SERENDIPIA

Vol. 5 N° 10 Julio – Diciembre 2016

Revista Electrónica del Programa de Cooperación Interfacultades

ISSN: 2443-44-34

Diseño y evaluación de instrumentos de medida para garantizar la calidad de la información en educación

Amalio Rafael Sarco Lira Barreto
Escuela de Educación
Universidad Central de Venezuela
Correo e: asarcolira@gmail.com

Resumen

El propósito de este artículo es divulgar la forma como pueden ser diseñados y evaluados los instrumentos de medida utilizados en el campo educativo a los fines de garantizar la calidad de la información que suministran. La compilación de un conjunto de definiciones sobre conceptos básicos utilizados tanto en psicometría como en edumetría permite la descripción de los procedimientos utilizados en el ambiente de la Teoría Clásica para la evaluación de las preguntas -nivel de dificultad y de discriminación de cada una, así como el análisis de sus distractores- y de las pruebas en general –validez y confiabilidad de la medida. Estos procedimientos son explicados y ejemplificados, comentando bondades y limitaciones. Se describen fórmulas para la definición de los niveles de dificultad y discriminación de las pruebas en su totalidad con sus respectivas recomendaciones de cuando utilizarlas, algunos de ellos poco difundidos en la literatura consultada. Se distinguen dos tipos de pruebas de acuerdo a su cobertura las de uso restringido y las de largo alcance, para cada una de ellas se hacen recomendaciones específicas sobre los planes de construcción y el cálculo de los índices necesarios, en la evaluación de los instrumentos se observaran variaciones derivadas de esta clasificación previa. Finalmente, a modo de ejemplo se analizan los resultados de una prueba de matemáticas utilizada en la UCV con fines de selección en el área de Ciencias de la salud, la consideración del conjunto de índices calculados permite concluir que es adecuada para los fines que se destina.

Palabras claves: Proceso de Construcción de Pruebas. Tablas de Especificaciones. Índices de Dificultad y de Discriminación. Validez y Confiabilidad. Calidad de Pruebas y Preguntas.

Introducción

Saber que tan buenas son las pruebas que construimos requiere de esfuerzo y dedicación, como todas las tareas académicas que nos proponemos en la universidad, quizás lo

primero que debemos establecer con absoluta claridad es el para qué las construimos, cuál es el propósito y a quién van dirigidas. El para qué podría ser evaluar conocimientos, habilidades o destrezas y el propósito, por ejemplo, el diagnostico o la selección, y con respecto a quién van dirigidas se refiere a si será utilizada en un aula de clase o si su aplicación será abierta a un distrito escolar, una entidad federal o región, a un conjunto de instituciones. Dependiendo de las respuestas que demos a estas interrogantes estableceremos cuáles son las características que debe tener el instrumento y seleccionaremos entre los procedimientos disponibles los que consideremos más adecuados para definir los estándares de calidad requeridos para cada índice. Aun así, lo más complicado no es la evaluación de la prueba, sino su construcción, a este punto dedicaremos un espacio al inicio del artículo, ya que no podemos esperar unos resultados buenos al evaluar un instrumento que ha sido construido de manera improvisada, sin observar los cuidados mínimos.

En una segunda parte se presentara la compilación de conceptos básicos y la descripción de los procedimientos de cálculo con sus respectivas recomendaciones de uso. Para la interpretación de los resultados veremos como el comportamiento de las muestras a quienes sometemos a las pruebas, permiten la generación de estadísticas con base a las cuales decidimos otorgarle estándares de calidad a los instrumentos y les conferimos a las mismas unas características que no se observan en el instrumento como tal.

Una tercera sección de este articulo estará dedicada a la aplicación de los índices para la evaluación de una prueba de selección utilizada en la UCV durante 2011, para lo cual se analizara los resultados de los diferentes índices considerados en este artículo, y, finalmente se elaborara una serie de conclusiones y recomendaciones tanto para el uso de los índices como para la prueba utilizada.

La calidad de los instrumentos de medición en el campo educativo

Uno de los objetivos de este trabajo es dar a conocer la forma como deben ser evaluados los instrumentos de medida utilizados en el campo educativo para determinar su calidad, describir los procedimientos que se utilizan, precisar los criterios que se aplican y los conceptos estadísticos básicos aplicados en el campo de la psicometría y de la edumetría bien sea que el análisis se practique en el ámbito de la Psicología o de la Educación, dado que hay variables cuyo comportamiento es importante analizar desde ambos campos con propósitos que, en algunos casos, pueden ser diferentes.

La mayor intensidad en el uso de instrumentos de medición en el campo educativo está asociada con la producción de datos que sirvan de base a la evaluación de los aprendizajes, sin embargo, interesa también la información que nos proporcionan sobre los intereses y gustos de los educandos, conocer sobre algunos rasgos de personalidad, y

en general, para develar los datos originales que requerimos en el área de la investigación educativa, de manera que resulta amplia la necesidad de uso de estos instrumentos en el ámbito educativo.

Los objetivos con los cuales los utilizamos suelen ser muy variados: cuantificar el aprendizaje de los contenidos o el grado de logro de los objetivos programáticos para medir los avances, como es el caso de los docentes en el aula al aplicar las pruebas de conocimientos específicos; para generar las calificaciones individuales, o diagnosticar las condiciones académicas de los estudiantes para definir planes de nivelación, tal como se propone en el documento El Ingreso Asistido, Sarco Lira (2010); seleccionar de un grupo a los más aventajados, los que muestren la mejor ejecución, como en las pruebas internas que se aplican en las universidades para asignar los cupos; o bien los que satisfagan los criterios establecidos en el proceso de selección, cuando se trata por ejemplo de la selección de personal para determinados cargos; de la misma manera son utilizados para develar información requerida en una investigación del área educativa, definir un perfil de los estudiantes con base a sus gustos, intereses, o características personales o socioeconómicas.

En este sentido es necesario puntualizar que a medida que se incremente el radio de acción de o se incremente el grupo al cual debamos aplicar el instrumento, habrá que ser más meticuloso en la observancia de ciertas normas o procedimientos técnicos, al docente de aula probablemente le bastara con verificar que las preguntas formuladas evalúan el contenido u objetivo programático tratado en determinado lapso, quienes apliquen pruebas diagnostico o de selección deberán disponer de un plan de construcción basado en los programas oficiales de las asignaturas y los investigadores se valdrán de jueces a quienes muestren sus cuestionarios y los objetivos de su investigación así como el informe metodológico para que esos grupos de expertos certifiquen que con esas preguntas se obtiene la información necesaria o pertinente para llevar adelante la investigación propuesta.

