Standardized evaluation of learning at UABC: Psycometric analysis innovation

Evaluación estandarizada de los aprendizajes en la UABC: innovación desde el análisis psicométrico

http://dx.doi.org/10.32870/Ap.v12n1.1698

Jorge Gustavo Gutiérrez Benítez*
Luis Alan Acuña Gamboa**

Keywords
Standardized evaluation; psychometric; methodology; educational quality; test

ABSTRACT

The evaluation has had a strong impact on teaching, which is why learning instruments that are valid and reliable are indispensable. The Faculty of Languages of the Universidad Autónoma de Baja California does not implement a technological tool for the development of standardized tests and their psychometric analysis. Therefore, this paper analyze the implementation of standardized tests referred to a criterion, based on the Item Response Theory. The used methodology required the integration of several committees of specialists to produce instruments that guided the construction of the test, such as the curricular relevance index table, the contents justification, etc. The test quality was evaluated using an innovating development that automatically quantified the psychometric criteria of validity and trustworthiness like the difficulty index of the items and the discrimination index. The results allow to observe the quality of the test identifying the yield of each item, as well as the yield and level of knowledge of the participant students sample, which contributes to the realization of more accurate assessments of their academic performance. It is concluded that the instrument showed utility for both formative and summative evaluation.

Palabras clave
Evaluación estandarizada; psicometría; metodología; calidad educativa; examen

RESUMEN

En virtud del fuerte impacto que ha tenido la evaluación en la enseñanza, resultan indispensables instrumentos de evaluación del aprendizaje que sean válidos y confiables. La Facultad de Idiomas de la Universidad Autónoma de Baja California no implementa una herramienta tecnológica para el desarrollo de exámenes estandarizados y su análisis psicométrico, de ahí que este trabajo analice la aplicación de exámenes estandarizados referidos a un criterio y basados en la teoría de respuesta al ítem. La metodología empleada requirió la integración de varios comités de especialistas para producir instrumentos que guiaran la construcción del examen, como la tabla de índices de relevancia curricular y la justificación de contenidos. La calidad del examen se midió mediante un desarrollo innovador que cuantificó automáticamente los criterios psicométricos de validez y confiabilidad, como el índice de dificultad de los ítems y el índice de discriminación. Los resultados permiten observar la calidad del examen e identificar el rendimiento de cada ítem, así como el nivel de dominio de la muestra de alumnos participantes, lo que contribuye a la realización de valoraciones más exactas de su desempeño académico. El instrumento mostró utilidad tanto para la evaluación formativa como para la sumativa.

Received: June 21, 2019 Accepted: November 11, 2019
Online Published: March 30, 2020

* PhD student of Innovation in Education Technology, Autonomous University of Querétaro. Full-time academic technician at Autonomous University of Baja California, Mexico. ORCID: https://orcid.org/0000-0003-3392-6398
** PhD in Regional Studies, Autonomous University of Chiapas. Research professor at Autonomous University of Chiapas, Mexico. ORCID: https://orcid.org/0000-0002-8609-4786

INTRODUCTION

Different educational purposes require of diverse tests and their use. When a test is not closely related to its purpose, it is difficult to make valid inferences from its results. Developing tests for the standardized assessment of learning (Fernandez, Alcaraz & Sola, 2017; Marquez, 2014) is a delicate task, as assessing with a wrongly designed instrument may have a negative impact both on the examinee and on the teacher; even to measure knowledge aspects other than the ones to be assessed is detrimental to the student.

To Hernandez, Ramirez & Gamboa (2018), one of the toughest challenges for educational institutions is the identification, through evaluation processes, of the capacities, knowledge and skills of students, so that they adapt to the plans, programs and educational methods aimed to improve the teaching and learning process. The foregoing makes the importance of methodological devices evident in the development or the preparation of tests, because, through them, the quality of the instrument may be assured, as well as its contribution in the attainment of valid and reliable information.

In higher education institutions of industrialized countries, it is common that assessments be used, designed and validated for admission purposes. In the United States, for example, since 1926, the Scholastic Aptitude Test has been used to enroll in bachelor studies; since 1949, the Graduate Record Examination, for post-degree; and since 1964, the Test of English as a Foreign Language to prove knowledge in the English language (Tirado *et al.*, 1997).

