在线测评：过去，现在和未来

[英]萨莉·乔丹（Sally Jordan）

（美国开放大学物理系，英国）

【摘要】文章阐述了广义上使用计算机的所有测评，重点说明了机考方式。测评任务的多样化、对即时反馈的诉求、阅卷的客观性以及节省资源的需要是撰写文章的动因。机考从早期单纯的多项选择题型到机读卡片等开始发展到了融合交互性的便于学生在家里使用的网络系统形式。这样设计精细的在线测评系统由大学、公司设计开发并作为虚拟学习环境的一部分使用。例如，选择题可以通过某些技巧，如基于信任度的评分来减少其不足。在课堂上尤其是在同伴讨论中使用电子复审系统（应答器）很有效果。学生设计的问题也会鼓励同伴之间围绕学习进行对话。

日趋精细的机考系统能够将数学题拆分成几律，并提供有针对性的、即时的反馈。文章也讨论了计算机代数以及简答型题目的测评。计算机分性测评选择采用学生对以前问题的回答来改变测评的后续形式，并更广泛的讲，机考测评包含了同伴测试、测评的电子文件夹、博客、维基和论坛。文章最后预测了在线测评的未来发展：在线测评可在MOOCs（大规模开放远程课程）中使用；可用于学习分析；在测评中要教学、测试与学习逐渐融合；解放了人力测评，并能更加真实地测评考试内容。

【关键词】在线测评；机考；综述
【中图分类号】G720 【文献标识码】A 【文章编号】1008-7648（2015）05-0008-011

一、概述

广义上的在线测评（JISC 2006）是指任何有计算机参与的测评，可以是终结性、形成性或诊断性测评。因此，其范畴包括网上提交的由辅导教师批改的作业，电子文件夹的测评，反思性博客，录制或声音文件的教师反馈。常见的是机考测评。其他同类含义的表述包括技术辅助的测评，计算机支持的或计算机协助的测评。

在线测评的早期研究文献（Conole & Warburton, 2005; Dikli, 2006; Heppleston 等, 2011; JISC, 2009; Kay & LeSage, 2009; Nicol, 2008; Ridgway 等, 2004; Ripley, 2007; Stödberg, 2012）主要聚焦于技术辅助的测评或反馈。尽管本文侧重于机考，但仍采用广义上的定义。因在线测评这个概念非常宽泛，且因篇幅限制，不能谈及所有的在线测评系统，因此本文精选了此项技术进行介绍（例如电子投票系统或“应答器”）文也讨论了在线测评和纸质试卷的选择，重点介绍了物理学和相关学科领域中开发、使用或评估过的在线测评系统。

二、动力及发展

有效的测评和反馈是指帮助学生在所选择的复杂学科领域里取得优异成绩的一种实践做法。学生参与有效的测评和反馈，这不但不会增加辅导教师的负担，还能帮助作为终身学习者的他们掌握学习技能，能够自信地继续学习。在这些目标的实现过程中，技术提供了巨大的支持。”（JISC2010, p.8）

在线测评是包含技术和测评方法的在线学习（Ashton&Thomas 2006, Gipps 2005）的自然组成
部分之一（Mackenzie 2009）。在线测评在设计上不断多样化并保证其真实性，比如，采用电子文件夹，模拟和互动游戏等方式，具有其他手段很难实现的测评技能（JISC 2010）。


研究表明定期的在线测试会提升年度考试成绩（Angus & Watson 2009）。在线测评增加了学生对课程的参与度，激发学习动力并帮助学生规划学习进度（Crebenik & Rust 2002, Jordan 2011）。学生可以根据网上作业检查对所学知识的理解从而制定未来学习计划，但是研究表明单纯的考试行为，即便没有任何反馈的考试，也比多余的学习资料更能提升后续的学习表现。这就是所谓的测试效应，有关该领域的研究可参阅 Roediger & Karpicke（2006）的文献。