Validez y Confiabilidad, bases de la calidad

El procedimiento de construcción del instrumento debe inspirar la suficiente confianza como para que el autor en algún momento lo considere concluido como para ser sometido a una revisión o evaluación por parte de los especialistas en el área. Dos preguntas copan la escena en esta fase del trabajo, la primera de ellas es si el instrumento sirve para los fines que orientaron su construcción, si mide lo que pretendemos medir, es decir, si tiene validez. La segunda pregunta está referida a la confianza que podemos tener de los resultados que su aplicación arroja, esto es si se mide el rasgo de manera consistente, si podemos considerar que el instrumento es confiable, Magnusson (1972). La

validez y la confiabilidad constituyen las dos características deseables básicas de cualquier instrumento de medida en cualquier disciplina.

Para el cálculo de los índices asociados con la validez y confiabilidad nos valemos de tres fuentes de información, en primer lugar el instrumento en sí mismo y su historia, cómo fue construido, qué teorías lo soportan, cómo se diseñó el plan de construcción o la tabla de especificaciones, qué formación académica tienen los especialistas que trabajaron en estas fases, qué relación tienen estos con el rasgo que se pretende medir. En segundo lugar el comportamiento que podemos observar en el grupo al cual aplicamos el instrumento, esto es si sus respuestas son consistentes, si mantienen una conducta similar a lo largo de la aplicación, si su conducta ante instrumentos similares es la misma, o más distante aun, por ejemplo su ejecución en actividades académicas futuras como la referencia al desempeño académico posterior. En tercer lugar la variable que mide el instrumento, para establecer si teóricamente constituye un rasgo estable en el corto tiempo o por el contrario, si en el corto tiempo el esfuerzo y dedicación de los candidatos les permite mejorar su ejecución ante instrumentos similares, esta instancia es muy importante para la interpretación de los resultados que arrojen los índices.

El orden en que fueron expuestas no explica la importancia de la fuente de información, su ordenamiento solo se hace con fines didácticos y explicativos. En este trabajo se intenta ilustrar con ejemplos prácticos diferentes situaciones en las cuales debemos recurrir al cálculo de estos índices, indicando siempre la fuente de información que corresponda, esto permitirá conocer cuando la evaluación sobre el instrumento es directa y cuando nos valemos de informaciones de otras fuentes para la medición de estas características deseables.

Los niveles de dificultad y de discriminación

En relación con los instrumentos diseñados para evaluar los aprendizajes, una vez aplicados, los especialistas focalizan su atención en la calidad de las preguntas, en relación con ellas atienden dos aspectos: qué tan fáciles o difíciles resultan al grupo y si logran en forma satisfactoria discriminar entre los integrantes del grupo que poseen el rasgo desarrollado y los que no lo tienen, Backhoff y otros (2000). En el primer caso la proporción del grupo que responde correctamente a la pregunta, generalmente expresada en porcentaje, constituye su índice de facilidad y el complemento con respecto al cien, su índice de dificultad, el porcentaje de sujetos que no logra responder satisfactoriamente a la pregunta.

En el caso del índice de discriminación, se determina la relación entre responder correctamente o no a la pregunta y la puntuación total alcanzada en la prueba, mediante el uso de dos coeficientes de correlación, el biserial para el cual se supone que las

respuestas forman una distribución continua que dicotomizamos y el coeficiente punto biserial para el cual las respuestas emitidas son correctas o incorrectas, Nelson (2001), una correlación positiva indicara que responder correctamente está asociado con altas puntuaciones en la prueba y responder en forma incorrecta se asociara con la obtención de bajas puntuaciones, si estos son los resultados la pregunta discrimina entre los que tienen el rasgo desarrollado y los que no lo tienen, y se considera que evalúa el mismo rasgo que la prueba en general. Obsérvese que estas características se las asignamos a las preguntas a partir del análisis del comportamiento del grupo.

En relación con los niveles de dificultad, en las tablas de especificaciones o plan de construcción de las pruebas se indica para cada pregunta el nivel estimado que corresponde, estos índices informan al constructor las características de cada una y deberá redactar cada pregunta de manera que ella ofrezca el nivel de dificultad previsto, sin embargo, estos índices son estimados, ellos expresan que podemos esperar del grupo en función de la experiencia del evaluador; el verdadero índice de dificultad se obtiene una vez aplicado el instrumento al determinar la proporción de respuestas acertadas y erradas, estas nos indicarán cuan fácil o difícil resulta la pregunta para el grupo al cual aplicamos el instrumento , la comparación entre la estimación y el dato empírico solo permite afinar las estimaciones próximas en relación con grupos similares.

Los niveles de discriminación no forman parte de las tablas de especificaciones, en parte porque son muy difíciles de estimar, por experiencia se conoce que la tendencia es a obtener niveles de discriminación satisfactorios más en preguntas de dificultad media que en las consideradas muy fáciles o muy difíciles. Esto resulta lógico, una pregunta estimada como muy fácil está redactada para que sea respondida correctamente al menos por el 80% del grupo, para que el índice de discriminación proporcionara una información importante se requeriría que el 20% de los participantes, que son los que respondieron incorrectamente, obtuviera las puntuaciones más bajas posibles, obtener un índice de discriminación alto en estas circunstancias sigue siendo posible pero poco probable. En el caso de las preguntas muy difíciles la posibilidad es la misma, solo que resulta un poco más probable que en el caso de las preguntas muy fáciles. Aun así, la discriminación es un rasgo necesario e imprescindible de cualquier prueba, tanto que la prueba que no discrimina podríamos decir que no tiene utilidad en ningún ambiente. Si elaboramos una prueba para evaluar la comprensión de la lectura, los reportes deben indicar quienes tienen el rasgo desarrollado y quienes no, las puntuaciones obtenidas en esta prueba deben evidenciarlo, al igual que una prueba clínica para el despistaje de una enfermedad deberá indicar al final, si el paciente sufre o no de dicha enfermedad. En caso contrario, la prueba no resulta útil para el diagnóstico, y, tampoco para el paciente a quien se la aplicamos, que no sabe lo que tiene pero tampoco sabe lo que no tiene.

En la revisión bibliográfica hecha para la redacción de este artículo se observó que los autores se concentran en los índices de dificultad y de discriminación de las preguntas

cuando hablan de la calidad de las pruebas, Thorndike y Hagen (1970) Nunnally y Bernstein (1995) y muy pocas son las referencias a los índices de la prueba en general, al respecto formularemos algunas recomendaciones sobre los procedimientos que podemos utilizar para estimar estos índices.

