新高考改革背景下不同版本试卷测量学指标的比较与监测

——以某年度英语试卷的分析为例*

章建石

摘要　在当前高考改革的背景下，不同版本高考试卷的使用，增加了考生水平、试题质量比较之间的难度，也不利于命题质量的改进和维护高考的公平。本文以英语学科为例，通过等值的有关设计，探索了对不同版本试卷在同一标准下进行难度、区分度等测量学指标分析的方法，为监测与比较不同版本的试卷质量以及提高命题水平，提供了重要参考。

关键词　难度；　区分度；　不同版本试卷；　高考

作者简介　章建石 / 教育部考试中心助理研究员　（北京　100084）

一、对大规模教育考试中试题及其难度的认识

在大规模教育考试中，试题是最基本的测量单元，它依据一定的考试目标和考试内容来创设一定的刺激情境，通过测试来获取学生的作答反应，并且“根据考生的作答反应来对考生的心理特质进行推测”。[1] 试卷是在综合考虑具体测量目标、考试内容、试题难易、题型、测试时间等多种因素的基础上，对试题进行筛选、排序和组合的结果，是试题有机、有序、有效的集合。在高考中，高质量的试卷不仅是科学选拔的前提，更关系到整个高考制度的公平。难度、信度、区分度等衡量试卷质量的测量学指标一直以来受到学生、教师、命题工作者甚至政府部门的广泛关注。因而，在高考结束后对试卷进行相应的统计分析，在整个命题过程中具有举足轻重的地位。在诸多的测量学指标中，试卷难度一直是关注的焦点，是评价命题质量的一个核心指标。在实践中，合理控制难度是命题者面临的一项艰巨任务，很多时候甚至成为行政部门明确提出的政策要求。

不同测量理论对难度的定义和计算方法有着较大差异。在经典测量理论

* 本文为国家教育考试科研规划2017年度课题的研究成果（批准号：GJK2017035）。

(Classical Test Theory,简称 CTT)中,难度实际上是试题的通过率,即难度值为答对某题的被试人数占被试总人数的比例。这里的通过率反映了试题本身和考生能力两方面的属性,或者说通过率至少是包括"考生能力有多高"和"试题本身有多难"这两个因素在内的函数。项目反应理论(Item Response Theory,简称 IRT)认为,试题难度具有一定的独立性,可以与被试能力值在同一个量尺上进行直观的比较。但其定义的难度值同样体现了考生能力和试题本身之间的复杂关系。因为该理论仍旧"将看不见的考生潜在特质与题目自身的属性结合在一起,体现微观层面的信息"。[2] 因而,不管采用哪种方法来对试题难度进行统计分析,其结果并没有对作为试题固有属性的难易情况进行精确的描述。也就是说,在试题还没有与考生发生"反应"之前,试题、试卷本身应当存在一个本体论意义上的"难易程度"。这一点在需要依托大量刺激情境的试题中表现得尤为明显,如在英语的阅读理解中,阅读材料和设问本身的难易程度。

在有关英语测试的研究中,对难度的理解和分析更加微观和细致。Ahmed 等认为试题难度存在于整个作答过程中,并把作答过程分为 5 个步骤,每个步骤都是一个"难度源"。[3] Leong 在 Osterlind 的试题定义基础上,提出了试题难度的四个组成部分,即内容难度(Content Difficulty)、刺激难度(Stimulus Difficulty)、任务难度(Task Difficulty)和预期答案难度(Expected Response Difficulty)。内容难度主要指试题所考查知识内容的难易水平;刺激难度是指"考生在理解单词、短语以及试题包含的其他信息过程中所感知的困难";任务难度体现在考生得出答案的过程中那些需要考生进行复杂推理、演算的问题上,推理越复杂,任务难度就越大;预期答案难度主要指命题者所预设的正确答案的复杂程度。[4] Yasuhiro Ozuru 等人进一步对标准化阅读测试的"难度位置"进行了分析,探讨了阅读材料难度和问题难度之间的关系。[5] 这里的阅读材料难度和问题难度即为 Leong 提出的刺激难度和设问难度。Irene Kostin 对托福考试的听力试题进行了难度方面的分析,结果表明词汇、材料主题和任务要求这三方面因素对难度的影响显著。[6] 其中,词汇、材料主题属于刺激难度,任务要求则属于设问难度。此外,McNamara,Kintsch,Songer & Kintsch;[7] Recht & Leslie;[8] Spilich,Vesonder,Chiesi & Voss[9] 的大量研究表明题材新颖程度、类别、设问方式等都会影响考后统计得出的试题难度值。需要说明的是,以上研究结论都是在母语为英语的教育考试中得出的,这些结论能否适用于把英语作为第二语言进行测试的考试项目中,还有待研究。另外,我国高考英语科命题还受到国家课程标准、教材、教学等多种因素的影响,对考试目标、内容选择、难度等方面均有一些特殊要求。本研究也将重点探讨试题的刺激难度与试题难度、区分度这两个特殊指标之间的关系。

