Science Journal of Education
2023; 11(4): 142-149
http://www.sciencepublishinggroup.com/j/sjedu
doi: 10.11648/j.sjedu.20231104.14
ISSN: 2329-0900 (Print); ISSN: 2329-0897 (Online)

Validity Criteria of a Standardized Test as an Opportunity for Efficient Assessment Created from the Teleological Perspective of Incremental Learning

Geanina Havârneanu

Department of Education Sciences, Faculty of Psychology and Educational Sciences, Alexandru Ioan Cuza University, Iasi, Romania

Email address:
geanina.havarneanu@uaic.ro

To cite this article:
Geanina Havârneanu. Validity Criteria of a Standardized Test as an Opportunity for Efficient Assessment Created from the Teleological Perspective of Incremental Learning. *Science Journal of Education*. Vol. 11, No. 4, 2023, pp. 142-149. doi: 10.11648/j.sjedu.20231104.14

Received: July 1, 2023; **Accepted**: July 18, 2023; **Published**: July 26, 2023

Abstract: This study aims to comprehensively present ways to determine the validity of a standardized test, highlighting essential criteria that can be widely used. Incremental learning, as conceived by Minsky, represents the construction of knowledge focused on the multitude of interactions between the cognitive components (which build inferences in a decision-making process) or metacognitive components (which support the solving and decision-making process); it requires a special type of feedback (motivational, affective, behavioral, cognitive) through a standardized assessment. Evaluation through standardized tools has unique requirements that are imposed to be able to provide feedback as eloquently and correctly as possible at the same time. One of the particular mandatory criteria is testing the validity of the test, with all its components: internal validity, external validity, content validity (highlighted by the value of the content validity coefficient and the value of the concordance coefficient), criterion validity (with its components competitive validity and predictive validity), and construct validity (which involves both a theoretical and an empirical approach). The examples built in this work illustrate how to calculate the concordance coefficient, Kendall's coefficient, respectively Cohen's coefficient for a set of results obtained by a group of students who two or more evaluators evaluated.

Keywords: Incremental Learning, Content Validity Coefficient, Inter-Evaluator Agreement Coefficient, Concordance Coefficient, Kendall Coefficient, K Cohen Coefficient

1. Introduction

The noematic correlation, in the Husserlian [19] phenomenological sense, between educational communication, which involves teaching or learning as a self-instructive process, and assessment, which operationalizes the level of skills, realizes diagnoses, makes predictions on student performance, constitutes an incremental process. Incremental learning focuses on the emerging philosophy [18] of organizing educational situations whose instructional design capitalizes on effective ways of reactualization, using operationalized structuring in microworlds of declarative, procedurals, conditionally, and strategic knowledge [30], in problem spaces [34], as well as in new, interconnected linguistic structures [38]. In Minsky's theory of incremental learning [28], learning construction is based on the diversity of interactions between sensations, perceptions, representations, knowledge, and plausible inferences that act as a functional network of assimilated symbolic information. The advantage of incremental learning is that it ensures a continuous flow of new knowledge in direct connection with the existing ones, constituting self-organized maps and building meta-heuristics [14], allowing the development of multidisciplinary, interdisciplinary, and transdisciplinary skills. The construction of meaning (at the syntactic, semantic, thematic, narrative level) and of understanding requires the ability to conceive, operationalize and use strategies for processing and integrating new data into the operative cognitive network that is constantly being restructured. Given that in communication, the speaker neglects certain information considered non-essential or assumed to be known, the conceptual processing tries to fill in these omissions, using his proper cognitive network at all its levels (recognition, understanding, application, analysis, synthesis, evaluation, creation [3, 2].) This instantiation, conceptual materialization, is often corrected by the concrete data of each situation because, sometimes, the information

received can lead to preconceived ideas, mainly due to the socio-emotional-cognitive background but also to the functional fixity [44], which predisposes to data interpretation mainly from the point of view of the denotative characteristics [10], and less of the implicit, insufficiently observable ones.

Currently, incremental learning is closely related to online learning, where advanced learning algorithms perform practical cognitive training that multiplies the chances of solving specific problem situations, significantly improving the accuracy and efficiency of the decision [6].

Standardized testing is the usual way to implement an assessment that promotes incremental learning when certain criteria are met. The first criterion is the relevance of the inferences that lead to a particular result by applying the test. Another criterion is the analysis of how the test was applied (ethical conditions, compliance with the norms in force, which offer the same opportunities to all participants, regardless of individual differences or particularities). Another important criterion is the efficiency of the methods by which the answers to the proposed test items were interpreted (according to the correction and scoring scale).

