语言测试

Language Testing

王振亚 编著

河北大学出版社
Hebei University Press

语言测试

Language Testing

王振亚 编著

河北大学出版社
Hebei University Press

图书在版编目(CIP)数据

语言测试=Language Testing/王振亚编著.保定：河北大学出版社，2009.5
ISBN 978-7-81097-366-3

Ⅰ.语… Ⅱ.王… Ⅲ.英语-测试 Ⅳ.H319.3
中国版本图书馆CIP数据核定(2009)第042383号

责任编辑：臧燕阳 Tel:0312-5921826 E-mail:zyyzmq@yahoo.com.cn
装帧设计：王占梅
责任印制：闻 利

出版：河北大学出版社（保定市五四东路180号）
经销：全国新华书店
印制：河北新华印刷一厂
规格：1/32(880mm×1230mm)
印张：11.25
字数：367千字
印数：0001～1000册
版次：2009年6月第1版
印次：2009年6月第1次
书号：ISBN 978-7-81097-366-3/H·73
定价：18.00元

前　言

本书是在给外国语言学及应用语言学专业的硕士研究生开设的语言测试课程讲义的基础上编写而成。本书主要分为三个部分。第一部分(第2—5章)介绍、评价主要的现代语言测试模型,包括心理测量学—结构主义语言测试模型、综合语言测试模型、交际语言运用测试模型,以及Bachman的语言测试模型。第二部分(第6—12章)介绍、简单评价词汇、语法结构、听力理解、阅读理解、口语、写作测试的主要方法以及语言测试的构卷与实施。第三部分(第13—15章)介绍与测试及测试研究相关的数据统计方法。

本书的目标读者是外国语言学及应用语言学专业的硕士研究生、英语教师、英语专业的师范生及其他对语言测试感兴趣的人士,力求为初次涉猎语言测试专业领域的目标读者全面介绍语言测试的基本理论概念、实践方法和研究手段。本书用英语编写的目的是为目标读者阅读语言测试的英语文献打下基础。

本书引用过的相关著作统一列入参考文献。在这里向这些著作的作者及出版机构表示感谢。

对本书中的错误与不足,望读者指正,以利改进。

王振亚

2008年10月于北京语言大学

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

Chapter 1 Fundamentals and types of tests

Ⅰ. Fundamentals of tests

1. The term "test" and the related terms

1) Test

A test can be defined as "a method for measuring a person's ability or knowledge in a given area" (Brown, 1994). A test is a method. There are established principles and procedures for the development, administration, scoring and evaluation of tests. There are widely used and carefully studied test item and task types in testing. A test is used to measure, to quantify individuals' skills or knowledge in either mathematically precise terms or broad and inexact terms. A test is normally prepared for a group of test takers or candidates. So the test developers should have profound knowledge about these candidates. And a test can measure only a small portion of human knowledge or ability. In language testing what is measured is general language proficiency, lexical and grammatical knowledge, or productive or receptive skills, all of which are non-observable, and therefore difficult to measure.

In testing the term is often used in three senses. Firstly, it may

refer to an assessment procedure consisting of s set of components prepared to measure the candidates' lexical knowledge, grammatical knowledge, and listening, reading, speaking and writing skills. In this sense, a test is synonymous to an examination. Secondly, the term may be used to refer to a single task or component designed to measure an area of knowledge or skill, such as lexical knowledge, grammatical knowledge, and listening, reading, speaking and writing skills. In this sense a test can be part of a complete assessment instrument. So a writing test or a cloze test can be included in a single examination. Thirdly, the term may refer to an assessment procedure that is relatively short and easy to administer, often devised for use in an institution or as part of a research program or for the test validation purposes. For example, the term test in "class progress test", "pilot test", "anchor test", etc. is often used in this sense.

2) Measurement, evaluation, assessment and examination

In literature on educational measurement and language testing, test or testing, measurement, evaluation, assessment and examination are synonymous to each other. There are similarities and differences between these terms.

