UDC 821.111(73):82-1:81'367

Shevelova-Harkusha Natalia
Kherson State Maritime Academy

ENGLISH LANGUAGE TESTING: PROBLEMS OF VALIDITY

Summary. This article describes test validity as the main characteristic of any test which enables to measure students' knowledge and is related to the accurate representation of the educational information and the interpretation of test scores. It is also stated about two major types of the test validity: content and construct validity as basic characteristics of the representativeness of the test content and the accuracy of the test measurement and final results. The article highlights items which show how to achieve the content and construct validity of the test. It represents some test statistics elements which can help to identify if the test is valid or not. Among them there are the percentage of average grade of all the marks, the median (middle) grade and standard deviation. The article states five principle criteria of the test among which there are difficulty or facility index of a test item, successfulness, random guess score, the intended weight and standard deviation.

Keywords: content and construct validity, facility index, random guess score, the intended weight, standard deviation.

Шевельова-Гаркуша Н.В.
Херсонська державна морська академія

ТЕСТУВАННЯ З АНГЛІЙСЬКОЇ МОВИ: ПРОБЛЕМИ ВАЛІДНОСТІ

Анотація. Ця стаття описує достовірність тестів як основну характеристику будь-якого тесту, що дає змогу виміряти знання учнів, виміряти, наскільки точним є поданням навчальної інформації та інтерпретація тестових балів. У статті зазначено також два основних типи обгрунтованості тесту: змістова і конструктивна валідність як основні характеристики репрезентативності тестового змісту і точність тестового вимірювання кінцевих результатів. У статті висвітлено питання, які показують, як досягти змістової та конструктивної достовірності тесту. Він представляє деякі елементи тестової статистики, які можуть допомогти визначити, чи є тест дійсним чи ні. Серед них є відсоток середньої оцінки всіх балів, середнє та стандартне відхилення. У статті викладено п'ять принципових критеріїв тесту, серед яких є індекс складності, успішність, бал випадковості, передбачувана вага та стандартне відхилення. У статі виявлено, що тест з високим обґрунтованням елементів буде тісно пов'язаний з передбачуваним фокусом тесту. Для багатьох сертифікаційних тестів це означає, що елементи будуть пов'язані з певною профорієнтацією студентів. Якщо тест має низьку достовірність, він не вимірює вміст, пов'язаний з майбутньою професією студентів та основними компетенціями, які вони повинні мати. Якщо це так, то немає ніякого обґрунтування для використання таких тестів та результатів випробувань за призначенням. Існує кілька способів оцінити дійсність тесту, що включає обґрунтованість контенту, конструктивну валідність, практичність і прогностичну валідність. Для того, щоб встановити, наскільки тест відповідає змістовій валідності необхідно перевірити чи відображає він навчальний план, наданий урядом і навчальним закладом. Учасники тестування повинні вивчати один і той же середній рівень володіння англійською мовою за навчальним планом. Також потрібно встановити, чи включає тест репрезентативний матеріалу, який має охоплювати весь комплекс вивчених одиниць. Це означає, що студенти, які проходять тестування, мають бути ознайомлені з граматикою, лексикою або фонетичним матеріалом, який представлений у тесті. Таким чином, матеріал, який викладався в класі, має відповідати матеріалу, який перевіряється.

Ключові слова: змістова та конструктивна валідність, індекс складності, успішність, бал випадковості, передбачувана вага та стандартне відхилення.

Problem statement. Validity is considered to be of great importance in language testing, and therefore, remains the central concept to all designs and research activities in the field of testing and assessment. Arguably, all researches in language testing are in some senses about validity and the process of validation. In this regard, it is the intent of the present research to investigate the validity of the English language tests. The research questions addressed concern finding out whether the tests are valid in terms of content and construct. The tests administered at this level are 'achievement tests', designed to measure the extent of learning in a prescribed content domain in accordance with explicitly stated objectives of a learning program [1, p. 67].

The problem of validity lies in difficulty of identification how well and accurately a test measures students' acquired abilities and knowledge up to its claims.