2. Literature Review

Learning assessment policies within educational institutions have been studied by Kane [20]. He identified two specific categories of evaluation organization. The first category highlights the institutional means of evaluating learning through empirical evidence – the ECD (Evidence-Centered Design) model conceived by Mislevy and colleagues [29-31]. The second category promotes institutional means of evaluating learning through structured models of interpretive analysis based on theoretical rationales – the IA (Interpretive Analysis) model, which was conceptualized by Kane [20].

A pertinent analysis of the ECD model [39] states that the assessment design has a first stage, that of identifying what should be assessed in terms of the competencies acquired by the student and the performance to be achieved, and a second stage, that of establishing the types of items that can determine the acquisition of specific behaviors and the achievement of the level of performance required. ECD-type design and assessment ensure the development of specific expertise in the field of study through collaborative learning, co-distributed expertise, and complex problems solving [36], which allow the assessment of higher cognitive levels from Bloom's taxonomy [2, 3].

Kane's IA model [20] represents a methodology that can be used to validate interpretive arguments in learning assessment, being summarized in the following table.

Table 1. *The argument for the interpretation of the level of acquisition of competence in relation to the score obtained by the student (adapted from Kane [20], p. 34).*

I1: Scoring (correlation performance - score awarded)	A1.1. Appropriate design of rules for scoring items and establishing criteria for passing the exam. A1.2. The application of the rules for scoring the items according to the specifications and establishing the criteria for passing the exam. A1.3. Unbiased scoring of items and objective establishment of exam passing criteria. A1.4. Adapting the measurement models involved in scoring the items and establishing the criteria for passing the exam.
I2: Generalization (starting from the particularities of the observed scoring to a universally valid scoring)	A2.1. Making a representative sample of observations to be able to generalize the scoring of the items. A2.2. Making a sufficiently large number of observations to control the random errors in scoring the items.
I3: Extrapolation (starting from the universally valid scoring to the specific scoring of the designed test instrument)	A3.1. The association of universally valid scoring with the specific scoring of the designed test instrument. A3.2. Elimination of systematic errors likely to affect extrapolation.
I4: Interpretation (starting from the specific scoring of the designed test instrument to the verbal description of the interpretation of the results as the level of acquisition of the targeted competence)	A4.1. The suitability at the theoretical status of the implications regarding the association of the obtained score with the level of acquisition of the targeted competence. A4.2. Empirical support of the implications regarding the association of the obtained score with the level of acquisition of the targeted competence

3. Data and Methodology

The design of standardized evaluation tests to be implemented involves checking the validity, reliability, replicability, and objectivity standards.

Essentially, validity depends more on how the test is applied than on the test itself, as all aspects and details of a measurement procedure can influence performance and, thus, what is being measured [9]. This is the main reason why validity conditions in a standardized test are necessary.

4. Validity of a Standardized Test

The validity of a test is the process of determining the extent to which descriptive, explanatory, or predictive interpretations lead to inferences based on its scores [40] obtained by applying appropriate instruments to determine the level of an aptitude or the acquisition of a skill. Messick [27] defined validity as the evaluatively integrated judgment of the degree to which empirical and rational theoretical evidence supports the accuracy and relevance of inferences based on the results of an evaluation. In educational testing, validity refers to the extent to which theory supported by empirical evidence demonstrates that interpretations of test results support their intended uses.

The validity of a test captures two facets of the same reality [4]:

1. Internal validity;
2. External validity.

Validity is a multi-vector construct [1, 25, 15] for the

establishment of which several strategies are designed to validate inferences made based on test scores [42] to assess the appropriateness of using the proposed test to measure a characteristic or competence. The components of validity are:

Content validity (highlighted by the value of the content validity coefficient and the value of the concordance coefficient) aims to verify declarative, procedural, strategic, and conditional knowledge [11, 41].

Criterion validity (with its competitive and predictive validity components) aims to test skills [9, 17, 41, 13].

Construct validity (which involves both a theoretical and an empirical approach) aims to verify the relevance of the transposition of a concept into an evaluative tool [5, 1, 8, 41].

4.1. The Internal Validity of a Test

The internal validity of a test [43] examines the extent to which inferences built based on test scores discriminate the actual existence of a causal relationship between the independent and dependent variables without the intervention of other factors.

