标准化考试简介

国家教育委员会学生管理司主编

高等教育出版社

标准化考试简介

国家教育委员会学生管理司主编

高等教育出版社

标准化考试简介

国家教育委员会学生管理司主编

*

高等教育出版社出版
新华书店北京发行所发行
北京印刷一厂印装

*

开本 787×1092 1/32 印张 4.375 字数 93,000
1985年10月第1版 1986年3月第2次印刷
印数20,201—70,200
书号7010·0661 定价0.94元

本　书　说　明

标准化考试是近几十年来国际上广为流行的考试方法，也是我国高考准备采取的一种形式。这本小册子概略地阐明了什么是标准化考试以及如何实施标准化考试；围绕标准化考试问题较为系统地介绍了教育测量学的基本知识，同时对我国近年在英语水平考试（EPT）中推行标准化的尝试，做了简要介绍。本书文字通俗易懂，深入浅出，适合于广大教师与教育行政干部阅读。对于各种考试的命题人员与组织者，对于从事教育测量与教育评价工作，以及有志于改革旧的考试方法的同志，均有一定的参考价值。

本书第一部分是请北京师范大学心理学系主任张厚粲同志编写、郑日昌和冯柏麟同志参加编写的；第二部分是请广州外国语学院院长桂诗春同志编写的。

国家教育委员会学生管理司

目　　录

第一部分

标准化考试的几个基本问题

一、什么是标准化考试

标准化考试是国际上广为流行的考试方法。近年来，“标准化”一词已逐渐为教育界人士所熟悉，有关部门开始研究推广标准化考试，并在局部地区、个别学科开展了试点工作。

什么是标准化考试？为什么考试要标准化？怎样使考试标准化？这是广大教师、学生乃至家长所关心的问题。本文试图对这些问题做一简略回答。

从物理测量谈起

在中国古代，人们把手掌伸开后的拇指与中指尖端之间的距离（俗称一拃）叫做一尺，较长的距离则用步来丈量。由于人手的大小、腿的长短各异，因此对测量结果无法比较。于是人们便采用一个固定的距离作为一尺的长度。现在为了国际交往的方便，我们在大多数情况下以米（俗称公尺）代替了市尺，这样在度量长度时便有了相同的尺度。这种统一度量衡的工作就是量具的标准化。

长度、重量都有绝对零点，对于它们的测量只要统一单位就可以了。而高度等却没有绝对零点，对于它们的测量除

了要有统一的单位外，还要有统一的参照点，也就是要确定一个相对零点。如测量地面高度以海平面为零点。

规定了统一的单位和参照点，並不能保证每个量具都符合此规定。譬如个别唯利是图的小贩在秤砣或秤盘上搞鬼，实际上是把绝对零点变成相对零点，通过改变参照点造成恒定误差以谋取不义之财，因此市场管理人员有必要对量具进行检查核准。

即使量具完全符合规定，在使用时不按严格程序操作，也会带来测量误差(多半为随机误差)。譬如售货时秤杆高低、秤砣线压在星里星外，或看错、算错等。

上述事例说明，在物理测量方面，为了便于比较和交流，为了减少误差，既需要量具本身的标准化（统一单位和参照点），也需要量具使用方法的标准化（统一操作程序）。

考试虽然是对心理现象（学习结果）的测量，却与物理测量有着同样的问题。

某生期末考试语文得80分，算术得85分，我们很难据此判断该生哪门课学得更好，因为很可能他的语文成绩在班内名列榜首，而算术却坐了末把交椅。即便两位学生在同一门课上得了同样的分数，如果他们来自不同学校，参加的是不同的考试，其成绩也很难比较，因为两个考试的难度很可能不同。

要对不同的考试加以比较，必须使它们具有相同的单位和参照点。

表面看来，各种考试分数都是以 1 分作单位，但以100为满分与以 120 为满分的考卷，其分值显然不等。有些老师在考试中用附加题来给学生加分，但在成绩册上並没注明其满

