Article

Journal-Economic History

December, 2019 Vol.3 No.5 21-30

Methodological proposal for the design of learning assessment instruments in educational research

Propuesta metodológica para el diseño de instrumentos de evaluación del aprendizaje en investigaciones educativas

BOCANEGRA-VERGARA, Netzahualcóyotl†*

Centro de Investigación e Innovación para el Desarrollo Educativo
Universidad Pedagógica de Durango

ID 1st Author: *Netzahualcóyotl, Bocanegra-Vergara* / **ORC ID:** 0000-0001-8292-8408

DOI: 10.35429/JEH.2019.5.3.21.30

Received August 11, 2019; Accepted December 25, 2019

Abstract

The objective of this document is to present a useful theoretical-methodological proposal for postgraduate thesis and researchers interested in evaluating the effect of an independent variable on the learning of the students of an institution. As an example, an instrument was designed to assess the learning achieved in the Spanish and mathematics subjects of students in the sixth grade of primary school through a test (EA6B3y4), previously worked in two specific phases: In the first, the model for evaluate the learning in the aforementioned subjects, the model for the design and the piloting of the test. In the second phase, the instrument is analyzed taking as reference the Classical Test Theory (TCT) to assess its attributes and proceed to calibration.

Assesment, Learning, Test

Resumen

El objetivo de este documento es presentar una útil propuesta teórico-metodológica para tesis de postgrado e investigadores interesados en evaluar el efecto de una variable independiente en el aprendizaje de los alumnos de una institución. A modo de ejemplo, se diseñó un instrumento para evaluar el aprendizaje obtenido en las asignaturas de español y matemáticas de los estudiantes de sexto grado de la escuela primaria a través de una prueba (EA6B3y4), previamente trabajada en dos fases específicas: En la primera, el modelo de evaluación el aprendizaje en las asignaturas antes mencionadas, el modelo para el diseño y el pilotaje de la prueba. En la segunda fase, el instrumento se analiza tomando como referencia la Teoría Clásica de Pruebas (TCT) para evaluar sus atributos y proceder a la calibración.

Abismmiento, Aprendizaje, Prueba

Citation: BOCANEGRA-VERGARA, Netzahualcóyotl. Methodological proposal for the design of learning assessment instruments in educational research. Journal-Economic History. 2019. 3-5: 21-30.

* Correspondence to Author (email: netza.bocanegra@durango.gob.mx)

† Researcher contributing as first author

Introduction

It is very common that young educational researchers and postgraduate students in education who go through research processes during the development of their thesis, seek to test the effect of any of the different incident variables on learning phenomena (such as motivation, self-regulation, a specific instructional design, among others), in this crossing they face the problem of implementing exams with acceptable properties for the exercise and validation of their hypotheses. Given this dilemma, it is necessary to establish a model of instrument construction that adapts to the circumstances of the thesis or the new researchers and that in turn offers reliable properties to determine the impact on learning, from a certain experimental treatment.

In addition to the above, different researchers specify the need to establish objective and clear criteria in this regard, since a set of theoretical, methodological and instrumental inaccuracies for the construction of these batteries can be found in the investigative future. In general, according to Barraza (2019, p. 12): Instruments are not used to collect information related to academic performance, failing that, the qualifications, or other indicators deemed appropriate by researchers, obtained institutionally (Álvarez et al. 2015; Bernal et al. 2018; Chiecher) are used centrally , Elisondo, Paoloni & Donolo, 2018; Díaz & Flores, 2018; Estrada, 2018; Gómez et al. 2015; González & Vera, 2018; Goñi, Ros & Fernández-Lasarte, 2018; Horna, 2018; Iniguez-Monroy, Aguilar - Salinas, De Las Fuentes-Lara & Rodriguez-Gonzalez, 2017; Regueiro et al. 2018; Sanz, Fernández-Martínez, Espada & Orgilés, 2018; Serrano et al. 2016; Trelles, Alvarado & Montánchez, 2018); on some occasions the grades or the average are reported directly by the students themselves (Del Rosal, Moreno-Manso & Bermejo, 2018; Marin et al. 2018; Roux & Anzures, 2015).