所以说，在线测评对提高学生的学习具有很大的帮助。然而，有趣的是，“客观题”这个词，尤其是用来描述多项选择题，反映了早期采用多项选择题的目的是否是希望使测评更客观。最早期的多项选择题可能是 e.I.桑代克的 α 和 β 测试，是第一次世界大战中美军用来测评新兵的服役情况（Mathews 2006）的。然而，20世纪，研究人员逐渐意识到写短文作为测评方法的局限性以后，多项选择题作为一种教学测验手段逐受受到追捧（E.R.Bacon 2003）。Ashburn（1938）注意到不同教材和短文的评分不同，这种现象让人担心，类似的发现有很多（例如 Millar 2005）。人工评分具有天生的不连续性，可能会影响对某个学生的期望而受到影响（Orrell 2008）。多项选择题具有客观性，而机评则具有连续性，这是多个评分者之间、或或是一个评分者在不同时段评分都难以确保的特点（Bull & McKenna 2004, Butcher & Jordan 2010）。

尽管设计出高质量的问题也会被当作重要的任务（Bull & McKenna 2000），但是，机考在提高可信度的同时，既能节省时间又能节省资源（Dermo 2007）。机评尤其适合学生人数较多的大班（Whitelock & Brasher 2006），能让使用者更加充分地利用时间，从这一点来说，机评具有附加价值（JISC 2010）。

20世纪，大规模的多项选择题通过读形式问题（如图1所示），学生在答题纸上标出每题所选答案。该系统（目前仍在使用）保证了测评的客观性，节省了资源，但是反馈的及时性的和学生的参与度没有显现出来。根据 Brown 等人（1999）所提高等教育实践做法的回顾，发现在线测试虽然也存在其他形式，但基本上都是“多项选择题”题同义词，设计者们仅仅是将测评从纸质版转移到了屏幕上，这已被证明是不恰当的做法。

图1 机读学生答题卡样式


模块物理教学软件（STOMP）测评系统的开发始于1995年（R.A.Bacon 2003）。最近的版本（Bacon 2011）是直接安装启用的QTIv2.1 版，问题和测试的互操作性(QTI)版设计的目的是
交换编写工具、试题库和测评系统之间的试题、测试和结果的数据（IMS Global Learning Consortium 2013）。

20世纪90年代，物理学和工程学本科段学生在数学准备越来越受到关注，因此，催生了诊断性测试（Appleby等人1997, Appleby 2007）。诊断性测试采用的是技能层次法，基于学生对前一个问题的回答，利用专业系统来确定下一个问题该如何设计。因此，诊断性测试也是早期适应性测试的一个案例（详见第三章第9节）。


在商业领域，Questionmark（以前叫做“question mark computing”）公司于1988年成立。Question Mark网络版（1995年启动实施）被看作是世界上首个商业互联网测评产品。其专业版于1993年启动实施，后来逐步被其认知版所取代（Kleeman 2013）。

随着越来越多的学习工具上线使用，大学和其他教育机构开始使用虚拟学习环境系统（VLEs），也通常被称为学习管理系统。大部分的VLEs都融合了测评系统。例如，Moodle学习管理系统（Moodle 2013）于2002年首次上线使用，它的测评系统紧接着在2013年上线（Hunt 2012）。Moodle和它的测评系统（详见第三章第6节）都是开放资源，在影响在线测评工具发展的理念上反映了重大的变化。

三、当前机考发展状况

1. 选择题还是建构题

Hunt（2012）确认了Moodle平台上使用的30多种不同的题型，例如拖拽题、计算题、数学题、判断正误题等题型更加多样化。但是Hunt从2,500多个Moodle站点收集了5,000多万份专门的调查问卷显示：正在使用的大约90%的题型都是选择题，例如，多选题或者给出备选答案供学生选择的拖拽题，远远超过需要学生自己组织回答的建构题。


