中国语文教育丛书

刘国正　顾黄初　章　熊　主编

中国当代写作与阅读测试

章　熊　著

四川教育出版社

中国语文教育丛书
刘国正　顾黄初　章　熊　主编

中国当代写作与阅读测试

章　熊　著

四川教育出版社
1995 年 · 成都

(川)新登字 005 号

责任编辑：吴晓桐　王积跃
封面设计：何一兵

中国当代写作与阅读测试　　章熊　著

四川教育出版社出版
（成都盐道街三号）
四川教育出版社发行　　四川省华印刷厂印刷

开本 8.0×1168毫米　1/32　印张 15　插页 5　字数 329 千
1995年 4 月第一版　　1995年 4 月第一次印刷
印数：1—2500 册

ISBN7—5408—2615—0/G・2512　　定价：12.50元

第一章 语文测试的一般概念

第一节 有关的测试学概念

第二节 语文测试的特点

在教育测量学中，“测验”、“考试”、“评价”等概念既有联系又有区别，而这些概念的内涵和教师们的通常理解又常常有所不同。例如在一些书里，“考试”指的是由任课教师依据自己的经验命题和评分，主观性较强；“测验”则对命题的过程、测试的实施、评分以及分数的解释都有严格的规定，有比较强的客观性。可是在人们的心目中和现在学校的常用语汇中，却把比较正式的考查称为“考试”，而把非正式的考查称为“测验”。这本书的目的不是进行一般的教育测量学的讨论，而是着眼于语文教育测量中的一些实际问题，着眼于实用性，所以不打算在这些概念问题上花费过多的篇幅，而取名为“写作和阅读能力的测试”。

在各文化学科的测试中，写作和阅读的测试可以说是最困难的，因为它所测量的对象是学生的精神产品，而且要作出综合评价。如果说“测验”和“考试”一般只是对学生知识水平的数量化，那么，“评价”则还要作出价值判断，它体现了定量分析和定性分析的结合，体现了教育测量的更高层次。上面说到的“综合性评价”指的是评价的因素很多，这些因素在整体中的作用并不相等，而且大都带有评估者的主观因素，因此问题是相当复杂的。甚至可以说，在写作和阅读能力的测试中，主观随意性是难以完全避免的。这些，读者在以后几章中将有所体会。

长期以来，我们的语文教学与测试可以说都是属于经验型的。在这些经验中蕴含着丰富的宝藏，因为我国可以说是世界上母语书面语教学历史最悠久、经验最丰富的国家。但是，经验型的教学与测试的最大弱点就是它的

随意性。现在，当我们进行理性的思考、吸收教育测量学的科研成果、使我们的经验得以升华的时候，有必要先了解一下有关语文测试的一般概念。

第一节 有关的测试学概念

●测试的分类○学业考试、水平考试、学能考试和诊断性考试○标准参照考试和常模参照考试○难度考试和速度考试○个别考试、小规模考试和大规模考试○其它●常用的基本概念和数学方法○质量分析的基本概念——a. 效度、b. 信度、c. 区分度○常用的计算方法——a. 难度、b. 均值、全距和均差、c. 标准差、标准分和变异系数、d. 相关系数

教育测量学是一门专业性很强的学科，不仅有自己的术语概念，而且有相应的数学方法。这些术语和数学方法是广大语文教师所不熟悉的，然而又是进行问题探讨时所不容回避的。为此，下面对一些有关的概念和知识作尽可能扼要的介绍。

测试的分类

随着测试的目的、要求、条件的不同，试卷的拟订、考试的管理、分数的解释等等也有所不同。因此，教师在设计测试的时候，应该大体上知道测试的种类。

对于比较复杂的事物，从不同的角度观察，就可以有不同的分类。下面简要介绍从不同角度对测试的几种分类。

1. 学业考试、水平考试、学能考试和诊断性考试

根据测试的目的和用途，可以分为学业考试（又称成绩考试）、水平考试（常用于人员选拔，在这种情况下，习惯称选拔考试）、学能考试（又称预估考试、潜能考试）和诊断性考试。

①学业考试

学业考试用来考查学生在一段时期内的学习状况。这种测试的特点是它与学生在这个时期的学习内容密切相关，考查的目的在于测量学生是否掌握了所规定的学习内容以及掌握的程度。如果说得通俗一点，这种考试在于测定学生的“昨天”。广大教师所熟悉的“毕业会考”就属于这种考试。

