【本文检索信息】Randy E. Bennett.教育测量的未来趋势[J].教育测量与评价,2019(3):3-14;18.

教育测量的未来趋势

[美]Randy E. Bennett

Bennett 博士是美国教育测量服务中心(Educational Testing Service, ETS)测量创新部首任 Norman O. Frederiksen 主任(2010 年至今),“国际教育评估协会”(International Association for Educational Assessment, IAEA)主席,(美国)全国教育测量学会(National Council on Measurement in Education,NCME)前任主席。1979 年入职 ETS,领导或者参与过许多重要的教育测量项目,包括探索计算机测试的“NAEP 技术支撑评估项目”(NAEP Technology Based Assessment Project)以及 ETS 的创新项目“针对学习、为了学习和作为学习的、以认知为基础的评估”(Cognitively-Based Assessment of,for,and as Learning,CBAL)。其研究专长是利用认知科学、技术和测量领域的创新,发明新的测量方法。

【摘 要】本文根据作者于2018年4月在纽约召开的(美国)全国教育测量学会(NCME)年会上的主席演讲稿修改而成。作者首先介绍了未来教育测量发展变化的11个可能特征、每个特征之所以重要的原因,以及应该如何看待这些变化。随后概述了未来教育测量领域不太可能发生变化的几个方面。最后对今后十年的教育发展进行了展望,并就这些发展对教育测量工作者可能产生的影响进行了讨论。

【关键词】评估;教育的未来;形成性评估;测量
【中图分类号】G40-058.1 【文献标识码】A 【DOI编码】10.16518/j.cnki.emae.2019.03.001

本文[①②]将描述未来教育测量在不同测试形式和应用方面的一些可能特征。笔者首先说明每个特征为什么重要,以及如何看待这些特征。所以,本文第一部分讨论未来教育测量可能发生的变化。第二部分比较短,讨论教育测量的哪些方面不会发生变化。最后,对未来十年的教育做了一个简单的展望,并讨论未来教育对测量的意义。

本文首先列出 11 项特征,有些特征显而易见,因为它们在教育测量中已然发生。但是,所有这些特征目前呈现的规模和范围仍然有限。笔者相信,这些特征在未来教育测量中的表现将会变得越来越明朗,将远远超出目前已有研究课题和高端运营环境的范畴。

未来教育测量的特征是:

*以技术为依托

*测量“新”构念

*建立在更深层次的认知和学习模型的基础上

*更充分利用复杂任务

*更“个性化”

*试图改善学习

*更好地考虑学生的背景

*“嵌入”教学活动并分布在不同时间

*采用自动评分
*把新的探索方法整合到建模和分析中
*提供更有效的测量报告

一、未来教育测量可能发生的变化

1.以技术为依托

利用电子技术传送测量的重要性至少表现在3个方面:第一,通过加速试题呈现和反应回收可以更有效和快捷地测试学生传统意义的学力(competence,学生所学的知识、技能和能力的统称);第二,可以对传统方法无法测量的、新的学力进行测试,例如用计算机进行写作、在有网络链接文本的条件下进行阅读、在虚拟空间与远距离同伴进行合作,以及对问题解决过程本身进行测量;第三,使人们收集和分析在线学习的"大数据"成为可能。

我们应该关注什么呢?我们应该关注国际性的、全国性的、全州性的主要测量课题。这类测量课题的前景、规模和资源占有优势,能够而且也确实把创新作为反映其合理性和示范作用的机制。

在国际教育评估中,第一个值得关注的是"国际学生评估计划"(Program for International Student Assessment,PISA)。2015年,PISA在57个国家或地区对40万15岁少年进行了测试,③[1]测试内容包括阅读、数学、科学、问题解决和金融素养。PISA的特殊之处在于该计划以技术为依托,测验试卷有90种语言,在管理、内容开发、测验题呈现、收集反应和分析数据方面存在大量挑战。

另一个值得关注的是在线"国际阅读素养研究进展"(Progress in International Reading Literacy Study),即ePIRLS。2016年,该在线测量利用学校的计算机对16个不同教育体系中大约85,000名4年级学生进行了测试。[2][3]ePIRLS的特点是测量在线信息阅读,有些人认为在线阅读与传统测量中的阅读理解有很大差别,当然这是一个仍有争议的问题。在ePIRLS中,阅读文本的选择包括网上链接的材料和存储在计算机不同文件夹里的材料,搜索阅读材料的方法与传统阅读测量中搜寻信息的线性加工完全不同。

在美国的全国性评估课题中,2017年美国"全国教育进展评估"(National Assessment of Educational Progress,NAEP)项目对全美4年级和8年级15万名学生的阅读和数学进行了测试[4],并对这两个年级中每个年级大约2万名学生的写作进行了评估[5]。NAEP的特点是测试人员把计算机带去各个学校,以期更好地控制由于各个学校计算机的不同而导致的学生操作上的差别。NAEP之所以能够采用这种方式,是因为它只是对为数不多的学校中比较少量的学生进行测试。NAEP运用计算机技术进行测试的方法具有特殊性,开始时进行了大量研究[6][7][8],随后才进行少量的操作性测试(2009年采用计算机互动进行科学测试[9]),再后来运用到全美样本的测试(写作、技术和工程素养测试[10][11]),最近扩展到州和全美样本的主要评估项目。

2018年起,"澳大利亚国家评估计划"(Australian National Assessment Program,NAPLAN)对3年级、5年级、7年级和9年级学生的在线素养和算术素养进行了评估。[12]NAPLAN使用学校的计算机和"经过批准使用的私人设备"进行测试。显然,这是唯一大型测量中采用这种测试方法的评估计划。

最后,在美国州一级的测试中,"加州学生操作和进步评估"(California Assessment of Student Performance and Progress,CAASPP)项目具有重要意义。2017年9月27日有美国媒体介绍了CAASPP的在线测试。该报道说:"加州对全州320万学生的测试进行得很顺利。在同一天时间里(2017年5月9日)50万学生参加了在线测试,这是有史以来单日在线测试学生最多的一天。"[13]CAASPP在线测试的成功说明以计算机技术为依托进行大规模测试是可行的。值得注意的是,从考生数量上看,世界上大多数K-12年级测验,以及美国许多K-12年级测验,仍然在用纸笔进行测试。