在选择题中，学生可以猜测答案，因此，老师们无从知晓学生学习到了什么（Crisp，2007）。Downing（2003）并不关注对选择题的分数，他指出，单独地通过猜题而通过整个考试是非常困难的。然而，Burton（2005）认为成功的猜题会对处于边缘学生的学习成果产生很大的影响。
Funk & Dickson (2011) 使用完全相同的题目设计了多项选择题和简答题两种版本从而进行对比研究。50名学生尝试着对两个版本的每个题目进行回答，其中一半的学生先完成了10道简答题作为50道多选题的预备考试，而另一半的学生先回答了50道多选题，作为多选题的后续又完成了10道简答题。在每个案例中，学生们在多项选择题中的表现远远好于回答简答题的表现（p<0.001）。然而，Ferraro（2010）发现，在多项选择题和开放式问答测验的分数之间有很高的关联度。一些学者称选择题特别适合特定学生群体，尤其是具有更多考试技巧的学生和更愿意冒险的学生（Hoffman 1967）。对两种题型的喜好也存在性别差异，例如，Gipps & Murphy（1994）发现在15岁年龄段的学生中，女孩不喜欢多项选择题，而男孩则喜欢选择题胜过简答题。Kuechler & Simkin（2003）发现母语为非英语的学生做多项选择题时在剖析词汇的细微差别方面会遇到困难。Jordan & Mitchell（2009）和 Nicol（2007）明确了在选择题和建构题的回答中存在根本不同的认知过程。


2. 基于信任度的评分和类似的方法

为了弥补学生们在多项选择题中猜测正确答案的讽刺，教师们采用了各种各样的方法。可采用单纯的否定打分（选择错误答案即减分）的方法，但是必须慎重（Betts 等，2009，Burton，2005）。

Ventouras et al（2010）构建了一种为同一题目使用“成对的”多项选择题的考试类型（但这明显不是给学生使用的），该类考试的得分规则是，如果两个问题都答对了，将会获得“奖励”。考试的结果几乎与建构型题目考试的结果没有任何差别。McAllister & Guidice（2012）介绍了另外一种考试方法，就是将一系列的问题与备选答案相结合，这会导致对很长的答案列表（他们的案例是为50个问题设计了60个备选答案），这样就大大降低了猜测答案的正确性。然而，一般来说，为一系列的问题找到同样可行的备选答案是很困难的。Bush（2001）介绍了一种“自由式多项选择题”的形式，在这种题型中，如果学生对某题的正确答案不确定，那么，他们可以选择不止一个答案来回答该问题。所述打分即是对多选答案扣分：每个正确答案是3分，每个错误答案被扣除1分，总分除以3。如果一个学生知道某题的正确答案，那么他/她可以得3分，也就是说该题他/她可以得100%。如果一个学生认为某题的正确答案是两个选项之一，则该题可获得（3-1）/3，也就是说67%，而不是0%或100%。如果一个学生认为正确的答案是三个选项之一，则该题可获得（3-2）/3，也就是33%，而不是100%或0%。毫无疑问，这种题型的打分方式是比较公平的，但是有些学生却认为这种方
式非常容易产生误解，而且会使学生们将重点放在“解题技巧上”而非对正确答案知识的理解上。

长期以来，有许多观点认为考试分数的可信度可以通过适当的方法对学生的学习成绩有所增加。Gardner-Medwin做了很多关于“以信任度为基础的”（或叫“以确信度为基础的”）评分，他指出这种方法不能支持一直可信赖或一直不可信赖的学生，却支持那些能够正确判断缘由的人。Gardner-Medwin使用了分数量和扣分原则（如表1所示）。

<table>
<thead>
<tr>
<th>信任度</th>
<th>C=1（低）</th>
<th>C=2（中）</th>
<th>C=3（高）</th>
</tr>
</thead>
<tbody>
<tr>
<td>如果正确即得分</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>如果错误即扣分</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
</tbody>
</table>


3. 应答器


4. PeerWise


5. 计算机辅助的数学学习 (CALM)和苏格兰测评项目 (PASS-IT): 一般聚焦于将问题分解成“多个步骤 (steps)”

