Test characteristics

Science, Statistics

Test characteristics Hossein Farhady Fundamental Concepts in Language Testing (4) Characteristics of Language Tests: Total Test Characteristics* Hossein Farhady University for Teacher Education Iran University of Science and Technology Introduction The first two articles in the series dealt with explaining two fundamental concepts in language testing, namely form of the tests and functions of the tests. The third paper was devoted to explaining the characteristics of an individual item. The processes of planning, preparing, reviewing, and pre-testing were discussed. In the pre-testing section of the previous article, the procedures for determining item characteristics including item facility, item discrimination, and choice distribution were also discussed. It should be clarified at the outset that if the items, which are the building blocks of a test, meet the criteria that have been introduced before, the whole test will be most likely acceptable. However, the assumption that good items will necessarily produce a good test may not always come true. Test developers should go one more step to determine the characteristics of the total test. This article, therefore, focuses on total test characteristics that include reliability, validity, and practicality. Reliability Reliability is one of the most important characteristics of all tests in general, and language tests in particular. In fact, an unreliable test is worth nothing. In order to understand the concept of reliability, an example may prove helpful. Suppose a student took a test of grammar comprising one hundred items and received a score of 90. Suppose further that the same student took the same test two days later and got a score of 45. Finally suppose that the same student took the same test for the third time and received a score of 70. What would you think of these scores? What would you think of the student? Assuming that the student's knowledge of English cannot go under drastic changes within this short period, the best explanation would be that there must have been something wrong with the test. How would you rely on a test that does not produce consistent scores? How can you make a sound decision on the basis of such test scores? This is the essence of the concept of reliability, i. e., producing consistent scores. Although the example mentioned above may demonstrate a very extreme case, it is not however, impossible. Reliability, then, can be technically defined as " the extent to which a test produces consistent scores at different administrations to the same or similar group of examinees". If a test produced exactly the same scores at different administrations to the same group, that test would be perfectly reliable. This perfect reliability, nevertheless, does not practically exist in reality. There are many factors influencing test score reliability. These factors range from examinees' differing mental 19 Test characteristics Hossein Farhady and physical conditions to the precision of the test items, and to the administration as well as scoring procedures. Therefore, reliability is " the extent to which a test produces consistent scores." This means that the higher the extent, the more reliable the test. Statistically speaking, reliability is represented by the letter " r", whose magnitude fluctuates between zero and one; zero and one demonstrate maximum and minimum degree of test score reliability. It should be mentioned that " R" is an independent statistical concept. It does not have anything to do with the content or the form of the test. It solely deals with the scores produced by a test. In fact, one can estimate " R" without having any information about the content of the test. Thus, when one talks about the reliability of a test, he refers to the scores and not to the content or the form of the test. Understanding the concept of reliability, one should next estimate “ R" which requires some statistical competency. In the following section, an attempt is made to explain the procedures to estimate “ r" in as non-technical terms as possible. Four methods of estimating reliability — (1) test-retest, (2) parallel forms, (3) split-half, and (4) KR-21 — will be briefly explained. 1. Test-Retest Method As the name implies, in this method a single test is administered to a single group of examinees twice. The first administration is called “ test" and the second administration is referred to as “ retest". The correlation between the two sets of scores, obtained from testing and retesting, would determine the magnitude of reliability. Since there is a time interval (usually more than two weeks) between the two administrations, this kind of reliability estimate is also known as “ stability of scores over time". Although obtaining reliability estimates through test-retest method seems very easy, it has some practical disadvantages. First, it is not very easy to have the same group of examinees available in two different administrations. Second, the time interval creates two obstacles. On the one hand, if it is very short, there might be practice effect as well as memorization effect carried over from the first administration. On the other hand, if the interval is too long, there will be the learning effect, i. e., the examinee’s state of knowledge will not be the same as it was in the first administration. To avoid these problems, other methods of estimating reliability have been developed. 