In the specific case of Mexico, in 1994, the National Center for the Assessment of Higher Education (Ceneval, by its acronym in Spanish) was created with the purpose of conducting a national examination, by way of indication, that would provide reliable and valid information on the knowledge and skills acquired by individuals who are beneficiary of the different educational programs; the test would be used as an indicator for participating higher education institutions. The concept of Ceneval was intended to see to country needs,

> of having a national test, by way of indication, prior to doing undergraduate studies [...] [that would allow] the institutions themselves, to governmental authorities and the society, generally, to assess the skills and basic knowledge of students aspiring to do an undergraduate degree and a general professional quality test [...] [that would enable] awareness of the relevance and pertinence of academic education of the new

> professionals in the country (National Center for Higher Education, 2017, p. 14).

The foregoing needs became into what is nowadays known as the National Entrance Test to Medium-Higher Education, National Entrance Test to Higher Education, National Entrance Test to Post-Degree Studies and General Undergraduate Degree Exit Test, unavoidable mechanism in the subject matter of selecting, promoting and obtaining academic tassels in the Mexican education system.

At present, the Basic Competency Test (Excoba, by its acronym in Spanish), formerly Exhcoba (Gongora, Rocha & Verver, 2015; Perez, Lazarrazolo & Backhoff, 2015), enables the inclusion of multimedia elements to enrich the educational experience and offer automatic results. This test is an innovative evaluation proposal on school competencies, as it departs from the multiple-choice format and approaches more "authentic or natural" forms to assess learning (Ferreyra & Backhoff, 2016); however, the Excoba structure is in line with the national curriculum, therefore it evaluates basic academic competencies specified in mandatory education syllabi.

The Faculty of Languages of the Autonomous University of Baja California (UABC, by its acronym in Spanish) does not have a technological tool that would help in the construction of standardized tests and which, in turn, may be valued by means of a psychometric analysis. Although there is software that enables the conduction of these analysis, prior knowledge is required by users; furthermore, it is necessary to build source archives in technical formats, which limits the use of experience personnel in these fields of knowledge. The foregoing opens the likelihood of innovation in the examination construction process to assess students, also to measure quality from a psychometric perspective in a more efficient way.

The following paragraphs detail, first off, the chosen method to lead the development of the test, which includes the description of instruments used for the construction and evaluation thereof. The numerical interpretation is explained for the quality values that are going to be measured and what is implied thereby. The results section shows the technical value of the psychometric analysis, which determines the quality and reliability of the exam, such as the difficulty index and discrimination. Additionally, data are presented with data obtained using three of the main instruments that lead to the preparation of the exam. In the last section, the relevance of

having standardized evaluation instruments is concluded as well as the advantages of having automated software for the application, the construction and psychometric analysis of the exams.

DESIGN

By referring to the general object of this research (designing an innovative proposal of a psychometric analysis for standardized learning examinations of the Faculty of Languages of UABC), the methodology employed in structuring this exam was formulated at the Institute of Investigation and Educational Development (Contreras, 2000; Contreras & Backhoff, 2004; Contreras, Encinas & De las Fuentes, 2005) based on the psychometric model originally proposed by Nitko (1994) to prepare large scale tests of criteria reference, guided by the curriculum.

Development of these exams required definition of certain stages of the methodology to prepare them in a standardized way and with quality. Moreover, for the last two stages in the process, which imply the analysis of the behavior of items, it is necessary to perform a specialized statistical analysis which requires *ad hoc* software design to do these calculations.

An important element in the methodology is the formation of committees that would be in charge of performing the activities for each phase in the development of the exam. The designing committee prepares the general design of the test; among these actions the following may be done: quoting the analysis of the curriculum, preparing the grid and analyzing curricular relevance. Another committee is responsible for the specifications for the construction of items and to detail the standard structure they should have.

The items committee is in charge of producing them in accordance with the features in the specifications. The opinion committee performs an evaluation of the different aspects, such as technical quality of the item, cognitive validity, format and edition, among others. The last committee is the computing analytics committee, which carried out psychometric evaluation processes of the items by processing information with a specialized software.

