



DISEÑO Y VALIDACIÓN DE UNA PRUEBA DE FÍSICA PARA EDUCACIÓN SECUNDARIA

ROSA FABIOLA PÉREZ SALAS

UNIVERSIDAD PEDAGÓGICA DE DURANGO
fabbiola_@hotmail.com

MARÍA CONCEPCIÓN SOSA ÁLVAREZ

UNIVERSIDAD POLITÉCNICA DE DURANGO
msosas74@gmail.com

RESUMEN

Los objetivos del presente estudio son: diseñar un instrumento para medir los aprendizajes esperados de la asignatura de Ciencias II (énfasis en Física), en el Bloque II y validar el instrumento desde una perspectiva clásica por medio de consulta a expertos y pruebas paramétricas. Para lo cual se realizó un estudio instrumental mediante el análisis de reactivos de tres pruebas piloto aplicadas 104 estudiantes de secundaria en diferentes momentos, para posteriormente construir la versión final de la prueba. Los resultados muestran índices de discriminación y dificultad adecuados cercanos al 0.5, así como un alto nivel de confiabilidad.

Palabras clave: aprendizaje de las ciencias, pruebas pedagógicas, índice de dificultad, índice de discriminación

INTRODUCCIÓN

El aprendizaje de los alumnos es el foco de atención de los diversos centros educativos, entre ellos las escuelas de educación básica, que plantean una educación centrada en los procesos de aprendizaje de los alumnos, que propicie ambientes de aprendizaje que favorecen experiencias que significativos y de interés para los estudiantes y que parta de sus saberes previos, para la construcción de los nuevos conocimientos (SEP, 2011a).

Esta propuesta de educación involucra aprendizaje significativo, el cual se define como el proceso a través del cual nuevos contenidos o conocimientos son relacionados de modo no





arbitrario y sustancial, es decir, no al pie de la letra con lo que el alumno ya sabe. En otras palabras sus conocimientos previos se relacionan con algún aspecto existente específicamente relevante de la estructura cognoscitiva del alumno, como una imagen, un símbolo, un concepto o una proposición (Ausubel, 1983).

El aprendizaje significativo ocurre cuando una nueva información "se conecta" con un concepto relevante ("subsursor") preexistente en la estructura cognitiva, esto implica que, nuevas ideas pueden ser aprendidos significativamente en la medida en que otras ideas estén adecuadamente claras y disponibles en la estructura cognitiva del individuo y que funcionen como punto de "anclaje" a las primeras (Ausubel, 1983).

Un ejemplo de lo anterior en física, es que si los conceptos de fuerza y velocidad se encuentran en la estructura cognitiva del estudiante, serán subsensores para nuevos conocimientos como las Leyes de Newton, funcionarían como puntos de anclaje; la interacción con la nueva información a su vez producirá una modificación en estos conceptos subsensores que evolucionarán para comprender otros conceptos de mayor complejidad.

El logro del aprendizaje significativo no es sencillo y sobre todo en el área de ciencias, por lo que este trabajo parte de la dificultad del aprendizaje de las ciencias mostrada por los estudiantes en comparación con otras áreas o asignaturas, prueba de ello son los resultados en pruebas estandarizadas como la Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE) y el Programa para la Evaluación Internacional de Alumnos (PISA) en donde los estudiantes en México han obtenido puntajes muy bajos, colocándose en niveles elementales e insuficientes (ENLACE, 2013; INEE, 2013).

Corroborar el nivel del aprendizaje de los estudiantes tampoco es tarea fácil. Existen varias opciones de evaluación y una de ellas son las llamadas pruebas escritas o exámenes, consideradas herramientas educativas de gran alcance, que pueden utilizarse para la obtención de evidencias de aprendizaje en los alumnos de educación básica del nivel secundaria (SEP, 2011b). Para que estas pruebas sean coherentes con los planteamientos constructivistas del programa de educación básica, deben incluir actividades con énfasis en los aspectos metodológicos, actitudinales, e incluso en la comprensión y manejo significativo de los contenidos conceptuales (Alonso, Gil & Martínez Torregoso, 1992).





