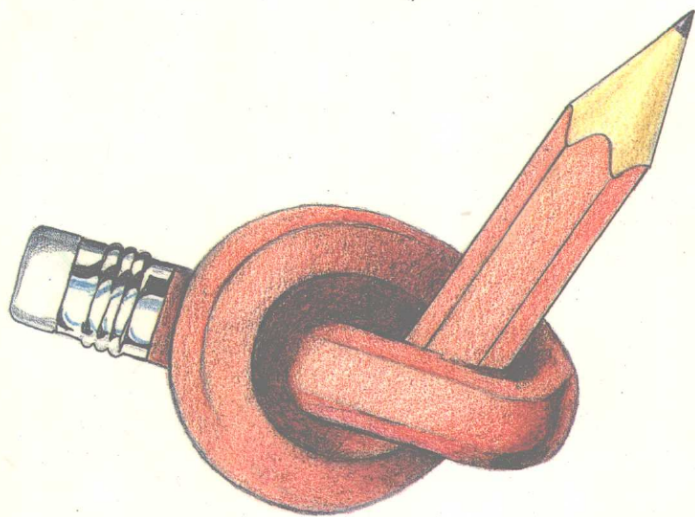


外语测试100问

A GUIDE TO FOREIGN
LANGUAGE TESTING

张端芬 编著



广东经济出版社

外语测试 100 问

张瑞芳 编著

责任编辑：徐雨苍

封面设计：胡 斌

责任技编：梁碧华

外语测试 100 问

张端芬 编著

出版发行 广东经济出版社

(社址：广州市环市东路水荫路 11 号 5 楼)

经 销 广东省新华书店

印 刷 番禺市印刷厂

(厂址：番禺市市桥镇环城西路工农大街 45 号)

规 格 787×1092 毫米 32 开本 4.25 印张

字 数 110,000 字

版 次 1997 年 5 月第 1 版

印 次 1997 年 5 月第 1 次

印 数 1—1000 册

标准书号 ISBN 7—80632—092—X/G·29

定 价 10.00 元

如发现印装质量问题，影响阅读，请与承印厂联系调换。

内容提要

本书系统介绍了外语测试的各种知识与技巧,内容包括外语测试的类型与特点、常用题型及其优缺点、命题方法与步骤、各种外语技能的测试手段与评分方法、试题分析测算的方法、教育统计学常识、分数的展示与解释方法、标准分的意义与类型、常用教育统计计算等。供各类学校的外语教师参考。

目 录

- [1] 为什么要进行外语测试 (Purpose of Foreign Language Tests)
- [2] 什么是水平测试 (Proficiency Tests)
- [3] 什么是成绩测试 (Achievement Tests)
- [4] 什么是分班测试 (Placement Tests)
- [5] 什么是语言禀赋测试 (Aptitude Tests)
- [6] 什么是诊断性测试 (Diagnostic Tests)
- [7] 什么是综合性测试 (Integrative Tests)
- [8] 什么叫分列式测试 (Discrete-point Tests)
- [9] 什么叫实力型测试 (Power Tests)
- [10] 什么叫速度型测试 (Speed Tests)
- [11] 什么是语言测试 (Linguistic Tests)
- [12] 什么是交际能力测试 (Communicative Competence Tests)
- [13] 什么是主观测试 (Subjective Testing)
- [14] 什么是客观测试 (Objective Testing)
- [15] 什么是常模参照测试 (Norm-referenced Tests)
- [16] 什么是标准参照测试 (Criterion-referenced Tests)
- [17] 什么是形成性评价 (Formative Evaluation)
- [18] 什么是总结性评价 (Summative Evaluation)

- [19] 什么是标准化考试 (Standardized Tests)
- [20] 一个优良的测试有哪些特点 (Characteristics of a Good Test)
- [21] 什么是测试的效度 (Validity)
- [22] 什么是内容效度 (Content Validity)
- [23] 什么是效标关联效度 (Criterion-Related Validity)
- [24] 什么是同期效度 (Concurrent Validity)
- [25] 什么是预测效度 (Predictive Validity)
- [26] 什么是构想效度 (Construct Validity)
- [27] 什么叫表面效度 (Face Validity)
- [28] 影响效度的因素有哪些 (Factors Influencing Validity)
- [29] 什么是测试的信度 (Reliability)
- [30] 如何判断试题的信度 (Methods of Estimating Reliability)
- [31] 影响测试信度的因素有哪些 (Factors Influencing Reliability)
- [32] 测试的信度和效度之间有什么关系 (Relation Between Reliability and Validity)
- [33] 什么是测试的实用性 (Practicality)
- [34] 什么是测试的区分性 (Discrimination)
- [35] 如何展示原始分数 (Raw Scores)
- [36] 什么是百分位 (Percentile Rank)
- [37] 什么是正态分布 (Normal Distribution)
- [38] 正态分布曲线下面积的分配规律如何 (Normal Distribution Curve)
- [39] 什么叫称名数据 (Nominal Data)
- [40] 什么叫顺序数据 (Ordinal Data)
- [41] 什么叫等距数据 (Interval Data)

