THE THEORY AND PRACTICE OF ENGLISH LANGUAGE TESTING

英语语言测试的理论与实践

端木庆一　编著

河南人民出版社

河南师范大学学术专著出版基金资助

THE THEORY AND PRACTICE OF ENGLISH LANGUAGE TESTING

英语语言测试的理论与实践

端木庆一　编著

河南人民出版社

图书在版编目(CIP)数据

英语语言测试的理论与实践 = The Theory and Practice of English Language Testing/端木庆一编著. -郑州:河南人民出版社,2005.12
ISBN 7-215-05872-7

Ⅰ.英… Ⅱ.端… Ⅲ.英语-测试-理论-高等学校-教材 Ⅳ.H319

中国版本图书馆 CIP 数据核字(2005)第 155157 号

河南人民出版社出版发行
(地址:郑州市经五路 66 号　邮政编码:450002　电话:65723341)
新华书店经销　　郑州文华印务有限公司印刷
开本　710 毫米×960 毫米　1/16　印张　14.75
字数　240 千字　印数　1-2 000 册
2005 年 12 月第 1 版　2005 年 12 月第 1 次印刷

定价:22.80 元

前　言

英语语言测试是应用语言学的一个重要分支。

测试研究涉及心理语言学、认知语言学、第二语言习得理论、普通语言学及教育测量学等学科领域。半个世纪以来，英语测试理论有了长足发展，测试学专著和理论文章颇丰，考试形式呈多样化涌现，这些表明了语言测试的活力方兴未艾。英语语言测试是升学、就业、晋升或出国学习等活动的重要测量手段；考试成绩与教学质量和教师教学成果直接挂钩，这些又体现了语言测试的必要性和重要性。

英语语言测试是英语教师专业和技能的一个重要组成部分，测试课也是高等师范院校高年级英语专业本科生、英语语言教育硕士研究生、全日制英语研究生必修课程之一，其重要性不言而喻。作者认真参阅了有关英语测试理论和实践方面的原版专著，编撰了本书，以期为广大英语教师提高专业知识技能和高校英语专业学生学习的需要，做出些许努力。

本书编写过程中，作者指导的部分研究生马艳芳、贺光辉、李美艳、张淑静收集了部分资料、编撰了有关章节，在此对他们表示感谢。

由于作者水平有限，加之时间仓促，书中问题在所难免，敬请读者给予批评指正。

谢谢！

端木庆一
河南师范大学外国语学院
2005 年 12 月

Contents

Introduction

Ⅰ. Types of Language Tests ····· 3

Ⅱ. Functions of Language Tests ····· 13

Chapter One Development of Testing

Ⅰ. The Three Phases of Language Testing ····· 19

Ⅱ. Schools of Testing in History ····· 21

Chapter Two Test Specifications

Ⅰ. What Are Test Specifications? ····· 29

Ⅱ. Who Need Test Specifications? ····· 29

Ⅲ. Writing Specifications for the Test ····· 31

Chapter Three Tasks of Testing

Ⅰ. A Framework of Language Test Task Characteristics ····· 39

Ⅱ. Some Detailed Explanations of Test-task Terms ····· 42

Ⅲ. Language Test Task in Pracice ····· 43

Chapter Four Qualifications for Examiners

Ⅰ. Introduction ····· 53

Ⅱ. Qualifications for Item Writing ····· 53

Ⅲ. How to Be a Good Item Writer? ····· 54
Ⅳ. How to Start Writing Test Items? ····· 56
Ⅴ. The Importance of Monitoring Examiners ····· 57

Chapter Five Listening Comprehension and Dictation

Ⅰ. Listening Comprehension ····· 63
Ⅱ. Dictation ····· 73

Chapter Six Writing Multiple-choice Items (MCI)

Ⅰ. Categories of MCI ····· 81
Ⅱ. Advantages and Disadvantages of MCI ····· 82
Ⅲ. Qualifications for MCI Writing ····· 84
Ⅳ. MCI Writing Moderation ····· 86
Ⅴ. Other Problems to Be Noticed in MCI Writing ····· 89

Chapter Seven Cloze Test

Ⅰ. Introduction ····· 95
Ⅱ. What Is a Cloze Test? ····· 96
Ⅲ. General Principles of Cloze Test ····· 102