Recent research and publications. The problem of test validity was investigated by different Ukrainian and foreign scholars, mainly Bachman L.F., Canale M., Cronbach L.J., Moskal B.M., Leydens J.A., Palmer A.S., Swain M. and many others.

The purpose of the article is to describe basic requirements for English testing composition and state major criteria how to identify test validity.

The objective of the study is, therefore, to examine how far the course objectives are reflected in the contents of the existing tests. Secondly, the study makes an assessment of how well these tests measure the abilities they are intended to measure. The findings reveal a great mismatch between what the tests aim at testing and what they actually test. A wide gap is found between the curriculum goals and the existing test format. The study also finds that the Higher Secondary language tests are largely unable to measure the constructs they are based on. The key recommendations to increase the content and construct validity of these tests include developing test specifications and designing syllabus in accordance with course objectives, using di-

rect tests and authentic tasks, sampling widely and unpredictably, arranging training programs for the language teachers, etc. [5, p. 562].

Language tests set out to measure specific abilities, for example, listening skills or knowledge of vocabulary. We want variation in test scores to be linked to variation in test taker ability, and for the test to distribute candidates as far and as widely as possible with the lowest ability candidate receiving the lowest score and highest ability candidate receiving the highest score. However, factors which are not linked to language ability can affect test scores and are therefore sources of measurement error. These factors might be linked to the test itself such as test methods, differences in the different forms of the test or differences in rater behaviour. They may be linked to the test conditions, for example, administrative procedures or time of day. Or they may be linked to test taker characteristics unrelated to language proficiency such as age, first language and extent of subject matter knowledge. While it is accepted that some measurement error is inevitable, test developers seek to minimize measurement error in the design of tests so that variations in scores match variation in candidate ability as closely as possible [6, p. 244].

Test Validity is the extent of how well and accurately a test measures students' acquired abilities and knowledge up to its claims. The Test Validity is the main characteristic of the test which enables to measure students' knowledge and which is related to the accurate representation of the educational information and the interpretation of test scores [3, p. 115].

Language tests require context. Reading and listening comprehension tests require written and spoken 'texts' for candidates to process and respond to, and speaking tests need to present audio, textual and/or visual prompts in order to elicit a speech sample. There are two major types of the test validity: content and construct validity.

Content validity is a characteristic of the representativeness of the test content. It means that this type of validity depends on what the test contains. A test has content validity built into it by careful selection of which items to include.

Construct validity is a characteristic of the accuracy of the test measurement and final results based on the structural or construct criterion. It is closely associated with the reliability and stability to the fact of being affected by occasional formal factors which can mitigate the test validity [2, p. 35].

To achieve the content validity the tested items must:

1) adequately reflect the **curriculum** provided by the government and the educational institution. Test-takers must be taught the same average level of English with the curriculum;

2) belong to the representative material that should be covered comprehensively. It means that examinees who are tested are acquainted with grammar, vocabulary or phonetic material which has been taught to the students at the lessons. Thus, the material that was taught in a class matches the material that is tested;

3) pertain to the active content, i.e. vocabulary and grammar. Active material comprises rules and words that learners understand and use in speaking or writing, whereas passive content refers to rules and words that learners understand but are not yet able to use. The use of tested active items makes the test more valid.

To attain the construct validity:

1) the tested items must have clear, short and unambiguous instructions to let students easily understand the task and not waste much time rereading it;

2) the questions or task must be short, equal in length and have no excessive information to distract students attention from the principle one;

3) tests must have more tasks with sufficient quantity of open answers along with multiple choice to decrease the reliability of the test because of random guesses;

4) tasks with multiple choice must have alternatives mutually exclusive not to perplex the students, if there several answers to choose it must be clearly stated in the instruction [1, p. 67];

5) the test must be rather long to enhance its reliability. The test with 20-35 tasks are considered being short, while the tests with more than 100 tasks tend to be pretty long and undesirable, because the longer test is the more mistakes can be done by students on account of psychological factor (fatigue, weariness and loss of motivation). The optimum quantity of tasks in tests tends to be around 40-60 to make it more reliable [3, p. 102].