二、作为刺激难度的阅读材料难易及其测评

在 Leong 提出的难度分析框架中,四个类别的难度都会影响试题在考后统

计分析中得出的难度值。结合我国高考英语科命题的现状，上述四个难度实际所发挥的影响存在较大差异。在内容难度上，国家课程标准、考试大纲决定了考试内容的统一性。因此，不同版本高考试卷之间在内容难度上差异不大。任务难度和预期答案难度受命题者的影响较大，因而难以进行定量的比较。相比之下，刺激难度对考生的作答影响较大。在一份试卷中，刺激的主要表现形式为题干，它是试题的主要组成部分，是集中体现考试内容与能力考查目标的载体，是获取考生作答反应最直接的“刺激”。在高考英语科阅读理解的测试中，作为阅读材料的题干，其重要性更是不言而喻。阅读材料的难度，就可以看作是最重要的刺激难度。

对阅读材料的难易水平进行相对客观的评价，一直是语言测试领域研究的重点和难点，在课程、教学改革实践中也受到广泛关注。20 世纪 90 年代以来，美国政府为了提高教育质量，一直致力于制定基础教育阶段统一的课程内容标准。2009 年，美国全国州长协会和各州教育长官委员会联合发起了“共同核心州立标准计划”（the Common Core State Standards Initiative）。2010 年 6 月，“共同核心州立标准”（Common Core State Standards，简称 CCSS）正式颁布，包括《数学标准》、《英语语言艺术和文学标准》。后者有专门章节对如何评价阅读文本难度进行了详细说明，提出文本难度评价的三维模型，分别为文本难度的定性维度、定量维度、读者和任务维度。定性维度指文章主旨大意、写作目的、语言结构、是否常见且清晰等；定量维度指词汇长度或词频、句子长度、文本连贯性等，并特别强调可以用相应的软件来进行分析；读者与任务维度主要指向与阅读文本有关的阅读者的动机、知识、经验以及阅读任务的特点。[10]

如何对阅读材料上述两个内在特性进行直观分析，以了解其难易水平，语言测试专家进行了积极的探索。在早期（20 世纪 40 年代左右），相关研究与阅读本文易读性（Readability）紧密联系。[11] 通过对文本易读性的分析，研究者希望找到影响文本难度的若干因素，并尽量用量化结果来对其进行描述。随后，一系列评价文本难易程度的公式不断出现，其中提出较早也较有影响力的是 Flesch 易读性公式。50 年代以后，影响文本难度的各种因素以及统计模型被不断挖掘。据初步统计，截至 70 年代，各种文本难度的计算公式总数超过了 200 个。[12] 在实践领域，针对文本难度的定性维度，美国的考试机构 ACT 提出了相应的分析指标体系，该体系包括关系、丰富性、结构、风格、词汇和写作目的等多个维度，每个维度给出三个等级的描述，由专业人员来进行主观评价。[13] 对于文本难度之定量维度，包括单词长度、单词频率、句子长度、语篇连贯性等，则可以使用不少专业性的软件进行分析。目前在语言国家比较常用的有蓝思测评（Lexile）、ATOS、Reading Maturity、Source Rater 等，这些软件主要由专业性的考试测评机构开发。值得强调的是，在这些测评方法中，蓝思测评在国际上最具影响力和公信力，在教育教学领域的使用也最为普遍。蓝思测评主要从语义难度（Semantic Difficulty）和语法复杂度（Syntactic

Complexity)[14]两个维度来衡量阅读文本的难度。测评结果是用数值来反映阅读能力的高低和文章的难度水平，测评结果范围通常在200—1 700 L之间，"L"是其独特的难度单位，数值越大表示文章难度越大。近年来，一些大规模教育考试项目已经开始关注蓝思测评在教育考试领域中的研究和实践应用。几年前，托福考试的主办方决定将其阅读分数联入了Lexile测评体系。

