Reliability Criteria of Standardized Test as a Form of Practical Assessment Created From the Entelechy Perspective Integrated Into Innovative Teaching

Geanina Havârneanu, Alexandru Ioan Cuza University, Romania

The Paris Conference on Education 2024
Official Conference Proceedings

Abstract
This study aims to present essential ways to determine the reliability of a standardized test. To this goal, we explain the most widely used essential criteria for the accreditation of reliability qualities of a standardized test. Standardized testing is integral to innovative teaching that captures essential elements, including offering a safe, inclusive, and beneficial competitive environment, which creates an operational cognitive background that promotes ethical intelligence, resilience, and the ability to make correct and quick decisions under challenging conditions. In a previous study (2022), we explained the relevance of using as many methods as possible to study a standardized test's validity. In this paper, we aim to analyze reliability estimation in different ways: test-retest reliability (stability coefficient); reliability estimated by alternative forms (equivalence coefficient); reliability calculated by the internal consistency/homogeneity of a test (internal consistency coefficient); inter-rater reliability (intra-class correlation coefficient); reliability estimated by item analysis. This analytical study concludes that the accredit of the reliability of a standardized test necessarily supposes the calculation of the coefficients studied in this article.

Keywords: Reliability, Validity, Standardized Test

Introduction

Innovative learning has broad perspectives. It is based on the education of values, attitudes, and behaviors and on pedagogical paradigm changes that envisage a real correspondence between society's future needs and the current ways of achieving the desired perspectives.

High-stakes standardized assessments are blamed because they can negatively impact how teachers provide and students learn (Kempf, 2016). Two factors cause this type of mentality. There is a tendency to practice teaching-learning only types of objective items with a single correct answer variant, items specific to a standardized test with low difficulty. It is also essential to design standardized tests that assume only items that activate critical thinking and access the higher cognitive levels (Bloom, 1968; Anderson, Krathwohl, et al., 2001).

A conceptual teaching-learning, which has as its central teleological perspective personal entelechy, viewed from intra- and inter-individual perspectives, involves the learning of human values, behaviors, and attitudes that determine the understanding of the concept of lifelong learning to ensure reaching the area of maximal development (Vygotsky, 1978).

Innovative teaching captures essential elements such as providing a safe, inclusive, and beneficial competitive environment for the child, creating an operational cognitive background that promotes ethical intelligence and resilience, and building the ability to make correct and quick decisions in unpredictable, novel, and complex conditions.

I believe the main problem is not the standardized tests themselves but how they are designed (Havârneanu, 2022c). The main element that must be taken into account is the type of evaluation, formative (structuring of the curriculum) or for certification (educational policies) (Nitko, Brookhart, 2011).

Literature Review

Recent studies indicate that examiners make unintentional errors when designing or administering tests, resulting in irrelevant test scores, which affects test reliability (Reed, Cummings, Schaper, Biancarosa, 2014).

The assessment tool should be based on competency criteria (Scallon, 2004), which must be requested and assessed. In designing the test instrument, the evaluator must emphasize the evaluation process, not the final result of the evaluation (Nitko, 1996). In the design of the items, there must be a match between the curriculum competencies requisite, the appropriate context through organized educational situations, and what is intended to be assessed. For this purpose, we follow the stages: contextualization and operationalization of objectives; planning the contents and their degree of difficulty depending on the cognitive level of the tested students; determining the types of items and their construction; the test administration and the analyses of the results, which determines adjustments regarding the difficulty of the verified contents, the number of items or how the statements of the items were designed (Gilles, Detroz, Crahay, Tinnirello, Bonnet, 2011).

Designing a test involves going through several stages: selecting the evaluation contents according to the curricular vision of the test; structuring the skills embodied in performance categories, well operationalized according to the teleological configuration of the test. The goals of evaluation are continuity, coherence, and interdisciplinarity; personal and relational

responsibility through peer and self-assessment (Burger, 2000); assessment for certification, progress, and transfer), as well as the creation of the specification matrix and the correction/rating scale of the proposed items (Havârneanu, 2022b).