La construcción de la prueba

Las pruebas de aplicación local o restringida, las pruebas de aula

Entre los instrumentos que conocemos como de lápiz y papel, se distinguen las pruebas que aplicamos en la aulas de clase, en una o varias secciones, en cualquier nivel educativo, de las que tienen aplicación a gran escala, esto es para utilizarla en varias instituciones, para una región o para todo el país, Navas (2013), las previsiones son distintas, comenzaremos con las primeras, las que aplica el maestro en el aula o las que aplicamos los profesores en educación media o educación superior. Estas pruebas de uso restringido para el grupo del aula generalmente se aplican por asignatura y evalúan un tema específico, el docente para elaborarlas deberá disponer al menos del programa de la asignatura, de los libros de texto recomendados y de los planes de clase que fueron desarrollados, esto le permitirá una visión completa de los objetivos que se tratan de lograr con el desarrollo del tema. El tema en cuestión deberá ser desglosado en tópicos o subtemas y contenidos, para cada uno de ellos deberá decidir cuantas preguntas formulara, o si una actividad propuesta abarca varios contenidos, generalmente se trata de pruebas cortas para responderlas en una o dos horas de clase por lo que las actividades propuestas son pocas, pudieran ser cinco. Los planes de clase y las lecturas recomendadas le suministran información sobre la complejidad recomendada para cada actividad. Una tabla sencilla en la que se vacié esta información es un buen soporte técnico en esta escala, puede contener el número de la pregunta o actividad, el tipo de pregunta, el nivel estimado de dificultad y la puntuación o peso dentro de la prueba, esto último dependerá de la escala de notas con la cual se trabaje. En la etapa de construcción de las preguntas o actividades, se debe ser cuidadoso con el lenguaje utilizado, se trata de garantizar que el mismo sea comprensible para los alumnos, la redacción de los planteamientos no debe ofrecer lugar a dudas, es necesario revisar las preguntas una vez formuladas, se deben tener presentes las estrategias utilizadas en el desarrollo de las clases, esto le ayudara a formular adecuadamente el planteamiento de los problemas.

Las pruebas de aplicación a gran escala

La elaboración de este tipo de prueba requiere la organización de un equipo de trabajo integrado por especialistas tanto en la asignatura o área de la cual se trate como de expertos en la construcción de pruebas, generalmente constituyen pruebas de alto impacto porque están relacionadas con la selección de candidatos a cupo o becas que ofrezca el Estado o la asignación de cargos en alguna empresa importante, ejemplifiquemos con una prueba de selección a la universidad. Debemos disponer de los programas oficiales de Educación Media y de los programas de las asignaturas de la carrera para la cual se seleccionan los candidatos, al menos los que correspondan al primer año de la carrera o los dos primeros semestres. Con ellos debemos armar una matriz que relacione los objetivos o temas de los programas de educación media (conjunto de partida) con los programas de educación superior (conjunto de llegada). La prueba deberá centrarse en las conexiones identificadas. Esta matriz deberá ser evaluada por el grupo de especialistas para ordenar las conexiones desde la más importante hasta la que lo es menos. En cuanto al tipo de preguntas generalmente se prefiere utilizar las de selección simple de varias alternativas (decisión del grupo), de la misma manera se debe acordar la extensión de la prueba, cuántas preguntas se formularan y de cuánto tiempo dispondrán los aspirantes para responder, son dos datos que permiten afinar la extensión. Con esta información y el conocimiento y adopción de una taxonomía de las actividades escolares, se construirá una Tabla de Especificaciones con al menos cuatro dimensiones: número de la pregunta, contenido o tema que evalúa, conducta esperada o nivel de actividad (taxonomía adoptada) y dificultad estimada. Si la prueba contendrá ítems de diferentes tipos será necesario incorporar una nueva dimensión para indicarlo. Se deben elaborar las normas para la aplicación por parte de los docentes y las de presentación por parte de los alumnos –instrucciones para responder la prueba-. Estas pruebas deben ser respondidas previamente por varios especialistas para asegurar su calidad, posteriormente deberá hacerse un muestreo piloto, seleccionar un grupo representativo de la población a la que se va a aplicar en forma definitiva para garantizar que se comprendan las instrucciones, afinar el tiempo de aplicación y obtener los estadísticos preliminares que permitan hacer los ajustes tanto en las preguntas y sus respuestas correctas como en los distractores.

En el caso de las pruebas que se aplican en el aula, que casi siempre incluyen preguntas de desarrollo, es necesario que el docente redacte respuestas modelo para cada pregunta o describa los elementos que deben estar presentes en las respuestas, para evitar la subjetividad en la apreciación de las respuestas emitidas y en la asignación de calificaciones.

Navas (2013), Profesora de la Universidad Nacional de Educación a Distancia de España, elaboró una lista de los objetivos propios de la evaluación en el aula y de la evaluación a gran escala:

EN EL AULA

a) Determinar si los estudiantes dominan un determinado concepto o habilidad.
b) Motivar a los estudiantes para que se involucren activamente en el Proceso de enseñanza-aprendizaje.
c) Conseguir que los estudiantes sean capaces de razonar y de aplicar los contenidos aprendidos.
d) Ayudar a desarrollar una actitud positiva hacia las asignaturas.
e) Informar a los padres de lo que saben y son capaces de hacer sus hijos.
f) Informar a los estudiantes de lo que saben y son capaces de hacer.
g) Informar a los estudiantes de lo que esperan de ellos.
h) Indicar a los estudiantes dónde han de centrarse para mejorar.
i) Elaborar el boletín de notas.
j) Evaluar la eficacia de los métodos pedagógicos.

A GRAN ESCALA

a) Identificar puntos fuertes y débiles de los estudiantes.
b) Determinar si los estudiantes cumplen los objetivos educativos establecidos.
c) Determinar cómo agrupar a los estudiantes.
d) Identificar estudiantes con necesidades especiales.
e) Comparar el rendimiento de determinados grupos de estudiantes con el promedio nacional.
f) Evaluar la eficacia de un nuevo curriculum.
g) Evaluar a profesores y directores de los centros.
h) Proporcionar información para la acreditación de los centros.
i) Comparar distintos centros escolares.
j) Distribuir recursos.

Esta lista ilustra la variedad de propósitos con los cuales pueden ser utilizados ambos tipos de pruebas.

Los conceptos básicos y los procedimientos de cálculo para los índices

Una vez probadas y corregidas las preguntas en relación con la claridad y precisión del lenguaje, su adaptación al nivel educativo que corresponde y el ordenamiento dentro de la prueba. Revisada y mejorada la versión experimental, la prueba es aplicada a la población para la cual fue diseñada, su corrección se hace al comparar las respuestas emitidas por cada participante con la tabla de respuestas correctas, pero antes de ofrecer los resultados alcanzados por los participantes es necesario evaluar el instrumento para estar seguros de la información que se va a suministrar sobre la ejecución de los participantes, hasta ahora con base al proceso cumplido en la construcción solo podemos asegurar su validez de contenido, que las preguntas evalúan lo mismo que la prueba en general, y que tanto las preguntas como las pruebas evalúan el rasgo para el que fueron diseñadas. Las estadísticas con base en las cuales se hace la evaluación del

funcionamiento de la prueba, corresponden a la participación de los usuarios, su ejecución ante el instrumento, y con ellas, le conferimos características al instrumento. Algunos especialistas hablan de evaluación post morten o post hoc dado que se realiza una vez aplicado el instrumento, Backhoff y otros (2000). Estos procedimientos están previstos en los cronogramas de aplicación de las pruebas de largo alcance, pero son igualmente recomendables para las de uso restringido.

