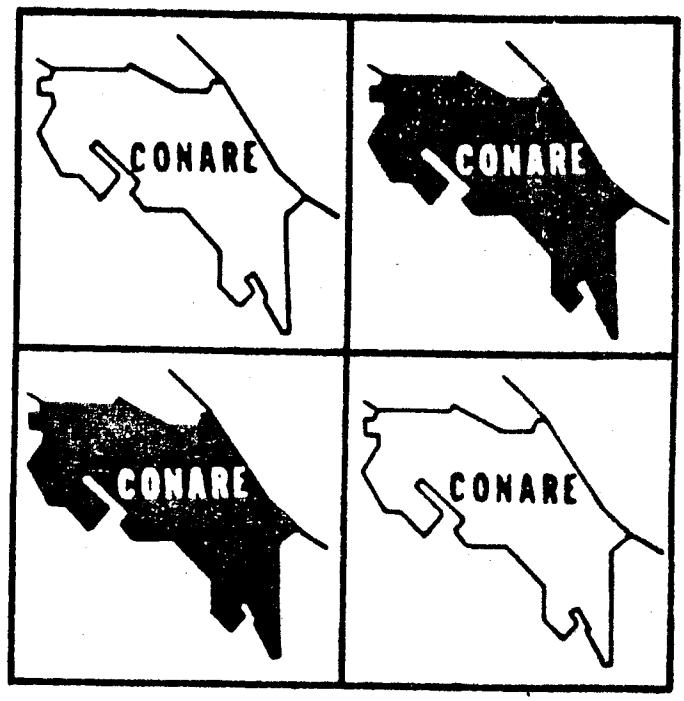
CONSEJO NACIONAL DE RECTORES
OFICINA DE PLANIFICACION DE LA EDUCACION SUPERIOR



INFORME FINAL DEL PROYECTO
PRUEBAS EXPERIMENTALES DE CONOCIMIENTOS

ING. MAYRA ALVARADO U.

DR. JUAN ML. ESQUIVEL A.

Junio. 1987

INFORME FINAL DEL PROYECTO PRUEBAS EXPERIMENTALES DE CONOCIMIENTOS

INDICE

./.

./.

INDICE DE CUADROS

CAPITULO IV: ANALISIS DE RESULTADOS

./.

PAGINA

./.

./.

INDICE DE ANEXOS

CAPITULO I

Introducción

El problema que se pretende abordar con el presente proyecto es el de la inadecuada preparación intelectual o académica del estudiante que accede a la Educación Superior.

Es un tema casi diario entre profesores universitarios y se ha convertido en la explicación más socorrida del ciudadano común, cuando se habla de la preparación, nivel de conocimientos, éxitos o fracasos educativos.

¿Se puede hacer algo, desde la Educación Superior, que no sea recibir - los estudiantes como vienen y tratar después de hacer frente a sus necesidades?.

Cualquier tipo de respuesta que se ensaye, ya sea en términos de admisión ya sea en términos instructivos, deberá apoyarse en un conocimiento preciso y sistemático del real nivel cognoscitivo de la población estudiantil.

Para obtenerlo, las pruebas de conocimientos referidas a criterios ofrecen una de las alternativas más prometedoras al permitir describir el estado actual de los conocimientos de un individuo.

Con ello, se podrá favorecer la eficiencia interna del Sistema de Educación Superior, al diagnosticar el dominio de conocimientos fundamentales en las áreas básicas (Matemática, Español, Ciencias y Estudios Sociales); al informar a las diferentes instancias interesadas y al estudiante mismo qué

./.

conceptos y habilidades domina o necesita aprender para satisfacer las exigencias universitarias.

Para hacerle frente al problema expuesto el Consejo Nacional de Rectores, por medio de la Comisión Interinstitucional de Estudios relacionados con la Admisión a la Educación Superior, se trazó la tarea de desarrollar pruebas de conocimientos, dentro del marco conceptual de la medición con referencia a criterios. Estas tienen carácter experimental y se desarrollaron para las siguientes asignaturas: Matemática, Español, Ciencias y Estudios Sociales. Estas pruebas pretenden diagnosticar los conocimientos que los alumnos de nuevo ingreso a cada universidad traen de su educación secundaria.

Para el desarrollo de este trabajo en primera instancia se hizo una revisión de la literatura existente en el país sobre la medición con referencia a criterios y se dieron los lineamientos metodológicos, para el desarrollo de las pruebas, que mejor se adaptaran a las condiciones del país, o sea, tomando en cuenta las limitaciones de recursos humanos, técnicos y financieros que se tendrían en cualquier institución de educación superior costarricense.

Objetivo

El objetivo general de este proyecto es la construcción y validación de pruebas de conocimientos referidas a criterios, que permitan diagnosticar el dominio de conocimientos mínimos, en las áreas básicas (Matemática, Español, Ciencias y Estudios Sociales), que tienen los estudiantes al egresar de la Educación Diversificada.

./.

Problemas de investigación

1.- ¿Se puede afirmar que la prueba construida es una prueba con referencia a criterios?

2.- ¿Cómo se escogió y especificó el dominio de conocimientos que incluiría la prueba, para que existiera evidencia de su validez?

3.- ¿Existe exactitud en la clasificación de los estudiantes como "másters" y "no másters" en un objetivo particular?

4.- ¿Existe consistencia en la clasificación de los estudiantes como "mas - ters" y "no másters" en un objetivo particular"

5.- ¿Los ítemes incluidos en la prueba, miden eficientemente el objetivo - para el cual fueron escritos?

Definiciones operacionales de las variables

Educación Diversificada: Bloque educativo constituído por dos o tres años (según la modalidad) posterior a la Educación General Básica que a su vez está constituida por 3 ciclos de 3 años cada uno.

Alumnos o estudiantes: Sujetos matriculados en las instituciones de Educación Superior Estatales (excepto la UNED) y en décimo año de la Educación Diversificada, en el curso lectivo de 1986.

./.

Sub-prueba: Conjunto de 5 ítemes que miden un mismo objetivo.

Prueba: Conjunto de 12 sub-pruebas (cada una correspondiente a un objetivo) que miden una asignatura.

Pruebas paralelas: Pruebas en las que la escogencia de los ítemes para cada sub-prueba se ha realizado mediante un muestreo aleatorio y estratificado, a partir de un conjunto de ítemes (de 15 a 20) que miden el objetivo.

Fórmula: Se refiere a cada una de las tres pruebas paralelas de una misma asignatura, se utilizaron las letras mayúsculas A, B y C para designarlas.

Número de aplicación: Se refiere al momento en que el estudiante contestó una prueba en particular, se utiliza solo en el caso de que el estudiante haya constestado dos pruebas en una misma sesión. La primera prueba contestada se distingue con el número 1 y la segunda con el número 2.

Nacionalidad: Se refiere al país de origen del estudiante que contestó la prueba. Todos los estudiantes NO COSTARRICENSES fueron catalogados como EXTRANJEROS.

Sexo: Se refiere a la condición MASCULINA o FEMENINA del sujeto que contestó la prueba.

Edad: Se refiere a la edad, en años cumplidos que tenía el estudiante, cuando constestó la prueba.

Año de egreso del colegio: Se refiere al año en el cual, el estudiante que contestó la prueba, finalizó la Educación Diversificada.

./.

Institución en que está matriculado: Se refiere a la institución de Educación Superior Estatal en la que está matriculado el estudiante (ITCR, UCR o UNA), que contestó la prueba.

Sede o Recinto: Se refiere al centro regional o sede de la institución de Educación Superior Estatal en el cual, el estudiante está matriculado.

Modalidad: Se refiere a los tipos de institución de la Educación Diversificada y el tercer ciclo de la Educación General Básica. En este estudio se consideran las siguientes categorías:

- Colegios académicos diurnos oficiales
- Colegios académicos diurnos particulares y semi-oficiales
- Colegios académicos nocturnos oficiales
- Colegios académicos nocturnos particulares y semi-oficiales
- Colegio técnico oficial
- Colegio técnico particualar y semi-oficial
- Bachillerato por madurez

Prueba de Ciencias: Prueba que evaluará los conocimientos de las asignaturas de Física, Química y Biología que se imparten en la Educación Diversificada.

Limitaciones

1.- A pesar de que se enfatizó en la búsqueda de profesionales con buena experiencia en la labor a desarrollar, la mayoría de los que se contrataron para la escritura de los ítemes tenían muy poca.

./.

2.- La calidad de los ítemes de algunas de las asignaturas no fue óptima y hubo necesidad de utilizarlos en esas condiciones.

3.- Por razones de tiempo y de los recursos disponibles, no se pudo revisar exhaustivamente, antes del análisis de congruencia, si todos los ítemes habían sido construidos siguiendo las especificaciones dadas en los objetivos amplificados.

4.- Lo ajustado del cronograma establecido y especialmente la falta de experiencia en la metodología empleada, nueva en el medio, impidieron que las pruebas pudieran ser aplicadas a los estudiantes de undécimo año de la Educación Diversificada, en vez de esto fueron aplicadas a estudiantes de primer ingreso de las universidades estatales.

5.- Los grupos criterio usados para algunos análisis (grupo instruido y no instruido) difieren en algunas características básicas.

6.- No se pudo utilizar ningún método de análisis basado en correlaciones, por carecer de los recursos computacionales del caso.

7.- Los recursos computacionales aportados para dar apoyo al proyecto no fueron los óptimos. Además de carecer de la tecnología apropiada, el procesamiento fue muy lento por la falta de disponibilidad de recursos humanos.

8.- El sistema de dar recursos, tanto humanos como financieros, al proyecto mediante los aportes de cada una de las universidades no fue eficiente. El tiempo y esfuerzo dedicado a "perseguir" dichos aportes fue muy alto y restó dedicación a los aspectos técnicos.

CAPITULO II

Revisión de la Literatura

Pruebas con referencia a criterios: definición utilidad y diferencia con el modelo de pruebas con referencia a normas

Definición

Desde que apareció el famoso artículo de Glaser en 1963, que introdujo el concepto de medición con referencia a criterios se ha dado una gran discusión sobre el significado de este concepto. Uno de los mayores aspectos de confusión se debió a la palabra "criterio", ya que para muchos individuos significaba estándar de ejecución, nivel de destreza o puntaje de corte. De esta manera, muchos autores han etiquetado sus "tests" como pruebas con referencia a criterios únicamente porque tenían un puntaje de corte por encima del cual, el rendimiento del estudiante se podía considerar adecuado (Popham, 1978a).

Popham y Husek (1969) aclararon la definición al establecer lo que llamaron la distinción básica entre medición con referencia a criterios y medición con referencia a normas. Definieron la primera como "...aquellas (medicio - nes) que se usan para establecer la condición de un individuo con respecto a algún criterio" (p. 2) contrastándola con la definición de medición con referencia a normas como "...aquellas que se emplean para establecer el rendimiento de un individuo con relación al rendimiento de otros individuos" (p. 2).

./.

Una definición más amplia es la ofrecida por Glaser y Nitko (1971): "Una prueba con referencia a criterios es aquella que se construye deliberadamente para que de ella resulten mediciones que sean directamente interpretables en términos de estándares de ejecución especificados previamente" - (p. 653). Continúan los autores ejemplarizando el término estándares de ejecución, al establecer que "son generalmente especificados al definir una clase de dominio de tareas que deban ser ejecutadas por un individuo" (p. 653).

Otra definición reiteradamente citada en la literatura es la ofrecida por Popham (1975): "Una prueba con referencia a criterios se usa para establecer la condición de un individuo con respecto a un dominio de conductas bien definido" (p. 130). Este concepto es parecido al ofrecido por Millman (1974) cuando define:

> "Para muchos objetivos instruccionales es posible describir, con un - alto grado de especificidad, el contenido de la población de ítemes - de la cual, los ítemes que aparecen en la prueba se han seleccionado al azar o de manera estratificada aleatoriamente. Una prueba así formada se llama prueba con referencia a un dominio" (p. 313-314).

Por lo tanto, si se acepta las definiciones de Popham y Millman no hay diferencias esenciales entre pruebas con referencia a criterios y con referencia a dominio. Es conveniente aclarar que existen otros conceptos también - empleados para distinguir pruebas con referencia a criterios, estos son pruebas con referencia a objetivos y pruebas de maestría (Mastery test). La primera se emplea para definir pruebas en que los ítemes están apareados con objetivos (Hambleton et al, 1978). El segundo término, comúnmente se refiere a pruebas con referencia a criterios, usados en programas educativos en que se prac

./.

tican modelos de enseñanza individualizados o de enseñanza para el dominio - (Mastery Learning Block, 1971).

La distinción primaria entre pruebas con referencia a criterios (o domi - nio) y pruebas con referencia a objetivos según Hambleton et al (1978) es la siguiente:

> "En una prueba con referencia a criterios, los ítemes son un grupo representativo de ítemes de un dominio de conductas claramente definidas que miden un objetivo, mientras que con una prueba con referencia a objetivos no se especifica el dominio de conductas y los ítemes no se consideran como representativos de ningún dominio de conductas" (p. 3)

Los autores van aún más allá, al establecer que las pruebas que se pueden comprar en el mercado en su mayoría deben correctamente catalogarse como - pruebas con referencia a objetivos.

Para finalizar este examen del concepto de pruebas con referencia a criterios, es importante señalar los cuatro supuestos que fundamentan la medición con referencia a criterios según Dilendik (1976):

> "El primer supuesto es que el propósito del maestro es originar, en - tantos estudiantes como sea posible, tanto aprendizaje como sea posible... El segundo, es que el éxito académico y la excelencia educativa no son mutuamente excluyentes... El tercer supuesto es que los - maestros conocen lo que quieren lograr... El cuarto supuesto y en mi experiencia, el más importante, es que los puntajes de la prueba sirven como un recurso de diagnóstico para el maestro y como un recurso de aprendizaje para los alumnos." (p. 92-93).

Origen

Las pruebas que genéricamente podemos llamar con referencia a criterios, se han desarrollado como respuesta a necesidades debidas a los nuevos pro -

./.

gramas educativos de enseñanza individualizada y al énfasis puesto en pedir al maestro ser responsable del aprendizaje del alumno (accuntability); en - contraposición con las pruebas con referencia a normas cuyas bases teóricas fueron desarrolladas para responder al movimiento psicológico de la medición de aspectos mentales del ser humano. Asimismo se puede establecer que las - raíces de la medición con referencia a criterios se encuentran en el paradig ma de diseño instruccional de Gagne y White (1978) que a su vez se fundamenta en el paradigma dado por Tyler (1949).

Diferencias entre los modelos de pruebas con referencia a normas y con referencia a criterios

Con respecto a las diferencias entre ambos modelos de medición, Greco (1974) establece las diferencias en cuanto a lo que los dos tipos de pruebas "hacen" y a "cómo se emplean" (p. 23) con base en los trabajos de Glasser (1963) y - Popham y Husek (1969) anteriormente expuestos. El concluye "que la distin - ción entre las dos, en la práctica, no está tanto en cómo se construyen, sino en la especificidad o estrechez del dominio que cubre la prueba y la forma completa como se determinan los niveles de rendimiento de ese dominio" - (p. 25).

Por su parte, Shepard (1979) establece que la característica esencial que distingue a una prueba con referencia a criterios de una referencia a normas, es la precisión con que se especifica el dominio de contenido y con que se - desarrollan ítemes para reflejar ese contenido.

./.

Otra diferencia básica, según Millman (1980), es la interpretación de los puntajes de una prueba. En aquellas con referencia a criterios los puntajes tienen sentido absoluto, mientras en las pruebas con referencia a normas, el significado deriva de las comparaciones con otros puntajes.

Se han enfatizado diferencias esenciales entre estos dos tipos de medición aunque se podría señalar otras, en aspectos técnicos, las cúales serán explicadas en el desarrollo de este trabajo de revisión bibliográfica.

Conviene examinar otra distinción que se hace entre pruebas de rendimiento académico que por ser general y por establecer un esquema de análisis resulta de gran utilidad. La distinción, la establecen Wardrop et al (1982) - entre pruebas que se usan para diferenciar y las que se emplean para medir. Las primeras son aquellas que se usan "para tomar decisiones de selección - cuando el acceso es limitado" (p. 2) y en las segundas el énfasis es en la - valoración absoluta, "...para diagnosticar bondades y debilidades y para seguir el progreso" (del estudiante) (p. 2).

Con el propósito de analizar si una prueba diferencia o mide, Wardrop et al (1982) definen cuatro características: generación de ítemes, revisión de pruebas e ítemes, valoración de la precisión y validación. En cada característica, establecen puntos en un continuo que va desde diferenciación hasta medición. De tal forma que al analizar cualquier prueba, se puede caracterizarla en cada aspecto y determinar el perfil de la misma.

./.

Usos

En cuanto al empleo de las pruebas con referencia a criterios, parece haber consenso entre diversos autores, que el propósito principal es el de diagnosticar y tomar decisiones sobre:

a) individuos (si alcanzan maestría o no en un dominio de conductas) y

b) tratamientos (eficacia de programas educativos).

En tanto que las pruebas con referencia a normas, tienen como fines primordiales seleccionar y servir como pruebas descriptoras amplias del rendimiento académico de individuos en extensas áreas de contenido (Popham 1975, 1978a, 1978b; Hambleton et al,1978; Sephard 1979; Greco 1974; Popham y Husek 1969; Block, 1971).

Resumen

Se ha hecho una revisión de las diversas definiciones de las pruebas con referencia a criterios, los aspectos que distinguen de otro tipo de pruebas y las diferentes clasificaciones que se pueden hacer de ellas. Se puede concluir en primer término, que las pruebas con referencia a criterios se distinguen de las pruebas normativas esencialmente en:

a) la definición específica y clara del dominio de conductas y

b) la interpretación de los puntajes de la prueba

Asimismo que la especificidad y estrechez con que se define el dominio de conductas distingue las pruebas con referencia a dominio de las de referencia a objetivos. También, se concluye que sus usos básicos son dos:

./.

a) diagnóstico (estimación del puntaje de dominio del estudiante) (Hambleton et al, 1978) y

b) toma de decisiones sobre individuos y tratamientos (asignación de los estudiantes a estados o categorías de maestría) (Hambleton et al, 1978).

Validez

La validez de una prueba se ha definido, tradicionalmente, como aquella característica por la cual la prueba mide lo que debe medir o cumple la función para la cual fue creada. Esta definición puede, muy bien, aplicarse a instrumentos con referencia a criterios.

Validez según Popham (1978):

Comúnmente la validez se ha estudiado de acuerdo con tres paradigmas diferentes: Validez de contenido, validez relacionada con el criterio y validez conceptual o de constructo. Según Popham (1978a) la validez de las pruebas con referencia a criterios se debe analizar desde tres puntos de vista: descriptiva, funcional y de selección del dominio.

a) Validez descriptiva. Se pretende con esta verificar hasta dónde la prueba mide el esquema descriptivo que se pretende medir. Para hacer tal verificación, es necesario establecer primero, si el esquema descriptivo (especificaciones de prueba, objetivos o forma de ítemes, etc.) logra comunicar a aquellas personas que interpretarán lo que significa el rendimiento del examinando en la prueba y a las que escribirán ítemes, las reglas para crear ítemes de tal manera que sean congruentes con ese mismo esquema descriptivo.

./.

Esto se logra de acuerdo con Popham (1978a) empleando personas que escriban ítemes por el esquema presentado y jueces que definan la homogeneidad de esos ítemes con respecto al esquema. Una vez que se ha verificado si el esquema cumple la función de comunicar eficazmente, se debe asegurar que los ítemes de la prueba ya construida sean congruentes con el esquema descriptivo. Nuevamente se sugiere el uso de jueces para establecer la congruencia entre los ítemes y el esquema. Se podría señalar el paralelismo entre la validez descriptiva y la validez de contenido tal y como se define tradicionalmente.

b) Validez funcional. Se define como la exactitud con la que la prueba con referencia a criterios satisface el propósito para el que se emplea. Esta clase de validez se podrá evidenciar al analizar detenidamente la función que se quiere dar a la prueba contrastándola con el esquema descriptivo de la prueba. Dependiendo del uso que se quiera dar a la prueba este tipo de validez puede ser o no importante, a veces es suficiente con saber que pueden o no pueden hacer los estudiantes y la validez descriptiva provee este tipo de información, sin embargo, es posible que se quiera medir predicción por ejemplo y debe evaluarse si la prueba es válida (validez funcional) para ello.

c) Validez de selección del dominio. Para obtener evidencia de esta clase de validez se debe responder a la pregunta: ¿Cómo se puede comprobar el buen juicio con que el dominio de conductas de la prueba se escogió? Como se es-

./.

tableció anteriormente el dominio de conductas se define en el esquema descriptivo de la prueba. Una forma de responder a la pregunta, es describiendo claramente qué características tenían las personas que seleccionaron el dominio y qué procedimientos y guías se les dieron para que realizaran dicha selección. Aunque no se esté interesado en establecer la existencia de algún constructo hipotético, que es la función de la validez conceptual tradicionalmente entendida, se puede notar que existen similitudes entre las técnicas que se emplean en evidenciar la validez de la selección del dominio y las de la validez conceptual.

Validez según Hambleton (1980):

En contraste con la posición con respecto a la validez analizada en los párrafos anteriores, Hambleton (1980) señala que las consideraciones de la validez de los puntajes de una prueba con referencia a criterios surge de tres etapas:

a) la selección de los objetivos (o competencias)
b) la medición de los objetivos incluidos en la prueba con referencia a criterios y
c) el uso de los puntajes de la prueba (p. 81).

Establece dos tipos de validez: de contenido y de constructo.

a) Validez de contenido. Dos elementos señala Hambleton (1980) como necesarios para establecer la validez de contenido, en primer lugar una descripción acertada del dominio de conocimientos que se pretende medir con la prue

./.

ba. En segundo lugar la consideración de tres características de los ítemes: a) validez, b) calidad técnica y c) representatividad.

La descripción detallada del dominio se puede hacer de varias formas: objetivos, objetivos amplificados, forma de ítemes y especificaciones del dominio entre otros.