Internal validity can be distorted by many causes, such as the expectations of the researcher (he may convey information or notes that suggest certain answers to students); students' expectations (they will try to give the answer they consider generally accepted, even if they have a different opinion); the subjective fluctuation of the measured values (the researcher may have fluctuations in the spirit of observation due to subjective or objective causes, and students may respond differently as a result of changes in their mental state or abilities, through learning or practice following the repetition of the assessment tool); the application to students of a different instructive-educational treatment for subjective reasons (the use of other records, the reduction of the number of students in a group through absenteeism at certain phases of the educational project or even transfer, abandonment or the partial application of the scheme to one of the subgroups); favorable sensitization to repeated testing or, on the contrary, the inoculation of test-refractory attitudes, statistical regression (as a result of a low randomization of student groups, or of using a selection based on high scores on specific tests); maturation (if the time interval between tests is longer, changes related to the mental evolution of the individual appear); psychophysiological or historical factors (the state of one's own biorhythm, illnesses, successes, disappointments); the interactions of hidden, non-experimental variables that may influence the accuracy of the causal link between the independent and dependent variables.

Internal validity can be improved by controlling extraneous and intermediate variables, using standardized instructions described at the time of test application, and eliminating items with ambiguous wording and effects due to the subjectivity of the investigator.

4.2. The External Validity of a Test

The external validity of a test [43] refers to the extent to which the inferences drawn from the test can be generalized situationally (ecological validity), socially (population validity), and temporally (historical validity).

External validity results can be distorted for various reasons. Ecological validity test values can be altered due to the Hawthorne effect [37], a type of reactivity in which individuals change some aspect of their behavior in response to the awareness of being observed. Social validity testing values may be distorted due to restrictions on the possibility of a random selection of the target group but also due to the more responsible reaction of the target group compared to that of the control group. The historical validity testing values can be distorted by the interference of multiple treatments, which implies the application of several evaluative tools within the same approach, or by a particular order of the items/ tests, which determines the production of physio-psychological effects that facilitate or, on the contrary, hinder the transfer of information.

External validity can be improved by setting experiments in a more natural setting, using a random selection of participants to generate a representative sample, and making subgroups to apply tests specific to different characteristics of the independent variable.

4.3. Content Validity

Content validity reflects the extent to which the objective, semi-objective, subjective, or performance skills items proposed in a test verifies the student's assimilation of relevant content, refer exclusively to learning related to organized educational situation and cover the targeted curriculum.

Several content validity measurement strategies aim to measure the following values:

1. Content validity coefficient;
2. Inter-evaluator agreement coefficient.

4.3.1. Content Validity Coefficient

The content validity coefficient is calculated by applying the experts' method. The experts' method specific to the Evans technique [11] has two variants of application:

1. Experts formulate a set of items to assess the extent to which the analyzed competence is mastered;
2. Experts judge the extent to which the items of the already constructed instrument capture the scope of acquiring the competence studied [41].

In the case of evaluating the content validity of the items of an already formulated instrument, the experts' method involves asking some specialists (at least seven people who excel in the respective field) to judge the extent to which the items of an evaluation instrument have made the measure of the competence studied. In this sense, a protocol is constructed for presenting the test and investigating to what extent the items designed in the proposed test actually assess what is intended. According to the Evans [11] technique, the protocol for determining the standardized test's validity by the experts' method is well structured. It includes a *prolegomena* in which the theoretical notions that describe the competence to be studied are explained (respectively the declarative, procedural, strategic, and conditional components included, but also the cognitive level to be reached, according to Bloom's taxonomy [3], updated by

Anderson [2]). The items constructed to assess the targeted skill's acquisition level are then presented. The items are accompanied by the justification of the choice of content and the way of structuring them according to certain criteria that faithfully reflect the acquisition of the targeted competence. The specification matrix, item correction scale, and associated score are also presented. Finally, a table is designed for each expert to evaluate the items' relevance and the extent to which the components of the created assessment tool capture the acquisition level of the skills to which the items refer.

The content validity coefficient is calculated using the formula [41]:

$$\text{CVC} = \frac{m - m/2}{m/2}, \quad (1)$$

Where m is the total number of experts, and m_e is the number of experts who consider the test, respectively, the item, to be representative.

This validity coefficient can have values between -1 and 1. The closer its value is to 1, the more validated the content of the item/test is from the perspective of its suitability as an evaluative element in establishing the degree of mastery of the competence in question.

4.3.2. Inter-Evaluator Agreement Coefficient

For statistical checks on the concordance between experts' opinions, the following methods are used that involve evaluation:

1. Concordance coefficient;
2. Kendall coefficient (τ);
3. κ Cohen coefficient;
4. Item-objective congruence index (I_{ik});
5. Item-objectives congruence index (I'_{ik}).

(i). Concordance Coefficient

The calculation of the concordance coefficient is established in the case of an evaluation tool already made, using the formula valid for a single item of the test:

$$\text{CC} = 1 - \frac{s^2 dif}{s^2 max}, \quad (2)$$

Where s^2_{dif} is the variance between expert evaluators, and s^2_{max} is the maximum possible variance between expert ratings:

$$s^2_{dif} = \frac{\sum dif^2 - \frac{(\sum dif)^2}{n}}{n-1}, \quad (3)$$

Where n is the number of experts, and dif represents the number of differences between the values obtained at the evaluations of n correctors for a certain item ordered in a statistical series.