The objective items test only lower cognitive levels (recognition, comprehension, application). The value of these item types increases by estimating superior cognitive levels if the multiple-choice item also has answer options, such as "no answer is correct", "all answers are correct", "not all answers are correct", "there are other correct possibilities" or "it is absurd" (Gilles, Lovinfosse, 2004). Using items as a teaching methodology is appropriate because today's students need an active learning process rather than traditional lectures (Twigg, Stoll, 2005).

Methodology

Different reliability measures vary due to their sensitivity to error sources and, therefore, need not be equal. Also, reliability is a property of the test results and is, thus, said to depend on the target group (Dawis, 1987).

Reliability is evaluated in five different ways (Gliner, Morgan, 2000):

1. Test reliability – repeated test (stability coefficient);
2. Reliability estimated by alternative forms (equivalence coefficient);
3. Reliability estimated by internal consistency/homogeneity of a test (internal consistency coefficient);
4. Reliability estimated by item analysis.

1. Test-Retest Reliability (Stability Coefficient)

Test-retest reliability evaluates the stability over time and the precision of the intended tool for assessing a construct. The magnitude of this type of reliability is miscalculated when repeated testing results are due to students' memorization of questions and answers and not to the qualities of the assessment tool, caused by students' familiarity with the questions. Therefore, the evaluator must ensure that the interval between two tests is reasonable (two to six weeks) to avoid this error. It is also essential that the target group is relatively homogeneous in terms of demographic, psycho-physiological, and prognostic characteristics. The empirical method of establishing the test-retest reliability coefficient is measured by calculating the stability coefficient, whose statistical indicator is the Pearson correlation coefficient between the scores obtained by the same target group on the same test at two different times, and must have at least 0.7 when the significance threshold is below 0.05 (Polit, 2014).

2. Reliability Estimated by Alternative Forms (Equivalence Coefficient)

Reliability estimated by alternative forms assumes that the subjects' results after applying a test are comparable to those obtained by the same subjects after applying another parallel test with similar items. Estimating this type of reliability requires the researcher to state the same items differently or change the order of the items within the same instrument randomly. The shortcomings of this method are that the two tests should administered simultaneously, one after the other, on the same day, and the conditions for administering the second test can be modified, demotivating and changing the students' physical-psychological state. The parallel form method is usually the most satisfactory way to determine reliability for well-conducted

tests because it indicates content equivalence and performance stability (Guilford, 1956). The statistical indicator of the equivalence coefficient is the Pearson correlation coefficient, with values between 0.80-0.90 (Anastasi, 1976).

3. Reliability Estimated by the Internal Consistency or Homogeneity of a Test (Internal Consistency Coefficient)

This type of reliability refers, on the one hand, to the extent to which all the items of the evaluation instrument relate to each other (have the same content and referential). On the other hand, to the extent to which each item refers to the score obtained by each individual, and here we mean both absolute consistency (the value of the individual's score) (Safrit, 1976) and relative consistency (the value of the individual's rank in the group) (Weir, 2005).

The empirical method of establishing the homogeneity magnitude involves calculating the internal consistency coefficient, which increases not only with the number of items but also with the number of response categories (Lozano et al., 2008). Several methods have been developed and are used to calculate the internal consistency coefficient, the most well-known of which are (Gliner, Morgan, 2000):

3.1. Subdivided Test Method;
3.2. The Kuder-Richardson method;
3.3. The method of calculating the coefficient α – Cronbach;
3.4. Inter-rater reliability.

3.1. Subdivided/ Split Test Method

The split test method has three variants:

3.1.1. Parallel Bisection Method;
3.1.2. Method of halving τ – equivalents;
3.1.3. The method of congeneric division.

3.1.1. Parallel Bisection Method

This method is a variant of split testing methods used when there is no alternative assessment tool or when the test step is repeated, but the results have not been completed. The technique consists of dividing the results of a test into comparable variances halves and obtaining their correlation coefficient. There are four ways of splitting into two equivalent halves the evaluative instrument designed with the items in the increasing order of their difficulties: by the first item/last item selection rule; by the even rank/odd rank item selection rule; by the permutation or by the rule of random selection of items.