La evaluación de las preguntas

Para la evaluación de las preguntas se consideran tres aspectos, el nivel de dificultad, el índice de discriminación y el análisis de los distractores, por razones de tiempo y espacio acá consideraremos los dos primeros dado que el análisis de los distractores debe realizarse uno a uno, una prueba de treinta preguntas tendría treinta respuestas correctas y ciento veinte distractores si cada pregunta ofrece cinco alternativas de respuesta.

El nivel de dificultad de la pregunta. Se refiere a cuan fácil o difícil resulta la pregunta al grupo que participa en la aplicación, a los aspirantes en nuestro caso. La mayoría de los textos indican que se determine el índice de facilidad **"p"** y que por complemento con respecto a la unidad o al cien según el caso se obtiene el índice de dificultad **"q"**, el procedimiento es el siguiente se cuenta el número de sujetos que responde correctamente a la pregunta y esto se divide entre el total del grupo, esta proporción de aciertos, índice de facilidad, se puede expresar como una proporción, en decimales, o en porcentaje, si la multiplicamos por cien, tal como se indica en la fórmula. Para obtener el índice de dificultad restamos de uno o de cien según el caso el índice de facilidad.

Cálculo del Índice de dificultad de las preguntas

$$\%Rc = \frac{NRc * 100}{Nmuestra} \quad \%Ri = \frac{NRi * 100}{Nmuestra}$$

$\%Rc$ = % de respuestas correctas. Índice de facilidad

$\%Ri$ = % de respuestas incorrectas. Índice de dificultad

$$Nmuestra = Total\ de\ personas\ que\ presento\ la\ prueba$$

$$\%Rc + \%Ri = 100 \quad \%Ri = 100 - \%Rc$$

La otra manera es contar todas las respuestas erradas y dividir su número entre el total de la muestra, el índice refleja la dificultad que les ofrece la pregunta a los participantes. La noción de transferencia de la conducta del grupo a las características de la prueba queda

evidenciada al suponer que de aplicarse esta versión del instrumento a individuos entrenados en el tema, a este nuevo grupo les resultara menos difícil cualquier pregunta contenida en el examen. Estos procedimientos pueden ser usados por el docente en el aula y también están considerados en los paquetes estadísticos para el análisis de ítems y de test (Lertap y Microitems) que son los recursos utilizados para corregir y evaluar las pruebas que se aplican a Gran Escala.

En relación con la interpretación de los resultados para cada pregunta existen varias tablas que sugieren cinco categorías para el nivel de dificultad: Muy difícil, Difícil, Promedio, Fácil y Muy Fácil. La diferencia entre ellas son los intervalos para las diferentes categorías, la más usada es la que propone un tamaño igual para cada categoría: MD de 0.80 a 1.00, D de 0.60 a 0.80, P de 0.40 a 0.60, F de 0.20 a 0.40 y MF de 0.00 a 0.20 elaborada por Crocker y Algina (1986) que permite la clasificación de los índices de dificultad tanto para las pruebas como para cada una de las preguntas.

Tabla I. Índices de Dificultad

Índices de Dificultad	**%***R correctas*	**%***R incorrectas*
Categorías	Rango de la Categoría	Rango de la Categoría
Muy Difícil	00……..20	80……100
Difícil	20……..40	60……..80
Promedio	40……..60	40……..60
Fácil	60……..80	20……..40
Muy Fácil	80……100	00……..20

El nivel de discriminación de la pregunta. El supuesto hecho con base en el procedimiento de construcción es que cada pregunta evalúa el mismo rasgo que evalúa la prueba, si lo damos por cierto entonces podemos suponer que responder correctamente a una pregunta en cuestión debe estar relacionado con obtener altos puntajes en la prueba y responder incorrectamente debe estar asociado a obtener bajos puntajes en la prueba total, si determinamos un valor de esa relación, si calculamos un índice de correlación, el resultado debe ser positivo con relación a la respuesta correcta y negativo con respecto a la respuestas incorrectas. La pregunta debe discriminar entre quienes de los participantes tienen el rasgo medido desarrollado y quiénes no (Tavella, 1978).

Qué puede hacer el docente de aula para conocer el poder discriminativo de las preguntas de sus pruebas: elaborar una lista de sus alumnos ordenada de mayor a menor por la calificación obtenida y dividirla en dos grupos de igual tamaño altos y bajos. Contar las respuestas correctas emitidas por cada grupo en la pregunta uno y restar del número de respuestas correctas del grupo alto la cantidad emitida por el grupo bajo y dividirla entre

el tamaño del grupo alto, esto le dará un índice de discriminación de la pregunta. Si ambos grupos tiene la misma cantidad de respuestas correctas el índice es cero y la pregunta no distingue entre los grupos que tienen desarrollado el rasgo del grupo que no lo tiene, si todos los del grupo alto responden correctamente y todos los del grupo bajo incorrectamente el índice será de 1.00. Con esta rutina puede diseñar una tabla en Excel para utilizarla con cualquier prueba que diseñe.

Procedimiento para determinar el Nivel de Discriminación de una pregunta en el aula

1.-Ordenar los alumnos de mayor a menor por la calificación obtenida en la prueba

2.-Dividirlos en grupos alto y bajo de igual tamaño

3.-Contar las respuestas correctas emitidas por cada grupo en la pregunta a evaluar y

4.-Aplicar la formula N.Dis= $\frac{\mathbf{NrcGA - NrcGB}}{\mathbf{NAGA}}$

Los programas de análisis de ítems y de test, que se utilizan en las pruebas de gran escala calculan el coeficiente de correlación punto biserial, comparando la media aritmética de las puntuaciones en el test total para los alumnos que respondieron correctamente le pregunta y le resta la media aritmética de los que respondieron incorrectamente, esta diferencia se divide entre la desviación típica del grupo total y se pondera por la raíz cuadrada del producto de **p** por **q**, que son la proporción de sujetos que respondió correctamente a la pregunta y la proporción de sujetos que respondió incorrectamente, respectivamente. Se espera que la correlación con la respuesta correcta resulte positiva y con las respuestas incorrectas negativa.

Coeficiente de Correlación Punto Biserial Rpb

Rpb= $\frac{\mu e - \mu f}{\sigma}\sqrt{p * q}$

Rpb*= Coeficiente de Correlación Punto Biserial*

μe = *Promedio de calificaciones de los sujetos asociados al éxito*

μf = *Promedio de calificaciones de los sujetos asociados al fracaso*

σ = *Desviación típica del grupo total*

p = *Proporción de sujetos asociados al éxito*

q = *Proporción de sujetos asociados al fracaso*

En relación a la interpretación del resultado obtenido, para el caso de la correlación punto biserial hay varias tablas en la que se sugieren distintos intervalos, una de las más usadas es la propuesta por Ebel y Frisbie (1986), en ella se indica que los índices iguales o superiores a 0.30 se consideran satisfactorios, de 0.20 a 0.30 regulares y por debajo de 0.20 deficientes. Sin embargo la recomendación general es la de revisar y corregir toda pregunta que tenga un índice de discriminación por debajo de 0.30 dado que la etapa de la construcción de preguntas resulta la más laboriosa de todas las descritas en la elaboración de una prueba.