The minimum number of differences is 0, which means that all n evaluators have the same grade as the corrected item in a test solved by a student. In this case, the minimum value of dif is dif=0, $(\sum dif)^2 = 0$, $\sum dif^2 = 0$, $s^2_{dif} = 0$, therefore CC=1, which signifies a maximum concordance between evaluators. The maximum number of differences between n-ordered evaluators in a statistical series is n-1, which means that all evaluators gave different grades to the corrected item in a test solved by a student. In this case, the maximum value of dif is dif=n-1, $(\sum dif)^2 = n^2 - 2n + 1$, $\sum dif^2 = n - 1$, so $s^2_{max} = \frac{n-1-\frac{n^2-2n+1}{n}}{n-1} = \frac{1}{n}$. If all evaluators gave different grades, then $s^2_{dif} = s^2_{max} = \frac{1}{n}$, consequently CC=0. So the concordance coefficient has values between 0 and 1.

A coefficient value of less than 0.3 indicates a very low agreement between expert opinions. A value of the concordance coefficient greater than 0.7 indicates a high consensus between the experts' opinions, so the content of the item/test is valid from the assessment of the targeted competence or characteristic [22].

Legendre [26] proposes the following formula for calculating the concordance coefficient valid for all test items:

$$\text{W} = \frac{12S' - 3m^2 n(n+1)^2}{m^2(n^3 - n) - mT}, \quad (4)$$

Where m is the number of evaluators, n is the number of items, T is a correction factor for equal grades, and S' is calculated with the formula:

$$S' = \sum_{i=1}^{n} x_i{}^2, \quad (5)$$

Where x_i is the sum of the grades the evaluator awards to those in the test. In addition, the correction factor T is calculated with the formula:

$$T = \sum_{k=1}^{g} (t_k^3 - t_k), \quad (6)$$

Where t_k i is the number of equal grades for item k out of the g items with equal grades. The sum is calculated for all groups of equal grades found in the m items.

Consider, for example, the table of scores of four evaluators on a ten-item test solved by a student:

Table 2. *Scores of four evaluators on a ten-item test solved by a student.*

***n* items** \ ***m* evaluators**	**The grade of the first evaluator**	**The grade of the second evaluator**	**The grade of the third evaluator**	**The grade of the fourth evaluator**	**The sum of grades per item (x_i)**	**Sum of squared grades per item (x_i^2)**	t_k	$t_k^3 - t_k$
Item 1	5	6	3	5	19	361	2	6
Item 2	10	4	8	2	24	576	0	0
Item 3	7	8	5	4	24	576	0	0
Item 4	8	10	9	2	29	841	0	0
Item 5	6	5	7	6	24	576	2	6
Item 6	9	7	10	7	33	1089	2	6
Item 7	3	3	2	8	16	256	2	6
Item 8	1,5	2	4	9	16.5	272.25	0	0

n items \ *m* evaluators	The grade of the first evaluator	The grade of the second evaluator	The grade of the third evaluator	The grade of the fourth evaluator	The sum of grades per item (x_i)	Sum of squared grades per item (x_i^2)	t_k	$t_k^3 - t_k$
Item 9	1,5	1	1	2	5.5	30.25	2	6
Item 10	4	9	6	10	29	841	0	0
W = (12·5418.5-3· 16 · 10 ·121); (16· 990 − 120) = 6942; 15720 = 0.441			$S' = 5418.5$				$T = 30$	

The value 0.441 indicates a low concordance between the correctors' ratings in the example.

(ii). Kendall Coefficient

Kendall coefficient [21] for two experts has values between -1 and 1 and is calculated according to the formula:

$$\tau = \frac{the\ number\ of\ pairs\ of\ concordant\ evaluations - the\ number\ of\ pairs\ of\ discordant\ evaluations}{the\ total\ number\ of\ pairs\ of\ evaluations} \quad (7)$$

If the first expert has the evaluations $X = \{x_1,x_2,...,x_n\}$ for the n items, and the second expert has the evaluations $Y = \{y_1,y_2,...,y_n\}$ for the n items, then any pair of observations made by the two researchers (which are elements of X and Y, respectively) for items i and j, where $i < j$ (so the pair of observations (x_i,y_i), respectively (x_j, y_j)) is considered concordant, either if $x_i<x_j$ and $y_i<y_j$, or if $x_i>x_j$ and $y_i>y_j$. In this case, we can formalize the calculation of Kendall's τ coefficient as follows [33]:

$$\tau = \frac{2}{n(n-1)}\sum_{i<j} sgn(x_i - x_j)sgn(y_i - y_j), \quad (8)$$

where $sgn(x_i - x_j)$ represents the sign of the difference between x_i and x_j, and the total number of pairs is $\frac{n(n-1)}{2}$.

