UNIVERSIDADE FEDERAL DA BAHIA

FIDEDIGNIDADE E VALIDADE DOS TESTES EDUCACIONAIS

RICARDO SERRÃO VINHAS
VALNEI DOS SANTOS SOUZA

Orientadora: PROF.ª LILIA COSTA

DEPARTAMENTO DE ESTATÍSTICA

JULHO 2006

Federal University of Bahia
Institute of Mathematics
Department of Statistics

Consulting Report
Discipline Statistics Laboratory

Advisor: Prof. Lilia Costa
Students: Ricardo Serrão Vinhas
Valnei dos Santos Souza

Salvador
July, 2006

Reliability and Validity Tests

1 - Introduction

The objective of this study is to evaluate the reliability and validity of the Occupational Certification Exam applied to literacy teachers in the state of Bahia.

The certification exam is an innovative and systematic evaluation process, through which it is recognized and attested that the teacher has the necessary skills, that is, skills, knowledge and pedagogical posture sufficient for the improvement of teaching in the state with regard to the performance of their professional activities, according to established standards.

The benefits of the certified professional are promotion in the career classes after passing the certification and recertification exam. For society, the benefit lies in the improvement in the quality of services offered by the government.

The Luis Eduardo Magalhães Foundation – center for modernization and development of public administration, through the Occupational Certification Agency (ACERT), is the first certifying organization for public management professionals in Brazil. To be a certifying agency, it is necessary to have technical capacity, total independence, impartiality, autonomy, credibility and national and international recognition.

ACERT develops, plans and coordinates the entire certification process with the effective participation of professionals occupying positions and functions to be certified, with notorious knowledge and extensive experience in their area of knowledge. These specialists are part of multiple technical committees, responsible for defining the standards of competence for the position or function; prepare the specifications of items, test questions and expectations of test answers; and establish the cut-off line for approval of certification candidates.

Occupational certification can be carried out through a single test or several tests, depending on the complexity of the position or function to be certified. The tests are developed based on competency standards, a document that contains the basic competencies for the performance of the position or function, which are in the public domain. No test is applied without being pre-tested with a group of professionals, with a level similar to that of the real population.

The entire evaluation process has its procedures, quality, coherence and precision validated by an occupational certification chamber, sovereign in decisions, formed by specialists, nationally and internationally recognized, with extensive knowledge in the segment whose professionals will be certified, and by representatives of the institution that hires the certification.

The main objective of the certification for literacy teacher is to contribute to improving the qualification of teachers, the improvement in the knowledge and performance of students and the standard of teaching.

The certification exam highlights professionals focused on objectives, with a critical-reflective and investigative pedagogical posture, with competence to measure the teaching-learning process and with attitudes appropriate to the context and exercise of the position.

From the statistical methodology of Reliability and Validity of the Tests, it is possible to measure the reliability and degree of consistency of the certification exam, that is, to verify whether the literacy teacher certified in this exam really proves to be efficient in the development of his educational activities.

2 - Methodology

2.1 - Reliability of Tests

The development of any scientific activity depends on the perfection of its measuring instruments. The problem in the area of physical sciences was presented by Kelvin (1883), when he stated: "When you can measure what you are talking about and express it in numbers, you have some knowledge; but when it cannot be measured, when it cannot be expressed numerically, knowledge, but thought, whatever the subject, has hardly progressed in the direction of the scientific stage."

The problem of measurement, and consequently, that of the perfection of measurements, has only become an object of concern in the sciences of man in more recent days, including in the educational area, with evident prejudice to its progress.

Educational measures are often used as a starting point for formulating judgments about students' school performance and other types of decisions that affect their lives and futures.

This judgment, however, is worthless if it is not based on reliable quantitative elements, which, in turn, depend on the degree of confidence of the instruments used. Without reliable instruments, reliable scores are not obtained, nor is it possible to express a judgment that is reliable (Wesman, 1952)

Any and all measurement is subject to a measurement error. Measures of educational interest are no exception to the rule; thus, it can be said that the score obtained (X_0) in a test is made up of two elements: the individual's true score (X_v) and the score corresponding to the error (Xe).

$$X0 = X_v + Xe \quad (1)$$

The errors that exist in the scores of a test, and that affect its consistency, are predominantly accidental errors and the greater the error, the less reliable the instrument. Its main characteristic is that, if several measurements are made, the errors tend to cancel each other out and, thus, the score obtained approaches the true one. Unfortunately, the measurements of individuals are few and the influence of error, when not controlled, can significantly affect the degree of consistency (reliability) of the scores.

