Validity and reliability of an online test on the phenomena of reflection and refraction of sound

Validez y confiablidad de un test en línea sobre los fenómenos de reflexión y refracción del sonido

http://dx.doi.org/10.18381/Ap.v11n2.1622

Jhonny Medina Paredes*
Mario Humberto Ramírez Díaz**
Isaías Miranda***

Keywords
Validation of instruments, physics education, distance education

ABSTRACT

This work presents the results of a validation and reliability process of a conceptual test on reflection and refraction phenomena of sound waves, a process that was carried out online using Knowledge Applied Technologies (TAC). The development of this process was done by using the Classical Test Theory as a theoretical framework and the use of a website for remote implementation. The use of TAC allowed the validation process to be more agile and extended to a larger sample, allowing application in several universities in Mexico, Colombia, Ecuador, and Chile. The obtaining and analysis of data to achieve the validation and reliability of the instrument was given by means of a system designed specifically for this work, which allowed the realization of the necessary statistics to obtain indicators such as difficulty, discrimination and reliability. Obtaining a test that can be applied and resolve online is a novelty in grounds of the Educational Physics.

Palabras clave
Validación de instrumentos, física educativa, educación a distancia

RESUMEN

Este trabajo presenta los resultados de un proceso de validación y confiabilidad de un test conceptual sobre fenómenos de reflexión y refracción de ondas sonoras, llevado a cabo en línea mediante tecnologías aplicadas al conocimiento (TAC). El desarrollo de este proceso se hizo por medio de la teoría clásica de los test como marco teórico y el uso de sitio web para la implementación a distancia. Las TAC ayudaron a que el proceso de validación fuera más ágil y se extendiera a una muestra mayor, lo que facilitó la aplicación en diversas universidades de México, Colombia y Chile. La obtención y el análisis de datos para la validación y confiabilidad del instrumento se dio por medio de un sistema diseñado específicamente para este trabajo, que permitió la realización de la estadística necesaria para lograr indicadores como dificultad, discriminación y fiabilidad. La obtención de un test que se puede aplicar y resolver en línea resulta una novedad en el medio de la física educativa.

Received: March 26, 2019
Accepted: July 18, 2019
Online Published: September 30, 2019

* Master in Educational Physics by the Instituto Polítécnico Nacional de México. Professor of the Basic Sciences University Teaching Center of the Universidad Austral de Chile Puerto Montt Campus, Chile.
** Ph.D. on Educational Physics Sciences by the Instituto Politécnico Nacional de México. Research Professor at the Advanced Technology and Applied Science Research Center, Legaria Unit, of the Instituto Politécnico Nacional. ORCID: https://orcid.org/0000-0002-3459-2927.
*** Ph.D. in Sciences specialized in Educational Math by the Instituto Politécnico Nacional de México. Research Professor in the Avanced Technology and Applied Science Research Center, Legaria Unit. ORCID: https://orcid.org/0000-0003-2076-7383

INTRODUCTION

According to the Standards for Educational and Psychological Tests, "a test is an assessment instrument or a procedure by which a sample of the behavior of the examinees on a specific domain is obtained and subsequently evaluated using a standardized procedure" (Martínez, Hernández and Hernández, 2006, p. 18). Even when inventories, scales, questionnaires and others are included in this definition, in this paper, the term *test* will represent a test with multiple choice questions with a sole key answer.

Since the 90's of the past century until now, different tests have been developed in physics to discover the degree of comprehension of certain physics concepts. Force Concept Inventory (Hestenes, Welss & Swackhamer, 1992), which objective is to assess the comprehension of velocity, acceleration and force from a Newtonian standpoint; Brief Electricity and Magnetism Assessment (Chabay & Sherwood, 2006), that evaluates the basic concepts of electricity and magnetism; Astronomy Diagnostic Test (Hufnagel, 2002), that analyzes the comprehension of astronomy concepts included in introductory astronomy courses for studies unrelated to science; Quantum Mechanics Conceptual Survey (McKagan, Perkins & Wieman 2010), that measures the comprehension of fundamental concepts of quantum mechanics; and the Bernoulli's Principle test (Barbosa, 2013), that seeks to measure the learning of Bernouilli's fluid dynamics principle in students of engineering, to mention a few.

Regarding waves, we can mention The Wave Concepts Inventory (Roedel *et al.*, 1998), that explores the visualization of waves, its definition and mathematical representation; Sound Concept Inventory Instrument (Eshach, 2014), that evaluates sound concepts in high school students and focuses on two aspects: the sound with material properties and sound with process properties; Ses Kavram Testi (Akarsu, 2015), which objective is to evaluate sound concepts studied in the last year of high school. All these instruments assess different physics concepts at different educational levels from a disciplinary standpoint.

It is important to consider that different studies in the area of Health sciences include physics contents in their curricula; for example: Medicine at the Pontificia Universidad Católica de Chile (s.f.); Medicine at Universidad de Sevilla (s.f.); and the Bachelor's degree in Bioimaging Production at Universidad de Buenos Aires (s.f.).

Based on the inclusion of physics science contents in their plans and programs of studies in the area of health sciences, and the need to assess them rigorously, we propose the creation of an instrument aimed at exploring the comprehension of physics concepts in students of health sciences. The plain question is: what concepts should be incorporated in an instrument of this type?