三、问题、研究设计与分析方法

(一) 问题与研究设计

研究问题一：我国高考英语试卷中阅读材料的难度处于什么水平？对此，本文选取两份不同的某年高考试卷，分别记为A卷(甲省使用)、B卷(乙省使用)，以及同一年份美国高校入选的两项考试：SAT样题和ACT样题，按照相同的顺序从其中选择均占有较大比重的阅读理解部分来进行分析。

研究问题二：不同阅读材料的综合难度，会对考后统计的试题难度和区分度产生什么影响？降低阅读材料的难度是否会影响必要的难度、区分度？在此需要特别强调的是，作为大规模选拔性考试，在我国现有的招生录取体制下，一份高考试卷需要对不同能力水平的考生进行合理区分，以满足若干批次的录取需要。因此，在命题过程中，保证高考试题必要的难度、区分度是必然要求。但是，即使在保证了必要的难度、区分度之后，由于各省在课程改革推进程度、高考录取率等方面存在较大差异，不同版本试卷的各项测量学指标之间并不能进行直接的比较。简单地说，在难度、区分度等指标上，各省有不同的"省情"。不同版本试卷在考后定量分析得到的结果并没有可比性，因为考生群体不一样，试卷也不一样，考试结果的使用情况(录取率)也不一样。对此，本文采用了等值中常用的"等组设计"，选取两所省属高校新入学的大一学生，综合报考科类、高考成绩、性别、所在院系等因素，将考生分成尽可能平行的两组，分别进行A卷、B卷的测试，同时，计算两份试卷中相关试题、题组的难度和区分度，再与其蓝思分析的结果进行比较。在保证样本群体一致的前提下，来分析阅读材料难度与试题难度、区分度的关系。

(二) 分析方法

针对问题一，使用蓝思测评的专业软件Lexile Analyzer进行分析，得到上述不同试卷阅读材料的蓝思值，即可进行难度值的比较。针对问题二，由于采用了"等组设计"，尽管在分组时考虑了影响群体一致性的各种因素，但两组考生的能力分布很可能还是不一样，有必要比较两组考生的实际能力值以进行必要的验证。为确保各个被试能力估计值在同一量尺上，在进行能力值估计时，采用矩阵抽样的同时估计方法，用Xcalibre 4.2进行分析。在试题难度和区分度的计算上，本文把每篇阅读理解作为一个题组，采用等级反应模型，用

Parscale 4.1 来估计试题参数。

需要特别指出的是，由于题组类试题违反了 IRT 局部独立性假设，在分析过程中一般不宜直接使用传统的 IRT 模型，否则，对被试能力和试题参数的估计都会存在偏差。[15] 对此，Sireci 等人提出可以将题组看成一个多级评分的超大项目（Super Item），并且可用多级评分模型来估计试题参数。[16] 在本次分析中，每一篇阅读理解有 n 个选择题，均为二元计分，考生答对 1 题得 1/n 分，这样考生在这篇阅读理解上的得分区间为 0 到 n 分，通过这一方式，比较方便地实现了将由若干项目组成的题组向一个超大项目的转换，随即就可采用传统常用的 IRT 模型来分析。

四、结果与分析

（一）等组设计中“平行”假设的检验

测试结束后，对两组考生在 A、B 卷上的得分进行了统计，结果表明，两组考生的卷面成绩、标准差均存在较大的差异。进一步对考生的能力值进行估计，结果见表 1。

表 1　考生能力值比较

	样本量	平均值	标准差	偏度	*Q*1	*Q*2	*Q*3	*IQR*
A 卷	2 335	0.01	0.96	−0.57	−0.57	0.11	0.71	1.28
B 卷	1 873	−0.01	0.98	−0.31	−0.69	0.04	0.76	1.45

从表 1 可见：两组考生的能力值相差很小，仅为 0.02，符合考生能力“平行”的假设。同时可以推论：因为两组考生的能力值差异很小，而两份试卷的观察分数相差较大，所以 A、B 两卷试题的难度应当存在较大差异。