En cumplimiento con los requerimientos anteriores “las pruebas de comprensión deben ser redactadas en un lenguaje diferente y deben ser presentadas en un contexto diferente al encontrado originalmente en el material de aprendizaje” (Ausubel, 2000, p.127). De esta manera se evita el riesgo de caer en una simulación mecánica de comprensión significativa, al plantear preguntas y problemas novedosos e incluso desconocidos para el estudiante, que implican una transformación máxima de los conocimientos existentes (Ausubel, 2000).

En el área de ciencias, se identificaron las siguientes investigaciones: dos en física, una a nivel preparatoria (García & Rentería, 2012) y otra a nivel universitario (Jandaghi, 2010); dos más fueron en el área de química (Jandaghi, G., & Shaterian, F., 2008; Martín, Díaz, Córdoba & Picquart, 2011) y una en biología (Arocena, Gascue, & Leymonié, 2010). Algunas otras investigaciones corresponden a pruebas matemáticas (López, 2009) y en ciencias médicas (Pérez, Acuña & Arratia, 2008). El único estudio que se localizó en México (Backhoff, Larrazolo, & Rosas, 2000) tuvo como propósito la validación de pruebas para selección de estudiantes para ingreso a un nivel universitario.

La revisión de la literatura arroja pocas investigaciones de índole instrumental en el área de física aunado a ello ninguno de estos estudios se llevó a cabo en México. Esta ausencia de investigaciones remite a la necesidad e importancia de realizar un estudio acerca del tema en cuestión. Por lo que esta investigación se plantea como objetivos: a) Diseñar un instrumento para medir los aprendizajes esperados en la asignatura de Ciencias II (énfasis en física), en el Bloque II y b) Validar el instrumento desde una perspectiva clásica por medio de consulta a expertos y pruebas paramétricas.

METODOLOGÍA

La presente investigación corresponde a un estudio instrumental, que de acuerdo con Montero y León (2007) “se consideran pertenecientes a esta categoría todos los estudios encaminados al desarrollo de pruebas y aparatos, incluyendo tanto el diseño (o adaptación) como el estudio de las propiedades psicométricas de los mismos” (pp.855-856).





El instrumento diseñado es una prueba académica, de acuerdo con Salkind (1999) “las pruebas de aprovechamiento sirven para medir los conocimientos en un área específica, y son las más comúnmente utilizadas cuando el resultado que se está midiendo es el aprendizaje”. La prueba aquí descrita se denominó Prueba de Ciencias II, Bloque 2 (PCII-B2). Está conformada por 21 ítems de opción múltiple, cada uno con cuatro opciones de respuesta. Tiene como propósito evaluar conocimientos y habilidades de la unidad o bloque II “Leyes del Movimiento” de la asignatura de Ciencias II (énfasis en Física) perteneciente al Programa de Estudios de Ciencias 2011 del nivel secundaria (SEP, 2011b). Se eligió esta unidad porque incorpora temas de difícil comprensión para los estudiantes de segundo grado de secundaria.

El procedimiento para la elaboración y validación de los reactivos de la prueba constó de las siguientes etapas: construcción (a cargo del investigador), validación (expertos construcción de reactivos), piloteo (aplicación de la prueba a estudiantes que cursaron el segundo grado de secundaria) y calibración (análisis y selección de reactivos por medio de pruebas psicométricas).

La construcción de reactivos, se realizó a partir de una Tabla de Validez de Contenido, esta tabla confiere los criterios que deben cumplirse para la elaboración de los reactivos, por medio de una matriz que contiene los aprendizajes esperados, contenidos del bloque y los reactivos correspondientes a cada uno de ellos. A esta validación se sumó un juicio por parte de expertos en construcción de reactivos.

El piloteo se llevó a cabo en tres momentos diferentes, con distintas muestras, de manera tal que tuvieran condiciones sociodemográficas similares: 1) La primera prueba se realizó en la Escuela Secundaria Técnica (EST) No. 62, localizada en la ciudad de Durango, a una muestra de 32 alumnos; 2) La segunda prueba se aplicó en la EST No. 67, ubicada en la misma ciudad, en este caso la muestra fue de 34 alumnos y 3) La tercera aplicación se llevó a cabo de nuevo en la EST No. 62, con una muestra diferente a la anterior de 35 alumnos.