Among the topics on physics related to health sciences (optics, hydrodynamics, electromagnetism, etc.), we decided to consider sound, more specifically the concepts of *reflection* and *refraction* of sound waves since these are basic phenomena that, besides being studied at university level, they are addressed in pre-university teaching, and since they constitute concepts necessary to a better comprehension of the physiological processes and instrumentation of clinical diagnosis. Hence, for example, audition is a process in which the reflection of sound is essential. In many situations, capturing a sound stimulus implies detecting the location of the source and, in order for this to occur, the hearing system "uses" sound reflection in parts of the body such as the shoulders and the folds of the pinna to generate changes in the sound spectrum being perceived and thus, obtain the source location in a specific plane.

Another example that shows how essential these phenomena are is the operation of the echograph. The ultrasound must refract through the skin to the internal organs and reflect on the organ under study where the incidence angle can be relevant in obtaining a good image.

As a background of this work, we made a bibliography search of this type of instrument, in physics as well as in health sciences, focused on the assessment of certain concept proper to each area (Medina y Ramírez, 2019). It was impossible to find an instrument (even more so in Spanish) that would allow us to assess the degree of comprehension of the phenomena of sound reflection and refraction in students of health sciences. Moreover, given the extension of the field of health sciences in Spanish-speaking countries, it is necessary for the process of validating an instrument and endowing it with reliability aimed at the phenomena of reflection and refraction of sound waves be conducted with samples of professors and students of different universities of the region. This situation is made easier with computerized tomography, more specifically those that imply the use of Internet and distance communication.

THEORETICAL ASPECTS ON THE VALIDITY AND RELIABILITY OF A *TEST*

The development of a test must combine two essential characteristics: validity and reliability (or dependability). In general validity refers to the use of the results obtained through a *test*, and the reliability to errors made in the measurements carried out through the test.

Regarding validity, applying a test cast a set of information through which it is possible to reach conclusions regarding what is being measured. These conclusions must be guaranteed by a series of tests and data (Muñiz, 1997); hence, it is more appropriate to say that the inferences based on the scores or results of the test and not the actual test should be validated.

Throughout the evolution of the concept of validity, different types have been mentioned; however, its current notion points out to a unique validity

from which different types of evidence can be obtained through a process (Martínez *et al.*, 2006). "The recommendations of the international commissions suggest five sources of validity evidence: content, response process, internal structure, relations with other variables and consequences of assessment" (Pedrosa, Suárez-Álvarez and García-Cueto, 2014, p. 4). More specifically, we will give a brief description of content validity since it has been well accepted in educational tests (Martínez *et al.*, 2006) and we will use it in this work. The research of Ding *et al.* (2006), McKagan *et al.* (2010) and Barbosa (2013) reflect the use of this type of validity.

The evidence of the validity of content can be defined as "the degree in which the content of the test represents a satisfactory sample of the proficiency to be assessed" (Martínez *et al.*, 2006, p. 222). According to Sireci (1998), it is possible to establish two methods to assess the validity of the content: the experts' opinion and the use of statistical indicators calculated based on the application of the instruments.

In order to determine the evidence of the validity of content by experts' opinion, it is essential to select the people adequately; this selection must consider the particularities and the experience these people possess regarding the proficiency being assessed in a test. The usual procedure to obtain the validity is to define the proficiency that will be assessed, to give details of the characteristics of the *test*, to specify the number of questions that will assess every content of the proficiency; and to define the format of the items and the answers. After doing so, the test is submitted to experts in the field (not involved in the development of items), who must assess if the questions are representative and relevant to the assessment of the proficiency. We recommend that the experts give their opinion of the reagents separately in order to avoid any possible bias.

Regarding the use of statistical indicators to obtain evidence of content, the majority uses some technique of multivariant analysis or the theory of generality even when these have been procedures sparsely exploited (Martínez *et al.*, 2006; Pedrosa *et al.*, 2014). As for reliability, there are five indicators in the classical theory of tests that are widely used to analyze the reliability of a test: difficulty index, discrimination index, point biserial coefficient, Kuder-Richardson reliability index and Ferguson delta index. The three first ones refer to items and the two last ones, to the test in its entirety.

ITEM DIFFICULTY INDEX (P)

It is usually defined as the sample proportion that responds correctly to a question. This is

$$P = \frac{A}{N} \ ,$$

where *A* represents the number of subjects that responded correctly to the item and *N* is the number of subjects that responded to it. It is worth noting that what is defined is the facility of the *test;* hence, if *A* equals *N* (all the sample participants responded correctly to the question), the *P* value is 1 and the question is quite easy; and if *A* is zero (nobody responded correctly to the question), the *P* value is zero and the question is very difficult.

There is no one criterion when assessing the difficulty of the questions of a test, and adopting it depends of the approach the testing administrator gives to the test. Thus, for example, García-Cueto (2005) points out that if most of the items are moderately difficult, generally, the assessments will show the best results. On the one hand, Tristan (2001) recommends that it is convenient to have reagents with different degrees of difficulty available in order to measure with greater accuracy the proficiency of every person.

It is also common to calculate the average difficulty index of the test as a whole that corresponds to the quotient between the sum of the difficulty indexes and the quantity of items in the test.

DISCRIMINATION INDEX (D)

This is a measure of the power of discrimination of an item, i.e., the capacity of an item to distinguish between subjects with good achievement and those with bad achievement. To calculate this indicator, we use a 50% - 50% method that consists in separating the sample into two groups: one formatted by the scores higher to the average and the other constituted by those lower to the mean. According to this method, the expression to determine the discrimination index is

$$D = \frac{N_s - N_I}{\frac{N}{2}}$$

where N_s is the number of students with scores higher than the mean that responded correctly to the item; N_I refers to the number of students with scores lower than the mean who responded correctly to the item; *N*, corresponds to the total of students who answered the question. The discrimination index will take the values from -1 to 1.