（二）A 卷、B 卷的全样本难度、区分度

根据经典测量理论，以甲、乙两省的全体考生的得分来计算两份试卷的难度值和区分度，A 卷的难度为 0.51，区分度为 0.37，B 卷的难度为 0.69，区分度为 0.46。在大规模考试中，如果试题的难度处于 0.3 与 0.7 之间，区分度大于 0.3 的话，则可以认为试卷具有较高质量。应当说，从全样本的统计指标来看，尽管 A 卷、B 卷的难度、区分度存在一定差异，但都在合理范围之内。实际上，由于试题、考试群体不一样，这里的统计结果只能说明两份试卷都满足了测量学的基本要求，数值上的差异并不足以用来评价试卷的优劣。

（三）不同试卷阅读材料难度值的比较

蓝思分析的结果表明，A 卷四篇文章的蓝思难度值分别为 1 200 L、1 300 L、1 010 L 和 1 550 L，B 卷的值分别为 800 L、1 070 L、950 L 和 1 260 L，ACT 样题的值分别为 1 120 L、1 460 L、1 310 L 和 1 250 L，SAT1 样题的值分别为 1 350 L、1 240 L、1 530 L 和 1 490 L（见图 1）。不同试卷的蓝思值区间见图 2。

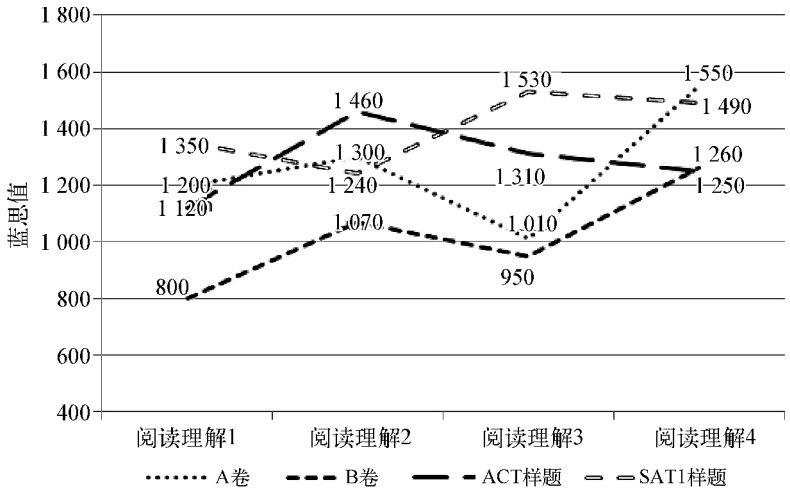

图 1　不同试卷的蓝思难度值

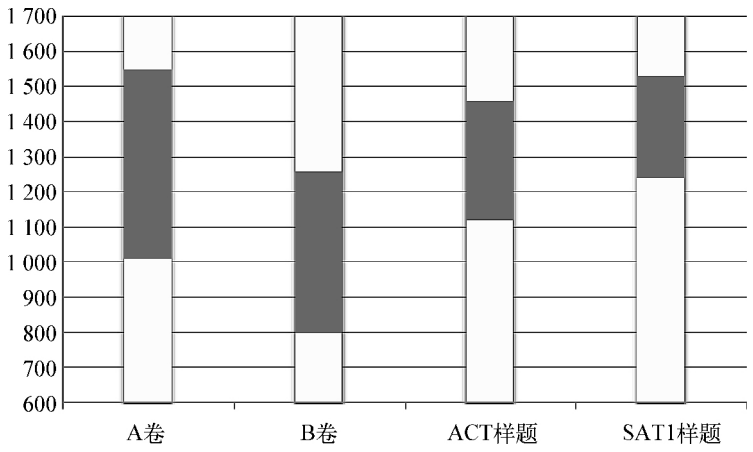

图 2　不同试卷的蓝思难度值区间

从图 1、图 2 可以看出，A 卷阅读理解材料的平均难度比 B 卷高出 265 L，即 A 卷阅读理解材料的难度明显高于 B 卷。就四份试卷的平均难度而言，SAT1 样题的平均难度最大，B 卷的平均难度最小，A 卷的难度与 ACT 样题相当。对于非母语国家的英语测试来说，A 卷在阅读理解的材料明显偏难。再者，ACT 样题和 SAT1 阅读理解材料的难度波动范围较小且相对比较稳定，而 A 卷、B 卷阅读理解材料在难度上存在较大的波动。

