On the issue of validity

Onik Mikayelyan, Sona Mikayelyan
Head of the Department of Assessment
National Institute of Education of RA, Yerevan
Office Address: 67 Tigran Mets street, Yerevan, RA
Mob.: +374 99 90 14 14
e-mail: mikons51@yahoo.com

Abstract: Assessment is a continuous process in the sphere of education. It has high stake influence on the future development of education. That's why the preciseness and fairness of assessment have become important factors. Mistakes in assessment, especially systematic errors, have bad impact on the students' performance particularly and on the assessment system in general.

The notion of validity is a valuable means for avoiding systematic errors in assessment. Use of validity as an important theoretical concept in the educational assessment process will give a chance to apply the assessment tools more correctly.

The importance of any theoretical concept is measured according to the extent its application helps to improve the quality of education. The effect of validity will become greater, in case it is possible to give a quantitative expression to it. This is the aim of this paper in which an attempt is made to prove the possibility of giving quantitative expression to the validity of a test.
Key words: validity, quality, quantity, practical, weight.

1. A Cognitive Approach to Validity

Validity is one of the characteristics of a human activity or of an event. As in reality reliability is related to consistency, validity is related to something being right or wrong.

Validity is an important and not fully explored concept. So, professionals keep exploring it. There are articles, manuals about modern approaches to validity. Here we are guided by the classical conception of validity.

In general, qualitative characteristic, as validity is, has its peculiar difficulty to be measured as a quantitative characteristic in terms of its degree of trueness.

Validity is an important theoretical concept. If the validity of a phenomenon doesn't receive quantity expression, it won't be possible to compare two similar phenomena or objects of the educational system in the sense of validity (Kane, M.T.2006. Messick, S. 1989, 1995. Stephen G. Sireci. 2007).

In general, without having quantity expression of validity it is impossible to make sound judgments about an event itself, not in comparison with other events. Therefore the idea of validity without quantity expression will not work purposefully and will be less usable also in education (Guion, R. M. 1980). Meanwhile it is obvious that the idea of validity is natural and if it is expressed in quantity, it will be applied both in life and in educational sphere widely (Cronbach, L. J. 1969).

Let's look at the following example where validity is used in life – validity of the human leg. The leg is considered fully valid if it is healthy and complete, i.e., it completely serves its purpose. The leg is fully invalid if it doesn't exist. The leg consists of toes, foot, hip, shin and muscles. If one lacks a toe, the other foot, the third lacks the hip or they exist but don't function properly, all these are the invalid forms of the leg or they are valid to some extent. It is possible to form general or particular scales for a leg which will categorize the degrees of defects and ascribe quantified values. Using this we can compare the validity of different defects of the legs or the category of validity of a leg.

To the extent to which it is possible to prove that this or that event or activity is true or false, it is possible to speak also about its validity. The clock can go right or wrong. Wrongness can be of different degrees, for example, one clock can be wrong for 1 second another can be wrong for a minute per day, etc. In this case the degree of error is not difficult to define but it is important because the further application of the clock depends on that factor (Wilson, N. 1997).

In the sphere of mental activity or education we will make an attempt to find mechanisms of similar judgments, for example, to compare validity of different tests compiled for the same aim or to define the degree (extent) of validity of a test or an item (Popham, W. J. 2008).

We cannot see the validity of an item or a task clearly as in case of physical values (Jo-Anne Baird 2010). If we divide a test or a task into parts according to the standards' requirements similar to the case of legs, it will be possible to make scales of weights or quantified expressions for each part. As doctors decide to assign a person disability pension depending on the degree of invalidity of a leg, the experts of testing, having the scale of the item's and the test validity, can make a decision about the test applicability or about refining the test if it is necessary. (Simon Wolming 2010).

Similarly, in the educational sphere if we define validity before applying a test it will rather be theoretical validity (Pamela A. Moss.1992). If the validity is defined by using the test results or the marks given by the teacher or pupils' opinion about face validity this will be less theoretical validity. The analysis of the practical validity can have impact on the decrease of standard errors and improvement of educational standards. Besides being used in analyses, practical validity is important for improvement and further usage of the test.