It is observed that all the experts' opinions are concordant (so either $sgn(x_i - x_j)$ =-1 and $sgn(y_i - y_j)$ =-1, or $sgn(x_i - x_j)$=1 and $sgn(y_i - y_j)$=1), then the value of the coefficient τ is 1. If all the experts' opinions are discordant (so either $sgn(x_i - x_j)$=-1 and $sgn(y_i - y_j)$=1, or $sgn(x_i - x_j)$=1 and $sgn(y_i - y_j)$=-1), then the value of the coefficient τ is -1. A higher value of Kendall's coefficient indicates a greater concordance between the experts' ratings. This means they apply similar standards when evaluating the content samples between the perspective of item concordance and the target competence [33].

Take the example of the table of grades of two evaluators on the test were solved by a student:

Table 3. *Grades of two evaluators on the test solved by a student.*

n items \ *m* evaluators	The grade of the first evaluator	The grade of the second evaluator
Item 1	5	6
Item 2	10	4
Item 3	7	8
Item 4	8	10
Item 5	6	5
Item 6	9	7
Item 7	3	3
Item 8	1.5	2
Item 9	1.5	1
Item 10	4	9

$\tau = \frac{1}{45}$ {[(-1) · 1+(-1) · (−1) +(-1) · (−1) + (−1) · 1 + (−1) · (−1) + 1 · (−1) + 1 · 1 + 1 · 1 + 1 · (−1)]+[1 · (−1) + 1 · (−1) + 1 · (−1) + 1 · (−1) + 1 · 1 + 1 · 1 + 1 · 1 + 1 · (−1)]+[(−1) · 1 +(-1) · (−1) + 1 · (−1) + (−1) · (−1) + (−1) · (−1) + (−1) · (−1) + (−1) · 1]+[1 · 1 + (−1) · 1 + 1 · 1 + 1 · 1 + 1 · 1 + 1 · 1]+[(−1) · (−1) + 1 · 1 + 1 · 1 + 1 · 1 + 1 · (−1)]+[1 · 1 + 1 · 1 + 1 · 1 + 1 · (−1)]+[1 · 1 + 1 · 1 + (−1) · (−1)]+[0 · (-1)+ (−1) · (-1)]+(−1) · (−1)}

$\tau = \frac{1}{45}$·[1-2+1+4+3+2+3+1+1]= $\frac{14}{45} = 0.311$

The value 0.311 indicates a low concordance between correctors' ratings in the example.

(iii). K Cohen Coefficient

κ Cohen coefficient [7] is calculated using the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

where p_o represents the observed relative agreement between expert opinions (which means a maximum difference of half a point plus or minus between the evaluators' marks), and p_e represents the hypothetical probability of random agreement between expert decisions. For two experts, the formulas are:

$$p_o = \frac{the\ number\ of\ agreements\ observed\ between\ evaluators}{N}, \quad (10)$$

$$p_e = \frac{1}{N^2}\sum_{i=1}^{g} x_i y_i, \quad (11)$$

where N is the number of items, g is the number of items on which evaluators have a relative agreement, x_i represents the grade offered by the first evaluator, and y_i represents the grade given by the second evaluator on item i, which is an item that shows relative agreement in grading.

The coefficient κ takes values between -1 and 1. If $\kappa < 0.00$, then the agreement is poor; if $0.00 \leq \kappa \leq 0.20$, then the agreement is slight; if $0.21 \leq \kappa \leq 0.40$, then the agreement is moderate; if $0.61 \leq \kappa \leq 0.80$, then the agreement is substantial, and if $\kappa > 0.80$ then the agreement is almost perfect. In practice, values of κ greater than 0.6 are accepted [23].

The approximate standard error of the κ coefficient is given by:

$$SE(\kappa) = \sqrt{\frac{p_0(1-p_0)}{n(1-p_e)^2}} \quad (12)$$

and the 95% confidence interval for the population value of the κ coefficient may be estimated by:

$$\kappa \pm 1.96\ SE(\kappa). \quad (13)$$

Take the example of the table of grades of two evaluators on the test solved by a student:

Table 4. *Grades of two evaluators on the test solved by a student.*

m evaluators / *n* items	The grade of the first evaluator	The grade of the second evaluator
Item 1	5	6
Item 2	10	4
Item 3	7	8
Item 4	8	10
Item 5	6	5
Item 6	9	7
Item 7	3	3
Item 8	1.5	2
Item 9	1.5	1
Item 10	4	9

$p_0 = \frac{3}{10} = 0.3$, $p_e = \frac{1}{100}(9 + 3 + 1{,}5) = 0.135$, $\kappa = \frac{0.3-0.135}{1-0.135}$, $\kappa = 0.190$. The value 0.190 indicates a very low concordance between the correctors' grades in that example.