为进一步了解我国高考英语试题阅读材料难度，我们对甲、乙两省连续三年英语试卷的阅读理解进行了蓝思分析，结果表明，甲省所用试卷连续三年的阅读理解材料的平均蓝思值分别为 1 060 L、1 275 L 和 1 265 L，三年的总平均值为 1 200 L，乙省所用试卷三年阅读理解材料中每年的平均蓝思值分别为 1 157 L、1 097 L 和 940 L，三年的总平均值约为 1 060 L。由此可见，整体而言，

甲省试卷的阅读理解材料难度要高于乙省试卷。另外，就单个省份年度之间的比较来看，甲省试卷的阅读理解材料难度在近三年有上升的趋势，而乙省试卷则在逐年下降，变化均比较明显。

（四）阅读材料难度与试题难度的比较

由上述分析可知，两组考生的能力值基本一致，可以看作是同一个考生群体。因此，考生参加A卷、B卷测试后各自计算出的难度值，就可以直接依据数值大小来进行比较，并据此进行优劣判断。结合两卷阅读理解材料的难度值，可以得到图3，其中，蓝思值在左侧纵轴上标示，试题的难度值在右侧纵轴上标示。

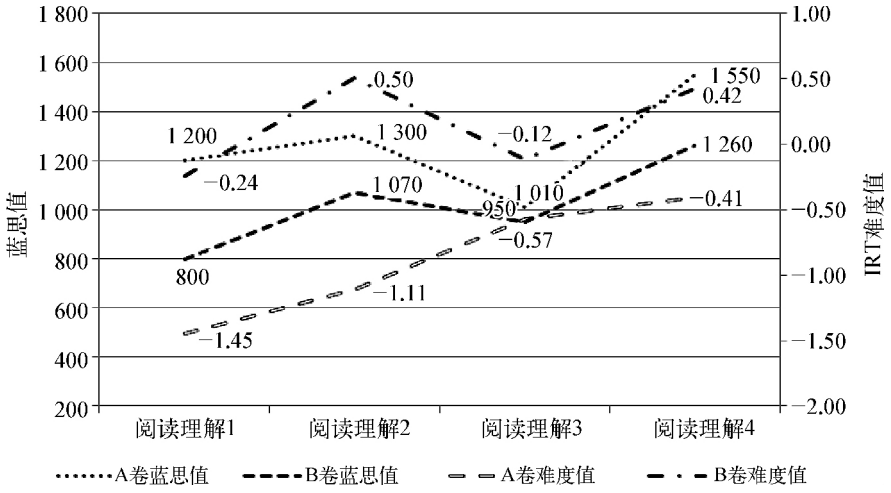

图3　A卷、B卷蓝思值和难度值比较

从图3中可以看出，A卷阅读理解的难度值分别为−1.45、−1.11、−0.57和−0.41，B卷阅读理解的难度值分别为−0.24、0.5、−0.12和0.42，显然，B卷阅读理解试题难度明显高于A卷。而就阅读理解阅读材料的难度来看，A卷则明显高于B卷。可以看出，A卷、B卷的阅读理解分别呈现出“高阅读材料难度，低试题难度”与“低阅读材料难度，高试题难度”的特点。相比较而言，在阅读理解试题的命制上，B卷选择了相对容易的阅读材料，但保证了较高水平的难度。A卷则相反，阅读材料较难，难度却较低。可以进一步推论：降低阅读材料难度，并不会以牺牲试题难度为代价，或者说，阅读理解试题的命制，在降低阅读材料难度的同时保证试题必要难度的做法是可行的。

（五）阅读材料难度与试题区分度的比较

同样，可以对阅读材料的难度与区分度进行比较。这里的区分度也是将每一篇阅读理解作为题组进行计算得出的，可以直接进行比较。结果见图4，其中，左侧纵轴表示蓝思值，右侧纵轴表示试题的区分度值。