除了这些测量计划带来的创新,我们还应该

关注测量技术对评估结果的影响。应该特别注意测试结果在以下几方面的差别:(1)从纸笔式到数字化测试形式的变化;(2)学生对计算机的熟悉程度的差别;(3)不同人群之间的差别;(4)技术变化在时间上的差别;(5)语言的差别。除此之外,我们还应该关注那些具有现实意义的、技术方面的、政策方面的,以及政治方面的和保存测量原始意义方面的需求(比如有人认为计算机测试和纸笔式测试对考生的操作有不同的影响,因而对NAEP计划的州与州之间的比较提出质疑[14])。最后,对于有些测量课题来说,电子技术还算是新技术,需要注意它们在具体执行过程中不要走歪[15][16],否则就会导致倒退。

2.测量"新"构念

未来教育测量的第二个特征是对新构念(construct)的测量。为什么?因为在教育领域和职业上的成功,以及作为优秀公民需要的东西比目前所测量的或教学的东西多得多。从个体水平来说,问题解决过程——可以利用技术促进测量[17]——社会情绪方面的"坚毅性"(grit),社会意识和自我意识[18],都是新的构念。从群体水平来说,通常提到的小组功能或者小组协作[19]也是新的构念。从组织水平来说,环境因素,比如课堂和学校氛围,对于管理和理解学生越来越重要。[20][21]

我们应该关注什么呢?应该关注教育决策者对这些构念进行测量的接受程度(比如,学校追责和大学录取中在何种程度上考虑这些新构念)。迄今为止,这方面评估结果的应用还非常有限,因为从直观上(事实也是如此)来看这类评估还不够完善[22],最大的担忧是这类构念容易受人为操纵的影响。也需要关注它们在形成性测量中应用的程度,在某些情况下这类新构念的测量也许更适合形成性测量。

3.建立在更深层次的认知和学习模型的基础上

未来教育测量的第三个特征是测量要建立在更丰富、更深层次的认知和学习模型的基础上。这些以理论为基础的模型对课程材料的组织比课程标准更合理。所以,这样的模型对于评估设计和项目编制能够提供更有用的指导。[23]当然,对这些模型进行实证、补充,有时甚至从数据中发现这些模型也很重要。这类模型也许能够对通过学习进步而产生的变化进行更有意义的测量,这里学习进步通常指某些特定学力达到熟练水平的程度。[24][25][26]

我们应该关注什么呢?应该关注根据这些模型所建构的测量的预测程度。[27][28]例如,当一个测验试图测量一个或者多个方面学习进步的时候,试题的难度与学习进步的水平相配吗?根据某个理论设计测验,每次施测时可以对该理论的假设进行评价。[29]例如,我们可以期望大多数学生的反应模式与理论模型所预测的反应模式是一致的。此外,还要关注这些模型对教师组织教学和指导他们在课堂上对学生学习进步的评价是否有用。[23][30]

4.更充分利用复杂任务

未来教育测量将更加充分地利用复杂的问题作为测验题。为什么?因为一个学科中具有熟练水平特征的活动通常包含拓展性问题解决情境。由于实际条件的局限,这样的问题解决情境在测量中很难得以重复。我们在测量中通常采用的任务和学科所定义的任务之间的脱节,是导致只有58%的美国公立学校学生家长认为目前的教育测验成功测量了儿童学习的原因。[31]过去我们在这个方面的努力是采用小论文、简单实验,或者收集学生档案使测量题近似于复杂问题解决的活动。[32][33]近年来,以计算机技术为基础的模拟和游戏形式来代替复杂问题解决活动,已经得到了应用或者被提出来了。[34][35]

我们应该关注什么呢?应该关注教育游戏和模拟中能够直接应用或者修改以后才能够应用的任务类型的设想和范例;关注与某些复杂活动近似的职业和专业评估[36];关注NAEP,因为这是一个少有的、在K-12年级测试中应用模拟任务及其他操作题的一个测量计划。在2009年的科学测量、2014年的技术和工程素养评估中的考生与计算机互动任务是NAEP在这方面的范例。[9][11]

但是，我们也应该认识到一些老问题的存在，包括测验题所覆盖的范围、对不同群体和个人的公平性、开发测量和评分的开销，以及考生完成这些试题所需要的时间[37][38]等。我们可以期许用新的方法（或者新近已经应用的方法）来减少这些问题的困扰。例如，有些测量课题把操作测验与较短的问题片段结合起来，这种方法已经在美国大学预修课程（Advanced Placement，AP）测量和 NAEP 测量中应用多年。一种较新的设想是运用结构化的操作测量题，也就是把一个大的任务拆分成一系列局部相对独立，而且比整个操作活动短小的问题。[39]另外，我们还可以期许更加成熟的开发工具的出现，例如，基于“证据为中心设计”的设计模式[40]，以及自动评分法的应用，这二者的结合可以降低测量前期和后期的开销。

5.“个性化”测量

未来的测验题在某种意义上来说会变得“个性化”。原因很简单，接受教育、接受评估的学生，具有不同水平和不同类型的学力、不同的背景和不同的兴趣。如果我们在评估中能够更好地包容这些差别，我们对学生知道什么和能够做些什么的特征的描述就可以更准确。这种思想，广义上说，其背景就是媒体通常所说的个性化学习运动的兴起。[41]

个性化测量有多个维度，所有维度都值得关注。第一个维度是惠及性（accessibility）。惠及性问题至少从 1938 年 SAT 第一次出现盲人版和大号字体版就开始了。[42]最近出现的“益智平衡测验”（smarter balanced）在测量过程中引入了许多内置辅助技术[43]，显著提高了测量课题所能提供的信息的标准。另一个更新的是美国研究生入学考试（GRE）中普通测验所开发的技术，该测验采用的是考生日常使用的人机交流技术（利用普通的商用显示屏，而非测量者自行开发的考生不熟悉的显示屏）。[44]

另一类个性化是适应性测量（adaptive testing）。适应性测量最简单的形式是使试题的难度与所要评估的技能水平相匹配，这项设想至少可以追溯到 1916 年斯坦福—比纳智力量表的开发。[45]1986 年这项设想通过项目反应理论（IRT）在美国大学委员会的计算机化大学预修课程测量中得到了实现[46]，这是最早的适应性测量之一。更现代化、更具有挑战性的适应性测量是不仅对考生的能力和试题的难度进行匹配，而且对试题的内容、考生的背景及其兴趣进行匹配。这样的匹配，不仅可以吸引考生的参与，而且可以减少过去操作测量“人—任务交流”中所固有的不公平。[37][47]当然，这个观点本身也引起了对测量公平性的质疑：如果考生不同意我们认为他或者她应该对什么感兴趣，怎么办？

