apertura
Volumen 11, número 2, octubre 2019, pp. 104-121 | e-ISSN 2007-1094 | Universidad de Guadalajara
http://dx.doi.org/10.32870/Ap.v11n2.1622

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

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

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

Recepción del artículo: 26/3/2019 | Aceptación para publicación: 18/7/2019 | Publicación: 30/9/2019

## 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 generó 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.

## *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

***Keywords***
*Validation of instruments, physics education, distance education*

* Maestro en Ciencias en Física Educativa por el Instituto Politécnico Nacional de México. Profesor del Centro de Docencia Superior en Ciencias Básicas de la Universidad Austral de Chile, Sede Puerto Montt, Chile | **Doctor en Ciencias en Física Educativa por el Instituto Politécnico Nacional de México. Profesor investigador en el Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Unidad Legaria, del Instituto Politécnico Nacional (CICATA-IPN), México. ORCID: https://orcid.org/0000-0002-3459-2927 | *** Doctor en Ciencias con especialidad en Matemática Educativa por el Instituto Politécnico Nacional de México. Profesor investigador en el Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Unidad Legaria, del Instituto Politécnico Nacional (CICATA-IPN), México. ORCID: https://orcid.org/0000-0003-2076-7383


## INTRODUCCIÓN

De acuerdo con los Standards for Educational and Psychological Tests, "un *test* es un instrumento evaluativo o procedimiento en el que se obtiene una muestra de la conducta de los examinados en un dominio especificado y posteriormente es evaluada y puntuada usando un procedimiento estandarizado" (Martínez, Hernández y Hernández, 2006, p. 18). Aun cuando en esta definición pueden incluirse inventarios, escalas, cuestionarios y otros, en este trabajo denominaremos *test* a una prueba de preguntas de opción múltiple de respuesta única.

En física, desde la década de los noventa del siglo pasado hasta la fecha, se han elaborado diversos *test* con el propósito de conocer el grado de comprensión de ciertos conceptos físicos. Podemos mencionar, entre otros: Force Concept Inventory (Hestenes, Welss & Swackhamer, 1992), cuyo objetivo es evaluar la compresión de los conceptos *velocidad*, *aceleración* y *fuerza* desde el punto de vista newtoniano; Brief Electricity and Magnetism Assessment (Chabay & Sherwood, 2006), que evalúa conceptos básicos de electricidad y magnetismo; Astronomy Diagnostic Test (Hufnagel, 2002), que analiza la comprensión de conceptos de astronomía incluidos en cursos introductorios de astronomía para carreras no relacionadas con la ciencia; Quantum Mechanics Conceptual Survey (McKagan, Perkins & Wieman 2010), que mide la comprensión de conceptos fundamentales de la mecánica cuántica; y el test de ley de Bernoulli (Barbosa, 2013), que busca medir el aprendizaje de la ley de presión hidrodinámica de Bernoulli en estudiantes de ingeniería.

En lo relativo a ondas, podemos mencionar The Wave Concepts Inventory (Roedel *et al.*, 1998), que explora la visualización de las ondas, su definición y su representación matemática; Sound Concept Inventory Instrument (Eshach, 2014), el cual evalúa conceptos de sonido en estudiantes de secundaria y se enfoca en dos aspectos: el sonido tiene propiedades materiales y el sonido tiene propiedades de proceso; Ses Kavram Testi (Akarsu, 2015), cuyo objetivo es evaluar conceptos de sonido estudiados hasta el último año de secundaria. Todos estos instrumentos valoran diversos conceptos físicos en distintos niveles educativos siempre desde un punto de vista disciplinar.

Es importante considerar que diversas carreras del área de ciencias de la salud incluyen en sus planes de estudio contenidos de física; por ejemplo: Medicina, en la Pontificia Universidad Católica de Chile (s.f.); Medicina, en la Universidad de Sevilla (s.f.); Licenciatura en Producción de Bioimágenes, en la Universidad de Buenos Aires (s.f.).

Con base en la inclusión de contenidos de ciencias físicas en los planes y programas de estudio de carreras del área de ciencias de la salud, y la necesidad de evaluarlos con rigor, proponemos la creación de un instrumento dirigido a indagar la comprensión de conceptos de física en estudiantes de ciencias de la salud. La pregunta natural es ¿qué conceptos incorporar en un instrumento de este tipo?

Como antecedente de este trabajo, hicimos una búsqueda bibliográfica de este tipo de instrumentos, tanto en física como en ciencias de la salud, enfocados en la evaluación de ciertos conceptos propios de cada área (Medina y Ramírez, 2019)

Entre los temas de física que guardan relación con las ciencias de la salud (óptica, hidrodinámica, electromagnetismo, etcétera), decidimos considerar el del sonido, en particular los conceptos *reflexión* y *refracción* de ondas sonoras. Esto, porque son fenómenos básicos que, además de ser estudiados en la formación universitaria, se abordan en la enseñanza preuniversitaria, y porque constituyen conceptos necesarios para una mejor comprensión de procesos fisiológicos e instrumentación de diagnóstico clínico. Así, por ejemplo, la audición es un proceso en el cual la reflexión del sonido es esencial; en muchas situaciones, captar un estímulo sonoro implica detectar la ubicación de la fuente y, para que esto ocurra, el sistema auditivo "hace uso" de la reflexión del sonido en partes del cuerpo, como los hombros y pliegues del pabellón auricular, para generar cambios en el espectro sonoro que es percibido y obtener, de ese modo, la ubicación de la fuente en un determinado plano.

Otro ejemplo que muestra lo esencial que son estos fenómenos es el funcionamiento de un ecógrafo; el ultrasonido debe refractarse a través de la piel hacia los órganos internos y reflejarse en el órgano que se desea estudiar, donde el ángulo de incidencia puede ser relevante en la obtención de una buena imagen.