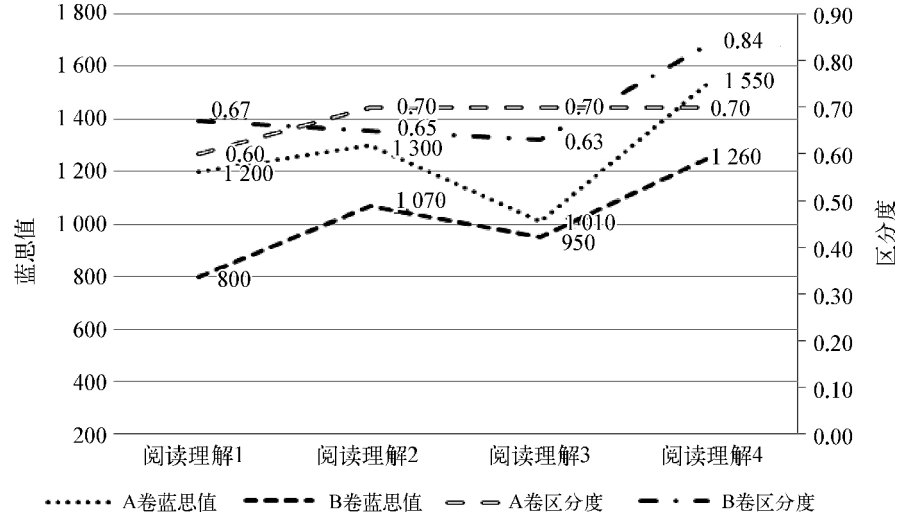

图 4　A 卷、B 卷蓝思值和区分度值比较

从图 4 中可以看出，A 卷四篇阅读理解的区分度分别为 0.60、0.70、0.70 和 0.70，B 卷四篇阅读理解的区分度分别为 0.67、0.65、0.63 和 0.84。在大规模教育考试中，试题的 IRT 区分度值一般在 0 到 2.50 之间，超过 0.50，就可认为比较理想了。按此标准，A、B 卷阅读理解的区分度是合适的，结合两卷阅读材料难度之间存在的差异，两卷的阅读理解呈现出“区分度均合理，阅读材料难度差异大”的特点。进一步可以推论：在保证阅读理解试题良好区分度的前提下，降低阅读材料难度的做法同样是可行的。

五、结论及讨论

通过与 SAT、ACT 阅读材料难度的对比可以看出，我国高考英语阅读理解试题所选取的阅读材料偏难，具体表现为阅读材料的句式较长、语法较难、生僻词较多。在以往经验中，刺激难度的提升有利于提高试题难度、区分度，但不是唯一的方式。考试内容的深浅、设问水平的高低、答案要求的繁简都是可以采纳的方法。另外，从测量理论层面来看，依靠增加阅读材料难度来提高试题难度、区分度的方式也不可取。在用大规模考试来对个体潜在特质进行测量时，一个重要的前提是：尽可能使个体是在某种潜在特质的支配下来进行作答，这样才能保证测量结果的准确性和精度。如果试题的刺激难度过大，被试可能无法做出反应或做出与刺激完全无关的反应，进而不能进行测量意义上的“行为抽样”，进而导致对这些“反应”的分析产生偏差，从而出现“测非所需”的结果，这既不符合科学测量的要求，也不利于公平。这一点对于潜在特质水平相对较低考生来说尤为明显。实际上，这个群体的考生很可能在进入

测量活动之前，就被较高的刺激难度拒之门外了。在这种情形下得出的测量结果，对他们来讲是仅仅是一个与测量目标没有关系的度量值而已。因而，对高考各科的命题来讲，需要迫切关注在确定的测量目标下，刺激难度对测量结果的影响。在命题实践中，试题素材的选择、试题情境的创设等尤其需要慎重，对试题质量的监测也需要密切关注与此相关的指标，确保考试对所有考生的公平、公正。

测试结果的使用方式会对测试工具的属性提出一定的外在要求，试题难度、区分度就是很好的例子。这一点在我国高考中体现得尤为明显，在世界教育考试行业中也颇具特色。虽然降低阅读材料难度的做法更为科学和公平，但应以保证现阶段国情所决定的试卷难度、区分度水平为前提。对此，本研究表明，在保证必要且合理的试卷难度、区分度的前提下，阅读材料难度可以降低。给定相同的阅读材料，即刺激难度相同，设问难度、预期答案难度，乃至考后统计出的包括试题信度、效度、难度、区分度在内的多个测量指标，都可能呈现出较大的差异。而上述所有指标所指，阅读材料的选择、设问的水平、答案的要求、对考生能力和试题测试学指标的预估，最终都体现在命题质量上。

