也有来自产业界的，比如强生公司和麦肯锡咨询公司的代表。最终设定了 6 个层级递进的掌握水平，分别为基本之下、基本（basic）、熟练（proficient）、优秀（accomplished）和杰出（advanced）。[22] 例如，达到基本水平的学生可能会读懂和解释直方图，但不一定能理解和解释散点图以及回归分析；达到优秀水平的学生应该能够精确地解释和分析定性和定量的证据，并能把这些信息整合到论据中；而达到杰出水平的学生应该能在理解文档细节的同时，具备创造力和综合能力，比如理解数据中的异常值以及样本量对结果的影响，从多个文档中综合信息等。标准设置和评分准则相结合，让 CLA 的结果解释更为清晰。可以说，标准化考试的严格程序让 CLA 的表现性评价的科学性更强。

CLA 是一种自愿性质的、低利害的学习成果评价项目。迄今 CLA 成绩还没有被正式地用来对学生个人或者学校进行任何性质的筛选、奖励或惩罚③。目前 CLA 的结果可以在两个层面上来报告和解释。在学生个体层次上，CLA 报告了学生的掌握水平、在所有测试参加者中的百分位数以及分析与解决问题、写作效能等各个分项的原始分数。在学校整体层次上，CLA 报告了参加测试的学生的平均分，以及学生成绩在不同掌握水平上的统计分布情况。更为重要的是，CLA 报告了教育评价领域内的新指标——学校的增值。所谓增值，指的是学校大四学生的 CLA 实际成绩与期望成绩的差值。提供增值的原因在于，CLA 是大规模学习成果评价项目，在各个高校之间进行比较是最基本的功能之一。一方面，参与 CLA 的高校之间性质和生源差异巨大，直接进行比较是不恰当的；另一方面，学生的进步或者说变化更能反映学校的教育质量。学校增值的计算步骤为：将参与学校大一学生的平均 CLA 分数对平均 SAT（俗称“美国高考”）分数做回归，计算各个学校的残差；将这一步骤重复至大四学生，计算各个学校的残差；学校的增值为两个残差之差。[23] 从 CLA＋开始，学生也可以利用技术手册来计算自己的增值。

六、几点思考

1. 表现性评价所关注的技能具有“可教性”吗？

如果批判性思维技能难以从大学的课程教学中获取，那么 CLA 评价结果的反馈便没有任何价值，表现性评价在高等教育内部质量保障体系中的作用也会大打折扣。实证研究或许能为这个问题提供答案。一项研究表明，大四学生的平均 CLA 成绩要

比大一学生的平均CLA成绩要高出1个标准差，但两者之间的SAT考试成绩差异只有0.16个标准差。这说明大四学生和大一学生除CLA成绩外不存在系统性的基线能力差异。[24]总体来说，学生在大学四年中CLA成绩获得了巨大的进步。唯一可以提供的解释因素是学生在课程教学中学到了CLA所评价的技能，取得了学习成果。另一项研究表明，CLA成绩与学生和学校投入之间有统计联系，比如CLA成绩与NSSE问卷中“与课堂中与其他学生一起完成项目”，“做作业或者在班级讨论中把不同课中所学的观点和概念综合到一起”，“从教师那里获得及时的反馈”等一系列有关大学课程教学的变量有统计上的正相关关系。[25]这也说明美国大学的通识教育对学生批判性思维能力的培养是有裨益的。总之，表现性评价所关注的技能是可教的，并且这种技能是“填鸭式”的应试教育难以传授的。真正的问题在于如何在大学的课程教学中提高学生的批判性思维能力。开发独立课程，在传统课程建设中有意识渗透和融入相关素材以及将批判性思维的知识技能融入日常教学实践[26]均为行之有效的方法。