Another important detail in the formation of these committees is that they are constituted by a teacher who is a specialist in the

subject, a teacher who is an expert in the curriculum, a teacher with years of experience in education and a teacher with full practice in assessment. These committees may, in turn, prepare or train other teachers so that they partake in the performance of the activities.

Based on the above, the sample in the administration of this test included 394 students, who are in the first semester of the common core of the Bachelor in Languages. The subject this test was designed for was Morphology of the second language, which is a series of Morphosyntax, both of them weighted with the highest curricular relevance for the common core. Gender of the population was 50% males and 50% females; 33% of students were selected from them with low performance (less than or equal to 6.9), 33% of students with a regular performance (between 7 and 8.9) and 34% of students with high performance (greater or equal to 9).

The variables measured to assess the quality of the test are established by the item response theory (Gomez, 2015; Mola *et al.*, 2013; Muñiz, 2010; Mayaute & Vazquez, 2010; Lord & Novick, quoted in Kramp, 2008), which addresses a model and a set of criteria that, if reached, guarantee accuracy in the students' proficiency measure on a specific learning. These criteria include the item difficulty index (*p*), the discrimination index, the item-total correlation and the option-total relation of distractors.

The difficulty index is measured by a range of values from 0 to 1; the closer the index value is to 1, the easier it is to answer the question, and vice versa. The same thing happens with item discrimination, as it measures between 0 and 1, and the discrimination will be better the closer this value is to 1. The item-total correlation should at least be placed in a value of .25 and, the closer it is to 1, the item will have greater relation to the highest likelihood of obtaining a good percentage in the whole test if answered correctly.

Instruments

The governing support in the construction of an exam is the subject description letter, a document which establishes the universe of knowledge to be assessed. Therefore, the test is known as an exam referred to a criterion (Leyva, 2011); criterion is specifically understood as the knowledge the exam is addressed to. Some of the instruments employed in the development are the result of the product of an execution of stages pertaining to the methodology to

prepare standardized tests, for example, the grid, justification of content and test specifications.

One of the essential specialized instruments for the final stages in the development is that software that enable a psychometric analysis. The instruments used were ITEMAN (Cechova, Neubauer & Sedlacik, 2014; Thoe, Fook & Thah, 2008) of Assessment Systems Corporation, as well as TAP, published by Brooks & Johanson (2003).

RESULTS

The grid is one of the products resulting from the application of methodology and which is fully useful for teachers. This instrument allowed us to fully appreciate how the thematic contents of the subject were interrelated, and to identify those with a greater impact to define other knowledge.

Figure 1 is an example of a small segment of the grid, where you can see how, by means of arrows, the thematic contents of the first unit are related with those of unit 2, and how they have a greater number of relations with the topics of previous and subsequent units. This type of contents give rise to a greater impact on the students' learning.

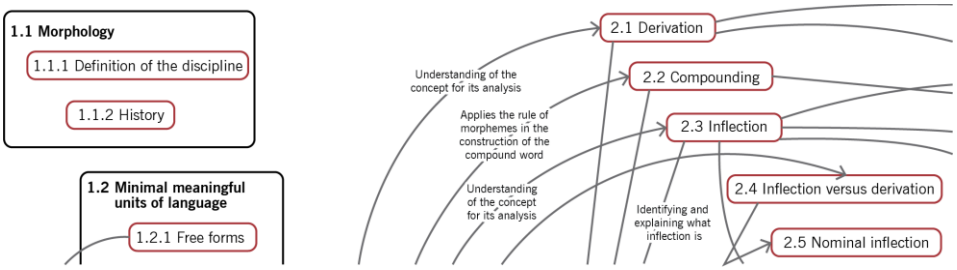


Figure 1. Example of the grid of the subject Morphology of the second language.

The product of the results of the curriculum relevance index (CRI) of the thematic contents is shown in Table 1, while their prioritization, from a greater CRI, is reported in Table 2, which

presents thematic contents as well as practice that is to be done in the subject.

Table 1 shows some of the criteria assessed to obtain the CRI. Each of these attributes for the thematic universe of the subject was evaluated and it was noted how some topics had such a low relevance that, if omitted, there would practically be no impact on the students' learning. This is interesting if attention is centered on designing subject curriculum, whose thematic content, in accordance with the above results, has little contribution to what the student should know and know how to do.

