Michael O'Sullivan: speech at

Education Supervision and Evaluation Seminar

Beijing University of Technology, 3 June 2015

(Chinese text)

Setting examinations that are fit for the future

如何设计适合未来的教育考试

“尊敬的各位代表，女士们，先生们：

非常荣幸受邀在今天的研讨会上发言。

我所在的剑桥国际考试委员会是全球唯一仍然隶属于大学的国际性考试机构。作为剑桥大学的一分子，我们拥有800余年历史，面向学生的考试一直是本校工作中不可或缺的一部分。过去150年，我们还发挥了独特的作用，为不同学习科目设计考试体系，以供世界各地的学校使用。

我们如今为160个国家和地区的1万多所学校提供服务，不仅提供考试和得到国际认可的资格证书，还面向学校提供课程大纲、学习资料、教师培训和支持，帮助它们为学生提供适合21世纪的教育。在新加坡等多个国家，我们参与提供公立学校体系中所采用的各种考试。我们的考试也在世界各地提供国际性课程的私立学校得到广泛采用。在埃及和哈萨克斯坦等多个国家，我们帮助政府开发面向21世纪的全新课程和考试。

今天在北京探讨考试可谓恰逢其时。下周，中国又将有数百万年轻人走进高考考场，面对人生中独一无二的机会。如果成功，他们将实现自己以及父母的梦想，进入中国一流的高等学府。

让我提醒一下诸位另一个更具历史性的日子。今年是中国科举考试被废除后的第110年，这个存在了长达1,300年的考试制度曾是中国封建时代精英统治阶层的重要遴选门槛。它也被广泛誉为如今地球上几乎每个国家都存在的现代公共考试制度的原型。

纵观历史，我们能够感受到重大考试制度对文明和社会产生的巨大影响。科举考试制度无疑在向整个华人世界推广共同的文学和哲学文明方面起到

了深远的积极作用。它培育了社会流动、英才教育等概念，以及勤奋学习的重要意义。然而，科举后来在中国饱受批判，不仅因为它缺乏对科学技术的关注，而且在一定程度上成为造就民族灾难的帮凶。

相比之下，现代高考格外强调科学和文化知识。然而随着社会的进步，对高考改革的呼声也越来越高。许多人会问，高考是否足够公平？高考能否真正考验学生亟待掌握的知识？它能否培育具有国际视野的人才？并非所有人对此持肯定态度。

教育评估领域的专业人士有时倾向于认为“考什么就教什么”。这种想法很有吸引力，当您试图推销考试时尤其如此。但毫无疑问，这种想法是错误的。如果正确，那么为了使学生掌握更多知识，只要提高考试的难度就足够了。实际上，这个错误在某些国家的教育部门非常普遍。

根据我的经验，这种想法更加准确的表述是“不考就不教”。因此，我们最好谨慎地设计考试，但在规划教育时，我们不能以考试作为起点。

剑桥认为，学校教育应该被视为课程内容、教学方法和考试评估构成的三角关系。教育的所有改进和改革必须兼顾这三方面的因素，孤立地改变其中某一个因素都将收效甚微。

教育的核心是课程内容，即我们希望学生所学习的内容。这通常表现为某种知识和技能体系，学生不仅要掌握核心知识点，理解不同知识点之间的关联，而且能灵活运用这些知识点，展开相关的计算、分析、批评、解释等活动。关于知识和技能哪一个更重要，最近出现了一些无谓的争论。事实上，我们在教育中所关注的一切事物都是技能和知识的复杂组合。阅读是一项技能，但必须具备语言结构和词语含义方面的知识。科学观察是一项技能，但必须知道如何以正确的方式寻找正确的事物。我们也许还会辩论空气和水哪一种物质对人类生存更加重要，这与探讨知识或技能哪一个在教育中更加重要如出一辙。

剑桥课程的任何科目都非常明确地阐述了所谓的“学习目标”：也就是对于学生在课程各个阶段所需学习的内容的详细描述。考试设计的基础是与学习目标挂钩的“考评目标”。借此，我们努力确保考试对必要的学习给予鼓励、认可和奖励，而非让考试本身成为一种目的。

采用这种方式推进的影响之一就是我们的考试很难评分。举例来说，由于化学课的学习目标包括开展科学实验的能力，因此某些考试必须在实验室中进行，并且通过观察进行评分。另举一例，针对18岁学生的历史课学习目标包括历史解读，所以对学生的考评必须通过论文进行，但需要由训练有素的考官来评分。采用这种方式对大批学生进行考核成本高昂而且困难