- [42] 什么叫比率数据 (Ratio Data)
- [43] 什么是算术平均数 (Mean)
- [44] 什么是加权算术平均数 (Weighted Arithmetic Mean)
- [45] 什么是中位数 (Median)
- [46] 什么是众数 (Mode)
- [47] 什么是两极差 (Range)
- [48] 什么是标准差 (Standard Deviation)
- [49] 如何计算标准差 (Computation of Standard Deviation)
- [50] 标准差有什么意义 (Meaning of Standard Deviation)
- [51] 什么是真分数 (True Score)
- [52] 什么是标准误 (Standard Error of Measurement)
- [53] 什么是猜题矫正公式及其使用时须注意的问题 (Guessing Correction Formula)
- [54] 多项选择题有哪些形式 (Forms of Multiple-choice Items)
- [55] 出多项选择题应注意哪些问题 (Advice on Writing Multiple-choice Items)
- [56] 多项选择题有哪些主要的优、缺点 (Advantages and Limitations of Multiple-choice Items)
- [57] 出正/误判断题时应注意哪些问题 (Advice on Writing True/False Items)
- [58] 正/误判断题的优缺点是什么 (Advantages and Limitations of True/False Items)
- [59] 如何出完形填空题 (Advice on Writing Cloze Tests)
- [60] 完形填空题有哪些形式 (Varieties of Cloze)
- [61] 完形填空题的优、缺点是什么 (Advantages and Limitations of Cloze)

- [62] 完形填空题如何给分 (Scoring of Cloze)
- [63] 什么是 C—测试 (C—test)
- [64] C—测试有什么优、缺点 (Advantages and Limitations of C—test)
- [65] 听写有什么优、缺点 (Advantages and Limitations of Dictation)
- [66] 听写是用来考什么能力的 (Dictation)
- [67] 简答题有什么优、缺点 (Advantages and Limitations of Short-answer Items)
- [68] 出简答题应注意哪些问题 (Advice on Writing Short-answer Items)
- [69] 出配对题应注意哪些问题 (Advice on Writing Matching Items)
- [70] 如何测试词汇 (Testing Vocabulary)
- [71] 如何测试语法 (Testing Grammar)
- [72] 如何测试听力 (Testing Listening)
- [73] 听力测试题常包括哪些方面的内容 (Content of Listening Tests)
- [74] 出听力考题应注意哪些问题 (Advice on Writing Listening Tests)
- [75] 口试有哪些常用的形式 (Varieties of Speaking Tests)
- [76] 口试如何评分 (Scoring of Speaking Tests)
- [77] 出阅读理解题应注意哪些问题 (Advice on Writing Reading Comprehension Tests)
- [78] 考阅读理解有哪些常用的题型 (Varieties of Reading Comprehension Tests)
- [79] 阅读理解主要有哪些提问类型 (Types of Questions in Reading Comprehension Tests)

- [80] 如何测试写作能力 (Testing Writing)
- [81] 如何给作文评分 (Scoring of Composition)
- [82] 如何对作文实施整体给分法 (Holistic Scoring)
- [83] 出外语测试题有哪些基本步骤 (Steps of Test Construction)
- [84] 命题的一般原则是什么 (General Rules of Test Construction)
- [85] 写题目说明有哪些要求 (Requirements for Test Instructions)
- [86] 为什么要进行试测 (Pretest)
- [87] 原始分有什么缺陷 (Limitations of the Raw Score)
- [88] 为什么需要有标准分 (Standard Scores)
- [89] 什么是标准九分 (Stanines)
- [90] 什么是 Z 分 (Z-score)
- [91] 什么是 T 分 (T-score)
- [92] 试题分析的内容和用途是什么 (Test Analysis)
- [93] 怎样确定试题的难度指数 (Item Difficulty)
- [94] 怎样确定试题的区分度指数 (Item Discrimination Index)
- [95] 怎样评价迷惑项的有效性 (Effectiveness of Distracters)
- [96] 什么是教学效果敏感指数 (Index of Sensitivity to Instructional Effects)
- [97] 什么是相关和相关系数 (Correlation)
- [98] 如何解释相关系数 (Interpretation of Correlation Coefficient)
- [99] 如何计算相关系数 (Computation of Correlation Coefficient)
- [100] 使用相关系数时应注意什么问题 (Caution in Explaining Correlation Coefficient)