Tabla II. Criterios clasificatorios de los índices de discriminación

Rango de Discriminación N.Dis	*Calidad o Categoría*	*Recomendación Técnica*
N.Dis ≥ 0.39	*Excelente*	*Conservar Resguardar*
De 0.38 a ≥ 0.30	*Bueno*	*Hay posibilidad de mejorar*
De 0.29 a ≥ 0.20	*Regular*	*Necesidad de revisar*
De 0.19 a ≥ 0.00	*Pobre*	*Revisar a profundidad*
N.Dis $\leq$ *-0.01*	*Pésimo*	*Descartar definitivamente*

Los paquetes estadísticos para el análisis de ítems y de test (Lertap y Microitems) traen la opción de utilizar un índice de discriminación distinto al descrito, se basa en la consideración de los grupos extremos del 27% superior e inferior de los alumnos en relación con la puntuación general alcanzada en la prueba, para cada pregunta se cuentan las respuestas correctas emitidas en cada grupo, se restan al número del grupo superior el número del grupo inferior y la diferencia se divide entre el total de cualquiera de los dos grupos.

La Evaluación de las Pruebas

Para la evaluación de las pruebas se debe estimar un nivel de dificultad y un índice de discriminación general, para ello podemos calcular la media aritmética de todos los índices calculados para las preguntas, dado que si mantenemos la misma estructura o formula en los cálculos las características estarían siendo evaluadas con base al mismo tamaño de muestra, todos procederían de la misma base. Este promedio sería una estimación bien de la dificultad de la prueba o del poder discriminativo, según el caso. Estos procedimientos están al alcance de los docentes en el aula. En el caso de las pruebas de gran escala los paquetes estadísticos nos ofrecen esos cálculos. Algunos programas traen la opción de calcular el nivel de dificultad en relación con las respuestas emitidas, no incorporan las omitidas, lo cual hace que la base de cálculo varíe de pregunta a pregunta,

en esos casos el promedio aconsejable desde el punto de vista estadístico es la media geométrica G que es la raíz enésima del producto de los **n** índices de dificultad o de discriminación calculados para cada pregunta.

Procedimientos para definir el Índice de dificultad y de discriminación para las pruebas en general

Como se indicó en el párrafo anterior, cuando la base de cálculo de los índices de dificultad de las preguntas es la misma, el procedimiento seria determinar la Media Aritmética de los índices de las preguntas

$$\%RcPt = \frac{\sum \%Rc.1 + \cdots \ldots + \%Rc.n}{Numero\ de\ Preguntas}$$

Cuando la base de cálculo de los índices de dificultad de las preguntas es variable, entonces determinamos la Media Geométrica de los índices de las preguntas

$$\%RcPt = \sqrt[n]{(\%Rc.1 * \cdots \ldots * \%Rc.n)}$$

Procedimientos similares pueden seguirse para calcular los índices de discriminación para las pruebas en su totalidad.

Esta información corresponde al promedio de dificultad y de discriminación de las preguntas de la prueba, la prueba en su extensión mantiene esas características, para la interpretación de los resultados pueden utilizarse las mismas tablas que se usan para la clasificación de las preguntas. Otro procedimiento sencillo que podría utilizar el docente en el aula y que podría informarnos del nivel de dificultad de una prueba, consiste en calcular el promedio de las calificaciones alcanzadas por los alumnos en la aplicación y ubicar ese valor en la escala de notas utilizadas, la cual previamente dividimos en cinco clases para categorizar los valores desde Muy difícil hasta Muy fácil. Este índice podría resultar más fácil de interpretar, hablaría con mayor propiedad sobre los resultados de la prueba en cuanto a su dificultad porque incorpora la escala de calificaciones y el comportamiento del grupo ante el instrumento.

Estos dos aspectos resultan de importancia capital para decidir sobre la calidad de las pruebas de gran escala. No se acostumbra ofrecer información sobre estos aspectos para las pruebas de aplicación local o restringida, sin embargo, los cuidados en la definición de los objetivos de la prueba y en la construcción de las preguntas le confieren un razonable criterio de validez tanto a la prueba como a las preguntas que la integran y por los índices de discriminación sabremos si ambas evalúan el mismo rasgo. Para evaluar la confiabilidad el procedimiento más sencillo consiste en dividir la prueba en mitades, índice de Spearman-Brown, (Nunnally y Bernstein, 1995), la partición que frecuentemente se usa es

ítems impares y pares, se calcula la correlación entre los puntajes obtenidos en ambas partes y el resultado del índice debe ser alto para indicar que los alumnos tienen un comportamiento similar en todo el recorrido de la prueba, que la calidad de su participación es similar en ambas partes, así obtenemos un indicador de consistencia interna del instrumento.

La confiabilidad medida por consistencia interna. No hay procedimientos sencillos para el cálculo de la confiabilidad de la prueba por consistencia interna, el cálculo de los índices utilizados Cronbach y Hoyt son bastante complejos lo cual hace imprescindible el recurso computacional, ambos evalúan la consistencia interna de la prueba al determinar si los participantes logran un comportamiento uniforme a través de todo el recorrido o trayecto de la prueba, estos procedimientos están incluidos en los paquetes estadísticos que permiten la evaluación de ítems y de Test, por ejemplo, Lertap, Microitems). El índice ALPHA de Cronbach constituye un resumen de las correlaciones calculadas entre todas las particiones posibles del test en mitades, la estimación de la varianza verdadera permite calcular una razón con respecto a la varianza total y es expresión de la precisión de la medida.

El índice de Hoyt constituye el resultado de un análisis de varianza que contrasta la variabilidad en los ítems con la variabilidad de los participantes, teóricamente considera todas las particiones posibles del instrumento, en ambos casos el límite superior del rango de resultados del índice es uno (1) y a medida que el resultado se aproxime a él, la calidad de la medición será mayor, la confianza en los resultados aumenta en la medida que el resultado tienda a uno. Prieto y Muñiz (2000) definieron una clasificación de los resultados para los resultados de estos índices: excelente para cualquier valor mayor o igual a 0.85, bueno para resultados entre 0.80 y 0.85, adecuado para valores entre 0.70 y 0.80, aceptable entre 0.60 y 0.70 e inadecuado para cualquier valor inferior a 0.60.