Como antecedente de este trabajo, hicimos una búsqueda bibliográfica de este tipo de instrumentos, tanto en física como en ciencias de la salud, enfocados en la evaluación de ciertos conceptos propios de cada área (Medina y Ramírez, 2019). No fue posible encontrar un instrumento (más aún en español) que permitiera evaluar el grado de comprensión de los fenómenos de reflexión y refracción del sonido en estudiantes de ciencias de la salud. Además, debido a la extención del campo de las ciencias de la salud en los

países de habla hispana, se hace necesario que el proceso de validar y dotar de confiabilidad a un instrumento dirigido a los fenómenos de reflexión y refracción de ondas sonoras se efectúe con muestras de profesores y estudiantes de diferentes universidades de la región. Esta situación se facilita con las TAC, en particular aquellas que implican el uso de internet y la comunicación a distancia.

## ASPECTOS TEÓRICOS SOBRE VALIDEZ Y CONFIABLIDAD DE UN *TEST*

La elaboración de un *test* debe reunir dos características esenciales: validez y confiabilidad (o fiabilidad). En términos generales, la validez hace referencia al uso de los resultados obtenidos a través del *test*, y la confiabilidad, a los errores cometidos en las mediciones realizadas por medio de este.

La elaboración de un *test* debe reunir dos características esenciales: validez y confiabilidad (o fiabilidad). En términos generales, la validez hace referencia al uso de los resultados obtenidos a través del *test*, y la confiabilidad, a los errores cometidos en las mediciones realizadas por medio de este

En relación con la validez, la aplicación de un *test* arroja un conjunto de información mediante la cual es posible lograr conclusiones respecto de lo que se está midiendo. Estas conclusiones deben estar garantizadas por una serie de pruebas y datos (Muñiz, 1997); por esta razón, es más apropiado hablar acerca de que las inferencias sobre la base de las puntuaciones o resultados de un *test* son las que deben ser validadas y no el *test* en sí mismo.

Aun cuando en la evolución del concepto de *validez* se ha hablado de distintos tipos, la noción actual apunta a hablar de una única validez de la cual se pueden obtener distintos tipos de evidencias a través de un proceso (Martínez *et al.*, 2006). "Las recomendaciones de las comisiones internacionales sugieren cinco fuentes de evidencia de validez: contenido, proceso de respuesta, estructura interna, relaciones con otras variables y consecuencias de la evaluación" (Pedrosa, Suárez-Álvarez y García-Cueto, 2014, p. 4). En particular, daremos una breve descripción de la validez de contenido, pues ha sido bien aceptada en *test* educativos (Martínez *et al.*, 2006) y es la que utilizaremos en este trabajo. Las investigaciones de Ding *et al.* (2006), McKagan *et al.* (2010) y Barbosa (2013) reflejan el uso de este tipo de validez.

La evidencia de validez de contenido se puede definir como "el grado en que el contenido del *test* representa una muestra satisfactoria del dominio que pretende evaluar" (Martínez *et al.*, 2006, p. 222). De acuerdo con Sireci (1998), es posible establecer dos métodos para estimar la validez de contenido: juicio de expertos y uso de indicadores estadísticos calculados con base en la aplicación de los instrumentos.

Para determinar evidencia de validez de contenido por juicio de expertos, es esencial una adecuada selección de las personas, que considere las particularidades y la experiencia que posean en relación con los dominios que se valoran en un *test*. El procedimiento usual para obtenerla es definir el dominio que se evaluará, detallar las características del *test*, precisar el número de preguntas que valorará cada contenido del dominio y definir el formato de los ítems y

las respuestas. Una vez hecho lo anterior, se debe colocar el *test* en manos de expertos en el tema (no involucrados en la confección de los ítems), quienes deben estimar si las preguntas son representativas y relevantes para la evaluación del dominio. Es recomendable que los expertos juzguen los reactivos por separado a fin de evitar posibles sesgos.

En relación con el uso de indicadores estadísticos para obtener evidencia de contenido, la mayoría emplea alguna técnica de análisis multivariante o la teoría de la generabilidad, aun cuando han sido procedimientos poco aprovechados (Martínez *et al.*, 2006; Pedrosa *et al.*, 2014). Acerca de la confiablidad, en la teoría clásica de los *test* existen cinco indicadores ampliamente utilizados para analizar la confiabilidad de un *test*: índice de dificultad, índice de discriminación, coeficiente de punto biserial, índice de confiablidad de Kuder-Richardson y la delta de Ferguson. Los tres primeros están referidos a los ítems y los dos últimos, al *test* en su conjunto.

### *Índice de dificultad del ítem (P)*

Usualmente, se define como la proporción de la muestra que responde de manera correcta una pregunta. Esto es

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

donde *A* es el número de sujetos que responde de forma correcta el ítem y *N*, el número de sujetos que lo contesta. Hay que notar que lo que se define estrictamente es la facilidad del *test*, pues si *A* es igual a *N* (todos los integrantes de la muestra responden de modo correcto la pregunta), el valor de *P* es 1 y la pregunta resulta ser muy fácil, y si *A* es cero (nadie responde de manera correcta la pregunta), el valor de *P* es cero y la pregunta resulta ser muy difícil.

No existe un único criterio a la hora de evaluar la dificultad de las preguntas de un *test*, y su adopción dependerá del enfoque que quiera

darle al *test* quien lo administre. Así, por ejemplo, García-Cueto (2005) señala que si la mayoría de los ítems tienen una dificultad media, por lo general se obtendrán mejores resultados en las evaluaciones. Por su parte, Tristán (2001) recomienda que, con la finalidad de medir el dominio de cada persona con mayor precisión, es conveniente disponer de reactivos con diferentes grados de dificultad.

Es común, también, calcular el índice de dificultad promedio del *test* en su conjunto, que corresponde al cociente entre la suma de los índices de dificultad y la cantidad de ítems del *test*.

### *Índice de discriminación (D)*

Es una medida del poder de discriminación de un ítem, esto es, la capacidad de un ítem para distinguir entre sujetos con buen rendimiento y sujetos con mal rendimiento. Para calcular este indicador, podemos proceder con el método 50% - 50%, que consiste en separar la muestra en dos grupos: uno formado por los puntajes superiores a la mediana y otro constituido por los puntajes inferiores a la mediana. La expresión para determinar el índice de discriminación según este método está dada por

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

donde $N_S$ es la cantidad de estudiantes con puntajes superiores a la mediana que respondieron correctamente el ítem; $N_I$ refiere la cantidad de estudiantes con puntajes inferiores a la mediana que respondieron correctamente el ítem; y $N$, al total de estudiantes que contestaron la pregunta. El índice de discriminación tomará valores que van desde -1 hasta 1.