Chapter Eight Testing of Reading Comprehension

Ⅰ. Introduction ····· 107
Ⅱ. What Is Reading? ····· 107
Ⅲ. Why Test Reading? ····· 112
Ⅳ. How to Test Reading? ····· 112
Ⅴ. How to Design Questions? ····· 115
Ⅵ. Question Types ····· 116
Ⅶ. How to Assess Questions? ····· 118
Ⅷ. How to Interpret Test Results? ····· 118

Ⅸ. Setting Tasks of Reading Comprehension Test ····· 122
Ⅹ. Practical Advice on Item Writing ····· 123

Chapter Nine Testing of Writing

Ⅰ. Introduction ····· 127
Ⅱ. Two Major Problems ····· 128
Ⅲ. Components of Writing ····· 128
Ⅳ. Types of Writing Tasks ····· 129
Ⅴ. Issues Related to Intermediate and Advanced Writing Tests ····· 130
Ⅵ. Selecting the Appropriate Writing Tasks ····· 131
Ⅶ. Writing Task Prompts ····· 136

Chapter Ten Oral Test

Ⅰ. Introduction ····· 147
Ⅱ. Setting the Tasks ····· 147
Ⅲ. Criterial Levels of Performance ····· 148
Ⅳ. Format ····· 150
Ⅴ. Advice on Planning and Conducting Oral Tests ····· 151
Ⅵ. Elicitation Techniques ····· 154
Ⅶ. Techniques not Recommended ····· 157
Ⅷ. Describing the Criterial Levels ····· 158
Ⅸ. Proficiency Descriptions ····· 159
Ⅹ. Training of Scorers ····· 162

Chapter Eleven Test Reliability and Validity

Ⅰ. Analysis of Test Reliability ····· 167
Ⅱ. Methods of Reliability Computation ····· 168
Ⅲ. Analysis of Test Validity ····· 174

Chapter Twelve The Interpretation and Analysis of Test Scores

Ⅰ. Distribution of Test Scores ····· *183*
Ⅱ. The Statistical Indexes of Describing the Central Tendency ··· *185*
Ⅲ. The Statistical Indexes of Describing the Dispersion ····· *189*
Ⅳ. Z Score ····· *195*
Ⅴ. Skewedness and Kurtosis ····· *197*
Ⅵ. Correlation Coefficient ····· *198*

Appendix Ⅰ Categories of Popular English Language Tests ····· *201*

Appendix Ⅱ Test Glossary ····· *206*

参考书目 ····· *229*

Introduction

This book is written for teachers of English who are responsible for teaching the language and drawing up tests of language ability, and for students of English who may be actively involved in learning the language. But it must be emphasized that the evaluation of student language performance for purposes of comparison or selection is only one of the functions of test. Although most teachers wish to evaluate individual performance, the aim of the classroom test is different from that of the external examination. While the latter is generally concerned with evaluation for the purpose of selection, the classroom test is concerned with evaluation for the purpose of enabling teachers to increase their own effectiveness by making adjustments in their teaching to enable certain groups of students or individuals in the class to benefit more. Tests may be constructed primarily as devices to reinforce learning and to motivate the student or primarily as a means of assessing the student's performance in the language. In the former case, the test is geared to the teaching that has taken place, whereas in the latter case the teaching is often geared largely to the test.

A great number of examinations in the past have encouraged a tendency to separate testing from teaching. Both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other. And there could be no science as we know it without measurement. Testing, including all forms of language testing, is one form of measurement, and good testing can be used as a valuable teaching device. Tests, to be useful, must provide us with reliable and valid measurements for a variety of purposes, and they have.

I. Types of Language Tests

Just as there are many purposes for which language tests are developed, so

there are many types of language tests. As has been known, some types of tests serve a variety of purposes while others are more restricted in their applicability. There are, however, many important broad categories of tests that do permit more efficient description and explanation. Many of these categories stand in opposition to one another, but they are at the same time bipolar or multi-polar in the sense that they describe two or more extremes located at the ends of the same continuum. Many of the categorizations are merely mental constructs to facilitate understanding.