En la búsqueda de información que nos permita establecer índices sobre la confianza que debemos tener en las mediciones que se hacen con base en pruebas, se han diseñado procedimientos o técnicas como el test retest, o la aplicación de pruebas paralelas, en ambos casos el cálculo de la correlación de Pearson nos proporciona un coeficiente sobre la estabilidad de la medida. La aplicación de estas técnicas requiere una cuidadosa planificación en la cual se garantice una selección adecuada de los tiempos que median entre ambas aplicaciones, evaluar la interferencia que pueda estar asociada a la memoria, demostrar el paralelismo de las pruebas o su equivalencia, determinar la disponibilidad de los recursos logísticos y económicos, humanos y materiales que garanticen la doble aplicación y evaluar si el rasgo medido es modificable o no en el corto tiempo, recordando la atención a los principios éticos que deben acompañar todos los actos evaluativos tanto psicológicos como educativos, estas circunstancias refuerzan la apreciación sobre la complejidad de la medición de la confiabilidad de los instrumentos de medición a gran escala.

La Validez establecida por criterio externo

Cuando se aplican instrumentos de medición con fines de selección, la evaluación de las preguntas y las pruebas no termina en forma inmediata, se debe esperar por los resultados numéricos obtenidos por los candidatos seleccionados para determinar la asociación entre las puntuaciones de la prueba y el criterio que se pretende predecir. En el caso de la selección a la Universidad el criterio seria el record académico de los estudiantes al final del primer año, el promedio de las calificaciones definitivas obtenidas, calcular la correlación entre ambas variables constituiría un coeficiente de validez predictiva de la prueba. Se esperan resultados positivos y moderados para afirmar que la prueba constituye un buen predictor del éxito académico en la universidad, recuérdese que los estudios de correlación se basan en la variabilidad conjunta de las variables y la selección nos ofrece una muestra sesgada de los resultados de la prueba, solo se consideran los mejores puntajes lo cual homogeneíza los puntajes de la muestra y esta pérdida de variabilidad afecta el valor del coeficiente y los resultados tienden a ser bajos en vez de altos. Desde el punto de vista teórico la manera de demostrar, o establecer la validez predictiva de los instrumentos de selección, consiste en evaluar las condiciones de entrada mediante las pruebas, permitir el ingreso de todos los aspirantes y al finalizar el primer año realizar los cálculos correspondientes, esto permitiría que los resultados reflejen de mejor manera el poder predictivo de las pruebas pero resulta impráctico ya que las posibilidades de ingresar toda la demanda son siempre inexistentes, es muy poco probable que se disponga de una capacidad instalada que lo permita.

Las pruebas de selección en la mayoría de los casos son conjuntos de pruebas que evalúan diferentes rasgos asociados teóricamente con las funciones u oficios a realizar en el futuro inmediato por lo que resulta conveniente determinar la forma en que cada una de las subpruebas explican el comportamiento de los alumnos en la variable criterio, en cuanto explican las subpruebas la variabilidad de la variable criterio, en nuestro caso el record académico. Esto se calcula mediante el uso de la Regresión Múltiple, declarando como variables independientes los resultados de las pruebas de selección y como variable dependiente el record académico del primer año de la carrera, el recurso inmediato es el Paquete Estadístico para las Ciencias Sociales SPSS por sus siglas en inglés, Pérez (2005).

Otros procedimientos que arrojan resultados sobre la validez predictiva es clasificar los alumnos en dos bloques con respecto a la variable criterio, record académico, altos y bajos o bien el 27% de más alto nivel académico y el 27% más bajo y calcular el índice de discriminación de la prueba, numero de respuestas correctas emitidas en relación con el grupo del criterio a predecir. Al establecer los grupos Altos y Bajos con respecto a la variable criterio y no a las puntuaciones en la prueba, la información que nos proporciona se refiere a la validez predictiva. Puede igualmente calcularse el coeficiente de correlación

punto biserial entre las puntuaciones en la prueba del grupo seleccionado y la dicotomía en el record académico aprobado-aplazado, ambos cálculos nos ofrecerían un coeficiente de validez de la prueba en su totalidad.

Evaluación de una prueba de Razonamiento Numérico-Lógico utilizada con fines de selección

En esta sección se aplicaran los índices expuestos en las páginas anteriores para la evaluación de la prueba de Matemática y Razonamiento Lógico, aplicada a los aspirantes a ingresar en carreras de Ciencias de la Salud durante el año 2011. La selección de los candidatos se llevó a cabo con el uso de siete pruebas, a saber: Química, Biología, Razonamiento Numérico - Lógico, Habilidad Verbal, Aptitud Espacial, Inventario de preferencias hacia el área de la Salud y el Inventario de Tendencias Conductuales, EDACI (2011).

Para la evaluación de la prueba y el análisis de las preguntas que la integran se utilizó el paquete estadístico Lertap, Nelson (2001). El número total de pruebas corregidas fue de 9.271, la extensión de la prueba es de 30 ítems de selección simple de cuatro alternativas de las cuales una es la respuesta correcta, se ofrecen los resultados del análisis de la prueba para los 392 candidatos seleccionados. Se seleccionó el 4.23% de los candidatos, la proporción es de un seleccionado por cada veinticinco candidatos. En la corrección se adjudica un punto por cada respuesta correcta, no se penalizan las respuestas incorrectas ni las preguntas no respondidas, dejadas en blanco, estas no modifican la puntuación que se alcance. No se utilizó la corrección por azar, esta rutina es optativa dentro del paquete estadístico.

Tabla III. Distribución de asignación según aplicación de pruebas

Conceptos	Numero	Porcentaje
Demanda	9271	100
Asignados	329	4.23
Preguntas	30	100
Alternativas	4	100

Nivel de dificultad de las preguntas y la prueba

Para cada pregunta se calcula la proporción de respuestas correctas en relación con el total de la muestra, esta información constituye el índice de facilidad, el complemento con respecto a uno o a cien, según el caso, es el índice de dificultad. Como la base sobre la cual se hacen los cálculos de las proporciones o porcentajes es el tamaño de la muestra el promedio que se calcula para determinar el índice de dificultad de la prueba es la media

aritmética. Según la tabla clasificatoria de los niveles de dificultad utilizada por Crocker y Algina (1986), de 5 categorías con intervalo de 20% cada una, 3 preguntas resultaron muy difíciles, menos del 20% de los candidatos respondió correctamente lo cual representa el 10% de la prueba. 13 preguntas resultaron difíciles lo cual equivale al 43.33% de la prueba, estas preguntas fueron respondidas correctamente por menos del 40% de los aspirantes. Más de la mitad de la prueba el 53.33% resultó difícil o muy difícil para los aspirantes asignados. 10 preguntas resultaron de dificultad promedio lo cual representa una tercera parte de la prueba y solamente cuatro preguntas resultaron fáciles para los admitidos, ninguna pregunta les resulto muy fácil. El promedio de estos índices de facilidad es de 40.17, la dificultad de la prueba en general, es de 59.83%, resultando el instrumento de promedio a difícil.