A reagent with a positive discrimination index indicates a group proficiency of better achievement, i.e., there are more students of this group that answered the question correctly than those of the group of lesser achievement; while a reagent with a negative discrimination index indicates the opposite, i.e., that the number of students of the group of

lesser achievement that answered the question correctly is greater than the number of students of the group of better achievement.

This implies the need to discard or revise the reagents with a negative discrimination index since they contradict the purpose of the index. The closer to 1, that is, the discrimination index of a question, greater will be its discriminatory "capacity".

There is an advantage in using the 50% - 50% calculation in all the students; however, the drawback is underestimating the discriminatory capacity of an item since the groups being considered are not too far apart. A way to address this shortcoming is to use the higher and lower percentiles than 25% in order to reduce the probability of underestimating the level of discrimination of the questions by including the individuals more "coherent" in their performance, notwithstanding that the totality of students is not considered. In such case, the expression to determine the discrimination index would be:

$$D = \frac{N_s - N_I}{\frac{N}{4}}$$

where N_s is the number of students with scores corresponding to 25% higher that responded correctly to the item; N_I refers to the number of students with scores corresponding to the 25% lower who responded correctly to the item; and N, is the total of students that answered the question.

An item with good discrimination has a value greater than or equal to 0.3. In the same way as with the difficulty index, it is possible to calculate the test average discrimination index by adding the discrimination indexes and dividing this sum by the number of items that makes up the test. The value recommended for this mean must also be greater than or equal to 0.3.

POINT BISERIAL COEFFICIENT (r_{PBS})

The point biserial coefficient is a measure of coherence of an item with the test as a whole and reflects the correlation between the scores of the students on a specific item and their scores on the entire test. The possible range for this indicator is [-1,1]. The interpretation is that if the correlation between an item and the test is highly positive, then it is more likely that the students with high scores respond correctly to the item than those with lower scores. If the correlation is negative, then the students with lower scores will tend to respond correctly to the question and it is probable that the item is defective.

The expression that allows to determine the point biserial coefficient is as follows:

$$r_{pbs} = \frac{\bar{X}_1 - \bar{X}}{\sigma_X}\sqrt{\frac{P}{1-P}}$$

Where $\bar{X}_1$ is the total average grade of the subjects that responded correctly to the item; $\bar{X}$ is the average of the total grade of the exam of the entire sample; σ_x is the standard deviation of the grades of the entire sample; and P is the difficulty index of the item. An item with a good reliability must have a point biserial coefficient greater than or equal to 0.2. It is possible to calculate the average of the point biserial coefficient by adding all the coefficients and dividing said addition by the number of items of the test. The adequate value is also greater than or equal to 0.2.

KUDER-RICHARDSON RELIABILITY INDEX (KR_{20})

The internal coherence is an evidence of dependability of a test as a whole and refers to the equivalence of the reagents when measuring the proficiency to be assessed. If the equivalence is sufficiently elevated, the items will be related with strength and will measure the proficiency in question with a similar degree. There is more than one method to assess the internal coherence; for example, the method of two halves or the covariance between items. However, the Kuder-Richardson coefficient is particularly useful because it is usable in situations of sole application of a test. The expression that allows to calculate this coefficient is as follows:

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} p_j q_j}{\sigma_X^2}\right)$$

Where $p_j q_j$ *is the variable of a* dichotomous variable, p_j is the proportion of individuals that responded correctly to item j and q_j is the proportion of whom responded incorrectly. The values acceptable for this indicator, in case of assessing a group, are those higher than 0.7.

FERGUSON'S DELTA (δ)

It measures the discriminatory power of the test as a whole by inquiring how widely total scores of a sample are distributed in the possible range (Ding *et al.*, 2006). The expression that allows to calculate this indicator is:

$$\delta = \frac{N^2 - \sum_{i=1}^{K} f_i^2}{N^2 - \frac{N^2}{K+1}}$$

Where N is the number of subjects that responded to the test; k is the number of items that makes up the test; and f_i is the number of occurrences of each one of the grades. A test with a good discrimination must cast a Ferguson's delta higher than 0.9.

The works of Ding *et al.* (2006), McKagan *et al.* (2010), Barbosa (2013), Barniol, Capos and Zavala (2018), and Zavala *et al.* (2019) reflect the use of these indicators.

THE DESIGN OF THE TEST

The test was designed according to the following process (Medina and Ramírez, 2019):

1) Design and application of a survey which was conducted as follows (Medina and Ramírez, 2016):

- Study the bibliography to know the state of the art. Multiple studies were conducted regarding sound.
- Consult practicing physics professors to gather useful information in planning a survey.
- Develop an open response survey regarding the reflection and refraction of soundwaves. At the beginning, the survey consisted of twelve questions which were submitted to the experts' assessment.
- Apply the survey to students in order to identify erroneous conceptions.

2) Using the information regarding the conceptions presented by the students, we began the process of designing a test that would allow us to verify the comprehension of the phenomena of sound reflection and refraction. We chose this type of instrument because, in spite of the difficulty that the development of the reagents and the time involved imply, it is reliable from the statistical standpoint and it allows to measure different learning achievements over a wide range of levels and topic areas (Aiken, 2003; López e Hinojosa, 2016).