La validez de los ítemes de la prueba se determina al establecer cuán bien reflejan los ítemes, los dominios de los que se derivan, en términos de su contenido. Existen dos formas de establecer esta validez: una involucra el juicio de especialistas, estos juicios analizan el grado de pareamiento entre los ítemes y el dominio que pretenden medir. El otro método consiste en la aplicación de las técnicas empíricas empleadas comúnmente en el desarrollo de preguntas para pruebas con referencia a normas. Con respecto a cuál de las dos formas, brevemente aquí aplicadas, es más conveniente, Hambleton (1980) afirma:

> "... creo que el primer método tiene más mérito... Hay al menos cuatro problemas involucrados en el empleo de procedimientos empíricos. Primero: la mayoría, sino todos, los procedimientos dependen de las características del grupo de examinandos y de los efectos de la instrucción, segundo, comúnmente se requieren técnicas sofisticadas y programas de computadoras que no están a disposición de los prácticos. Tercero, cuando se usan estadísticas de ítemes para seleccionar preguntas para una prueba con referencia a criterios, el que desarrolla el test corre el riesgo de obtener un grupo de ítemes no representativo de los dominios que miden los objetivos incluidos en la prueba. Finalmente, los métodos empíricos en muchas ocasiones requieren datos de pretest y de postest en los mismos ítemes" (p.87)

Sin embargo, Rovinelli y Hambleton (1977) encuentran que los métodos empíricos tienen utilidad si el constructor está interesado en identificar -

./.

ítemes que tienen defectos y así mejorarlos. Otro problema probable al emplear técnicas empíricas es la posible falta de variabilidad de los puntajes, lo que produciría ítemes con bajos índices de discrimación. Sin embargo numerosos investigadores reportan suficiente variabilidad para permitir tales cálculos.

Tres técnicas para la recolección y análisis, de los juicios de los ítemes dados por los especialistas, son examinados por Hambleton (1980). La primera se fundamente en valorar cada ítem en una escala de tres puntos (+1, -1 y 0) según mida, no mida el objetivo o el juez esté inseguro. Estos valores sirven para establecer un índice de congruencia ítem-dominio. La segunda emplea una escala de calificaciones, en la que la tarea consiste en juzgar la calidad con que la pregunta mide el dominio. En la tercera técnica, a los jueces especialistas se les da dos listas, una con preguntas y la otra con la especificación del dominio y se les pide que pareen preguntas con dominio (objetivos).

La segunda característica citada por Hambleton (1980) es la calidad técnica de los ítemes. Con este propósito se deben revisar, empleando listas de cotejo, las características de formato y redacción de las diferentes clases de ítemes.

La representatividad de los ítemes se establece después de la selección de ellos para la versión final de la prueba. Nuevamente es necesario emplear el juicio de especialistas para responder a la pregunta: ¿Son éstos ítemes representativos del dominio establecido en las especificaciones de la prueba (u objetivos)?

./.

b) Validez de constructo (conceptual). Es, en la prueba con referencia a criterios, validez de las descripciones y las decisiones que se hacen con los puntajes de las pruebas (Hambleton, 1980). La utilidad de esos puntajes para describir los niveles de rendimiento de un estudiante y así tomar decisiones debe determinarse haciendo investigaciones de la validez de constructo de la prueba. Entre las técnicas que se pueden emplear en estas investigaciones se señalan: el análisis de escalogramas de Guttman, análisis factorial, estudios de validez predictiva, estudios experimentales y análisis de las posibles fuentes de invalidez.

Es muy importante ofrecer la opinión de otros estudiosos con respecto al concepto de la validez de constructo. Linn (1979) escribe "Preguntas de validez son preguntas de la validez de la interpretación de la medida... Es por lo tanto, la interpretación más que la medida lo que se valida. Los resultados de las mediciones tienen muchas interpretaciones que difieren en su grado de validez y en el tipo de evidencia que se requiere en el proceso de validación" (p. 109). Por su parte Messick (1975) al apuntar la importancia de la validez de la interpretación de los puntajes en contraposición con el peso que se le ha dado a la validez de contenido dice: "El mayor problema... es que la validez de contenido...se enfoca sobre formas de test en vez de los puntajes del mismo, sobre instrumentos en vez de mediciones" (p. 960-961).

Otros aspectos y enfoques de la validez.

Por su parte Crambert (1977) hace una revisión interesante del problema de

./.

validez de las pruebas con referencia a criterios y llega a las siguientes - conclusiones:

a) La validez de contenido es el resultado del cumplimiento estricto del procedimiento por el que se establece el dominio de la prueba. Otras técnicas podrían agregar evidencias sobre la validez y un mejor entendimiento de lo - que se mide.

b) El juicio subjetivo es inherente al proceso de establecer la validez de - contenido.

c) "Debido a la importancia del contenido "per se" en darle sentido al rendimiento en estas pruebas, hay gran énfasis en la fuerza de la relación que - pueda ser inferida entre el contenido de la prueba y los objetivos o especificaciones subyacentes al test" (p. 7).

d) El cuidado en la construcción de los ítemes y la rigurosidad en el proceso de juicio de los especialistas son dos diferencias con los procesos tradicionales de obtener evidencia de la validez de contenido.

e) Empíricamente la validez de contenido podría obtenerse correlacionando el resultado de un ítem con los resultados de aquellos que miden un mismo objetivo. Estas correlaciones deben ser mayores que las de este ítem con otros que miden otros objetivos, tal como lo sugieren Klein y Kosecoff (1973).

f) El cálculo de un índice de validez predictiva puede carecer de interés debido a los propósitos de estas pruebas y por la posible falta de variabilidad.

./.

Benson (1981) examina el concepto de validez de contenido estableciéndolo como la especificación formal del universo de tareas que una prueba intenta medir, pero manifiesta que es muy importante considerar la estructura misma de la prueba. Los elementos estructurales considerados cruciales son: es - critura, formato y nivel de lectura de los ítemes, así como las indicaciones de la prueba.

Benson y Crocker (1979) estudiaron la influencia del formato y nivel de - lectura de los ítemes en el rendimiento de jóvenes de noveno y décimo año ma triculados en un curso de ciencias de la salud. El formato y el nivel de - lectura sí influenciaron en el rendimiento; de esta manera los jóvenes tuvie ron mejor rendimiento en los ítemes de pareo que en los de falso o verdadero o de selección.

Las limitaciones que existen cuando se usan objetivos u otras formas de - definir las especificaciones de contenido en las pruebas con referencia a - criterios es analizada por Zwarts (1982). Establece que en ningún caso se - hace una verdadera unión entre la instrucción y las pruebas y por lo tanto - no se puede garantizar la validez de contenido, opina que se debe consultar a los especialistas en curriculum y mejorar la forma de muestreo de los íte- mes para mejorar el establecimiento de la validez de contenido. Asimismo, - analiza la necesidad de obtener evidencia de la validez de constructo, ya que más importante que la "perfomance" en la muestra de tareas de la prueba, son las capacidades subyacentes en esa "perfomance". Investigar la relación en- tre la "perfomance" y esas capacidades es encontrar la validez de constructo.

./.

Sobre la validez de decisión que es una forma de validez de constructo, - Lang (1982) indica que debe medirse a través de la validez de los ítemes y - de la interpretación de los resultados (congruencia, puntaje de corte).

Rose (1984) introduce otro concepto de validez, la validez instruccional que define como la propiedad con que los profesores enseñan las habilidades que se miden en la prueba. Considera que debe ser parte del desarrollo de una prueba con referencia a criterios, dar evidencia o demostrar el - grado en que ésta mide lo que los profesores han enseñado en el aula, pues - no se puede asumir que los objetivos sean enseñados uniformemente en todas - las aulas. Las ventajas que menciona de hacerlo así son las siguientes:

1) Cuando se descubre que no existe concordancia entre los objetivos y lo - que se enseña en el aula cabe tomar alguna de las siguientes opciones: eliminar alguno de los objetivos o dejar los objetivos y averiguar porqué los - profesores no los enseñan, para darles la asistencia necesaria.

2) El entender la relación existente entre los ítemes y el proceso de instrucción relacionado con esos ítemes aumenta la posibilidad de comprender mejor los análisis psicométricos.

3) Ayuda a reestablecer a los profesores como parte del proceso de valoración.

Rose da además el método empleado para medir la validez instruccional en un estudio realizado con estudiantes de noveno año, para medir los conoci - mientos de matemática general.

./.

Finalmente cabe manifestar la observación hecha por diversos autores en el sentido que no existe suficiente investigación y consecuentemente, suficiente literatura en el campo de la validez de las pruebas con referencia a criterios (Hambleton, 1980, Hambleton and Novick, 1973, Popham, 1978a).

Resumen

Se puede resumir la literatura revisada sobre la validez de las pruebas con referencia a criterios estableciendo que:

a) el aspecto distintivo se encuentra en la importancia primordial que tiene el establecimiento de la validez de contenido (o descriptiva) de la prueba, ésta se lleva a cabo mediante el empleo del juicio de especialistas que juzguen la congruencia de las preguntas con las especificaciones u objetivos, su representatividad y su calidad técnica,

b) también es importante señalar que el proceso de validar el dominio de conocimientos es básico para mejorar la interpretación de los puntajes producidos por estas pruebas. Otro aspecto a considerar es la preocupación de que las pruebas se utilicen con el propósito para el que fueron creadas,

c) asimismo, la obtención de evidencia de la validez de constructo, la menos investigada, debe ser considerada en el desarrollo de pruebas con referencia a criterios, pues esta evidencia mejorará la interpretación de los puntajes,

d) otros enfoques señalan la importancia de obtener evidencia de la validez instruccional de las pruebas y de considerar como elemento importante de jui

./.

cio de validez, la estructura misma de la prueba.

Confiabilidad

Si al estudiar la validez de las pruebas con referencia a criterios se encuentra el investigador con un número reducido de publicaciones sobre este aspecto, lo contrario le sucede al adentrarse en el estudio de la confiabilidad, pues la literatura además de muy abundante, es muy heterogénea. Debido a este hecho, el estudioso se ve obligado a hacer una selección que pueda ofrecerle un cuadro general del estado de desarrollo de este aspecto.

El primer artículo en que se discute la confiabilidad es de Popham y Husek (1969). Los autores establecen que los procedimientos empleados clásicamente para la determinación de la consistencia interna y la estabilidad no son apropiados en el estudio de pruebas con referencia a criterios, debido, principalmente, a la ausencia de variabilidad que puede darse en la distribución de puntajes de la aplicación de las pruebas con referencia a criterios. Ellos no ofrecen alternativa concreta alguna. La sugerencia de Haladyna (1974) es que para que el evaluador pueda artificialmente tener variabilidad de los puntajes, debe unir los puntajes de las pruebas dadas, antes y después de la instrucción, y esto hace posible la aplicación de las formas clásicas de calcular confiabilidad. También Hambleton y Novick (1973) se refieren al empleo de aplicaciones de la teoría clásica manifestando que la interpretación de estos índices debe hacerse con cuidado o descartarse su uso del todo, basándose en la consideración de que la correlación representa un escogimien

./.

to inapropiado como técnica estadística.

Otra de las primeras contribuciones al estudio de la confiabilidad fue ofrecida por Livingston (1972a), quien propuso una fórmula derivada de la teoría clásica fundamentada en que, el propósito de las pruebas con referencia a criterios es el de discriminar el puntaje de dominio (1) de cada examinando del puntaje de corte ("cut-off score"). Por lo tanto se definieron, las variaciones, que en la teoría clásica son acerca de la media, como variaciones del puntaje del dominio ("domain scores"), acerca del puntaje de corte. Este artículo de Livingston provocó una reacción en varios autores y sus consecuentes respuestas de Livingston (1972b, 1972c). Las limitaciones señaladas por Harris (1972), Hambleton y Novick (1973), Hambleton (1974), Shalvelson, Block y Ravitck (1972) pueden resumirse así:

a) El error estándar de medición es el mismo si se aplica la fórmula de Livingston o las fórmulas clásicas.

./.

---

(1) Hambleton, et al (1978) afirman acerca del puntaje de dominio: -- "El problema básico, dado el puntaje de un estudiante en un conjunto de ítemes que miden un objetivo, es estimar el puntaje proporcional acertado por el estudiante como si se le hubiesen administrado todos los posibles ítemes que miden el objetivo. Ese puntaje "estimado" se conoce como puntaje de dominio, puntaje de nivel de funcionamiento o verdadero puntaje proporcional acertado. Un examinando tiene un puntaje de dominio definido para cada uno de los objetivos medidos por la prueba con referencia a criterios" (p.5). Existen varios métodos para "estimar" (cálculo aproximado) el puntaje de dominio de un estudiante.

b) La propuesta en que se fundamenta Livingston para el cálculo de la confiabilidad de las pruebas con referencia a criterios, no es tan importante como el hecho de asignar al examinando, al mismo lado del puntaje de corte después de la aplicación de pruebas paralelas o de una prueba dos veces; por lo consiguiente para Hambleton y Novick (1973) este índice tiene poca utilidad. - Los resultados empíricos obtenidos por Hambleton (1974) dan apoyo a la posición anterior.

c) La tercera limitación está en el reporte de un índice de confiabilidad para el puntaje total de la prueba, cuando existen ítemes relacionados con objetivos diferentes. Shalvelson, et al (1972), establecen por primera vez, - que el valor de la confiabilidad debe darse para cada subprueba, constituida por un número "n" de ítemes midiendo un solo objetivo.

No todos los autores estuvieron de acuerdo con estas limitaciones, Brennan y Kane (1977) por ejemplo, trabajaron en derivar una medida de la confiabilidad a partir del concepto de Livingston que se señaló como la segunda limitación.

Según varios autores (Hambleton et al (1978), Brennan (1980), Subkoviac - (1980), Schaefer y Gross (1983) la confiabilidad de las pruebas con referencia a criterios puede analizarse desde tres puntos de vista:

a) Confiabilidad de las estimaciones de los puntajes de dominio: consistencia del puntaje de un estudiante si se repite la aplicación de una misma -

./.

prueba, sin hacer referencia a un puntaje de corte particular.

b) Confiabilidad de las decisiones de la clasificación de maestría o dominio: consistencia en la clasificación de los estudiantes como "masters" o como "no masters" en la aplicación repetida de una misma prueba.

c) Confiabilidad de los puntajes de pruebas con referencia a criterios: estabilidad de las desviaciones de los puntajes de los estudiantes con respecto al puntaje de corte, si se repite la aplicación de la prueba al mismo grupo.

Confiabilidad de las "estimaciones" de los puntajes de dominio.

Varios métodos se pueden emplear para obtener un valor de la confiabilidad de las "estimaciones" de los puntajes de dominio. Uno de ellos es el uso del error estándar de medición. Como se ha señalado en la literatura (Lord and Novick, 1968) el error estándar puede calcularse siempre que existan formas paralelas de una prueba, en el sentido clásico del término y este es muy útil en la interpretación de puntajes resultantes de la aplicación de pruebas con referencia a normas o criterios. Frecuentemente en el desarrollo de estas últimas su construcción se lleva a cabo seleccionando ítemes al azar de un grupo de ítemes relacionados con un objetivo, estas pruebas se les denomina pruebas paralelas aleatorias o nominales. Dada esta característica entonces Crombach, et al (1972) desarrollan formas de cálculo del error estándar de medición basándose en la teoría de generabilidad ("generalizability theory") que libera y extiende los conceptos de la teoría clásica en este

./.

campo. Más adelante Brennan y Kane (1977, 1978) y Brennan (1980) amplían el trabajo de Crombach, et al (1972) y desarrollan un índice de confiabilidad, que es independiente del puntaje de corte y que puede ser usado para "esti - mar" la precisión de los puntajes de dominio individuales, basándose siempre en la teoría de la generabilidad y en una definición de error particular.

Otra forma de determinar la confiabilidad de la estimación de los puntajes de dominio, según los trabajos de Millman (1974) y Hambleton, Swaminathan y Algina (1976) consiste en hacer uso del error estándar de estimación derivado del modelo binomial de pruebas ("binomial test model"). Este error constituye la desviación estándar de los errores de medición para un examinando que tiene un puntaje de dominio "x" a través de administraciones de "n" muestras de - ítemes obtenidas al azar de un grupo de ítemes. Según Hambleton, et al (1978) esta forma del error estándar tiene ventajas sobre el error estándar de medi ción porque "es menos conservador... y el efecto de la longitud de la prueba en la precisión de los estimados pueden ser estudiadas más fácilmente... es relativamente más fácil de calcular" (p. 19.)

Confiabilidad de las decisiones de la clasificación de maestría

Enfoque de la confiabilidad que hace énfasis en la consistencia con que los individuos son clasificados como "masters" o "no masters" a través de - una prueba aplicada en forma repetida.

Los primeros métodos fueron propuestos por Carver (1970). El primer proce dimiento requiere la administración de una misma prueba a dos grupos simila

./.

res y la comparación del porcentaje de examinandos que fueron clasificados - por encima del puntaje de corte. En el segundo método, a un mismo grupo se le aplican dos pruebas paralelas y se compara el porcentaje de alumnos que - se clasifican por encima del puntaje de corte ("masters"). Así en ambos métodos, cuanto más similares sean los porcentajes, más confiable es la prueba. A diferencia de la metodología tradicional, usada para determinar la confiabilidad, que se fundamenta en la reproducción de los puntajes individuales, los procedimientos de Carver se basan en la reproductividad de las distribuciones de puntajes. Esto es una limitación ya que una prueba puede ser no - confiable y producir iguales porcentajes de masters. Los estudiantes clasificados como masters en la primera aplicación podrían no ser los mismos que los de la segunda. Según Hambleton, et al (1978) "proverá (su procedimiento) sólo la forma más débil de evidencia de confiabilidad con referencia a criterios; esto es, sus condiciones son necesarias, pero no suficientes para establecer la confiabilidad de las pruebas" (p. 20-21).

Subsecuente a la propuesta de Carver Hambleton and Novick (1973) sugieren un método más sensitivo a la consistencia de las clasificaciones individuales y es que la proporción de individuos consistentemente clasificados como "masters" y "no masters" en dos pruebas (una misma prueba repetida o dos - pruebas paralelas) puede ser utilizada como índice de confiabilidad; para - ello proponen un índice P_o que es la proporción antes mencionada y que se - puede describir también como la proporción de decisiones observadas que están "en acuerdo", de ahí que varios autores lo denominen índice de acuerdo.

./.

Aunque P_o es de fácil cálculo, Swaminathan, Hambleton y Algina (1974) manifiestan que tiene una limitación, P_o no toma en cuenta la proporción de clasificaciones correctas que ocurren al azar y consecuentemente, sobrevalora la consistencia de las decisiones. Ellos sugieren que se emplee el índice Kappa de Cohen (1960) como medida de confiabilidad:

$$K = (P_o - P_c) / (1 - P_c)$$

donde:

$$P_o = \sum_{K=1}^{m} P_{kk} \; ; \quad P_c = \sum_{k=1}^{m} P_{.k} P_{k.}$$

P_{kk} es la proporción de examinandos clasificados en el mismo estado de dominio K en las dos administraciones; $P_{.k}$ y $P_{k.}$ representan la proporción de examinandos asignados al estado de dominio en la primera y segunda administraciones respectivamente.

Es importante señalar que el coeficiente Kappa (K) es afectado por:

- El puntaje de corte: K es menor para puntajes de corte en los extremos (muy altos o muy bajos)

- La heterogeneidad de los puntajes: a mayor variabilidad mayor es el valor de K.

- Longitud de la prueba o número de ítemes: a mayor número de ítemes, mayor es el valor de k; esta relación no es directamente proporcional, pues para valores muy grandes de n (# de ítemes) el aumento de k es menor.

./.

El coeficiente Kappa varía de 0 a 1 inclusive. El valor menor se da cuando la información de la prueba no contribuye a la exactitud del proceso de decisión de maestría. Cuando los puntajes no muestran suficiente variabilidad, el valor de α_{21} (Kuder-Richardson) puede ser cero o negativo. Si es negativo, el valor de K también lo será, en cuyo caso debe reemplazarse con el valor más pequeño positivo que se haya estimado para la confiabilidad; Huynh (1978).

El coeficiente K es muchas ocasiones difiere muy poco del coeficiente de correlación de Pearson para datos dicotómicos, y del coeficiente phi (Keid y Roberts 1978).

Huynh (1976a) propone un método para calcular P_o y K. con una única aplicación de una prueba. Asume que los puntajes verdaderos en la prueba deben distribuirse como una distribución beta, esta suposición comúnmente se da pues las distribuciones beta pueden tomar diferentes formas dependiendo de los diferentes valores de los parámetros α, β y n. Una segunda condición que asume Huynh (1976a) es la siguiente: si las pruebas de n ítemes fuesen repetidamente administrados a un individuo, la distribución de puntajes resultante se asume que es binomial y dichos puntajes pueden ser simulados utilizando un modelo matemático. Este supuesto se cumple si existen tres condiciones:

a) los ítemes se califican dicotómicamente 0 ó 1,

b) los ítemes son estadísticamente independientes, de tal forma que el resultado de uno no determine el resultado de los otros, y

./.

c) los ítemes tengan igual dificultad.

De estas, la que es más difícil de cumplir es la tercera, pues en la práctica las pruebas tienen ítemes de diferente dificultad, sin embargo, la comparación de este modelo con otros más complejos que toman en cuenta las diferencias de dificultad de los ítemes es favorable. La violación de la condición C lo que hace es producir estimaciones ligeramente conservadoras de la confiabilidad (P_o o K) (Berk (1980)).

Los procedimientos de cálculo que se requieren para determinar el índice de Huynh son tediosas y consumen mucho tiempo si se hacen en forma manual. El mismo autor propone otro método más sencillo en sus cálculos que requiere el cumplimiento de ciertas condiciones, además para disminuir el trabajo operatorio, Huynh ha tabulado los diferentes valores de los índices de consistencia P_o y Kappa (K).

Aunque menos sofisticada matemáticamente Subkoviak (1976, 1980) propone el coeficiente de acuerdo P_o, definido como la sumatoria de las probabilidades de clasificaciones de maestría consistentes, de los examinandos, en formas paralelas. Para estimar el coeficiente asume que las dos distribuciones son binomialmente idénticas e independientes y que la regresión de los puntajes verdaderos en los puntajes observados ès lineal. Este índice provee información tanto individual como grupal y al igual que el de Huynh se puede obtener con una sola administración de una prueba y produce resultados similares a los de éste. Subkoviak (1980) señala con respecto a este -

./.

último punto:

> "...pero en la práctica, los dos procedimientos generalmente producen resultados similares. Esto no es del todo inesperado, pues los su - puestos en los dos métodos son básicamente equivalentes a pesar de - las apariencias externas" (p. 142).

También Marshall-Haertel (1976) presentan un método de cálculo de un índice de acuerdo P_o, que también requiere una sola administración y asume que, si un individuo es examinado repetidamente, la distribución de sus puntajes observados tendría forma binomial.