These types of incidents limit the credibility of the results and the decision-making based on these findings. Therefore, it will be pertinent to establish a consistent and functional model that allows addressing the previously described shortcomings and, in turn, improve the research processes in areas of low impact on education.

The pedagogical test described in this study was generated at the methodological level based on an adaptation of internationally recognized evaluation and psychometric models. The study model was based on the work of Pérez (2010) who designed a low-impact criterial benchmark evaluation instrument at the institutional level at the Autonomous University of Baja California.

In turn, the Nitko model (1994) was taken as a reference for the development of national evaluations of criterial reference and normative reference aligned to the curriculum for the certification and selection of students. On the other hand, the model adapted by Contreras (2009) was taken into account, to develop tests of this same type.

Likewise, the methodology for the construction of tests, both criteria and regulations, proposed by James Popham (1990), of the manual for the development of tests, proposed by Steven Downing and Thomas Haladyna (2006), the recommendations to establish learning objectives in evaluations of Gallardo (2009) and Marzano and Kendall (2007) and finally, the recommendations for this type of evaluation described by the American Educational Research Association [AERA] American Psychological Association [APA] and the National Council on Measurement in Education [NCME] (2014) through the document entitled: Standards for Educational Psychological Testing.

The methodological models mentioned in the previous paragraph, conceive the curriculum as the basis from which it is decided: What contents are essential to evaluate, the procedure necessary to evaluate, the development of the test or instrument that will be used for said evaluation, the application of the test and the analysis of the results.

Table 1. Proposal for the design of the evaluation of the learning of the third and fourth bimester in sixth grade of primary education (EA6B3y4), which is shaped as an adaptation of the Nitko models (1994); Contreras (2009) and Pérez (2011), shows in a general way the procedures used for the design of the test, the reasons for conducting under this model lead to the analysis of the mentioned references.

Phase I	Moments	Activities
Logistics, design, validation and Exam pilot.	1.1 Definition of domain of results that pretend the curriculum	-Constitute the Examination Coordinating Committee (CCE). -Establishment of the quality standards of the evaluation. -Establish the Exam Design Committee (CDE) or the main designer. -Make a first analysis of the curriculum. -Determine the universe of content to evaluate.
	1.2 Analysis of Resume	-Analyze the curriculum. -Develop the referent of contents to evaluate.
	1.3 Development of the plan evaluation	-Develop the specifications of the items -Design the structure of the exam.
	1.4 Production, validation and item piloting	-Select and coordinate the Development Committee of Items (CIS) or failing that the main designer. -Develop items according to specifications. -Evaluate the congruence item-specification-curriculum and possible biases in the items (procedures performed informally) in the table evaluated contents (supplemented) -Structure the first version of the exam. -Ship the exam to consult experts -Analyze recommendations. -Determine content validity. -Pilot the EA6B3y4 Analysis of the technical quality of items
Phase II	Stages	Procedures
Settings items of exam	2.1 Review of Items	.TCT analysis -Analyze item failures. -Adjust the items according to the type of fault.
	2.2 Structuring of the model of exam	-Structure the final version of the exam (if the calibration justifies it) -Apply the test to the selected samples. Analysis

Table 1 Proposal for the design of the evaluation of the learning of the third and fourth bimester in sixth grade of primary education (EA6B3y4)

As you can see, in this table two phases are revealed for the development of the test: On the one hand, the logistics, design, validation and piloting of the exam; and on the other the settings of exam items. Subsequently, the aspects of each of the phases will be addressed.

Methodology

The present investigation corresponds to the quantitative approach, and is of an instrumental type, in accordance with the categories set forth by Montero and León (2005) since all studies aimed at the development of tests and devices, including both the design (or adaptation) as the study of their psychometric properties.

The collection-accumulation of evidence of validity of an academic performance test consists of going through a process that, through different statistical procedures supported in the Classical Theory of the tests and carried out to a database built from the application of the instrument in question, determines the psychometric properties of its measurement process and the scores obtained in its application In the present study, reliability and evidence of validity, based on the content and internal structure, which are considered as psychometric properties of the designed instrument are revealed. The sample consisted of 81 students of sixth grade of primary education in the City of Durango, Mexico, was constituted under non-random procedures for accessibility to informants.