3) The step following the design of the test was to formulate a method that would help collect suggestions from a group of experts in order to make the necessary modifications to grant content validity to the instrument through the opinion of the experts (Hernández and Mendoza, 2018).

4) After designing the test, we proceeded to implement it online with the help of CTs so that both experts and students respectively could assess and respond to the test. The details are shown in the following section.

IMPLEMENTATION OF THE TEST ONLINE

The fundamental idea in collecting qualitative and quantitative information by means of a test was to obtain a wide panorama of a

phenomenon through a sample under real and various conditions. In this specific case, we wanted to apply the instrument in different geographic latitudes. This compelled us in thinking of creating a web system that would be accessible globally.

Our intention was to preserve the concept of the test to avoid behaviors such as the copy of the responses, included on the Internet, so we applied the following restrictions:

- The test will have a limited response time.
- If a user decides to leave the test, there will be a pause in time so the user will resume responding the next time he enters the system.
- The questions will be displayed randomly for every user.
- After responding to a question, it will be impossible for the user to make any correction.
- After completing the test, it will be impossible for the user to access the question again.
- This test was developed to be answered by students of different educational levels and areas of knowledge; however, ten questions were considered exclusively for students of the health sciences field.

In order to achieve the foregoing and to keep a record of the responses obtained throughout the application of the survey, it was necessary to create a data model to identify the students, know their source of information, nationality, area of studies and the institution to which they belong.

Regarding the statistical analysis of data, we designed a tool within the system itself to cast the responses of every student according to the following variables: name, question answered, area of knowledge, gender and institution. We also wanted that the system project a statistical visualization of the data obtained in a graph.

Moreover, since we also wanted that the data collected from the application of the test could be analyzed in other systems of information technology, the system added the functionality of exporting the data obtained in a .csv format.

As indicated above, our intention was that the test could be applied in different countries; hence, it was necessary to develop a software that would allow the users to access at any time or geographic location and, at

the same time, preserve the data integrity. Therefore, a web system was the most adequate option to meet these requirements.

Even when there are different frameworks that facilitate the development of a platform, such as Angular or React, it is important to highlight that the system was intended as an extension of an already existing web site that connect to other type of technologies among which the Java Web.

On the other hand, we should mention that PHP is one of the languages with greater support within the commercial web servers existing currently. Therefore, we decided to work with this language as the main server tool. To support the storage and management of the data collected, we proposed the MySQL motor, a database managing system that allows the creation of visualizations, transactions and procedures natively stored.

With the specifications above, the work team decided to use the JavaScript and PHP languages to develop the digital version of the test since these are languages that improve the quality of the interfaces and possess sufficient documentation; and to use the CSS and HTML technologies to build the site. The web site is located at the following address: http://physics-education.tlamatiliztli.net/index.php.

Figure 1 contains the log-in screen of the system.

Figure 1. Log-in screen to the system of application and assessment of the test.
Source: Website.

After the implementation of the system on the Web site, we proceeded to step 3 of the process of assessment described above.

CONTENT VALIDITY BY MEANS OF EXPERTS' OPINION

The group of experts was composed of eight academics of different universities (Chilean and Mexican) with postgraduate studies in

Phonoaudiology, Physics, Educational Physics and Educational Innovation. The group was set up according to a criterium based on the area of expertise: physics, teaching of physics and health sciences with the purpose of obtaining "feedback" from the three areas involved in the research and, thus, enrich its content.

The experts, by logging into the system, had access to the questions of the test first proposal (See Figure 2). In regard to every one of the questions, the request was to assess the level of their difficulty on a scale of 0 to 10, where 0 meant that the question was very easy, and 10, very difficult – sentence completion method (Hodge & Gillespie, 2003) – and, at the same time, the experts would submit an opinion on each one of the questions they considered relevant (Hernández and Mendoza, 2018).

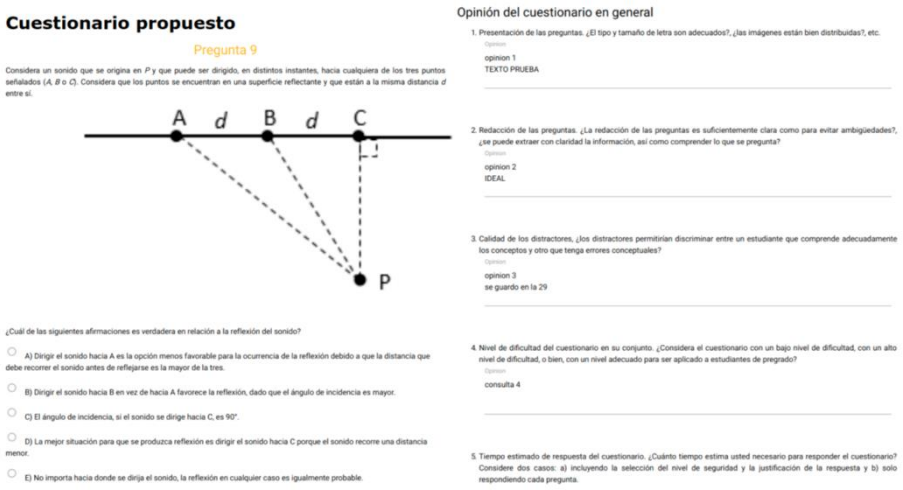


Figure 2. View of the test provided by the system to be assessed by the experts.
Source: Web site.

Besides the analysis of each question, we requested the experts to answer the following questions regarding the test as a whole:

- Questions presentation. Are the type and size of letters adequate? Are the images well distributed? etc.