Diversas comparaciones empíricas se han hecho entre los métodos de cálculos de los índices P_o y K. Subkoviak (1978, 1980), en un estudio que involucró 1.586 estudiantes que contestaron formas paralelas de pruebas de 10, 30 y 50 ítemes, concluye lo siguiente:

a) El método de Swaminathan et al (1974) es el más simple de cálculo y produce "estimaciones" no sesgadas; aunque tiene la desventaja que requiere dos administraciones y los errores de estimación tienden a ser grandes para grupos pequeños (menores de 40).

b) Las ventajas y desventajas de los métodos de Huynh, Subkoviak y Marshall-Haertel son similares; tienen la ventaja de requerir solamente la administración del test en una sola ocasión y producir "estimaciones" con pequeños errores estándar para grupos pequeños. Las desventajas están en las estimaciones sesgadas que producen para pruebas de pocos ítemes y en lo tedioso que resulta su computación. Las "estimaciones" sesgadas para pruebas de pocos ítemes

./.

son diferentes para cada uno de los tres índices: el de Huynh produce índices subestimados, mientras que el de Subkoviak produce valores sobreestimados para puntajes de corte altos y valores subestimados para puntajes de corte bajos (Algina y Noe, 1978) y Marshall por su parte produce valores sobreestimados en los puntajes de corte medios e índices subestimados para puntajes de corte extremos .

Por su parte Huynh (1981) indica que el índice de acuerdo (P_o) es la proporción combinada de examinandos clasificados consistentemente como masters y como no masters (si hay sólo 2 categorías) en las dos administraciones, mientras que Kappa (K) expresa la propiedad con que los puntajes de la prueba aumentan la consistencia de las decisiones, más allá de lo esperado por el azar.

Huynh y Saunders (1980) compararon los índices P_o y K empleados según los procedimientos de Hambleton y Novik (1973) y Swaminathan, Hambleton y Algina (1974) con los índices P_o y K producidos por el procedimiento de Huynh (1976a) ellos concluyen que: "los resultados indican claramente que el estimado de una sola administración (beta-binomial) de P_o se comporta adecuadamente con una cantidad despreciable de sesgo negativo; un grado moderado de sesgos negativos (acerca del 10 por ciento) muestra el estimado beta-binomial para el índice Kappa" (p. 357). Además, aunque los estimados beta-binomiales sean derivados, asumiendo que los ítemes son homogéneos en contenido y dificultad, los datos del estudio por ellos reportados muestran que los sesgos de estos estimados, no dependen de ninguno de estos supuestos.

./.

Por su parte Peng y Subkoviak (1980) y Peng (1979) hicieron estudios con datos reales simulados comparando los dos métodos aproximados (de más fácil cálculo) de los estimados K y P_o de Huynh (1976a). En ambos casos se muestra que el procedimiento aproximado de cálculos menos complejos (aproximación simple normal) produce estimados más correctos, de valores exactos de confiabilidad.

Apoyándose en Subkoviak (1980), y en un estudio realizado, Lang (1982) indica que los coeficientes P_o y K son sensitivos a diferentes tipos de consistencia de la decisión de maestría-no maestría: P_o representa la proporción total de clasificaciones consistentes que ocurren por cualquier razón; mientras que en el coeficiente k está corregido al azar. Por lo tanto la escogencia de P_o o K depende de si se quiere la consistencia total o sólo la de la prueba.

Indica además que la confiabilidad de la consistencia de las decisiones es sensitiva a la densidad de los puntajes en las cercanías al puntaje de corte y si el puntaje de corte se mueve en alguna dirección con respecto a la media (aumentándolo o disminuyéndolo) la confiabilidad aumentaría (P_o).

Cerca del punto de mayor densidad P_o adquiere su valor más bajo por esta razón es necesario reportar otros datos como la media y el puntaje de corte. Cuando el puntaje de corte se acerca a la media, la probabilidad de hacer decisiones inconsistentes en ambas administraciones es muy alta.

D. R. Digvi (1980) también indica que para interpretar un coeficiente o

./.

índice de confiabilidad debe tenerse como información adicional, la media y la varianza de los puntajes ya que el coeficiente K es mayor cuando el puntaje de corte está muy cercano a la media (Huynh 1976) y P_o es mayor cuando el puntaje de corte se aparta de la media (Subkoviak 1976).

Confiabilidad de los puntajes de prueba con referencia a criterios.

Esta confiabilidad se refiere a la estabilidad de las desviaciones de los puntajes con respecto al puntaje de corte, en dos pruebas paralelas o en una doble aplicación de una misma prueba.

Brennan y Kane (1977) establecieron una medida de la confiabilidad, a la que llamaron índice de seguridad, para pruebas con referencia a criterios, basándose en la teoría de la generalidad.

El índice de seguridad ("dependability index") $\Phi(\lambda)$, donde λ es el puntaje de corte, es un índice similar al planteado por Livingston (1972a) con la diferencia que toma en cuenta la definición de error inherente al propósito de las pruebas con referencia a criterios que desean distinguir entre el puntaje de cada examinado y un puntaje de corte. El error está dado por:

$$\Delta = (X_{pi} - \lambda) - (\mu_p - \lambda) = X_{pi} - \mu_p$$

donde:

Δ = es el error para un examinando

X_{pi} = es el puntaje observado promedio para un examinando en una muestra de ítemes

./.

μ_p = el puntaje universo para la persona p.

λ = es el puntaje de corte.

Esta definición de la varianza de error es la principal distinción entre este enfoque y los anteriores. Livingston (1972), se fundamenta en el error definido dentro de la teoría clásica de los tests, además de la diferencia ya - antes señalada en la definición de pruebas paralelas. El índice de seguridad Φ (λ) tiene varias características:

a) incorpora la varianza de error, definida de acuerdo con la definición anterior de error

b) será diferente para diferentes valores de λ

c) tiene como límite superior el valor uno (Brennan, 1980, p. 203).

También Brennan (1980) establece el índice de seguridad con un propósito general que es una derivación del anterior, definida como su límite inferior. Es interesante notar que al calcular este último índice y los índices Kuder-Richarson 20 y 21 se da la siguiente relación: KR-21<Φ<KR-20. Del mismo modo, es importante señalar que Brennan (1979) ofrece un programa de computadora para la ejecución de este análisis de acuerdo con la teoría de generabilidad.

Berk (1980a) también establece comparaciones entre los diversos índices de confiabilidad dentro de las tres corrientes para solucionar el problema de la confiabilidad dadas por Hambleton et al (1978) y discutidos anterior-

./.

mente es este capítulo. Algunas de sus principales recomendaciones se puede resumir de la siguiente forma:

a) El índice P_o debe ser empleado para pruebas con referencia a criterios - en las que existe un puntaje de corte absoluto y para pruebas que tienen - subtest cortos o producen baja varianza en los puntajes.

b) El índice K es más útil para pruebas donde los puntajes de corte relativos se establecen de acuerdo con las consecuencias de que un porcentaje dado de estudiantes aprueben o no.

c) En cuanto a los dos índices dados por Brennan (1977, 1980) y Livingston (1972a) para calcular la confiabilidad de los puntajes de pruebas con referencia a criterios así como para K y P_o, se recomienda, que para una mejor interpretación de los índices, se reporte junto a ellos el puntaje de corte, la longitud de la prueba, la media, el estimado de la varianza de error y las especificaciones de la prueba (Berk, 1980; Lang, 1982; Huynh, 1978; Digvi, 1982).

Otros índices. Frasier y Raeth (1980) estudian la adopción del K de Cohen (1960) como un índice de consistencia interna de las pruebas con referencia a criterios. Proponen dividir la prueba en dos mitades cada una con conductas iguales.

Por su parte Popham (1978) ataca el concepto de confiabilidad siguiendo la clasificación clásica de: estabilidad, equivalencia y consistencia interna. Con respecto al paradigma de la estabilidad, aconseja la administra

./.

ción de la prueba dos veces después de un programa instruccional, arreglar los datos de un cuadro de 2x2 con los niveles de "masters" y "no masters" en las administraciones de la prueba y aplicando luego el coeficiente phi o el análisis de chi cuadrado o simplemente usando el porcentaje de decisiones - correctas. Para la aplicación del paradigma de la equivalencia, Popham - (1978) sugiere un método para obtener el promedio de la correlación entre - formas de una misma prueba, con base en una única aplicación de la prueba y la simulación, por medio de métodos computarizados, de una serie grande de posibles muestras aleatorias de dos subpruebas cada una con la mitad de los ítemes de la prueba original. Finalmente con respecto al empleo de medidas de consistencia interna Popham (1978) escribe:

> "Consecuentemente, para las pruebas con referencia a criterios, mejor concebimos los estimados de consistencia interna como un vehículo para verificar la homogeneidad derivada de un grupo de ítemes de un - test. Conceptualmente, los métodos de consistencia interna no son particularmente útiles cuando se piensa en la consistencia de medición. de una prueba con referencia a criterios" (p. 155).

Resumen

La confiabilidad de las pruebas con referencia a criterios se puede establecer desde varios puntos de vista:

a) Confiabilidad de los valores estimados de los puntajes de dominio, con - métodos tales como los de Crombach (1972), Millman (1974), Hambleton, Swaminathan y Algina (1976).

b) Confiabilidad de las decisiones de clasificación de dominio o maestría, para esta forma de estimar la confiabilidad existen numerosos índices de -

./.

acuerdo, empezando por el más simple dado por Carver (1970) hasta los más complejos en su derivación matemática como el de Huynh (1976)

c) Confiabilidad de los puntajes de pruebas con referencia a criterios que se refiere a la estabilidad de las desviaciones con respecto al puntaje de corte con el índice $\Phi(\lambda)$ de Brennan (1977) basado en la teoría de la generabilidad y el de Livingston (1972)

d) Confiabilidad respetando los paradigmas clásicos de estabilidad, equivalencia y consistencia interna, como lo proponen entre otros Popham (1978) - Haladyna (1974) y Livingston (1972a).

La selección de uno o más de estos puntos de vista y del índice respectivo es una decisión en que deben considerarse varios aspectos entre los que podemos señalar:

a) La naturaleza de las alternativas de decisión; sobre individuos o sobre programas

b) Si interesa solamente clasificar a los estudiantes como "masters" y "no masters" o si se interesa conocer las diferencias de grado (con respecto al puntaje de corte) tanto de los "masters" como de los "no masters"

c) La varianza de los puntajes, aspecto esencial en la aplicación de índices clásicos

d) La disponibilidad de administración de pruebas paralelas o la posibilidad de examinar con una misma prueba dos veces a un grupo de individuos

./.

e) La disponibilidad de programas de computación

f) El número de ítemes en cada subprueba (ítemes que miden un mismo objetivo o especificación).

Cabe señalar que dos estudios comparativos de índices clásicos de consistencia interna e índices como los revisados en este trabajo concluyen lo siguiente:

a) Moyer y Fishbein (1977) indican que Kappa parece estar relacionado con la homogeneidad de los ítemes en una prueba, medida por KR-20

b) Downing y Mehrens (1978) después de comparar empíricamente los dos índices de Huynh (1976), el índice de Livingston (1972a), el índice de Subkoviak (1974) y los índices de Kuder-Richarson 20 y 21, señalan que el coeficiente KR-21 es útil para pruebas con referencia a criterios para el investigador que no tiene acceso a facilidades de cómputo sofisticado; también concluye que todos los coeficientes excepto el de Subkoviak dan resultados semejantes, Lovett (1978) concluye que el índice KR-21 tiende a ser más válido cuando hay baja variabilidad entre medias de ítemes, y el índice KR-20 cuando la variabilidad es alta.

Berk (1980), indica que los enfoques de confiabilidad b y c (pág. 20) no son óptimos para todas las aplicaciones y recomienda que se use el enfoque de Brennan y Kane (c) si la importancia radica en el grado de "maestría" o "no maestría" y el (b) (métodos de varios autores) si no es así y solo inte

./.

resa la clasificación de los estudiantes como "masters" o "no masters" independientemente de su cercanía o alejamiento del puntaje de corte.

Hambleton et al (1978) por su parte recomiendan que independientemente del enfoque utilizado, la información relacionada con la confiabilidad debe ser reportada para cada objetivo.

Lang (1982) hace énfasis en la importancia del puntaje de corte en la clasificación de maestría.

Desarrollo, selección y análisis de ítemes

Varios autores, (Haladyna (1980), Hambleton (1979), Enright (1982)) coinciden en que para construir y validar una prueba con referencia a criterios es necesario llevar a cabo los siguientes pasos:

1.- Definir el propósito de la prueba y preparar o seleccionar los objetivos conductuales u otro esquema descriptivo, que se pretenda medir y sus especificaciones.

2.- Dar las especificaciones necesarias para la construcción de la prueba: número de ítemes, tipo de preguntas, uso de la prueba, condiciones de aplicación, vocabulario que debe emplearse.

3.- Confeccionar ítemes que midan los objetivos seleccionados para formar la prueba (o pruebas si se necesitan formas paralelas) y hacer una edición preliminar de los mismos.

./.

4.- Hacer una valoración sistemática de los ítemes para determinar su "congruencia" con el objetivo respectivo, su calidad técnica y su representatividad.

5.- Descartar y/o ajustar los ítemes de acuerdo a los resultados del punto anterior.

Para los dos últimos autores es necesario también:

6.- Montar la(s) prueba(s)

7.- Establecer los estándares que permitan interpretar los resultados.

8.- Aplicar la(s) prueba(s)

9.- Hacer la valoración de la validez y de la confiabilidad de las pruebas y recopilar datos normativos si se considera necesario.

Uno de los autores propone además como pasos adicionales, la preparación de manuales (uno técnico y otro para el usuario) y la recopilación periódica de información técnica adicional.

En este subtítulo se revisará bibliografía correspondiente a los pasos - del dos al quinto de los citados anteriormente.

Especificaciones del dominio de conocimientos de la prueba y desarrollo de los ítemes.

Diversos autores han tratado este tema y es uno de los aspectos de las - pruebas con referencia a criterios que más atención está recibiendo actual-

./.

mente. Popham (1980) hace una revisión de las estrategias de especificación del dominio, empieza por establecer que este es el paso más importante en el desarrollo de las pruebas con referencia a criterios, dice él: "una prueba con referencia a criterios que no describa sin ambiguedades exactamente lo - que está midiendo, no ofrece ninguna ventaja sobre las medidas con referen - cia a normas... Note que una condición requisito, para dar una interpreta - ción exacta de lo que significa el rendimiento de un estudiante en una prueba, es una descripción clara de la naturaleza de los ítemes de la prueba" - (p. 16). Berk (1979a) por su parte ofrece varios argumentos contra la definición de estrategias empleando solamente un esquema del contenido, un grupo de objetivos, una tabla de especificaciones o un cuadro de balanceo. Estos argumentos son:

a) Cualquiera de las anteriores especificaciones producen una definición ambigua del dominio

b) La sujetividad se involucra mucho en la composición de estas especificaciones pues la selección de tópicos y objetivos es arbitraria reflejando solo la conceptualización de un investigador

c) Están abiertos para interpretaciones diferentes

d) Son inadecuadas para la escritura de ítemes ya que los grupos de ítemes que se desarrollen con base en estas especificaciones reflejarán los sesgos e idiosincracias de cada escritor.

./.

Cuatro estrategias analiza Popham (1980):

1) Objetivos conductuales

2) Formas de ítemes

3) Objetivos amplificados

4) Especificaciones IOX [1/] del test.

De los primeros afirma que son limitados pues son abreviados, no constriñen suficientemente al escritor de ítemes y dejan en sus manos muchas decisiones. Las formas de ítemes fueron desarrolladas originalmente por Hively, Patterson y Page (1968); estas son reglas detalladas para crear ítemes que se esperan por naturaleza sean homogéneos. Son las formas de ítemes entonces, un proceso que tiene las siguientes características:

a) Genera ítemes, con una estructura sintáctica fija

b) Contiene uno o más elementos que varían y

c) Define una clase de frases para el ítem al especificar los grupos de reemplazos para los elementos que varían.

En una forma de ítem se detallan los siguientes elementos:

a) Título descriptivo, ej: resta, hecho básico, minuendo menor que 10

b) Muestra de un ítem, ej: 13-6

c) Forma general, ej: A-B, y

d) Reglas para generar ítemes, ej: a=1a; B=b; (a<b) ε U; {H, V}, donde A y

./.

1/ IOX significa Instructional Objetives Exchange (Intercambio de objetivos para instrucción).

B son numerales, a y b son dígitos, U grupo (1, 2, ... 9) y {H, V} forma horizontal o vertical.

Los objetivos amplificados son versiones más elaboradas de un objetivo conductual. Se podría decir que representan el término medio entre un objetivo conductual y las formas de ítemes. Los objetivos amplificados están constituidos por:

a) Objetivos
b) Un ítem de muestra
c) Elementos de estímulo
d) Las alternativas de respuesta
e) Criterio de corrección, según Roid y Haladyna (1980)

Las especificaciones IOX fueron desarrolladas por Popham (1978) y tienen cuatro componentes, con un quinto adicional:

1) Descripción general y breve de la conducta a especificar; puede ser un objetivo conductual

2) Item de muestra; es una pregunta ilustrativa que refleja los atributos de las conductas

3) Atributos de estímulo ("stimulus attributes"); son una serie de afirmaciones que intentan delimitar la clase de material de estímulo que se encontrará el examinando; se establecen los factores que podrían delimitar la composi-

./.

ción de un grupo de ítemes.

4) Atributos de respuesta ("response attributes"); están constituidas por una serie de afirmaciones que intentan delimitar la clase de respuesta que el alumno escogerá o establecer los estándares explícitos por los que la respuesta construida por el estudiante se juzgará

5) Suplemento de especificaciones; esta parte es adicional y dependerá del contenido a ser medido y del evaluador. Se usa para detallar más los atributos del contenido a medir (en el apéndice A se ofrece un ejemplo de unas especificaciones IOX).

Además de las anteriores estrategias Berk (1978a, 1979a) desarrolla un mecanismo para generar especificaciones o ítemes basado en la teoría estructural de facetas. Las frases de mapeo ("mapping sentences") es el elemento básico de esta estrategia; están compuestas por partes variables y fijas. La parte fija se parece a un ítem y se compone de categorías llamadas facetas, éstas son las dimensiones del contenido dentro de las que variarán los ítemes potenciales. Cada faceta está, a su vez, compuesta de elementos de faceta que definen el contenido específico a medir y se presentan a manera de una lista de términos. Los posibles patrones de combinación entre los elementos de faceta generados por una serie de reglas constituyen el diseño de faceta. El producto de los diseños de faceta, llamados perfiles semánticos, sirven a su vez para constituir la base para el desarrollo de ítemes.

Otra estrategia de especificación y desarrollo de ítemes es

./.

la ofrecida por Bormuth (1970), quien propone una técnica para escribir ítemes que evalúen el aprendizaje de material en prosa. Esta técnica está constituida por una serie de reglas que le dicen al escritor de ítemes cómo transformar segmentos de material de instrucción en prosa, en preguntas. A esta técnica se le llama transformación de ítemes (Berk, 1979a). Bormuth (1970) estableció dos clases de transformaciones:

a) Itemes derivados de frases y
b) Itemes derivados de relaciones entre frases

Diversos autores, entre ellos Roid y Haladyna (1978), Finn (1975) y Roid (1979) ampliaron el trabajo de Bormuth (1970) al desarrollar un método para escribir ítemes de selección múltiple para material de aprendizaje en prosa. Este método puede dividirse en tres pasos básicos:

1) Análisis del texto y selección de frases
2) Transformación de frases en preguntas
3) Generación de alternativas para el formato de selección múltiple.

Asimismo, Roid y Haladyna (1980) reportan el trabajo de Markle y Tiemann (1978) referido a la investigación del empleo de conceptos en la enseñanza y la medición. Tiemann y Markle (1978) dan guías para el análisis de conceptos; el estudio en mención tiene varias etapas:

1) Establecimiento de los atributos críticos del concepto
2) Identificación de los atributos variables

./.

3) Generación de listas de ejemplos correctas y ejemplos erróneos para enseñanza y para exámenes. Los ítemes para una prueba se pueden generar al escoger al azar ejemplos correctos y erróneos variando sistemáticamente los atributos críticos y variables.

Finalmente existen varias estrategias, relativamente nuevas, basadas en el empleo de la computadora. Millman (1980) ofrece un buen resumen de cuatro estrategias: banco de ítemes, medición adaptiva, algoritmos y transformaciones linguísiticas.

Para resumir las características de las estrategias sucintamente revisadas en este trabajo, es conveniente examinar cuidadosamente lo que escriben Roid y Haladyna (1980):

> "La mayor limitación de los métodos actualmente disponibles para escribir ítemes con la mejor técnica, es que no se pueden aplicar, indiscriminadamente, a cualquier área de contenido y a cualquier nivel cognitivo. Cada método parece tener una aplicación particular. Los métodos de formas de ítemes y los métodos similares basados en el empleo de la computadora para escribir ítemes, se han aplicado principalmente an áreas de ciencias y matemáticas...Los métodos de Bormuth (1970), Finn (1975) y otros (ej. Roid, Haladyna y Shaughnessy, Nota 12) en su forma actual parece que siguen siendo aplicables principalmente a las áreas, para las cuales fueron originalmente desarrolladas: comprensión de lectura y memoria"(p. 309-310).

Por su parte Berk, R. A. (1979a) establece que:

> "Los perfiles de las estrategias sugieren que el rigor y precisión de las especificaciones son inversamente relacionadas a su practicabilidad. Las transformaciones de ítemes, formas de ítemes y algoritmos son capaces de generar dominios de ítemes finitos y el muestreo de dominio parece ser muy impráctico...Los objetivos amplificados, las especificaciones IOX del test y las frases de mapeo que son asociadas con la conceptualización de un dominio de ítemes infinito, -

./.

tienen el más grande potencial para el uso de maestros y evaluadores. Desafortunadamente, todas las estrategias excepto los objetivos amplificados y las especificaciones IOX del test han mostrado ser eficaces sólo en una área de contenido, esto es en lectura (transformaciones de ítemes), matemáticas (formas de ítemes, algoritmos) o conductas afectivas (mapeo de frases)". (p. 4)

Selección de ítemes

La selección de ítemes es parte del proceso de validación de contenido de la prueba. Tal y como se explicó en la revisión del concepto de validez, los ítemes desarrollados pasan un examen de especialistas para determinar su congruencia con los objetivos (validez descriptiva), su calidad y su representatividad (Hambleton, 1980). Para Berk (1980b) los análisis de discriminación y dificultad contribuirán a mejorar la selección de los ítemes. Para él, los ítemes que sean congruentes, que discriminen (validez de decisión) entre grupos de "masters" y "no masters" o entre grupos de pre y post instrucción, y que tengan alta dificultad para los grupos sin enseñanza y baja para los grupos después de la enseñanza, deben ser seleccionados para formar el grupo de ítemes de entre los cuales, aleatoriamente se seleccionarán los que constituirán las formas paralelas de la prueba.