Table 1. Assessed attributes to obtain the curriculum relevance index

Content	Contribution to the accomplishment of the competence of the unit 20%	Planning (amount of implicit content) 10%	Credit hours (assigned or estimated for learning) 10%	Disciplinary relevance 20%	Index of curricular relevance
1.1.1 Definition of the discipline	0.20	0.10	0.10	0.20	0.600
1.1.2 History	0.13	0.07	0.07	0.13	0.400
1.2.1 Free forms	0.20	0.10	0.07	0.13	0.533
1.2.2 Signs and morphemes	0.20	0.10	0.10	0.20	0.633

In view that the purpose of these tests is to assess proficiency of a specific knowledge in possession of students, a large number of items is used of thematic contents, with a greater CRI, because this type of topics required the formation of a larger amount of knowledge and, therefore, this assumes a higher proficiency level. These were some of the results obtained in the first stages of the methodology; however, for the following stages, the results were more technical, as they implied an evaluation of a higher weight in

the development of the test and in measuring quality by means of psychometric analysis.

The psychometric analysis resulting from the pilot application showed the different quality criteria in the item, which were to be measured. In total, 63 items were applied for the Morphology of the second language test; for each, an opinion was given with the purpose of determining whether they were accepted or not, and the reasons of this opinion. General descriptive statistics of the test are shown in Table 2.

Table 2. Test descriptive statistics

Items	Mean of correct answer	Standard deviation	Minimum score	Maximum score	Median of p value	Median of rpBis	Alpha
63	40.211	7.242	27	61	0.638	0.212	0.796

Among the most important data in Table 2, are the medium value of *p* or the test difficulty index; this value was placed in .638, which is a sign that this is a medium difficulty test with a slight tendency to be easier that common. It is also highlighted that, in average, 40 of the 63 questions are correct and the attainment of a score close to perfect as 61 correct answers are reached.

Now then, Table 3 details the results of some items. It is noted, for example, item 25, whose difficulty is very low, that is, is very easy to answer; this is concluded because the *p* value of the item is .789. As mentioned in the previous paragraph, the closer this value is to 1, the easier it is to answer the item or the question.

Table 3. Psychometric analysis of the items

Analysis	IT25	IT26	IT28	IT48
Difficulty (*p*)	Easy item con p=.789	Hard difficulty with .395	Very hard item with *p*=.289	Very hard item with .21
Discrimination	Regular discrimination with .26	Very good discrimination .53	Good discrimination with .33	Negative discrimination with .16

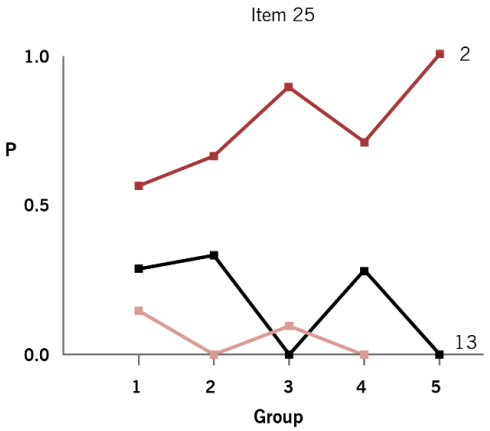
rpBis	Good with .264	Very good with .432	Low with .132	Negative with .271
Distractor	The distractors worked properly. Differences of -.173 y -.091 were found.	The distractors worked properly. Differences of -.173 y -.355 were found.	The distractors worked properly. Differences of -.064 y -.264 were found.	The distractors were chosen by the high and low groups but in equal proportion

The discriminatory power of the item is of a common nature, with a value of .26; furthermore, if a correct answer is given the trend tends to relate to good scores throughout the exam (rpBis of .264).

Another factor noted in item 25 is distractors or incorrect answers. They were chosen in all the cases and in a greater proportion by low groups (students with low performance, averages lower than or equal ty 6.9) or those who obtained a grade at the end of the exam. In order to know whether an item is a good discriminant or not, the values reached by rpBis for the correct answer ought to be greater than .2; incorrect answers or distractors ought to be negative, in the event they are positive, they ought to have lower values than those of the correct answer. Additionally, it should be considered that all of the distractors are to be chosen.