海里特-瓦特大学从1985年开始实施计算机辅助的数学学习项目 (CALM) (CALM 2001), 各种不同的机考系统也部分源于该项目。包括CUE、交互式历年论文 (Interactive Past Papers)、苏格兰使用信息技术的考试项目 (PASS-IT)、在线考试 (i-assess)和NUMBAS (Foster 等, 2012)。有些系统在具有高风险的终结性考试中使用，但是核心目的仍是支持学生的学习 (Ashton 等, 2006b)。从早期的观察来看，建构类题型受到了欢迎，因为此类题型可以给学生提供提示 (Beever & Paterson 2003)。
CALM考试系统的特征之一是使用“分解步骤”，即允许将问题分解为可管理的步骤，从而为不能继续考试的学生提供帮助（Beever & Paterson, 2003；Ashton 等，2006a）。图表2（a）中显示的是一个问题，一个学生选择没有任何中间步骤的帮助来回答问题，这样，在终结性考试中，他就可获得满分。同样地，如图2（b）所示，学生可点击“分解步骤”按钮，问题即可被分解成不同的步骤，这样，学生在终结性考试中通常能获得部分分数。

McGriene等（2002）对比了学生采用CUE考试体系中三种不同格式（没有分解步骤，必须使用分解步骤，或选择性使用分解步骤）参加模拟考试的结果和就同样内容采用纸质考试的结果。在这种情况，采用分解步骤不存在扣分的情况。没有使用分解步骤的总分要低于使用分解步骤所获得的分数。而两者的分数均低于纸质考试的分数。他们的结论是，“这意味着没有分解步骤的情况下，纸质考试分数体系不能被当下所使用的机考框架体考试体系的表现在总结报告。对学生来说，看到自己的表现情况能够帮助培养独立学习的能力。这种分析也适合循环问题的设计过程。例如，如果学生在某一特定问题中持续使用“分解步骤”，这意味着学生解决该问题是有困难的，也许是因为问题设计得不好，也许是由学生理解所造成的。

PASS-IT考试系统的研究与开发现在已经在完全融入到了SCHOLAR项目中，在苏格兰的所有中学里，SCHOLAR项目（2013）在中低年级到高年级学习资料中将形成性考试作为核心。

6. OpenMark和Moodle：倒反反馈

开普通过CD光盘成功向学生提供互动问题。2002年开始成功地使用在线平台（Jordan 等，2003；Ross 等，2006），继而之后，2005年开普实施了OpenMark系统。开普的学生人数很多，因此，对在线测评的投资是很有意义的。在远程学习中，及时有针对性的反馈是十分重要的。

图3展示的是一套典型的OpenMark的问题，这是一套截图，每个都显示了学生问题回答的尝试过程。这个案例包含的原理是：

(1) 强调反馈；
(2) 强调交互性（多次尝试能够使学生对收到的反馈迅速回应）；
(3) 支持宽度的交互（提供一系列种类的问题，目的是“充分使用现代多媒体计算机的全部功能创造参与性考试” ）（Butcher 2008）。

另外，OpenMark考试系统的设计使兼职学生能根据自己的时间，以适合自己日常生活的的方式来完成测试，这意味者考试可以在任何一个点被打断，而且只要联网便可在任何地点继续完成（Butcher，2006，2008）。

2002年开始，开普在课程中引入了互动机考（iCMAs）。在2012年8月，大约有60门独立的课程中设置了630 000多个互动机考（Butcher 等，2013），这其中有，四分之一的考试已经成为课程的正式考试办法（如例，作为终结性考试或作为入门考试）。在终结性考试里，如果每次尝试不成功，那么就会被扣分。如果每次尝试有一部分是正确的，那么也将会获得适当的分数和反馈。

开普学院是第一个使用OpenMark互动机考
评估特定的数值。这是一种合理的方法，但是会导致将错误的答案评为正确的（如：\(X = 2\) 和 \(X = 3\) 时，结果都是 4）。OpenMark 采用的是字符串匹配法。这种方法很有效（比如图 3 所示对学生做出有针对性的反馈），但是却要依赖于对题目设置者，因此要考虑所有正确答案以及尽管结果相同但是却有错误的情形。


STACK 作为一个独立的体系于 2004 年首先发布使用，但是直到 2012 年该体系才应用到 Moodle 问题类型中（Butcher 等，2013）。Moodle 系统开发人员将其重要在于对学生答案的反馈和监控方面，这也是 STACK 体系中重要的内容。在这种情况下，STACK 采用了一种“提出问题”式的、有针对性的反馈，即当问题能够给出具体的、有针对性的反馈时，比如当学生在做题时遇到困难时，系统会给出具体的反馈。