Tabla IV. Nivel de dificultad de las preguntas

CATEGORIAS	Nº de Preguntas	Porcentaje
Muy Difícil	3	10
Difícil	13	43.33
Promedio	10	33.33
Fácil	4	13.33
Muy Fácil	0	0

Nivel de discriminación de las preguntas y la prueba

Se calcula mediante el uso del coeficiente de correlación punto biserial, que compara el promedio de calificaciones alcanzadas por los que responden correctamente con el promedio de los que responden de manera incorrecta en cada pregunta, esta diferencia se divide entre la desviación típica y la razón se pondera por la raíz cuadrada del producto de las proporciones de acierto y fracaso en cada pregunta. Para obtener el índice de discriminación de la prueba en general se calcula la media aritmética de los indicas calculados para cada pregunta dado que todos fueron calculados en relación con la misma base. De acuerdo con la tabla de clasificación de índices de discriminación ideada por Ponce y Granel (2005), de cuatro categorías con diferentes intervalos, 19 de las 30 preguntas tienen índices muy altos de discriminación, mayores a 0.35, 8 índices altos entre 0.25 y 0.35, 2 ítems son de discriminación media, entre 0.2 y 0.25 y 1 es de baja discriminación, valores menores a 0.2. De esta manera el índice general de la prueba, el promedio aritmético, alcanza a 0.386 por lo que se tipifica el instrumento como de alta discriminación, la tendencia que se describe es que responder bien a las preguntas está asociado con la obtención de altos puntajes en la prueba, las preguntas evalúan el mismo rasgo que el instrumento en general.

Tabla V. Nivel de discriminación de las preguntas

CATEGORIAS	Nº de Preguntas	Porcentaje
Muy Alto	19	63.33
Alto	8	26.67
Medio	2	6.67
Bajo	1	3.33

La confiabilidad del instrumento medida por consistencia interna

El índice de Hoyt que evalúa la consistencia interna del instrumento es de 0.81 el cual de acuerdo con la tabla de clasificación propuesta por Prieto y Muñiz (2000) se considera buena, esto indica que la medición realizada tiene una buena precisión, se puede confiar en los resultados alcanzados en la medición del rasgo. El error de estimación es de 2.32 que corresponde aproximadamente a un 8% de la puntuación máxima posible que es de 30 puntos.

Las estadísticas básicas

Los índices de tendencia central, posición, dispersión y forma nos permiten definir la distribución de los puntajes alcanzados por el grupo, a través de ellos podemos hacer algunas consideraciones en relación con el comportamiento del grupo ante el instrumento. Estos índices se relacionan con el rango de las calificaciones posibles y en algunos casos se evalúan por contraste con las características del modelo de distribución más frecuentemente utilizado en el campo educativo la Distribución Normal. La media aritmética fue de 12.05 puntos, este es el punto de la escala donde la distribución hace equilibrio, la mediana resulto ser de 10 puntos que marca el límite del primer tercio de la escala, el de los puntajes más bajos, esta medida indica que la mitad del grupo obtuvo calificaciones por debajo de ese límite y la otra mitad valores superiores a y el modo o moda resulto ser la puntuación 9, este es el valor que más se observa el que tiene la mayor frecuencia. Todos estos valores se interpretan en la escala de la cual provienen en este caso de cero a treinta puntos. Obsérvese que los tres valores están en torno al límite superior del primer tercio de la escala lo cual indica que la prueba tiende a resultar difícil al grupo de alumnos. Regresemos a la media aritmética esta es 12.05 de 30 puntos posibles, al expresarla en porcentaje, esto es dividirla sobre su base y multiplicarla por cien obtenemos 40.17%, los logros del grupo corresponden a ese porcentaje, lo que el grupo no logro fue el complemento 59.83%, esta es otra manera de establecer los índices de facilidad y de dificultad del instrumento, los porcentajes coinciden exactamente con los promedios calculados en la sección de niveles de dificultad porque cada respuesta correcta acredita un punto en la calificación, este es un procedimiento que puede utilizar

un docente en su aula para determinar cuan fáciles o difíciles resultan las evaluaciones a su grupo de alumnos. La desviación típica es de 5.421, este es el promedio de la distancia a que se encuentran los valores en relación con su media aritmética, debiera tender a 5 que es un sexto del rango de calificaciones que es de treinta para que la distribución de calificaciones tuviera un recorrido de seis desviaciones típicas que es aproximadamente el recorrido de la distribución normal. El coeficiente de variación que nos indica que porcentaje de la media aritmética es la desviación típica es de 44.99%, esto es así porque la media aritmética se considera baja en relación con las expectativas es de 12.05 y de ser el punto central del recorrido debió ser 15, se estima que el valor de este coeficiente debe estar entre 30 y 35%.

Tabla VI. Estadísticas básicas utilizadas en el análisis

ÍNDICES	Resultados	Porcentajes	ÍNDICES	Resultados	Porcentajes
MEDIA	12.05	40.17	DESV.TIPICA	5.421	18.07
MEDIANA	11	36.67	VARIANZA	29.3872	
MODA	9	30	COEF.VARIA		44.99
Xi MAYOR	29	96.67	ASIMETRIA	0.672	
Xi MENOR	2	6.67	CURTOSIS	-0.28	
RANGO	27	90	LOGRADO		40.17
CENTIL 25	5	16.67	N0 LOGRADO		59.83
CENTIL75	10	33.33	ERROR ESTIMA.	2.32	8

En cuanto a los valores de posición, el menor valor observado es de 2 y el mayor 29 lo cual hace que el recorrido contenga 29 valores posibles. El cálculo de los cuartiles indica que el 25% de los alumnos obtuvo calificaciones por debajo de los 8 puntos, el siguiente 25% se encuentre entre las calificaciones 8 y 10 puntos. Por debajo de los 10 puntos, que representa el primer tercio de la escala, se encuentra la mitad de los seleccionados, el valor del cuartil tres es de 16 puntos que es casi la mitad de la escala, y el 25% superior de los alumnos seleccionados se distribuye entre los 16 los 29 puntos que fue la máxima nota alcanzada. Estas medidas nos dicen que hay una concentración de alumnos hacia los valores menores de la escala lo cual indica que la prueba resulto más difícil de lo esperado. El índice de asimetría Alpha sub tres que se calcula con base a los desvíos de los valores en relación con la media es positivo 0.677, lo cual indica que hay una alta concentración hacia los valores bajos y la distribución esta coleada hacia los valores altos, esta característica no es mala en relación con el uso de los resultados con fines de selección ya que esta se inicia por los valores altos que indican una mejor ejecución, complicaría la selección si la cola de la distribución esta hacia los volares bajos porque la concentración se observaría hacia los valores altos que es por donde comenzamos la selección.