Análisis de ítemes

Este se lleva a cabo para mejorar la selección de los ítemes. Según Berk (1978b, 1980b) el análisis de ítemes se hace mediante la revisión de los siguientes aspectos: congruencia, estadísticas, selección y revisión. Con respecto a los estadísticos indica que los pasos a seguir incluyen la selección de los grupos de criterio, obtención de la información informal de par

./.

te de los estudiantes y el cálculo de índices de dificultad, discriminación y homogeneidad; se revisará cada uno de estos pasos descritos por Berk. La selección de grupos debe fundamentarse en el propósito de la prueba. En la mayoría de los casos, una prueba con referencia a criterios se emplea para identificar "masters" de "no masters", por lo consiguiente comúnmente se requerirán dos grupos de individuos, sea con estudiantes que han y no han recibido enseñanza sobre los contenidos que la prueba pretende medir, o sea - con grupos pre y postinstitución; a estos grupos se les conoce frecuentemente como grupos criterios. Asimismo se necesitan dos grupos criterio para - el cálculo de casi todos los índices. Las dos estrategias de selección de los grupos: grupo de pre y postinstrucción, y grupos con y sin enseñanza - tienen ventajas y limitaciones, aunque por razones de economía, de tiempo y practibilidad se recomienda la primera estrategia.

La retroalimentación informal de los estudiantes se puede obtener fácilmente después de aplicada la prueba. Con este propósito, se llevan a cabo discusiones o entrevistas individuales con el grupo de estudiantes. Las - preguntas que se hacen a los estudiantes en grupo o en forma individual, deben versar sobre respuestas incorrectas, comprensión de palabras y claridad en la redacción de las preguntas. Esta retroalimentación es importante porque prevee información que no se puede obtener de un examen cuantitativo de los ítemes.

La medida de dificultad de los ítemes es el porcentaje de personas que - responden el ítem correctamente. Este índice es equivalente a la media arit

./.

mética del ítem multiplicado por 100. Este índice puede variar de 0 a 100. Los estimados de dificultad deben obtenerse para ambos grupos de criterio, ya que los índices de un ítem dado, obtenidos con cada grupo, se podrían emplear para su selección.

El índice de discriminación del ítem mide cambio en rendimiento, si es - entre pretest y postest, o diferencias entre los grupos con enseñanza o sin ella. En la literatura se encuentran once formas diferentes de calcular el índice de discriminación. Berk (1980a, 1980b) los agrupó de acuerdo con su practibilidad y complejidad . Los cuatro primeros los juzga como conceptuales y de cálculos simples, pero con fundamento estadístico. Estos son los pro - puestos por Cox y Vargas (1966), Klein y Kosecoff (1976) y Roudabush (1973). Todos ellos usan la proporción como estadístico. Cox y Vargas (1966), lo - definen como la proporción de estudiantes que responden al ítem correctamente en el postest menos la proporción que lo responden correctamente en el - pretest. El siguiente índice (Klein y Kosecoff (1976)) se define como la - proporción de educandos que responden en forma correcta en el grupo instruido menos la proporción que lo responden correctamente en el grupo no instruido, Para Roudabush (1973) la discriminación es la proporción de estudiantes que contestaron el ítem incorrectamente en el pretest y correctamente en el postest. Todos estos índices tienen valores en el rango de -1 a +1.

Vale la pena destacar entre estos, el índice "B" de Brennan (1972) y Hsu (1971) que emplea la proporción de "masters" y "no masters" de un solo grupo instruido como base de cálculo y él usa puntaje de corte para definir -

./.

los "masters" y los "no masters". Según Berk (1980) tiene las desventajas de que: "la validez del puntaje de corte es una condición necesaria para la - validez del estadístico del ítem y la interpretación del índice no es ortodoxo". (p. 62).

Por otra parte existen cuatro estadísticos para medir la homogeneidad; - con ellas se intenta verificar estadísticamente que los ítemes que se juzgaron congruentes con un objetivo, se comportan de tal manera después de una administración de la prueba o de administraciones sucesivas. Berk (1980b) opina que las condiciones que se asumen para buscar la homogeneidad o sea - "que los ítemes deben dar idénticos índices de dificultad o puntajes de cambio son cuestionables. Esta "homogeneidad" puede ser no realística y, de - hecho indeseable..." (p. 64).

Finalmente Berk (1980), sugiere técnicas para revisión de los ítemes, - cuando estos muestran índices de valores no óptimos, basados en el análisis de las respuestas dadas por los estudiantes (frecuencia en cada alternativa de respuesta del ítem).

Otras consideraciones técnicas

Como W. J. van der Linden (1982) indica, el establecer el puntaje - de corte y la longitud de la prueba de forma que permitan tomar decisiones óptimas, es un problema clásico en una prueba con referencia a criterios. A continuación se hará mención de cada uno de estos temas.

Longitud de la prueba

Se entiende por la longitud de la prueba el número de ítemes que miden -

./.

cada objetivo o especificación del test. Esta característica está directamente asociada con la utilidad de los puntajes de una prueba con referencia a criterios. Las pruebas muy cortas producen estimados de puntajes de dominio muy imprecisos y por lo tanto, las decisiones de maestría o dominio serán inconsistentes para pruebas paralelas o para dos administraciones de una misma prueba. Existen varios métodos para determinar la longitud de la prueba. Fhaner (1974) introduce el concepto de zona de indiferencia, esta se da al sustituir el puntaje de corte por un intervalo o rango (Π_o y Π_1). Él propone que se escoja un valor mínimo de "n" y un valor de "c" para los que las probabilidades de producir clasificaciones erróneas sean mínimas. En esta misma línea de pensamiento se encuentra Wilcox (1982), que propone evitar que el establecimiento de "n" sea hecho arbitrariamente, propone un método llamado "respuesta hasta tanto correcta" fundamentado en el trabajo de Fhaner (1974).

Por otra parte, tanto van der Linden (1982), Millman (1972, 1973) y Hsu (1980) emplean el método del error binomial para relacionar el puntaje de corte con la longitud de la prueba. De esta manera se pueden lograr longitudes óptimas conociendo las pérdidas asociadas a los errores falso-negativo y falso-positivo. Millman (1972, 1973) ofrece tablas en las cuales dado un valor de Π, n y c se encuentra la probabilidad con que una persona, con un determinado puntaje en la prueba, es clasificada correcta o incorrectamente. (c es el puntaje de corte del objetivo).

Para Berk (1979b) cuatro son los factores esenciales para determinar

./.

cuántos ítemes deben construirse para una prueba. Estos factores son:

a) Importancia y tipo de decisiones que se harán con los resultados

b) Importancia y énfasis asignado a los objetivos

c) Número de objetivos

d) Limitaciones prácticas.

Tomando en consideración la investigación hecha en puntajes de corte y - confiabilidad, él recomienda que se empleen entre 5 y 10 objetivos, en aquellos casos en que se tomen decisiones de aula y entre 10 y 20 objetivos si las decisiones se emplean a nivel de región o nacional, ya que en el aula - las decisiones tomadas pueden cambiarse o corregirse si existiera error y a nivel de región o nacional es difícil hacerlo.

Hambleton, Hutten y Swaminathan (1976) en un estudio empírico en que comparan métodos de obtener los puntajes de dominio y su efecto en varios factores (entre ellos la longitud del test) concluyen que un número de ítemes igual a ocho da "suficiente base para evaluar el dominio del estudiante o para tomar decisiones de instrucción para los datos de pruebas con referencia a criterios" (p. 62).

Por su parte Popham (1978) afirma lo siguiente: "Para simplificar un poco, para muchas de las situaciones educativas en las que se emplearán pruebas con referencia a criterios, ya sea el modelo binomial o el Bayesiano - dictan que la prueba debe consistir de 10 a 20 ítemes por dominio conductual" (p. 101).

./.

Por otra parte, muchas de las pruebas que se encuentran en la bibliografía no tienen tantos ítemes por objetivo como recomiendan los autores señalados anteriormente. Por ejemplo, Poggio y Glasnapp (1980) en el desarrollo del programa de pruebas de maestría, empleado en el Estado de Kansas, usaron tres ítemes por objetivo; Sheehan y Davis (1979) desarrollaron una batería de pruebas con referencia a criterios de matemáticas, en las que emplearon cuatro ítemes por objetivo. En el país, las pruebas desarrolladas por Esquivel, Peralta y Delgado (1984) en matemáticas y por Esquivel y Quesada (1985) en ciencias está constituidas por tres ítemes por objetivo.

Puntaje de corte

Hambleton (1978) define el puntaje de corte como "un punto en la escala de puntajes de una prueba que se utiliza para clasificar a los individuos dentro de dos categorías,que reflejan diferentes niveles de pericia o habilidad con respecto a un objetivo particular medido en la prueba" (p. 279).

El mismo autor (Hambleton, 1980) establece, en una catalogación de los métodos de definición de puntajes de corte, que los mismos se basan en:

a) Contenido de los ítemes
b) Puntaje al azar y muestreo de ítemes
c) Datos empíricos de grupos de "masters" y "no masters"
d) Procedimientos teóricos
e) Medidas de criterio externo y
f) Consecuencias educativas

./.

Es muy importante hacer notar que todos los métodos involucran juicio y son arbitrarios. Como Popham (1978a) lo manifiesta muy bien, dicha arbitrariedad no significa juicio caprichoso, sino más bien juicio meditado, además, en la vida muchas cosas se hacen arbitrariamente como: estándares de salud, de incendios o de la conservación del ambiente (Popham, 1978a, Hambleton, 1978).

Huynh (1980) indica que muchos de los procedimientos estudiados para el establecimiento de los puntajes de corte se pueden clasificar dentro de las siguientes tres categorías:

- Comparación con la ejecución de otros individuos (usando NRT)
- Revisión del contenido de los ítemes (tal como el de Nedelsky)
- Consideración de las consecuencias en que se incurre si se da una clasificación errónea (Hambleton, Swaminathan, Algina y Coulson (1978) hacen una revisión de algunos de estos procedimientos).

De acuerdo con Hambleton (1980) los métodos se pueden catalogar en tres clases:

a) Métodos de juicio
b) Métodos empíricos
c) Métodos combinados

En los primeros, los ítemes se analizan y se juzga cuál sería el rendimiento de una persona con capacidad de maestría mínima. Entre estos métodos tenemos: Nedelsky, Angoff, Ebel y Jaeger.

./.

Al referirse a los métodos de juicio, Francis y Holmes (1983) indican que los mismos se pueden dividir en dos clases: los primeros, involucran un juicio con respecto al contenido u otros aspectos de la prueba. Los segundos, contienen juicios acerca de los individuos o grupos de individuos.

Los métodos empíricos sugeridos principalmente por Livingston (1975) emplean una serie de funciones lineales o semilineales para establecer el efecto de la exactitud de la decisión sobre un estándar o puntaje de corte. También Veldhnyzen (1982) propone un método que utiliza la función "utilidad" por medio de la cual ordena las posibles consecuencias de una clasificación errónea y de ella parte para establecer un puntaje de corte, tal que las inferencias que se hagan sean las óptimas. Llama a su procedimiento "MAXIN". Huynh (1980) utilizó tanto esta función de utilidad como otras como la lineal y la cuadrática para establecer "c" de tal manera que minimice el error de clasificación de un individuo; otra contribución importante es la aportada por van der Linden que sugiere la utilización de métodos fundamentados en las teorías de Bayes y Neyman Pearson, para calcular los puntajes de corte, para obtener el mismo propósito de decisiones óptimas con el mínimo de error posible.

Los métodos combinados, así llamados por mezclar datos empíricos con juicios, utilizan grupos de individuos con los que se recogen datos; pero también se emplean jueces para juzgar la ejecución de los estudiantes. Los autores de estos métodos son Berk (1976), Lieky y Livingston (1977) y Popham (1978d). Una excelente crítica de Glass (1978) al establecimiento de están

./.

dares y sus métodos, provocó la reacción de varios autores, entre ellos, - Popham (1978b), Block (1978) y Gross (1982).

Sheeham y Davis proponen un método, en el que utilizan la probabilidad - de que un alumno responda correctamente un ítem cuando su respuesta la hace por adivinación. Este valor de probabilidad se multiplica por el número de ítemes y se le suma un número determinado de desviaciones estándar. En el caso particular de los autores, recomiendan sumar dos desviaciones estándar, basados en la opinión de Gulliksen (1950) que dice: "Un puntaje que está - dentro de una o dos desviaciones estándar del puntaje aleatorio, no debe - ser interpretado como que signifique conocimiento de la materia del examen" (p. 128). Cabe destacar que esta decisión es de juicio, por lo que se ha - clasificado este método, como un método combinado.

Se puede resumir este concepto diciendo que aunque existe una fuerte polémica sobre la utilidad de los puntajes de corte, "estos aunque involucren juicio y arbitrariedad son preferibles al no uso de estándares del todo, en términos de aprendizaje de los estudiantes y del desarrollo de programa de instrucción" (Block, 1978, p. 295). Para finalizar cabe indicar la recomendación de Hambleton (1980) para que se empleen las técnicas de Ebel, Nedelsky y Angoff. En la comparación empírica de estas técnicas hechas por Poggio, Glasnapp y Eros (1981) se concluye que: "El uso de un único método para establecer el estándar de rendimiento es arbitrario y que la literatura existente y los datos presentes no dan sustento a la superioridad de uno cual - quiera de los cuatro métodos investigados" (p. 18).

./.

Estimación del puntaje de dominio

Hambleton, et al (1978) afirman acerca del puntaje de dominio:

> "El problema básico, dado el puntaje de un estudiante en un conjunto de ítemes que miden un objetivo, es estimar el puntaje proporcional acertado por el estudiante como si se le hubiesen administrado todos los posibles ítemes que miden el objetivo. Ese puntaje "estimado" se conoce como puntaje de dominio, puntaje de nivel de funcionamiento o verdadero puntaje proporcional acertado. Un examinado tiene un puntaje de dominio definido para cada uno de los objetivos medidos por la prueba con referencia a criterios" (p. 5).

De acuerdo con el examen de este concepto que hacen Hambleton, et al (1978) existen cinco métodos de "estimar" (cálculo aproximado) el puntaje de dominio de un estudiante. Debe entenderse que existirá un puntaje de dominio para cada objetivo que es medido por "n" ítemes. Los cinco métodos son:

a) Estimado de la proporción correcta de ítemes: es el más simple ya que únicamente se divide el puntaje de la prueba entre el número de ítemes, aunque este método da una estimación sin sesgo, es poco confiable cuando el número de ítemes en que se basa es muy pequeño.

b) Estimado del modelo clásico II: se pretende aplicar la teoría clásica de las pruebas con el estimado de regresión del puntaje verdadero.

c) Estimado del modelo Bayesiano II: emplea una solución bayesiana para estimar el puntaje de dominio de un conjunto de examinados.

d) Estimado de la media marginal: utiliza una modificación al método anterior para estimar el puntaje de dominio de un estudiante en particular

./.

e) Estimados cuasi bayesianos: son modificaciones al método del modelo Bayesiano II.

Otra forma de estimar los puntajes de dominio es el empleo de uno de los modelos de rasgos latentes ("latent trait models"), según lo señalan Hambleton y Cook (1977).

CAPITULO III

Metodología

En este capítulo se dan detalles de la construcción de las pruebas de conocimientos, dentro del modelo de pruebas con referencia a criterios, referidas a asignaturas que se imparten en la Educación Diversificada (Español, Matemáticas, Estudios Sociales, Física, Química y Biología; éstas tres últimas integradas en una sola prueba que se denominará de "Ciencias") así como de los procedimientos seguidos para obtener evidencias sobre la validez y confiabilidad de las mismas.

Fuente de datos

Las pruebas se aplicaron a una muestra de grupos de estudiantes de primer ingreso de las Instituciones de Educación Superior Estatales (excepto en la UNED). En cada una de las instituciones se definió un tamaño de muestra proporcional a la matrícula esperada en ella (10%) y se realizó un muestreo aleatorio y estratificado de los grupos según el horario de asistencia a lecciones de los mismos. Estos grupos reciben clases en bloques de cuatro horas que pueden ser en la mañana, en la tarde o en la noche.

La administración se realizó en sesiones de cuatro horas cada una y se aplicaron las pruebas de dos asignaturas en cada sesión. Las pruebas se aplicaron a un total de 3517 estudiantes; en el Anexo N°1 se puede ver la distribución de estudiantes por asignatura, fórmula e institución de procedencia.

./.

Con fines de investigación de ciertos índices, las pruebas también se aplicaron a una muestra de colegios de educación secundaria de San José. - Se escogieron dos colegios académicos diurnos oficiales, uno académico - diurno particular, dos profesionales diurnos oficiales y uno académico noc turno oficial. Esta distribución correspondió a la muestra más pequeña - que pudiera representar la distribución de las diferentes modalidades en el nivel nacional. En cada uno de los colegios escogidos se solicitaron - cuatro grupos y en cada uno de ellos se aplicaron las pruebas de una de - las asignaturas, en una sesión de 2 horas; se trabajó con un total de 711 estudiantes; en el Anexo Nº1 se puede ver la distribución de estos estudian tes por colegio, asignatura y fórmulas.

Procedimientos

Especificaciones de la prueba

Una vez realizada la revisión bibliográfica, tanto la metodología como algunas características básicas de las pruebas fueron establecidas de tal forma que se adaptaran a las condiciones reales en que se desarrollarían - las mismas.

Dichas características se irán mencionando en cada uno de los apartados metodológicos correspondientes.

Desarrollo de la especificación del dominio

La selección y especificación de los conocimientos mínimos que se medirían, se hizo siguiendo el siguiente procedimiento:

Para cada una de las asignaturas se formó un equipo de 3 profesores (en total 18 profesores ya que "Ciencias" es una composición de Física, Química y Biología) con experiencia tanto en la Educación Diversificada como en la universitaria, así como en la confección de objetivos cuando fue posible.

Con base en su experiencia como profesores, una explicación pormenorizada del proyecto y la lectura de documentos que junto con los programas oficiales de la asignatura, le fueron entregados para ese fin, cada uno elaboró una lista de 20 objetivos y un ítem de muestra por objetivo, que representaran, en su criterio y con base en los programas oficiales en la Educación Diversificada, los conocimientos mínimos que un estudiante debería poseer al finalizar la educación secundaria. Para el caso de Física, Química y Biología, cada profesor elaboró una lista de siete objetivos.

Posteriormente los tres profesores de cada asignatura, discutieron las listas en conjunto y elaboraron a partir de ellas una única lista de veinte objetivos con su respectivo ítem de muestra. Para el caso de Física, Química y Biología la lista constó de siete objetivos cada una. (Anexo N°2).

A estos objetivos se adicionó una escala de valoración de cuatro puntos y un instructivo para su empleo (Ver Anexo N°3); este documento fue presentado, al mismo tiempo que se justificó y motivó para el trabajo, a 250 profesores que impartían lecciones (y tenían experiencia de por lo menos tres años) en las diferentes modalidades de la Educación Diversificada, en las 18 regiones educativas del país (definidas así por el MEP), para que calificaran la importância que en su criterio tiene el contenido de cada obje-

./.

tivo, en los conocimientos que debe poseer un estudiante al egresar de la Educación Diversificada. Para esta etapa se contó con la colaboración de la Dirección Regional del MEP y dos equipos de dos profesionales cada uno que realizaron las giras y entrevistas, a quienes se entrenó para llevar a cabo las visitas.

Una vez recolectadas todas las calificaciones dadas por los profesores, se calcularon las medias aritmêticas de cada objetivo y se escogieron los doce objetivos con medias más altas; en el caso de Química, Física y Biología, se escogieron cuatro objetivos de cada una.

La especificación del dominio de conocimientos, indispensable en una prueba con referencia a criterios, se hizo mediante la técnica de "amplificación de objetivos" por considerar que los mismos permiten dar suficiente claridad a la definición del dominio, sin requerir demasiados recursos para su desarrollo. La amplificación fue realizada por uno de los profesores de cada grupo, a quien se instruyó para esa tarea mediante un documento confeccionado con base en unos ejemplos de amplificación dados por Popham (1978) (Ver Anexo N°4).

Aunque existen otras técnicas para la especificación de los objetivos, todas excepto los objetivos amplificados y las especificaciones IOX, han demostrado ser eficientes solo en área de contenido (Berk, (1979)), y estas últimas aunque más extensas y profundas, demandan una gran cantidad de recursos.

./.

Al seguir los procedimientos descritos para obtener las listas de objetivos y las especificaciones de los mismos, se puede afirmar que existe evidencia de la validez de selección del dominio (Popham, (1978)) y que se ha cumplido cabalmente con uno de los elementos necesarios para establecer la validez de contenido (Hambleton, (1980)).

Desarrollo y validez de los ítemes y validez descriptiva de la prueba

Para la escritura de los ítemes con los cuales se medirían cada uno de los doce objetivos, se escogieron cuatro personas con experiencia en escritura de ítemes, por asignatura, quienes escribieron cinco ítemes de selección para cada objetivo. Los ítemes de selección estaban compuestos (tal como se les indicó) por un enunciado y cuatro o cinco alternativas de respuesta, siendo una de éstas la respuesta correcta o clave. Antes de iniciar la escritura de los ítemes cada una de las personas contó con documentos referentes a la escritura de ítemes de selección y con los objetivos amplificados.

Para evaluar, si los ítemes medían el objetivo para el cual fueron creados o sea para validar su conguencia, fueron sometidos al juicio de ocho especialistas en cada asignatura (profesores de las universidades) a quienes se solicitó que dieran un valor de + 1 si consideraba que el ítem medía el objetivo en cuestión, un 0 (cero) si estaba indeciso y un - 1 si creía que no lo medía y debía tomar cada uno de los doscientos cuarenta (240) ítemes y compararlo con cada uno de los objetivos conductuales (no amplificados), o sea què, a cada juez en las asignaturas de Español, Matemática y

./.