Las tres muestras, seleccionadas de forma no probabilística, dan una totalidad de 102 alumnos del 3º grado de educación secundaria, cuyas características sociodemográficas se describen a continuación:

Piloteo 1: a) 42.4% pertenecen al género masculino y 57.6% al género femenino

b) El 84.8% tiene una edad de 14 años y el 15.2% 15 años





Piloteo 2: a) 38.9% pertenecen al género masculino y 61.1% al género femenino

b) El 94.4% tiene una edad de 14 años y el 5.6% 15 años

Piloteo 3: a) 40% pertenecen al género masculino y 60% al género femenino

b) El 88.6% tiene una edad de 14 años y el 11.4%, 15 años

RESULTADOS

La primera versión de esta prueba se envió a juicio de expertos para validación de contenido y de estructura metodológica. De acuerdo con el juicio de tres expertos en cuanto al contenido, la totalidad de los ítems fue correspondiente al bloque elegido de la asignatura de Ciencias II del Plan de Estudios de Ciencias 2011. Por otra parte, la revisión metodológica de otros tres expertos sugirió correcciones en cuanto a redacción de los ítems y la puntuación ortográfica. Tales observaciones fueron tomadas en consideración para la reformulación y corrección de preguntas.

Posterior a de cada una de las aplicaciones de pilotaje, se realizó el análisis de reactivos, “técnicas y procedimientos matemáticos encaminados a verificar la calidad y pertinencia de los reactivos o preguntas de una prueba” (Chávez & Saade, 2010, p.15). Esta etapa permite la selección de los reactivos idóneos para incorporarse a las versiones del examen y la estimación de las características psicométricas del examen considerado como una unidad (Ceneval, 2013). El análisis de reactivos está basado en los parámetros de la Teoría Clásica del Test, referente básicamente a dos indicadores de las características psicométricas de cada uno de los reactivos:

- El *índice de dificultad (ID)*, se refiere a que tan fácil o difícil resulta el reactivo para la población. Es la proporción de examinados que contestaron bien el reactivo (Salkind, 1999).
- El *índice de discriminación (Id)*: qué tan eficiente es el reactivo para ayudar a diferenciar entre quienes saben y quiénes no. Es la fracción de los examinados del grupo alto que contestaron correctamente el reactivo menos la fracción de los examinados del grupo bajo que contesto correctamente el reactivo.

Para realizar el cálculo de estos índices es preciso dividir el total de puntajes de la prueba en un grupo alto y un grupo bajo, considerando el 27% de los puntajes más altos y el 27% de los puntajes más bajos (Salkind, 1999). En este estudio, en el primer pilotaje se tomaron 9 alumnos





para cada grupo, superior e inferior, en el segundo y tercer pilotaje los grupos inferior y superior estuvieron conformados de 10 estudiantes.

El valor óptimo del índice de dificultad del reactivo depende de varios factores, por tanto no existe una clasificación consensuada al respecto, por lo que para interpretar los resultados se toma como referente a Backhoff et al. (2000), quienes proponen agruparlos de la siguiente manera: altamente difíciles (< 0.32), medianamente difíciles (de 0.33 a 0.52), dificultad media (de 0.53 a 0.73), medianamente fáciles (de 0.74 a 0.86) y altamente fáciles (> 0.86).

Los niveles de discriminación se presentan en la tabla 1. Un índice de discriminación de 1.00 indica que el reactivo discrimina a la perfección, valor que casi nunca se obtiene. Por otro lado si el índice es -1.00, significa que todos los del grupo bajo lo contestaron correctamente y nadie del grupo alto contesto bien, situación que no se debe de presentar.

Tabla 1. Evaluación de los índices de discriminación

Índice de discriminación	Evaluación del ítem
≥ 0.40	Los ítems discriminan muy bien
0.30 – 0.39	Razonablemente buenos pero sujetos a mejoras
0.20 – 0.29	Ítems marginales, necesitan ser mejorados
≤ 0.19	Ítems pobres, deben ser eliminados o modificados para mejorar

Fuente: Traducción de Ebel y Frisbie (1991).