Measurement refers to the process of quantifying, using numerical values to indicate, the characteristics of individuals, such as language proficiency, lexical knowledge, and reading skills. Normally, in measurement only quantitative data are used. Therefore, measurement is in contrast with purely qualitative descriptions, in which the focus of attention is not comparison between individuals. A test may consist of a set of questions to each of which there is a correct or ac-

ceptable answer. In measurement tests can be used to collect quantitative data. But questions whose answers cannot be judged as correct or acceptable may also be included. For example, according to the answers given by a subject to a number of questions we may say that the subject is introvert or extravert, without knowing whether the subject's answers were true or not. In this sense measurement is more inclusive than testing. Testing is a special form of measurement.

In evaluation such qualitative methods as interview, questionnaire, and observation can be used to gather information systematically. The purpose for which evaluation is conducted in educational institutions is to determine the effectiveness of educational programs and to determine the future of the programs. In evaluation tests are frequently used to gather data too. For example, pre-tests and post-tests are often administered to determine the progress the students have made and achievement tests are often used to gather information about the students' achievements in the educational programs.

Assessment is a term often used interchangeably with testing. But it can be used more broadly to encompass the gathering of educational data, including test data, for the purpose of evaluation. In assessment such instruments as interview, case study, questionnaire, and observation are often used. More narrowly assessment is used to indicate assessment procedures which do not involve tests. Assessment may be conducted on the educational achievement of an individual learner or a group of learners. And assessment can also be conducted about the effectiveness of educational programs, including curriculum, teaching methods, teaching materials, educational re-

此为试读,需要完整PDF请访问: www.ertongbook.com

sources, teaching plans and the teaching personnel.

Examination is a term generally used synonymously and often interchangeably with "test". Even though there might not be any clear distinction in meaning between these two terms, one may be preferred over the other in certain contexts. For example, the term examination is more likely to be used for syllabus-related assessment. The assessment format may also influence people's choice between these two terms: an objective, discrete-point assessment procedure is more likely termed as test, while a subjective, direct assessment procedure is more likely to be called "examination".

2. Problems in language testing

Even though language testing has been practiced and carefully studied for a long time, there are still problems in this branch of applied linguistics. Some major ones are presented below.

1) No single approach to language testing is universally accepted. There are competing approaches to language testing. For example, the psychometric-structuralist, integrative, communicative approaches are all modern approaches in the field of language testing. They all play important roles in language teaching and testing. All of them have strong points and weaknesses as well. None of them can be said to be better than the others.

2) Language testing is usually based on limited samples of behavior. A language test normally lasts only a couple of hours, being able to take only a rather limited sample of the test taker's behavior. And we know that there will be a sampling error whenever we use a sample to estimate the characteristics of the population. But in language testing situations we have to rely on a limited sample of behav-

ior for the assessment of the construct we are interested in.

cause there will always be some error when we use a sample to estimate the characteristics of the population, when we rely on a limited sample of the test taker's behavior to infer what the test taker knows about and can do in his target language. The measurement obtained or the scores we assign to the test taker according to his performance on the test can never provide us with the precise information about the test taker's knowledge or ability measured by the test. There is always a measurement error no matter how carefully the test is prepared, administered and scored.

4) The units of the scores are not well defined. If the units of the scores obtained are well defined, we will know the exact meanings of the scores and the distance between any pair of adjacent scores will be exactly the same. But in many testing situations we do not know the exact meaning of the scores we assign to the test taker and in no testing situations we can say for sure that the difference between any pair of scores with an equal interval is the same.

5) Language is used as the medium to measure the test taker's abilityto use the language. This is a problem specifically found in language testing situations. In other types of educational measurement language might be used as the medium to measure the test taker's ability to attack the test problems that are not about language. This means that the scorer will attend to exclusively the content when scoring the test taker's performance on the test. But this is not possible in language testing situations. For example, in scoring the test taker's performance on a composition test of writing it is

illegitimate for the scorer to focus exclusively on the content of the composition. He or she is expected to attend to both the content and the expression in scoring the composition. In addition to the ideas and the organization of the ideas, the mechanical techniques of writing, the choice of words and structures, the textual cohesion and coherence, the stylistic appropriateness of the composition should be assessed as well. This is by no means easy.