分分数，这就容易使人产生错觉，仍把它看作是百分制的分数。即使两个考试都采用百分制，由于难度不同，其分值也是不等的。

人的绝大多数心理现象是没有绝对零点的。即使一个学生数学考试得了零分，我们也不能说他的数学知识为零。可见，考试分数的零点是相对的，是教师根据需要通过调整试题难度确定的。这样，不同的考试因为难度不同便有不同水平的零点。这种因参照点不同使不同考试分数无法比较的现象是显而易见的。

把不同考试的原始分数直接相加，依据总分决定取舍，是人们常犯的错误。实际上，不同的考试分数，因为参照点和单位不同，是不具可加性的。例如，1984年高考，数学试题较难，考生的分数普遍偏低，这样数学在总分中占的比例便相对小些，这对择优录取是很不利的。

考试的原始分数非但不能直接相加，也不能直接加以解释。例如，某生的成绩册上报告期末考试的数学成绩为85分，作为孤立的分数，它几乎没有任何意义。据此我们既无法得知与其他科相比该生数学学得如何，也无法知道与其他同学相比他的成绩高低，更无法知道该生对数学教材掌握了多少，哪些已经掌握，哪些没有掌握。在报告和解释考试结果时人们常把分数绝对化，把分数的微小差异看作是有意义的，甚至以1分之差决定取舍。实际上有些分数差异是由误差因素造成的，把它们当作有意义的差异来解释是不合适的。

上述例证，不但说明了考试标准化的必要，也说明了心理测量的标准化比物理测量的标准化更加复杂，更加困难。

考试的误差来源

任何一种分数变异，只要与考试目的无关並使得结果不准确，便可认为是一种误差因素。考试中常见的误差来源于三个方面：试卷内部、考试过程、考生本身。

试卷内部的误差主要来源于题目取样。当取样缺乏代表性，试题偏于某一方向时，不但不能对学生作出全面考察，达不到测量目的，而且会把教学引向歧途；当试题数量太少，考生可能碰巧准备到或没准备到某题，考试成绩受机遇影响较大。除题目取样不当可引起误差外，其他一些因素，如题目用词的模棱两可，对解题要求说得不清，题目过难引起猜测，时限太短使考生仓促作答等，也都可成为误差的来源。

与考试过程有关的误差因素主要来源于以下几方面：1)物理环境。考场的光线、温度、通风、桌面好坏、空间阔窄以及周围安静与否等均会对考生发生影响。2)主试者方面。主考与监考人员的年龄、性别、言谈举止、表情动作等均能影响考试结果。倘若不按规定实施，如故意制造紧张气氛，给予特别协助或暗示，以及计时错误等都会带来较大误差。3)意外干扰。在考试情境复杂，特别是当考生人数较多时，容易发生出乎意料的干扰或分心事件。例如停电，有人迟到、生病、作弊、计时表停了、临时发现个别试卷印刷不清或装订错误等，无论哪种情况都会引起不安和扰乱，导致成绩不准确、不可比。4)评分计分。评分不客观以及合成分数、登记分数出错也是常见的误差。特别是论述题，由于评分标准难以掌握，加之阅卷者的身份、偏好、态度、情绪等因素的影响，评分误差几乎是不可避免的。1983年，我们将五份高考语文

考卷复制后请全国各省市、自治区阅卷组分头评阅，结果有四份考卷不同阅卷组评分相差30分左右，最高达33分。象高考这样重大严肃的考试尚且有如此大的评分误差，平时考试的误差就可想而知了。