1. 为什么要进行外语测试

(Purpose of Foreign Language Tests)

我们进行外语测试的最主要目的就是衡量受测者使用外语的能力。与长度、重量之类的物理测量不同，对人的外语能力的测量是无法直接进行的，只能借助于外语测试这个工具来间接地进行考查，根据受测者对试题作出的反应，推论出其外语水平的状况，从而得出一定的结论。根据不同的情况，外语测试可作如下不同的用途：

(1) 用来将学生按水平分类，编入不同的班级，以利于因材施教，取得较好的教学效果。

(2) 用来诊断学生掌握外语的情况，使教师和学生都知道薄弱点在何处，并据此采取补救措施。

(3) 用来考查受测者的语言禀赋，挑选最有语言天赋的学生进行特定的训练。

(4) 用来考查学生的学业成绩，衡量学习进步的幅度和对所学课程的掌握情况。

(5) 用来评价教学效果。这种测试常用于科研目的，例如研究某种教材或教学方法的优劣。还可用来判断教学是否成功，以利于将来改进教学与大纲。

(6) 用来筛选人员。通过测试来判定受测者是否具有所要求的水平，以此决定其能否参加某种类型的学习或工作。尤其是当名额少于申请者时，可通过考试决定取舍。

(7) 用来鉴定资格。表明受测合格者具有从事相应的外

语工作或晋升某种职称的资格。

2. 什么是水平测试 (Proficiency Tests)

水平测试是用来衡量受测者综合运用外语能力的测试。它的目的是确定受测者目前是否具有所需的外语水平。水平测试的特点是它并不根据任何一个教学大纲或任何一本教材命题。要考查的不是对所教的课程学得怎样，而是根据未来对外语的使用要求考查受测者目前的外语熟练程度。典型的水平测试例子有中国的 EPT，英语四、六级考试，美国的 TOEFL 和英国的 ELTS。衡量受测者某种语言技能（例如口语）水准的测试也属水平测试。水平测试一般都有特定的测试目的。例如，TOEFL 是专为筛选申请就读美国大学的外国学生而设计的。由于目的不同，各种水平测试的内容和针对点也会有所不同。

3. 什么是成绩测试 (Achievement Tests)

成绩测试的目的是评价、衡量学生在掌握所学教材方面取得的进展。通过成绩测试，我们能确定学生在其所在班级中的相对等第，了解其掌握知识的程度和应达到的学业标准之间的差距，并达到巩固所学知识的目的。成绩测试的例子有期末考试、中学毕业会考以及某门课程结束时所进行的考试。评价教学效果的测试也是成绩测试。例如作研究用的后

测试等。

4. 什么是分班测试 (Placement Tests)

我们都知道学生程度不齐。教师难教，学生也难学。所以常常需要将学生按程度分班，以利教学。分班测试就是为这一目的而设计的。它要考查的是学生目前现有的语言能力，所以水平测试的题目常可作分班测试用。有时上一个学期的成绩测试结果也可作为下一个学期的分班依据。由于分班测试往往时间紧迫，没有太多的时间准备试题和改卷，所以常选用间接测试的客观试题，如多项选择题和完型填空等。

5. 什么是语言禀赋测试 (Aptitude Tests)

语言禀赋测试的目的是评估受测者在语言学习方面的禀赋或潜在能力以预测其在学习外语时取得成功的可能性。禀赋高的人学得就快、就好。语言禀赋测试一般是用一种人造的语言来进行测试，所以受测者参试时的外语程度不影响测试结果，有的人也许从未学过外语也可以考得较好。测试的内容主要有：语音与符号的配对能力、辨认能力、记忆力、对语法结构的感受力以及归纳能力等等。但目前对语言禀赋测试的实际效果仍有争议。

6. 什么是诊断性测试 (Diagnostic Tests)

诊断性测试是为教学服务的。它不是考察学生的总体语言水平,而是针对所学的某些具体内容和语言知识进行测试,以了解学生的困难所在。我们可以对某些学习难点进行诊断性测试。如果结果显示学生对此已经掌握,则不必再花太多功夫。通过诊断性测试,可以查明学生的困难之所在以及产生这些困难的原因,使学生知道哪里需要多加努力,而教师则可有针对性地采取补救措施。通过分析学生的错误,教师的教学也就更加有的放矢。诊断性测试宜采用分列式测试,哪里有问题,一目了然。若采用综合性测试,则不易查明问题所在。比如采用听写的形式,就不容易分清造成考得不好的原因到底是因为听不懂还是因为不会写。

7. 什么叫综合性测试 (Integrative Tests)