The approximate standard error of κ coefficient is SE(κ) $= \sqrt{\frac{0.3(1-0.3)}{10(1-0.135)^2}} = 0.167$, and the 95% confidence interval for the κ coefficient is (-0.137, 0.517).

(iv). Item-Objective Congruence Index (I_{ik})

The item-objective congruence index [35] is used to obtain the meaning of content validity, the fact that each item (s) fulfills a single objective (k); it is calculated using the formula restructured by Crocker and Algina [8] in the form of the item-objective index:

$$I_{ik} = \frac{N}{2N-2}(\mu_k - \mu), \quad (14)$$

Where I_{ik} is the item-objective congruence index, N is the total number of items, μ_k is the average of expert ratings for item i in the context of objective k, and μ is the average of expert ratings for item i from the perspective of all objectives.

(v). Item-Objectives Congruence Index (I'_{ik})

The formula for calculating the item-objective congruence index has been adjusted for several objectives. It is designed by Crocker and Algina (apud [45], p. 169) in the form of the item-objectives index:

$$I'_{ik} = \frac{N\mu_k-(N-p)\mu_l}{2N-p}, \quad (15)$$

where I'_{ik} is the item-objectives congruence index, i is the item constructed for the set of k objectives, N is the total number of objectives, p is the number of objectives associated with i, μ_k is the average score of the experts for the item i in the context of the k valid objectives, and μ_l is the average score of the experts for item i in the context of the other l objectives (where $k+l = p$).

4.4. Criterial Validity

Criterion validity assesses how accurately a test measures the outcome for which it was designed; in other words, it refers to the extent to which a test reflects the existence in a student of a characteristic or an acquisition according to a given criterion. Criterion validity has components: predictive/empirical validity and competitive/concurrent validity [13, 16, 17]. Concurrent validity is used when test scores and criterion variables are measured at the same time. Predictive validity is used when criterion variables are measured after obtaining test scores. Criterion validity, through its two structural components (related to competitiveness and predictability), is influenced by the chosen criterion's nature and the target group's characteristics. This is why the declaration of the validity coefficients of a test must be accompanied by the corresponding specification, a clear description of the criterion test, and the target group on which the evaluation was carried out.

4.4.1. Predictive or Empirical Validity

Predictive validity expresses the power of the instrument to predict future characteristics and purchases. In the case of an external predictor, the validity coefficient represents the correlation between the results obtained by applying two assessment instruments designed to measure competency based on different operational objectives. Therefore it is recommended that the instrument or battery of tests that have proven predictive for one group of students be counter-validated by applying it to another group of students. In the case of an internal prediction, the validity coefficient expresses the consistency between the different parts of a test or a battery of tests, respectively, the predictive value of each part of the test for its other subdivisions [24].

The study of empirical validity uses triangulation, a technique by which the researcher makes inferences of precision by interpreting the results regarding the targeted competence he obtained through several methods [12]. Predictive validity is established by working together several types of triangulation [32]: statistical (by varying some temporal, spatial, and group subdivisions), analogical (by comparing with the results obtained by other researchers), theoretical (by combining several theories which refer to the targeted aspects), methodological (by using various data collection techniques) or debate type (by calling for preliminary discussions with other researchers regarding the adequacy of the interpretations made). The regression coefficient can be used to evaluate the validity coefficient (the correlation coefficient between two series of results obtained by correlative testing) [8].

4.4.2. Competitive or Concurrent Validity

Competitive validity consists in comparing either the results obtained by applying two assessment instruments designed to measure a competency-based on different operational objectives or the results obtained by using a proposed instrument and another instrument made by experts (whose validity was previously determined and this is raised). Correspondence between the two types of measurements (scores on the proposed test and on a correlative test, which measures an equivalent competence, respectively, on the proposed test and an expert test)

indicates that the test is valid because it measures what is proposed. In this case, the validity coefficient (the correlation coefficient between two series of results obtained by correlative testing) expresses the correlation between the instrument and the criterion.

4.5. Construct Validity

Construct validity refers to how well a characteristic or an acquisition of a targeted competence is translated into an instrument.

Assessing construct validity requires both a theoretical and an empirical approach. Reviewing specialized bibliography that includes previous experimental data on the same skill and precise construction of hypothetico-deductive reasoning is as essential as procedures based on empirical data.