Un reactivo con índice de discriminación positivo indica un dominio del grupo de mejor rendimiento, es decir, que son más los estudiantes de este grupo que contestan bien a la pregunta que los del grupo de menor rendimiento, mientras que un reactivo con un índice de discriminación negativo revela lo contrario, esto es, que la cantidad de estudiantes del grupo de menor desempeño que contesta correctamente la pregunta es mayor que el número de estudiantes del grupo de mejor rendimiento que lo hace.

Lo anterior implica la necesidad de descartar o someter a revisión reactivos con índice de discriminación negativo, pues contradicen el propósito del índice. Entre más cercano a 1 sea el índice de discriminación de una pregunta, mayor será la "capacidad" discriminatoria de esta.

Utilizar el cálculo 50% - 50% tiene la ventaja de considerar a todos los estudiantes, pero puede traer consigo el inconveniente de subestimar la capacidad discriminatoria de un ítem, ya que los grupos considerados son poco extremos. Una forma de subsanar lo anterior es utilizando los percentiles superior e inferior del 25%, con el propósito de reducir la probabilidad de subestimar el nivel de discriminación de las preguntas al incluir a las personas más "consistentes" en su desempeño, no obstante que no considera la totalidad de los estudiantes. En tal caso, la expresión para determinar el índice de discriminación sería:

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

donde $N_S$ es la cantidad de estudiantes con puntajes correspondientes al 25% superior que respondieron correctamente el ítem; $N_I$ es la cantidad de estudiantes con puntajes correspondientes al 25% inferior que respondieron correctamente el ítem; y $N$ es el total de estudiantes que contestaron la pregunta.

Un ítem con buena discriminación es aquel con un valor mayor o igual a 0.3. De igual forma que con el índice de dificultad, es posible calcular el índice de discriminación promedio del *test* al

sumar los índices de discriminación y dividir esta suma por la cantidad de ítem que conforma el *test*. El valor recomendado para este promedio también debe ser mayor o igual a 0.3.

### *Coeficiente de punto biserial ($r_{pbs}$)*

El coeficiente del punto biserial es una medida de la consistencia de un ítem con el *test* en su conjunto y refleja la correlación entre los puntajes de los estudiantes en un ítem en particular y sus puntajes en el *test* completo. El rango posible para este indicador es [-1,1]. La interpretación es que, si la correlación entre un ítem y el *test* es altamente positiva, entonces es más probable que los estudiantes con altos puntajes respondan de manera correcta el ítem que aquellos con puntajes inferiores. Si la correlación es negativa, entonces los alumnos con menores puntajes tenderán a responder en forma correcta la pregunta y es probable que el ítem sea defectuoso.

La expresión que permite determinar el coeficiente de punto biserial está dada por

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

donde $X_1$ es la calificación total promedio de los sujetos que contestan correctamente el ítem; $X$ es el promedio de la calificación total en el examen de toda la muestra; $\sigma_X$ es la desviación estándar de las calificaciones de toda la muestra; y $P$ el índice de dificultad del ítem. Un ítem con una buena confiabilidad debe tener un coeficiente de punto biserial mayor o igual a 0.2. Es posible calcular el promedio de los coeficientes de punto biserial sumando todos los coeficientes y dividiendo esta suma por la cantidad de ítem del *test*. El valor apropiado es también mayor o igual a 0.2.

### *Índice de confiabilidad de Kuder-Richardson ($KR_{20}$)*

La consistencia interna es una evidencia de la fiabilidad de un *test* en su conjunto y hace referencia a la equivalencia de los reactivos al medir el dominio que se pretende evaluar; si la equivalencia es lo suficientemente elevada, los ítems estarán relacionados con fuerza y medirán el dominio en cuestión con un grado similar. Existe más de un método para evaluar esta consistencia interna, por ejemplo, el método de dos mitades o la covariancia entre los ítems; sin embargo, el coeficiente de Kuder-Richardson resulta ser en especial útil porque es utilizable en situaciones de aplicación única de un *test*. La expresión que permite calcularlo está dada por

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

donde $p_j q_j$ es la varianza de una variable dicotómica: $p_j$ es la proporción de personas que contestan en forma correcta el ítem $j$ y $q_j$ es la proporción de quienes lo responden incorrectamente. Los valores aceptables para este indicador, si es que se va a evaluar a un grupo, son aquellos superiores a 0.7.

### *Delta de Ferguson ($\delta$)*

Mide el poder discriminatorio del *test* en su conjunto, al indagar qué tan ampliamente se distribuyen las puntuaciones totales de una muestra en el rango posible (Ding *et al.*, 2006). La expresión que permite calcular este indicador es

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

donde $N$ es la cantidad de sujetos que responden el *test*; $K$, el número de ítems que conforman el *test*; y $f_i$, el número de ocurrencias de cada una de las calificaciones. Un *test* con una buena discriminación debe arrojar un delta de Ferguson superior a 0.9.

Los trabajos de Ding *et al.* (2006), McKagan *et al.* (2010), Barbosa (2013), Barniol, Capos y Zavala (2018), y Zavala *et al.* (2019) reflejan el uso de estos indicadores.

## DISEÑO DEL *TEST*

Para el diseño del *test*, el proceso fue el siguiente (Medina y Ramírez, 2019):

1) Diseño y aplicación de una encuesta, la cual se llevó a cabo de esta manera (Medina y Ramírez, 2016):
   - Estudio bibliográfico para conocer el estado del arte. Arrojó múltiples estudios relativos al sonido.
   - Consulta a profesores de física en ejercicio, como información útil a la hora de planear una encuesta.
   - Elaboración de una encuesta de respuesta abierta relativa, principalmente, a los fenómenos de reflexión y refracción de ondas sonoras. En un principio se compuso de doce preguntas, las cuales fueron sometidas a evaluación de expertos.
   - Aplicación de la encuesta a estudiantes con el propósito de identificar concepciones erróneas.