还有一种个性化测量是让考生自己选择试题。美国大学预修课程（AP）测量中的美国史考试 B 部分的第二节测验属于这一类[48]，允许考生从有关 3 个历史时期（例如南北战争、第二次世界大战、越南战争）的 3 个试题中任选一个进行论述，但是测量的是相同的推理技能。对于这种个性化的试题，人们担心的是如果考生选择得不好，测量效度会不会有问题。过去有关这个问题的研究没有明确的结论。[49]例如，Bridgeman 等人[50]要求学生事先自选一个认为自己会答得较好的论述题，然后要求考生对两个论述题都作答，结果发现多数学生自选得不错，但是 10 个学生中有 3 个学生在自选题上的得分低于他们没有选择的论述题上的得分。

最后一种个性化测量的形式是自己选择目标或者课程标准。简单的例子是允许考生自己选择把哪些课程的分数送给大学招生处。复杂的例子包括美国大学预修课程中艺术工作室三维设计组合的第二部分：作品深度（concentration），这个测量“没有优选（或不可接受的）风格或内容”。[51]考生送交的作品可以是有形或无形的雕塑作品、建筑模型、金属作品、玻璃作品、组装品、操作作品、组合品和三维纺织/纤维艺术品。更极端的情况是，评估的作用在于衡量学校教师与学生协商的具体的目标。对大学预修课程的艺术工作室测量来说，师生的这种协商是在共同的、高水平的目标和评估标准的条件下进行的，这些目

标和评估标准针对的是特定的学生。在这类测量中，测量组织者往往会制定出详细而通用的评分标准，并且会对评分人员进行严格的培训，以应对评分的挑战。

6.试图改善学习

传统的美国州立追责测量（accountability test）是通过给决策机构提供信息（例如，通过测量了解某所学校需要特别关注、某项课程的教师需要特别加以培训），来间接提高学生学习的水平。但是，这种测量的价值越来越受到质疑，很多公众对追责测量抱有负面的态度，轻者认为这是浪费教学时间，重者认为测量对学生有害无益。[31]公众的这种态度导致奥巴马总统于2015年削减了联邦政府要求的测量时间。[52]“益智平衡评估联盟”（Smarter Balanced Assessment Consortium）、“大学和职业预备评估协作组织”（Partnership for Assessment of Readiness for College and Careers，PARCC），以及其他许多州由此缩减了学生的测试时间。有些州的学生则选择完全退出测量。[53]例如，在纽约州，2017年有19%的学生拒绝参加州政府的测量。[54]

我们需要关注什么呢？需要关注测量的编制，不仅要测得好，还要将测量题设计成教师的教学模型并且使之能够指导学生的学习。[55]我们要寻找能够帮助学生更多地了解重要议题、引发其偶然学习的测量。我们还需要关注向学生提供优质反馈、鼓励学生对问题解决过程进行自我反省的测量（例如，学生怎样撰写论文或者怎样进行模拟科学实验）。

7.更好地考虑学生的背景

未来的测量会更好地考虑学生背景的影响。大规模总结性评估的设计是超越“背景”的，无视个体或者群体的社会方面、学习方面和教学方面的影响，使有关测量的推论在不同的背景条件下都具有普遍的意义。一个学生或者一个群体对一个测验题如何操作是一个事实，但是这些学生“为什么”（why）这样操作是一种“解释”，做出这种解释时需要知道他们的背景，然而现在我们只是试图利用学校、教师和学生回答背景问卷的材料，从最基本的层面说明学生背景的影响。

在电子学习环境中嵌入评估非常值得关注，因为把评估嵌入学习环境会使测量结果更加具有实际的价值，评估所需要的内容、知识、工具与学习环境相同。[56]这种嵌入实际上是使评估的环境和教学合二为一。

另外值得关注的是，测量课题本身试图变得更加“嵌入”。

8.“嵌入”教学活动并分布在不同时间

为什么测量课题会变得（或者看起来变得）更“嵌入”呢？上面讲到的一个原因是为了更好地考虑学习环境的因素，而另一个原因则是测量课题要避免被电子学习（或者其他技术）公司取代。如果别的产业能够提供更简单、成本更低的替代品而成为测量课题的竞争对象，测量课题就可能被替代或者被取消。[57]

嵌入式测量是什么意思呢？一种意思是作为一种跨时间、随意地（但是广泛地）对学生在学校（或者其他任何地方）正在做什么的行为进行取样的机制。这里说“取样”（sampling），因为我们无法完整地记录所发生的一切事情；这里说样本是“随意的”（causal），因为它并不是按照一定的设计，根据某个确定的质量标准，对某方面的能力在一定的熟练范围内提供证据。这只是一种偶然的取样（incidental sampling）。但是这种偶然取样的范围很广，可以提供大量的数据。如果所有学生在教学活动中采用同样的电子学习环境，或者把学生在所有学习环境中的数据都收集起来，数据量就相当可观了。在极端情况下，如果学生的学习活动完全变成了工具性学习——把传感器、相机、麦克风安装在学习室里，用它们把学生的每次按键和鼠标动作都记录下来，那么，这种完整的工具性学习就会导致“大数据”的产生，而大数据又会导致什么呢？

也就是说，我们应该怎样应用随意地（但是范围广泛地）收集起来的行为样本的大数据呢？一种方法是对样本进行描述，也就是用样本行为说明学生正在做什么或者正在学习什么。这种描述比较容易，也可以是很有价值的描述，这使我

们能够精准地描述一个课堂与另一个课堂、一名教师与另一名教师、一所学校与另一所学校、一个学区与另一个学区，以及一个群体与另一个群体之间的不同，从而能够把“学生进步”这样的增值结果与随意机制联系起来。可惜的是，所有这些目前都无法做到。

这些随意收集起来的行为样本的另一个用途是对这些行为进行推理，即通过这些行为样本对学生知道什么和能够做什么，推断出能够进行相互比较的结论。由于数据具有随意性，我们很难据此做出有意义的推理。

然而，把测量“嵌入”到教学活动之中的第二种意思是，在某个点上把设计好的事件插入课程之中，记录会发生什么。如果每个这样的有针对性的探测都采用我们称之为“学习挑战”的形式，那么这样的挑战就可能发生在一堂课或者几堂课之中，也可能发生在试题中、游戏中、模拟任务中，或者其他操作任务中，即可能发生在任何一种符合测量预期目的和用途的收集证据的机会之中。当然，这种定期性的评估所能提供的行为样本的范围少得多。这种定期性的评估也可能与学习环境不太协调，因为这种评估来自学习环境之外。但是这种评估是经过“设计”的，人们可以争辩说，这样的评估相比随意收集的学生的行为样本，能够提供更有力和更可比的证据，也更有可能得出学生知道了什么和能够做什么的推论。