- Drafting of questions. Were the questions sufficiently clear to avoid any ambiguity? Can the information be extracted clearly? and, Is the content of the question well understood?

- Quality of the distractors. Will the distractors allow discriminating between a student that understands the concepts adequately and another that has conceptual errors?

- Level of difficulty of the questionnaire as a whole. Do you consider that the questionnaire has a low level of difficulty, a high level of difficulty, or, an adequate level of difficulty to be applied to undergraduate students?

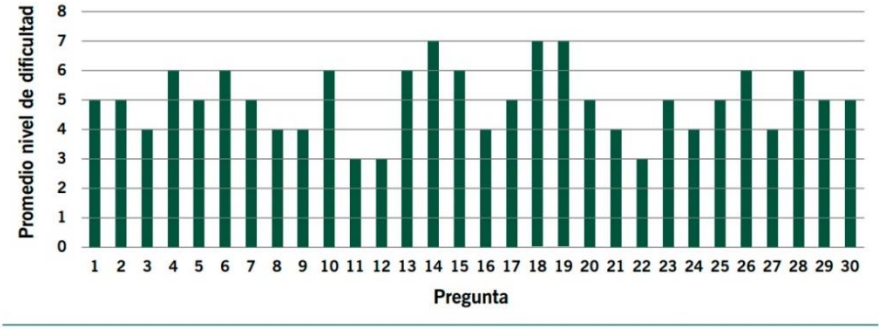
- Estimated response time of the questionnaire. How much time do you consider necessary to answer the questionnaire? Let us consider two cases: a) including the selection of security level and the justification of the response and, b) responding only to every question.

- Other comments you deem relevant.

From the analysis of the questionnaire conducted by the experts, we obtained the following results:

a) Level of difficulty

The average difficulty of each one of the questions assigned by the experts is shown in the following Graph.



Graph. The average difficulty level in comparison to the number of questions.

The level of difficulty assigned to the questions by the experts varies within a range of four points in which none of the questions are too easy or too difficult. According to the experts' opinion, 10% of the questions were easy and 10% were difficult, and the remaining questions were in average difficult. Five is the average difficulty of the test as a whole. From a qualitative standpoint, the results above are favorable considering that the test have an adequate level of difficulty.

b) Comments on each question and the test as a whole

The responses to these questions corresponded to the adequate presentation of the questions, both letter and images; that the writing was sufficiently clear and it allowed to understand what was requested in every question. It was only suggested to modify some reagents that might

generate ambiguity and, among other cases, improve the metrics of these to give them homogeneity; that the distractors helped discriminating between the students that did understand the concepts and those that did not. It was suggested to pay attention to the items that contained too heterogeneous distractors; that the level of difficulty of the test was adequate for undergraduate students and that the questions implied different degrees of difficulty and involved different cognitive dimensions. The average response time indicated by the experts, including the selection of the security level and the justification was of 100 minutes, while the average time to answer the questions without selecting the security level nor the justification was of 64 minutes.

The comments were used to improve the first version of the test and to obtain a second version validated by the experts which was put online so the students responded on the same web site. Once registered, the students logged in as users and answered the test. If the student belong to the area of medical biological sciences, thirty questions were displayed with a maximum response time of 65 minutes and if they belong to the physics mathematics area, the questions were 20 with a maximum response time of 44 minutes. In both cases, there was a countdown timer and the questions were displayed randomly. Every question had to be answered before going to the next one.

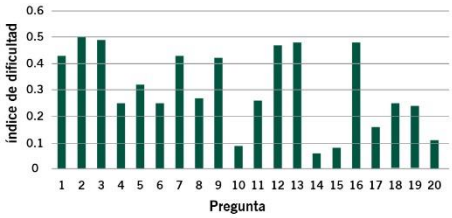
RELIABILITY

The second version of the test was implemented online to a total of 288 students of the bachelor's degree of universities in Chile, Colombia and Mexico. From the 288 students, 233 were from the area of physics mathematics, and 55 were from the medical biological area; 121 were female and 167, male. Having only 55 students in the medical biological area produced inconclusive results; this is the reason why we decided to conduct the analysis for the 20 general questions with the 288 subjects of the sample. From the 20 general questions, the first ten assessed aspects of sound reflection.

The indicators presented the following values:

a) Difficulty index (*P)*

Pregunta	1	2	3	4	5	6	7
P	0.43	0.50	0.49	0.25	0.32	0.25	0.43
Pregunta	8	9	10	11	12	13	14
P	0.27	0.42	0.09	0.26	0.47	0.48	0.06
Pregunta	15	16	17	18	19	20	
P	0.08	0.48	0.16	0.25	0.24	0.11	



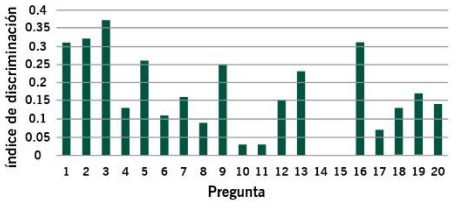
Source: Self development.