However, the assumption of segregation into strictly parallel elements is too restrictive (Webb, Shavelson, Haertel, 2006). There could be more than one way to divide a test. Each split-half date gives a different reliability value. The complete reliability report is a summary on a synoptic table, such as:

Result per item	Total number of items	α – Cronbach coefficient	SEM Standard error of measurement	Division in halves random	First / last split	Even / odd division	Spearman-Brown random	Spearman-Brown first / last	Spearman-Brown even /odd

Table 1. Correlation between the three rules of application of halving methods and the coefficient α – Cronbach

The table can be automatically generated using the Iteman system.[1]

The method does not throw errors if the items are classified in order of their difficulty; the items are segregated into two parts by bringing together similar items (targeting the same competence and the same content) in one half and singular items in the other half.

The correlation coefficient of the halves of the test is used in the calculation of the internal consistency coefficient, corrected by the Spearman-Brown (1910) formula (Anastasi, 1976, pp. 115-116; Gliner and Morgan, 2000, pp. 314-315):

(1) $$\rho_{total} = \frac{2\rho_{12}}{1+\rho_{12}},$$

where ρ_{12} is the Pearson correlation coefficient between the two halves chosen from among the items of the evaluative instrument.

Cho (2016) criticizes the fact that it is not specified in the assumptions of the calculation of the Spearman-Brown formula that it is assumed that the halves are chosen so that their variances are equal (so-called parallel halving). Parallel-item tests have means, variances, and inter-correlations between equal items (Gulliksen 1950). Cho suggests the use of the following systematic formula, equivalent to the Spearman-Brown type, for calculating the internal consistency coefficient by the split-halves in parallel (SP) method:

(2) $$\rho_{SP} = \frac{4\rho_{12}}{4\rho_{12}+2(1-\rho_{12})}.$$

This is still useful, although it is less often used after developing the formula for calculating the internal consistency coefficient by the method into τ-equivalent halves used when the variances of the split halves are not equal.

3.1.2. Method of Halving τ – Equivalents

Cho (2016) proposes the calculation of reliability by the method of halving with unequal variations of the parts, using the systematic formula of the coefficient of internal consistency by the technique of halving τ - equivalents (split-halves and total - ST):

(3) $$\rho_{ST} = \frac{4\rho_{12}}{\sigma^2},$$

where σ^2 is the variance of the integral test.

It is noted that the methods of calculating the internal consistency coefficient by the parallel halving method and the equivalent halving method have the hypothesis that the segregations of the test items are made so that each part has the same number of items (Cho, 2016).

3.1.3. The Method of Congeneric Division

Calculating the internal consistency coefficient by the congeneric division method mitigates the assumption that the test items are segregated so that each part has the same number of

[1] https://assess.com/iteman/

items. Raju (1970) devised a formula in which he took into account the fact that the length of each part of the test is known, while Angoff (1953) and Feldt (1975) took into account the fact that the length of each part of the test is proportional to the sum of variances, respectively, with the sum of the covariances (the products of the scores obtained on the homologous items of the two correlated test parts).

The Angoff - Feldt formula for calculating the internal consistency coefficient by the congeneric division method is:

(4) $$r_{AF} = \frac{4\sigma_{12}}{\sigma^2 - \frac{(\sigma_1^2 - \sigma_2^2)}{\sigma^2}},$$

where σ_1^2 is the variance of the first part of the test, σ_2^2 is the variance of the second part of the test, σ_{12} is the covariance between the two parts of the test, and σ^2 is the variance of the entire test.

3.2. The Kuder-Richardson or Rational Equivalence Method

The Kuder-Richardson method of calculating internal consistency estimates the homogeneity of the items used in the test.

Homogeneity between items can be affected by two types of errors:

- content sampling (all items are chosen from an extended item base related to the content to be evaluated, therefore, they are too homogenous);
- the heterogeneity of the competencies the items refer to is too high.

The more homogeneous the range of skills the items test, the greater the inter-item consistency. If the researcher is aware that the scope of the competencies studied is heterogeneous, the heterogeneity of the test should not be considered significant. Instead, the items describing the same competence should be homogeneous. In other words, the inter-item consistency of a skill tested by the instrument must be high.