因此，我们应该关注，通过这两种嵌入方式收集数据的测量公司、电子学习公司和技术公司这类实体，尤其应该关注把随意地、完全嵌入学习活动的评估所获得的数据信息，与经过设计的、与教学背景有较大差别的评估方法所获得的信息结合起来，也就是使两种方法获得的信息进行互补。研究二者不一致的原因，也许可以提高两种方法的质量。此外，在连续不断地记录教师和学生的行为时，涉及个人隐私的安全，很多家长、教育工作者以及教育决策人员对此表示了担忧，对于这一现象，我们也需要关注。[15][16]

9.采用自动评分

未来教育测量的第九个特征是，未来的评估可能更倚重于自动评分（automated scoring）。这种评分将允许采用更复杂的测量任务，减少评分时间及其财力支出，提高效率，同时对学生的操作提供更详细的反馈。

自动评分中的“黑箱”算法（algorithm）绝对值得特别关注。这些算法通常可以合理预测采用人工方法对作文或者演讲样本进行评分或者分级的精确性。[58]黑箱算法也许对形成性测量更合适，因为在形成性测量中做出错误决定的后果通常不太严重，决定本身也容易得到纠正。相对来说，黑箱算法对具有高风险（high-stake）的决策带来的问题更多，因为他们采用的算法高深莫测或者干脆就是他们自己的独家算法——所以称之为“黑箱”。换句话说，我们无法知道他们是如何给每个人进行评分的。在预测时，他们可能采用相关法，但是并不考虑那些相关因子与所要测量的构念是否关联。这样一来，提高相关因子的地位就能够提高测量分数，但并不一定提高了所要测量的能力。在许多自动评分系统中，作文长度通常直接或间接地用作自动评分的例子，例如增加组织好的、没有语法错误的、空洞的文字，就能提高考生的分数，但其实这并不意味着考生的写作能力更好。

在黑箱技术方面，我们应该关注欧盟《通用数据保护条例》（*General Data Protection Regulation*[59]）这样的法律法规。这个条例受到了大量的关注，因为它导致许多美国公司修改了他们的保密政策，并通知给了数百万消费者。鲜为人知的是，截至 2018 年，《通用数据保护条例》赋予了个人要求解释对其产生重要影响的算法决策的权力。[60]人们可以想象，在给不给消费者贷款时，算法决定规则是多么重要。《通用数据保护条例》等法规的颁布实施，促进了“可解释人工智能”（Explainable Artificial Intelligence，XAI）的兴起。[61][62]

根据《通用数据保护条例》的要求，自动评分的目标是使评分过程能够进行解释，而且评分结果与所测构念的定义必须一致。例如，一个测量议论性写作技能的测验，其自动评分就应该

侧重于有关论文的反应特性，如所用推理支持所持观点、支持每个推论的证据的强度，以及对可能相反的观点进行反驳的质量。具有人工智能方面经验的人都知道，对机器来说，这一类分析在目前看来是有多么困难。但是，正如在电影《红粉联盟》(*A League of Their Own*)中 Tom Hans 对 Madonna 说的："如果不难，每个人都会做；正是因为难，事情才伟大！"请注意这"伟大"二字。

10.把新的探索方法整合到建模和分析中

未来教育测量的第十项变化，是需要把新的探索方法整合到建模和分析之中。在线学习和在线评估会导致新的数据类型的出现，其中，一个突出的例子是有关测量过程数据的出现。[63]在技术支撑的评估中，我们可以记录考生每个活动的类型、活动发生的时间、活动延续的长度，以及该活动前后考生做了什么，所有这些都有可能是有用信息。相对来说，经典测量学模型适合更简单的数据，如二项选择题或多重记分题的反应，这些反应具有更严格的假设。

值得一提的是，有几个方面的进步值得关注，包括教育数据挖掘（educational data mining，EDM）和学习分析（learning analytics，LA），二者都关注大数据在教学中的运用和推动学习科学的发展。[64]最后，我们应该继续关注统计学，因为现代测量学中的一些重要内容就是从统计学中发展起来的[65]，而且上述领域中新方法出现的主要驱动力也还是统计学。

11.提供更有效的测量报告

最后，未来的教育评估最好能够提供更有效的测量报告。这有三个方面的原因。首先，分数报告是分数使用者体验的一个重要部分，考生参加测量的最初动机当然会认为测量结果是有用的。其次，测量报告的正面影响远未得到充分利用。测量报告是指导教师、学生、家长以及政策制定人员的思想和行动，提高测量质量的可能途径。最后，与测量的其他方面相比，测量报告的进步非常缓慢。目前常见的操作测量的报告，包括加利福尼亚州学生操作和进展评估课题中用电子手段给上百万考生发送的测量报告、"益智平衡测量"和普通 GRE 测验中的适应性测量报告、NAEP 和美国医疗执照考试（Medical Licensing Examination）中的模拟任务报告，以及 GRE 写作部分、TOEFL 测验、"益智平衡测量"中的自动记分报告，我们很少看到创新。

我们应该关注什么呢？应该着重报告那些更适合考生具体情况或是更贴近评估用途的内容。例如，我们应该关注重播操作的过程，体育运动教练经常用这种方法分析运动员或者运动队做了什么，据此帮助他们提高运动水平。重播学生的操作过程对问题解决可能有用，因为问题解决过程也是评估其操作的一部分。制定一项科学实验计划和进行一项科学研究[66]就是一个很好的例子：实验的最终结果当然可以是评分的依据，而获得结果的过程是否符合规范以及能够经受质询的程度，也可以作为评分的依据。即使考生操作的过程并非评判的内容，重播操作也可能具有一定价值。如果学生最终的作品存在缺陷，那么查看操作过程也许可以得到一些改正的建议。以写作为例，在观察论文写作过程的重播中，我们可以清楚地看到有些学生拿到作文题后很快就开始写作，没有计划，没有修改，他们在计算机上敲完最后一个键后马上就发送了出去，没有留时间再读一遍。观察和反思这些行为对这些学生和教师来说都是有帮助的。同样，评论群体的典型行为样本也可以帮助决策者和公众更具体地理解测量结果的意义。