When considering the first ten reflection questions, the mean was 0.34 and, the average of the refraction questions was 0.26

b) Discrimination index (*D)*

- Calculation with the 50% - 50% method

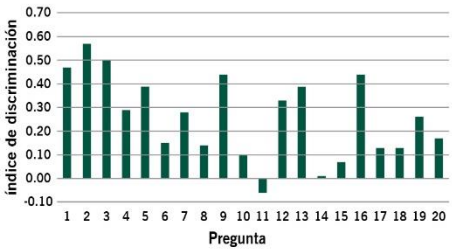
Pregunta	1	2	3	4	5	6	7
D	0.31	0.32	0.37	0.13	0.26	0.11	0.16
Pregunta	8	9	10	11	12	13	14
D	0.09	0.25	0.03	0.03	0.15	0.23	0.00
Pregunta	15	16	17	18	19	20	
D	0.00	0.31	0.07	0.13	0.17	0.14	



Source: Self development.

- Calculation with the 25% - 25% method

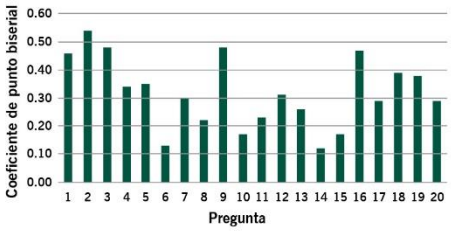
Pregunta	1	2	3	4	5	6	7
D	0.47	0.57	0.50	0.29	0.39	0.15	0.28
Pregunta	8	9	10	11	12	13	14
D	0.14	0.44	0.10	-0.06	0.33	0.38	0.01
Pregunta	15	16	17	18	19	20	
D	0.07	0.44	0.13	0.13	0.26	0.17	



Source: Self development.

c) Point biserial coefficient *(r_{pbs})*

Pregunta	1	2	3	4	5	6	7
r_{pbs}	0.46	0.54	0.48	0.34	0.35	0.13	0.30
Pregunta	8	9	10	11	12	13	14
r_{pbs}	0.22	0.48	0.17	0.23	0.31	0.26	0.12
Pregunta	15	16	17	18	19	20	
r_{pbs}	0.17	0.47	0.29	0.39	0.38	0.29	



Source: Self development.

d) Kuder-Richardson reliability index (KR_{20})

Since the purpose of the test was to assess aspects related to the reflection and refraction of sound waves and the Kuder-Richardson reliability index is the indicator of the homogeneity of the test, it is convenient to calculate an index for the set of questions on reflection (KR_{20B}), which are the ten first questions, and to calculate an index for the set of questions on refraction (KR_{20A}), which is the second set of ten questions. The values for each one of those were KR_{20A} = 0.24 and KR_{20B} = -0.14.

e) Ferguson's Delta (δ)

Likewise, for the Kuder-Richardson reliability index, we calculated a delta for the set of questions on reflection (δ_A) and a delta for the set of questions on refraction (δ_B). The values were δ_A= 0.87 and δ_B = 0.81.

ANALYSIS OF RESULTS

As we have mentioned, the validity was obtained by means of the evidence of content validity and, said evidence, in turn, was obtained through the assessment of experts. This validity was reflected in the second version of the test which was administered to the 288 students that participated in the study in order to confer reliability to the instrument. The results of the respective indicators will be analyzed next.

Regarding the values of the difficulty index of every question, we found that questions 4, 6, 8, 10, 11, 14, 15, 17, 18, 19 and 20 were very difficult, and the 10, 14 and 15 were extremely difficult. Questions 1, 2, 3, 5, 7, 9, 12, 13 and 16 were moderately difficult. None of these questions were easy.

The optimum level of difficulty of an item is 0.5; however, since it is almost impossible to achieve that all the questions have such degree of difficulty, several criteria have been proposed to select the questions for a test adequately. One of those criteria was to use a range of difficulty from 0.4 to 0.6 as an interval that acknowledges 0.5 as the optimal level. A second criterion is to consider a range between 0.3 to 0.9 and eliminate the very easy elements (over 0.9) and the very difficult (under 0.3); this is because the very easy questions and the very difficult questions do not contribute to the discriminatory capacity of a test (Doran, 1980).

A third criterion is to combine questions from different degrees of difficulty based on the following criterion:

Table 1. Dificulty levels

Concept	Range of values	Percentage of items
Very easy	0.85 – 1.00	15
Moderately easy	0.60 – 0.85	35
Moderately difficult	0.35 – 0.60	35
Very difficult	0.00 – 0.35	15

Source: R. Doran (1980, p. 97). Washington: National Science Teachers Association.

By considering the second criterion already pointed out, we could select questions 1, 2, 3, 5, 7, 9, 12, 13 and 16, for the test.

Regarding the discrimination index, Ebel and Frisbie (1991) display the following classification for the discrimination of the items:

Table 2. Discrimination index of the items.

Discrimination Index	Assessment of the item
0.40 and more	Very good item
0.30 to 0.39	Reasonably good but possibly subject to improvements
0.20 to 0.29	Marginal item that generally requires improvements
under 0.20	Poor item, must be discarded or improved through revision

Source: R. Ebel y D Frisbie, 1991, p. 232, Englewood Cliffs, NJ: Prentice-Hall.

From the above, we infer that an item with an index greater than or equal to 0.30 represents a good discrimination; hence, according to the values obtained, questions 1, 2, 3 and 16 have great discrimination, following the

50% - 50% method; however, if the 25% - 25% method is followed, the questions having good discrimination between the students with good achievement and the students with bad achievement are questions 1, 2, 3, 5, 9, 12, 13 and 16. Questions 4 to 7 have values close to 0.30. We noticed that the questions discriminate better in more extreme groups of achievement.

In regard to the point biserial coefficient, considering that an adequate value for this indicator is the one greater or equal to 0.2, questions 1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 16, 17, 18, 19 and 20 present a good coherence. Items with values lesser than 0.2 must not necessarily be discarded and they "can even remain in a test, but they must be few of them" (Ding *et al.*, 2006, p. 3); hence, questions 10 and 15 could be considered in the test since their values are relatively close to 0.2.