Logistics, design, validation and test piloting

The evaluation of the learning of the third and fourth bimester in sixth grade of primary education (EA6B3Y4) is presented as a theoretical-methodological proposal aligned to the curriculum since the 2011 Curriculum for Basic Education in Mexico, can be recognized from a reference trial criterial characterized as low impact (Nitko, 1994; Popham, 1990 & Ravela, 2006).

Bearing in mind that as a new thesis or researcher, the tools and relationships to build an academic test validation committee are not available on different occasions, the recommendation of Barraza (2010) is proposed in this proposal, which proposes to rely on the known academics in what he calls the critical friend.

When the individual educational agent cannot contact other agents, and participate in a network, it is necessary to locate a partner or friend who accepts to serve as a friend critical. The main function of the critical friend is to serve as an interlocutor to discuss, analyze and reflect, jointly, on the actions that are developed for the elaboration of the Project or Proposal.

Who can be a critical friend?

When referring to the critical friend I don't know Are you thinking of a specialist or expert, do not think of an educational agent with teacher or doctor degree, you simply think of a classmate and / or friend who covers.

The following features (Barraza, 2010, p. 32):

- Be willing to listen, or if necessary read, what the educational agent has to well share with him either way.

- Show respect for the logic of action developed by the educational agent.

- Have availability of time to share and support the experience that the educational agent.

The Examination Coordinating Committee (CCE) was constituted by Dr. Juan Manuel Coronado Maqueros, Dr. Omar David Almaraz and a server. (The two primary school teachers and Doctors of Science for Learning), who took care of the process in the definition and development of the proposal. The establishment of the quality standards of the evaluation was retaken based on the technical criteria for the development and use of educational evaluation instruments, 2014-2015 of the INEE (2014). In the phase that has to do with the Exam Design Committee (CDE) it was decided to design the instrument individually, but respecting the technical criteria and the parameters established by the CCE.

Subsequent activities related to the analysis of the curriculum, as well as the delimitation of the benchmark to evaluate and the design and conduct of the exam were worked by the main designer except for the development of specifications of the items which were designed by the CCE.

It can be pointed out that all the expected learning for bimester III and IV of Spanish were approached in accordance with Table 2:

Characteristics of the evaluated contents of Spanish by item. It will be observed that they were taken into account according to the technical characteristics of the reagents, four sections in four columns of the table: Expected learning and content according to the Basic Education Curriculum (2011), Specifications according to the CENEVAL guidelines (2013) and the processing levels according to the New Taxonomy of Marzano and Kendall (2007).

Regarding this last criterion, it can be mentioned that only the first four levels of processing (NdP in the table) were taken into account: Recovery (1), Comprehension (2), Analysis (3) and Use of knowledge (4), corresponding to the cognitive system and located in the first and second domain of knowledge: information and mental procedures, which will not be taken into account specifically for the analysis since the main interest is focused on the relationship of the question to the level of processing which evokes both in the subject of Spanish and mathematics.

Item	Expected learning	Contents	Specification	NoP
1	It establishes the order of the events reported (succession and simultaneity).	Succession and simultaneity, and cause and consequence relationships in historical accounts.	Identify the cause-consequence relationship in historical accounts	3
2	Infers dates and places when the information is not explicit, using the clues that the text offers.	Inference of dates and places from the clues offered by the text itself.	Infer dates and places from a historical text	1
3	It recognizes the function of historical accounts and uses the characteristics of formal language when writing them.	Characteristics and function of historical accounts.	Identify the characteristics of historical accounts when writing them.	2
4	Write a text in paragraphs, with conventional cohesion, spelling and punctuation.	Regular spelling patterns for times past (accentuation in the third person singular in the past simple, endings in co-past, derivations of the verb to have).	Use regular spelling patterns in tenses as copretérito.	4
5	Recognize the structure of a play and the way it differs from stories.	Characteristics of plays (similarities and differences with stories).	Recognize the differences and similarities of plays and stories by reading them.	3