We use the Kuder-Richardson formula (apud Gliner, Morgan, 2000) to calculate inter-item consistency:

(5) $$r = \frac{n}{n-1}\left(1 - \frac{\sum p_i q_i}{\sigma^2}\right),$$

where n is the number of test items, $\sum p_i q_i$ is the correct sum of the products of the proportion of answers to item i in the test (p_i), and the proportion of wrong answers to item i in the test (q_i) (i is from 1 to n - the total number of test items), and σ^2 is the total variance of the test results (Ebel, 1967).

The Kuder-Richardson formula uses the error variance of a respondent with an average score from the sample, and this fact overestimates the error variance of respondents with high or low scores (Colledani, Anselmi, Robusto, 2019).

Instruments containing multiple-choice items do not lend themselves to this type of internal consistency analysis.

The rational equivalence method has the advantage of not retesting the target group, thus eliminating the transfer effect (fluctuations in individual abilities caused by environmental or physical conditions that are minimized) and the practice effect (the difficulty of constructing parallel test forms). The disadvantages are that the division can be done in several ways, and the correlation coefficient in each case can be different. Furthermore, since the test is administered only once, chance errors may affect the two subgroups of items similarly and thus tend to make the reliability coefficient too high.

3.3. The Method of Calculating the Internal Consistency Coefficient α – Cronbach

Internal consistency assesses the consistency of results between items in a test. The most common measure of internal consistency is the α – Cronbach coefficient (a generalization of the Kuder – Richardson method), which is usually interpreted as the average of all possible partition coefficients of test items (Cortina, 1993).

The formula for calculating the internal consistency coefficient α – Cronbach is:

$$(6) \qquad \alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right),$$

where n is the total number of items, σ_i^2 is the variance associated with item *i*, and σ^2 is the total variance of the results obtained following the application of the evaluation tool.

In the analysis of the variance of an item, if it does not fit, it can be removed, which can follow the reliability but sometimes leads to the reporting of the reliability at the group level as higher than the reliability at the population level (Kopalle, Lehmann, 1997). Eliminating less reliable items must be done according to statistical studies (in which the entire target group is divided and then cross-validated) and on theoretical and logical grounds (Kopalle, Lehmann, 1997). Suppose it is desired to increase the reliability of the test by adding items. In that case, you must consider maintaining the homogeneity of the test, which means that new items refer to the same target competence as the existing ones and order the items according to the difficulty level.

The values of the α - Cronbach coefficient recommended for an optimal level of reliability must comply with the George - Mallery grid (2003). It would be best if you also considered stadium research when calculating the value of Cronbach's α coefficient, which should be 0.5-0.7 at the early stage of research, around 0.8 at the stage of applied research, and a minimum of 0.9, when you have to make an important decision (Nunnally, 1978).

3.4. Inter-rater Reliability

Inter-rater reliability refers to the agreement between ratings by two or more researchers applying the same instrument to the same students. Evaluators can be randomly selected, but it is also recommended to involve experts by using experts simultaneously with randomly selected evaluators. Inter-rater reliability can be determined by calculating the following coefficients:

3.4.1. Intra-class correlation coefficient;

3.4.2. The concordance correlation coefficient.

3.4.1. The Intra-class Correlation Coefficient

The intraclass correlation coefficient assesses the consistency or reproducibility of quantitative measurements made by different raters using the instrument applied to the same students.

We suppose that we know a set of data related to the values obtained by a student in tests evaluated by two experts when applying a test format of N items. Thus, N unordered data values of pair type (x_n, y_n) are obtained, where x_n represents the student's grade in an evaluation of item n given by the first evaluator, and y_n is the student's grade on item n given by the second evaluator for $n = 1, ..., N$. The intra-class correlation coefficient r proposed initially by Fisher (1954) is:

$$r = \frac{1}{Ns^2}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y}), \quad (7)$$

where

$$\bar{x} = \frac{1}{2N}\sum_{n=1}^{N} x_n \quad (8)$$

$$\bar{y} = \frac{1}{2N}\sum_{n=1}^{N} y_n \quad (9)$$

$$s^2 = \frac{1}{2N}\{\sum_{n=1}^{N}(x_n - \bar{x})^2 + \sum_{n=1}^{N}(y_n - \bar{y})^2\} \quad (10)$$

Since in the denominator for the calculation of s^2, the number of degrees of freedom is 2N −1, the calculation of the value of s^2 becomes unbiased and objective. Also, since in the denominator for the calculation of r, the number of degrees of freedom is N−1, the calculation of the value of r becomes fair and unbiased if it is known. The intraclass correlation coefficient for unordered pairwise data takes values in the range [−1, +1].