2. Validity of a Test to be Used in the Educational Process

Our goal here is to discuss the possibility of measuring validity (Cunningham, G. K. 1998, Michael K. Russell, Peter W. Airasian 2012) of a test used in the educational process.

If, coming out from the goals, for a test and its items a scrupulous and qualified (skilled) scale is constructed and weights are ascribed to its steps, then it will be possible to decide the extent to which each item and therefore the whole test serves its goal. Thus the subjectivity in the process of defining validity will be diminished essentially. The different tests (items) that are made for the same goal but by different test designers can both be valid, but if the goal is scrupulously scaled, then it will be possible to ascribe weights to the tests (items) constructed by different test designers and hence it will be possible to compare their degrees of validity. If we define a number expression of validity for each item of a test, then the arithmetical mean of these numbers can be treated as the test validity. Or, we can use another combination of those numbers depending on our goal. Another

approach for counting test validity, by assigning weights for each item, can be used in case if our aim is to define validity with more precision. The number corresponding to each item's validity belongs to the segment [0; 1], (Croker, L. Algina J. 2008).

For the tests that are used in the educational process the following types of validity are important: curricular validity, content validity, face validity and instructional validity (Payne, D. A. 2003a., 2003b.).

Test validity can be defined either before giving the test, or after giving it to students. The decision depends on further usage of the test, i.e., whether the test must be used more than one time.

By having a test, its aim and specifications of tasks, an independent expert checks the quality of tasks and the extent to which the tasks serve the aim. Unfortunately, this process has a great deal of subjectivity as it depends on the professional quality, taste of the expert and other circumstances.

However, it is not only the grade given by an expert is taken into account, but also other components (in our case, there are three components). By means of blending estimates of validity from different sources, the final validity becomes as objective as possible for that subjective concept.

2.1. Practical validity

The concept of validity is getting more subtly evaluated depending on the degree of importance of assessment results' accuracy in particular community (Christina Wikström 2010). However, there is a wide gap between theoretical validity and its practical use. It is a serious problem, which needs a solution. There are many trials to find metodology for practical usage of validity. Modern approaches of finding the quantitative expression of validity sometimes are obtained by such a complicated means, that they lose their chances to be used in practice. We discuss validity of a test which is used in the educational process. Validity considered from this point of view can impact on the improvement of the standards, curriculum and the methods of instruction. Hence, the value of this validity is great.

Validity is called practical validity if it becomes measurable by use of a method.

For putting validity into practical use, it is important to mention for which concrete purpose the validity is considered (Paul Black. 2010).

The degree to which a test serves the anticipated goal is decided by an independent expert. Although the concept of validity is an absolute concept, practically it is subjective and hence relative. As there is no objective means to decide the degree of validity, it depends on the personal characteristics of the expert who defines the validity of the test.

The role of assessment validity is very crucial in improving the quality and effectiveness of education in terms of curriculum and educational standards. The standards are requirements to meet by the teachers and students to provide progress in education.

To use practical validity we suggest to keep back from the theory in favour of its application in practice as much as possible, which is more valuable. With this consideration in mind, and with the aim to improve and further use the test, it is possible to involve teachers, as subject specialists to measure content validity and students – to measure face validity. In order to check pupils' mastery of any theme, the tester, the specialist defining the content validity and the teacher are led by the subject standards, curriculum and the text-book (Cronbach, L. J., & Meehl, P. E. 1955. Guion, R. M. 1977). In the process of defining practical validity it is purposeful to take into account teacher's opinion, as the teacher is the professional who knows what the achievements of the pupils are. It means that the marks given to the test written by an independent specialist should be compared with the marks given by the teacher for the same theme. If the test is for one-time use, the independent specialist defines face validity before applying the test. If the test is for further use it is preferable to have the test-takers' feedback. Test takers give their opinion about each item of the test: to what extent the item is understandable.