6	Use verbs to introduce indirect discourse in narratives and dimensions.	Verbs to introduce indirect discourse in narratives and dimensions.	Use verbs to introduce indirect discourse in narratives.	4
7	Use question marks and exclamation marks, as well as dimensions to show intonation in dramatization.	Question marks and exclamation marks to emphasize intonation.	Use exclamation marks to emphasize intonation	3
8	Interpret a text properly when reading it aloud.	Narrative voices in plays and stories.	Recognize intonation in a story from narrative voices	2
9	Identify the structure of the opinion letters.	Characteristics and function of formal and opinion letters.	Identify the characteristics of opinion letters through their writing	1
10	Identify the differences between expressing an opinion and referring a fact.	Ways to write an opinion based on arguments.	Recognize the characteristics of the opinions argued through their structure	3
11	Contrast information from texts on the same topic.	Differences and similarities in the treatment of the same subject.	Distinguish differences and similarities in a given topic	4
12	Use logical connectives to link the paragraphs of a text.	Use of logical connectives to link the paragraphs of a text (unlike, on the contrary, also, on the other hand, however, among others).	Use logical connectives to link paragraphs in a text.	3
13	Recognizes various practices for the treatment of discomforts.	Cause and consequence relationships between the origin of an upset and its treatment.	Relate the cause and consequence of some discomfort with its probable treatment.	2
14	Meet and appreciate different cultural and linguistic manifestations of Mexico.	Literary expressions of Mexican traditions.	Recognize literary expressions belonging to the Mexican tradition.	1
15	Understand the meaning of songs from the oral tradition.	Meaning of the texts of the Mexican oral tradition (songs in indigenous language).	Understand the meaning of songs in the indigenous language through their characteristics in writing.	2
16	Identify some differences in the use of literary resources between Spanish and some indigenous language.	Linguistic diversity of the country.	Identify differences in the use of literary resources between Spanish and some indigenous language	3
17	Identify words and expressions that indicate time and space in personal letters.	Words and expressions that denote time and space in personal letters from the date of the letter and the sender's information.	Recognize words that denote time and space in personal letters	2
18	Know the data structure of the postal and / or electronic addresses of the recipient and sender.	Data structure of the conventional and / or electronic addresses of the recipient and sender.	Identify the structure of the data to send a conventional email	1
19	Adapt the language to target known recipients.	Production of written texts considering the potential recipient.	Make writings adapting the language depending on the recipients	3
20	Complete data forms effectively to obtain a service.	Characteristics of the forms for opening an e-mail account.	Complete forms for opening an email.	4

Table 2 Characteristics of the evaluated contents of Spanish by item

In the case of mathematics, in addition to addressing all the expected learning, all the topics of reflection of each of the axes were also taken into account according to the Curriculum (2011), taking the example as a reference for its structuring and distribution from the previous table. The expert consultation phase was basically used to establish whether the items of the instrument adequately represent the construct to be measured. (Barraza, 2007). This procedure is defined as an informed opinion of people with experience in the subject, which are recognized by others as qualified experts in the subject, who can give information, evidence, judgments and assessments. The identification of the people who will be part of the expert judgment is a critical part in this process, against which Skjong and Wentworht (2000) propose the following selection criteria: (a) Experience in making judgments and making decisions based on evidence or expertise (degrees, research, publications, position, experience and awards among others), (b) reputation in the community, (c) availability and motivation to participate, and (d) impartiality and inherent qualities such as self-confidence Same and adaptability. In order to carry out this activity, 4 experts in the field of evaluation were requested to take into account under a previously designed scale the relevance of the items in the designed instrument. The names of the experts who were asked to perform the judge in response to confidentiality will not be mentioned, but the following information is explained in this regard.

- INEE Area Chief. Who is an expert in design and validation of evaluation instruments.

- Head of the Evaluation Department of the Ministry of Education in a federal entity in Mexico.

- Collaborator in the design, validation and reagent construction committees of CENEVAL.