The reliability of a test (r_{Xx}) refers to the stability of its results, i.e., the degree of consistency of the scores. If a test is applied to the same group a large number of times, the results are expected to be the same, as long as the group does not change. If each time the test is applied, under certain conditions, the scores are different for the same group, one cannot have confidence in the instrument, because there will be no consistency in the measurements.

The greater the error, the less reliable the instrument. The important thing is to minimize the influence of the error so that the score obtained (X0) is closer and closer to the true score (X_v). This is not always possible in practice, hence the need to estimate the degree of reliability in order to know the confidence that the scores deserve.

Just as the score obtained (1) was broken down into its elements, it can also be assumed that the correlation between the true score (X_v) and the score relative to the error (Xe) is null, decompose its variance (Downie and Heath, 1965):

$$s_0^2 = s_v^2 + s_e^2 \quad (2)$$

so the variance of the true score (s_v^2) is

$$s_v^2 = s_0^2 - s_e^2 \quad (3)$$

Reliability is the relationship between the variance of the true score and that of the score obtained, i.e.,

$$r_{xx} = \frac{s_v^2}{s_0^2} \quad (4)$$

Substituting the value of s_v^2 In equation (4), we have

$$r_{xx} = \frac{s_0^2 - s_e^2}{s_0^2} \quad (5)$$

If the test is perfectly reliable, that is, $s_e^2 = 0$, there will be no error in the scores, in this way.

$$r_{xx} = \frac{s_0^2}{s_0^2} = 1 \quad (6)$$

It is the perfect reliability for an equally perfect instrument. Dividing the elements of equation (2) by the variance of the scores obtained (s_0^2), obtains:

$$\frac{s_0^2}{s_0^2} = \frac{s_v^2}{s_0^2} + \frac{s_e^2}{s_0^2} \tag{7}$$

and, by transposition,

$$\frac{s_v^2}{s_0^2} = 1 - \frac{s_e^2}{s_0^2} \tag{8}$$

Equation (4) shows that the reliability (r_{Xx}) is the relationship between s_v^2 and s_0^2; Therefore, equation (8) becomes:

$$r_{xx} = 1 - \frac{s_e^2}{s_0^2} \tag{9}$$

Thus, from a statistical point of view, the reliability is equal to 1 minus the part of the total variance that is the variance of the error.

Operationally, reliability can be defined as the correlation coefficient between two sets of scores obtained independently, in parallel forms of the test, for the same group (Ebel, 1965).

2.1.1- Methods for Calculating Reliability

Reliability refers to the accuracy (consistency, stability) of a test measurement. Precision is a general term, synonymous with reliability; consistency describes the reliability associated with the formula for the distribution of the scores; stability refers to the stability of the trace or characteristic over time (Standards, 1960).

The calculation of reliability requires the comparison of two measures, at least, since operationally, it is a measure of correlation. There are different methods to calculate it. Only the most common were used in this study.

- **Half method** (consistency coefficient)

It is used the split-half method when a single form of the test is applied in a single session and only the influence of sampling is desired, but not that of the variation of responses.

Reliability is calculated by correlating the scores obtained in the two halves of the test. The halves are organized with the sum of the scores in the odd items and the sum of the scores in the even items. The main assumption in the halves method is that the two resulting new

tests are reasonably equivalent; otherwise, the coefficient will be an underestimate of reliability.

If the test is large, different criteria can be adopted to organize the halves; Consequently, there will be fluctuations in the coefficient of trustworthiness. Although different criteria can be used, one should avoid dividing the test in half, simply on the basis of 50% of the first items. The items in a schooling test are arranged in ascending order of difficulty; Therefore, the second half is usually the most difficult. The two forms obtained would not be equivalent and the scores of the second half would suffer the effects of fatigue and practice, which would be reflected in the reliability coefficient. Halves formed by even and odd items are more recommended.

The coefficient obtained translates the correlation of a test that has been reduced by half of its original length; thus, to calculate the reliability of the entire test, the Spearman-Brown prophecy formula must be applied. The coefficient is a measure of sampling consistency.