Finalmente se evalúa la altura relativa que alcanza la distribución, el índice Alpha sub cuatro, el resultado fue de -0.208 lo cual indica que la distribución es más baja de lo normal en cuyo caso el índice seria cero, se clasifica como platicúrtica. Hay mayor dispersión en el conjunto de valores de la que se considera normal, tal como apuntábamos en párrafos anteriores.

A manera de conclusión, los comentarios finales

Los aspectos a evaluar para garantizar la calidad de las pruebas son los niveles de dificultad y discriminación de las pruebas y sus preguntas. Es necesario garantizar el rigor metodológico en la consultas de los materiales y documentos que sirven de base a la planificación y diseño de ambos tipos de prueba. Se deben reportar los índices de validez y confiabilidad del instrumento, estos últimos resultan indispensables en las pruebas de largo alcance. Para las pruebas de aula se recomiendan procesos sencillos que proporcionan información importante al docente sobre el comportamiento de sus alumnos ante las pruebas.

Todo docente de aula cuando se acercan los lapsos de evaluación revisa los textos utilizados, piensa en la extensión de la prueba a aplicar, imagina los niveles de dificultad de las preguntas, ordena los tópicos o temas en función de su importancia, la invitación a través de este artículo no es a trabajar más, sino a hacerlo mejor, vaciar todo lo que piensa y decide en una Tabla de Especificaciones que sirva de soporte a la prueba que aplica, esto convierte un cuestionario en una prueba, sin ella solo tenemos un conjunto de preguntas en un papel. Acostumbrémonos a utilizar los recursos técnicos que posteriormente van a facilitarnos el trabajo. Si en cada lapso diseñamos la prueba, la planificamos y vaciamos sus características en una Tabla de Especificaciones, si conservamos las estadísticas de estos modelos de prueba, en los años venideros podemos planificar el perfeccionamiento de esas pruebas. No se trata de aplicar siempre el mismo instrumento sino de evaluar con instrumentos equivalentes, las preguntas deben ser distintas. En la medida de sus posibilidades pueden incorporar los cálculos más sencillos para obtener la información básica que les permita evaluar las pruebas con las que examinan a sus alumnos.

Para los docentes responsables de las pruebas de largo alcance, es indispensable garantizar la calidad de los instrumentos dado el uso que tienen los resultados y la intervención social que significa el conjunto de propósitos con el cual se aplican los instrumentos. El problema se inicia porque no podemos asumir responsabilidades individuales, estas pruebas requieren del trabajo en equipos multidisciplinarios para ofrecer garantías sobre la calidad de los instrumentos. Uno de los elementos de máxima importancia es que estas pruebas requieren generar en la población que atienden un alto nivel de confianza, esto significa la realización de esfuerzos adicionales para atender

debidamente los reclamos que planteen los participantes, asegurar que quienes reclaman salgan convencidos de la transparencia del proceso y de la pulcritud en el manejo de la información. Una sana practica es la publicación de los planes de construcción de las pruebas para que los usuarios se informen previamente sobre los tópicos en que van a ser evaluados, de ser posible indicar los procedimientos a utilizar en la corrección y la calificación en cada prueba.

En diferentes escalas las pruebas deben permitir la creación de un banco de preguntas, a medida que se avanza en el uso de las diversas versiones este banco se irá ampliando lo cual facilita el trabajo a futuro, la idea no es incrementar el trabajo la idea es hacerlo bien o mejor de lo que lo hacemos para facilitar el trabajo subsiguiente y disponer de suficiente información que nos permita garantizar una adecuada calidad de los instrumentos. El uso de estas buenas prácticas nos hará cada vez mejores evaluadores.

En relación con la prueba de razonamiento numérico y lógico con la cual se ejemplifico la interpretación de los índices para evaluar la calidad de la misma, presentada por 9.271 aspirantes a ingresar en la UCV, área de Ciencias de la Salud, facultades de Medicina, Odontología y Farmacia, en el año 2011. Los resultados para el grupo seleccionado indican que la prueba de treinta itemes puede ser usada con fines selectivos, ofrece un nivel de dificultad alto 59.83%. El poder discriminativo resulta muy alto para la prueba total, el índice es 0.39%. La consistencia interna evaluada por el índice Hoyt fue de 0.81 indicativo de una buena precisión en la medida. Considerando los aspectos analizados se recomienda su uso como instrumento de selección, porque tiene un buen nivel de dificultad moderada a difícil, su discriminación como instrumento es muy alta Y la precisión de la medida es alta por lo que ofrece confianza en la medición del rasgo. Es una prueba de buena calidad.

Referencias

Backhoff, E., Larrazolo, N. y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa*, 2 (1). www.redalyc.org/pdf/155/15502102.pdf. Consultado el 07 de enero, 2017 en: http://redie.uabc.mx/vol2no1/cont.-backhoff.html

Crocker, L., y Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing.

Ebel, R.L. y Frisbie, D.A. (1986) *Basic concepts in items and test analysis*. Consultado el 08 de diciembre de 2016. https://www.google.co.ve/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=Ebel+y+Frisbie+ (1986).

EDACI 2011. UCV. Facultades de Medicina, Odontología y Farmacia, Archivo digital de Hojas de Respuestas.

González, J. (2009). Asignación de aspirantes a través del Sistema Nacional de Ingreso a la Educación Superior (SINIES), UCV 2008. *Revista de Pedagogía*. vol.31, n.88 pp. 39 – 60.

Magnusson, D. (1972). *Teoría de los Test*. México: Trillas.

Navas, M.J. (2013) La medición en el ámbito educativo. Universidad Nacional de Educación a Distancia. España. *Revista digital del Colegio Oficial de Psicología de Madrid*. Psicología Educativa. Volumen 18. Número 13, pp 15-28. Año 2013.

Nelson, L.R. (2001). *Item Analysis for Tests and Surveys. Using Lertap 5.* Perth: Curtin University of Technology.

Nunnally, J. e I.J. Bernstein (1995). *Teoría Psicométrica*. México: McGraw-Hill.

Pérez, C. (2005). *Técnicas Estadísticas con SPSS 12*. Madrid: Pearson Prentice Hall.

Ponce, M. y P. Granel (2005). *Reporte Técnico de la Prueba Interna de la Facultad de Humanidades*. Mimeografiado. UCV.

Prieto, G. y Muñiz, J. (2000) Un modelo para evaluar la calidad de los test utilizados en España. *Revista Papeles del Psicólogo* N° 77. pp. 65-71.

Sarco Lira, A. (2010) El Ingreso Asistido. Calidad en la selección/Equidad en el ingreso. *Docencia Universitaria*. Vol. 12 N°2. Año 2011. UCV. SADPRO. Caracas.

Tavella, N. (1978). *Análisis de los Ítems en la construcción de Instrumentos Psicométricos*. México: Trillas.

Thorndike, R y Hagen, E. (1970). *Test y técnicas de Medición en Psicología y Educación*. México: Trillas.