芬兰阿尔托大学的开发小组认为及时的反馈是 STACK 最好的特征（Sangwin 2013），Sangwin 认为“不是所有使用计算
机代数系统的（CAS）体系都能够让教师对反馈进行编码”感到非常惊讶。图4阐释了Sangwin (2013)所具体描述的STACK的部分复杂特征。

问题的变式。因为所有的计算都是通过计算机代数系统来执行的，因此，只要稍微努力就可以创作出一个问题的多种不同变式。所以，在上面显示的例子中，变式可能要求学生区分其他函数合起来的结果而生成的任何一个函数。然而，最好的做法是检验并使用难度相同的变式。

问题设置者的选择。问题设置者可以决定，例如，是否接受隐式乘法，是插入点号还是乘号来表示乘法。设置者也可以选择是否接受所有代数方法上相等的答案。所有前述内容如图4所示，但是，要回答“简化U3U5”这个问题，那么答案“U3U5”就是不可接受的。

审核与考试是分开的。学生通过键盘输入的答案会通过系统“理解”并得到显示，在评分之前会检查句子的正确性。而这一过程又使得有些答案有机会欺骗计算机代数系统，在这种情况下，包含“Diff”命令（该命令告诉计算机代数系统区分原始功能）的答案就会被拒绝。

8．简答题和小论文

在使用计算机代数系统生成和检测的数学题的同时，简答题评分软件也被引入到建构题中。简答题通常是指需要一两个句子来回答，后面紧跟着判断。Jordan (2012b)严格地限制了简答题不超过20个字，一方面警告学生需要回答什么，另一方面鼓励学生正确回答和非正确答案内容的时问出现。Mitchell et al (2002)最先明确了正确答案内容中包含的不正确因素是简答题自动评分系统中可能存在的一个非常严重的问题。


图5 一道PMatch题目的正确答案

英国开放大学OpenMark系统中使用了智能测评技术和PMatch答案匹配法，而PMatch（模式匹配法）与STACK一样，都是Moodle问题类型之一，即时的、定制式的反馈在考试中非常重要。关于学生对简答自由文本问题的参与度和所提供的反馈，Jordan (2012b)曾开展过非常详细的评估。

机考可以得到准确评分，好于或至少等同于人工评分（Butcher & Jordan 2010, Jordan 2012a），但此问题尚未被充分利用。Jordan(2012a)收集了几百个学生答案和评分的需求，认识到此类问题被广泛使用的时间长度，对设计答案匹配所需的时间。她建议研究应该聚焦于使用机器学习、设计答案匹配规则上，聚焦于调查不同学生对类似问题给出相同答案的频率；如果他们的答案相似，那么，就具有分享问题的可能性。

些系统做过综述。更为智能的系统尚在开发之中，如 OpenEssayist（Van Labeke，2013）侧重于提供反馈帮助学生提升论文写作技能。写作样式的习惯用单代理模式的系统一直以来饱受批评，如 Perelman（2008）就在培训3个学生写小论文时利用这种单代理模式的窍门，使用长词、引用名句（不管多么的不相关）从而获得机考的高分。Condon（2013）认为直到计算机能够对写作样式的做出有意义的评价后，才能使用这种考试方式。

9. 有效使用问题

根据 Hunt（2012）描述，机考系统包含三部分：

（1）问题引擎，将每个问题呈现给学生，给学生的答案评分并给出恰当的反馈；
（2）试题库；
（3）检测系统，将单独的问题与完整的测试相结合（有可能在测试阶段给出反馈）。

因此，除了考虑问题类型外，考虑问题的组合方式也很必要。


适性测验（经常被描述为“计算机适性测验”）是根据学生对以前问题的答案来判断其学习能力，从而使接下来所设计的问题能够保持在恰当的水平上（Crisp，2007）。Lilley 等（2004）发现学生在适性测验中并没有处于不利地位，而且他们乐意不回答那些自认为简单的问题。计算机适性测验的问题通常是试题库中选取，使用统计工具来分配难易度（Gershon，2005）。因此，大部分系统变得很复杂并且依赖大量的、具有标准化试题的试题库。Pyper & Lilley（2010）描述了一个比较简单的“flexilevel系统，该系统采用固定的分支技术来选择下一个问题以及在哪个阶段呈现给学生。