The Spearman-Brown prophecy formula allows us to calculate the new reliability of a test when it is increased (or decreased) by a certain number. of times. Its general expression is

$$r_{nn} = \frac{nr_{xx}}{1+(n-1)r_{xx}} \tag{10}$$

rnn = reliability for the test increased n times

n= number of times the test is increased

rxx = reliability of the test

When it comes to estimating the reliability of a test with twice the number of items of the original test, a situation that applies to the halves method, the general Spearman-Brown formula turns into:

$$r_{nn} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1+r_{\frac{1}{2}\frac{1}{2}}} \tag{11}$$

rnn = reliability of the entire test

$r_{\frac{1}{2}\frac{1}{2}}$ = Trust of the Half

- **Alpha Coefficient** (α)

Obtaining the alpha coefficient requires the calculation of two parameters, namely: the total variance of the test (s_T^2) and the variance of each item individually (s_i^2). This formula implies that the alpha index will be higher when the specific variance of each item is small and the variance that they together produce is large. It means that the sum of the variances of the individual items is reduced and the variance between them increases, which is the one that guarantees congruence (internal consistency) between the items of the same test. This means that, at the extreme, all the variance produced by the test (s_T^2) would be covariance. Thus, the formula below shows that if all items vary in the same way, that is, if there is no variance between the items individually, the alpha will be equal to 1; that is, that the items will be totally homogeneous, in fact identical, producing exactly the same variance. As such an event is not to be expected, the alpha will give as much congruence or covariation as the items have within the test. The alpha coefficient ranges from 0 to 1, 0 indicating total absence of internal consistency of the items and 1 indicating the presence of 100% consistency.

V(X1+X2) = V(X1) + V(X2) + COV(X1, X2)

This method reflects the degree of covariance of the items with each other, serving as an indicator of the internal consistency of the test itself. Its formula is as follows:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right) \tag{12}$$

where

n = number of items

$\sum s_i^2$ = sum of variances of n items

s_T^2 = total variance of the test scores.

2.2 – Validity of the Tests

The problem of validity is often considered *retrospective*, that is, if the test, after its application, proves to be an efficient measure of what was intended to be measured, it is considered to have validity. Those who do so start from the unknown and run the risk of reaching null results, which almost always happens. Building an item is an art, but structuring the test is a scientific activity that presupposes the formulation of hypotheses and the establishment of adequate means to verify their veracity or falsity. The problem of the validity of a test actually begins at the moment when one thinks of constructing it and persists throughout the process of elaboration, application, correction and interpretation of the results.

The statement that a test is valid when it measures what it is intended to measure is too generic and requires further explanation. Validity is not a general characteristic, it is not a problem of being or not being, it is not a matter of all or nothing, because validity exists in different degrees (Ebel, 1965). Validity refers to what the test measures; Consequently, the word validity cannot be used in isolation and presented abstractly. The test is not valid in general, but in relation to a certain purpose (Adkins, 1947) and for a particular group with specific characteristics; As such, it may not be valid for other purposes and other different groups. It is valid for a certain course and a certain teacher, but it is not valid for other courses and other teachers (Wood, 1969); hence, therefore, it is necessary to characterize the type of validity of the test.

- **Content Validity**

Content validity – also called curricular, sample or logical validity – is, among the different types of validity, the one that most interests tests in the educational area. A test is considered to have content validity when it constitutes a representative sample of knowledge and behaviors acquired during the educational process.

The word content usually provokes several interpretations, which reflect, in a way, the confusion that many still make about content in the area of knowledge and educational objective. Content validity refers not only to the representativeness of the content taught, but also to the behavior developed.

Content validity is not statistically determined, it is not expressed by a correlation coefficient, but rather results from a judgment of different examiners, who analyze the representativeness of the items in relation to the content areas and the relevance of the following objectives. Thus, to demonstrate that a test has content validity, judges must:

1- Identify relevant behaviors that are representatively sampled;

2- Identify content areas, also representatively sampled.

The planning of the test has a great influence on the validity of the content, as it is at this moment that a representative sample of knowledge and behaviors is organized. The problem, however, is not easy to solve. It requires the definition and delimitation of the universe of the two variables, in order to extract a sample from it. One device used by test builders, especially the tests that are going to be standardized, is to imagine a theoretical, ideal universe and then organize the sample. Although the procedure has its logic, the solution is not fully satisfactory.