除了重播操作过程，我们还应该努力探索采用简单、直观和能够让使用者直接参与的方式（例如，用游戏的成分）来报告测量结果。PISA 2015 科学测量结果采用了互动报告屏的方式。互动报告屏根据各个参与国家和地区的操作水平，用不同颜色表示了 9 种类型，让人一看就明白哪些国家和地区参加了测试，其操作在成就水平的什么位置。在计算机屏幕上点击该国家或地区，可以看到任何一个参加了 PISA 测试的国家或地区在这项测试以及该轮测试中其他测量的详细报告。

二、未来教育测量哪些方面不太可能发生变化

上文所述是笔者认为未来教育测量将会发生变化的方面。那么，哪些方面不会变呢？

首先，测量的基本特性不太可能发生变化。总体来说，测量包括4个方面的特性。第一，创造机会使我们能够对我们希望知道的学力进行观察，并收集证据。狭义而言，这包括测量设计、项目编制和施测。第二，把我们观察到的证据与考生个体、群体，或者组织机构有意义的特征联系起来（并估计与这些特征联系的不确定性）。这就是测量模型的问题。第三，在决策或者分数报告方面进行交流，应用测量的结果。第四，评价证据收集的机会、特征，以及决策的质量和影响，即评价测量的效度。从这些意义上来说，未来测量将保持不变。

其次，测量所针对的重大社会问题也不会发生变化。这些问题包括收集教育体系有效性的材料、掌控主要社群之间在教育成就上的差异、给学生个体提供资源分配的信息（例如，录入顶尖大学），以及帮助学校提高学习和教学质量。这些问题长久以来就伴随着测量的发展。毫无疑问，这些问题还将继续存在。

再次，教育测量潜在的社会价值也不会改变。[67]这些社会价值包括效度、公平性、可比性和可重复性。可重复性就是信度或“再给我看看”。如果我们不能用行为的另一个样本、另一位评分人，或者在另一个测试场合重复出相似的结果，测量的结果就不可信。

最后，形成性评估和总结性评估的差别也不会改变。当然，经典的总结性评估也能够提供一些（次要的）形成性信息（例如，公布测验项目以后让考生重复其操作[68]）。同样，在某种条件下，形成性评估也可以提供学习质量的证据，对判断学生知道了什么和能够做什么做出（次要的）贡献。有人认为，可以用课堂学习中收集到的测量的量化数据完全替代重要的总结性评估。[34][69][70][71]笔者认为这不太可能，原因如下。

第一个原因是，不同的学区、同一个学区的不同学校，甚至同一所学校的不同班级，所收集的学生的操作材料的类型和数据的质量存在着很大的差别。这种差别的根源在于，具体的学习目的不同，达到这些目的所采用的课程不同，在这些课程中所采用的电子化资料不同，以及根据不同渠道所收集的证据的类型和质量不同。这些巨大差别减少了在不同学生之间、课堂之间、学校之间以及学区之间对学生的熟练水平进行比较推理的可能性。

第二个原因是，决策者和公众都希望，评估学校的质量至少要在某种程度上独立于被评估的学校。人们普遍认为，测量的内容与某所学校日常学习的内容越接近，当采用这个测量内容作为学生到另一所学校或另一个环境以后所学内容优劣的指标时，这个指标的价值就越低。另外，测量越受地域局限，对测验结果的解释也就越受局限。

第三个原因是基于隐私和政治方面的考虑，这些因素还将继续左右学生和教师的行为。美国国内接连不断的私人公司和政府泄漏大量个人信息的事件[72][73][74]以及数据的不当使用[75]，已经导致公众对学校、州和联邦教育机构，以及教育公司的不信任[76]。

最后一个原因是对教和学可能产生潜在的负面影响。学生学习时通常需要参加实验，而参加实验就无法避免会有失败，也有成功。[77]经常地或连续不断地收集学生和教师的操作材料，其可能的一个负面影响就是遏制他们在学习中接受挑战的态度和习惯。同样值得担忧的是，无休止的测量可能导致学生丧失练习和扎实学习的机会。从某种程度上来说，只要测验能够很好地代表学习的内容标准，教师围绕这些标准进行广泛教学，帮助学生为总结性评估做好准备，这就应该是一项理想的活动。可惜，这种思想很大程度上由于把测量操作与教师奖惩联系起来而受到了损害，比如把教师的升迁、任期、奖金以及解雇等奖惩措施，与学生的测量操作捆绑在一起。这样的捆绑使教学的重点变得日趋狭隘，增加了学

生和家长的焦虑，从而导致公众反对测量。[53]由于美国州政府评价教师时已经减少或者取消了对学生测量分数的要求[78]，如果州政府的评估能够更有效地代表教育标准，也许会出现一种利用总结性评估鼓励好的教和学的更加平衡的教学方法。

三、总结

本文的要点是，未来教育测量将会迎来以下变化：

*测量我们认为重要的学力（例如，增加社会—情绪学习）；

*为观察这些学力的证据而设计的机会的性质（按照更好的学习模型编制更复杂的任务；采用更多的背景观察）；

*怎样把收集到的证据与考生的特征联系起来（例如，用新的数学模型）；

*怎样通过更好、更具有互动性的报告在决策中交流测量结果；

*怎么评价测量的质量和影响，例如，更大程度地关注测量对教和学的正面作用，以及公众对测量的积极反映。

同时，我们认为测量的基础、测量所针对的重大社会问题、测量所隐含的社会价值，以及基本的测量类型将不会发生变化。

展望十年后的教育，我们会看到许多可观的变化。这些变化肯定会影响教育的目的（例如，培养学生成为更好的公民）、与现有目的相关的目标，以及教和学的方法。由于职场工作性质会不断发生变化，这必然影响未来教育的目标。与此同时，技术的不断发展也必然会影响未来的教学方法。显然，教育测量必须适应这些变化，否则它就会被淘汰。如要维持测量的合理性，测量工作者必须跳出小圈子，进行实验，接受变化，同时不断关心测量的核心价值和基本原则。否则，如果我们追逐虚妄的时尚，短期内可能有利可图，令人兴奋，长此下去却会失却教和学的意义。

注释：

①原文发表在（美国）全国教育测量学会（NCME）主办的《教育测量：问题和实践》杂志（*Educational Measurement: Issues and Practice*）2018年第4期，原标题是*Educational Assessment: What to Watch in a Rapidly Changing World*，本刊转载时略有删减，并得到Wiley出版公司和作者的许可。作者感谢Bob Mislevy、Andreas Oranje和Rebecca Zwick对本文初稿的评论。

②本文中文翻译由刘育明完成。译者为美国教育测量服务中心（ETS）教育测量学者，博士。

③该信息来自2017年11月21日作者与M. Von Davier的私人交流。

参考文献：

[1]Office of Economic Cooperation and Development (OECD).PISA Technical Report[EB/OL].[2018-10-10].http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

[2]Institute for Education Sciences (IES). (n.d.). Progress in International Reading Literacy Study (PIRLS)): Countries[EB/OL].[2018-10-10].https://nces.ed.gov/surveys/pirls/countries.asp.