②水平考试

水平考试用来考查学生在某些知识或技能方面所达到的水平，从而测定他们所达到的水平是否是已完成某些特定的要求。如果是像升学那样的选拔性考试，则可以根据所甄别水平的高低而决定是否录取。这种考试和成绩考试不同，它不需要结合学生前一阶段所学的具体内容，而以学生现在所达到的水平为准。它的目标是建立能够适用于各种考生（例如来自不同学校、使用不同课本，等等）的共同标准。由于考生来源、条件等方面的不同（例如重点学校与非重点学校、城市和乡村，等等），这种考试并不意味对他们前一阶段的学习进行鉴定。倘若说得通俗一点，这种考试在于测定学生的“今天”。广大教师瞩目的高考就属于这种考试。

③学能考试

学能考试用以了解学生的潜在能力。说得通俗一点，它在测定着学生的“明天”。这种考试在我国尚未得到充分的发展，只在少数领域开始使用，例如语言的学能考试，它可以用来预测考生能否很好地学习另一种语言。

④诊断性考试

诊断性考试用以了解考生能否使用某种知识或者是否具有某种技能。这种考试往往需要专门设计，在我国也尚未得到充分发展。不过上面谈到的成绩考试和水平考试也常常可以起到诊断性的作用。

2. 标准参照考试和常模参照考试

现在的考试一般都以“分”为计算单位，但是分数只是一个数据，它并不能直接说明问题。比如说，某所学校的某学生平均成绩为 85 分，另一个学生在另一所学校平均成绩为 82 分，这两个学生究竟孰优孰劣，是难以判断的；再如某生期末考试语文得 80 分，数学得 85 分，这两个分数也是不能直接比较的，因为这个学生的语文成绩可能在全班居中上游，而数学成绩实际上只居下游。因此对分数需要加以解释。

在物理测量中，我们用“尺”、“米”等来测量长度，用“斤”、“公斤”等测量重量。长度和重量有“绝对零点”，因此只需要有统一单位就可以完成测量任务。高度则不然，它没有绝对零点，因此对它的测量除了要有统一单位外，还需要有统一的参照点，例如测量地面高度以海平面为参照点，这样的参照点就是“相对零点”。

教育测量和物理测量存在着同样的问题，而且更为复杂，因为它是对心理现象（学习结果）的测量。人的绝大

多数心理现象是没有绝对零点的，这样，不同的考试因为难度不同就有不同水平的零点。不同的相对零点（参照点）就使不同考试的分数不能直接比较，也不能直接加以解释。

这样，从解释分数的方法来看，就有标准参照考试和常换参照考试之分。

①标准参照考试

标准参照考试又称目标参照考试，它用以测量考生是否达到某些事先决定的目标或标准。这种考试完全以一定目标为准，例如汽车驾驶员的考试，应试者可以百分之百通过，也可以一个也没有通过。

标准参照考试成绩的衡量，可以只分“及格”与“不及格”两等，也可以在是否及格的基础上再进一步区分优劣，例如可以用百分制来计算。上面说过的学业考试一般都属于标准参照考试。

这种考试有明确的及格标准，我国所习惯的“60分及格”的观念就是从这种考试中产生的。它已经形成了很强的心理影响，甚至对评估行为起着一定的约束作用。例如作文考试的评分，人们就往往习惯于在心目中先把规定的分值折合成百分制，再把各等次的划分与“60分及格”的观念相对应，所以评分中的“趋中倾向”（参阅本书第二章）并非向分数段的中点聚拢，而是在高于中点的分值上聚合，使分数分布形成明显的负偏态。不管是什么样的考试都要统计“及格率”，也是这种观念的一种反映。这种观念的存在，有时会干扰评分标准的执行，特别是对常换参照考试。

②常换参照考试

常换参照考试是把一个考生的成绩放在考生团体中来衡量，也就是和其他同类考生的成绩相比较，从而判断该考生的水平。它通过建立百分位的常换或标准分的常换来起选拔考生的作用，所以称为常换参照。我们通常采用的“排名次”的做法，通过该考生在团体中的位置来判断其成绩的优劣，就是一种常换参照的方法。

既然在常换参照考试中分数并不能直接反映考生的成绩，我们就需要探求对这种分值给予科学解释的方法。自然界中许多变量的概率都服从正态分布，大量实践证明，人的能力也符合正态分布，考试分数的分布从整体上看也是符合正态分布的，因而它是关于考试的教育测量学研究的基本前提之一。从这点出发，就形成了对分数的意义加以科学解释的一系列概念和方法，关于这些，下文将加以扼要说明。