Estudios Sociales le correspondió emitir dos mil ochocientos ochenta jui - cios. En el caso de Química, Física y Biología en que el número total de ítemes para cada asignatura era ochenta y debía compararlos con los cuatro objetivos de cada una, a cada especialista le correspondió emitir trescien tos veinte juicios.

Para cada ítem se calculó el índice de congruencia ítem-objetivo de Ham bleton y Rovinelli (1977) y se aceptaron como congruentes aquellos ítemes con un índice igual o mayor a .75. El índice empleado es uno de los tres índices sugeridos en la literatura para revisar la congruencia (Hambleton, 1980), y el puntaje de corte es sugerido también por ese autor. En los - cuadros del 1 al 4 del capítulo IV se puede ver el número de ítemes por ob jetivo y asignatura que fueron considerados congruentes.

Para evaluar la calidad técnica de los ítemes se escogieron dos jueces por asignatura (también profesores de las universidades), con experiencia en la escritura de ítemes y conocimiento de la asignatura en cuestión. Se entregó a cada juez la totalidad de los ítemes de cada asignatura así como una hoja de cotejo para cada ítem. Esta actividad se llevó a cabo en for- ma paralela al análisis de congruencia, por razones de tiempo. En el Anexo N°5 se puede ver el número de ítemes por objetivo y asignatura que fueron considerados aceptados, en la primera revisión. Los ítemes son sugerencias de modificación por parte de los jueces de calidad técnica fueron devueltos a los escritores de ítemes para su corrección o sustitución tantas veces - como fue necesario; se hizo una evaluación del tipo de modificación -

./.

efectuado y algunos de estos ítemes junto con los nuevos (sustituciones) - fueron sometidos al análisis de congruencia nuevamente.

Una vez revisados todos los ítemes se construyeron tres pruebas paralelas mediante un muestreo aleatorio estratificado. Los criterios de estratificación se tomaron de los objetivos amplificados. Cada subprueba (correspondiente a cada uno de los objetivos) quedó constituída por cinco ítemes que es el límite inferior aceptable de ítemes que establece Berk (1979).

Las tres pruebas de cada asignatura fueron suministradas a grupos de estudiantes de primer ingreso de las universidades estatales así como de décimo año de la Educación Diversificada. Las instrucciones utilizadas para llevar a cabo la aplicación de las pruebas se pueden ver en el Anexo N°6.

En algunos de los grupos a los cuales se administró la prueba, se incluyó dentro del proceso de administración, una discusión con el grupo para recibir retroalimentación por parte de los estudiantes sobre, la claridad en la redacción de las preguntas, la comprensión del vocabulario y aspectos relacionados con el contenido y forma de las preguntas en general, con el fin de detectar ítemes defectuosos.

Establecimiento del puntaje de corte

Para el establecimiento del puntaje de corte se escogió un método empírico que se tenía disponible, que es el propuesto por Sheeham y Davis (1979), en el cual se emplea el puntaje aleatorio del ítem, multiplicado por el número de ítemes que mide determinado objetivo y a este valor se suman dos desviaciones estandar de la distribución de puntajes al azar.

Análisis de los ítemes

Congruencia ítem objetivo

Como se dijo anteriormente para llevar a cabo este análisis se empleó - la fórmula dada por Hambleton y Rovinelli (1977).

Dificultad:

Con el fin de revisar los ítemes, después de las aplicaciones de las - pruebas, se calculó la dificultad de los ítemes de ambas poblaciones (universitarios y estudiantes de Educación Diversificada), utilizando para ello la proporción de estudiantes que contestaron correctamente el ítem multi - plicada por cien.

Discriminación:

Para el análisis de discriminación se utilizó:

a) El índice B de Brennan (1972), con la muestra de estudiantes universitarios, ya que es el único índice que se puede aplicar con una sola administración de la prueba.

b) El índice de Klein y Kosecoff (1976) con una submuestra de los estudiantes universitarios como grupo instruido y la muestra de estudiantes de décimo año como grupo no instruido. La submuestra mencionada se obtuvo mediante un muestreo aleatorio y estratificado y tanto el tamaño de la sub - muestra como los criterios de estratificación utilizados se basaron en las características de la muestra de estudiantes de décimo año, en cuanto a -

./.

edad, sexo, tipo de colegio de procedencia, prueba realizada y año de egreso del colegio, para dentro de lo posible hacer comparables dichos grupos; en el Anexo N°7 se pueden ver los criterios de estratificación utilizados. Este índice fue escogido por su practicidad (Berk, 1980), y para fines investigativos ya que la diferencia entre los grupos criterio, es una limitación para su uso.

Selección y revisión de ítemes

Con los datos obtenidos en el análisis de ítemes (Discriminación) y el listado de frecuencias de respuesta, con que los estudiantes, escogieron las diferentes alternativas de respuesta del ítem, se clasificaron los ítemes en 3 grupos:

1) Buenos
2) Defectuosos
3) Descartados

Se consideraron como buenos aquellos ítemes con índice de discriminación superior a 30 y todas las alternativas de respuesta con una frecuencia de respuesta superior al 1%.

Se consideraron defectuosos los ítemes que cumplían con la primera condición pero no con la segunda y "descartados" los que no cumplían con la primera condición.

./.

Los ítemes defectuosos fueron revisados y se corrigieron aquellos en el que el número de alternativas de respuesta fueron cinco y de ellas sólo una tuviera una frecuencia de respuesta inferior al 1%, en cuyo caso se eliminó dicha alternativa. Los ítemes defectuosos que no se podían corregir - por no cumplir con las condiciones descritas fueron descartados (Berk 1980)

Con los ítemes resultantes de todo este proceso, se obtuvieron dos pruebas paralelas por asignatura mediante un muestreo aleatorio y estratificado, según los objetivos amplificados ya establecidos.

Además con fines de investigación, para cada una de las pruebas se construyeron dos fórmulas, en la primera se ordenaron los ítemes en orden ascendente de dificultad, según el índice correspondiente y en la segunda se hizo una selección aleatoria de los mismos.

Análisis de confiabilidad

Por falta de los recursos de computación necesarios y de grupos criterio (pre y pos-instrucción) el análisis de confiabilidad se hizo utilizando un método que se basara en una sola aplicación de la prueba (grupo instruido), para lo cual se utilizó el método propuesto por Huynh (1976), que utiliza el procedimiento de Keats y Lord (1962) para simular los puntajes de una segunda aplicación. Se prefirió este método al de Subkoviac (1976), que da resultados similares (Subkoviac, 1980), por basarse en un modelo matemático más fino.

Se obtuvieron también los índices 20 y 21 de Kuder Richardson que son recomendados en la literatura para aquellos casos en que no se tenga acceso a facilidades de cómputo sofisticado. (Downing y Mehrens (1978)).

./.

Otros análisis

A pesar de que las pruebas construidas no se utilizarán para tomar decisiones, con fines de investigación se calculó el índice propuesto por Hambleton (1978) (también escogido por su facilidad de cómputo) para evaluar la validez de decisión, tomando como grupos criterio (instruido y no instruido) los antes descritos. Los datos recolectados para la muestra de estudiantes universitarios presentaron suficiente variabilidad por lo que, también se utilizaron para hacer un análisis de normas.

Instrumentos

Escala para validar el dominio de conocimientos

Como parte del proceso de validación del dominio de conocimientos, la solicitud planteada a los profesores de la educación diversificada para valorar la importancia de los objetivos que se les presentaron, se hizo mediante una escala de cuatro puntos, los cuales se desglosan a continuación:

1. No tiene importancia no vale la pena exigirlo.
2. Tiene poca importancia; no es grave si el estudiante lo ignora.
3. Es importante vale la pena exigirlo.
4. Es imprescindible; todos los estudiantes deberían saberlo.

Dentro del documento se incluía un ejemplo de utilización de la escala y la lista de objetivos de la asignatura en cuestión.

Esta escala fue construida con la colaboración de dos especialistas en evaluación. (Anexo N°3).

./.

Guía para amplificar objetivos

Para la instrucción de los profesores que amplificarían los objetivos - se elaboró un pequeño documento confeccionado con base en los ejemplos de amplificación encontrados en la literatura revisada.

El documento consta de una breve descripción de lo que es un objetivo - amplificado y las consecuencias de su utilización, cómo está constituido un objetivo amplificado y una descripción del contenido de cada una de las - partes de que consta. (Anexo N°4).

Guía para analizar la calidad técnica de los ítemes

Para el análisis de calidad técnica de los ítemes por parte de los especialistas se utilizó la hoja de cotejo que aparece en el Anexo N°8, la - cual está constituida por tres partes; en una primera parte hay una serie de espacios, para incluir el objetivo y el ítem que se analizará, así como la identificación del revisor y la fecha. En la segunda parte se dan una serie de preguntas acerca de la calidad técnica del ítem y los espacios correspondientes para que el revisor marque la que considere pertinente. La tercera parte está destinada para las sugerencias y juicio final del revisor.

Instrucciones para la aplicación de las pruebas

Para la aplicación de las pruebas fue necesario elaborar un paquete de instrucciones tanto para el profesor encargado de aplicarlas como para los estudiantes.

./.

Las instrucciones para los profesores estan constituidas por una serie de normas para el desarrollo del proceso, que incluye las actividades previas a la aplicación, la aplicación en sí y las actividades posteriores a ella.

Las instrucciones para los estudiantes están constituidas por una serie de indicaciones para comunicar a los estudiantes el por qué y cómo se debe llevar a cabo el proceso.

Ambos paquetes de instrucciones tienen como objetivo garantizar la estandarización de las aplicaciones así como el establecimientos de los mecanismos de seguridad necesarios para mantener la confidencialidad de las pruebas.

Estas instrucciones fueron elaboradas tomando como base las instrucciones utilizadas por la Universidad de Costa Rica y el Instituto Tecnológico de Costa Rica en sus respectivos procesos de admisión. En el Anexo Nº6 se pueden ver las instrucciones utilizadas para cada una de las muestras (para la aplicación en la Educación Diversificada hubo necesidad de modificar algunos aspectos y se aprovechó para incluir sugerencias y observaciones surgidas de las primeras aplicaciones).

Análisis de datos

La respuesta al primer problema de investigación (¿Se puede afirmar que la prueba construida es una prueba con referencia a criterios?), se dio en la descripción del procedimiento que se empleó para construir las pruebas.

./.

La respuesta al segundo problema (¿Cómo se escogió y especificó el dominio de conocimientos que incluiría la prueba, para que existiera evidencia de su validez?), que se refiere a la validez de selección del dominio; (Popham, 1978) y a un elemento de la validez de contenido, según Hambleton (1980), se respondió también mediante la descripción de los procedimientos seguidos para la escogencia y especificación del dominio.

Para dar respuesta al tercer problema (¿Existe exactitud en la clasificación de los estudiantes como "másters" y "no másters" en un objetivo particular?) que refleja, según Berk (1976), la validez de decisión (constructo), que a su vez depende en gran medida del poder de discriminación que tengan los ítemes que se incluyeron en la prueba, o su validez, se utilizó el índice B de Brennan (1972) y el de Klein y Kosecoff (1976) y el método propuesto por Hambleton (1978) para evaluar la validez de decisión.

Para dar respuesta al cuarto problema que se refiere a la consistencia de la clasificación de los estudiantes como másters y no másters (confiabilidad) se utilizó el índice propuesto por Huynh (1976).

La respuesta al quinto problema (Los ítemes incluídos en la prueba ¿miden eficientemente el objetivo para el cual fueron escritos?) que se refiere a la congruencia ítem-objetivo y es la base de la validez de contenido según Hambleton (1978) (descriptiva según Popham (1978)), se buscó mediante los procedimientos descritos para revisar la calidad técnica de los ítemes y su representatividad así como mediante el juicio de especialistas, cuyos resultados fueron cuantificados utilizando el índice de Hambleton y Rovinelli (1977).

./.

Para dar respuesta al último problema que se refiere a la consistencia interna de cada subprueba (homogeneidad) se utilizaron los índices 20 y 21 de Kuder Richardson.

Para el análisis de normas se utilizó estadística descriptiva.

Los análisis se hicieron empleando programas de computadora confeccionados por funcionarios del Centro de Cómputo de la Universidad Nacional, de la Oficina de Planificación de la Educación Superior e Instituto de Investigaciones para el Mejoramiento de la Educación Costarricense.

CAPITULO IV

Análisis de resultados

En este capítulo se dan a conocer y se comentan los resultados obtenidos en cada una de las etapas del proyecto. Primero se analizará lo relacionado con el desarrollo de las pruebas y posteriormente lo relacionado con la prueba piloto (aplicación de las pruebas a sendas muestras de estudiantes).

Desarrollo de las pruebas

En este apartado se hará énfasis en los análisis de congruencia y calidad técnica, ya que los procesos de validación fueron descritos con detalle en la metodología.

En los cuadros números 1, 2, 3 y 4 se presentan los resultados del análisis de congruencia de los ítemes de cada una de las asignaturas, por objetivo. En tres de las cuatro asignaturas, el número de ítemes eliminados por falta de congruencia según el parámetro establecido de .75, fue menor o igual a un cinco por ciento del total de ítemes construidos para la asignatura en cuestión. Si se hace una revisión por objetivo, para esas asignaturas, se podrá ver que sólo en un caso (objetivo #1 de Ciencias) se eliminan 3 ítemes por incongruencia, en la mayoría de los casos se eliminan sólo 1, 2 o no se elimina ninguno.

En Estudios Sociales el número de ítemes eliminados corresponde a un 15%.

./.

CUADRO N°1

INDICE DE CONGRUENCIA DE LOS ITEMES DE

ESPAÑOL POR CATEGORIAS SEGUN OBJETIVOS

	INDICE DE CONGRUENCIA						
OBJETIVO N°	.74 Y MENOS	.75-.79	.80-.84	.85-.89	.90-.94	.95-100	N° DE ITEMES
1		1	-	2	2	15	20
2		-	2	-	6	14	22
3		-	-	1	1	18	20
4		-	2	1	5	13	21
5		-	1	3	11	8	23
6	-	-	1	4	-	17	22
7	1	7	2	11	-	-	21
8	-	-	3	6	2	10	21
9	-	4	5	2	1	7	19
10	1	1	2	4	11	3	22
11	1	6	1	2	4	11	25
12	2	2	1	-	3	12	20
TOTAL	5	21	20	36	46	128	256
%	1.9	8.2	7.8	14.1	18.0	50	100

CUADRO N°2

INDICE DE CONGRUENCIA DE LOS ITEMES DE

MATEMATICAS POR CATEGORIAS SEGUN OBJETIVOS

	INDICE DE CONGRUENCIA						
OBJETIVO N°	.74 y MENOS	.75-.79	.80-.84	.85-.89	.90-.94	.95-100	N° DE ITEMES
1	1	-	1	-	3	15	20
2	-	3	4	4	5	4	20
3	-	-	-	1	-	19	20
4	1	-	10	1	-	8	20
5	-	-	-	4	3	13	20
6	-	-	-	2	-	18	20
7	1	1	-	-	4	14	20
8	-	-	-	-	-	20	20
9	1	-	4	-	5	10	20
10	-	-	1	1	6	12	20
11	-	-	-	3	1	16	20
12	1	1	1	-	17	-	20
TOTAL	5	5	21	16	44	149	240
%	2.1	2.1	8.7	6.7	18.3	62.1	100

CUADRO Nº3

INDICE DE CONGRUENCIA DE LOS ITEMES DE CIENCIAS (BIOLOGIA, FISICA Y QUIMICA) POR CATEGORIAS SEGUN OBJETIVOS

	INDICE DE CONGRUENCIA						
OBJETIVO Nº	.74 y MENOS	.75-.79	.80-.84	.85-.89	.90-.94	.95-100	Nº DE ITEMES
1	3	6	2	2	2	6	21
2	2	2	2	1	2	11	20
3	1	8	1	1	4	5	20
4	2	10	3	4	2	-	21
TOTAL	8	26	8	8	10	22	82
%	9.7	31.7	9.7	9.7	12.2	26.8	100
5	-	2	-	1	4	13	20
6	-	2	5	1	8	4	20
7	2	4	-	4	-	10	20
8	-	1	2	4	3	10	20
TOTAL	2	9	7	10	15	37	80
%	2.5	11.2	8.7	12.5	18.8	46.2	100
9	2	13	2	3	-	-	20
10	-	1	-	1	-	18	20
11	-	-	-	1	3	16	20
12	-	-	-	6	2	12	20
TOTAL	2	14	2	11	5	46	80
%	2.5	17.5	2.5	13.7	6.2	57.5	100
TOTAL GENERAL	12	49	17	29	30	105	242
%	5	20.2	7	12	12.4	43.4	100

CUADRO N°4

INDICE DE CONGRUENCIA DE LOS ITEMES DE ESTUDIOS SOCIALES POR CATEGORIAS SEGUN OBJETIVOS

OBJETIVO N°	INDICE DE CONGRUENCIA						
	.74 Y MENOS	.75-.79	.80-.84	.85-.89	.90-.94	.95-100	N° DE ITEMES
1	-	5	2	6	-	7	20
2	7	6	-	3	-	4	20
3	2	4	2	5	-	8	21
4	4	9	7	1	1	-	22
5	3	-	-	6	1	10	20
6	3	6	4	5	1	1	20
7	4	8	6	2	-	-	20
8	2	3	4	5	-	6	20
9	5	6	6	3	-	-	20
10	3	8	5	2	1	1	20
11	1	5	2	11	-	1	20
12	2	3	2	8	-	5	20
TOTAL	36	63	40	57	4	43	243
%	14.8	2.6	16.5	23.4	1.6	17.7	100

CUADRO N°5

N° DE ITEMES ANTES Y DESPUES DEL ANALISIS DE CONGRUENCIA Y CALIDAD TECNICA POR ASIGNATURA SEGUN SUBPRUEBAS

OBJETIVO N°	ESPAÑOL			MATEMATICA			CIENCIAS			ESTUDIOS SOCIALES		
	N° IO	N° IF	%	N° IO	N° IF	%	N° IO	N° IF	%	N° IO	N° IF	%
1	20	20	100	20	19	95	21	17	81	20	19	95
2	22	20	91	20	19	95	20	17	85	20	13	65
3	20	20	100	20	20	100	20	17	85	21	14	66.7
4	21	21	100	20	19	95	21	17	81	22	14	63.6
5	23	22	95.7	20	17	85	20	20	100	20	16	80
6	22	20	91	20	20	100	20	20	100	20	14	70
7	21	20	95.2	20	19	95	20	18	90	20	16	80
8	21	20	95.2	20	19	95	20	20	100	20	16	80
9	19	19	100	20	19	95	20	18	90	20	15	75
10	22	17	77.2	20	18	90	20	19	95	20	17	85
11	25	24	96	20	15	75	20	19	95	20	16	80
12	20	18	90	20	19	95	20	20	100	20	18	90
TOTAL	256	241	94.1	240	223	93	242	222	92	243	188	77.4

NOTA: a) En el caso de Ciencias, los primeros cuatro objetivos corresponden a Biología, los segundos cuatro (del 5 al 8) a Física y los últimos a Química.

b) N° IO = Número de ítemes originales
N° IF = Número de ítemes finales

Esto podría significar que para el caso de esta asignatura, la amplificación de objetivos no fue la óptima para todos los objetivos.

En el cuadro N°5 se da, por objetivo y asignatura el número de ítemes - eliminados conjuntamente por la calidad técnica y la congruencia.

Es importante hacer notar que en todos los casos la eliminación de íte - mes, debido a su calidad técnica, fue mayor que la provocada por el análi - sis de congruencia.

Nuevamente es Estudios Sociales la asignatura que muestra mayor porcenta_ je de ítemes eliminados (23%) mientras que en las otras tres asignaturas, - la cantidad de ítemes eliminados corresponde a un porcentaje inferior o - igual al 8%.

Conclusión

Tanto por los procedimientos seguidos para la validación de las pruebas (descritos en el capítulo de la metodología), como por los resultados obte- nidos en los análisis de congruencia y calidad técnica se puede afirmar que existe evidencia de la validez de selección del dominio (Popham (1978)) y - de la validez de contenido (Hambleton (1980)) para cada una de las asignatu_ ras.

Prueba Piloto

En este apartado se da una descripción de las muestras de estudiantes a quienes se aplicaron las pruebas y de los resultados obtenidos en dichas -

./.

aplicaciones en cuanto a: puntaje de corte, índices de dificultad, discriminación, validez, confiabilidad y el análisis normativo de los resultados.

Descripción de las muestras

Muestra principal

La muestra principal estuvo constituida por 3.512 estudiantes y se describirá con base en algunas características como son sexo, nacionalidad, edad, tipo de colegio de procedencia, año de ingreso a la universidad, año de egreso de la educación secundaria, e institución en que está matriculado.

Dentro de cada una de las características anteriores, los 3.512 estudiantes muestreados se ubican en los siguientes valores:

- Sexo: 2.041 estudiantes muestreados son hombres (58,1%) y 1.471 mujeres - (41,9%).

- Nacionalidad: 3.389 estudiantes muestreados (96,5%) son costarricenses y 123 no lo son (3,5%).

- Edad: 1.586 estudiantes o un 37,8% de los muestreados tienen 17 años o menos, 1.206 (34,3%) entre 18 y 20 años, 398 (11,3%) entre 21 y 23 años, 137 (3,9%) entre 24 y 26 años, 52 (1,5%) entre 27 y 29 años, 49 (1,4%) entre 30 y 32 años, y 84 (2,4%) 33 y más años.

- Colegio de procedencia: la mayoría de los estudiantes proceden de colegios académicos diurnos oficiales (1.712 ó 49,99%), siguen en orden de importancia, los de colegios académicos diurnos particulares y semioficiales (774 ó - 22,60%), técnicos oficiales (570 ó 15,64%), académicos nocturnos oficiales (332 ó 9,69%), académicos nocturnos semioficiales y particulares (21 ó - 0,61%) y técnicos particulares (16 ó 0,47%).

./.

- Año de ingreso a la universidad: un 83,74% (2.941 estudiantes) de los estudiantes muestrados ingresó a la universidad en 1986, un 6,58% (231 estudiantes) en 1985, y un 9,68% (340 estudiantes) en 1984 o antes.

- Año de egreso de la educación secundaria: en 1985 se egresaron 2.298 de los estudiantes muestrados (65,4%); en 1984, 395 (11,3%) y antes de 1984, 819 (23,3%).