2) Con la información acerca de las concepciones que presentaron los estudiantes, iniciamos el proceso de diseño de un *test* que permitiera averiguar la comprensión de los fenómenos de reflexión y refracción del sonido. Optamos por este tipo de instrumento porque, a pesar de la dificultad que implica la elaboración de los reactivos y el tiempo que esto involucra, es confiable desde un punto de vista estadístico y permite medir logros de aprendizaje diversos en un amplio rango de niveles y áreas temáticas (Aiken, 2003; López e Hinojosa, 2016).

Con la información acerca de las concepciones que presentaron los estudiantes, iniciamos el proceso de diseño de un *test* que permitiera averiguar la comprensión de los fenómenos de reflexión y refracción del sonido

3) El paso siguiente al diseño del *test* fue la formulación de un método que ayudara a recoger sugerencias de un grupo de expertos, con el propósito de hacer las modificaciones necesarias que le otorgaran al instrumento validez de contenido por medio del juicio de expertos (Hernández y Mendoza, 2018).

4) Una vez diseñado el *test*, procedimos a su implementación en línea con ayu-

da de las TAC a fin de que, tanto los expertos como los estudiantes, lo pudieran evaluar y responder, respectivamente. Los detalles se muestran en la sección siguiente.

## IMPLEMENTACIÓN EN LÍNEA DEL *TEST*

La idea fundamental de la recopilación de información cualitativa y cuantitativa por medio de un *test* es tener un panorama amplio de un fenómeno a través del muestreo en condiciones reales y variadas; para el caso particular que esta investigación plantea, era deseable poder aplicar el instrumento en diferentes latitudes geográficas. Esto nos obligó a pensar en la creación de un sistema web que fuera accesible de manera global.

Con la intención de conservar el concepto del *test* de tal modo que se eviten comportamientos como la copia de respuestas, incluso en internet, planteamos las siguientes restricciones:

- El *test* tendrá un tiempo límite para ser respondido.
- Si un usuario decide salir del *test*, se pausará el tiempo para reanudarlo la próxima vez que ingrese al sistema.
- Las preguntas serán desplegadas de manera aleatoria a cada usuario.
- Una vez que el usuario haya contestado una pregunta, ya no podrá corregir su respuesta.
- Una vez que el usuario haya terminado de contestar el *test*, no podrá acceder a las preguntas otra vez.
- Este *test* fue pensado para que alumnos de diferentes niveles educativos y áreas del conocimiento puedan contestarlo; no obstante, existe un total de diez preguntas que se consideraron solo para estudiantes del área de ciencias de la salud.

Para lograr lo anterior y llevar un control de las respuestas obtenidas a lo largo de la aplicación de la encuesta, fue necesario crear un modelo de datos para identificar a los alumnos, conocer su información de procedencia, nacionalidad, área de estudios e incluso la institución a la que pertenecen.

En cuanto al análisis estadístico de los datos, diseñamos una herramienta dentro del propio sistema que arroja las respuestas de cada alumno en función de las siguientes variables: nombre, pregunta respondida, área del conocimiento, género e institución. También, buscamos que el sistema proyectara una visualización estadística de los datos obtenidos en forma de gráfica.

Además, con la idea de que los datos producto de la aplicación del *test* pudieran ser analizados en otros sistemas de tecnologías de la información, el sistema añade la funcionalidad de exportar los datos obtenidos a un formato .csv.

Como señalamos, nuestra intención era que el *test* pudiera ser aplicado en diferentes países, por lo que fue necesario desarrollar un *software* que permitiera a los usuarios acceder sin importar la hora ni la ubicación geográfica y que, de igual forma, mantuviera la integridad de los datos. Por tal razón, un sistema web fue la opción más adecuada para el cumplimiento de esos requerimientos.

Aun cuando hay diversos *frameworks* de trabajo que facilitarían el desarrollo de la plataforma, como Angular o React, es importante resaltar que el sistema fue pensado como una extensión a un sitio web ya existente que acopla otro tipo de tecnologías, entre las que se destaca Java Web.

Por otro lado, debemos mencionar que PHP es uno de los lenguajes con mayor soporte dentro de los servidores web comerciales que existen en la actualidad. Por lo anterior, decidimos trabajar con este lenguaje como principal herramienta en el servidor. Para dar soporte al almacenamiento y la gestión de los datos que se recopilen, propusimos el motor MySQL, sistema gestor de base de datos relacional que permite la creación de vistas, transacciones y procedimientos almacenados de manera nativa.

Con las especificaciones anteriores, el equipo de trabajo decidió utilizar los lenguajes JavaScript

y PHP para la elaboración de la versión digital del *test*, debido a que son lenguajes que mejoran la calidad de las interfaces y poseen bastante documentación; y usar las tecnologías CSS y HTML para la construcción del sitio. El sitio web se alojó en la siguiente dirección: http://physics-education.tlamatiliztli.net/index.php.

Ya con el sistema implementado en el sitio web, procedimos al paso tres del proceso de evaluación descrito. La figura 1 contiene la pantalla de ingreso al sistema.

## VALIDEZ DE CONTENIDOS POR MEDIO DEL JUICIO DE EXPERTOS

El grupo de expertos estuvo integrado por ocho académicos de diferentes universidades (chilenas y mexicanas) con posgrados en Fonoaudiología, Física, Física Educativa e Innovación Educativa; se configuró siguiendo un criterio basado en el área de experticia: física, enseñanza de la física y ciencias de la salud, con el propósito de que la "retroalimentación" obtenida procediera de las tres áreas involucradas en la investigación y así enriqueciera su contenido.

Al ingresar al sistema, los expertos tenían acceso a las preguntas de la primera propuesta del *test* (ver figura 2, página siguiente). En relación con cada una de estas, la solicitud fue que evaluaran el nivel de dificultad de la pregunta en una escala de 0 a 10, donde 0 significó que la pregunta era muy fácil y 10, que era muy difícil –método de completación de frases (Hodge & Gillespie, 2003)– y, al mismo tiempo, que entregaran una opinión que consideraran pertinente de cada una de ellas (Hernández y Mendoza, 2018).