The type of sample to be organised is also a problem. Theoretically, a random sample would be more suitable, but in practice, the nature of the test requires that the sample be representative of a few specific areas, emphasized during the educational process.

The central problem, of content validity, is, therefore, that of the representativeness of the sample, which, in educational tests, is controlled through specification tables. It is a direct validity, as opposed to other types of validity, which are classified as indirect, because they are represented quantitatively.

- **Predictive Validity**

Also called empirical or statistical validity – it is the degree of correlation between the scores of a test and other measures of performance (criterion), obtained independently of the first test. If the correlation between the scores of the predictive test (X) and the scores of the criterion variable (Y) is high, it is said that the test is valid for the purpose for which it is intended and that it has predictive validity.

Validity scores do not necessarily have to be the scores of a test, they can be grades, passing or failing the course, success or failure in an activity, mistakes made in a task, etc. (Adkins, 1947)

The basic element in establishing predictive validity is the criterion. Without a criterion that is perfectly defined in its characteristics, that is reliable and that is also valid, it is impossible to say what the test actually measures, because its characteristics are defined by the criterion.

The coefficients of predictive validity, often expressed by the moment coefficient – Pearson's product, are almost never high, and those that exceed 0.65 are rare. There are, in fact, no absolute standards. Validity of 0.10 and 0.20 are so inconsequential that the tests should not be used for prediction purposes; However, the validity of 0.45, which many establish as the minimum acceptable value, is not always obtained. Under certain conditions, a test with a validity just above 0.20 has selective values, if no other instrument is available, and there may be a failure of the examinee in his subsequent activity, if the test is not applied, or the examinee is rejected without the application of the test (Guilford, 1946)

- **Construct Validity**

Construct or concept validity is considered the most fundamental form of validity of psychological instruments, and rightly so, since it is the direct way to verify the hypothesis of

the legitimacy of the behavioral representation of latent traits and, therefore, is exactly in line with the psychometric theory defended here.

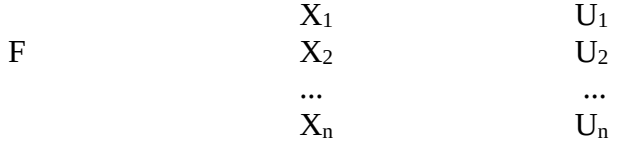
The concept of construct validity was elaborated with the now classic article Cronbach and Meehl (1955), Construct validity in psychological tests, although the concept already had a history under other names, such as intrinsic validity, factorial validity and even face validity. These various terminologies demonstrate the confused notion that construct had. Although they have tried to clarify the concept of construct validity, Cronbach and Meehl still define it as the characteristic of a test as a measurement of an attribute or quality, which has not been "operationally defined". They acknowledge, however, that construct validity called for a new scientific approach. In fact, defining this validity in the way they defined it seems a bit strange in science, given that concepts not operationally defined are not susceptible to scientific knowledge. Concepts or constructs are scientifically searchable only if they are at least amenable to adequate behavioral representation. Otherwise, they will be metaphysical and not scientific concepts. The problem lies in the fact that, synthesizing the general attitude of the psychometricians of the time, in order to define construct validity, the authors started from the test, that is, from the behavioral representation, instead of starting from the psychometric theory that is based on the elaboration of the construct theory (latent traits). The problem is not to discover the construct from a legitimate, adequate representation of the construct. This approach requires a much closer collaboration than exists between psychometricians and Cognitive Psychology.

The construct validity of a test can be worked from several angles: the analysis of the behavioral representation of the construct, the analysis by hypothesis, the IRT information curve, in addition to the statistical falsetto of the estimation error of the Classical Test Theory (TCT).

Factor Analysis

Factor analysis has the logic precisely to verify how many common constructs are necessary to explain the covariances (or intercorrelations) of the items. The correlations between the items are explained, by factor analysis, as resulting from source variables that would be the causes of these covariances. These source variables are the latent constructs or traits that Psychometrics talks about.