En la tabla 2 se muestra el concentrado de los resultados de las tres aplicaciones. En la primera aplicación la gran mayoría de los ítems eran de alta dificultad para los estudiantes, no discriminaban adecuadamente y la confiabilidad era baja, razón por la cual se procedió a la reformulación del examen. Para ello se utilizó el análisis de distractores, que no es otra cosa que contar el número de veces que se seleccionó cada distractor. Al encontrar que demasiados examinados del grupo superior o muy pocos del grupo inferior seleccionaron un distractor determinado, se procedió a modificarlo o reemplazarlo.

Tabla 2. Resultados generales de las aplicaciones piloto





Piloteo	Índice de dificultad medio	Índice de discriminación medio	Confiabilidad Alfa de Cronbach
1º	0.37	0.18	0.588
2º	0.39	0.23	0.744
3º	0.40	0.28	0.809

En la segunda prueba, hubo un considerable incremento en la confiabilidad, pero los índices sólo tuvieron un ligero aumento, por la cual se procedió nuevamente a la modificación de los distractores. En la tercera aplicación los cambios fueron más perceptibles tanto en los índices como en la confiabilidad. Para conformar la versión final de la prueba con reactivos satisfactorios se tomaron en cuenta tres criterios: 1) un valor de discriminación mayor a 0.40, 2) se evitaron en la medida de lo posible aquellos ítems muy fáciles o muy difíciles, 3) por último se valoró la inclusión del reactivo, para que fuera un número equitativo de ítems en cada aprendizaje esperado.

La distribución de los reactivos de acuerdo a su nivel de dificultad se muestra en la tabla 3. Los valores alejados de la dificultad media (0.5) no se recomiendan puesto que los reactivos con alto grado de dificultad o con un grado extremadamente bajo, no aportan información importante a la medición. Una regla general que se utiliza para la selección de reactivos es aceptar aquellos que se encuentren cerca de la mitad de la escala (Chávez & Saade, 2010). El nivel medio de dificultad de la prueba final es de 0.49 y su discriminación media es de 0.48, valores muy cercanos a la media. El rango de dificultad de los reactivos fue de 0.30 a 0.75, solamente se incluyeron dos reactivos altamente difíciles y uno medianamente fácil, debido a que representaban aspectos importantes de las habilidades a evaluar.

Tabla 3. Distribución de los reactivos de acuerdo con su nivel de dificultad

Nivel de Dificultad	N	%
Altamente fáciles (> 0.86)	0	0.0
Medianamente fáciles (0.74 - 0.86)	1	4.7
Dificultad media (0.53 - 0.73)	9	42.9
Medianamente difíciles (0.33 - 0.52)	9	42.9
Altamente difíciles (< 0.32)	2	9.5





La medición de la confiabilidad se realizó por alfa de Cronbach obteniendo un valor de 0.869. De acuerdo con Ebel y Frisbie (1991), las pruebas estandarizadas más publicadas han demostrado producir resultados que tienen fiabilidades en el rango de 0.85 a 0.95, valores considerados como altamente aceptables.

CONCLUSIONES

La propuesta de un examen o prueba académica para la evaluación del aprendizaje, es una idea criticada en la actualidad en donde se proponen diversas evaluaciones más encaminadas al tipo cualitativo. Sin embargo, estas herramientas son utilizadas actualmente para las evaluaciones estandarizadas, que plantean preguntas que propician la reflexión por parte del estudiante así como la aplicación de los conocimientos en situaciones diversas de la vida cotidiana, van más del tipo tradicional de preguntas memorísticas a preguntas con énfasis en el manejo significativo de conocimientos.

Para asegurar una evaluación de este tipo, más completa y que involucre diversas habilidades y conocimientos, es necesario diseñar instrumentos de alta calidad que proporcionen resultados confiables y válidos. La única manera de conseguir una prueba de esta naturaleza es calcular los índices de dificultad y discriminación, para tener un control sobre los reactivos de dicho instrumento. Esto implica tiempo y dedicación a fin de afinar y adecuar los ítems a las necesidades de la evaluación.

En el caso del presente estudio, fueron necesarias dos reformulaciones del examen original para lograr un alto nivel de confiabilidad e índices de discriminación y dificultad aceptables. Es importante señalar que la mayoría de los ítems eliminados corresponden al tipo de pregunta en el que se requiere del conocimiento de fórmulas para la resolución de problemas. Esto no es del todo lamentable, puesto que los reactivos que conforman la versión final de la prueba van dirigidos a aplicaciones cotidianas de la física, aspecto fundamental que destaca el Plan de Estudios 2011 y requerido para un aprendizaje significativo.