A method for finding practical validity and expressing it by numbers is described here. The concept of validity used in practice this way will contribute to the high quality of education in which we were conceived in the result of an experiment.

2.2. An Experiment for Defining Practical Validity

In this example, the test validity is defined taking into account the fact that the test will be used further, so the validity is measured after giving the test to students and using the results of assessment.

In Yerevan, Armenia, there was a piloting in Mathematics with participation of 274 students from 10 schools. The aim of the piloting was to define the degree of content validity of the tests which were composed to check the mastery and application of the theme “Polynomial of one variable”.

Two specialists developed tests for the same aim independently. The time allotted to tests was 45 minutes.

Subjectivity was obvious just in that stage. A test developed by one specialist consisted of 7 tasks; the other test developed by the second specialist consisted of 11 tasks.

An independent expert or a teacher having the standards for this theme, its content, the test and knowing the aim of the test and the scale of its assessment, defines the validity (relevance) of each item and for the whole test in the [0; 1] interval. The expert writes his or her professional opinion. These activities are implemented until piloting.

If it is necessary the aim of the test and the quality of the tasks can also be discussed in this stage. In the result, some parts can be modified which should be discussed with the tester. If any, a new opinion should be written (Wiggins, G. 1992).

The answers to the question ‘To what extent do the tests correspond to the aim?’ were 0.85 for one and 0.74 for other test. So, the validities of both tests vary.

The correlation of the marks given to pupils for the test and the marks given by the teacher for the same theme is 0.621. This number shows the degree to which the test scores correlate with students’ grades, or concurrent validity (David Mott. 2008). A special approach is needed in case, when the correlation is negative. These cases must be discussed separately.

So in order to define practical validity of the test we got three coefficients corresponding to three components.

1. The coefficients given by the expert according to the variants are $e_1 = 0.85$, $e_2 = 0.74$.

2. The coefficients of the correlation of the teacher's marks and the test scores are approximately the same, so we took them equal $r = 0.62$.

3. The degree of the test's face validity according to test takers' feedback, which must be in the segment [0;1], are $f_1 = 0.8$, $f_2 = 0.7$.

Studying the results it should be noted that in two tests, the results in three components go with each other. This gives us a ground to rely on the usage of three components in defining test validity. If the corresponding components differ, some analysis should be done and consequences should be driven.

Coming out of the importance and status of these three components, weights should be defined. For the components conditionally α_1, α_2 and α_3 weighting coefficients are suggested correspondingly, so that $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The most valuable is the coefficient given by the expert, so its weight is the highest, the next place is for teacher's coefficient, and the face validity has the least weight. In this experiment we ascribed $\alpha_1 = 0,6$, $\alpha_2 = 0,3$, $\alpha_3 = 0,1$.

The final validity is defined by the following formula

$$v = \alpha_1 \cdot e + \alpha_2 \cdot r + \alpha_3 \cdot f \ .$$

By means of this formula and our data, validities of the two tests were counted: $v_1 = 0,776$ and $v_2 = 0,7$. As it can be seen, the degrees of validity given by the experts are correspondingly higher than the final practical validities of the tests. This is because of grades given by the teacher, which are much higher than pupils' grades for the test in reality.

Valuable inferences are made based on the results we got from this experiment (Popham, W.J. 2003).

So the quantitative expression of validity of a test can be defined both before and after (practical validity) using the test. In both cases the main issue is scaling of the goal of the test and of each item precisely and correctly. In this paper an attempt is made to represent a method for finding a test's practical validity. However it may be not the unique method for that purpose (Angoff, W. H.1988. Cronbach, L. J. 1969, 1971. Friedman, S. J., Frisbie, D. A. 1993).

The evidence of the knowledge that has been taught and skills is important in the instructional/teaching process. Considering its importance, it is necessary to include a separate component, which shows to what extent the test items are checking the evidence. By means of that component it is grounded how the student is capable or competent to apply what has been taught. In practice, it is more rational to apply this component with end-of-semester test or a wider comprehensive test. This component refers to every single task and at last to the overall test. The degree of evidence of the task is defined by an expert by ascribing a number from [0; 1] interval – the number of the sector that, as the expert thinks, shows the degree of evidence (which can be e.g. 0,7), and the evidence of the overall test – as their average. It is a better performance indicator of test purpose and the requirements set up in standards.