Item 25 is the ideal which is sought in an exam, as it correctly meets the quality standards for all the value criteria mentioned. The chart shows the item behavior.

Chart. Item 25 behavior



In the chart, the three possible answers are represented by red, black and pink lines. Axis *x* or horizontal line symbolizes low, medium and high groups, while *y* or the vertical line, is the population. The red line is for the correct answer and, as noted in the chart, it was chosen with greater proportion by the high group, which would be placed between categories 4 and 5; choosing it was lower by the low group, which is placed between categories 1 and 2. The question was of a very low difficulty, 30 of 38 persons had a correct answer, but even so, the item had a greater proportion of correct answers for the high group.

The results shown were obtained from the psychometric analysis performed by using the SIEXAES software, which automatically generated the technical reports with the descriptive analysis of the exam generally –as shown– and the reliability and validity criteria established as quality indicators, the difficulty index of each item, the discriminatory power, the item-total correlation and the function of each distractor. The charts were prepared in the basis of the analysis performed by using the ITEMAN software.

DISCUSSIONS

Once the methodology for the construction of the standardized exam was applied and the psychometric analysis performed, there was clear indication on how some test items had acceptable quality indicators, whereas others still needed to be corrected. The process of designing and constructing the exam provided enough information to determine whether the instrument was valid and reliable, but not only by analyzing the exam, but also the technological tool used for its construction, application and subsequent analysis. This last item will shortly allow that an answer be given to the main objective of the research: designing a technological proposal for the psychometric analysis of standardized tests.

The proposed tool is intended to remove the difficulties of a teacher to carry out a psychometric analysis of a test, because the software commonly employed for these purposes requires of the experience of users, both to handle it and to interpret the results, since the reports produced are of technical nature.

Thus, once of the proposed phases in the methodology have been applied, a series of products was obtained which allowed us to better know the universe of knowledge that was to be assessed in the exam, as well as to identify, in a graphical and quantifiable manner, the importance of each of the thematic contents of the specific subject.

On the other hand, formation of the different committees that partook in designing the exam was a great contribution to the work, as each of the elements of the proposal presented herein was supported with greater accuracy. The committees were formed from the experience of teachers partaking in the fields of general design of the test, specifications, items, opinion and the computer analytic section, with which more objective opinions were obtained from the exam which enriched the justification stages of contents, the preparation of item specifications, the evaluation thereof, among other stages.

There still are elements that are to be improved, for example, it was noted that of the 63 items comprising the test, there are deficiencies in 12 of them; from these, 8 are for the discrimination index, so the answers considered by these items are to be analyzed. Likewise, more items are still to be produced for the text with the purpose of making different versions of it.

It is also important to track the observations made during the preparation of the table with the curriculum relevance index and the grid on those thematic contents that contribute little to students' learning, as this type of situations have a direct repercussion on the quality of the subject curriculum design. Similarly, continuity should be given to the suggestions and evaluations made by participating students and teachers who made on the SIEXAES software for the improvement of the tool quality and to fulfill the evaluated usability criteria in a more efficient manner, such as the size of objects, colors, navigation buttons, among other elements.

As the educational aspect is resumed of this work, there was an option to use exams referred to a criterion, since the main intention of the exam is to explore the proficiency level of students on a clearly delimited universe of knowledge, that is, to measure the skills and knowledge mastered by them, or not, in a specific field of knowledge. As mentioned, the criterion to guide the development of the exam is in the curriculum of the relevant subject, which specified the thematic contents to be assessed and the sequence thereof. Every decision on what is to be assessed and the manner in which it is done

shall be oriented as established in the curriculum. Therefore, there is an advantage to describe accurately and clearly what is to be measured.

Through this type of exams, individual scores are obtained associated only to the student's performance, and not as a function of group performance where the student is part of, as assumed by normative assessment. Furthermore, the particular score of each student provides the opportunity to design individual strategies for the improvement, which is complicated with normative tests.

The main point of the criteria evaluation is the representativeness of the element in respect to the measuring universe, understood in the educational aspect as consistency between the item and the objective; thus, the student may be certain that what is to be assessed actually is what he/she learned and not a different aspect of knowledge unrelated to his/her learning, whereas to the teacher, this means certainty that the score obtained in a test exactly reflects the student's learning in relation to what has been taught in class and in agreement with thematic contents specified in the curriculum.