Regarding the reliability index, the criteria are not unique and they can vary according to the assessors and the purpose of a test. The following table summarizes some criteria widely accepted (Doran, 1980):

Table 3. Reliability index

Index Values	Reliability criterion
0.95 – 0.99	Very high, very seldomly found
0.90 – 0.95	High, sufficient for the assessment of individuals
0.80 – 0.90	High, could be considered for individual assessment
0.70 – 0.80	Good, sufficient for a group measurement, not for individuals
Under 0.70	Low, useful only for averages or surveys

Source: Doran, 1980, p. 104, Washington: National Science Teachers Association

The values of the index for the set of questions on reflection were 0.24 and for the set of questions on refraction were of -0.14. In both cases, the values were too low; however, the reliability indexes were influenced by a series of factors such as the test extension, difficulty and discrimination of the questions, as well as the range of the abilities of the subjects of the sample. By eliminating questions with difficulty indexes "far" from 0.5 (considered the optimum level) and with "too low" discrimination index, the value of the reliability index raises.

On the other hand, Adams and Wieman (2011) claim that the instruments designed to measure multiple concepts can have a low Cronback alpha (which for this test is equivalent to the Kuder-Richardson reliability index since the instrument is dichotomous), because these concepts can be independent. This is the case of the test that, despite it assesses only two phenomena (reflection and refraction), these involve more than one concept.

The values of Ferguson's delta were 0.87 for the set of questions on reflection and 0.81 for the set of questions on refraction. The values are close to the desired number which is 0.9.

Lastly, even when the complete analysis was done for the 20 general questions, we present the values for the three first indicators (those referring to every item and not to the test as a whole) obtained for the specific questions on health sciences.

a) Difficulty index (*P*)

Pregunta	21	22	23	24	25
P	0.27	0.29	0.42	0.55	0.27
Pregunta	26	27	28	29	30
P	0.27	0.04	0.35	0.20	0.27

Source: Self development.

Questions 23, 24 and 28 were moderately difficult and the remainder were difficult, except for question 27 that was particularly complicated.

b) Discrimination index (*D*)

Calculated with the 25% - 25% method.

Pregunta	21	22	23	24	25
D	0.29	0.29	0.36	0.51	0.36
Pregunta	26	27	28	29	30
D	0.36	0.07	0.29	0.15	0.07

Source: Self development.

Questions 23, 24, 25 and 26 presented good discrimination; 21, 22 and 28, a 0.29 discrimination, so they should not be discarded *a priori*. Questions 27 and 30 had low discrimination.

d) Point biserial coefficient (r_{pbs})

Pregunta	21	22	23	24	25
r_{pbs}	0.38	0.29	0.49	0.40	0.41
Pregunta	26	27	28	29	30
r_{pbs}	0.21	0.22	0.26	0.25	0.27

Source: Self development.

All the questions show a coefficient higher than the minimum accepted, i.e., there is greater possibility that the students with better scores answer the questions well.

CONCLUSION

The object pursued in this research was partially met since we could develop a conceptual research test on sound phenomena focusing on health sciences but not in the terms originally proposed. It was possible to obtain the validity of the instrument through the content evidence, results that were already published (Medina y Ramírez, 2019); however, the reliability was reached partially only given the reduced sample.

The "reduced" sample we used allowed us to conduct a more coherent analysis of the 20 first questions and a weaker analysis of the last ten questions all focusing on health sciences. Questions 1, 2, 3, 5, 9, 12, 13 and 16 met the reliability standards without any problem. However, some remaining questions can be considered; question 4, even though it was difficult, has an adequate point biserial coefficient and a 0.29 discrimination; question 7 was moderately difficult with an adequate point biserial coefficient and a 0.28 discrimination; question 19 presented a high difficulty, with an adequate point biserial coefficient and a discrimination of 0.26. The remainder must be revised in depth.

Regarding the Kuder-Richardon index that measures the test coherence, the values obtained were low which could be explained by the fact that, while it is true, the test only measures two phenomena, but there are different elements involved.

The construction of a test that follows an online work process has advantages, but at the same time, it has some additional requirements. Among the advantages, we should mention the "timelessness" since, in theory, the students can respond to the questionnaire at any time. Furthermore, in this specific case, the system saves the responses automatically and the timer is activated only when the student answers. This allows the individuals responding to the questionnaire to exit the system if necessary.

It is valid to note that the online application prevents spending an enormous amount of money on paper and ink besides avoiding the possible loss or deterioration of printed material. The input of data in a program of statistical analysis is also very simple since the responses can be recovered, e.g., on an Excel screen. The additional requirement lies in how to generate a cooperative environment in which those involved feel motivated in participating, this could be facilitated in an in-class situation in comparison with a virtual situation in which the contact with participants can be scarce.

As future research, it is necessary to review in depth the questions that did not reached the appropriate values to the parameters in order to reformulate them adequately and reduce the phenomena to only one, whether reflection or refraction, which would facilitate a greater homogeneity and the development of new tests that would assess every concept separately.

Likewise, a future research would be to develop a test that actually addresses health sciences where reliability is reached using the item response theory, not because the classic theory of the tests does not suffice but rather given the richness of the additional information obtained through an analysis conducted under this theory. To do so, we need a more considerable sample than the one obtained in this research, generally above 500 (Martínez *et al.*, 2006).