Además del análisis de cada pregunta, les solicitamos a los expertos que respondieran a las siguientes preguntas relativas al *test* en su conjunto:

- Presentación de las preguntas. El tipo y tamaño de letra ¿son adecuados?, ¿las imágenes están bien distribuidas?, etcétera.


**Figura 1.** Pantalla de ingreso al sistema de aplicación y evaluación del *test*.
Fuente: sitio web.

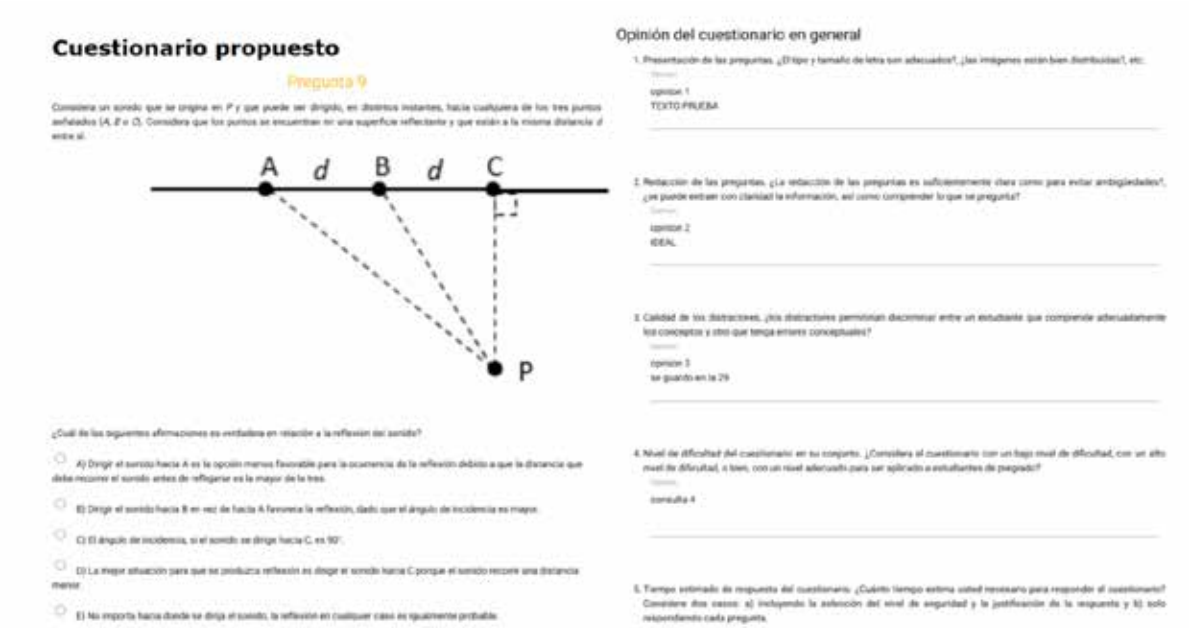


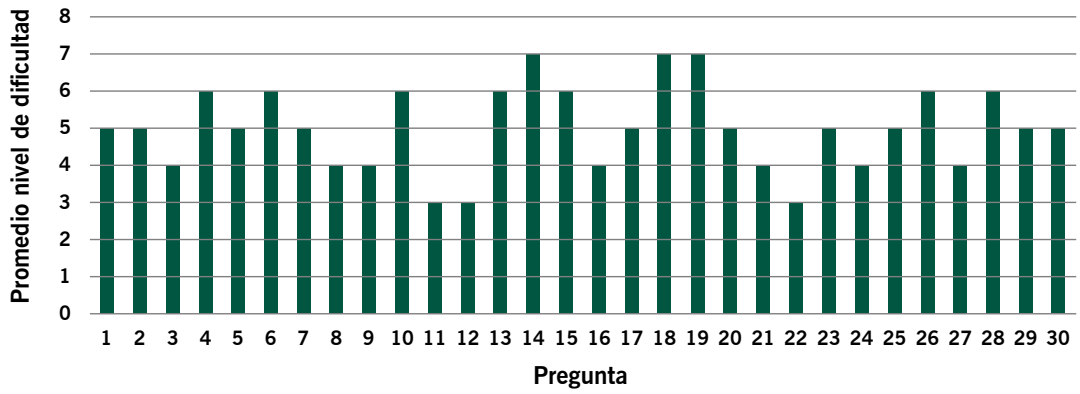

**Figura 2**. Vista del test para la evaluación de los expertos proporcionada por el sistema.
Fuente: sitio web.

- Redacción de las preguntas. ¿La redacción de las preguntas es suficientemente clara como para evitar ambigüedades?, ¿se puede extraer con claridad la información, así como comprender lo que se pregunta?
- Calidad de los distractores. ¿Los distractores permitirían discriminar entre un estudiante que comprende adecuadamente los conceptos y otro que tenga errores conceptuales?
- Nivel de dificultad del cuestionario en su conjunto. ¿Considera el cuestionario con un bajo nivel de dificultad, con un alto nivel de dificultad, o bien, con un nivel adecuado para ser aplicado a estudiantes de pregrado?
- Tiempo estimado de respuesta del cuestionario. ¿Cuánto tiempo estima usted necesario para responder el cuestionario? Considere dos casos: a) incluyendo la selección del nivel de seguridad y la justificación de la respuesta y b) solo respondiendo cada pregunta.
- Otros comentarios que crea pertinentes.

Del análisis del cuestionario aplicado por los expertos, obtuvimos los siguientes resultados:

a) Nivel de dificultad

El promedio de dificultad de cada una de las preguntas asignadas por los expertos se muestra en la gráfica 1 (página siguiente). El nivel de dificultad asignado varía dentro de un rango de cuatro puntos, en el cual ninguna pregunta es demasiado fácil, ni demasiado difícil. A juicio de los expertos, un 10% de las preguntas son fáciles, un 10%, difíciles, y el resto es de dificultad media. El promedio de dificultad del *test* en su conjunto es de cinco. Desde un punto de vista cualitativo, los resultados anteriores son favorables para considerar que el *test* tiene un nivel de dificultad adecuado.