In a different instance, it was necessary for the research to recover the opinion of participating teachers and students in the pilot test, for which a satisfaction survey was conducted for a sample based on the convenience criterion; that is, they wanted to give their comments and perceptions voluntarily about this test. The following comments were obtained:

> Student 1: I liked the multiple-choice format, it is easier to answer the exam.
>
> Student 2: The questions were clear and without wasting words, like tricky questions.
>
> Student 3: I feel that the test really put what I knew about the subject to the test, I believe it covered every topic seen.
>
> Teacher 1: A test which follows a quantifiable method to determine whether a question in the test is well done or not gives me reassurance to know that if one of my students gives a correct answer to said question it is because he/she really knows that the question is about and, in a sense, tells me that my learning strategies have helped him to give correct answers.

> Teacher 2: The final score of the test lets me know if I have generally taught well or not what is assumed to be seen in the subject, thinking that the class conforms to what is said in the description letter of the subject.
>
> Teacher 3: In the beginning, I was skeptical about the quality of the test, but once I noted that each topic was assessed in accordance to the experience of more than one teacher, which justified each question, who were not the same teachers who prepared the whole test but that different committees participated during the whole process, and that calculation was made subsequently to show what ought to be and what ought not to be in the test, it made me realize that the score obtained by a student in the test was actually the score he deserved, that this number really meant what he knew about the subject.

By the comments above, the importance of having standardized test has been made manifest. Regardless of whether there is an "n" number of versions of a test, each equally assess a student's learning; there are no more difficult or easier tests. Thus, the teacher and the student are reassured that the instrument is valid and reliable.

CONCLUSIONS

In this first approximation, the software provided data with which important inferences may be made; for example, one of them is related with the number of average correct answers in the test. In this sense, group 2 had a better number of correct answers in average, as group 1 was surpassed by almost ten points. However, in the last unit of the subject, group 1 obtained a better average (6.37) as compared to group 2 (5.53).

In the previous case, it is the type of situations that the academic coordination areas need to identify to make decisions both preventive and remedial in the learning process of students. The intention to collect these data is to foster the analysis process to be aware of the cause for these result variations, which may be a sign of situations like absenteeism of the teacher during that period, some wrongly employed learning strategies, among other factors.

An example of these actions or decisions by these coordination areas may be training course for teachers, remarks or sanctions for malpractice or, to the contrary, recognitions for good performance. By technical and specialized reports which the SIEXAES software generates, the teacher may identify, regarding the thematic content, what deficiencies there are of his/her students or aspects of knowledge specifically where they are not proficient and, therefore,

foster the improvement of and the development of his/her teaching competencies. Likewise, a student may see his/her performance in greater detail, identify what areas he/she masters and what areas he/she does not and, above all, be certain that the instrument he/she is assessed with is reliable, that the areas of knowledge he/she is proficient in are directly related with the subject are actually assessed.

It is worth mentioning that a quality education shall not be the object of measurement only by the types of evaluation instruments they use, as this implies other processes which, similarly, ought to be submitted to assessments. The test is, on its own, a sample of assessment tasks which represent proficiency of a content, specifically. In this sense, for example, learning strategies employed in the course, the relevance of thematic contents that are the object of study, the cognitive levels reached, the instructor's teaching strategies, among many other elements which, by way of synergy, may result in a quality education, are discussed.

Learning based on a constructivist approach, for example, gives greater importance to the teaching and learning process than to the actual contents; this, therefore, implies that what student ought to do is more assessed than what he ought to know. The foregoing may be the cause for some thematic contents of the subject had very low curriculum relevance indexes, topics which, as mentioned above, may be omitted and which would not have greater impact on a student's learning. It may be the case that, when designing the thematic content of the subject, what the student should be capable of was prioritized in relation to what he ought to know.

As the assessment is considered from a cognitive approach, it is required that the purpose of all the instruments to be used is of a cognitive end, which implies that the same items of a test reflect the relationship with these objectives. The results obtained in the technical analysis showed aspects related with this approach; for example, the difference of proficiency attained in unit three between two groups was notorious, if we start from the assumption that the students in both groups had the same hours of dosing to learn said topics. A first approximation to identify the cause for this difference may be the type of learning strategies employed by teachers, or rather, the irregular and improper hours of dosing which, again, shows a possible wrong design in the curriculum.