2. 大规模评价中表现性任务的缺陷在哪里？

除了评分人需要经过严格训练外，表现性评价的缺陷主要表现在如下两点。第一，评分一致性或者说信度不足。表现性任务题量少，答题时间长，需要采用主观评分，因此存在评分不一致的可能。CLA已经从双人评分过渡到人工评分和机器评分相结合。智能作文评分系统通过潜在语义分析(LSA)技术进行机器评分，但这种机器评分是在“学习”大量人工评分的基础上实现的，所以从本质上来说仍然是主观评分。人工与机器评分的一致性系数为0.77[27]，虽然比双人评分有所改善，但信度依然难以达到较为理想的水平。第二，评价偏见出现的可能性较大。评价偏见指的是学生由于性别、族群、社会经济地位、宗教信仰或其他特征，在评价中受到冒犯或不公平的惩罚。比如，CLA的表现性任务涉及许多职业情境，这对于经常参加实习工作的发达国家大学生来说可能非常熟悉，而我国大学生“主动实践”的机会仍然较少，因此，表现性任务对我国大学生存在“跨文化”偏见的可能性较大。再比如，考查推理能力、批判性思维的表现性任务对于某些群体的学生可能存在偏见。事实上，CLA近年的全国报告显示，理工科学生的分数要显著高于人文社科和商科；男生的分数要显著高于女生；非西裔白人学生的分数要显著高于其他族群学生[28]。这些差距或许不能完全用优势群体在高校中所获得的批判性思维能力较多来解释。

3. 我国的高等教育是否需要表现性评价？

评价向来在教育教学中起到“牛鼻子”的作用。学生总是倾向于学习和掌握那些被正式评价的知识与技能，这一点无论在哪个教育阶段都是适用的。尽管我们不希望如此，但高利害的评价确实决定着学生学什么、怎么学；教师教什么，如何教。如果CLA成绩被美国许多企业用来(哪怕是辅助性地)筛选申请者，这对美国的大学本科教学一定是一个重要的冲击力量。我国高等教育是否需要表现性评价呢？笔者认为非常需要。一方面，相对于美国大学生来说，我国的大学生可能更加缺乏表现性评价所聚焦的能力。钱颖一教授曾提出，“当前绝大多数中国学生缺乏好奇心、想象力和批判性思维能力；特别是不敢于、也不善于提出问题。”[29]正是这一点促成了清华大学等许多高校的本科教学改革。另一方面，相对于美国来说，我国对高等教育学习成果的评价更为滞后。有研究表明，美国大学中“以学习者为中心的评价”在大学组织变革中正发挥重要作用[30]，而表现性评价正是其中最有代表性的方法。我国当前还缺乏有关大学学习成果的大规模评价实践，即使是课程教学评价也存在诸多缺陷。更为严重的是，大学教师在学生学业评价中做好“规定动作”的情况都不容乐观，对表现性评价这种要求更高的“自选动作”更是掌握不足。全面提高教育质量已成为我国高等教育发展的中长期目标，而评价又是质量保障中的重要环节。鉴于评价的牵引作用，高校需要建立统一的评价考试发展机构，诊断学业评价的质量，提供给教师相关的评价方法培训，提升评价能力。一方面，要参考CLA的表现性评价对学校的通识教育课程学习质量进行统一的评价；另一方面，要鼓励教育评价研究人员与各学科教师之间的合作，把表现性评价应用于具体的专业中。

(本文在写作过程中，得到了华中科技大学教育科学研究院郭卉、朱新卓、蔺亚琼、李函颖等同事的建议，硕士研究生周兰兰承担了部分资料的翻译工作，在此一并感谢！)

注释：

① 事实上，越是声望高的高校，其学生的平均GPA越高，也就是说名校更易发生分数膨胀现象。参见ROJSTACZER, S. & HEALY, C. (2012). Where A Is Ordinary: The Evolution of American College and Uni-

versity Grading，1940-2009. New York：Teachers College Record.

② 参见 ARUM，R.，& ROSKA，J.（2011）. Academy adrift：Limited learning on our campuses. Chicago：University of Chicago Press，该著作引起了美国高等教育界对于大学学习成果的广泛讨论。

③ 因此，CLA 面临着学生参与率不高，学生样本量不大等问题。虽然没有与各种形式的问责和筛选联系起来，但媒体似乎对 CLA 分数作为用人单位筛选大学毕业生的作用很感兴趣。

参考文献：

[1] 岑逾豪. 本科教学中的高阶学习：问题、实践和挑战[J]. 复旦教育论坛，2014，(12)：47-53.

[2] 刘献君. 抓住四个关键问题 加强大学本科课程建设[J]. 中国高等教育，2013，(17)：40-43.

[3] 赵德成. 表现性评价：历史、实践及未来[J]. 课程·教材·教法，2013，(2)：97-103.