Factor analysis also says that a smaller number of latent traits (source variables) is sufficient to explain a greater number of observed variables (items), as shown in the figure below:



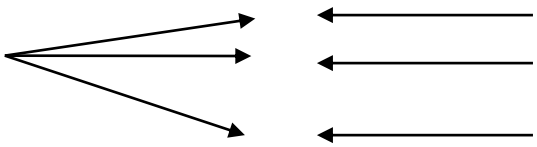


Figure 1 – Representation of the factorial model

The model in figure 1 shows that n variables (X) can be explained by a factor common to all variables (F) plus a specific factor for each of them (U). So that each variable has its equation expressed in terms of these two factors. For instance:

$$X1=a1F+d1U1$$

A1 is the saturation, correlation, and covariance (called factor loading) of the variable X1 in factor F. It represents the percentage of relationship it has with the factor, i.e., how much of a percentage it constitutes as a representation of the factor (latent trait); it indicates, in other words, whether it is a good behavioral representation of the latent trait. In addition, the factor loadings are the ones that determine the correlation between the empirical variables themselves. Thus, the correlation between X1 and X2 is defined by a1 to$_{2.}$

Thus, the construct validity of a test is determined by the magnitude of the factor loadings (which are correlations ranging from -1 to 1) of the variables in the factor, which are the behavioral representation of this factor, which, in turn, is the latent trait for which they were initially elaborated as an empirical representation. These factor loadings represent the fundamental part of the true score (V) of the equation of classical Psychometrics T=V+E. We say fundamental part, because another part of the V is constituted by the specific contribution of the item (contained in the U factor of the factorial model) to the empirical score T of the test. In fact, the total variance of an item or variable can be decomposed into common variance, specific variance, and error variance.

The common variance represents what the variables of the test have in common (expressed by the intercorrelations between them) and which is collected in the factor loadings in the common factor F and it is this that constitutes the question of the validity of the test, that is, when the latent trait (factor F) is empirically represented by the variables (items). The rest of the variance of the items is collected in the so-called uniqueness (U) of each item, which represents both what is specific to each of them and the measurement errors. These last two aspects of variance (specificity and error) are grouped into a single concept, namely, uniqueness, because they do not contribute to the validity of the test, since it is the portion of the item that does not constitute a representation of the latent trait.

If there were no difficulties with the factor analysis model, it would constitute a complete empirical demonstration of the construct validity of a test, since it would provide the

exact expression of how much the test would be representing the latent trait. But unfortunately, factor analysis poses some major problems. Two reasons are the main concern in this regard. First, the factorial model is based on exclusively linear equations between variables and factors. Secondly, there is the rotation of the axes, for which there is no objective criterion, other than the psychological interpretability of the factors.

2.3 – Reliability and Validity

Trustworthiness is only one aspect of validity. If a test is not reliable, it cannot be valid; However, being trustworthy is not enough to guarantee validity. Reliability is a necessary condition, but not sufficient to ensure the validity of a test. Thus, an educational test that measures only a few content areas and/or few relevant behaviors may present consistent results, but will not be valid, due to lack of representativeness of the sample.

Reliability is important for validity, but by itself it does not guarantee this characteristic of the test.

3 – Result

To analyze certification consistency, we initially calculated the reliability of each test separately, where each test represents the following situation:

TCE-1: test offered to Professionals with a permanent link with the State, working from the 1st to the 4th grades of EF (Children's Literacy - Daytime) in schools of the State Education Network of Bahia or in municipal schools or those associated with the SEC and that have a complete teaching course, complete normal or complete higher education course in pedagogy.

TCE-2: test offered to Professionals with a permanent link with the State, working from the 1st to the 4th grades of EF (Children's Literacy - Daytime) in schools of the State Education Network of Bahia or in municipal schools or those associated with the SEC and who have a complete teaching course, complete normal or complete higher education course in pedagogy and to candidates absent or not approved in the 1st Exam.

TCE-3: test offered to Professionals with a permanent link with the State, working from the 1st to the 4th grades of EF (Children's Literacy - Daytime) in schools of the State Education Network of Bahia or in municipalized/affiliated schools with the SEC and that have a complete teaching course, complete normal or complete higher education course in pedagogy and to candidates absent or not approved in previous versions.

TCE-4: test offered to teachers with undergraduate training in Pedagogy, Complete or Complete Normal Teaching Course, member of the effective staff of the state public teaching of Elementary Education (1st to 4th grade) of Bahia and who is in effective exercise of teaching activities, corresponding to the attributions of the position they hold, according to Law No. 8,480 of October 24, 2002 and Decree No. 8,451 of February 13, 2003, that regulates it, as well as corresponding to the level of integration in the career.