Los resultados de este estudio deben incidir en la forma común en que los docentes construyen los exámenes, se requiere de un análisis constante de los reactivos que se formulan para las diversas pruebas, de esta forma será posible identificar áreas de debilidad y se





contribuirá a mejorar la evaluación y tomar decisiones para el logro del aprendizaje de los estudiantes.

BIBLIOGRAFÍA Y REFERENCIAS

- Alonso, M., Gil, D., & Martínez Torregosa, J. (1992). Los exámenes de física en la enseñanza por transmisión y en la enseñanza por investigación. *Enseñanza de las ciencias* 10 (2), 127-138.
- Arocena, R., Gascue, C., & Leymoní, J. (2010). Evaluación y validación de pruebas parciales de opción múltiple de un curso universitario de primer año. *Revista Electrónica de Enseñanza de las Ciencias*, 9(3), 737-756.
- Ausubel, D. (1983). Teoría del aprendizaje significativo. *Fascículos de CEIF*.
- Ausubel, D. (2000). *The Acquisition and Retention of Knowledge: A Cognitive View*. Nueva York: Springer Science + Business Media.
- Backhoff, E., Larrazolo, R., & Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *REDIE. Revista Electrónica de Investigación Educativa*, 2(1). Recuperado de <http://www.redalyc.org/articulo.oa?id=15502102ayo,2000>
- Ceneval (2013). *Metodología Ceneval*. México: Autor.
- Chávez, C., & Saade, A. (2010). *Procedimientos básicos para el análisis de reactivos. Cuaderno técnico 8*. México: CENEVAL. Recuperado de <http://www.ceneval.edu.mx/ceneval-web/file.do?idFile=7494>
- Ebel, L.R., & Frisbie, D.A. (1991). *Essentials of Educational Measurement* (5ª ed.). Englewood Cliffs, NJ: Prentice Hall.
- ENLACE (2013). *Resultados de años anteriores*. Recuperado de http://www.enlace.sep.gob.mx/ba/resultados_antteriores/
- García, J.J., & Rentería, E. (2012). La medición de la capacidad de resolución de problemas en las ciencias experimentales. *Ciência & Educação (Bauru)*, 18(4), 755-767.
- Jandaghi, G. (2010). Assessment of Validity, Reliability and Difficulty Indices for Teacher-built Physics Exam Questions in First Year High School. *Arts and Social Sciences Journal. Educational Research and Review*, 5(11), 651-654.





- Jandaghi, G., & Shaterian, F. (2008). Validity, Reliability and Difficulty Indices for Instructor-Built Exam Questions. *Journal of Applied Quantitative Methods*, 3(2), 151-155.
- INEE (2013). México en PISA 2012 (1ª Ed.) México: Autor.
- López, P. (2009). Construcción y validación de una prueba para medir conocimientos matemáticos. *Horizonte Pedagógico* 1, 29-39.
- Martín, N., Díaz, C., Córdoba, G. & Picquart, M. (2011). Calibración de una prueba química por el modelo de Rasch. *Revista Electrónica de Investigación Educativa*, 13(2), 132-148. Recuperado de: <http://redie.uabc.mx/vol13no2/contenido-martindiazetal.html>
- Montero, I., & León, O.G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology*, 7(3), 847-862.
- Pérez, J.H., Acuña, N. & Arratia, E.R. (2008). Nivel de dificultad y poder de discriminación del tercer y quinto examen parcial de la cátedra de cito-histología 2007 de la carrera de medicina de la UMSA. *Cuadernos – Hospital de Clínicas*, 53(2), 16-22. Recuperado de <http://www.revistasbolivianas.org.bo/pdf/chc/v53n2/v53n2a03.pdf>
- Salkind, N.J. (1999). *Métodos de Investigación* (3ª ed.). México: Prentice Hall
- SEP (2011a). *Plan de estudios 2011. Educación Básica*. México: Autor.
- SEP (2011b). *Programas de estudio 2011. Guía para el Maestro. Educación Básica. Secundaria. Ciencias*. México: Autor.