**Gráfica 1.** Promedio del nivel de dificultad frente al número de pregunta.
Fuente: elaboración propia.

### *b) Comentarios a cada una de las preguntas y al test en su conjunto*

Las respuestas a estas coincidieron en que la presentación de las preguntas era adecuada, tanto en letra como en imágenes; que la redacción era suficientemente clara y permitía comprender lo que se solicitaba en cada pregunta; solo se sugirió modificar algunos reactivos que podrían generar ambigüedad y, en otros casos, mejorar la métrica de estos para dotarlos de homogeneidad; que los distractores ayudaban a discriminar entre aquellos estudiantes que comprenden los conceptos y los que no; se sugirió prestar atención a los ítems que contuviesen distractores demasiado heterogéneos; que el nivel de dificultad del *test* es adecuado para estudiantes de pregrado y que las preguntas implicaban diferentes grados de dificultad e involucraban dimensiones cognitivas distintas. El promedio del tiempo de respuesta señalado por los expertos, incluyendo la selección del nivel de seguridad y la justificación, fue de 100 minutos, mientras que el promedio del tiempo si solo se respondieran las preguntas sin seleccionar el nivel de seguridad ni la justificación fue de 64 minutos.

Los comentarios fueron utilizados para mejorar la primera versión del *test* y obtener una segunda versión validada por los expertos, la cual fue puesta en línea para que los estudiantes respondieran en el mismo sitio web. Una vez registrados, los estudiantes debieron ingresar como usuarios y responder el *test*. Si el estudiante era del área de ciencias médico-biológicas, se desplegaban 30 preguntas con un tiempo máximo de respuesta de 65 minutos y si era del área de físico-matemáticas, las preguntas eran 20 con un tiempo máximo de respuesta de 44 minutos. En ambos casos se presentaba un contador regresivo de tiempo y las preguntas fueron mostradas al azar; cada una de estas tenía que responderse antes de avanzar a la siguiente (ver gráfica 1).

## CONFIABILIDAD

La segunda versión del *test* fue administrada en línea a un total de 288 estudiantes de licenciatura de universidades de Chile, Colombia y México. De los 288 estudiantes, 233 fueron del área físico-matemáticas y 55, del área médico-biológicas; 121 fueron mujeres y 167, hombres. Contar solo con 55 estudiantes del área médico-biológicas hizo que los resultados fueran poco concluyentes, razón por la cual decidimos hacer el análisis para

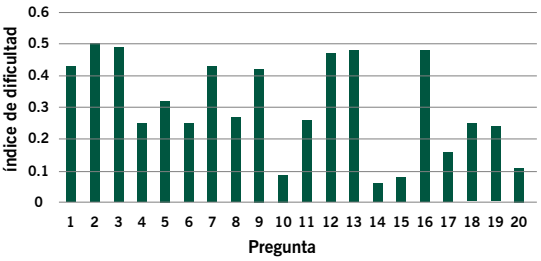
las 20 preguntas generales con los 288 sujetos de la muestra. De las 20 preguntas generales, las primeras diez evaluaron aspectos de la reflexión del sonido y las diez preguntas restantes evaluaron aspectos de la refracción del sonido.

Al considerar las primeras diez preguntas de reflexión, la media sería 0.34 y el promedio de las preguntas de refracción, 0.26. Los indicadores presentaron los valores que se encuentran en las tablas y gráficas que se presentan a continuación.

### *a) Índice de dificultad (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 | |

Fuente: elaboración propia.

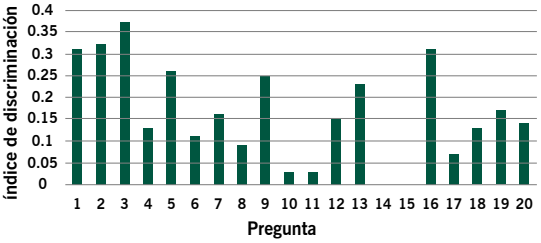



### *b) Índice de discriminación (D)*

- Cálculo con el método 50% - 50%

| 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 | |

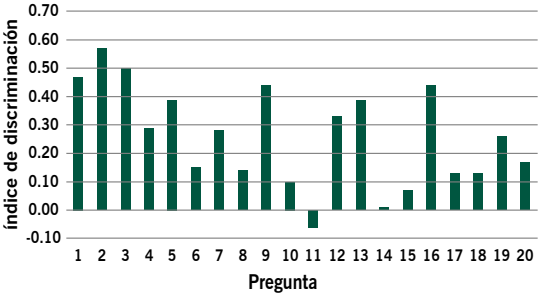
Fuente: elaboración propia.




- Cálculo con el método 25% - 25%

| 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 | |

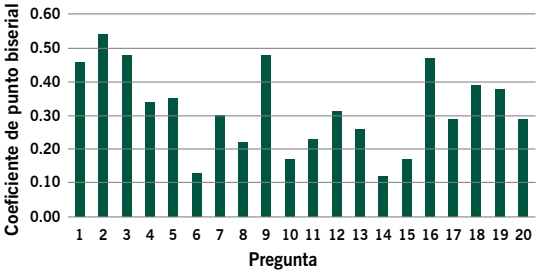
Fuente: elaboración propia.




### *c) Coeficiente de punto biserial ($r_{pbs}$)*

| 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 | |

Fuente: elaboración propia.



### *d) Índice de confiablidad de Kuder-Richardson ($KR_{20}$)*

Debido a que el *test* busca evaluar aspectos relacionados con la reflexión y la refracción de ondas sonoras y el índice de confiablidad de Kuder-Richardson es un indicador de la homogeneidad del *test*, es conveniente calcular un índice para el conjunto de preguntas sobre reflexión ($KR_{20A}$), que son las diez primeras preguntas, y calcular un índice para el conjunto de preguntas sobre refracción ($KR_{20B}$), que son las segundas diez preguntas. Los valores para cada uno de estos fueron $KR_{20A}$= 0.24 y $KR_{20B}$= -0.14.