[3]Mullis, I. V. S., & Prendergast, C.O.. Developing the PIRLS 2016 Achievement Items. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.).Methods and Procedures in PIRLS 2016 [Z]. Chestnut Hill, MA: IEA, 2017.

[4]National Assessment Governing Board (NAGB). (n.d. a). 2017 NAEP Mathematics and Reading Assessments [EB/OL].[2018-10-10]. https://www.nationsreportcard.gov/reading_math_2017_highlights/.

[5]Institute for Education Sciences (IES).Writing Assessment[EB/OL].(2018-10-10). https://nces.ed.gov/nationsreportcard/writing/.

[6]Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F..Problem Solving in Technology-Rich Environments: a Report from the NAEP Technology-Based Assessment Project (NCES 2007-466)[EB/OL].[2018-10-10]. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466.

[7]Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B.. Online Assessment in Writing.In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project (NCES 2005-457)[EB/OL]. Washington, DC: National Center for Education Statistics, US Department of Education[2018-10-10]. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457.

[8]Sandene, B., Bennett, R. E., Braswell, J., & Oranje, A..Online Assessment in Mathematics. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), Online Assessment in Mathematics and Writing:

Reports from the NAEP Technology-Based Assessment Project (NCES 2005-457)[EB/OL]. Washington, DC: National Center for Education Statistics, US Department of Education[2018-10-10].http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457.

[9]Institute for Education Sciences (IES).Hands-On Tasks(HOT)and Interactive Computer Tests(ICT)for the 2009 Science Assessment[EB/OL].[2018-10-10]. https://nces.ed.gov/nationsreportcard/tdw/sample_design/2009/2009_samp_nation_science_hot_ict.aspx.

[10]Institute for Education Sciences(IES).The Nation's Report Card: Writing 2011[EB/OL].[2018-10-10]. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012470.

[11]National Assessment Governing Board (NAGB).(n.d.b). 2014 NAEP Technology and Engineering Literacy(TEL)[EB/OL].[2018-10-10]. https://www.nationsreportcard.gov/tel_2014/.

[12]National Assessment Program (NAPLAN).FAQs[EB/OL].(2018-04-20). https://www.nap.edu.au/online-assessment/FAQs.

[13]California Department of Education (CDE).State Schools Chief Tom Torlakson Announces Results of California Assessment of Student Performance and Progress Online Tests (Release #17-67a)[EB/OL].(2017-09-27). https://www.cde.ca.gov/nr/ne/yr17/yr17rel67a.asp.

[14]Barnum, M..Did Computer Testing Muddle This Year's NAEP Results? Testing Group Says No; Others Are Unconvinced[EB/OL].(2018-04-10). https://www.chalkbeat.org/posts/us/2018/04/10/did-computer-testing-muddle-this-years-naep-results-testing-group-says-no-others-are-unconvinced/.

[15]Herold, B..Online State Testing in 2018: Mostly Smooth, with One Glaring Exception. Education Week[EB/OL].(2018-05-09).https://www.edweek.org/ew/articles/2018/05/09/online-state-testing-in-2018-mostly-smooth.html.

[16]Herold, B.. How (and Why)Ed-Tech Companies Are Tracking Students'Feelings. Education Week[EB/OL].(2018-06-12).https://www.edweek.org/ew/articles/2018/06/12/how-and-why-ed-tech-companies-are-tracking.html?cmp=eml-enl-eu-news1&M=58515710&U=1605540.

[17]Zhang, M., & Deane, P..Process Features in Writing: Internal Structure and Incremental Value over Product Features (RR-15-27)[EB/OL].[2018-10-10]. https://doi.org/10.1002/ets2.12075.

[18]Belfield, C., Bowden, B., Klapp, A., Levin, H., Shand, R., & Zander, S..The Economic Value of Social and Emotional Learning[EB/OL].[2018-10-10] http://blogs.edweek.org/edweek/rulesforengagement/SEL-Revised.pdf.

[19]Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., …, Von Davier, A..Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress [EB/OL].[2018-10-10]. https://nces.ed.gov/nationsreportcard/pdf/researchcenter/collaborative_problem_solving.pdf.

[20]Cohen, J., Pickeral, T., & McCloskey, M..The Challenge of Assessing School Climate [EB/OL].[2018-10-10]. http://www.ascd.org/publications/educational-leadership/dec08/vol66/num04/The-Challenge-of-Assessing-School-Climate.aspx.

[21]Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A.. A Review of School Climate Research[J]. Review of Educational Research, 2013(83): 357-385.

[22]Duckworth, A. L., & Yeager, D. S..Measurement Matters: Assessing Personal Qualities Other than Cognitive Ability for Educational Purposes[J]. Educational Researcher, 2015, 44(4): 237-251.

[23]Bennett, R. E., Deane, P., & Van Rijn, P.W..From Cognitive-Domain Theory to Assessment Practice[J]. Educational Psychologist, 2016(51): 82-107.

[24]Corcoran, T., Mosher, F. A., & Rogat A..Learning Progressions in Scienc: an Evidence-Based Approach to Reform[M].New York: Consortium for Policy Research in Education (CPRE), 2009.

[25]Daro, P., Mosher, F. A., & Corcoran, T..Learning Trajectories in Mathematics: a Foundation for Standards, Curriculum, Assessment, and Instruction(RR-68)[M]. Philadelphia: CPRE, 2011.

[26]Deane, P. D., & Song, Y..The Key Practice, Discuss and Debate Ideas: Conceptual Framework, Literature Review, and Provisionial Learning Progressions for Argumention (RR-15-33)[EB/OL].[2018-10-10]. https://doi.org/10.1002/ets2.12079.

[27]Van Rijn, P. W., Graf. E. A., & Deane, P..Empirical Recovery of Argumentation Learning Progressions in Scenario-Based Assessments of English Language Arts[EB/OL].[2018-10-10]. https://doi.org/10.1016/j.pse.2014.11.004.