来自考生本身的误差因素既有心理方面的，也有生理方面的。1)考生对考试的动机不同，会影响其注意力、持久力、作答态度、反应速度等，从而影响考试成绩。2)临考前或考试中的焦虑（一种紧张的不愉快的情绪体验）也会影响考生的成绩。过高的焦虑会使工作能力降低，注意力分散、思维变得狭窄、刻板，记忆中储存的东西提取不出来。但一点焦虑都没有，也不是好事，内驱力过小的考生往往采取满不在乎的态度，因而成绩大多较低。只有适度的焦虑会使人的兴奋性提高，注意力增强，提高反应速度，从而对考试成绩产生积极影响。3)学生的应考经验也会影响成绩。有些人经历过多次考试，发展了应付考试的技能，他们在觉察正确答案与错误答案的细微差别，合理分配时间，以及适应新的试题形式等方面具有丰富的经验，因此常比那些能力相差不多、但缺乏考试经验和技巧的人获得更高的分数。4)生病、疲劳、失眠等生理因素也会影响考试成绩而带来误差。

心理测量的复杂性，决定了误差来源的多样性。考试的标准化，就是为了控制这些因素，以减少误差。

考试标准化的含义

国际上提倡标准化考试已有几十年的历史，但对什么是标准化考试却至今没有一个严格的科学定义。目前在国内对标准化考试有种种误解，有人认为标准化考试就是

由专门的测验机构编制並组织实施的考试（区别于教师自编的测验）；有人认为标准化考试就是采用选择题等能客观评分的题目进行的考试（区别于不能客观评分的问答题、论述题考试）；也有人认为标准化考试就是采用统计学上的标准分来记分、合分的考试；还有人认为标准化考试就是用标准参照点（常模）来解释分数的考试等等。上述看法均不全面，也就是说，以上几点都不是标准化考试的充分条件，有的甚至不是必要条件。

我们认为，标准化考试是按照系统的科学程序组织、具有统一的标准、並对误差作了严格控制的考试。

此定义的三层意思是紧密联系的。只有按照一套严格的科学程序来组织考试，才能有统一的比较标准（即相同的单位和参照点），才能最大限度地减少误差，使测量尽可能准确、可靠。

考试是一个系统的过程，每个环节都可能带来误差，因此对考试的每个环节都要标准化。具体包括试题编制的标准化，施测过程的标准化，评分记分的标准化，分数合成的标准化以及分数解释的标准化等。

标准化考试的试题是由有关专家根据一定的教育目标集体编制的。这些专家不但要精通本学科的知识，还要受过心理与教育测量学方面的训练。在编题前要制定编题计划，以保证题目对知识和能力两个维度均具有代表性。对于征集来的题目也要经过学科专家与测量专家审查修订。所有题目都要经过预测和统计分析，取得难度、区分度等有关资料，只有经过实践检验各方面符合要求的题目才能存入题库备用。在拚配试卷时，题目的难易和排列顺序要得当，以符合学生

的心理特点。标准化考试的题目一般都有多套等值的复本。

标准化考试对实施手续与施测条件的控制是很严格的。考试手册中对考场设置、收发试卷手续、对考试的说明、主试者的态度、考生注意事项、以及如何计时、意外事件如何处置等均有明确规定，任何人不得随意改变。总的要求是，无论什么人、在什么时候、什么地点使用同一测验，都必须做同样的事，说同样的话，以保证施测条件与实施手续客观化，避免环境与各种偶然因素对考试成绩的影响。

标准化考试大多采用选择题等能客观记分的题目，此种题目记分比较简单。只要把考生的答卷直接送入阅卷机就可为每人评出分数。在必须采用论文式题目时，则要定出详细的评分准则，并且采用一些评分技巧，以克服各种无关效应，使评分尽可能准确客观。

一次考试可能包括几个部分或几个学科，有时为了做出一个判断或决策，我们还要把考试分数与其他资料（如平时成绩、作业、论文等）结合起来使用，这就会遇到如何合成分数的问题。是用原始分数直接合成还是转化为具有相等单位的标准分数后再合成？对不同变量做等量加权还是差异加权？采用不同的合成方法，结果不尽相同，据此所做判断的准确性，决策的有效性当然也不会相同。因此标准化考试对合成分数的方法均有明确规定，尽可能采用最有效的方法。