### *e) Delta de Ferguson ($\delta$)*

De igual forma, para el índice de confiablidad de Kuder-Richardson, calculamos una delta para el conjunto de preguntas sobre reflexión ($\delta_A$) y una delta para el conjunto de preguntas sobre refracción ($\delta_B$). Los valores fueron $\delta_A$= 0.87 y $\delta_B$= 0.81.

## ANÁLISIS DE RESULTADOS

Como mencionamos, la validez se obtuvo por medio de la evidencia de validez por contenido y esta, a su vez, a través de la evaluación de expertos. Esta validez se vio reflejada en la segunda versión *test*, la cual fue administrada a los 288 estudiantes que participaron en el estudio a fin de conferir confiabilidad al instrumento. Los resultados de los indicadores respectivos se analizan a continuación

En relación con los valores del índice de dificultad de cada pregunta, encontramos que las preguntas 4, 6, 8, 10, 11, 14, 15, 17, 18, 19 y 20 resultaron ser muy difíciles, y la 10, 14 y 15 tuvieron una dificultad extrema. Las preguntas 1, 2, 3, 5, 7, 9, 12, 13 y 16 resultaron con una dificultad moderada. Ninguna de estas fue fácil.

El nivel óptimo de la dificultad de un ítem es 0.5, pero como es casi imposible lograr que todas las preguntas tengan tal grado de dificultad, se ha propuesto más de un criterio para seleccionar adecuadamente las preguntas para un *test*. Uno de ellos es utilizar un rango de dificultad desde el 0.4 al 0.6 como un intervalo que recoge el nivel óptimo de 0.5. Un segundo criterio es considerar un rango que va desde 0.3 a 0.9, y eliminar los elementos muy fáciles (sobre 0.9) y los muy difíciles (bajo 0.3). Lo anterior obedece a que las preguntas muy fáciles y las preguntas muy difíciles no contribuyen a la capacidad discriminatoria de un *test* (Doran, 1980).

Un tercer criterio es combinar preguntas con diferentes grados de dificultad según el criterio de la tabla 1.

**Tabla 1.** Niveles de dificultad

| Concepto | Rango de valores | Porcentaje de ítems |
|---|---|---|
| Muy fácil | 0.85 – 1.00 | 15 |
| Moderadamente fácil | 0.60 – 0.85 | 35 |
| Moderadamente difícil | 0.35 – 0.60 | 35 |
| Muy difícil | 0.00 – 0.35 | 15 |

Fuente: traducido de Doran R. (1980). *Basic Measurement and Evaluation of Science Instruction*, p. 97. Washington: National Science Teachers Association.

Al considerar el segundo criterio señalado, podríamos seleccionar para el *test* las preguntas 1, 2, 3, 5, 7, 9, 12, 13 y 16.

Respecto al índice de discriminación, Ebel y Frisbie (1991) exponen la clasificación para la discriminación de los ítems en la tabla 2.

**Tabla 2.** Índice de discriminación de los ítems

| Índice de discriminación | Evaluación del ítem |
|---|---|
| 0.40 y más | Muy buen ítem |
| 0.30 a 0.39 | Razonablemente bueno, pero posiblemente sujeto a mejoras |
| 0.20 a 0.29 | Ítem marginal que, por lo general, requiere mejoras |
| Por debajo de 0.20 | Ítem pobre, debe ser descartado o mejorado mediante una revisión |

Fuente: traducido de Ebel, R. & Frisbie, D. (1991). *Essentials of educational measurement*, p. 232. Englewood Cliffs, NJ: Prentice-Hall.

De lo anterior, deducimos que un ítem con un índice mayor o igual a 0.30 aporta una buena discriminación; por ello, de acuerdo con los valores obtenidos, el 1, 2, 3 y 16 tienen una buena discriminación, siguiendo el método 50% - 50%; sin embargo, si se sigue el método 25% - 25%, las preguntas que aportan una buena discriminación entre estudiantes con buen rendimiento y estudiantes con mal rendimiento son la 1, 2, 3, 5, 9, 12, 13 y 16. Con valores cercanos a 0.30 están la 4 y 7. Observamos que las preguntas discriminan mejor en grupos más extremos de rendimiento.

En lo relativo al coeficiente de punto biserial, al considerar que un valor adecuado para este indicador es aquel que es mayor o igual a 0.2, las preguntas 1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 16, 17, 18, 19 y 20 presentan una buena consistencia. Los ítems con valores menores de 0.2 no deben, de modo necesario, descartarse y "aún pueden permanecer en una prueba, pero debe haber pocos de ellos" (Ding *et al*., 2006, p. 3), por lo cual las preguntas 10 y 15 podrían ser consideras en el *test*, ya que sus valores son relativamente cercanos a 0.2.

Referente al índice de confiablidad, los criterios no son únicos y estos pueden variar de acuerdo con los evaluadores y el propósito de un *test*. La tabla 3 resume algunos criterios ampliamente aceptados (Doran, 1980).

**Tabla 3.** Índice de confiabilidad

| Valores del índice | Criterio de confiabilidad |
|---|---|
| 0.95 – 0.99 | Muy alto, rara vez encontrado |
| 0.90 – 0.95 | Alto, suficiente para la evaluación de individuos |
| 0.80 – 0.90 | Alto, podría considerarse para la evaluación individual |
| 0.70 – 0.80 | Bueno, suficiente para la medición grupal, no para individuos |
| Por debajo de 0.70 | Bajo, útil solo para promedios o encuestas grupales |

Fuente: traducido de Doran, R. (1980). *Basic Measurement and Evaluation of Science Instruction*, p. 104. Washington: National Science Teachers Association.

Los valores del índice para el conjunto de preguntas sobre reflexión fueron de 0.24 y para el conjunto de preguntas sobre refracción, de -0.14. En ambos casos, los valores resultaron demasiado bajos; sin embargo, los índices de confiablidad están influenciados por una serie de factores, como extensión del *test*, dificultad y discriminación de las preguntas, así como rango de habilidad de los sujetos de la muestra. Al eliminar preguntas con índices de dificultad "lejanos" a 0.5 (considerado el nivel óptimo) y con índice de discriminación "demasiado bajos", sube el valor del índice de confiablidad.