- Institución en que está matriculado: la mayor parte de los estudiantes muestrados están matriculados en la Universidad de Costa Rica (2.546 estudiantes ó 72,5%), siguen en importancia relativa los matriculados en la Universidad Nacional (636 estudiantes ó 18,1%) y los matriculados en el Instituto Tecnológico de Costa Rica (330 estudiantes ó 9,4%).

Muestra de estudiantes de IV ciclo de la Educación Diversificada

Esta muestra estuvo constituida por 711 estudiantes, procedentes de los siguientes tipos de colegio:

267 estudiantes de colegios académicos diurnos oficiales, 37,5%

133 estudiantes de colegios académicos diurnos particulares, 18,5%

229 estudiantes de colegios técnicos oficiales, 32%

82 estudiantes de colegios técnicos nocturnos, 12,5%

La distribución de estos estudiantes por sexo, edad y nacionalidad es la siguiente:

- Sexo: 411 de los estudiantes, un 58% eran varones y 300 (42%) eran mujeres

- Nacionalidad: 651 de los estudiantes, un 92% eran costarricenses y 60 es-

./.

tudiantes (8%) extranjeros.

- Edad: 308 estudiantes tenían 15 años (43%), 204 estudiantes tenían 16 años (29%), 151 estudiantes (21%) tenían de 17 a 20 años y 48 estudiantes eran - menores de 15 años y mayores de 20 años (7%).

Puntaje de corte

En el cuadro N°6 se pueden ver los puntajes de corte obtenidos para cada subprueba, fórmula y asignatura.

CUADRO N°6

PUNTAJES DE CORTE (C) POR ASIGNATURA, FORMULA Y SUBPRUEBAS

SUBPRUEBA N°	MATEMATICAS			CIENCIAS			ESPAÑOL			ESTUDIOS SOCIAL.		
	A	B	C	A	B	C	A	B	C	A	B	C
1	2.78	2.78	2.78	3.17	3.17	3.17	3.17	3.17	3.17	2.78	2.85	2.85
2	2.94	2.94	2.94	3.17	3.17	3.17	2.78	2.78	2.78	2.78	2.78	2.85
3	2.85	2.85	2.85	3.17	3.17	3.17	2.78	2.78	2.78	2.85	2.85	2.85
4	2.85	2.85	2.85	3.17	3.17	3.17	2.78	2.78	2.78	3.01	3.01	3.01
5	3.10	3.10	3.10	2.78	2.40*	2.78	2.78	2.40*	2.85	2.78	2.85	2.78
6	2.85	2.47*	2.47*	2.78	2.78	2.78	2.78	2.78	2.78	2.94	2.94	2.94
7	2.78	2.78	2.78	2.78	2.40*	2.78	3.17	3.17	3.17	3.01	3.01	2.94
8	2.85	3.04	2.85	2.78	1.98**	2.78	3.17	2.94	3.17	2.85	2.78	2.85
9	2.85	2.85	2.85	3.17	3.17	3.17	3.17	3.17	3.17	2.94	2.94	2.94
10	2.94	2.94	2.94	3.17	3.17	3.17	2.78	2.78	2.78	2.78	2.85	2.85
11	3.01	3.01	3.01	3.17	3.17	3.17	3.17	3.17	3.17	2.78	2.78	2.78
12	2.85	2.85	2.85	2.72*	3.17	3.17	3.17	3.17	3.17	2.78	2.85	2.78

* Las subpruebas correspondientes a estos puntajes de corte estaban formadas por 4 ítemes y no por 5 como la mayoría.

** La subprueba correspondiente a este puntaje de corte estaba formada por 3 ítemes solament

./.

Si no se toman en cuenta los casos anotados en el cuadro como excepciones, en cuanto al número de ítemes por subprueba (que es uno de los datos en que se basa el cálculo del puntaje de corte), se puede notar que en - Español, Ciencias y Matemáticas, los puntajes de corte de las tres fórmulas son iguales, no así para Estudios Sociales que tiene siete subpruebas en las que alguna de las fórmulas tiene un puntaje de corte diferente.

El cálculo del puntaje de corte utilizado se basa principalmente en el puntaje aleatorio del ítem, que a su vez es una función del número de opciones de respuesta de los ítemes. Partiendo de lo anterior, pueden explicarse las diferencias encontradas en Estudios Sociales ya que en esta asignatura se tuvo que hacer un muestreo estratificado de los ítemes, con un número muy limitado de estos para cada subprueba; esto obligó a poner, en alguna de las fórmulas, el conjunto de ítemes con diferente número de opciones de respuesta. Nótese que las variaciones se distribuyen entre las tres fórmulas (2 en A, 3 en B y 2 en la C).

Dificultad

Los índices de dificultad de los ítemes, por subprueba y asignatura, - obtenidos como resultado de la aplicación de las pruebas a los 3.512 estudiantes de primer ingreso en las universidades, se muestran en los cuadros numerados del 7 al 10.

En el cuadro N°7 se puede ver que todos los ítemes de Matemáticas tienen un índice de dificultad inferior a 70 (en una escala de 0 a 100) lo - cual, indica que ninguno de los ítemes resultó de fácil resolución para - la población a la que se aplicaron las pruebas.

./.

Por otra parte, del total de estos ítemes, un 71.3% muestran índices inferiores a 30 que se pueden interpretar como difíciles para esa población.

En Español (cuadro N°8), los porcentajes de ítemes con índices superiores e inferiores a 50 son 44.7% y 55.3% respectivamente. En este caso las proporciones de ítemes que resultaron "fáciles" y "difíciles" para la población fueron muy semejantes, agrupándose la mayoría (58%) en un rango intermedio (índices de 30 a 70).

CUADRO N°7

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE MATEMATICAS

POR CATEGORIAS DE INDICE DE DIFICULTAD SEGUN SUBPRUEBAS

		INDICE DE DIFICULTAD						
SUBPRUEBA N°	N° DE ITEMES	0-9	10-19	20-29	30-39	40-49	50-59	60-69
1	15	-	8	7	-	-	-	-
2	15	-	4	6	5	-	-	-
3	15	-	4	5	3	2	1	-
4	15	-	4	7	4	-	-	-
5	15	1	4	2	6	2	-	-
6	13	-	-	4	5	3	1	-
7	15	-	8	6	1	-	-	-
8	15	1	6	7	1	-	-	-
9	15	1	3	7	4	-	-	-
10	15	-	2	5	7	1	-	-
11	15	1	7	6	-	-	-	1
12	15	-	5	6	2	1	1	-
TOTAL	178	4	55	68	38	9	3	1
%	99.9	2.2	30.9	38.2	21.3	5.1	1.6	.56

CUADRO N°8

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE ESPAÑOL

POR CATEGORIAS DE INDICE DE DIFICULTAD SEGUN SUBPRUEBAS

		INDICE DE DIFICULTAD									
SUBPRUEBA N°	N° DE ITEMES	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-100
1	15	1	-	-	-	2	2	4	3	1	2
2	15	-	1	1	2	4	2	3	2	-	-
3	15	1	-	4	2	1	4	-	1	2	-
4	15	1	2	1	1	5	3	1	1	-	-
5	14	-	-	2	1	4	4	2	1	-	-
6	15	-	-	1	-	1	1	3	2	4	3
7	15	-	1	3	4	6	1	-	-	-	-
8	15	-	-	-	1	-	2	5	2	4	1
9	15	-	-	1	-	3	5	-	3	3	-
10	15	-	-	1	2	4	3	2	1	2	-
11	15	1	1	1	2	1	3	4	1	1	-
12	15	1	2	6	-	1	2	-	1	1	1
TOTAL	179	5	7	21	15	32	32	24	18	18	7
%	99.9	2.8	3.9	11.7	8.4	17.9	17.9	13.4	10.0	10.0	3.9

CUADRO Nº9

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE CIENCIAS

POR CATEGORIAS DE INDICE DE DIFICULTAD SEGUN SUBPRUEBAS

		INDICE DE DIFICULTAD									
SUBPRUEBA Nº	Nº DE ITEMES	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-100
1	15	-	-	3	6	3	3	-	-	-	-
2	15	-	3	2	4	2	3	-	-	1	-
3	15	-	1	-	3	7	2	1	1	-	-
4	15	-	1	2	2	1	3	3	2	1	-
TOTAL	60	-	5	7	15	13	11	4	3	2	-
%	100	-	8.3	11.7	25.0	21.7	18.3	6.7	5.0	3.3	-
5	14	-	1	3	5	3	-	2	-	-	-
6	15	-	-	-	-	3	3	3	3	3	-
7	14	6	4	1	-	1	2	-	-	-	-
8	13	-	1	4	4	4	-	-	-	-	-
TOTAL	56	6	6	8	9	11	5	5	3	3	-
%	100	10.7	10.7	14.3	16.1	19.6	8.9	8.9	5.4	5.4	-
9	15	-	-	5	3	2	3	2	-	-	-
10	15	-	4	7	4	-	-	-	-	-	-
11	15	-	5	8	2	-	-	-	-	-	-
12	14	1	2	5	5	1	-	-	-	-	-
TOTAL	59	1	11	25	14	3	3	2	-	-	-
%	100	1.7	18,6	42.4	23.7	5.1	5.1	3.4	-	-	-
TOTAL GENERAL	175	7	22	40	38	27	19	11	6	5	-
%	99.9	4.0	12.6	22.9	21.7	15.4	10.8	6.3	3.4	2.8	-

NOTA: Las 4 primeras subpruebas corresponden a Biología, las 4 siguientes (de la 5 a la 8) a Física y las últimas a Química.

En el caso de Ciencias, expuesto en el cuadro Nº9, aproximadamente tres cuartas partes (76.6%) de los ítemes resultaron con índices de dificultad inferiores a 50.

Si vemos las subpruebas correspondientes a Biología y Física podemos notar que la proporción de ítemes con índices inferiores a 50 disminuye un poco (67 y 61% respectivamente), mientras que en Ouímica se da una situación similar a la de Matemáticas ya que no hay ítemes con índices superiores a 69, aunque en este caso la proporción de ítemes, que se pueden calificar como difíciles para esa población (índices de 30 o menor) es un poco menor (62.5%).

Nótese que en ninguna de las subpruebas de esta asignatura hay ítemes que hayan resultado totalmente fáciles para la población que realizó las pruebas. (No se ubica ningún ítem en categoría de 90 a 100).

En Estudios Sociales también se da (cuadro Nº10) que la mayoría de los ítemes (63%) tienen índices inferiores a 50; en este caso esta proporción es menor que en Matemática y Ciencias y además si hay ítemes, aunque pocos (1.1%) que se ubicaron en la última categoría.

Se puede concluir que en todas las asignaturas los ítemes tendieron a resultar "difíciles", para la población escogida como muestra, para la prueba piloto, especialmente en el caso de Mátemáticas y Ouímica.

CUADRO Nº10

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE ESTUDIOS SOCIALES

POR CATEGORIAS DE INDICE DE DIFICULTAD SEGUN SUBPRUEBAS

		INDICE DE DIFICULTAD									
SUBPRUEBA Nº	Nº DE ITEMES	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-100
1	15	1	-	2	2	2	3	1	-	2	1
2	15	2	1	4	3	4	1	-	-	-	-
3	15	-	1	1	-	-	4	1	2	5	1
4	15	-	-	1	3	1	2	6	1	1	-
5	15	2	6	1	1	3	1	-	-	1	-
6	15	-	4	1	2	1	3	4	-	-	-
7	15	-	2	3	5	2	2	-	1	-	-
8	15	-	2	5	2	2	-	1	3	-	-
9	15	-	2	3	1	1	-	2	4	2	-
10	15	-	-	3	4	5	2	-	1	-	-
11	15	-	5	4	1	3	-	-	-	2	-
12	15	-	2	3	3	2	2	2	1	-	-
TOTAL	180	5	25	31	26	26	20	19	13	13	2
%	99.9	2.8	13.9	17.2	14.4	14.4	11.1	10.6	7.2	7.2	1.1

Discriminación

Berk (1.980) indica que en pruebas con referencia a criterios, es suficiente con que el índice de discriminación del ítem sea positivo, para su selección. Sin embargo ya que es un criterio que refleja la calidad del ítem, se trató en lo posible seleccionar aquellos ítemes con índices de discriminación superiores a 45, que es un criterio que se puede considerar conservador. En aquellos casos en que las especificaciones de los objetivos amplificados, hicieron necesario incluir ítemes con índices menores, se utilizó un valor menos conservador de 30 como límite.

./.

En los cuadros del N°11 al N°14 se pueden ver los índices de discriminación por subprueba y asignatura obtenidos mediante la fórmula dada por Brennan (1972) (que se calcula con los datos de una sola administración de la prueba).

La asignatura con mayor porcentaje de ítemes con índices superiores al límite conservador (45) fue Ciencias con un 76.5% (cuadro N°13), le sigue con un 71.4% Matemáticas (cuadro N°11), después está Español (cuadro N°12) con un 44.2% y finalmente Estudios Sociales con un 36.6% (cuadro N°14).

Si se toma en cuanta el límite menos conservador de 30 el orden de importancia relativa de las asignaturas cambia así: Matemáticas (93.3%), Ciencias (92%), Estudios Sociales (79.4%) y Español con un 78.2%.

Se puede concluir que la calidad de los ítemes en cuanto a discriminación se refiere, fue alta ya que por un lado no hubo ítemes con índice de discriminación negativo y, por otra parte, en todas las asignaturas la proporción de ítemes con índices superiores al límite (no conservador) fue superior a las tres cuartas partes del total de ítemes.

El índice de discriminación fue calculado también utilizando el método de Klein y Kosecoff (1976) que requiere de los datos de dos grupos criterio: instruido y no instruido. Como se dijo en la metodología este índice se escogió por su practicidad y para fines investigativos ya que la diferencia existente entre los grupos criterio, es una limitación para su uso e interpretación. En los cuadros del N°15 al N°18 se muestran los resultados obtenidos por subprueba y asignatura. En este caso un índice de discriminación igual a cero puede interpretarse como falta de conocimiento de la asignatura en ambos grupos (Berk 1980).

./.

CUADRO N°11

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE MATEMATICAS

POR CATEGORIAS DE INDICE DE DISCRIMINACION SEGUN SUBPRUEBAS

		INDICE DE DISCRIMINACION												
SUBPRUEBA N°	N° DE ITEMES	0-29	30-35	36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80	81-85	86-90
1	15	2	2	1	2	2	2	—	—	2	1	—	1	—
2	15	2	1	—	1	1	4	4	2	—	—	—	—	—
3	15	1	2	—	5	—	2	3	1	1	—	—	—	—
4	15	—	—	2	1	3	2	2	2	1	2	—	—	—
5	15	1	1	—	1	1	5	—	2	2	1	—	—	1
6	13	—	—	—	—	4	3	3	3	—	—	—	—	—
7	15	—	—	1	2	2	2	3	2	1	2	—	—	—
8	15	2	1	2	2	3	1	—	—	1	1	1	1	—
9	15	2	—	—	3	2	1	1	3	2	1	—	—	—
10	15	1	1	1	—	3	3	3	2	1	—	—	—	—
11	15	1	—	2	1	—	2	—	—	1	2	2	3	1
12	15	—	3	1	—	2	1	4	4	—	—	—	—	—
TOTAL	178	12	11	10	18	23	28	23	21	12	10	3	5	2
%	99.8	6.7	6.2	5.6	10.1	12.9	15.7	12.9	11.8	6.7	5.6	1.7	2.8	1.1

CUADRO N°12

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE ESPAÑOL

POR CATEGORIAS DE INDICE DE DISCRIMINACION SEGUN SUBPRUEBAS

SUBPRUEBA N°	N° DE ITEMES	INDICE DE DISCRIMINACION										
		0-29	30-35	36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80
1	15	6	1	-	1	4	3	-	-	-	-	-
2	15	2	2	2	3	2	1	2	1	-	-	-
3	15	4	3	-	1	3	1	1	-	2	-	-
4	15	3	2	1	1	3	3	-	2	-	-	-
5	14	2	3	-	-	1	4	3	1	-	-	-
6	15	3	-	4	1	2	-	2	1	1	-	1
7	15	1	1	1	1	3	3	3	2		-	-
8	15	1	3	-	8	1	1	1	-	-	-	-
9	15	5	3	1	2	2	1	-	-	1	-	-
10	15	3	1	2	3	3	2	1	-	-	-	-
11	15	4	-	5	1	1	2	-	2	-	-	-
12	15	5	-	2	2	-	1	-	1	3	-	1
TOTAL	179	39	19	18	24	25	22	13	10	7	-	2
%	100	21.8	10.6	10.0	13.4	14.0	12.3	7.3	5.6	3.9	0	1.1

CUADRO N°13

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE CIENCIAS

POR CATEGORIAS DE INDICE DE DISCRIMINACION SEGUN SUBPRUEBAS

		INDICE DE DISCRIMINACION											
SUBPRUEBA N°	N° DE ITEMES	0-29	30-35	36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80	81-85
1	15	1	1	1	2	2	1	4	2	1	-		
2	15	2	-	1	1	5	1	1	2	1	1	-	
3	15	1	3	2	2	2	-	3	2	-	-	-	
4	15	3	-	2	2	2	3	2	1	-	-	-	
TOTAL	60	7	4	6	7	11	5	10	7	2	1		
%	99.8	11.7	6.7	10.0	11.7	18.3	8.3	16.7	11.7	3.3	1.7		
5	14	-		-	1	5	1	3	1	3	1	-	
6	15	2		1	2	4	1	2	2	2	-	-	
7	14	1		-	-	1	-	2	2	-	5	2	
8	13	-		-	-	3	1	1	3	4	-	-	
TOTAL	56	3	-	1	3	13	3	8	8	9	6	2	
%	99.9	5.3	-	1.8	5.3	23.2	5.3	14.3	14.3	16.1	10.7	3.6	
9	15	-	-		3	1	4	3	3	-	-	1	-
10	15	-	-		1	2	1	-	2	3	5	1	-
11	15	2	1		-	1	-	4	1	1	3	1	1
12	14	2	-		1	-	1	1	4	2	2	1	-
TOTAL	59	4	1	-	5	4	6	8	10	6	10	4	1
%	100	6.8	1.7	-	8.5	6.8	10.2	13.5	16.9	10.2	16.9	6.8	1.7
TOTAL GENERAL	175	14	5	7	15	28	14	26	25	17	17	6	1
%	99.9	8.0	2.8	4.0	8.6	16.0	8.0	14.8	14.3	9.7	9.7	3.4	.6

NOTA: Las 4 primeras subpruebas corresponden a Biología, las cuatro siguientes (de la 5 a la 8) a Física y las últimas a Química.

CUADRO N°14

DISTRIBUCION DE LOS ITEMES DE LAS PRUEBAS DE ESTUDIOS SOCIALES

POR CATEGORIAS DE INDICES DE DISCRIMINACION SEGUN SUBPRUEBAS

		INDICE DE DISCRIMINACION							
SUBPRUEBA N°	N° DE ITEMES	0-29	30-35	36-40	41-45	46-50	51-55	56-60	61-65
1	15	4	3	1	2	3	1	1	-
2	15	2	-	2	3	1	5	1	1
3	15	3	1	2	-	5	3	-	1
4	15	2	2	6	2	1	2	-	-
5	15	5	1	-	-	3	3	1	2
6	15	3	3	4	2	3	-	-	-
7	15	2	1	2	1	4	1	4	-
8	15	3	2	2	3	3	1	1	-
9	15	4	5	1	5	-	-	-	-
10	15	2	2	1	4	4	2	-	-
11	15	4	1	1	3	3	3	-	-
12	15	3	2	3	4	1	1	1	-
TOTAL	180	37	23	25	29	31	22	9	4
%	100	20.6	12.8	13.9	16.1	17.2	12.2	5.0	2.2

./.

CUADRO N°15

INDICE DE DISCRIMINACION (KLEIN y KOSECOFF) DE LAS PRUEBAS

DE MATEMATICAS POR CATEGORIAS SEGUN SUBPRUEBAS

		INDICE DE DISCRIMINACION									
SUBPRUEBA N°	N° DE ITEMES	< 0	0	.01-.05	.06-.10	.11-.15	.16-.20	.21-.25	.26-.30	.31-.35	.36 o >
1	15	2	2	7	1	2	-	1	-	-	-
2	15	1	-	7	3	1	-	-	1	2	-
3	15	-	1	6	6	1	-	-	1	-	-
4	15	-	-	8	3	-	1	3	-	-	-
5	15	-	4	4	7	-	-	-	-	-	-
6	13	-	-	-	6	2	-	3	1	-	1
7	15	1	-	5	6	2	1	-	-	-	-
8	15	1	1	8	3	-	1	1	-	-	-
9	15	2	1	4	5	1	1	-	1	-	-
10	15	-	1	4	3	4	1	1	1	-	-
11	15	1	2	7	2	3	-	-	-	-	-
12	15	-	-	5	6	2	1	-	-	1	-
TOTAL	178	8	12	65	51	18	6	9	5	3	1
%	100	4.5	6.7	36.5	28.6	10.1	3.4	5.1	2.8	1.7	.6

CUADRO N°16

INDICE DE DISCRIMINACION (KLEIN Y KOSECOFF) DE LAS PRUEBAS DE ESPAÑOL POR CATEGORIAS SEGUN SUBPRUEBAS

	INDICE DE DISCRIMINACION			
SUBPRUEBA N°	< 0	0	.01-.05	N° DE ITEMES
1	1	4	10	15
2	1	4	10	15
3	-	3	12	15
4	1	8	6	15
5	-	2	12	14
6	-	6	9	15
7	-	4	11	15
8	-	4	11	15
9	1	5	9	15
10	1	4	10	15
11	-	7	8	15
12	1	6	8	15
TOTAL	6	57	116	179
%	3.4	31.8	64.8	100

CUADRO N°17

INDICE DE DISCRIMINACION (KLEIN Y KOSECOFF) DE LAS PRUEBAS DE CIENCIAS POR CATEGORIAS SEGUN SUBPRUEBAS

	INDICE DE DISCRIMINACION				
SUBPRUEBA N°	< 0	0	.01-.05	.06-.09	N° DE ITEMES
1	-	2	11	2	15
2	2	4	9	-	15
3	-	1	13	1	15
4	-	1	13	1	15
5	-	2	9	3	14
6	-	-	8	7	15
7	-	7	6	1	14
8	-	1	11	1	13
9	-	3	11	1	15
10	1	3	11	-	15
11	-	7	8	-	15
12	1	1	11	1	14
TOTAL	4	32	121	18	175
%	2.28	18.3	69.1	10.3	100

CUADRO N°18

INDICE DE DISCRIMINACION (KLEIN Y KOSECOFF) DE LAS PRUEBAS DE ESTUDIOS SOCIALES POR CATEGORIAS SEGUN SUBPRUEBAS

SUBPRUEBA N°	<0	0	.01 - .05	.06 - .09	.1 - .15	N°ITEMES
1	5	5	2	1	2	15
2	7	4	3	1	-	15
3	8	1	6	-	-	15
4	5	3	4	2	1	15
5	5	8	1	1	-	15
6	5	5	5	-	-	15
7	3	4	8	-	-	15
8	3	2	8	2	-	15
9	6	2	4	1	2	15
10	6	1	7	1	-	15
11	4	6	5	1	-	15
12	5	3	5	-	2	15
TOTAL	62	43	58	10	7	180
%	34.4	23.9	32.2	5.5	3.9	100

Los ítemes con índices de discriminación inferiores a cero no sobrepasaron el 5% para Matemáticas (cuadro N°15), Español (cuadro N°16) y Ciencias (cuadro N°17), mientras que en el caso de Estudios Sociales fue de un 34.4%.