[28]Graf, E. A., & Van Rijn, P. W.. Learning Progresions as a Guide for Design: Recommendations Based on Observations from a Mathematics Assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), Handbook of Test Development(2nd Ed.)[M].New York: Routledge, 2016.

[29]Bejar, I. I.. A Generative Approach to Psychological and Educational Measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test Theory for a New Generation of Tests[M]. Hillsdale, NJ: Lawrence Erlbaum, 1993: 323-359.

[30]Heritage, M.. Learning Progressions: Supporting In-

struction and Formative Assessment[EB/OL].[2018-10-10]. http://www.k12.wa.us/assessment/ClassroomAssessmentInte -gration/pubdocs/FASTLearningProgressions.pdf.

[31]Phi Delta Kappan (PDK).The 49th Annual PDF Poll of the Public's Attitudes toward the Public Schools[EB/OL]. [2018-10-10]. http://pdkpoll.org/assets/downloads/PDKna-tional_poll_2017.pdf.

[32]Camp, R..The Place of Portfolios in Our Changing Views of Writing Assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. Choice in Cognitive Measurement[M]. Hillsdale, NJ: Erlbaum, 1993: 183-212.

[33]Wolf, D. P.. Assessment as an Episode of Learning. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. Choice in Cognitive Measurement[M]. Hillsdale, NJ: Erlbaum, 1993: 213-240.

[34]Gee, J. P., & Shaffer, D. W..Looking Where the Light Is Bad: Video Games and the Future of Assessment[EB/OL]. [2018-10-10]. http://edgaps.org/gaps/wp-content/uploads/EDge-Light.pdf.

[35]Mayrath, M.C., Clarke-Midura, J., Robinson, D.H., & Schraw, G. (Eds.).Technology-Based Assessments for 21st Century Skills[M]. Charlotte, NC: Information Age, 2012.

[36]Dillon, G. F., & Clauser, B. E.. Computer-Delivered Patient Simulations in the United States Medical Licensing Examination (USMLE)[J]. Simulation in Healthcare, 2009, 4(1): 30-34.

[37]Shavelson, R. J., Baxter, G. P., & Gao, X.. Sampling Variability of Performance Assessments[J]. Journal of Educational Measurement, 1993(30): 215-232.

[38]Stecher, B., & Klein, S..The Cost of Science Performance Assessments in Large-Scale Testing Programs[J]. Educational Evaluation and Policy Analysis, 1997, 19(1): 1-14.

[39]Deane, P. D., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M., Bennett, R. E., Sabatini, J., & Zhang, M.. The Case for Scenario-Based Assessment of Written Argumentation[Z]. Reading and Writing, 2018.

[40]Mislevy, R. J., & Haertel, G. D..Implications of Evidence-Centered Design for Educational Testing[J]. Educational Measurement: Issues and Practice, 2006, 25(4): 6-20.

[41]Burnette, II, D.. States Take Steps to Fuel Personalized Learning[EB/OL].(2017-11-07). https://www.edweek.org/ew/articles/2017/11/08/states-take-steps-to-fuel-personalized-learning.html.

[42]Saretsky, G.. SATs for the Blind Offered 45 Years Ago[J].Examiner, 1983, 13(7): 3.

[43]Smarter Balanced Assessment Consortium.Smarter Balanced Assessment Consortium: Usability, Accessibility, and Accommodations Guidelines[EB/OL].[2018-10-10]. https://portal.smarterbalanced.org/library/en/usability-accessibility-and-accommodations-guidelines.pdf.

[44]Educational Testing Service(ETS). Accommodations for Test Takers with Disabilities or Health-Related Needs[EB/OL].[2018-10-10]. https://www.ets.org/gre/revised_general/register/disabilities?WT.ac=rx28.

[45]Becker, K..History of the Stanford-Binet Intelligence Scales: Content and Psychometrics (Stanford-Binet Intelligence Scales, Fifth Edition Assessment Service Bulletin No. 1)[EB/OL]. [2018-10-10].https://www.hmhco.com/~/media/sites/home/hmh-assessments/clinical/stanford-binet/pdf/sb5_asb_1.pdf?la=en.

[46]Ward, W. C.. The College Board Computerized Placement Tests: an Application of Computerized Aadaptive Testing[J]. Machine-Mediated Learning, 1988(2): 271-282.

[47]Linn, R. L., & Burton, E..Performance-Based Assessment: Implications of Task Specificity[J]. Educational Measurement: Issues and Practice, 1994, 13(1): 5-8.

[48]College Board.AP United States History Practice Exam: from the Course and Exam Description [EB/OL].[2018-10-10].https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-united-states-history-ced-practice-exam.pdf.

[49]Powers, D. E., & Bennett, R. E..Effects of Allowing Examinees to Select Questions on a Ttest of Divergent Thinking[J].Applied Measurement in Education, 1999(12): 257-279.

[50]Bridgeman, B., Morgan, R., & Wang, M..Choice among Essay Topics: Impact on Performance and Validity[J]. Journal of Educational Measurement, 1997(34): 273-286.

[51]College Board.Studio Art Course Description[EB/OL]. [2018-10-10]. https://apcentral.collegeboard.org/pdf/ap-studio-arts-course-description.pdf?course=ap-studio-art-3-d-design.

[52]Associated Press.Obama Calls for Capping Time Devoted to Standardized Tests [EB/OL].(2015-10-24).http://www.pbs.org/newshour/rundown/obama-calls-cap-class-time-devoted-standardized-tests/.

[53]Bennett, R. E.. Opt Out: an Examination of Issues (RR-16-13)[EB/OL]. (2016-04-25). http://dx.doi.org/10.1002/ets2.12101.

[54]Tyrrell, J.. It's Test Time for NY Students in Grades 3-8[EB/OL].(2018-04-07). https://www.newsday.com/long-island/education/ela-math-opt-out-test-refusals-1.17901520.

[55]Bennett, R. E..Cognitively Based Assessment of, for, and as Learning: a Preliminary Theory of Action for Summative and Formative Assessment[J].Measurement: Interdisciplinary Research and Perspectives, 2010(8): 70-91.

[56]Bennett, R. E.. The Changing Nature of Educational

Assessment[J].Review of Research in Education, 2015(39): 370-407.

[57]Christensen, C. M.. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail[M]. Boston, MA: Harvard University Press, 1997.

[58]Bennett, R. E., & Zhang, M.. Validity and Automated Scoring. In F. Drasgow(Ed.), Technology and Testing: Improving Educational and Psychological Measurement [M].New York: Routledge, 2016: 142-173.