从考试中直接得到的原始分数要转化到一个有确定参照点（相对零点）和单位的量表上去才有意义，这种转化后的分数叫导出分数。导出分数有两种，一种叫常模参照分数，另一种叫标准参照分数。前者是把个人的分数与其他人的比较，以所在团体的平均分数（即常模）作参照点，根据个人

分数距平均分数的远近确定个人在团体中所处的位置。后者是把个人分数与教育目标比较，以某种可接受的最低标准作参照点，看一个人的成绩是否达到标准或对某一个指定范围的内容或技能掌握了多少，换句话说，是看一个人知道什么，能做什么。标准化考试在向当事人（教师、学生、家长、教育行政部门及用人单位等）报告分数时都采用导出分数，而且把每人的分数看作是一个区间，而不是一个确切的点，並指出落在某一个区间的可能性有多大。在解释分数时，还考虑到对当事人可能产生的心理影响，並尽可能提供指导与帮助。

标准化考试的种类

标准化考试种类繁多，名称各异，从不同角度可做出不同分类。

从考试性质上分有成就考试与能力倾向考试两类。成就指的是经过一定的教育或训练后所学到的东西，是在一个比较明确的、相对限定的范围内的学习结果。成就考试又可分为单科成就考试与综合成就考试两种。能力倾向指的是学习的能力，是在给予适当的机会时获得某种知识或技能的能力，此种能力是在一定的遗传素质基础上生活中各种经验累积的结果。美国大学招生用的学能测验即属后一种考试。

从考试要求上分有难度考试与速度考试。前者包含各种不同难度的题目，由易到难排列，其中有些题目几乎所有考生都解答不了，但作答时间较为充裕，因此测量的是解答难题的最高能力。后者题目较为容易，但数量多並严格限制时间，主要测量反应速度。多数考试是以上两者的结合。

从考试材料上分有文字考试与非文字考试。前者所用的是文字材料，考生用文字作答。后者所用的材料是图形、实物等，无需使用文字作答。

从考试对象上分有个别考试与团体考试。前者一次只能考一人，后者可同时考许多人。

从考试时机上分有进展性考试与总结性考试。前者在教学进行中实施，后者在教学结束后实施。

从解释分数的方法上分有参照常模的考试与参照标准的考试。前者是把考试分数与常模作比较，后者是把考试分数与某种标准作比较。

从功能上看，标准化考试可以有许多种，常见的有：1）选拔考试(例如大学招生)；2)安置考试（例如按能力和知识水平分班或分配工种)；3）准备性考试（主要用来测量考生对于完成某项学习或工作任务是否作好了准备，亦即是否具有所要求的最低能力，是用作预测的水平考试)；4）诊断性考试(主要用来确定学生学习困难之所在)；5）证书考试(是总结性的水平考试，主要用于对正规或非正规的学历给予承认，或用于为某种职业者发执照)；6）用作研究工具的考试(国际教育成就评价协会举办的一系列标准化考试即属此种)。

对标准化考试的要求

好的标准化考试应具备以下三个特性：

1．可靠性

可靠性又称作信度，指的是考试分数的稳定性与一致性。对同一群学生实施同一个考试，在考生知识、能力水平未变的情况下，多次考试结果应该稳定、一致，否则便不可信。

标准化考试都要求有较高的信度，并采用一定的统计指标加以估计。

2．有效性

有效性又称作效度，指的是一次考试是否测到了所要测的东西，是否达到了测量目的。一个考试对它所要测量的东西测量得越正确，便越有效。标准化考试必须有较高的效度，这是衡量考试质量的主要指标。

3．实用性

实用性指考试是否易于实施，是否省时，是否易于评分，分数是否容易解释，是否有复本可用，以及是否省钱等。标准化考试在保证有效可信的前提下，还应具有方便实用的特性。

对标准化考试的评价

推行标准化考试有四点好处：1）减少无关因素对考试目的的影响，使测量准确可靠。2）使不同的考试分数具有可比性。3）同一套测验有多个复本可以反复使用，较为经济。4）可用来校准其他考试。