[4] GUSKEY T R. "It wasn't fair!" Educators' Recollections of Their Experiences as Students with Grading [C]//Paper Presented at the Annual Meeting of the American Educational Research Association，San Francisco，CA，2006.

[5][9] FRY H，KETTERIDGE S，MARSHALL S. A Handbook for Teaching and Learning in Higher Education：Enhancing Academic Practice（3rd edition）[G]. New York：Routledge，2009：132，14.

[6][11][16] 安德森，索斯尼克. 布鲁姆教育目标分类学40年的回顾[M]. 谭晓玉，袁文辉等，译. 上海：华东师范大学出版社，1998：76，22，73.

[7] 戴维迈尔斯. 心理学（第9版）[M]. 黄希庭等，译. 北京：人民邮电出版社，2013：286.

[8] SHEPARD L A. The Role of Assessment in a Learning Culture[J]. Educational Researcher，2000，29(7)：4-14.

[10] 布鲁姆，等. 教育目标分类学第一分册：认知领域[M]. 罗黎辉，丁证霖，石伟平等，译. 上海：华东师范大学出版社，1986：26-27.

[12] SHAVELSON R，RUIZ-PRIMO M A，LI M，ET AL. Evaluating New Approaches to Assessing Learning [EB/OL]. [2015-03-02]. http://www.cse.ucla.edu/products/Reports/R604.pdf.

[13] It Takes More than a Major：Employer Priorities for College Learning and Student Success：Overview and Key Findings | Association of American Colleges & Universities[EB/OL]. [2015-03-03]. http://www.aacu.org/leap/presidentstrust/compact/2013SurveySummary.

[14] POPHAM W J. 促进教学的课堂评价[M]. 国家基础教育课程改革促进教师发展与学生成长的评价研究项目组，译. 北京：中国轻工业出版社，2003：16，59.

[15] CHUN M. Performance Tasks and the Pedagogy of Broadway[J]. Change：The Magazine of Higher Learning，2012，44(5)：22-27.

[17] BENJAMIN R. Leveling the Playing Field From College To Career[EB/OL]. [2015-02-02]. http://cae.org/images/uploads/pdf/Leveling_the_Playing_Field__From_College_To_Career.pdf.

[18] GATES B. Academically Adrift：Limited Learning on College Campuses[EB/OL]. [2015-02-02]. http://www.gatesnotes.com/Books/Academically-Adrift.

[19] ARUM R，ROKSA J. Aspiring Adults Adrift：Tentative Transitions of College Graduates[M]. IL：Chicago：University of Chicago Press，2014.

[20] OECD. AHELO Feasibility Study Report Volume 3 Further Insights[M]. Paris：OECD Publishing，2013：16.

[21] Council for Aid to Education — How Do I Prepare for CLA+？[EB/OL]. [2015-02-02]. http://cae.org/students/college-student/how-do-i-prepare-for-cla/.

[22] ZAHNER D. CLA+ Standard Setting Study Final Report [EB/OL]. [2015-03-02]. http://cae.org/images/uploads/pdf/cla_ss.pdf.

[23][24] KLEIN S，FREEDMAN D，SHAVELSON R，ET AL. Assessing School Effectiveness[J]. Evaluation Review，2008，32(6)：511-525.

[25] CARINI R，KUH G，KLEIN S. Student Engagement and Student Learning：Testing the Linkages[J]. Research in Higher Education，2006，47(1)：1-32.

[26] 陈振华. 批判性思维培养的模式之争及其启示[J]. 高等教育研究，2014，(9)：56-63.

[27] CLA+ Technical FAQs[EB/OL]. [2015-03-02]. http://cae.org/images/uploads/pdf/CLA__Plus__Technical_FAQs.pdf.

[28] CLA+ National Results，2013-2014[EB/OL]. [2015-03-02]. http://cae.org/images/uploads/pdf/CLA_National_Results_2013-14.pdf.

[29] 钱颖一. 经济学家应致力于教育改革[EB/OL]. [2015-03-02]. http://magazine.caijing.com.cn/20150202/3813182.shtml.

[30] WEBBER K L，TSCHEPIKOW K. The Role of Learner-centred Assessment in Postsecondary Organizational Change[J]. Assessment in Education：Principles，Policy & Practice，2013，20(2)：187-204.

（本文责任编辑　许　宏）