This requirement to the test validation will have a positive backwash effect on designing a task in particular and on learning and raising standards in general

The final validity is defined by the following formula

$$v = \alpha_1 \cdot e + \alpha_2 \cdot r + \alpha_3 \cdot f + \alpha_4 \cdot ev .$$

For the components conditionally α_1, α_2, α_3 and α_4 weighting coefficients are suggested correspondingly, so that $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$.

Here *ev* means evidence, α_4 - weighting coefficient of evidence. The coefficient values are defined according to the degree of importance.

Example. For the item which provides direct proof of the truth of an assertion ev=1. In the case of open question ev=0.

Such a mechanism of application of validity will encourage test writers to choose each item of the test more scrupulously so that it corresponds to the aim of testing and to the educational standards.

Reference

1. Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.

2. Christina Wikström. (Jan 1, 2010) The concept of validity in theory and practice. Assessment in Education: Principles, Policy & Practice
3. Croker, L. Algina J. (2008), Introduction to Classical and Modern Test Theory, Cengage Learning.
4. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.
5. Cronbach, L. J. (1969). Validation of educational measures. *Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton*, NJ: Educational Testing Service, 35-52.
6. Cronbach, L. J.(1971) Test Validation.
7. Cunningham, G. K. (1998), *Assessment in the Classroom: Constructing and Interpreting Texts*, London-Washington, D.C., The Falmer Press.
8. David Mott. (2008), Test Validity Revisited Again. pps, www.vatd.org/
9. Friedman, S. J., and Frisbie, D. A. (1993). The validity of report cards as indicators of student performance. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.
10. Guion, R. M. (1977). Content validity–The source of my discontent. *Applied Psychological Measurement, 1*, 1-10.
11. Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 385-398.
12. Jo-Anne Baird. (2010), Editorial,The theory-practice gap. Assessment in Education: Principles, Policy & Practice. Vol. 17, No. 2,May, 113-116.
13. Kane, M.T. (2006). " Validation " . In *Educational measurement,* , 4th ed. , Edited by: Brennan, R.L. 17 – 64 . Westport, CT : American Council on Education/Praeger
14. Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 634-694.
15. Messick, S. (1989) 'Validity', in LINN, R.L. (ed.) *Educational Measurement,* (3rd Ed.), New York: Macmillan, pp. 13–104.
16. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.
17. Michael K. Russell, Peter W. Airasian. (2012). Classroom Assessment, McGraw-Hill, NY.,.

18. Pamela A. Moss. (1992). Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment // Review of Educational Research, Vol. 62, No. 3. pp. 229-258.
19. Paul Black. (2010), Validity in teachers' summative assessments, Assessment in Education: Principles, Policy & Practice. Volume 17, Issue 2.
20. Payne, D. A. (2003a.), *Applied Educational Assessment*, Canada, Wadsworth Group.
21. Payne, D. A. (2003b.), *Instructor's Manual for Applied Educational Assessment*, Canada, Wadsworth Group.
22. Popham, W.J. (2003), *Test Better, Teach Better: The Instructional Role of Assessment*, USA, ASCD.
23. Popham, W. J. (2008). All About Assessment/A Misunderstood Grail. *Educational Leadership, 66*(1), 82-83.
24. Simon Wolming. (2010),The concept of validity in theory and practice Assessment in Education: Principles, Policy & Practice. Volume 17, Issue 2.
25. Stephen G. Sireci. (2007), On Validity Theory and Test Validation. Educational Researcher, Vol. 36, No. 8, pp. 477–481
26. Wiggins, G. (1992). Creating tests worth taking. Educational Leadership, 44(8), 26-33.
27. Wilson, N. (1997) Educational standards and the problem of error. Education Policy Analysis Archives, Vol 6 No 10.