Upon considering these first results and in spite that several aspects of the test are still to be calibrated, the research provided relevant information: it identified the level of proficiency they have in the subject at hand, it disclosed that for a thematic content in particular, there was a marked difference between one group and the other, and it pointed at likely deficiencies in the subject curriculum as topics are included with no curricular relevance or wrongly employed thematic sequences.

The test, in addition to providing training assessment of the students, also does it for the summative assessment; for example, it may help in predicting the success of a student in subsequent courses in reference to the subject, above all when the student obtains high proficiency levels in thematic contents that are connected with subsequent learning and, in a way, the student's guarantee that, when sitting his/her test, he/she possesses a specific mastery.

It is a known fact that it is important for an educational institution to obtain information on its students' performance, to assess attained learning and to compare said achievements with established goals. For this reason, it is essential to have learning assessment instruments that are valid and reliable; that is, that they are correctly designed and provide assurance on what has been assessed in terms of knowledge and skills for which it was formulated.

REFERENCES

Brooks, Gordon & Johanson, George. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement*, *27*(4), 303-304. https://doi.org/10.1177/0146621603027004007

Cechova, Ivana; Neubauer, Jiri & Sedlacik, Marek. (2014). Computer-adaptive testing: Item analysis and statistics for effective testing, in *European Conference on e-Learning*. Academic Conferences International Limited.

Centro Nacional de Evaluación para la Educación Superior. (2017). *Origen y evolución del Ceneval*. http://www.ceneval.edu.mx/documents/20182/49855/OrigenEvolucionCeneval_2018/f8406659-7d28-4960-9ec1-964d90c76e4c

Contreras Niño, Luis Ángel. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California* (tesis de maestría). Instituto de Investigación y Desarrollo Educativo: Universidad Autónoma de Baja California, Campus Ensenada. http://iide.ens.uabc.mx/images/pdf/tesis/MCE/Tesis%20MCE%20Luis-Angel-Contreras-Nino.pdf

Contreras Niño, Luis Ángel & Backhoff Escudero, Eduardo. (2004). Metodología para elaborar exámenes criteriales alineados al currículo, in Sandra Castañeda (ed.), *Educación aprendizaje y cognición, teoría en la práctica*. México: Manual Moderno.

Contreras Niño, Luis Ángel; Encinas Bringas, José Álvaro & De las Fuentes Lara, Maximiliano. (2005). Evaluación colegiada del aprendizaje en la Universidad Autónoma de Baja California: el caso del examen de Matemáticas I de la Facultad de Ingeniería Mexicali, in *Memoria del VIII Congreso Nacional de Investigación Educativa*. México: COMIE.

Fernández Navas, Manuel; Alcaraz Salarirche, Noelia & Sola Fernández, Miguel. (2017). Evaluación y pruebas estandarizadas: una reflexión sobre el sentido, utilidad y efectos de estas pruebas en el campo educativo. *Revista Iberoamericana de Evaluación Educativa*, *10*(1), 51-67. https://doi.org/10.15366/riee2017.10.1.003

Ferreyra Martínez, María Fabiana & Backhoff Escudero, Eduardo. (2016). Validez del generador automático de ítems del examen de competencias básicas (Excoba). *Relieve-Revista Electrónica de Investigación y Evaluación Educativa*, *22*(1), 1-16. https://doi.org/10.7203/relieve.22.1.8048

Gómez Rada, Carlos Alberto. (2015). Diseño, construcción y validación de un instrumento que evalúa clima organizacional en empresas colombianas, desde la teoría de respuesta al ítem. *Acta Colombiana de Psicología*, (11), 97-113. https://editorial.ucatolica.edu.co/index.php/acta-colombiana-psicologia/article/view/482

Góngora Ortega, Javier; Rocha Hernández, Tania Marlene; Verver, Ingrid Verenice. (2015). La prueba Exhcoba como predictora para la deserción y reprobación en medicina. *Revista de la Escuela de Medicina "Dr. José Sierra Flores" Universidad del Noreste*, *29*(1), 16-24. http://www.une.edu.mx/Resources/RevistaMedicina/2015/Vol29No1.pdf#page=16