Test validity is deduced from the correlation between the testees' results (successful performance of the test or the test failure) and the outer criteria of the test.

Validity is generally considered the most important issue in educational testing because it concerns the meaning placed on test results. Though many textbooks present validity as a static construct, various models of validity have evolved since the first published recommendations for constructing education tests [4, p. 192].

Test validity can itself be tested/validated using tests of inter-rater reliability, intra-rater reliability, repeatability(test-retest reliability), and other traits, usually via multiple runs of the test whose results are compared. Statistical analysis helps determine whether the differences between the various results either are large enough to be a problem or are acceptably small.

The successful performance of a test shows how many students have completed it successfully. All tasks are differentiated according to the percentage of students' successful passing or performance of tasks. To count this percentage the amount of students must be multiplied by 100 and divided by the total amount of students [4, p. 3].

Thus, the range between 100%-60% of performed tasks indicates the test being successfully passed; the range from 59% to 0% denotes that the test is very complicated with extremely sophisticated tasks which were not passed successfully because of different reasons. The tasks with marginal results should be deleted or insistently recommended to be remediated (altered).

To check if the test is passed successfully the test statistics elements can be investigated:

• ***the average grade*** of all the marks must be within the range of 50-75%;

• ***the median (middle) grade*** – a middle point between the highest and the lowest score.

• ***standard deviation*** – between 12-18%; if the percentage is less, the score are too bunched up, that is almost all students are passed or failed the test [3, p. 74].

Validity is arguably the most important criteria for the quality of a test. A validity scale, in psychological testing, is a scale used in an attempt to measure reliability of responses, for example with the goal of detecting defensiveness, malingering, or careless or random responding. On a test with high validity the items will be closely linked to the test's intended focus.

For many certification and licensure tests this means that the items will be highly related to a specific job or occupation. If a test has poor validity then it does not measure the job-related content and competencies it ought to. When this is the case, there is no justification for using the test results for their intended purpose. There are several ways to estimate the validity of a test including content validity, concurrent validity, and predictive validity. The face validity of a test is sometimes also mentioned [3, p. 121].

There are 5 principle criteria of the test:

1. Difficulty / Facility index of a test item is the average score on the items, expressed as a percentage. This percentage designates a group of testees that chooses the correct response. It is contingent on the type of knowledge being tested by a particular item and the intellectual skill demanded. The item difficulty index ranges from 0 to 100; the higher the value, the easier the question. The task is regarded being "easy" if the index is 85% or above; "moderate" (medium) if it is between 51 and 84%; and "hard" if it is 50% or below. The level of simplicity of all tasks can easily be checked in the tables with students' results of each definite test at Moodle Platform [4, p. 193].

2. Successfulness. The previous characteristic is closely associated with the criterion of success. The successful performance of the tasks shows how many students answered the question or task successfully. All tasks can be differentiated according to the percentage of successful passing or performance of tasks. Thus, the range between 100%-90% of performed tasks indicates items being very simple which were answered by almost all students; 89%-66% – points out simple tasks; 65%-35% – specifies the tasks of moderate simplicity; 34%-11% – denotes difficult questions; 10%-6% – itemizes very complicated tasks; 5%-0% – singles out extremely sophisticated tasks. The tasks with marginal results should be deleted or insistently recommended to be altered [2, p. 34].

Thus, to augment test validity the test must:

1. reflect the curriculum of the educational institution;
2. comprise the representative material taught by students at the lessons;
3. be based on the active content, related to the topics;
4. have mutually exclusive alternatives;
5. include options pertaining to the same topic of approximately the same difficulty.
6. involve short, equal in length instructions and the alternatives with no excessive information;
7. be of the sufficient amplitude (dimension) with around 40-60 tasks;
8. be presented with the gradual augmentation of the difficulty level, from simple tasks to more sophisticated.