The opinion and argument of the experts served to accumulate evidence of apparent validity in some cases and content validity. Contreras (2000), emphasizes that the quality of a criterial test is judged and constantly contrasted with the educational objectives intended by the curriculum. The above makes sense when considering that the validity of content is a highly relevant indicator for the development of criterial evaluations. The concept of validity in this regard refers to the degree to which the value judgments made in the evaluation are adequately supported by empirical evidence and are effectively related to the "referent" defined for the evaluation.

To carry out the consultation activity, the experts analyzed the instrument taking into account different criteria such as sufficiency, clarity, coherence and relevance taken into account as indicators inherent in the discrimination process (Escobar & Cuervo, 2008) through five large aspects: Relationship with the curriculum, Writing and spelling, Technical quality of the items, Temporary indicated and Plausibility of distractors. From this they made a discrimination by item on a scale of 0 to 3 (where 0 is unacceptable, 1 regularly acceptable, 2 acceptable and 3 very acceptable), determined the importance of the items from an average value and provided suggestions for its adjustment according to different arguments as was the case.

The results are shown regarding the averages in the consultation of experts reflect as final average 2.63, considered in the scale of Barraza (2007), as a strong validity.

The evidence of content validity was complemented previously described by calculating the agreement between judges. In this regard it is taken into account that the consultation of experts is still considered valid with a high degree of subjectivity, however, the procedure to quantify the data and obtain final scores entails a rigorous and statistically reliable procedure. This same procedure was used to apply a new model (Kappa Analysis) to determine the agreement between judges. The results in this regard showed from the 40 cases analyzed at a value of .706 and a typical asymptotic error a significance of .000. Based on the foregoing and according to the value of the Kappa analysis, the reference is considered satisfactory according to the scale of Cerda and Villarroel (2008).

Exam item settings

Various criteria of technical quality and psychometric standards were established for the development of the test. In principle, it is necessary to recognize that the test was conceived as a small-scale evaluation because it was only applied in three primary education groups (Aiken, 1996). However, since the development of the test began, the technical quality standards for the development of large-scale national and international tests were mostly adopted.

In general, the organization of quality control areas built by Pérez (2010) was taken to ensure the quality of the development of criteria exams. These control areas are similar to those proposed by Nitko (1994) in its model. In particular, there are three control areas: a) quality of the content of the test items, b) technical quality of each item and c) quality of the test scores. Below, the areas of quality control and the standards established for the development of this evaluation are shown.

After the instrument was made, we proceeded to the production, validation and piloting of items that integrated the test. Based on the specific specifications of the items, the main designer developed 40 items trying to proceed according to the quality standards established by international organizations such as the AERA, APA and the NCME, and thus consolidate the validity of the test. To address the items empirically, two pilots were carried out with 40-item instruments, taking as samples two Primary Schools of the City of Victoria of Durango, Durango, taken into consideration for having contextual and organizational attributes of great similarity.

It is necessary to recognize that the main designer developed the items according to the specifications, in turn applied the exams so no applicator training was needed.

After having piloted the items and the test model, the results were captured and their analysis was carried out, using the information contained in the answer sheets and using software for the formation of the base and the analysis of data such as SPSS version 22 and Microsoft Excel

The purpose of the psychometric analysis of the items of EA6B3 and 4 through the Classical Test Theory was to calibrate and assess them in the light of the technical quality standards adopted in the present study.

The psychometric indicators that were promptly analyzed are the difficulty index, the discrimination index, the high-low percentage in the test results, the correlation coefficient of contrasted groups and the Kuder Richardson internal consistency coefficient (KR-20).

The procedure for obtaining the first three psychometric indicators mentioned was generated from 3 main equations. The first that was required to obtain the reagent difficulty index was equation (1):

$$pi = Ai/Ni \qquad (1)$$

In this equation (1) pi is the reagent difficulty index, Ai is the amount of successes in the reagent and Ni is the amount of successes plus the amount of errors in the reagent. Equation (2) that was used to obtain the discrimination index (high-low) was:

$$Di = \frac{GAi - GBi}{N_{grupo\ mayor}} \qquad (2)$$

In this equation (2) Di is the discrimination index of reagent i, GAi is the amount of reagent successes of 33% of those who obtained the highest test scores, GBi the amount of reagent successes of 33 % of examinees who obtained the lowest scores on the exam, and N is the number of people in the largest group (GAi or GBi). The equation that was used to obtain the internal consistency coefficient (KR-20) of the instrument was:

$$KR\text{-}20\ es\ [n / n-1] * [1- (\Sigma p * q) / Var]$$

where:

n = sample size for the test,

Var = variance for the test,

p = proportion of people who pass the article,

q = proportion of people who fail the article.