We use tests to obtain information. The information that we hope to obtain will of course vary from situation to situation. It is possible, nevertheless, to categorize tests according to a small number of kinds of information being sought. This categorization will prove useful both in deciding whether an existing test is suitable for a particular purpose and in writing appropriate new tests where these are necessary. Here we will discuss in the following sections different tests according to their uses, scoring methods or score interpretations.

1. According to the uses of tests, we have

1) *Aptitude Test(Prognostic or Predictive Test)*

Aptitude tests are most often used to measure the suitability of a candidate for a specific program of instruction or a particular kind of employment. For this reason these tests are often used synonymously with intelligence tests or screening tests. A language aptitude test may be used to predict the likelihood of success of a candidate for instruction in a foreign language. The modern language aptitude test is a case in point. Frequently vocabulary tests are effective aptitude measures, perhaps because they correlate highly with intelligence and may reflect knowledge and interest in the content domain.

2) *Diagnostic Test(Formative or Progress Test)*

Diagnostic test is used to identify students' strengths and weaknesses. They are intended primarily to ascertain what further teaching is necessary. At the level of broad language skills this is reasonably straightforward.

We can be fairly confident of our ability to create tests that will tell us that a student is particularly weak in, say, speaking as opposed to reading in a

language, and we may analyze samples of a student's performance in writing or speaking in order to create profiles of the student's ability with respect to such categories as "grammatical accuracy" or "linguistic appropriacy". But it is not so easy to obtain a detailed analysis of a student's command of grammatical structures, something which would tell us, for example, whether she/he had mastered the present perfect/past perfect tense distinction in English. In order to be sure of this, we would need a number of examples of the choice the student made between the two structures in every different context which we thought was significantly different and important enough to warrant obtaining information on. A single example of each would not be enough, since a student might give the correct response by chance. As a result, a comprehensive diagnostic test of English grammar would be vast (think of what would be involved in testing the model verbs, for instance). The size of such a test would make it impractical to administer in a routine fashion. For this reason, very few tests are constructed for purely diagnostic purposes, and those that there are do not provide very detailed information.

The lack of good diagnostic tests is unfortunate. They could be extremely useful for individualized instruction or self-instruction. Learners would be shown where gaps exist in their command of the language, and could be directed to sources of information, exemplification and practice. Happily, the ready availability of relatively inexpensive computers with very large memories may change the situation. Well-written computer programs would ensure that the learner spent no more time than was absolutely necessary to obtain the desired information, and without the need for a test administrator. Tests of this kind will still need a tremendous amount of work to produce. Whether or not they become generally available will depend on the willingness of individuals to write them and of publishers to distribute them.

3) *Achievement Test* (*Attainment Test*)

Most teachers are unlikely to be responsible for proficiency tests. It is much more probable that they will be involved in the preparation and use of achievement tests. In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how

successful individual students, groups of students, or the courses themselves have been in achieving objectives. They are of two kinds: final achievement tests and progress achievement tests.

Final achievement tests are those administered at the end of a course of study. They may be written and administered by ministries of education, official examining boards, or by members of teaching institutions. Clearly the content of these tests must be related to the courses with which they are concerned, but the nature of this relationship is a matter of disagreement amongst language testers.

In the view of some testers, the content of a final achievement test should be based directly on a detailed course syllabus or on the books and other materials used. This has been referred to the "syllabus-content approach". It has an obvious appeal, since the test only contains what it is thought that the students have actually encountered, and thus can be considered, in this respect at least, a fair test. The disadvantage is that if the syllabus is badly designed, or the books and other materials are badly chosen, then the results of a test can be very misleading. Successful performance on the test may not truly indicate successful achievement of course objectives. For example, a course may have as an objective the development of conversational ability, but the course itself and the test may require students only to utter carefully prepared statements about their home town, the weather, or whatever. Another course may aim to develop a reading ability in France, but the test may limit itself to the vocabulary the students are known to have met. Yet another course is intended to prepare students for university study in English, but the syllabus (and so the course and the test) may not include listening (with note taking) to English delivered in lecture style on topics of the kind that the students will have to deal with at university. In each of these examples, which are based on actual cases, test results will fail to show what students have achieved in terms of course objectives.