适性测验的另外一个用法是创造“迷宫”，在某些测验中，后面问题依据学生对前面问题的回答，没有必要给出“正确”或其他答案。Wyllie & Weight（2010）设计了临床决定迷宫，依据各种信息资源，模仿要获得的决定，从而确定如何治疗一位患有精神疾病的老年人。这种迷宫的类型为提高机考的真实性提供了一种方法。

海里特-瓦特大学计算机辅助数学学习项目团队成员正在寻找办法，希望在考试中加入模拟元素（Ashton & Thomas，2006），例如，采用分层模式，在一边的屏幕上加入实践练习，另一边加上问题（Thomas & Milligan，2003）, 分层的目的是要提高互动性，“在学习环境中学习”来检测学生（Ashton 等，2006b, p.125）。

另一个将教学与考试相结合的努力尝试是为教科书匹配题库，最著名的一款此类产品就是培生集团的“精通”系列，如“精通物理”（Pearson，2013）。尽管此类问题通常被描述为“家庭作业”，但是老师们对如何使用这些问题保留了自己的选择权。培生系统中所提供的对学生提问的分析工具也在学生给出错误信息时起到了引导作用。例如，Wale & Birch（2012）使用了“及时性”教学模式，学生首先通过讲座进行自我学习，然后参与在线测试，测试结果在几小时后知晓。Wale & Birch 使用精通物理，但是发现学生对“家庭作业模式”下的问题（没有提示）并不能很好地理解，所以现在他们使用“辅导模式”（带有提示）并在必要的地方加上反馈。

10. 超越测验的在线测评


在第三和第四部分讨论了使用PeerWise系统让学习创作和复习问题，但是通常来说同伴测评是指由同伴来评价学生的作业。这种方法可以节省教师
资源，无教师再给学生提供额外的反馈，但是Honeychurch等（2013）指出同伴测评的真正价值“不在于反馈本身（反馈作为一种产品），而是进行反馈的过程”。技术能够支持同伴测评，更适合大规模的班级和在线学习环境（Luxton-Reilly，2009，Honeychurch等，2013）。


Sim&Hew（2010）确认博客是电子文件夹很自然的一部分，学生通过博客可以分享学习经验并进行反思。Cruchill（2009）鼓励学生参与到博客活动中，并将其作为课程评估的一部分。博客活动受到了学生的欢迎，并有超过一半的学生表示即使不作为课程评估的一部分，也会继续自己的博客活动。

Caple&Bogle（2013）热衷于使用维基来评价小组协作活动。小组中的任何一名成员都可以修改维基页面，其优势是每次修改都被记录下来并归于做修改的特殊使用，这意味着不但可以评估小组活动，也可以评估个人活动。网上论坛也是很有用的协作工具，但是Conole&Warburton（2005）明确论坛中“估量”不同互动的难度。

四、在线测评的前景

基于以上对在线测评文献及相关教育技术发展的回顾，我们来预测在线测评的未来发展并讨论可能存在的困难。

1. 大规模开放在线课程(MOOCs)

2013年夏天，Cathy Sandeen在《测评研究与实践期刊》上撰写了MOOCs专题，吸引了我们的注意，并由此在教育领域掀起了前所未见的热情、实验、讨论和辩论（Sandeen 2013, p.11）。无论MOOCs的未来是什么，积极的一面是此类课程能够促使设计者考虑采用恰当的测评方法，测评大规模的学生非正式和在线学习。

MOOCs最初提供给学生时是免费的并且无学分。在这个阶段，‘Masters’（2011）指出，“在MOOCs课程中，测评并不能激发学习；学习者的目的是学习的驱动力”，这一论断是合理的。然而，大部分的MOOCs课程都会基于一定程度的参与或者取得的成绩签发某种程度的“徽章”。如果要衡量所获得的成绩，那么就需要某种程度的测评。另外，形成性考核提供了参与的可能性并激发学生积极参与，所有的这些因素都可能提高大部分MOOCs课程的完成率。