Σ = summarize (add). In other words, multiply each question p by q, and then add all. If you have 10 elements, multiply p * q ten times, then add those ten elements to get a total.

To analyze the percentage of highs and lows in the test, it is necessary to divide the total test scores into a high group and a low group, considering 33% of the highest scores and 33% of the lowest scores. . In this study, in the first pilot 10 students were taken for each group, upper and lower and taking into account that a group with the same number of students was piloted, the same reference was taken.

In Table 3: First pilot, the result of the analysis of items of the EA6B3y4 after its first phase of piloting the sample of students is shown. In this first result, the psychometric indicators are identified: total successes per item, difficulty index, discrimination index (low-high), the correlation coefficient and the internal consistency coefficient. Previously, some indicators for the analysis from the TCT are presented below.

N = Total number of students evaluated
AL = High average
K = Total reagents
BA = Low average
T = Total average of correct answers
ID = Discrimination index
KR-2O = Reliability coefficient
IF = Difficulty index

KR-20= .776	N=30		K= 45	T=27
ITEM	**IF**	**AL**	**BA**	**ID**
1	0.87	0.9	0.9	0
2	0.87	0.9	0.9	0
3	0.71	1	0.5	0.5
4	0.9	1	0.8	0.2
5	0.84	1	0.9	0.1
6	0.35	0.6	0.1	0.5
7	0.32	0.5	0.1	0.4
8	0.9	1	0.8	0.2
9	0.68	0.7	0.7	0
10	0.74	0.9	0.7	0.2
11	0.35	0.5	0.3	0.2
12	0.32	0.5	0.1	0.4
13	0.39	0.4	0.6	-0.2
14	0.48	0.9	0.1	0.8
15	0.81	1	0.6	0.4
16	0.45	0.8	0.2	0.6
17	0.9	1	0.9	0.1
18	0.26	0.4	0.3	0.1
19	0.42	0.5	0.3	0.2
20	0.77	1	0.5	0.5
21	0.74	0.9	0.5	0.4
22	0.97	1	1	0
23	0.19	0.4	0.2	0.2

24	0.68	1	0.5	0.5
25	0.87	1	0.7	0.3
26	0.77	1	0.7	0.3
27	0.68	0.7	0.7	0
28	0.65	0.7	0.7	0
29	0.84	0.9	0.9	0
30	0.84	1	0.8	0.2
31	0.19	0.3	0	0.3
32	0.61	0.8	0.6	0.2
33	0.26	0.2	0.3	-0.1
34	0.55	0.9	0.5	0.4
35	0.39	0.5	0.4	0.1
36	0.45	0.6	0.4	0.2
37	0.65	0.9	0.5	0.4
38	0.77	1	0.4	0.6
39	0.65	0.9	0.5	0.4
40	0.84	0.9	0.9	0
TOTAL	0.62	0.77	0.53	0.24

Table 3 First pilot

As can be seen in the previous table, 22 items show some type of irregularity, of which those below the quality standards in the discrimination index stand out (Ebel & Frisbie, 1986) equivalent to P> 0.30 and the Correlation coefficient which, as it should be, confirms the index score previously described especially in items 13 and 33 when negative values are found.

The total difficulty index could be considered as an attribute in this pilot when scoring on average at .62 but when analyzed separately, 16 items can be found with an IP outside the established range (> 0.30 and <0.80). Once the values were analyzed by means of some TCT criteria, calibration was carried out through two activities: Review of the items and the distractors and adjustment according to the results. During this stage, the main designer made a thorough analysis of the psychometric indicators of the items and the exam model. Based on the result of the analysis in the first pilot, some of these indicators were identified that did not meet the established quality standards. The objective of the review was to identify the type of failure presented by the items to make the corresponding adjustments. Therefore, a new pilot was carried out with another group also of 30 students, leaving again 40 reagents (modified in some cases) in order to keep the construct valid. The results of the second pilot are presented in Table 4: Second pilot, in which significant changes can be seen from the analysis made for decision making according to the difficulty of the items, the characteristics and plausibility of the distractors (response options), to address these shortcomings.