综合性测试考察的是受测者理解和运用语言的综合能力。因此各种各样的语言构成成分和技能都将同时受到检验。例如外语面试,受测者不仅要听得懂对方的问话,还要运用所掌握的语言知识和技能准确地表达自己的意思。又例如作文,受测者的许多技能都同时受到考查,如词汇、语法、拼写、构思、组织材料等。其他的例子还有完型填空、听写、作文等。

8. 什么叫分列式测试 (Discrete-point Tests)

在分列式测试中，每个题目只考一个语法构成成分的某一点，如某个音素、某个单词、某条语法规则等。此外，每个题目必须是独立存在的，与前后的其他题目无关。更换其中一题不会影响学生对其他题目的回答。最典型的分列式测试题为多项选择题。但看完一篇阅读材料后考理解能力的多项选择题则不属于分列式测试。

分列式测试的长处在于：(1) 针对性强。学生无法回避所考内容，尤其适宜作诊断性测试用。(2) 评分客观、准确。不受人为因素的影响。(3) 易于更换不合适的题目。便于在需经难度检验的大规模测试中使用。

9. 什么叫实力型测试 (Power Tests)

在实力型测试（也称难度测试）中，考生有足够的时间做完所有的题目。个别考生无法完成全部答题是因为不会做，做不出，而不是因为时间不够。这类测试的目的是要使考生充分展示其掌握的语言知识和能力的程度，因此题目的编排往往由易到难。如果把难的题目放在前面，有的考生就会因题目一开始就这么难而失去信心，放弃继续做下去的努力。或者因为在难题上费时过多而无法做完全部题目，致使后面容易做的也没有做。这样就会影响测试结果的准确性。

10. 什么叫速度型测试 (Speed Tests)

速度型测试所给的答题时间非常有限，考生必须尽可能快地完成答题。一般只有极少数人或根本没有人能够做完全部题目。它所测试的不仅仅是语言知识和理解能力，而且还有运用这些知识和能力的速度。通常题与题之间的难度相当。如考快速阅读就是一例。

11. 什么是语言测试 (Linguistic Tests)

语言测试考查的是受测者对语言各构成成分掌握的精确程度。例如语法是否正确、发音是否标准、词汇搭配是否得当等等。

12. 什么是交际能力测试

(Communicative Competence Tests)

交际能力测试注重的是学生使用语言进行交流的能力。重要的不是语言的精确性，而是通过语言来理解和传达信息的能力。其中也包括在各种语言环境中正确使用语言的能力。比如，应征面试就是一种交际能力测试。这时发音是否标准、语法是否规范都不是最主要的，只要不影响意思的表达，可

以接受就行。重要的是能听懂问话，并作出恰当的回答，语言流利，使用域正确。

13. 什么是主观测试 (Subjective Testing)

主观测试指的是需要评分者作出主观判断的测试。例如，作文就是一种典型的主观测试。

14. 什么是客观测试 (Objective Testing)

客观测试指的是评分方式上的客观。评分人不需要作出自己的主观判断，只需根据答题是否与正确答案相同来给分。评分结果不会因评分人不同或评分时间不同而改变。最典型的客观测试题是多项选择题。

15. 什么是常模参照测试 (Norm-referenced Tests)

当我们判断一个分数是好还是不好时，主要有两种解释方法：一个是参照常模，另一个是参照标准。常模指的就是一群类型相同的人的水平的平均值。常模的功用就是给测试分数提供比较的标准，即给测试分数提供参照点，从而使人们可以对测试分数加以解释。常模参照测试就是参照其他考

生的情况来解释某人的得分的测试。在这种情况下，衡量的标准是相对的，而不是绝对的。常模参照测试的目的是要显示考生间外语水平上的差异，提供相对的等第顺序，所以它必须能将分数拉开，以确保可靠的区分性。如果在一个常模参照测试中，人人都得高分，那它决不可能是一个成功的测试。水平测试属于常模参照测试。尤其是大规模的标准化测试，因为要保持每次测试的难度相等，具有可比性，单靠题目上下功夫很难做到，而考生群体的水平则是相对稳定的，所以得靠常模参照测试来达到这一要求。

16. 什么是标准参照测试

(Criterion-referenced Tests)

判断一个分数好坏的另一个方法，就是将考生的表现（分数）与事先定好的标准对比来决定他是否具备某种能力。这样的测试就叫作标准参照测试。这种测试所要达到的目的就是表明某人是否达到了既定的标准，而不关心考生间的成绩差异。如果在一次标准参照测试中，人人都得高分，是值得庆贺的事，因为这表明人人都达到了既定的目标，学习掌握得很好。标准参照测试的例子有单元测试、诊断性测试等。

17. 什么是形成性评价 (Formative Evaluation)

形成性评价是指在教学过程中为了解教学的结果和学生