For each test, a procedure analogous to that described below was performed.

Calculation of the reliability of the TCE-1 by the halving method.

The test contains 64 items, applied to 5152 teachers. The test was corrected by assigning 1 (one) to the items answered correctly and 0 (zero) to the items answered incorrectly. The matrix of the answers and the raw scores of each student, as well as other elements for analysis, are presented in Table 1.

Table 1 – Response matrix of 5152 teachers
to a 64-item test.

Items / Prof.	1	2	3	4	5	6	7	...	62	63	64	X	And
93	1	0	0	0	0	0	0	...	1	0	1	16	11
1083	1	1	1	1	1	1	1	...	1	1	1	29	30
1090	1	1	1	0	1	0	1	...	1	1	1	21	21
1465	1	0	1	1	1	1	1	...	1	0	0	29	26
2172	1	0	1	1	1	1	1	...	1	1	1	29	30
2196	0	1	1	0	1	1	1	...	0	0	0	21	17
2790	1	0	1	0	1	1	1	...	1	0	1	20	21
2837	1	0	1	1	1	1	1	...	0	1	0	22	18
2844	1	1	1	0	1	1	1	...	1	1	1	31	25
3131	1	0	1	1	1	1	1	...	0	0	1	21	20
3186	1	0	1	1	1	0	1	...	1	1	1	22	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
145411	0	0	1	0	1	1	1	...	1	0	1	20	18
145428	1	1	1	1	1	0	1	...	1	1	0	21	16
145435	1	0	0	0	1	0	1	...	0	0	1	13	8
145442	0	1	1	0	1	0	1	...	1	1	1	16	12
145459	1	0	1	1	1	0	1	...	1	0	1	27	20
145466	1	1	0	0	1	0	1	...	0	1	0	18	15
Hits	4309	2403	4064	2594	4152	2430	4554	...	3363	3556	3594		
p	0,84	0,47	0,79	0,50	0,81	0,47	0,88	...	0,65	0,69	0,70		
q	0,16	0,53	0,21	0,50	0,19	0,53	0,12	...	0,35	0,31	0,30		
Why	0,14	0,25	0,17	0,25	0,16	0,25	0,10	...	0,23	0,21	0,21		

Note: the teachers are, in the tables presented, represented by their registration number.

The test was divided in half and two new tests were organized based on the scores of the odd (X) and even (Y) items, subsequently calculating the correlation between the sets of pairs of scores and, subsequently, the reliability estimated by the Spearman-Brown formula.

Table 2 - Calculation of the correlation coefficient of the scores in Table 1 for the purpose of estimating the reliability of the test (halving method)

Professor	X	And
93	16	11
1083	29	30
1090	21	21
1465	29	26
2172	29	30
2196	21	17
2790	20	21
2837	22	18
2844	31	25
3131	21	20
3186	22	19
⋮	⋮	⋮
145411	20	18
145428	21	16
145435	13	8
145442	16	12
145459	27	20
145466	18	15
Sum	104383	95949

The characteristics of variables X and Y were calculated and are as follows:

N = 5152

$\overline{X}$ = 20,26

$\overline{Y}$ = 18,62

Sx = 5,49

And = 5,90

$\sum XY$ = 2082072

Substituting the above values in the formula for calculating the correlation coefficient, based on the raw scores, we have:

$$r_{xx} = \frac{\frac{\sum XY}{N} - \overline{X}\overline{Y}}{S_x S_y}$$

$$r_{xx} = 0{,}83$$

The reliability for the test with half the number of items is 0.83. Applying the Spearman-Brown prophecy formula (10), to estimate the reliability of the test with twice the number of items, that is, the entire test, we have:

$$r_{n(2)} = \frac{2x0{,}83}{1 + 0{,}83} = 0{,}91$$

The reliability of the test is 0.91, which means that 91% of the variance in the scores depends on the true variance and 9% on the variance of the error.

Calculation of the reliability of the TCE-1 by the Alpha Coefficient.