When the number of correctors increases, the following formula is applied to calculate the intra-class correlation coefficient (Harris, 1913).

$$r = \frac{K}{K-1}\cdot\frac{N^{-1}\sum_{n=1}^{N}(\overline{x_n}-\bar{x})^2}{s^2} - \frac{1}{K-1}, \quad (11)$$

where K is the number of evaluators, N is the number of items, and $\overline{x_n}$ is the average of the marks given by the K evaluators obtained by the student on the n^{th} item.

3.4.2. The Concordance Correlation Coefficient

The concordance correlation coefficient assesses reproducibility (the degree of agreement between a series of measurements made with the same assessment tool when individual measurements are made by changing one or more conditions) or inter-rater reliability.

We know a set of data related to the values obtained by a student in tests evaluated by two experts when applying a test format of N items. The concordance correlation coefficient is calculated using the formula:

$$\widehat{\rho_c} = \frac{2s_{xy}}{s_x^2+s_y^2+(\bar{x}-\bar{y})^2}, \quad (12)$$

where

$$\bar{x}=\frac{1}{N}\sum_{n=1}^{N} x_n \quad (13)$$

(14) $\bar{y}=\frac{1}{N}\sum_{n=1}^{N} y_n$

the variance is:

(15) $s_x^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$

The covariance is:

(16) $s_{xy} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})\,(y_n - \bar{y})$.

It was observed that the concordance correlation coefficient values are almost identical to the intra-class correlation coefficient values. Comparisons of these two coefficients on different data sets found only minor differences between the two correlations, most often at the third decimal place (Nickerson, 1997).

4. Reliability Estimated by Item Analysis

Clarifying item announcements, repeated measures (Cortina, 1993), and complex item analysis can establish reliability.

Several methods can do the complex analysis of the items:

4.1. Formal item analysis;
4.2. Rasch analysis to identify non-representative items;
4.3. Informal methods of item analysis;
4.4. The relationship between reliability and test length.

4.1. Formal Item Analysis

Formal item analysis, which involves calculating item difficulty and discrimination indices, is considered the most effective way to increase reliability.

The difficulty coefficient of an item (Anastasi, 1976) is calculated as the percentage of subjects who solve an item correctly. Items that are too easy or too difficult from the perspective of the skills involved in formulating an answer do not provide relevant information about the students and are eliminated in the test review stage. From a strictly statistical point of view, the ideal item would be the one that is solved correctly by 50% of the subjects.

The discrimination coefficient (Anastasi, 1976) indicates how an item differentiates high and low performers. It is calculated as the difference between the percentage of subjects who correctly solved the analyzed item in the top fifth of the ranking (the first 20% of subjects) and the percentage of subjects who correctly solved the analyzed item in the bottom fifth. The value of the discrimination coefficient must meet the condition of being at least 25%.

4.2. Rasch Analysis to Identify Non-representative Items

In a Rasch analysis (Lans et al. 2018), items that do not usefully contribute to a measurement can be identified by reviewing the so-called representativeness statistics[2], which apply to each item separately. If an item clearly does not fit after many tests, it is most effective to remove it from the test and replace it with another representative item.

[2] MNSQ Item Outfit, MNSQ Item Infit

Because measurements with perfect reliability are invalid (Cho, Kim, 2015), sacrificing validity to increase reliability results in the validity attenuation paradox (Loevinger, 1954). For high content validity, each item should be constructed to represent the content to be measured comprehensively. However, repeatedly asking the same question in different ways is often used just to increase reliability (Streiner, 2003).