Hernández Madrigal, Mónica; Ramírez Flores, Élfego & Gamboa Cerda, Silvia. (2018). La implementación de una evaluación estandarizada en una institución de educación superior. *Innovación Educativa*, *18*(76), 149-170. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-26732018000100149&lng=es&tlng=es

Kramp Denegri, Uwe. (2008). Equivalencia entre los modelos de análisis factorial de los ítems y teoría de respuesta a los ítems en la evaluación de las propiedades psicométricas de los instrumentos de medición psicológica. *Revista Peruana de Psicometría*, *1*(1). https://www.academia.edu/3673226/Equivalencia_entre_los_modelos_de_an%C3%A1lisis_factorial_de_los_%C3%ADtems_y_teor%C3%ADa_de_respuesta_a_los_%C3%ADtems_en_la_Evaluaci%C3%B3n_de_las_propiedades_Psicom%C3%A9tricas_de_los_instrumentos_de_Medici%C3%B3n_psicol%C3%B3gica

Leyva Barajas, Yolanda Edith. (2011). Una reseña sobre la validez de constructo de pruebas referidas a criterio. *Perfiles Educativos*, *33*(131), 131-154. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982011000100009

Márquez Jiménez, Alejandro. (2014). Las pruebas estandarizadas en entredicho. *Perfiles Educativos*, *36*(144), 3-9. http://www.redalyc.org/articulo.oa?id=13230751001

Mayaute Escurra, Luis Miguel & Vásquez Delgado, Ana Esther. (2010). Análisis psicométrico del test de matrices progresivas avanzadas de Raven mediante el modelo de tres parámetros de la teoría de la respuesta al ítem. *Persona*, (13), 71-97. http://www.redalyc.org/articulo.oa?id=147118212004

Mola, Débora Jeanette; Saavedra, Bianca Analía; Reyna, Cecilia & Belaus, Anabel. (2013). Valoración psicométrica de la Psychological Entitlement Scale desde la teoría clásica de los tests y la teoría de respuesta al ítem. *Pensamiento Psicológico*, *11*(2), 19-38. https://ri.conicet.gov.ar/bitstream/handle/11336/23949/CONICET_Digital_Nro.4290b666-0f89-4d7f-8738-01898530318f_A.pdf?sequence=2

Muñiz Fernández, José. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo*, *31*(1), 57-66. http://www.redalyc.org/articulo.oa?id=77812441006

Nitko, Anthony. (1994). A model for development curriculum-driven criterion- referenced and norm-referenced examination for certification and selection of students, in *Conference of Education, Evaluation and Assessment for the Association Studies of Educational Evaluation in Sudafrica (ASEESA)*. Sudáfrica. https://eric.ed.gov/?id=ED377200

Pérez Morán, Juan Carlos; Larrazolo Reyna, Norma & Backhoff Escudero, Eduardo. (2015). Análisis de la estructura cognitiva del área de habilidades cuantitativas del Exhcoba mediante el modelo LLTM de Fisher. *Revista Internacional de Educación y Aprendizaje*, *3*(1), 25-38. https://journals.epistemopolis.org/index.php/educacion/article/view/584

Tirado Segura, Felipe; Backhoff Escudero, Eduardo; Larrazolo Reyna, Norma & Rosas Morales, Martín. (1997). Validez predictiva del Examen de Habilidades y Conocimientos Básicos (Excoba). *Revista Mexicana de Investigación Educativa*, *2*(3), 67-84. http://www.comie.org.mx/revista/v2018/rmie/index.php/nrmie/article/download/1057/1057

Thoe, Ng Khar; Fook, Fong Soon & Thah, Soon Seng. (2009). Use of ICT tool for Item Analysis of a Science Performance Test. *Journal of Educational Technology*, *9*(1), 5-15. http://www.mjet-meta.com/resources/V9N1%20-%20NKT%20-%202009%20-%20ICT%20-%20Online.pdf

HOW TO CITE

Gutiérrez Benítez, Jorge Gustavo & Acuña Gamboa, Luis Alan. (2020). Standardized evaluation of learning at UABC: Psycometric analysis innovation. *Apertura*, *12*(1), pp. 118-131. http://dx.doi.org/10.32870/Ap.v12n1.1698