当前，我国高考改革进入了一个新时期，多数省份将逐步恢复使用全国卷，但全国卷也分为不同的卷种，以适应不同省份的教育发展状况和课程改革的进度。工欲善其事，必先利其器。作为人才选拔的工具，不管试卷有几个版本，都需要一些共同的约束和技术规范，其中，测量学的指标就是重要的参数，并且是命题质量改进的重要依据。目前，对试题、试卷的分析基本上针对某个具体版本的试卷，并没有考虑不同版本试卷之间的关系，这使得在命题环节预设不同版本试卷的难度、区分度的水平以及不同省份在选择哪个版本试卷上，均缺乏科学的证据。本文仅以英语学科为例做了相应的分析，有一些技术和方法是各学科均可参考的，如不同版本试卷比较所需的等值设计以及具体的分析方法。当然，更多的分析及其结果解读还需要结合学科的特点。

参考文献：

[1] Steven J. Osterlind. Constructing Test Items: Multiple-choice, Constructed-response, Performance, and Other Formats [M]. New York: Kluwer Academic Publishers, 1998: 2.

[2] 冯艳宾，马洪超. 关于经典测量理论和项目反应理论中难度和区分度的探讨 [J]. 中国考试，2012(4): 10-14.

[3] Ahmed, A. & Pollitt. Curriculum Demands and Question Difficulty [C]. Slovenia: IAEA Conference paper, 1999.

[4] Leong, S. C. On Varying the Difficult of Test Items [C]. Singapore: IAEA Conference paper, 2006.

[5] Yasuhiro Ozuru, Michael Rowe, Tenaha O'reilly & Danielle S. Mcnamara. Where's the Difficulty in Standardized Reading Tests: The Passage or the Question? [J]. Behavior Research Methods, 2008(4): 1001-1015.

[6] Irene Kostin. Exploring Item Characteristics that are Related to the Difficulty of TOEFL Dialogue Items [R]. New Jersey: ETS, RR-79, 2004: 7.
[7] McNamara, D., Kintsch, E., Songer, N. & Kintsch, W. Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text [J]. Cognition and Instruction, 1996(1): 1-43.
[8] Recht, D. & Leslie, L. Effect of Prior Knowledge on Good and Poor Readers' Memory of Text [J]. Journal of Educational Psychology, 1988(1): 16-20.
[9] Spilich, G., Vesonder, G., Chiesi, H. & Voss, J. Text Processing of Domain Related Information for Individuals with High and Low Domain Knowledge [J]. Journal of Verbal Learning and Verbal Behavior, 1979(3): 275-290.
[10] Common Core State Standards Initivative. Common Core State Standards for English Language Arts & Literacy in History, Social Studies, Science, and Technical Subjects [S]. Appendix A: Research Supporting Key Elements of the Standards, Glossary of Key Terms, 2011: 4.
[11] Kevyn, C. T. & Jamie Callan. A Language Modeling Approach to Predicting Reading Difficulty Proceedings [C]. Boston: the HLT/NAACL 2004 Conference, 2004: 193-200.
[12] John, J. P. Readability [M]. Boston: Houghton Mifflin, 2002: 2.
[13] ACT, Inc. Reading between the Lines: What the ACT Reveals about College Readiness in Reading [R]. Iowa City, 2006(2): 63.
[14] Stenner, A. J., Horabin, I., Smith, D. R. & Smith. The Lexile Framework [M]. Durham, NC: Metametrics, 1988.
[15] Yen, W. Scaling Performance Assessments: Strategies for Managing Local Item Dependence [J]. Journal of Educational Measurement, 1993(30): 187-213.
[16] Sireci, S. G., Wainer, H. & Thissen, D. On the Reliability of Testlet-Based Tests [J]. Journal of Educational Measurement, 1991(28): 237-247.

The Comparison and Monitoring of the Measurement Indicators in Different Versions of Tests in the Context of the New Reform of Gaokao

ZHANG Jianshi
(The National Education Examinations Authority, Beijing, 100084, China)

Abstract: Currently, different versions of tests of Gaokao are in use. It is difficult to compare the ability of different testees and the item quality of different tests. The paper explores a set of methods to analyze the measurement indicators such as difficulty, discrimination under the same criterion by equating design. By doing so, it enables the comparison of quality between different versions of tests and provides concrete reference to item development.

Key Words: Difficulty; Discrimination; Different versions of testing papers; Gaokao

(责任校对: 杨秀秀)