有人对标准化考试怀有疑虑，认为这种考试会成为教学的指挥棒，使教师教测验，学生学测验。考试只要用作为选拔学生和评价教学的工具，考试的指挥棒作用就是不可避免的，古今中外，概莫能外，我国的高考在这个问题上表现得尤为突出。考试对教学的指挥棒作用是客观存在的，问题是如何使它指挥得当，不使教育工作误入歧途。一般说来，标准化考试是由有关专家采用科学方法集体编制的，内容取样具有代表性，重点、难点掌握得当，而且题目数量较多，因

此能更准确，更全面地反映教学大纲和教材的要求，用这种考试来评价和指挥教学当然比其他考试更有价值。

标准化是相对的，只有水平高低之分，而不是“全有”或“全无”的性质。没有一种考试做到了完全的标准化，也没有一种考试一点没有标准化，以我国目前的高考来说，通过多年实践，我们已经取得了丰富的经验，从命题、施测、评分到录取，有一套较为严格的组织管理办法，对各种影响测验分数的误差因素做了一定控制。但总的看来，我国高考的标准化水平还比较低，在考试科学化方面还有许多的工作要做。

标准化是手段，不是目的。不同的考试目的，对标准化水平有不同要求。一般说来，考试越重要，规模越大，对标准化要求越高。因为标准化测验编制困难，实施费钱，因此平时课堂考试只要借鉴标准化的主要原则和思想，尽可能减少测量误差就行了，对于编制和实施程序不必过于拘泥。

标准化考试不是万能的，认为只要实现了标准化，考试中的一切问题就都解决了的想法是不切实际的。标准化水平再高，也不可能把所有误差都排除。为此我们需要多次考试，多种考试，而不能只凭一次考试、一种考试定终身。要把考试分数与从其他方面来的信息结合起来，只有这样才能对一个学生、一位教师、一所学校或一个地区的情况做出较为全面、正确的评价。

标准化考试是在工业社会中随着教育的标准化（学校教育、班级授课等）应运而生的，这是历史的进步。《第三次浪潮》一书的作者托夫勒认为，“第二次浪潮最为人们所熟悉的原则，就是标准化”。在这本畅销世界的书中他指出：在工业

社会中，“不仅劳动逐渐标准化，而且雇用办法也不断地标准化了。为了准备青年进入劳动力市场，教育家设计了标准化的课程，标准化的智力测验，学校升级原则、入学条件、学分计算也都标准化了”。但作者同时又指出，第二次浪潮的所有原则“条条都受到了第三次浪潮的冲击”。标准化考试是一定历史时期的产物。随着新的技术革命的到来，计算机的广泛应用，出现了由集中化教学转向分散式教学的趋势。坐在计算机前学习的学生，教学与考试已经溶为一体。近年来，在某些发达国家出现了机器教学与变通式测验。计算机对不同学生进行不同水平的教学，实施不同难度的考试。具体方法是：显示出一个题目由被测者作答，如果答对了则下一题便更难些，反之则根据错误类型，在相应方面显示一个较为容易的题目。此种变通式测验每个人接受的是不同的题目，这与量具要统一这个标准化的重要原则是相悖的。但这种非标准化恰恰是这种考试的长处，能更为准确地反映每个人的真实情况，它与那种用一套固定模式衡量所有人的作法又是一个进步。但目前我们离那一步尚有不小距离，那是远景和发展趋势，眼下我们仍有必要推行标准化测验。

二、标准化试卷的编制程序

标准化考试一般由专门的考试机构编制试卷并组织实施。为了保证考试的独立性和专业性，世界上许多国家和地区都有这种机构，如美国的教育测验中心、英国的伦敦职业考试中心、日本的大学入学考试国家中心以及香港考试局等。其中美国教育测验中心是世界上最大的考试机构，成立于1948