	KR-20= .787	**N=30**	**K= 40**	
Num.	IF	High	LOW	ID
1	0.62	0.86	0.29	0.57
2	0.71	0.86	0.43	0.43
3	0.62	1.00	0.14	0.86
4	0.67	1.00	0.57	0.43
5	0.52	0.86	0.43	0.43
6	0.43	0.86	0.14	0.71
7	0.33	0.57	0.14	0.43
8	0.62	0.86	0.29	0.57
9	0.57	0.71	0.29	0.43
10	0.76	0.86	0.57	0.29
11	0.57	0.71	0.14	0.57
12	0.62	0.86	0.14	0.71
13	0.38	0.57	0.29	0.29
14	0.48	0.86	-	0.86
15	0.76	1.00	0.43	0.57
16	0.48	0.86	0.14	0.71
17	0.76	1.00	0.57	0.43
18	0.52	0.57	0.29	0.29
19	0.57	0.71	0.29	0.43
20	0.57	1.00	0.43	0.57
21	0.76	0.86	0.43	0.43
22	0.71	1.00	0.43	0.57
23	0.43	0.71	0.29	0.43
24	0.52	1.00	0.14	0.86
25	0.57	0.86	0.29	0.57
26	0.57	1.00	0.29	0.71
27	0.67	0.86	0.43	0.43
28	0.48	0.57	0.14	0.43
29	0.67	1.00	0.43	0.57
30	0.52	0.71	0.14	0.57
31	0.62	0.86	0.29	0.57
32	0.62	0.71	0.29	0.43
33	0.43	0.57	0.14	0.43
34	0.57	1.00	0.14	0.86
35	0.62	1.00	0.57	0.43
36	0.57	0.71	0.29	0.43
37	0.62	0.86	0.29	0.57
38	0.57	1.00	0.14	0.86
39	0.71	1.00	0.43	0.57
40	0.71	1.00	0.43	0.57

Table 4 Second Pilot

Conclusions

In the course of this review article, a general model for the construction of reagents has been evaluated with the intention of designing a test. For the present work different operative phases were established from the model of Pérez (2010), Nikto (1994) and Contreras (2009).

As can be seen, the results of the pilings (particularly the second one) show a significant adjustment according to the attributes of EA6B3 and 4 since the total scores of the difficulty and discrimination indices point to a greater extent towards .50, respecting the criterion> 0.30 and <0.80 Featured in advance.

It is worth recalling that two reagents are slightly below the quality criterion, of which reagent 13 attracts attention, which had already presented a low index of discrimination since the first pilot. In summary terms, it can be said that the test meets the quality standards for low-scope instruments and therefore, its use in research and as a regular application of tests is feasible. The general summary of results taking into account that in the two pilots 40 reagents were worked, is presented below:

On the average difficulty index, the quality standard p> 0.30 and <0.80 must be recognized, which in the first pilot showed 0.62 while in the second one, 0.51, finding no significant problem from the first application. The criterion that did show a need for adjustment in the first pilot was the discrimination index, a medium since the quality standard emphasized p> 0.35 and in the first pilot it remained at 0.24, while for the second pilot it reached 0.37.

The reliability of the instrument was determined from the KR-20 statistic (> 0.70), to be in the second pilot at 0.787 and although it might seem low, it is within the criteria stipulated at the methodological level and is seen as respectable from the scale of Barraza (2007). With the above, quality criteria are required for the design of tests for educational and research purposes.