Table 3 – Matrix for calculating the Alpha Coefficient

Professor	X	And	**T**	**t=$T-\overline{T}$**	**t²**
93	16	11	27	-11,88	141,24
1083	29	30	59	20,12	404,64
1090	21	21	42	3,12	9,71
1465	29	26	55	16,12	259,72
2172	29	30	59	20,12	404,64
2196	21	17	38	-0,88	0,78
2790	20	21	41	2,12	4,48
2837	22	18	40	1,12	1,24
2844	31	25	56	17,12	292,95
3131	21	20	41	2,12	4,48
3186	22	19	41	2,12	4,48
⋮	⋮	⋮	⋮	⋮	⋮
145411	20	18	38	-0,88	0,78
145428	21	16	37	-1,88	3,55
145435	13	8	21	-17,88	319,85
145442	16	12	28	-10,88	118,47
145459	27	20	47	8,12	65,86
145466	18	15	33	-5,88	34,63
Sum	104383	95949	200332	-	610789

Where

T = Xi + Yi;

–

T = ∑ T/N

The characteristics of the variables T and t and t²:

n = 64

$\sum s_i^2$ = 13,34

s_T^2 = 610789/N

Substituting the above values in the formula for calculating the alpha coefficient (12) we have:

$$\alpha = \frac{n}{n-1}(1-\frac{\sum s_i^2}{s_T^2})$$

$$\alpha = \frac{64}{63}(1-\frac{13,34}{118,55})$$

$$\alpha = 0,90$$

The reliability of the test is 0.9, which means that 90% of the variance in the scores depends on the true variance and 10% on the variance of the error.

Therefore, for both methods, whether through the Half Method or the Alpha Coefficient, the test showed good reliability.

The following table shows the summary of the Reliability tests applied to the TCE-1, TCE-2, TCE-3, and TCE-4 exams.

Table 4 – Summary of Reliability Tests

	Half Method	Spearman-Brown	Alpha Coefficient
TCE-1	83%	91%	90%
TCE-2	75%	86%	84%
TCE-3	75%	86%	85%
TCE-4	83%	91%	92%

From the analysis of the results, we concluded that the exams applied to literacy teachers are really reliable.

Study of the Validity of TCE-1 and TCE-2

According to a widely accepted basic document from the American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999) test validity research procedures can be divided into five categories:

1. Content-based evidence (Content Validity):

It collects data on the representativeness of the test items by investigating whether they consist of comprehensive samples of the domain that is intended to be evaluated with the test.

2. Evidence based on relationships with other variables (Predictive validity):

It collects data on the correlation patterns between test scores and other variables by measuring the same construct or related constructs (convergence) and with variables measuring different constructs (divergence). It also provides data on the predictive capacity of the test of other facts of direct interest (external criteria) that have importance in their own right and are associated with the direct purpose of the use of the test (e.g., success at work).

3. Evidence based on internal structure (Construct Validity):

It collects data on the structure of the correlations between items evaluating the same construct and also on the correlations between subtests evaluating similar constructs.

4. Evidence based on the response process:

It collects data on the mental processes developed in the performance of the tasks proposed by the test.

5. Evidence based on the consequences of the test:

It examines the intended and unintended social consequences of the use of the test to verify that its use is having the desired effects according to the purpose for which it was created.

The validity of category 1, or content validity, applied to the certification exam for literacy teachers is done by specialist professionals who are part of multiple technical committees, responsible for defining the standards of competence for the position or function, preparing the specifications of items, test questions and expectations of test answers, and establishing the cut-off line for approval of candidates for certification.

The entire evaluation process has its procedures, quality, coherence and precision validated by an occupational certification chamber, sovereign in decisions, formed by specialists, nationally and internationally recognized, with extensive knowledge in the segment whose professionals will be certified, and by representatives of the institution that hires the certification. All items that make up the certification exam undergo this type of validity.

The validity of category 3, or construct validity, was performed through factor analysis of the TBI-1 and TBI-2 exams, with the objective of verifying how many common constructs are necessary to explain the covariances (or intercorrelations) of the items. The correlations between the items are explained, by factor analysis, as resulting from the constructs or latent traits that would be the causes of these covariances.

Factor analysis was performed using the TESTFACT software, in which, in addition to the inspection of the eigenvalues of the tetrachoric correlation matrix, the database was reduced in three factors by the PROMAX method.

The factor loadings of each item in factors C1, C2 and C3 were ordered to identify the type of subject most present in each of them.

In the TCE-1 for factor C1, #Assunto2# represents 52% of the items. For factor C2, #Assunto1# represents 58% of the items, and for factor C3, #Assunto3# represents 100% of

the items. Which means that questions that address the same subject have a high correlation and, therefore, have validity.