[59]European Commission. (n.d.).2018 Reform of EU Data Protection Rules[EB/OL].[2018-10-10]. https://ec.europa.eu/commission/priorities/justice -and -fundamental -rights/data -protection/2018 -reform -eu -data -protection -rules_en.

[60]Meyer, D.. AI Has a Big Privacy Problem and Europe's New Data Protection Law Is about to Expose It[EB/OL]. (2018-05-25). http://fortune.com/2018/05/25/ai-machine-learning-privacy-gdpr/.

[61]Kuang, C. Can A.I. Be Taught to Explain Itself? New York Times Magazine[EB/OL].(2017-11-21). https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

[62]Maglieri, G., & Comande, G.. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation[J].International Data Privacy Law, 2017(7): 243-265.

[63]Cope, B., & Kalantzis M.. Big Data Comes to School: Implications for Learning, Assessment, and Research[Z]. AERA Open, 2016.

[64]Siemens, G., & Baker, R. S. J. d..Learning Analytics and Educational Data Mining: towards Communication and Collaboration. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge [EB/OL].[2018-10-10]. http://www.upenn.edu/learninganalytics/ryanbaker/LAKs%20r-eformatting%20v2.pdf.

[65]Braun, H. I.. Research on Statistics. In R. E. Bennett & M. Von Davier (Eds.), Advancing Human Assessment: the Methodological, Psychological, and Policy Contributions of ETS [Z]. Cham, Switzerland: Springer Open, 2017.

[66]National Research Council.A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas[M].Washington, DC: The National Academies Press, 2012.

[67]Messick, S.. Foundations of Validity: Meaning and Consequences in Psychological Assessment[J].European Journal of Psychological Assessment, 1994, 10(1): 1-9.

[68]Bennett, R. E.. Formative Assessment: a Critical Review[J]. Assessment in Education: Principles, Policy and Practice, 2011(18): 5-25.

[69]Bennett, R. E.. Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing[EB/OL].[2018-10-10]. https://www.ets.org/research/policy_research_reports/pic-reinvent.

[70]Pellegrino, J. W., Chudowsky, N., & Glaser, R.. Knowing What Students Know: the Science and Design of Educational Assessment[M]. Washington, DC: National Academy Press, 2001.

[71]Tucker, B.. Grand Test Auto: the End of Testing[EB/OL].[2018-10-10]. http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php.

[72]Fung, B.. Equifax's Massive 2017 Data Breach Keeps Getting Worse[EB/OL].(2018-03-01). https://www.washingtonpost.com/news/the-switch/wp/2018/03/01/equifax-keeps-finding-millions-more-people-who-were-affected-by-its-massive-data-breach/?utm_term=.6a07f261799d.

[73]Larson, S.. Every Single Yahoo Account Was Hacked-3 Billion in All [EB/OL].(2017-10-04). http://money.cnn.com/2017/10/03/technology/business/yahoo-breach-3-billion-accounts/index.html.

[74]McCoy, K.. Target to Pay $18.5M for 2013 Data Breach That Affected 41 Million Consumers[EB/OL].(2017-05-23).https://www.usatoday.com/story/money/2017/05/23/target-pay-185m-2013-data-breach-affected-consumers/102063932/.

[75]Granville, K.. Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens[EB/OL].(2018-03-19).https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

[76]Herold, B.. InBloom to Shut Down Amid Growing Data-Privacy Concerns[EB/OL].(2014-04-21). http://blogs.edweek.org/edweek/DigitalEducation/2014/04/inbloom_to_shut_down_amid_growing_data_privacy_concerns.html.

[77]Kapur, M. Productive Failure in Mathematical Problem Solving[J]. Instructional Science, 2010(38): 523-550.

[78]Loewus, , L.. Are States Changing Course on Teacher Evaluation?[EB/OL].(2017-11-28) https://www.edweek.org/ew/articles/2017/11/15/are-states-changing-course-on-teacher-evaluation.html.

(下转第18页)

[6]Stocking, M.L., & Lord, F.M.. Developing a Common Metric in Item Response Theory[J].Applied Psychological Measurement, 1983(7):201~210.

[7]Haberman, S. & Yang, Z.. Regression-Based Simultaneous Linkage of a Large Number of Test Forms via Item Response Theory[C]//Paper Presented at the Annual Meeting of the National Council on Measurement in Education (NCME).New Orleans, Louisiana, USA, 2011(4):7~11.

[8]Yang, Z. & Haberman, S.. Impact of Linking Designs on Simultaneous Linking[C]//Paper Presented at the Annual Meeting of the National Council on Measurement in Education (NCME). Chicago, Illinois, USA, 2015.

[9]杨志明.高考原始分合成:问题与改进思路[J].教育测量与评价,2015(10):61-64.

[10]Cizek, J., Gregory, & Bunch, B. Michael. Standard Setting:a Guide to Establishing and Evaluating Performance Standards on Tests[M].Thousand Oaks:Sage Publications, 2010.

[11]American Educational Research Association(AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). Standards for Educational and Psychological Testing[M]. Washington, DC: AERA, 2014.

[12]杨志明.高中学业水平考试等级设定的若干方法[J].教育测量与评价,2016(10):4-9.

Three Approaches for Maintaining the Stability of the College Entrance English Test across Multiple Administrations

Yang Zhiming

Abstract: Allowing students to take the College Entrance English Test (CEET)multiple times within a year marks great progress in gaokao reform in China. The inconsistency of the CEET in terms of the form difficulty, however,may lead to big problems in making college admission decisions if using raw scores only. The equating method,which is the typical approach for maintaining form stability in the testing industry,as well as the standardized expert judgment approach derived from the Angoff method of standard setting,and the mini-pilot pretest method of using a small representative sample based on validity evidences are discussed in this paper.

Keywords: test equating, standard setting, test validity, College Entrance English Test

责任编辑/王彩霞

(上接第 14 页)

Educational Assessment: What to Watch in a Rapidly Changing World

Randy E. Bennett

Abstract: This paper is a written adaptation of the Presidential address I gave at the NCME annual conference in April 2018. The paper describes my thoughts on the future of assessment. I discuss eleven likely characteristics of future tests and,for each characteristic,why I think it is important and what to watch with respect to it. Next,I outline what is unlikely to change. The paper concludes with a comment about the probable state of education ten years on and what that state might mean for members of the assessment community.

Keywords: assessment, educational futures, formative assessment, testing

责任编辑/王彩霞