El orden de importancia relativo de las asignaturas en cuanto a los ítemes de discriminación positivo es decir ítemes que podrían ser seleccionados para formar las pruebas finales (según Berk -1980-) es el siguiente: Matemáticas con un 89%, Ciencias con un 79%, Español con 65% y finalmente Estudios Sociales con sólo un 42%.

Validez de decisión

Se utilizó el índice propuesto por Hambleton (1978). Se debe insistir en que:

a) Este índice se obtuvo solamente con fines de investigación pues las pruebas no van a ser utilizadas para tomar decisiones.

b) El cálculo del índice requirió de los datos de dos grupos criterio (como en el caso del índice de discriminación de Klein y Kosecoff) y estos grupos muestran diferencias que impiden hacer una interpretación rigurosa.

c) El índice de validez de decisión, que da una medida del acierto que se tuvo en clasificar a los estudiantes como "masters" y no "masters", se aplica generalmente a un proceso de instrucción sistematizado. En el caso que nos ocupa no se puede considerar que se está realizando una medición de un proceso sistematizado pues las características de los estudian

./.

tes a quienes se aplicaron las pruebas (en cuanto a instrucción se refiere) son muy heterogéneas. Esto se puede afirmar desde el momento en que se conoce que proceden de diferentes colegios y zonas geográficas, que recibieron los cursos de diferente manera, etc.

En el cuadro N°19, se muestran los resultados obtenidos por subprueba, fórmula y asignatura.

CUADRO N°19

INDICE DE VALIDEZ DE DECISION(*) DE LAS PRUEBAS POR FORMULA SEGUN SUBPRUEBAS

SUBPRUEBA N°	MATEMATICAS			CIENCIAS			ESPAÑOL			ESTUDIOS SOC.		
	A	B	C	A	B	C	A	B	C	A	B	C
1	.51	.54	.49	.58	.53	.56	.51	.56	.60	.58	.64	.40
2	.65	.54	.53	.54	.54	.50	.50	.64	.61	.58	.44	.82
3	.53	.55	.58	.58	.56	.54	.64	.51	.66	.57	.47	.59
4	.60	.54	.56	.61	.54	.69	.53	.55	.52	.64	.46	.91
5	.49	.52	.49	.61	.52	.59	.64	.64	.61	.50	.33	.92
6	.66	.57	.62	.74	.67	.67	.63	.53	.53	.61	.47	.80
7	.52	.55	.53	.63	.49	.49	.53	.53	.53	.60	.43	.88
8	.53	.51	.48	.56	.55	.57	.59	.64	.65	.71	.51	.91
9	.57	.51	.54	.62	.51	.53	.55	.53	.61	.65	.62	.68
10	.60	.62	.58	.53	.49	.50	.54	.60	.62	.56	.42	.86
11	.52	.51	.49	.51	.49	.51	.57	.53	.52	.55	.48	.92
12	.60	.54	.58	.59	.55	.52	.50	.55	.47	.63	.49	.89

(*) La validez de decisión según Hambleton es la sumatoria de las decisiones correctas en cada grupo, es decir la suma de la proporción de estudiantes de la muestra de universitarios con $x > C$ y de la proporción de estudiantes de la muestra de estudiantes de secundaria con $x < C$. (Hambleton, 1978).

Tomando en consideración las limitaciones antes descritas se puede ver que los índices obtenidos son altos por lo menos en tres de las asignaturas: en Matemáticas y Ciencias todos los índices son superiores o iguales a .49. En Español, excepto por un caso, todas las subpruebas muestran índices superiores o iguales a .50. Estudios Sociales es la asignatura que muestra índices más bajos y aún en este caso tres cuartas partes de las subpruebas muestran índices superiores o iguales a .50.

Confiabilidad

Para medir la confiabilidad se utilizó el método propuesto por Huynn (1976) que se basa en los datos de una única aplicación de las pruebas. Los índices K (Kappa) resultantes para cada subprueba, fórmula y asignatura se muestran en los cuadros numerados del 20 al 23.

Antes de realizar el análisis de los índices obtenidos debe recordarse que:

1.- El coeficiente K pretende medir la consistencia en la clasificación de los estudiantes como "masters" o como "no masters" en la aplicación repetida de una misma prueba o en su efecto con los datos de una única aplicación más la simulación, mediante un modelo matemático, de los datos de una segunda aplicación, que es el caso que nos ocupa.

Una de las condiciones para poder llevar a cabo esa simulación es, entre otras cosas, que los ítemes tengan igual dificultad y en caso de no poder lograrlos se producen estimaciones conservadoras de la confiabilidad (Berk, 1980).

./.

CUADRO Nº20

INDICE DE CONFIABILIDAD (KAPPA) Y OTROS PARAMETROS

DE LAS PRUEBAS DE ESPAÑOL POR SUBPRUEBAS SEGUN FORMULA

SUBPRUEBA Nº	PUNT. DE CORTE (C)			MEDIA			DESV. EST.			VARIANZA			R21			KAPPA (K)		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
1	3.17	3.17	3.17	2.94	3.39	3.26	1.00	1.21	1.05	1.00	1.47	1.12	-.26	.32	-.01	.48	.97	.77
2	2.78	2.78	2.78	1.91	3.02	2.37	1.11	1.39	1.29	1.24	1.95	1.67	.06	.48	.31	.71	.54	.06
3	2.78	2.78	2.78	2.59	1.67	2.58	1.27	1.03	1.35	1.63	1.07	1.83	.29	-.04	.39	.65	-.03	.49
4	2.78	2.78	2.78	1.80	1.94	2.55	1.19	1.35	1.21	1.44	1.83	1.48	.25	.43	.19	.34	.62	.26
5	2.78	2.40	2.85	2.33	1.93	2.63	1.31	1.14	1.21	1.74	1.31	1.47	.35	.31	.18	-.07	.93	.71
6	2.78	2.78	2.78	3.41	3.65	3.87	1.16	1.14	.80	1.36	1.32	.65	.25	.31	-.43	-.35	-.37	.41
7	3.17	3.17	3.17	1.90	1.83	1.92	1.17	1.27	1.26	1.38	1.63	1.61	.18	.36	.33	.99	.75	0
8	3.17	2.94	3.17	4.34	3.20	3.10	.89	1.26	1.17	.81	1.60	1.39	.36	.35	.19	.25	-.49	.39
9	3.17	3.17	3.17	3.09	2.53	3.18	1.10	1.20	.82	1.21	1.45	.68	.03	.17	-.41	.49	.49	.83
10	2.78	2.78	2.78	2.56	2.76	2.72	1.36	1.25	1.10	1.85	1.58	1.22	.40	.27	-.02	.39	-.48	.76
11	3.17	3.17	3.17	2.71	2.10	2.62	1.11	1.21	1.00	1.25	1.48	1.02	0	.22	-.27	*	.56	0
12	3.17	3.17	3.17	1.91	2.73	1.14	1.00	.91	.94	1.01	.83	.90	.21	.61	.02	.39	.53	.82

CUADRO N°21

INDICE DE CONFIABILIDAD (KAPPA) Y OTROS PARAMETROS

DE LAS PRUEBAS DE MATEMATICAS POR SUBPRUEBAS SEGUN FORMULA

SUBPRUEBA	PUNT. DE CORTE (C)			MEDIA			DESV. EST.			VARIANZA			R21			KAPPA (K)		
N°	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
1	2.78	2.78	2.78	.90	1.02	.76	.87	1.08	.88	.77	1.17	.78	.05	.38	.21	-.70	.85	.02
2	2.94	2.94	2.94	1.47	.97	1.44	1.25	.97	1.28	1.57	.95	1.65	.42	.22	.47	.74	.66	.64
3	2.85	2.85	2.85	1.29	1.40	1.53	1.02	1.15	1.20	1.06	1.34	1.46	.12	.30	.34	-.26	.34	.74
4	2.85	2.85	2.85	1.18	1.09	1.44	1.28	1.12	1.33	1.66	1.27	1.79	.57	.41	.53	-.39	.82	.27
5	3.10	3.10	3.10	1.60	1.43	.99	1.09	1.07	1.07	1.19	1.16	1.15	.10	.14	.38	.29	.68	.07
6	2.85	2.47	2.47	1.83	1.35	1.54	1.48	1.11	1.23	2.20	1.25	1.52	.59	.37	.50	.72	.71	.59
7	2.78	2.78	2.78	.83	1.25	.99	.93	1.28	1.14	.87	1.66	1.30	.25	.54	.48	.75	-.47	.79
8	2.85	3.04	2.85	1.24	.98	.95	1.02	.98	.90	1.06	1.98	.82	.15	.24	.07	.27	.99	-.64
9	2.85	2.85	2.85	1.29	1.11	1.30	1.28	1.06	1.29	1.65	1.14	1.67	.52	.30	.52	-.74	.87	-.74
10	2.94	2.94	2.94	1.34	1.79	1.33	1.18	1.43	1.23	1.41	2.07	1.52	.38	.55	.44	-.15	-.66	.72
11	3.01	3.01	3.01	1.48	.90	.96	1.03	.92	.94	1.07	.86	.89	.03	.17	.16	.52	.18	.13
12	2.85	2.85	2.85	1.39	1.04	1.47	1.15	1.11	1.32	1.33	1.25	1.76	.30	.42	.51	.88	.81	.61

CUADRO N°22

INDICE DE CONFIABILIDAD (KAPPA) Y OTROS PARAMETROS

DE LAS PRUEBAS DE CIENCIAS POR SUBPRUEBAS SEGUN FORMULA

SUBPRUEBA N°	PUNT. DE CORTE (C)			MEDIA			DESV. EST.			VARIANZA			R21			KAPPA (K)		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
1	3.17	3.17	3.17	2.13	1.64	2.09	1.23	1.26	1.21	1.52	1.59	1.48	.24	.38	.22	.53	.33	.42
2	3.17	3.17	3.17	1.61	2.30	1.84	1.14	1.04	1.19	1.31	1.09	1.42	.20	-.17	.22	.13	.22	.70
3	3.17	3.17	3.17	2.42	2.13	2.20	1.24	1.20	1.21	1.54	1.45	1.48	.23	.19	.20	.16	.98	.78
4	3.17	3.17	3.17	2.72	1.91	2.99	1.09	1.17	1.30	1.20	1.37	1.69	-.04	.17	.36	.23	.91	.93
5	2.78	2.40	2.78	1.62	2.04	1.56	1.51	1.26	1.45	2.31	1.59	2.12	.65	.49	.61	.95	.74	.11
6	2.78	2.78	2.78	3.15	3.79	2.71	1.44	1.28	1.49	2.10	1.64	2.24	.55	.55	.55	0	.93	.21
7	2.78	2.40	2.78	1.74	.60	.41	1.17	.78	.92	1.38	.61	.85	.22	.21	.69	.09	.62	.91
8	2.78	1.98	2.78	1.88	.85	1.49	1.41	.88	1.47	1.99	.78	2.19	.51	.32	.65	-.38	.41	-.45
9	3.17	3.17	3.17	2.84	1.38	1.86	1.63	1.24	1.25	2.66	1.54	1.57	.67	.43	.32	.53	.53	.83
10	3.17	3.17	3.17	1.48	1.04	1.17	1.32	1.12	1.07	1.76	1.26	1.15	.51	.43	.27	.37	.71	.41
11	3.17	3.17	3.17	1.01	1.15	1.33	1.05	1.05	1.09	1.11	1.11	1.20	.34	.25	.23	.47	.78	.90
12	2.72	3.17	3.17	1.08	1.48	1.25	1.00	1.21	1.13	1.00	1.48	1.29	.28	.37	.34	.28	.19	.21

CUADRO Nº23

INDICE DE CONFIABILIDAD (KAPPA) Y OTROS PARAMETROS DE LAS PRUEBAS DE ESTUDIOS SOCIALES POR SUBPRUEBAS SEGUN FORMULA

SUBPRUEBA Nº	PUNT. DE CORTE (C)			MEDIA			DESV. EST.			VARIANZA			R21			KAPPA (K)		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
1	2.78	2.85	2.85	1.91	2.85	3.07	1.06	1.14	.93	1.13	1.31	0.87	-.05	.08	-.45	.51	.59	.36
2	2.78	2.78	2.85	1.65	1.59	1.61	1.35	1.04	1.02	1.83	1.09	1.06	.49	1.0	-.03	.40	.56	-.12
3	2.85	2.85	2.85	3.46	3.49	2.84	1.15	1.00	1.03	1.33	1.02	1.08	.24	-.04	-.17	-.54	.87	-.36
4	3.01	3.01	3.01	3.17	2.55	2.62	1.20	1.14	1.19	1.45	1.31	1.43	.24	.05	.15	.86	.92	.36
5	2.78	2.85	2.78	1.38	1.90	1.07	1.0	.93	.90	1.02	.88	.82	.02	-.42	-.03	-.34	.16	-.50
6	2.94	2.94	2.94	2.26	2.14	1.90	1.11	1.18	.97	1.25	1.41	.96	.01	.16	-.28	.86	.60	.25
7	3.01	3.01	2.94	2.15	1.67	1.60	1.07	1.04	1.01	1.16	1.09	1.04	-.07	-.02	-.05	.92	.28	.79
8	2.85	2.78	2.85	2.38	1.89	1.78	1.17	1.18	1.10	1.37	1.40	1.22	.11	.20	.07	.77	.87	.57
9	2.94	2.94	2.94	2.08	2.85	2.87	.94	1.00	1.03	.90	1.02	1.08	-.43	-.25	-.16	.53	.28	-.07
10	2.78	2.85	2.85	2.07	2.33	1.78	1.22	1.23	1.12	1.49	1.53	1.26	.23	.23	.11	.80	.77	.81
11	2.78	2.78	2.78	1.53	1.66	1.71	1.09	.94	.95	1.20	.89	.91	.14	-.30	-.29	.72	.50	0
12	2.78	2.85	2.78	2.41	2.23	1.78	1.10	1.13	1.16	1.22	1.28	1.36	-.02	.04	.19	.70	.28	.81

2.- Cuando los puntajes no muestran suficiente variabilidad, el valor de R_{21} (Kuder Richardson) puede ser cero o negativo. Si es negativo el valor de K también puede serlo, en cuyo caso debe reemplazarse por el valor más pequeño positivo que se haya estimado para la confiabilidad (Huynn, 1978). En los cuadros antes mencionados no se realizó ese reemplazo.

3.- El coeficiente es afectado por el puntaje de corte, la heterogeneidad de los puntajes y la longitud o número de ítemes de la subprueba respectiva.

Por lo anotado anteriormente se deduce la dificultad que existe en hacer una interpretación unidireccional de los índices obtenidos. Si puede notarse en los cuadros que la consistencia en la clasificación de los estudiantes, para una subprueba y asignatura particular, varía de una fórmula que muestra una alta consistencia (superior o igual a 70) y también en la mayoría de ellas hay por lo menos dos fórmulas que muestran una consistencia superior o igual a 50. Ya que la calidad (discriminación, congruencia, dificultad, etc.) de los ítemes que conforman cada subprueba es lo que le da a la misma su poder de discriminación entre "masters" y "no masters" y dados los resultados obtenidos en esas características (ya detallados anteriormente) se puede afirmar que los índices de confiabilidad son congruentes con ellos.

Análisis normativo

En el Anexo #9 se encuentran los resultados normativos obtenidos en la calificación de las pruebas, por subpruebas y asignaturas.

CAPITULO V

Conclusiones y recomendaciones

Discusión

Es importante una vez concluido el proyecto, hacer una serie de reflexiones sobre las decisiones metodológicas en que se basó el mismo, que permitan juzgar adecuadamente, la labor realizada.

Un sistema de enseñanza-aprendizaje que emplee consecuentemente la medición con referencia a criterios, se fundamenta en los siguientes supuestos:

1. Que la gran mayoría de los estudiantes de un grupo es capaz de mostrar dominio de los conocimientos enseñados, siempre que cada estudiante:

a) Tenga el tiempo que necesita para aprender.

b) Reciba la enseñanza óptima según su aptitud.

c) Tenga un grado de incentivación adecuado.

2. Que la materia a aprender esté dividida en pequeñas unidades, estructuradas de tal manera, que el estudiante conozca claramente lo que debe aprender (objetivos específicos).

3. Que se emplean continuamente diversos medios para retroalimentación al estudiante, sobre el logro de los objetivos cognoscitivos específicos (medición y evaluación formativa).

4. Que la evaluación del curso se realiza con base en el logro de los objetivos. Pudiéndose cambiar de esta manera la calificación numérica por un concepto dicotómico (ganó o perdió el curso).

./.

Los supuestos antes descritos no se dan en su totalidad, en los siste - mas de enseñanza de nuestro medio, y esto se refleja en alguna medida en - los resultados de la aplicación de las pruebas aquí investigadas.

Por otra parte, el estudio se fundamentó en los procedimientos dados - por diferentes autores, para la construcción de pruebas con referencia a - criterios y en cada una de las etapas se escogió una de las opciones metodológicas disponibles, escogencia que respondió tanto a los objetivos del proyecto, como a las limitaciones del mismo, en cuanto a recursos temporales, financieros, humanos y bibliográficos.

El primer paso en la construcción de estas pruebas, es el establecimiento de los conocimientos que se pretende medir (esquema descriptivo), procedimiento que de primera impresión parecería sencillo y de hecho es muy poco citado en la literatura; no obstante, se encuentran dificultades de orden práctico, como por ejemplo, la especificidad o amplitud con que se definan y la forma en que se hacen. Es necesario hacer notar que este esquema descriptivo es de vital importancia, pues sirve de base, tanto para que - los especialistas escogidos para establecer la validez de selección del dominio, emitan sus juicios, así como para que los encargados respectivos, - confeccionen las especificaciones del mismo.

Analizadas las estrategias para la especificación del esquema descriptivo existentes, que son muy variadas y en algunos casos muy específicas para ciertas áreas del conocimiento, se considera que la estrategia escogida para realizar este trabajo, (que es aplicable a diferentes áreas del conocimiento) con las limitaciones que lo rodearon, fue la mejor, sin embargo

./.

sería recomendable experimentar con otras estrategias para cada área de conocimiento.

La evidencia de la validez descriptiva (o de contenido) se obtuvo mediante el juicio de congruencia ítem-objetivo conductual, que tiene la limitación de que el objetivo conductual podría permitir interpretaciones diferentes de los jueces. La literatura no es explícita en estudios comparativos, del efecto que podría tener, en el juicio de congruencia, diferentes esquemas descriptivos. Esta área representa una veta para futuras investigaciones.

De igual forma se puede afirmar, que el área de la validez de constructo de este tipo de pruebas, no está suficientemente clara, ni tratada con amplitud en la literatura. Una opción que se da, es el cálculo del índice de validez de decisión que, en esta investigación en particular, se obtuvo con base en grupos criterio no experimentales. Estos grupos presentaron deficiencias cualitativas y por tanto la interpretación de los resultados de este índice es limitada.

Rose (1984), establece la importancia de tener evidencia sobre la validez instruccional de las pruebas. Por la forma en que fue validado el dominio de conocimientos de este proyecto, así como por la población a la que estaba destinada la prueba, no es pertinente obtener evidencia de este tipo de validez, ya que no se está midiendo el resultado de un sistema de enseñanza-aprendizaje homogéneo y delimitado.

Este proyecto constituyó un primer esfuerzo en la construcción de pruebas con referencia a criterios, dirigidas a la población de último año de

./.

la Educación Diversificada, por lo que el banco de ítemes construido fue el primero en su género. Este banco de ítemes debe ampliarse y mejorarse, especialmente en cuanto a la homogeneidad de las dificultades de los ítemes, esto permitiría una mayor propiedad en la construcción de las pruebas paralelas, mediante el muestreo aleatorio de los ítemes.

Con respecto al uso de hojas de cotejo, para revisar la calidad técnica de los ítemes, se considera que este instrumento fue útil únicamente en aquellos casos en que la persona que lo utilizó tenía experiencia y conocimiento en confección y revisión de ítemes.

La confiabilidad de las pruebas con referencia a criterios, es tratada con mayor amplitud que cualquier otro aspecto técnico, en la literatura. De los tres enfoques con que se trata el tema de la confiabilidad, en este trabajo solamente se aplicó una medida, dentro de uno de ellos.

Existen varias limitaciones para que se diera esta situación, entre ellas, se puede señalar, la falta de una verdadera situación experimental (un verdadero grupo criterio no instruido) y la carencia de los recursos computacionales (software) del caso.

Estas mismas limitaciones prevalecieron en la selección del índice de discriminación: se dio el caso que de once posibles índices, solo se podía utilizar con propiedad uno de ellos, debido a la ausencia de grupo criterio no instruido.

Nuevamente se considera necesario recomendar una mayor experimentación en ambos campos.

./.

Aunque en el estudio bibliográfico de la confiabilidad se pueden identificar algunos autores, los nuevos que favorecen el uso e interpretación de las fórmulas clásicas de la confiabilidad, también se encuentra una gran - mayoría de autores, que no recomiendan la interpretación de esas fórmulas clásicas, razón por la que, en este trabajo, se calculan pero no se interpretan los índices de consistencia interna clásica, de Kuder-Richardson.