The most used construct validity assessment procedures based on empirical data analysis are [1, 41]:

1. the correlation between the results obtained by applying the designed test and those obtained by using other tests that measure the same targeted competence;
2. the correlation between the results obtained through the application of the designed evaluative tool and those obtained through the application of the measurement tools of other competencies than the one studied;
3. factorial analysis;
4. the study of the effect of certain experimental variables on the results obtained by students;
5. the constructs-methods matrix [5], through which convergent validity (the extent to which two tests measuring the same competence can be correlated) and discriminative validity (the extent to which two tests measuring other competencies than the one studied do not can be correlated with the question test). The analysis results in the centralizing matrix that includes correlations between tests measuring the same target competence, tests measuring the target competence, tests measuring different competencies from the target competence, and correlations between the latter when using different measurement methods.

5. Conclusions

The construction of a standardized test created from the teleological perspective of incremental learning is a complex process aimed at going through several stages. One of the essential stages is establishing the extent to which the created test meets the validity criteria, so necessary in recognizing the test as an effective assessment tool that can be applied in specific conditions with correct, reliable results. A complex analysis of the validity of a standardized test involves the study of all its components: internal validity, external validity, content validity (highlighted by the value of the content validity coefficient and the value of the concordance coefficient), criterion validity (with its components competitive validity and predictive validity), and construct validity (which involves both a theoretical and an empirical approach).

References

[1] Anastasi, A. (1976). *Psychological testing*. New York, U. S. A.: Mac Millian Publish, Co., Inc.

[2] Anderson, L. W., Krathwohl, D. R. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York, U. S. A.: Addison Wesley Longman, Inc.

[3] Bloom, B. S., Krathwohl. D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners.* Handbook 1: Cognitive domain. New York, U. S. A.: Longmans.

[4] Campbell D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin. 54 (4): 297–312.

[5] Campbell, D. T., Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 56 (2): 81–105.

[6] Carpenter, G. A., Grossberg, S., Rosen, D. B. (1991). Fuzzy Art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*. 4 (6): 759-771.

[7] Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*. 70 (4): 213–220.

[8] Crocker, L. Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, U. S. A.; Holt, Rinehart and Winston Inc.

[9] Cronbach, L. J. (1971). Validation test. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC, U. S. A.; American Council on Education.

[10] Davis, G. A. (2003). Prompting middle school science students for reflection: Generic and directed prompts. *The Journal of the Learning Sciences*. 12.

[11] Evans, J. D. (1985). *Invitation to psychological research*. U. S. A.; New York. CBS College Publishing. 74-82: 232-254.

[12] Fielding, N., Fielding, J. (1986). Linking Data: Qualitative Research Methods. London, U. K.; *Sage*. 4.

[13] Gall, M. D., Gall, J. P., Borg. W. R. (2007). *Educational research: an introduction*. The 8th Edition Pearson Education, Inc. Boston, U. S. A.; 192-227.

[14] Gepperth, A., Hammer, B. (2016). Incremental learning algorithms and applications. *European Symposium on Artificial Neural Networks (ESANN)*. Bruges, Belgium.

[15] Gilles, J. -L. (2002). Spectral quality of standardized university tests – Development of edumetric indices for the analysis of the spectral quality of evaluations of university student achievements and application to the MOHICAN checkup ’99 tests (PhD thesis in education sciences) (Qualité spectrale des tests standardisés universitaires – Mise au point d’indices édumétriques d’analyse de la qualité spectrale des évaluations des acquis des étudiants universitaires et application aux épreuves MOHICAN checkup ’99 (Thèse de doctorat en sciences de l'éducation)). Université de Liège, Liège, Belgique.

[16] Gilles, J. -L., Detroz, P., Crahay, V., Tinnirello, V., Bonet, P. (2011). The ExAMS platform, an "assessment management system" to instrument the construction and quality management of learning assessments. In Blais, Jean-Guy (Ed.) *Evaluation of learning and information and communication technology* (La plateforme ExAMS, un "assessment management system" pour instrumenter la construction et la gestion qualité des évaluations des apprentissages. In Blais, Jean-Guy (Ed.) *Evaluation des apprentissages et technologie de l'information et de la communication*). Québec, Canada: Presses de l'Université Laval. 2: 11-40.

[17] Gliner, J. A., Morgan G. A. (2000) *Research Methods in Applied Settings: An Integrated Approach to Design and Analysis.* New Jersey, U. S. A.; Lawrence Erlbaum Associates.

[18] Hartmann, E. (1923). Category theory (*Kategorienlehre*). Edited by Fritz Kern Philosophical Library 72 a/b/c. XXXII. 978-3-7873-2894-9.