与标准参照考试不同，常换参照考试是没有“及格线”的。不少人在像全国高考这样的常换参照考试中也来计算“及格率”，这是一种误解。

对“标准参照”和“常换参照”的认识是研究语文测试的一个重要问题。

3. 难度考试和速度考试

根据考试的要求，测试可以分为“难度考试”和“速度考试”。

难度考试目的在于测量考生解答难题的最高能力，作答的时间比较充裕，试题的拟订着眼于难度。难度考试的试题并不都是难度非常大的，它一般包含不同难度的

题目，由易到难排列，其中有些题目则几乎所有的考生都解答不了。

速度考试的目的在于测量考生的反应速度，或者某种技能的熟练程度（例如打字），一般题目比较容易，但时间限制相当严格，所设计的题量可以使几乎所有的考生都难以完成，而以完成的数量（必须解答正确）为衡量成绩的标准。

难度考试和速度考试也可以结合起来。这样的试卷既可以测量考生解答难题的能力，又可以测量他们的思维反应速度和技能的熟练程度，测试的功能比较全面，但在这两个方面的测量精确度则要比上述两种考试低一些。在这种情况下，题量过大和难度过高对测试的结果是没有意义的，甚至会起到相互干扰的作用，因此试卷的编制比较复杂。一般的经验是掌握“两个 75%”的原则，即题量控制在 75%的考生能够答完，难度控制在一般考生能够答出其中的 75%。

对“难度考试”和“速度考试”的认识与试题的拟订、试卷的编制有着密切关系。

4. 个别考试、小规模考试和大规模考试

考试规模的大小对于考务的管理、题型的选择、评分标准的研究、成绩的评定等许多方面都有着重大的影响。

个别考试是规模最小的考试，每次只测一人，考试材料灵活多样，考试方法也灵活多样，可以利用文字材料，也可以利用实物、录音磁带、录像等等，可以口试，也可以笔试。这种考试可以比较全面地评价一个考生，但常常要求测评者有较高的专业素养和随机应变的能力。

小规模考试和大规模考试都属于团体考试，它们可同时考许多人。规模的“大”和“小”之间并没有明确的界限，一般是把考生属于同一群体（例如班级）。由同一测试者进行测评（例如同一任课教师）、参加测试人数较少的称为小规模考试，而把考试对象包括不同群体，由不同的测试者进行测评，参加测试人数较多的称为大规模考试。

考试规模越大，考试的组织与管理越为复杂。主要的问题是：一、如何保证考试公平合理，二、如何有效地控制评分误差。

倘若考生来自学习条件不同的群体（例如使用的课本不同），命题就要注意防止试题偏向于某一群体；倘若考试在不同的监测人员管理下进行，就要注意使不同的监测人员步调一致，包括临场指导语的一致，避免因为管理程序甚至言语的暗示性影响到考试的结果；如果考试在不同的场所或跨地区进行，则还要注意到环境的差异是否会影响到考试的成绩……总之，为了考试公平合理，大规模考试要求有严格的科学管理程序，要求施测过程的标准化。

如果试卷由不同的人员评定，就特别要注意由于阅卷人员的差异所引起的评分误差问题。在这种情况下，一些在小规模考试中常用的题型在大规模考试中就要受到限制，命题人员就要力求试题的客观化，制订详细而明确的评分标准，还要设想可能出现的各种情况，以限制和减少评分的主观随意性；阅卷过程也要实行严格的科学管理，控制阅卷流程中可能出现的差错，并且寻求必要的监测和平衡措施……总之，为了最大限度地减少误差，必须

有统一的比较标准(即相同的单位和参照点),还要研究评分记分的标准化,分数合成的标准化以及分数解释的标准化,等等。

本书后面所探讨的,主要是在大规模考试中所出现的问题。

5. 其它

上面所介绍的考试分类,都是和当前的语文测试有密切关系的。考试还可以从其它角度进行分类,例如:

①文字考试与非文字考试

文字考试所用的是文字材料,考生用文字作答;非文字考试所用的材料是图形、实物等等,考生无需用文字作答。

②分立式考试与综合式考试

分立式考试用来测量学生是否掌握某项技能或能力的特定因素,例如语言考试的语音考察;综合式考试则比较全面地考查有关的各个方面。

③进展性考试与总结性考试

进展性考试用于教学过程当中,总结性考试用于教学结束之后。

此外,根据考试的用途,还可以分为"入学考试"、"安置考试"、"证书考试"、"录用考试"等等,不胜枚举。

常用的基本概念和数学方法

概念是思维的武器,任何一门学科的研究,都需要有自己的概念系列;为了分析的精确性,教育测量还有一些常用的数学方法。不过对于现在的众多教师来说,这些概念和方法可能是陌生的,甚至是不容易