REFERENCES

Adams, W. & Wieman, C. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, *33*(9), 1289-1312. https://doi.org/10.1080/09500693.2010.512369

Aiken, L. (2003). *Tests psicológicos y evaluación*. México: Pearson Education.

Akarsu, B. (2015). Ses Kavram Testi. *Journal of European Education*, *5*(1), 23-30. Recuperado de: https://toad.halileksi.net/sites/default/files/pdf/ses-kavram-testi-toad.pdf

Barbosa, L. (2013). Construcción, validación y calibración de un instrumento de medida del aprendizaje: test de ley de Bernoulli. *Revista Educación en Ingeniería*, *8*(15), 24-37. Recuperado de: https://www.educacioneningenieria.org/index.php/edi/article/viewFile/301/161

Barniol, P.; Campos, E. y Zavala, G. (2018). La prueba conceptual de electricidad y magnetismo: análisis de confiabilidad y estudio de las dificultades más frecuentes. *Enseñanza de las Ciencias*, *36*(2), 167-192. https://doi.org/10.5565/rev/ensciencias.2456

Chabay, R. & Sherwood, B. (2006). Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research*, *2*(1), 7-13. Recuperado de: https://www.physport.org/assessments/assessment.cfm?A=BEMA

Ding, L.; Chabay, R.; Sherwood, B. & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research*, *2*(1). https://doi.org/10.1103/PhysRevSTPER.2.010105

Doran, R. (1980). Basic *Measurement and Evaluation of Science Instruction*. Washington, DC: National Science Teachers Association

Ebel, R. & Frisbie, D. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Eshach, H. (2014). Development of a student-centered instrument to assess middle school student's conceptual understanding of sound. *Physical Review Special Topics - Physics Education Research*, *10*(1). https://doi.org/10.1103/PhysRevSTPER.10.010102

García-Cueto, E. (2005). Análisis de los ítems. Enfoque clásico, en J. Muñiz, A. M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (eds.), *Análisis de los ítems*. Cuadernos de Estadística número 30. Madrid: Editorial La Muralla.

Hernández, R. y Mendoza, C. (2018). *Metodología de la Investigación: las rutas cuantitativa, cualitativa y mixta*. Ciudad de México: McGraw-Hill.

Hestenes, D.; Wells, M. & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141-158. https://doi.org/10.1119/1.2343497

Hodge, D. & Gillespie, D. (2003). Phrase completions: An alternative to Likert scales. *Social Work Research*, *27*(1), 45-55. https://doi.org/10.1093/swr/27.1.45

Hufnagel, B. (2002). Development of astronomy diagnostic test. *Astronomy Education Review*, *1*(1), 47–51. https://doi.org/10.3847/AER2001004

López, B. e Hinojosa, E. (2016). *Evaluación para el aprendizaje: alternativas y nuevos desarrollos*. México: Trillas.

Martínez, M., Hernández, M. y Hernández, M. (2006). *Psicometría*. Madrid: Alianza Editorial.

McKagan, S., Perkins, K. & Wieman, C. (2010). Design and validation of the quantum mechanics conceptual survey. *Physical Review Special Topics - Physics Education Research*, 6(2), 020121. https://doi.org/10.1103/PhysRevSTPER.6.020121

Medina, J. y Ramírez, M. (2016). Obtención y clasificación de ideas previas sobre fenómenos sonoros: estudio en alumnos universitarios de carreras de ciencias de la salud. *Latin-American Journal of Physics Education*, *10*(3). Recuperado de: http://www.lajpe.org/sep16/3305_Medina_2016.pdf

Medina, J. y Ramírez, M. (2019). Construcción de un test sobre fenómenos sonoros orientado a estudiantes de ciencias de la salud. *Innovación Educativa*, *10*(79), 79-98. Recuperado de: https://www.ipn.mx/assets/files/innovacion/docs/Innovacion-Educativa-79/Construccion-de-un-test-sobre-fenomenos-sonoros-orientado.pdf

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Pedrosa, I.; Suárez-Álvarez, J. y García-Cueto, E. (2014). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción Psicológica*, *10*(2), 3-18. https://doi.org/10.5944/ap.10.2.11820

Pontificia Universidad Católica de Chile. (s.f.). Física para ciencias biomédicas. Recuperado de http://catalogo.uc.cl/index.php?tmpl=component&option=com_catalogo&view=programa&sigla=FIS119M

Roedel, R.; El-Ghazaly, S.; Rhoads, T. y El-Sharawy, E. (1998). Wave concepts inventory –an assessment tool for courses in electromagnetic engineering, en *Proceedings-Frontiers in Education Conference*, *2*, 647-653. https://doi.org/10.1109/FIE.1998.738761

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*(1/3), 83-117. https://doi.org/10.1023/A:1006985528729

Tristán, A. (2001). *Análisis de Rasch para todos*. México: Ceneval.

Universidad de Buenos Aires. (s.f.). Licenciatura en Producción de Biomimágenes. Recuperado de: http://www.uba.ar/download/academicos/carreras/bioimagenes.pdf

Universidad de Sevilla. (s.f.). Física Médica. Recuperado de: http://www.us.es/estudios/grados/plan_172/asignatura_1720006

Zavala, G.; Barniol, P. y Tejeda, S. (2019). Evaluación del entendimiento de gráficas de cinemática utilizando un test de opción múltiple en español. *Revista Mexicana de Física*, *65*(2). https://doi.org/10.31349/RevMexFisE.65.162