[19] Husserl, E. (1950). Guiding ideas for a pure phenomenology and a phenomenological philosophy (Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie). Walter Biemel. *Husserl Gesámmelte Werke*. Germania: Kluwer Academic Publishers.

[20] Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing, T. M. Haladyna (Eds.). *Handbook of test development*. Mahwah, New Jersey, U. S. A.; Lawrence Erlbaum Associates. 131-153.

[21] Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika.* Oxford, U. K.; Oxford University Press. 30 (1/2): 81-93.

[22] Lamata M. T., Pelaez J. I. (2002). A method for improving the consistency of judgments. *Int. J. Uncertain. Fuzziness.* 10: 667–686.

[23] Landis, J. R., Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*. 33 (2): 363-374.

[24] Laugier, H, Piéron, H, Toulouse, É, Weinberg, D. (1934). *Docimological studies on the improvement of exams and competitions*. (*Etudes docimologiques sur l'amélioration des examens et concours*). Paris, France: Conservatoire National des Arts et Métiers.

[25] Leclercq, D. (1993). Validity, Reliability, and Acuity of Self-Assessment in Educational Testing. In: Leclercq, D. A., Bruno, J. E. (eds) Item Banking: *Interactive Testing and Self-Assessment*. NATO ASI Series. 112. Berlin, Germany: Heidelberg. Springer.

[26] Legendre, P. (2010). Coefficient of Concordance. *Encyclopedia of Research Design*. New Jersey: Salkind ed. SAGE Publications Inc. 1: 164-169.

[27] Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement*. Washington, DC, U. S. A: Amercian Council on Education and Macmillan: 13-103.

[28] Minsky, M., Papert, S. (1972). Progress Report on Artificial Intelligence. *AI Memo. 252. MIT Artificial Intelligence Laboratory*. Cambridge, Massachusetts, U. S. A.

[29] Mislevy, R. J., Almond, R. G., Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report 03-16). New Jersey, U. S. A.; Princeton: Educational Testing Service.

[30] Mislevy, R. J., Haertel, G. (2006). Implications of evidence-centered design for educational testing. Menlo Park, CA, U. S. A.; SRI International.

[31] Mislevy, R. J., Steinberg, L. S., Almond, R. G., Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar, R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing*. Mahwah, New Jersey, U. S. A.; Erlbaum. 15–48.

[32] Muchielli, A. (2002). (coord.). Dictionary of Qualitative Methods in the Humanities and Social Sciences (Dicționar al metodelor calitative în științele umane și sociale). Iasi, Romania: Editura Polirom.

[33] Nelsen, R. B. (2001) [1994]. Kendall *tau* metric. *Encyclopedia of Mathematics*. Helsinki, Finland: EMS Press.

[34] Newell, A., Simon, H. A. (1972). *Human Problem Solving*. New Jersey, U. S. A.; Prentice-Hall.

[35] Rovinelli, R. J., Hambleton, R. K. (1977) *On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity*. Tijdschrift Voor Onderwijs Research. 2: 49-60.

[36] Rupp, A. A., Gushta, M., Mislevy, R. J., Shaffer, D W. (2010). *Evidence-centered Design of Epistemic Games: Measurement Principles for Complex Learning Environments*. JTLA. 8 (4).

[37] Sandy, S. (1980). *The Hawthorne effect.* (1st ed.) Lawrence, Kansas, U. S. A.: Tansy Press.

[38] Schank, R. (1972). *Conceptual Dependency: A Theory of Natural Language Understanding.* Cognitive Psychology. 3: 552-631.

[39] Shute, V. J., Masduki, I., Donmez, O. (2010). *Conceptual Framework for Modeling, Assessing and Supporting Competencies within Game Environments*. Technology Instruction Cognition and Learning. 8 (2): 137–161.

[40] Silva, T. (1993). *Toward an understanding of the distinct nature of L2 writing: the ESL research and its implications*. TESOL Quarterly. 27: 665- 677.

[41] Stan, A. (2002). *Testul psihologic. Evolutie, constructie, aplicatii*. Iași, România: Editura Polirom.

[42] *Standards for Educational and Psychological Testing*. (2013). Washington, DC, U. S. A.: American Educational Research Association.

[43] Streefkerk, R. (2022). Internal vs. External Validity. Understanding Differences and Threats. Scribbr. Retrieved April 17, 2023, from https://www.scribbr.com/methodology/internal-vs-external-validity/

[44] Székely, L. (1950). *Productive processes in learning and thinking*. Acta Psychologica. 7: 388–407.

[45] Turner, R. Carlson, L. A. (2003). *Indexes of Item-Objective Congruence for Multidimensional Items.* International Journal of Testing. 3 (2): 163-171.