2. Parallel Forms Method In order to remove some of the problems inherent in the test-retest method, experts have developed the parallel forms method. In this method, two parallel forms of a single test are given to one group of examinees. The correlation between the scores obtained from the two tests is computed to indicate the reliability of the scores. This method has an advantage over the test-retest method in that there is no need for administering the test twice. Thus, the problem of examinees' knowledge undergoing changes does not exist in this method. Nevertheless, this method has a major shortcoming. That is, constructing two parallel forms of a test is not an easy task. There are certain logical and statistical criteria that a pair of parallel forms must meet. 20 Test characteristics Hossein Farhady Therefore, most teachers and test developers avoid this method. Due to the complexity of the task, they prefer to use other methods of estimating reliability. 3. Split-Half Method In test-retest method, one group of examinees was needed for two administrations. In parallel forms method, on the other hand, two forms of a single test were needed. Each of these requirements is considered a disadvantage. To obviate these shortcomings, the split-half method has been developed. In this method, a single form of a test is given to a single group of examinees. Then each examinee's test is split (divided) into two halves. The correlation between the scores of the examinees in the first half and the second half will determine the reliability of the test scores. The only problem with this method is how to divide the test items into two halves. The best way is to use odd and even items to form each half, i. e., items numbered 1, 3, 5, 7, etc. will constitute the first half, and items numbered 2, 4, 6, 8, ... will form the second half. 4. The KR-21 Method The previously mentioned methods to estimate test score reliability require a statistical procedure called ' correlation'. Majority of teachers and non-professional test developers, however, are not quite familiar with statistics. Thus, they may have some problems in using statistical formulas and interpreting the outcome of statistical analyses. To overcome these problems, two statisticians — named Kuder and Richardson — developed a series of formulas to be used in statistics. One of these formulas is used to estimate test score reliability through simple mathematical operations. The formula is called KR-21, in which K and R refer to the first initials of the two statisticians and 21 refers to the number of the formula in the series. This formula is used to estimate the reliability of a single test given to one group of examinees through a single administration. This method requires only the testers and teachers to be able to calculate two simple statistical parameters. These parameters are (1) the mean and (2) variance. The methods of computing the mean and variance are explained in almost all introductory statistics books. However, for the purposes of clarification, a brief explanation of each parameter will be given here. For further information, interested readers are to consult statistics books. (1) The Mean: The mean, commonly known as the average, is the most frequently used concept in statistics. It simply refers to a single score that best represents the __ __ scores of a group. If each score is symbolized as X, then the mean (represented by X and read X bar) will be computed by adding up all Xs and dividing the sum by the number of scores (represented by N). To represent the sum of scores, the Greek letter (Î£), read 'sigma' is used in statistics. Thus the statistical formula to compute the mean would be: __ Î£X X = –– N 21 Test characteristics Hossein Farhady This simply means that add up all scores (Î£X) and divide it by the number of scores (N). A numerical example may be helpful. Consider the scores of fifteen students who took a language test: 98 97 95 93 90 89 89 84 82 82 78 73 70 60 50 To determine the mean score of the test, add the fifteen scores, that is, Î£X = 1230. __ Then, divide it by N, 15, to give X = 82. (2) The Variance: The variance, represented by the letter V refers to the variation of scores around the mean. Although the formula for computing variance may seem cumbersome, it is not actually difficult. To avoid complexities, the formula will be explained as follows: __ Î£$(X-X)^2$ V = –––– N-1 The formula means to do the following operations: __ 1. Compute the mean (X) 2. Compute the deviation scores by subtracting the mean from each single score __ (X-X). __ 3. Square every deviation score $(X-X)^2$ __ 4. Add up all deviation scores squared Î£$(X-X)^2$ 5. Divide the result of step 4 by N-1 In order to clarify the computational procedures, a numerical example is given below. Consider the scores of ten subjects on a short grammar test: 3, 2, 3, 4, 5, 5, 5, 6, 6, 8. To compute the variance we follow the instructions given before: 1. Compute the mean. __ Î£X 47 X = –– = –– = 4. 7 N 10 2. Compute the deviation scores. 3. Square each deviation score. 22 Test characteristics Hossein Farhady X 3 2 3 4 5 5 5 6 6 8 __ X 4. 7 4. 7 4. 7 4. 7 4. 7 4. 7 4. 7 4. 7 4. 7 4. 7 __ X-X -1. 7 -2. 7 -1. 7 -0. 7 0. 3 0. 3 0. 3 1. 3 1. 3 3. 3 __ $(X-X)^2$ 2. 89 7. 29 2. 89 0. 49 0. 09 0. 09 0. 09 1. 69 1. 69 10. 89 __ 4. Add up all deviation scores squared Î£$(X-X)^2$ = 28. 10 5. Divide the result of step 4 by N-1 __ Î£$(X-X)^2$ 28. 10 V = –––– = ––– = 3. 