理解的；而对于从事教育测量的人们来说，这些又不过是普通的常识，他们所缺少的，是对于学科特点的认识。考虑到本书的两类不同的读者，以下只介绍与本书内容有关的概念和公式，而且尽量作扼要而通俗的解释。

1．质量分析的基本概念

对试卷、试题质量的分析，有三个基本概念，即：效度、信度、区分度。

①效度（有效性）

在教育测量中，效度是一个非常重要的概念。效度所涉及的问题，是考试能不能真正测量出它所要测量的东西。试题的效度常常不能从表面现象得到反映。例如有时一道看起来好像是测试物理知识和能力的题目，其实测试的是考生的数学能力；再如小学的四则运算题；如果使用了学生不熟悉的词语，它所测试的实际上就成了学生的词汇量和文字认知能力。像这样的题目，也许各项数量指标都还符合要求，但它的效度却应该说是很低的。以语文测试而言，现在常采用提供一定材料来作文的办法，但倘若学生不能正确理解材料的内容，作文势必出现偏差，这时候，试题所考核的就不是考生的文字表述能力而是对材料的理解了。类似这种情况，命题时是必须认真注意的。

效度的名目繁多，各种效度侧重的问题各不相同。最常遇到和使用的效度主要有两类，即“内容效度”和“效标效度”；此外，从研究学科测试的角度来看，值得注意的还有“结构效度”。

甲、内容效度

内容效度要求试题能充分地体现所要测试的内容。为此，就要对所测试的项目作全面而细致的分析，命题前制订出详尽的蓝图（如“双向细目表”）。考试之后，还应该根据题目分析的结果以及其它方面的资料、经验对考试内容做进一步的审核。任何一种测量工具都只是对一定的目的来说才是有效的，所以内容效度只是一个相对的概念。可以说没有一个试卷的编制者能够设计出一份把考生的所有知识和能力都测出来的试卷，因此，不能笼统地说某份试卷是否有效，而应说这份试卷在哪些方面是有效的。内容效度没有定量描述的方法，它没有数量指标。

关于内容效度存在着不同的观点和争论。一种观点认为，对所测试的项目，内容的覆盖面越全越好，因为只有内容覆盖得全，才能防止由于考生存在着强项、弱项而出现试卷偏向，影响考试的公平合理性。持反对意见者则认为每一个测试项目的内容结构不同、重要性不同、所反映的能力层次不同，只有把握住最关键、最重要的内容，才能更深刻地反映出考生对所测项目的把握程度。其实，这个问题的处理要看考试的性质，比如成绩考试，内容的覆盖面可以全一些，可以更好地反映学生的学习状况；选拔考试则应该更注重基本能力，以便了解考生对测试项目的本质方面。不过，各种考试都不宜过于求全，而应该注意区分轻重主次。尤其是像语文这样的综合性很强的学科，它所涉及的知识面很多，由文字而语言，由语法、修辞而文体知识、文学知识、文化常识……考试如果面面俱到，教学就容易面面兼顾，从而分散师生精力，不利于教

学导向。

乙、效标效度

效度的高与低，需要有一定的参照点加以检核，尤其是能力的测试。实际上能力都是无法测量的，因此，我们只能确定一个或几个能反映能力的标准，然后做间接的比较。这种人为确定的效度标准简称"效标"。效标效度是可以量化的，其测算的方法通常以实测分数与效标分数之间的相关系数来表示（相关系数的计算公式见下文）。

某个公认的标准测验上所得的分数、历来的学习成绩、有经验的教师评定等都可以成为效标。在我国，目前还缺乏具有权威性的教育测量方面的效标，所以常常用学生在校的学习成绩作为参试的效度效标。然而从本书第二章可以知道，这样的效标并不是很可靠的，尤其是语文测试，因此亟需国家教育主管部门尽快编制各种量表或其它量具来解决这个矛盾。在作文测试中，用专家团体评分的平均值作为效标，经过多次试验，证明具有比较高的稳定性。

丙、结构效度

结构效度要求测试的结果与测试内容的能力结构一致。结构效度可以用因素分析法加以测算。因素分析就是把一些具有错综复杂的因素归结为数量较少的几个综合因素（公因素），并且用这少数几个因素解释能力结构。通过因素分析（对因素负荷的测算）可以了解不同的试题与相应能力的关系，而且可以知道各公因素之间的适当比例。例如英语 EPT 考试试卷各部分因素负荷的测算（图表 1—1）。