3. Random guess score shows the likelihood (probability) of the correct answer to the question by means of guess when students give the answers randomly. This criterion is associated with the reliability of the test. The lower percentage the more reliable the test or task is. For instance, open questions usually have *lower index* which means that they are *more* reliable and less vulnerable to random guesses, while tasks with the *highest percentage* of random guess scores designates that the task is weak and *unreliable* because of the ability to guess. Thus, tasks that use some form of multiple choice and Yes/No questions tend to be of low reliability with a high percentage of random guess scores. These tasks mustn't be numerous in the test. The results from 100%-70% shows that the tasks are answered [3, p. 132].

Thus, to augment test validity the test must:

1. be of different types with sufficient quantity of open answers along with multiple choice tasks;
2. provide students with at least three or more options in each multiple choice task.

4. The intended weight is a weight of a task which is expressed in percentage from the whole mark of the test. If all tasks are estimated in an equal way (1/2 scores), the intended weight will be the same for all of the tasks. In case if more difficult tasks are marked by higher intended scores, then different tasks will have different estimation (measurement) – different intended weight. The less the intended weight of the task is, the more simple it is. It is usually compared to the ***effective weight*** of the tasks, which designates the effectiveness of the questions according to the facility index [6, p. 256].

Thus, to augment test validity the test must:

1. be valued for the performance of all the options;
2. be valued by the same number of scores.

5. Standard deviation is a measure of the dispersion of student scores on that item. That is, it indicates how "spread out" the responses were. The item standard deviation is most meaningful when comparing items which have more than one correct alternative and when scale scoring is used. For this reason it is not typically used to evaluate classroom tests.

Reliability is one of the most important elements of test quality. It has to do with the consistency, or reproducibility, or an examinee's performance on the test. For example, if you were to administer a test with high reliability to an examinee on two occasions, you would be very likely to reach the same conclusions about the examinee's performance both times. A test with poor reliability, on the other hand, might result in very different scores for the examinee across the two test administrations. If a test yields inconsistent scores, it may be unethical to take any substantive actions on the basis of the test. There are several methods for computing test reliability including test-retest reliability, parallel forms reliability, decision consistency, internal consistency, and interrater reliability. For many criterion-referenced tests decision consistency is often an appropriate choice [4, p. 194].

Average inter-item correlation is a subtype of internal consistency reliability. It is obtained by taking all of the items on a test that probe the same construct (e.g., reading comprehension), determin-

ing the correlation coefficient for each pair of items, and finally taking the average of all of these correlation coefficients. This final step yields the average inter-item correlation [1, p. 72].

Thus, validity is the main characteristic of any test which enables to measure students' knowledge and is related to the accurate representation of the educational information and the interpretation of test scores. There are two major types of the test validity: content and construct validity which should be constantly verified. To check test validity more efficiently it is necessary to make sure your goals and objectives are clearly defined and operationalized. Expectations of students should be written down. It also needs to match the assessment measure to the goals and objectives. Additionally, the teacher should have the test reviewed by faculty at other schools to obtain feedback from an outside party who is less invested in the instrument. It is necessary to get students involved; have the students look over the assessment for troublesome wording, or other difficulties. If possible, the teacher should compare his/her measure with other measures, or data that may be available.

References:

1. Bachman L.F., Palmer A.S. (1981). The construct validation of the FSI oral interview. *Language Learning*, vol. 31(1), pp. 67–86.
2. Canale M., Swain M. (1990). Theoretical Basis of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, vol. 1, pp. 1–47.
3. Cronbach L.J. (1991). Test validation. *Educational Measurement,* 2nd ed., 443 p.
4. Moskal B.M., Leydens J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation,* vol. 7(10), pp. 192–205.
5. Brier J. (1997). Cultural Understanding through Cross-Cultural Analysis. *Pathways to Culture : Readings on Teaching Culture in the Foreign Language Class*, vol. 21, pp. 561–569.
6. Zhongliang C. (2010). On the Applications of Modern Educational Technology in Maritime English Teaching from the Perspective of Constructivism. *English Language Teaching*, vol. 3, No. 3, pp. 244–248.