Por otro lado, Adams y Wieman (2011) argumentan que los instrumentos diseñados para medir múltiples conceptos pueden tener un bajo alfa de Cronbach (que para este *test* es equivalente al índice de confiabilidad de Kuder-Richardson debido a que el instrumento es dicotómico), porque estos conceptos pueden ser independientes. Es el caso de esta prueba que, a pesar de que evalúa solo dos fenómenos (reflexión y refracción), estos involucran más de un concepto.

En lo referente a los valores de la delta de Ferguson, estos fueron de 0.87 para el conjunto de preguntas sobre reflexión y de 0.81, para el conjunto de preguntas sobre refracción. Los valores resultan cercanos al número deseado, que es 0.9.

Finalmente, aun cuando el análisis completo se realizó para las 20 preguntas generales, mostramos los valores de los tres primeros indicadores (los relativos a cada ítem y no al *test* en su conjunto) obtenidos para las preguntas específicas de ciencias de la salud.

### *a) Índice de dificultad (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 |

Fuente: elaboración propia.

La dificultad de las preguntas 23, 24 y 28 resultan moderadas, y el resto, difíciles, a excepción de la 27, que fue en particular complicada.

### *b) Índice de discriminación (D)*

Cálculo con el método 25% - 25%

| 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 |

Fuente: elaboración propia.

Las preguntas 23, 24, 25 y 26 presentan una buena discriminación; la 21, 22 y 28, una discriminación de 0.29, por lo que no deben descartarse *a priori*. Las preguntas 27 y 30 resultaron con baja discriminación.

### *c) Coeficiente de punto biserial ($r_{pbs}$)*

| 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 |

Fuente: elaboración propia.

Todas las preguntas presentan un coeficiente superior al mínimo aceptado, es decir, hay mayor probabilidad que los alumnos con mejores puntajes contesten bien las preguntas.

## CONCLUSIONES

El objetivo de esta investigación se cumplió de manera parcial, debido a que se logró elaborar un *test* de indagación conceptual sobre fenómenos sonoros orientado a ciencias de la salud, pero no en los términos originalmente planteados. La validez del instrumento fue posible obtenerla a través de la evidencia de contenido, resultados que, además, ya fueron publicados (Medina y Ramírez, 2019); sin embargo, la confiabilidad se alcanzó solo de modo parcial, en virtud de la muestra reducida que se reunió.

La muestra "reducida" con que trabajamos permitió hacer un análisis más consistente de las 20 primeras preguntas y un análisis menos fuerte con las últimas diez preguntas orientadas por completo a ciencias de la salud. Las preguntas que cumplen los estándares de confiablidad sin ningún problema son la 1, 2, 3, 5, 9, 12, 13 y 16. No obstante, algunas preguntas restantes pueden ser consideradas; la 4, si bien resultó difícil, tiene un coeficiente de punto biserial adecuado y su discriminación es de 0,29; la 7 reporta una dificultad moderada, un adecuado coeficiente de punto biserial y una discriminación de 0.28; la 19 presenta una dificultad alta, un coeficiente de punto biserial adecuado y una discriminación de 0.26. El resto debe ser revisado en profundidad.

En relación con el índice de Kuder-Richardon, que mide la consistencia del *test*, los valores que se obtuvieron fueron bajos, lo que podría explicarse por el hecho de que, si bien es cierto, el *test* mide solo dos fenómenos, los elementos involucrados son diversos.

La construcción de un *test* que sigue un proceso de trabajo en línea tiene ventajas, pero a la vez cierta exigencia adicional. Entre las ventajas, podemos mencionar la "atemporalidad", pues en teoría los estudiantes pueden contestar el instrumento en cualquier instante. Además, en este caso particular, el sistema guardaba automáticamente las respuestas y tenía un contabilizador de tiempo que se activaba solo cuando el estudiante contestaba, lo que permitía a quienes respondían salir del sistema si fuera necesario.

Es válido hacer notar que la aplicación en línea evita un enorme gasto de papel y tinta, además de posibles pérdidas o deterioro del material impreso. El ingreso de datos en algún programa de análisis estadístico también resulta más sencillo debido a que las respuestas se pueden rescatar, por ejemplo, en una planilla de Excel. La exigencia adicional está en cómo generar un ambiente cooperativo en que los involucrados se sientan motivados a participar, situación que en forma presencial podría facilitarse, en comparación con una situación virtual, en la que el contacto con los participantes puede ser escaso.

Como trabajo futuro, es necesario una revisión en profundidad de las preguntas que no alcanzaron valores apropiados de los parámetros, con el propósito de reformularlas adecuadamente, y también reducir los fenómenos a uno solo, ya sea reflexión o refracción, lo que facilitará una mayor homogeneidad y la elaboración de nuevos *test* que evalúen a cada concepto por separado.

Asimismo, un trabajo a futuro es elaborar un *test* que, en efecto, se oriente a las ciencias de la salud, donde la fiabilidad se alcance utilizando la teoría de respuesta al ítem, no porque la teoría clásica de los *test* sea insuficiente, sino por la riqueza de la información adicional que entrega un análisis con aquella teoría. Para esto, se requiere una muestra considerablemente mayor de la que se obtuvo en este trabajo, en general superior a 500 (Martínez *et al*., 2006).

## REFERENCIAS BIBLIOGRÁFICAS


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, D. C.: National Science Teachers Association.

Ebel, R. & Frisbie, D. (1991). *Essentials of educational measurement*. Englewood Cliffs, N. J.: 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.

Hesten, 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

Este artículo es de acceso abierto. Los usuarios pueden leer, descargar, distribuir, imprimir y enlazar al texto completo, siempre y cuando sea sin fines de lucro y se cite la fuente.

## CÓMO CITAR ESTE ARTÍCULO:

Medina Paredes, Jhonny; Ramírez Díaz, Mario Humberto y Miranda, Isaías. (2019). Validez y confiablidad de un test en línea sobre los fenómenos de reflexión y refracción del sonido. *Apertura, 11*(2), pp. 104-121. http://dx.doi.org/10.32870/Ap.v11n2.1622