学习MOOCs课程可以很容易地通过完全在线的机考、协作式和同伴测评方式来运作。机考可以低成本、快捷地传输，但存在低质量测评的风险。为避免这一风险，MOOC体系需要提供多种类的问题，提供有意义的、即时的反馈，允许学生多次尝试问题。另外一个重要的因素是，不同的学生最好接收到不同的问题，这样就需要题库或者问题的多变性。最后，在提供了高质量的工具后，需要保证MOOC创作者经过训练设计出高质量的问题，教师们不应成为“被忽略的学习者”（Sangwin&Grove，2006）。

2. 学习分析和测评分析

学习分析可以被定义为“对学习者及其所处学习环境的数据的测量、收集、分析和报告，目的是理解并其发生环境并对其最优化”（Ferguson，2012，p.305），Redecker等（2012）建议我们应该“超越测评的范畴”，在测评中使用学习分析。在网络环境中收集到的学生互动数据使得掌握学生的真实互动成为可能，从而不需要加入独立的测评环节。Clow（2012）指出学习分析系统甚至可以在非正式的情景中提供如测评一样的反馈。

广义上来讲，学习分析能够向教师告知学生的学习情况，Ellis（2013）呼吁进行“测评分析”（分析测评数据），指出测评在高等教育中无处不在，而在网络学习环境（尤其是社交媒体）中，缺少学生互动测评。Jordan（2013）阐述过网络学习环境（尤其是社交媒体）中开展测评的潜在可能，她分析了学生在机考中的表现，解释了学生的误解，但是也展示了更多深层次参与的内在驱动力。
通过研究两组学生，她发现了对相同作业上存在不同的完成模式，能够将不同模式与不同学生、不同工作量建立联系。

3. 计算机、评估和学习之间界限不断模糊

如果如上文所建议的那样，我们使用学习分析作为测评，使用学习分析来帮助学生学习，那么，测评与学习之间的界限就越来越模糊了。同样，更加复杂专业的系统应该应用适性测验，为个性化的形成性考试和诊断性考试（也许也包括终结性考试）提供机会，从而能够更好地支持学生们的练习。

也许，使用移动设备的网上教学能够将教学资源嵌入测评之中，反之亦然。在教学最恰当的时候提出问题，无论何时何地，学生都可以进行测试。考试形式的作业再也不是一个独立的个体。课本内容呈现过程中设置问题是一种常见做法，但是现在这种问题实现了互动。学生可以学习到一系列的教学资源从而帮助他们回答问题，而任何“标准答案”都隐藏起来，直到学生提交了自己的答案才会显示。

4. 计算机的恰当使用

Phil Butcher 在2013年的eSTeM（2013）年度论坛上回顾了英开机考的使用，“现在是什么样呢？我是否有建议要开始将计算机当成一个计算机来使用？”他指的具体问题是在Moodle平台中引入了STACK问题类型（Butcher 等，2013）。然而，还有一些近期案例说明计算机能够精准地评分，例如Coderunner（Lobb 2013）通过运行学生所写的编码来测评计算机编程技能。除了用计算机来计算和评估测试本身外，它有可能会利用技术提高问题的质量，例如，从学生答案到自由文本简答问题，采用机器学习来设计答案匹配规则（Jordan 2012a）。

然而，计算机仅当其适用时才能被采用。有时，混合的方法更加有效，数学学科个人学习过程中半自动分析能够自动地监控学生的互动，如果有必要，会将这些互动发送给辅导教师，由其填写详细的反馈（Herdin & Schroeder, 2012）。Butcher & Jordan（2010）建议计算机无法“识别的”简答题应该进行人工评分。当前，还有一些测评任务（例如实验报告、小论文、论文）对机评来说仍具挑战。采用小论文评分软件做形成性考试是合理的，但是问题是能否广泛应用于终结性考试，就如McGuire 等（2002）十年前认知的那样，我们不应错误地认为将问题分解为多个步骤就测评了问题解决过程中的所有技能。概括起来，我们应该利用计算机做其擅长的。解放辅导教师人工评分的部分繁琐劳动，从而使他们去测评那些必须由他们本人去判断真实性的部分。

【参考文献】

http://oro.open.ac.uk/38536/