The alternative approach is to base the test content directly on the objectives of the course. This has a number of advantages. For example, it makes it possible for performance on the test to show just how far students have achieved those objectives. This in turn puts pressure on those

responsible for the syllabus and for the selection of books and materials to ensure that these are consistent with the course objectives. Tests based on objectives work against the perpetuation of poor teaching practice, something which course-content-based tests, almost as if part of a conspiracy, fail to do. It is my belief that to base test content on course objectives is much to be preferred: it will provide more accurate information about individual and group achievement, and it is likely to promote a more beneficial backwash effect on teaching.

4) *Placement Test*

Placement test, as its name suggests, is intended to provide information which will help to place students at the stage of the teaching program most appropriate to their abilities. Typically it is used to assign students to classes at different levels as well as to screen them with extremely low English proficiency from participation in regular university instruction.

Placement test can be bought, but this is not to be recommended unless the institution concerned is quite sure that the test being considered suits its particular teaching program. No one placement test will work for every institution, and the initial assumption about any test that is commercially available must be that it will not work well.

The placement tests which are most successful are those constructed for particular situations. They depend on the identification of the key features at different levels of teaching in the institution. They are tailor-made rather than bought off the peg. This usually means that they have been produced "in house". The work that goes into their construction is rewarded by the saving in time and effort through accurate placement.

5) *Proficiency Test*

Proficiency test is designed to measure people's ability in a language regardless of any training they may have had in that language. The content of a proficiency test, therefore, is not based on the content or objectives of language courses which people taking the test may have followed. Rather, it is based on a specification of what candidates have to be able to do in the language in order to be considered proficient. This raises the question of what we mean by the word "proficient".

In the case of proficiency tests, "proficient" means having sufficient command of the language for a particular purpose. An example of this would be a test designed to discover whether someone can function successfully as a United Nations' translator. Another example would be a test used to determine whether a student's English is good enough to follow a course of study at a British or American university. Such a test may even attempt to take into account the level and kind of English needed to follow courses in particular subject areas. It might, for example, have one form of the test for arts subjects, another for sciences, and so on. Whatever the particular purpose to which the language is to be put, this will be reflected in the specification of test content at an early stage of a test's development.

There are other proficiency tests which, by contrast, do not have any occupation or course of study in mind. For them the concept of proficiency is more general. British examples of these would be the Cambridge examinations such as EILTS (English as International Language Test Syndicate) and Oxford EFL examinations (Preliminary and Higher), and American examinations such as TOEFL (Test of English as a Foreign Language) and GRE (Graduate Recording Examination). The function of these tests is to show whether candidates have reached a certain standard with respect to certain specified abilities. Such examining bodies are independent of the teaching institutions and so can be relied on by potential employers to make fair comparisons between candidates from different institutions and different countries. Though there is no particular purpose in mind for the language, these general proficiency tests should have detailed specifications saying just what it is that successful candidates will have demonstrated that they can do. Each test should be seen to be based directly on these specifications. All users of a test (teachers, students, employers, etc.) can then judge whether the test is suitable for them, and can interpret test results. It is not enough to have some vague notion of proficiency, however prestigious the testing body concerned.

Despite differences between them of content and level of difficulty, all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken. On the other hand, as we saw

that such tests may themselves exercise considerable influence over the method and content of language courses. Their backwash effect may be beneficial or harmful. The effect of some widely used proficiency tests is more harmful than beneficial. However, the teachers of students who take such tests, and whose work suffers from a harmful backwash effect, may be able to exercise more influence over the testing organizations concerned than they realize.

2. According to different scoring methods of tests, we have

1) *Objective and Subjective Tests*

The distinction between *objective test* and *subjective test* here is between methods of scoring, and nothing else. If no judgment is required on the part of the scorer, then the scoring is objective. A multiple choice test, with the correct responses unambiguously identified, would be a case in point. If judgment is called for, the scoring is said to be subjective. There are different degrees of subjectivity in testing. The impressionistic scoring of a composition may be considered more subjective than the scoring of short answers in response to questions on a reading passage.

The following diagram shows us the degrees of subjectivity of different tests:

Multiple choice (True/False questions) — Objective
↓
Dictation
↓
Cloze test
↓
Paraphrase or Explanation
↓
Bilingual-translation
↓
Composition — Subjective

2) *Direct and Indirect Tests*

Testing is said to be *direct* when it requires the candidate to perform