12 N-1 9 Computing the magnitudes of the mean and variance, we are now ready to put these values in the KR-21 formula and get the reliability of the test scores. The formula is as follows: __ __ K X (K-X) KR-21 = [–––] [1- ––––––] K-1 KV __ In this formula, K refers to the number of items in the test, X represents the mean of test scores, and V is the variance of test scores. Again, a numerical example follows: Suppose a one-hundred-item test is administered to a group of students. The mean and variance computed to be 65 and 100, respectively. Reliability of the scores will be computed using the KR-21 formula: __ __ K X (K-X) KR-21 = [–––] [1- ––––––] K-1 KV 100 65(100-65) KR-21 = [–––] [1- –––––] = 100-1 (100)(100) 100 52275 [–––] [1- –––] = 99 10000 100 100 (–––) (1-. 23) = (–––) (. 77) = . 78 99 99 23 Test characteristics Hossein Farhady The procedure may seem a little complex, but with some practice, it will prove easy and very useful. This method is especially valuable for those who do not have a strong statistical background. From the four methods of estimating reliability, KR-21 method is the most practical and commonly used one. Therefore, it is recommended that teachers and administrators use this method. After covering the first characteristic of a good test, i. e., reliability, and the ways of estimating reliability, the next section is devoted to explaining the second characteristic, namely, validity. Validity Validity is certainly the most important single characteristic of a test. If not valid, even a reliable test does not worth much. The reason is that a reliable test may not be valid; however, a valid test is to some extent reliable as well. Furthermore, where reliability is an independent statistical concept and has nothing to do with the content of the test, validity is directly related to the content and form of the test. In fact, validity is defined as " the extent to which a test measures what it is supposed to measure". This means that if a test is designed to measure examinees’ language ability, it should measure their language ability and nothing else. Otherwise, it will not be a valid test for the purposes intended. As an example, suppose a test of reading comprehension is given to a student and on the basis of his test score, it is claimed that the student is very good at listening comprehension. This kind of interpretation is quite invalid. That particular score can be a valid indication of the student’s reading comprehension ability; however, the same score will be an invalid indication of the same student’s listening comprehension ability. Thus, a test can be valid for one purpose but not the other. In other words, a good test of grammar may be valid for measuring the grammatical ability of the examinees but not for measuring other kinds of abilities. Thus, validity is not an allor-non purpose phenomenon, but a relative one. In order to guarantee that a test is valid, it should be evaluated from different dimensions. Every dimension constitutes a different kind of validity and contributes to the total validity of the test. Among many types of validity, three types — face validity, content validity, and criterion-related validity — are considered important. Each will be discussed briefly. 1. Face Validity: Face validity refers to the extent to which the physical appearance of the test corresponds to what it is claimed to measure. For instance, a test of grammar must contain grammatical items and not vocabulary items. Of course, a test of vocabulary may very well measure grammatical ability as well; however, that type of test will not show high face validity. It should be mentioned that face validity is not a very crucial or determinant type of validity. In most cases, a test that does not show high face validity has proven to be highly valid by other criteria. Therefore, teachers and administrators should not be very much concerned about the face validity of their tests. They should however be careful about other types of validity. 2. Content Validity: Content validity refers to the correspondence between the content of the test and the content of the materials to be tested. Of course, a test 24 Test characteristics Hossein Farhady cannot include all the elements of the content to be tested. Nevertheless, the content of the test should be a reasonable sample and representative of the total content to be tested. In order to determine the content validity of a test, a careful examination of the direct correspondence between the content of the test and the materials to be tested is necessary. This would be possible through scrutinizing the table of specifications explained in the previous article. Although content validity, like face validity, is determined subjectively, it is, however, crucial for the validity of the test. Therefore, subjectivity should not imply insignificance. It is just the only that way the content validity of a test can be determined. 3. Criterion-Related Validity: Criterion-related validity refers to the correspondence between the results of the test in question and the results obtained from an outside criterion. The outside criterion is usually a measurement device for which the validity is already established. In contrast to face validity and content validity, which are determined subjectively, criterion-related validity is established quite objectively. That is why it is often referred to as empirical validity. Criterion-related validity is determined by correlating the scores on a newly developed test with scores on an already-established test. As an example, assume that a new test of language proficiency, called ‘ ROSHD’ is developed by a group of teachers. In this case, the criterion must also be a language proficiency test. Assume further that ‘ TOEFL’ is selected as the outside criterion. In order to determine the criterion-related validity of ‘ ROSHD,’ these two tests should be administered to a group of students and the two sets of scores be correlated. The degree of correlation is the validity index of the ‘ ROSHD’ test validated against ‘ TOEFL.’ It means that to the extent that the two tests correlate, they provide the same information on examinees’ language proficiency. Criterion-related validity is of two major kinds. If a newly developed test is given concurrently with a criterion test, the validity index is called concurrent validity. However, when the two tests are given within a time interval, the correlation between the two sets of scores is called predictive validity. Both concurrent and predictive validity indexes serve the purposes of prediction. The magnitude of correlation predicts the performance of the examinees on the criterion measure from their performance on a newly developed test or vice-versa. The question may arise that how the criterion measure itself is validated. Of course, the answer is that it is validated against another valid test. The question can go back to the very first validated test. In this sense, criterion-related validity is a relative concept. That is, a test is valid in comparison to another test which itself is validated against still another test. No matter how tedious validation procedures might be, it is an inevitable part of test construction process. Test developers must go through the validation process since face validity and content validity are not sufficient indicators for a test to be considered valid. Furthermore, after determining the reliability and validity of a test, the test developer should pay attention to still one more characteristic of a test, namely, practicality. 25 Test characteristics Hossein Farhady Practicality The last characteristic of a good test is practicality. It refers to facilities available to test developers regarding both administration and scoring procedures of a test. As far as administration is concerned, test developers should be attentive to the possibilities of giving a test under reasonably acceptable conditions. For example, suppose a team of experts decide on giving a listening comprehension test to large groups of examinees. In this case, test developers should make sure that facilities such as audio equipments and/or suitable acoustic rooms are available. Otherwise, no matter how reliable and valid the test may be, it will not be practical. Regarding the scoring procedures of a test, one should pay attention to the problem of ease of scoring as well as ease of interpretation of scores. For instance, assume that composition tests are excellent indicators of language ability. Would it be possible to use it in large scale administrations? How would the compositions be scored? How long would it take to score them? All these questions relate to the practicality of the test in terms of scoring. Therefore, test developers should be very careful in selecting and administering a test. The test should be practical, i. e., it should be easy to administer, easy to score, and easy to interpret the scores. Conclusion This article dealt with characteristics of a good test including reliability, validity, and practicality. First, the concepts were defined, and then the methods of estimating reliability and determining different types of validity were discussed. Finally, the concept of practicality was touched upon. The procedures for test construction may seem tedious. However, regardless of the complexity of the tasks in determining the reliability, validity, and practicality of a test, these concepts are indispensable parts of test construction. It means that in order to have an acceptable and defendable test, upon which reasonably sound decisions can be made, test developers should go through planning, preparing, reviewing, and pretesting processes. Furthermore, both item and test characteristics should be determined. On the basis of the results of pre-testing and reviewing, test developers should make all necessary modifications on the test. Only then can a test be used for practical purposes. Only then can one make fairly sound decisions on the lives of people. Without determining these parameters, nobody is ethically allowed to use a test for practical purposes. Otherwise, the test users are bound to make inexcusable mistakes, unreasonable decisions and unrealistic appraisals. Although an attempt was made to avoid technical explanations, content was not sacrificed for oversimplification. The writer’s only hope is that teachers, administrators, and test users assume great responsibility regarding testing procedures. All of us should be aware of the fact that using a test of which the reliability and validity indexes are not known simply equals an academic crime. Making unjustified decisions on the basis of an invalid and/or unreliable test score will have unpredictable but serious consequences upon the lives of the examinees. 26 Test characteristics Hossein Farhady Therefore, all of us must be careful with test construction procedures and more cautious about our judgments at all academic levels, starting from the elementary school up to the highest levels of educational careers. I hope these recommendations will materialize as soon as possible. * This is the revised version of the paper printed in Roshd Foreign Language Teaching Journal (1986). 2 (2 & 3). Tehran, Iran. 27