4.3. Informal Methods of Item Analysis

Methods to increase reliability before data collection include removing ambiguity from the wording of the items being measured, constructing items only from curriculum known to the students, increasing the number of items (without destroying measurement effectiveness), using a scale that is known to be highly reliable, pretesting, excluding or modifying items that proved unreliable in the pretest. Methods to increase reliability after data collection are eliminating unreliable items (accompanied by a theoretical justification) and using a reliability coefficient as accurately as possible.

4.4. The Relationship Between Reliability and Test Length

Considering that following the pretest, it is indicated to make changes not only in the restructuring of the wording of some items to eliminate ambiguities and make them more coherent and easier for students to understand, but also in the structure of the evaluation tool by removing or introducing new items, it was necessary to analyze the reliability of the new instrument obtained, depending on the new number of items.

In this sense, the Spearman-Brown formula indicating the relationship between reliability and test length is used to estimate the possible change in reliability/precision when changing the size of the test by removing or adding items for different reasons:

(17) $$r_{xx} = \frac{nr}{1+(n-1)r},$$

where r_{xx} is the reliability estimate coefficient after changing the length of the test, in this new number of items, from the revised version of the test, and is the correlation coefficient calculated between the original and the revised form of the test. In this case, the formula for calculating the standard error of measurements (SEM) is:

(18) $$SEM=\sigma\sqrt{1-r_{xx}},$$

where SEM is the standard error of the measurements, σ is the standard deviation of the results obtained following the administration of the test. In addition, the formula for calculating the 95% confidence interval for obtaining the actual T-test result is:

(19) $$95\%CI = X \pm 1{,}96 \cdot SEM,$$

where 95%CI is the 95% confidence interval for obtaining an actual test result, *T*, *X* is the estimated value of a student's actual test result, ±1.96 the two points on the standard curve that include 95% of the values obtained by students on the test and SEM is the standard error of the measurements. After calculating the coefficient, we can say that there is a 95% chance that the accurate *T* result obtained by the targeted student is between the values *X*-1.96 ·SEM and *X*+1.96 ·SEM.

Conclusions

Calculating reliability requires considering complex factors that can change and, depending on them, choosing the correct method(s).

It is also important not to confuse the reliability with reliability or the validity of a test. The fidelity of a test refers to the degree to which a research study accurately reflects or captures the conditions and procedures of the real-world phenomenon being studied. The reliability of a standardized test assumes that this test produces the same accurate, reproducible, and consistent results when administered multiple times, diachronically, longitudinally, to the same group of students. Validity predicts that an instrument measures the characteristic it is supposed to measure. Reliability is a necessary condition for its validity, meaning that if repeated measurements made by applying an assessment instrument are consistent, the instrument will likely be valid. Validity is a sufficient condition for reliability, meaning a valid test is also reliable. In other words, while a reliable test may provide helpful information from a validity perspective, an unreliable test is certainly not valid (Murphy, Davidshofer, 2005).

References

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

Anderson, L. W., Krathwohl, D.R. (eds.) (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives. New York, US: Longman.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrik.* 18(1): 1-14.

Bloom, B. S. (1968). Taxonomie des objectives pedagogiques. Domaine cognitive. Montreal Lavalle, Canada: Education Nouvelle.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Burger, D. (2000). Assessment and Accountability. PREL Briefing Paper. Pacific Resources for Education and Learning. USA: Honolulu, US.

Cho, E., Kim, S. (2015). Cronbach's coefficient alpha: well-known but poorly understood. *Organizational Research Methods*. 18(2): 207–230.

Colledani, D., Anselmi, P., & Robusto, E. (2019). Using multidimensional item response theory to develop an abbreviated form of the Italian version of Eysenck's IVE questionnaire. *Personality and Individual Differences, 142,* 45–52. https://doi.org/10.1016/j.paid.2019.01.032

Cortina, J.M., (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*. 78(1): 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.

Dawis, R.V. (1987). *Scale Construction*. Journal of Counseling Psychology, 34, 481-489.

Ebel, R. L. (1967). The Relation of Item Discrimination to Test Reliability. *Journal of Educational Measurement*. *4*(3): 125–128.

Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika*. 40(4): 557-561.

Fisher, R. A. (1954). *Statistical Methods for Research Workers (Twelfth ed.). Edinburgh: Oliver and Boyd.*

George, D., Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston, MA: Allyn, Bacon.

Gilles, J.-L., Detroz, P., Crahay, V., Tinnirello, S., Bonnet, P. (2011). In Blais, Jean-Guy (Ed.) La plateforme ExAMS. In *Evaluation des apprentissages et technologie de l'information et de la communication - Tome 2.*

Gilles, J.-L., Lovinfosse, V. (2004). Utilisation du cycle SMART de gestion qualité des evaluations standardisées. Proceedings of World Education for Educational Research (WAER).

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.

Guilford, J. P. (1956). Fundamental Statistics in Psychology and Education. McGraw-Hill.

Gulliksen, H. (1950). The reliability of speeded tests. *ETS Research Bulletin Series*.1: 1-16.

Harris, J. A. (1913). On the Calculation of Intra-Class and Inter-Class Coefficients of Correlation from Class Moments when the Number of Possible Combinations is Large. *Biometrika.* 9 (3/4): 446–472.

Havârneanu, G. (2022a). Validity Criteria of a Standardized Test as an Opportunity for Efficient Assessment Created from the Teleological Perspective of Incremental Learning. Science Journal of Education 11(4):142-149.

Havârneanu. G. (2022b). The Architecture of a Standardized Mathematical Creativity Test for Developmental Evaluation. Proceedings of CBU in Social Sciences. 3:43-49.

Havârneanu G. (2022c). Création de tests standardisés à l'aide de la plateforme SMART.

Kempf, A. (2016). *The Pedagogy of Standardized Testing: The Radical Effects of Educational Standardization in the US and Canada*. Palgrave: MacMillan.

Kopalle, P. K., Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes*. 70(3): 189–197.

Kuder, G. F., Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*(5), 493–504. https://doi.org/10.1037/h0058543

Lozano L. M., García-Cueto E., Muñiz J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology* 4 73–79.

Murphy, K. R., Davidshofer, C. O. (2005). Psychological Testing: Principles and Applications (6th Ed.). *Upper Saddle River, New Jersey.: Pearson/Prentice Hall.*

National Research Council (2003). Assessment in Support of Instruction and Learning: Bridging the gap between large-scale and classroom assessment. USA: The National Academies Press.

Nickerson, C. A. E. (1997). A Note on A Concordance Correlation Coefficient to Evaluate Reproducibility". Biometrics. 53 (4): 1503–1507.

Nitko, A.-J. (1996). *Educational Assessment of Students*. Merrill. Virginia University.

Nitko, A.-J., Brookhart, S.-M. (2011). *Educational Assessment of Students*. Publisher, Pearson/Allyn and Bacon, *2011*; ISBN, 0131382888, 9780131382886.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York, U.S.A: McGraw-Hill.

Polit, D.F. Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research* 23, 1713–1720 (2014).

Raju, N. S. (1970). New formula for estimating total test reliability from parts of unequal length. *Proceedings of the 78th Annual Convention of APA*. 5: 143-144.

Reed, D. K., Cummings, K. D., Schaper, A., Biancarosa, G. (2014). Assessment Fidelity in Reading Intervention Research. *Review of Educational Research*, *84*(2), 275-321.

Safrit, M. J. E. (1976). *Reliability Theory*. Washington, DC: American Alliance for Health, Physical Education, and Recreation.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent (Montréal): éditions du renouveau pédagogique.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Stone-MacDonald, A., Pizzo, L., Feldman, N. (2018). Using Checklists to Improve the Fidelity of Implementation of Standardized Tests. *Springer, Cham*.

Twigg, C., Stoll, C. (2005). Technology as teacher. *The Chronicle of Higher Education*, pp.12-14.

Vygotsky, L. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Webb, N. M., Shavelson, R. J., Haertel, E. H. (2006). *Reliability Coefficients and Generalizability*.

Weir, J.P. (2005). Quantifying Test-retest Reliability using the Intra-class. Correlation Coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19(1), 231– 240.