重重，同时有关分数更正的申诉必然层出不穷，而且必须公平地加以处理。但是，我们仍然在广泛采用这些方法，因为我们相信，与考试设计中的成本或便利相比，对有用的学习给予鼓励和奖励更加重要。然而，我必须承认，针对中国的高考广泛运用此类方法也许不符合实际。在我看来，高考的规模以及至关重要的公平性，对这项考试的设计提出了诸多限制。

下面，容我说明一下对于考试的另一种常见的批评，该批评认为考试中难以评估的某些内容，例如学生的科学实践能力或深入研究某项主题的能力，对于教师来说却更加易于评估。这种批评通常还包括一个观点，即学生在考试中的表现可能随着考试当天不可预知的原因而发生改变， 而教师却了解学生的真正实力。对于所有此类原因，剑桥认同由教师承担的评估在很多情况下可能比通过考试开展的评估更有效。但是，对于大学入学遴选等某些关系重大的决策，或者对于任何旨在衡量教师和学生表现的评估，我们认为依据教师对学生的考核来建立可靠的评估体系并不现实。

最后，我将简要评述一下教育评估领域目前的两个重大趋势，并且思考它们对于我之前提出的问题有何意义。

首先是对所谓“21世纪技能”的评估。剑桥已经对可以找到的有关21世纪技能的所有文献资料进行了分析解读，并且我们自己也开展了多项研究。我们发现，尽管并非所有此类技能都真正属于21世纪，但包括解决问题和团队合作在内的许多技能无疑都对学生在大学和毕业之后取得成功具有重要意义。因此这样的技能应当给予鼓励，其中一种鼓励方法就是在考试中进行考核，我们许多人也正在为此努力。但必须注意的是，在目前的研究阶段，即使在数学的领域中，人们对“解决问题”的真正含义也没有达成一致意见。“团队合作”的准确含义同样也没有取得广泛的共识。因此，我们应当谨慎地宣称自己有能力对学生的解决问题或协作技能进行评级和打分。务必采取一种实验性的方法。“实验性”不仅意味着先尝试各种事物，还意味着对某一项理论努力证伪，并在经过大量实验表明其可能基本正确后方才予以接受。并非所有教育考试提供机构都能照此行事。

第二个趋势是数字技术在教育评估中越来越普遍的应用。整体上，这是一个非常积极的趋势，为将来更好、更快、更低成本、更灵敏地实施学生评估带来了巨大潜力。例如，计算机适应性测试和虚拟现实都为教育评估开启了新的可能性。不难想象，未来的教育和评估将100%实现数字化，剑桥已经在进行这方面的测试。

但是，如果方式不当，数字化评估也存在某些风险。其中之一是有可能出现高频度、低成本、评分仓促的学生评估。在美国，由于主要依赖越来越频繁的低成本标准化测验，雄心勃勃的“No Child Left Behind”（“有

教无类”）运动演变成一场灾难。这种拔苗助长的方法最终分散精力，打击士气，在教育方面收效甚微。孩子们的学习节奏肯定不会相同，过多地检验这一点毫无意义，同时频繁考核不在教学范围中的内容无疑也毫无意义，这恰好证明了“考什么就教什么”的错误性。

如今借助数字化评估，对部分内容的检验十分简单易行，然而如果数字化评估实施不当，另一种风险则是那些不易借助数字化方式评估的重要技能未能得到足够检验。例如，针对长篇幅书面答案的高品质机器评分仍处于早期开发阶段，并且在多数情况下尚无法应用于重大考试。我们应当注意的是，急于采用仍需进一步研究的数字技术 – 可能降低评估的质量。

最终，优秀的考试设计取决于考试的目的，这一点必须精确清晰地加以确认。有时，同一项考试被应用于过多的不同目的：评估学生，开展遴选，评估教师和学校。一般而言，考试的目的越简单和明确，设计得当的可能性越大。

对于高考改革，我认为高考的目的非常明确，就是中国的大学入学考试，为考试提供支持的各种现有技术也正逐步完善，因此在设计上取得显著进步是完全有可能的。但是，出于多种原因，我认为对高考的“颠覆式”改革也许并不明智。突然的全面变革在整个社会看来必然有失公平。考试发生突然的大规模变革，也将很难确定学生的表现是进步还是退步。并且任何创新举措也都需要得到评估，只有在评估的基础上，这些举措才能进一步完善。我抛砖引玉，提出以下策略供参考：

- 进一步明确学习目标，为考试设计提供支撑
- 加强研究，依据学生们在大学的学术表现，确立考试各组成部分的预测性效度
- 提高可选择性，使学生能参加他们希望在大学学习的相关科目的考试，而不仅是一般性科目的考试
- 逐渐普及数字化评估，以改善评估的灵敏度和效率。

女士们，先生们，预祝今天在北京的重要讨论取得成功！再次感谢给我这次发言的机会。”