References

Aiken, L. (1996). *Tests psicológicos y evaluación.* México: Prentice Hall Hispanoamericana.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Barraza, A. (enero, 2007) Confiabilidad? *Investigación Educativa Duranguense.* Recuperado de dialnet.unirioja.es/descarga/articulo/2292993.pdf (6)

________ (septiembre, 2007). La consulta a expertos como estrategia para la recolección de evidencias de validez basada en el contenido. *Investigación Educativa Duranguense*, *2*(7) Recuperado de http://www.upd.edu.mx/PDF/Revistas/InvestigacionEducativa

________ Elaboración de Propuestas de Intervención. Durango: UPD. Disponible en http://www.upd.edu.mx/PDF/Libros/ElaboracionPropuestas.pdf

________ (2019). *Validación de pruebas de rendimiento académico.* Durango: UPD. Disponible en http://www.redie.mx/librosyrevistas/libros/validacionpruebas.pdf

Cerda, J. y Villarroel, L. (enero-febrero, 2008) Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. *Revista Chilena de Pediatría*. *79*(1). Disponible en: http://www.scielo.cl/pdf/rcp/v79n1/art08.pdf

Contreras, L. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California.* Tesis de maestría publicada en línea. UABC. Recuperado de http://iide.ens.uabc.mx/blogs/mce/files/2010/09/Luis-Angel-Contreras-Nino.pdf

Ebel, R.L. y Frisbie, D.A. (1986). Essentials of Education Measurement. Englewood Cliffs, NJ: Prentice Hall.

Gallardo, K. (2013). *Evaluación del aprendizaje: retos y mejores prácticas*. Monterrey, México: Editorial Digital Tecnológico de Monterey

Gallardo, K. E., Gil, M. E., Contreras, B., García, E., Lázaro, R. A. y Ocaña, L. (julio-diciembre, 2012). Toma de decisiones para la evaluación formativa: el proceso de planeación y determinación de sus mecanismos. *Sinéctica,* (39). Recuperado de http://www.sinectica.iteso.mx/index.php?cur=39&art=39_08

Gallardo, K. y Gil, M.A. (2011). *Incorporación de la Nueva Taxonomía como Referente para el Diseño de Herramientas de Evaluación del Aprendizaje Conducida en una Materia de Posgrado en Entornos Virtuales*. Memorias del XI Congreso Nacional de Investigación Educativa. Monterrey, Nuevo León. Recuperado de http://www.ruv.itesm.mx/convenio/catedra/recursos/material/cn_27.pdf

Gallardo, K. y Valenzuela, J. (2014). Evaluación del desempeño: acercando la investigación educativa a los docentes. Revista de evaluación educativa, 3 (2). Recuperado de: http://revalue.mx/revista/index.php/revalue/issue/current

Haladyna, T. (2004). *Developing and validating multiple-choice test items* (3ª ed). New York: Routledge.

Marzano R.J. y Kendall, J.S. (2007). *The new taxonomy of educational objectives*. California, EE.UU.: Corwnin Press.

Marzano, R.J. (2000). *Designing a new taxonomy of educational objectives*. Thousand Oaks, CA: Corwin Press.

Montero, I. y León, O. (2005). Sistema de clasificación del método en los informes de investigación en psicología, Internacional Journal of Clinical and Health Psychology, 5, (1), 115-127.

Nitko, A. (1994, julio). *A model for developing curriculum-driven criterion-referenced and norm-referenced national examinations for certification and selection of students.* Trabajo presentado en la Asociación para el Estudio de Evaluación de la Educación en la Conferencia del Sur de África (ASEESA)

Pérez, J. (2010). *Evaluación criterial del área metodológica de la carrera de Psicología de la UABC*. Tesis de maestría publicada en línea. Universidad Autónoma de Baja California. Recuperado de http://iide.ens.uabc.mx/blogs/mce/files/2010/11/Tesis-Master-Juan-Carlos-P.-M.-2010.pdf

Popham, W. (1990). Modern educational measurement: practical guidelines for educationals leaders. Michigan: Allyn and Bacon

Ravela, P. (2006). *Fichas didácticas para comprender las evaluaciones educativas*. Chile: PREAL

Skjong, R. & Wentworth, B. (2000). *Expert Judgement and risk perception*. Recuperado de http://research.dnv.com/skj/Papers/SkjWen.pdf.