These problems are not likely to be removed from language testing. The best we can do is to reduce or minimize their influence upon language testing. The educational measurement and language testing specialists have developed useful instruments for the evaluation and improvement of test items and tests as a whole.

3. The true score

The true score is the hypothetical error free score a person would obtain, reflecting the test taker's true ability or knowledge in relation to the test in question. Technically, it is the average of the scores a person would earn in many applications of the same test. Even though the true score can never be observed, it can be estimated. Since language testing has the problems discussed in the preceding section and other problems, it is impossible for a test to be free from measurement errors (indicated by the magnitude of the error score). As a result, there is always a distance between the true score and the observed score, the score we actually assign to a test taker according to his or her performance on the test. This distance is termed as the error score. The relationship between the true score, the observed score and the error score can be presented graphically by an equation:

$Xt = Xo + Xe$

Xt = the true score; Xo = the observed score; Xe = the error score

What should be noted is that the true score can be higher or lower than the observed score. Therefore, the error score can be a positive or a negative number.

4. The characteristics of a good test

All educational measurement and language testing specialists agree that a good test should have three qualities: validity, reliability and practicality.

Validity refers to the degree to which a test measures what it is supposed to measure. A test of reading comprehension, for example, should measure the test taker's reading comprehension ability. That is, the scores obtained from this test should provide us with an accurate representation of the test taker's reading comprehension ability. Apparently, if a composition is scored according to the number of words it contains, the score cannot provide an accurate indication of the test taker's composition writing ability which cannot be equated with the speed at which the test taker writes words. The most commonly discussed types of validity are content validity, construct validity, concurrent validity and predictive validity. These types of validity and the relevant validation methods will be presented in Chapter 16.

Reliability refers to the degree to which a test gives consistent results. Ideally, if there is no measurement error, a test should give the same results no matter when and where it is administered and by

whom it is scored as long as it is administered to the same group of test takers or test takers who are of the same ability level. To reduce or minimize the measurement error will improve the reliability of a test. Reliability is commonly classified into test reliability and rater reliability. Three methods are commonly used to improve test reliability or reduce or minimize the measurement error of tests:

(1) standardizing the testing conditions;

(2) lengthening the test;

(3) constructing better test items since the test taker tends to respond to such items more consistently.

Rater reliability refers to the level of agreement between two or more independent raters in their judgments of the test performance of a test taker or a groups of test takers and the extent to which a particular rater is consistent in using the measurement scale to evaluate different test takers' performance. Rater reliability is one of the major concerns of the direct speaking and writing tests. Rater training is a method commonly used to try to improve the rater reliability of a test. A new statistical technique, the multi-faceted Rasch analysis, is likely to improve the rater reliability of tests.

A valid test must be reliable. The error score cannot provide us with the information we are interested in. However, a reliable test may not necessarily be valid. A test that can measure consistently may not measure what it should measure. Reliability is a necessary but not a sufficient quality of a test.

Practicality refers to the degree to which a test is applicable to a

certain language testing situation. In theory practicality may not be as important as validity and reliability. But in practice it can never be ignored. The importance of practicality stems from the recognition that however valid and reliable a test may be, if it is not practical to administer it in a given situation then it will not be used in that situation. Practicality covers a range of issues: the cost of test development, maintenance, administration and scoring, ease of scoring, time required to administer and score the test, ease of administration, equipment required, etc.

A good test is a test that is valid, reliable and practical.

Ⅱ. Types of tests

1. Tests classified according to their educational purposes

Language tests can be classified according to their different features. For example, they can be classified into proficiency, achievement, progress, diagnostic, aptitude and placement tests according to the educational purposes for which they are prepared and used.

1) Proficiency tests

Proficiency tests are tests that measure how much of a language someone has learned. They are not based on any language teaching syllabuses or programs. The American TOEFL test, the British-Australian IELTS test and the Chinese PETS test are typical standardized English proficiency tests. They are not based on any course of instruction. Proficiency tests often measure the test takers' general language proficiency in relation to a particular real world purpose. For example, TOEFL is used to measure the English language profi-

此为试读,需要完整PDF请访问: www.ertongbook.com