In the TCE-2 for factor C1, #Assunto1# represents 52% of the items. For factor C2, #Assunto2# represents 45% of the items, while for factor C3, #Assunto3# represents 100% of the items. This shows that the exam is valid because the questions with the same subject have a high correlation.

The validity of category 5 measures the social consequences of the exam, that is, it verifies whether the exam is accepted and/or important for the professionals involved in the area of education.

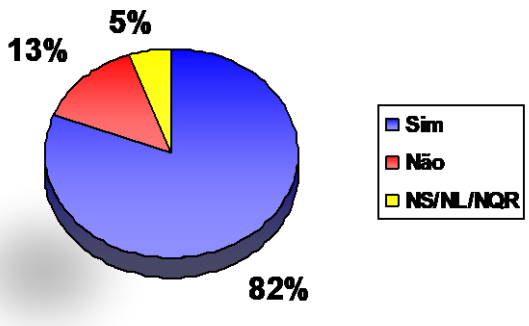
This validity was made from the study of issues related to the improvement of the quality of public education, the professional evaluated and the performance of the students in the Evaluation Form of the Certification Exam for State Literacy Teachers applied to 1367 professionals in the area of education, including teachers, principals, vice-principals and pedagogical coordinators.

The results of the questions on the importance of the Occupational Certification System are described below.

- Improvement of the quality of public education:

Do you consider the Occupational Certification System for Education Professionals important for improving the quality of public education?

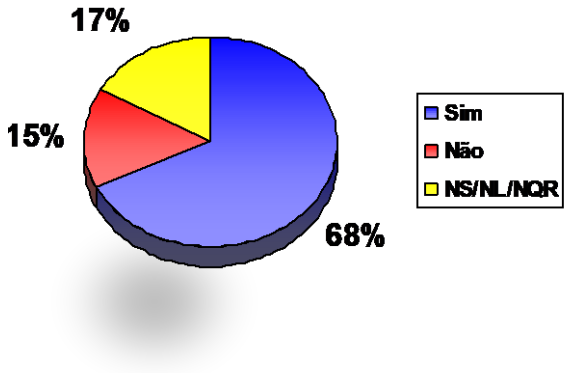
Answers	#	%
Yes	1110	81,2%
No	182	13,3%
NS/NL/NQR	75	5,5%
Total	1367	100,0%



- Improvement of the quality of the professional:

Has the Certification Exam improved your performance in the area of Education?

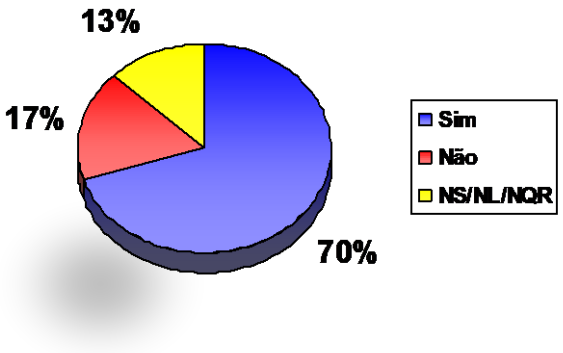
Answers	#	%
Yes	927	67,8%
No	211	15,4%
NS/NL/NQR	229	16,8%
Total	1367	100,0%



- Improved student performance:

Can being certified contribute to improving the performance of your students?

Answers	#	%
Yes	960	70,2%
No	231	16,9%
NS/NL/NQR	176	12,9%
Total	1367	100,0%



The majority of professionals (82%) consider the certification system important for improving the quality of public education. The professionals believe that the certification process contributes to improving the performance of their activities in the area of education and also to the performance of the students, which signals a good validity of the exam.

4 – Conclusion

Based on the theory of Reliability and Validity of Tests and with the use of statistical technology, it can be concluded that the Occupational Certification Exam applied to literacy teachers in the State of Bahia by the Luis Eduardo Magalhães Foundation has coherence, precision and, therefore, are reliable and valid to attest that the teacher has the necessary skills, that is, skills, knowledge and pedagogical posture sufficient for the improvement of teaching in the state with regard to the performance of their professional activities, according to established standards.

5 – Bibliographic References

- American Educational Research Association, American Psychological Association,Nacional Concil on Measurement in Education (1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association

- Psychometry
Theory of Tests in Psychology and Education, Luiz Pasquali.