Cabe señalar, como Popham (1978) lo establece, que estos índices, pueden ofrecer evidencia de la solidez de cada subprueba.

En el análisis de ítemes, la estrategia de analizar, después de la prueba piloto, la frecuencia con que cada distractor fue seleccionado, por los alumnos, resultó de mucha utilidad para mejorar la selección de los ítemes, lo que concuerda con lo recomendado por Berk (1980).

La longitud de cada subprueba, (número de ítemes por objetivo) fue una decisión arbitraria (en el sentido que Popham (1978) le da a este término) fundamentada en:

a) Los reportes de autores que han elaborado pruebas señalan una longitud - de subprueba no mayor de 5 ítemes.

b) El número de objetivos incluidos en la prueba que limitan la longitud - de cada subprueba.

c) La interpretación que se daría a los resultados (descriptiva, no para - tomar decisiones sobre cada individuo o grupos de individuos). Un área de investigación amplia para un futuro, estaría en ahondar en las técnicas que la literatura señala para optimizar la longitud de cada subprueba.

./.

El método escogido, para obtener el puntaje de corte, se seleccionó, tomando en cuenta, el grado de objetividad que este ofrecía y las limitaciones temporales para poder aplicar otros métodos. Este es otro campo que ofrece interesantes perspectivas de investigación para el futuro.

Finalmente, existe un problema inherente a cualquier tipo de prueba: el nivel de lectura de los ítemes. Tanto en este caso como en cualquier otro, es un factor que puede afectar sensiblemente la validez de la prueba tal como lo afirma Benson (1981); por la limitación de tiempo y de disponibilidad de un instrumento para medir este nivel de lectura, en este trabajo se obvió el análisis de dicho problema.

Conclusiones

1.- El procedimiento seguido para la construcción y validación de las pruebas permite afirmar con certeza que las mismas son pruebas con referencia a criterios que permiten diagnosticar conocimientos mínimos de estudiantes que egresan de la Educación Diversificada.

2.- El procedimiento seguido en la definición y especificación de los contenidos ofrece evidencia de la validez y selección del dominio de las pruebas (Popham 1978). Asimismo el procedimiento y los resultados obtenidos en el análisis de congruencia ítem-especificación del contenido (objetivo amplificado en este caso) y calidad técnica, sustenta la validez descriptiva (Popham 1978) o validez de contenido de Hambleton (1980).

./.

3.- En general se puede afirmar que la exactitud en la clasificación de los estudiantes como "masters" y "no masters" mostrada por cada una de las subpruebas es aceptable, en algunos casos muy buena (Hambleton 1980).

4.- En cuanto a la consistencia en la clasificación de los estudiantes como "masters" o "no masters" se tiene lo siguiente:

a) El índice Kappa varía de una subprueba paralela a otra para un mismo objetivo.

b) El resultado anterior es consistente con la calidad (discriminación, congruencia, dificultad, etc.) de los ítemes que conforman cada subprueba paralela.

c) A partir de las subpruebas paralelas analizadas se puede construir una prueba para cada asignatura, constituida por subpruebas con altos índices Kappa de confiabilidad (superiores a 70).

5.- Las medidas de calidad de los ítemes indican una alta proporción de ítemes buenos por asignatura (77% y más).

Recomendaciones

Se considera importante recomendar en general, que OPES continúe haciendo investigación en este campo, especialmente en los siguientes aspectos:

Recomendaciones sobre metodología empleada

./.

1.- Someter los objetivos ya amplificados al criterio de un grupo de expertos con el propósito de validar la amplificación de objetivos.

2.- Los jueces empleados para el análisis de congruencia deben ser supervisados durante todo el tiempo que empleen para llevar a cabo el análisis.

3.- Debe aumentarse el número de ítemes de base a construir para prever la disminución de los mismos debido al análisis de calidad técnica y congruencia.

4.- Obtención adicional de evidencias de confiabilidad dentro de los dos enfoques no contemplados en este proyecto, a saber: confiabilidad de las estimaciones de los puntajes de dominio y confiabilidad de los puntajes de pruebas con referencia a criterios.

5.- Dadas las limitaciones señaladas para las pruebas de Estudios Sociales construidas en este proyecto, en caso de querer utilizarlas, es conveniente mejorarlas en los siguientes aspectos:

a) Revisión de la especificación del dominio de conocimientos (objetivos amplificados).

b) La revisión de estos objetivos provocaría la necesidad de obtener nueva evidencia sobre la validez de la congruencia ítem-objetivo.

c) Aumento de una base de ítemes que dé posibilidad de incrementar el banco actual y que se ajuste a las posibles nuevas descripciones de contenidos.

./.

Recomendaciones administrativas

1.- El proyecto debe contar con:

a) Un presupuesto propio y no mediante la aportación en especie de cada una de las instituciones, situación que en este caso acarreó múltiples inconvenientes.

b) Una dirección académica y una administrativa que permita una utilización óptima del recurso humano en cada uno de esos campos.

2.- Dentro del presupuesto deben contemplarse recursos para:

a) Contratación de apoyo técnico por períodos determinados, de acuerdo con las necesidades del desarrollo del proyecto (asistentes de investigación, programación, digitación).

b) Aumento en el pago de jueces y constructores de ítemes.

Recomendaciones para Investigación Básica

Además de las recomendaciones antes señaladas que están relacionadas directamente con el desarrollo de las pruebas, se considera importante contribuir a la fundamentación teórica de las pruebas con referencia a criterios por lo que se recomiendan algunos posibles estudios:

a) Comparación de la eficacia de diferentes esquemas de especificación del dominio de conocimientos.

b) Comparación de la eficacia del empleo de objetivos amplificados versus objetivos conductuales en el análisis de congruencia ítem-objetivo.

./.

c) Comparación de diferentes procedimientos para obtener el juicio en la - congruencia ítem-objetivo.

d) Comparar la validez de diferentes técnicas, empleadas en la obtención - del puntaje de corte.

BIBLIOGRAFIA

Algina, J., and Noe, M.J. A study fo the accuracy of sukoviak's single-administration estimate of the coefficient of agreement using two truescore estimates. Journal of Educational Measurement, 1978, 15, 101-10.

Benson, J. A redifinition of content validity. Educational and Psychological Measurement, 1981, 41.

Benson, J. and Crocker, L. The Effects of Item Format and Reading Ability on Objective-Test. Performance: A question of vality. Educational and Psychological Measurement. 1979, 39.

Berk, R.A. Determination of optimal cutting score in criterion-referenced measurement. Journal of experimental Education, 1976, 45, 4-9.

Berk, R.A. The aplication of structural facet theory to achievement test construction. Educational Research Quaterly, 1978, 3, 62-72 (a).

Berk, R.A. A consumers' guide to criterion-referenced test statistics. Measurement in Education, 1978, 9, 1-8 (b).

Berk, R.A. A critical review of content domain specifications/item generation strategies for criterion-referenced tests. Paper presented at the annual meeting of the American Education Research Association, San Francisco, 1979. (ERIC Document Reproduction Service N°ED170382) (a).

Berk, R.A. Some guidelines for determining the lenght of objective-based criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, 1979. (ERIC Document Reproduction Service N°ED170381) (b).

Berk, R.A. A consumer' guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 17, 323-49 (a).

Berk, R.A. Criterion Referenced Measurement. The State of the Art, The John Hopkins University Press, Baltimore and London, 1980.

Block, J.H. Criterion-referenced measurement Potential. School Review, 1971, 79, 289-298.

Block, J.H., ed. Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston, 1971.

Block, J.H. Standars and criteria: A response. Journal of Educational Measurement, 1978, 15, 291-95.

Bormuth, J.R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

./.

Brenann, R.L. A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

Brennan, R.L. GAPID: A fortran IV Computer program for generalizability - analysis with single-facet designs. ACT Thecnical Bulletin N°34. Iowa City, Iowa: American College Testing Program, 1979.

Brennan, R.L. Applications of Generalizability theory. In R.A. Berk (Ed.) Criterion-referenced measurement: the state of the art. Baltimore, Maryland: The John Hopkins University Press, 1980.

Brennan, R.L. and Kane, M.T. An index of dependability for mastery tests. - Journal of Educational Measurement, 1977, 14, 277-89.

Brennan, R.L. and Kane, M.T. Signal/noise ratio domain-referenced tests. - Psychometrika, 1978, 42, 609-25, Errata, 1978, 43, 289.

Brown, F. Principle of educational and psychological testing. New York, N.Y.: Holt, Renehart and Winston, 1976.

Burton, N.W. Societal Standards. Journal of Educational Measurements, 1978, 15, 263-71.

Carver, R.P. Special problems in Measuring change with psychometric devices. In Evaluative research: Strategies and methods. Washington: American Institutes for research, 1970.

Cohen, J.A. Coefficient of agreement for nominal scales. Educational and - Psychological Measurement, 1960, 20, 37-46.

Cox, R.C., and Vargas, J.S. A comparison of item selection techniques for - norm-referenced and criterion-referenced.tests. Paper presented at the annual meeting of the National Council on Measurement in Education, 1966.

Crambert, A.C. Estimation of Validity for criterion-referenced tests. Paper presented at meeting of the American Educational Research Association, - New York, April 1977. (ERIC Document Reproduction Service N°ED151418).

Crehan, K.D. Item analysis for teacher-made mastery tests. Journal of Educational Research, 1974, 11, 255-62.

Crehan, K.D. Item analysis for teacher-made mastery tests. Journal of Educa tional Measurement, vol. 11, N°4, Winter 1974.

Crombach, L.J., et al. The dependability of behavioral measurement: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Curlette, W.L. and Stallings, W.M. Ten issues in criterion-referenced Testing: A response to commonly heard criticims. The Clearing House, vol. 53, nov. 1979.

./.

Digvi, D.R. Group dependence of some reliability indices for mastery test. Applied Psychological Measurement, Vol. 4, N°2, pp. 213-218, Spring 1980.

Dilendik, J.R. Assumptions underlying criterion referenced assesment are - educationally sound. Education, 1976, 99, 89-96.

Downing, S.M., and Mehrens, W.A. Six single-administration-reliability coeffi cients for criterion-referenced tests: A comparative study. Paper presented at the Annual Meeting of the American Educacional Research Asso - ciation. Toronto, Ontario, Canada. March, 1978.

Downing, S.M., and Mehrens, W.A. Six single-administration coefficients for criterion-referenced tests: A comparative study. Paper presented at the Annual Meeting of the American Educational Association. Ontario, 1979. (ERIC Document Reproduction Service N°ED161929).

Ebel, R.L. Evaluation and Educational objectives. Journal of Educational Measurement, 1973, 10, 273-79.

Enright, B.E. Criterion-referenced tests: A guide to separate useful from - useless. Paper presented at the Annual International Convention of the Council for Excepcional Children, Houston, Texas, April 1982.

Esquivel, J.M. and Quesada, L. The development, validation and administration of the criterion-referenced science battery for general education students in Costa Rica. Paper presented at the Annual Covention of National Asso ciation for Research in Science Teaching. French-Licks Springs, Indiana, 1985.

Esquivel, J.M., Peralta, T. y Delgado, V. Desarrollo y validación de Pruebas de Conocimientos Mínimos en Matemática y su aplicación en una muestra na cional de escuelas y colegios. Revista de la Universidad de Costa Rica, Educación, 1984, 7, 125-134.

Finn, P.J. A question writing algorithm. Journal of reading behavior, 1975, 4, 341-67.

Frasier, G.M., and Raeth, P.G. An internal consistency estimate for criterion-referenced tests. A paper presented at the Annual Meeting of the Natio- nal Council and Measurement in Education, New York, 1977. (ERIC Document Reproduction Service N°ED 137 359).

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1973, 18, 519, 521.

Glaser, R. and Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (ED), Educational Measurement. (2da. ed.) Washington: Ameri- can Council on Education, 1971.

Greco, T.H. Is there really a difference between criterion-referenced and - norm-referenced measurement. Educational Technology, 1974, 22-25.

./.

Green, K.E. Subjective Judgment of Multiple-Choice Item Characteristics. Educational and Psychological Measurement, 1983, 43.

Gross, L.J. Standars and criteria: A responde to Glass'criticism to the Nedelsky technique. Journal of Educational Measurement, 1982, 19, 159-162.

Haertel, E. Detection of a skill dichotomy using standarized achievement tests items. Journal of Educational Measurement, 1984, 21, 59-72.

Haertel, E., and Calfee, R. School achievement: thinking about what to test. Journal of Educational Measurement, 1980, 20, 119-132.

Haladyna, T.M. Effects of different samples of item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

Hambleton, R.K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

Hambleton, R.K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-90.

Hambleton, R.K. Test score validity and standard-setting methods. In R.A. Berk (ED.). Criterion-referenced Measurement: The state of the art. Baltimore, Maryland: The John Hopkins University Press, 1980.

Hambleton, R.K. and Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R.K., and Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.K., Hutlen, L.R., and Swaminathan, H.A. A comparison of several methods of assesing students mastery in objective-based instruction programs. Journal of Experimental Education, 1976, 45, 57-64.

Hambleton, R.K., Swaminathan, H., and Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D.N.M. the gruyter, and L.J. Th. van der Kamp (Eds.). Advances in psychological and educational measurement. New York: Wiley, 1976.

Hambleton, R.K., et al. Criterion-referenced testing and measurement: a review of technical issues developments. Review of Educational Research, 1978, 48, 1-48.

Harris, C.W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

./.

Hively, W., Patterson, H.L., and Page, S.A. A "universe defined" system of arithmetic achievement tests: Journal of Educational Measurement, 1968, 5, 275-90.

Horodezky, B. and Labercane G. Criterion-referenced tests as predictors of reading performance. Educational and Psychological Measurement, 1983, 43.

Hsu, T.C. Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hsu, L.M. Determination of the number of items and passing score in a mastery test. Educational and Psychological Measurement, 1980, 40.

Huynh, H. On reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264 (a).

Huynh, H. Statiscal considerations of mastery scores Psycometrika, 1976, 41, 65-78.

Huynh, H. The Kappamax reliability index for decisions in domain-referenced testing. Paper presented at the annual meeting of the american Educational Research Association, New York, 1977 (ERIC Document Reproduction Service NºED 154004).

Huynh, H. and Saunders, J.C. Accuracy of two procedures for estimating reliability of mastery tests. Journal of Educational Measurement, 1980, 17, 351-58.

Huynh, H. Adequacy of asymptotic normal theory in estimating rehability for mastery tests based on the beta-binomial model. Journal of Educational statistics, vol. 6, Nº3, pp. 257-266, Fall, 1981.

Jencks, C., and Crouse, J. Aptitud vs. achievement: should we replace the SAT? The Public Interest, Vol. 67, Nº21, 35, Spring 1982.

Kane, M.T., and Brennan, R.L. Agreement coefficients as indices of dependability for damain-referenced tests, ACT. Technical Bulletin Nº28. Iowa City. Iowa: American College Testing Program. (ERIC Document Reproduction Service NºED 185076).

Kim, J.O. Factor analysis, statistical package for the Social Sciences, University of Iowa.

Klein, S.P., and Kosecoff, J.B. Issues and procedures in the development criterion-referenced tests. Princeton, N.J.: Educational testing Service, 1973. (ERIC Document Reproduction Service NºED 083284).

Klein, S.P., and Kosecoff, J.B. Issues and procedures in the development of criterion-referenced tests. In W.A. Mehrens (ED.) Readings in measurement and evaluation in education and psychology. New York: Holt, Rinehart and Winston, 1976.

./.

Lang, H.G. Criterion-referenced test in science: An investigation of reliability, validity, and standards-setting. Journal of research in Science Teaching, vol. 19 N°8, pp. 665-674, 1982.

Linn, R.L. Issues of validity in measurement for competency-based programs. In M.A. Bunda and J.R. Sandees (ED.) Practices and problems in competency-based measurement. Washington, D.C.: National Council on Measurement in Education.

Livingston, S.A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26 (a).

Livingston, S.A. A reply to Harriesl. "An interpretation of Livingston's - reliability coefficient for criterion-referenced tests". Journal of Educational Measurement, 1972, 9, 31 (b).

Livingston, S.A. Reply to Shavelson, Block and Ravitch's. "Criterion-refe - renced testing: Comments on reliability". Journal of Educational Measurement, 1972, 1, 139-140 (c).

Livingston, S.A. A utility-based aprroach to the evaluation of pass/fail testing decisions procedures. Report N°COPA-75-01. Princeton, N.J.: Center for Occupational and Proffesional Assessment, Educational Testing Service, 1975.

Livingston, S.A., and Wingersky, M.S. Assesing the reliability of tests used to make pass/fail decision. Journal of Educational Measurement, 1979, - 16, 247-60.

Lord, F.M., and Novick, M.R. Statistical theories of mental test scores. - Reading Mass: Addison-Wesley, 1968.

Lovett, H.L. Criterion-referenced reliability estimated by ANOVA. Educational and Psychological Measurement, 1977, 37, 21-29.

Lovett, H.L. The effect of violating the assumption of equal item means in - estimating the Livington coefficient. Educational and Psychological - Measurement, 1978, 38, 259-51.

Marshall, J.L., and Haertel, E.H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Manus - cript University of Wisconsin, 1976.

Mellenbergh, G., and Van der Linden, W.J. Selecting Items for Criterion-Referenced Tests. Evaluation in Education. Vol.5, pp. 117-190, 1982.

Messick, S.A. The standard problem: Meaning and values in measurement and - evaluation. American Psychologist, 1975, 30, 955-66.

./.

Miller, H.G., and Reed, G.W. Constructing higher level multiple choice questions covering factual content. Educational Technology, 1973, 39, 42.

Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in education: Current aplications. Berkeley, California: McCut - chan Publishing Co., 1974.

Millman, J. Computer-based item generation. In R.A. Berk (Ed.), Criterion - referenced measurement: The state of the art. Baltimore, Maryland: - The John Hopkins University Press, 1980.

Millman, J. Reliability and validy of Criterion Referenced Test Scores. New Directions for Testing and Measurement, vol. 4, 1979.

Millman, J., and Popham, W.J. The issue of item and test variance for criterion-referenced test: A clarification. Journal of Educational Measurement, 1974, 11, 137-38.

Moyer, J.E., and Fishbein, R.I. A comparison of Kuder-Richarson formula 20 - and kappa as estimates of the reliability of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977. (ERIC Document Reproduction Service, - Nº ED 139.821).

Nitko, A.J. Distinguishing the many varieties of criterion referenced tests. Review of Educational Research, 1980, 50, 461-485.

Peng, C.J. An investigation on Huynh's normal approximation procedure (Docto ral dissertation, University of Wisconsin, 1979). Dissertation Abstracts International, 1979, 40, 4546 A.

Peng, C.J., and Sukoviac, M.J. A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, 1980, 17, 359-68.

Poggio, J.P., and Glassnapp, D.R. Report of research findings: The Kansas - Competency Testing Program-1980. Topeka, KS: Kansas State Department - of Education, 1980.

Poggio, J.P., Glasnapp, D.R., and Eros, D.S. An empirical investigation of - the Angoff, Ebel and Nedelsky standards setting methods. Paper presen - ted at the annual meeting of the American Educational Research Associa - tion, Los Angeles, 1981.

Popham, W.J., and Husek, T.R. Implications of criterion-referenced measure - ment. Journal of Educational Measurement, 1969, 6, 1-9.

Popham, W.J. Evaluation in Education. University of California. Los Ange - les, 1974.

./.

Popham, W.J. Educational Evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1975.

Popham, W.J. Normative data for criterion-referenced test? Phi Delta kappan. 1976, 57, 593-94.

Popham, W.J. Criterion referenced-measurement. University of California, Los Angeles, 1978.

Popham, W.J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1978 (a).

Popham, W.J. As always, provocative. Journal of Educational Measurement, 1978, 15, 297-300 (b).

Popham, W.J. Well-crafted criterion-referenced tests. Educational Leadership, 1978, 91-95 (c).

Popham, W.J. Setting perfomance standars. Los Angeles: Instructional Objectives Exchange, 1978 (d).

Popham, W.J. The case of criterion-referenced measurement. Educational researcher, 1978, 7, 6-10 (e).

Popham, W.J. Domain specification strategies. In R.A. Berk (Ed.). Criterion-referenced Measurement: The state of the art. Baltimore, Maryland: The John Hopkins University Press, 1980.

Priestley, M., and Nassif, P.M. From here to validity. Developing a conceptual framework for test item generation in criterion-referenced measurement. Educational Technology, 1979, 27-32.

Ravid, R. Presentation of Procedures for Development of a Second Language Achievement Test. Foreigh language annals, 16, N°3, 1983.

Reid, J.B., and Roberts, D.M. A Monte Carlo comparison of Phi and Kappa as measures of criterion-referenced reliability. Paper presented at the annual meeting of the American Educational Association, Toronto, 1978. (ERIC Document Reproduction Service N°ED 159226).

Roid, G.H. The technology of test-item writing. In H.F. O'Neill, Jr. (Ed.). Procedures for instructional systems development. New York: Academic Press, 1979.

Roid, G.H., and Haladyna, T.M. A comparison of objective-based and modified Bormuth item writing techniques. Educational and Psychological Measurement, 1978, 38, 19-28.

Roid, G.H., and Haladyna, T.M. The emergence of an item writing technology. Review of Educational Research, 1980, 50, 293-314.

./.

Rose, J., and others. Instruccional validity: Merging Curricular, Instruc - cional and Test Development Issues. Paper presented at the annual mee - ting of the American Educational Research Association, New Orleans, April 1984.

Roudabush, G.E. Item selection for criterion-referenced tests. Paper presen ted at the annual meeting of the American Educational Research Associa - tion, New Orleans, 1973.

Rovinelli, R.J., and Hambleton, R.K. On the use of coritent specialists in - the assesment of criterion-referenced test item validity. Dutch Journal of Educational Research, 1977, 2, 49-60.

Safrit, M.J., and Stamm, C.L. Reliability estimates for criterion-referenced measures in the psychomo for domain. Research Quaterly for Exercise and Sport, 1980, 51, 359-68.

Scandura, J.M. Structural approach to behavioral objective and criterion-refe renced testing. Educational technology, 1977, 20-25.

Schaefer, M.M. and others. A Comparison of Reliability Estimates form Single and Double Administrations of Criterion-Referenced Tests. (ERIC Educatio nal Resurses Information Center).

Schmidt, W.H. Content bias in achievement tests. Journal of Educational - Measurement, 1983, 20, 166-178.

Shalvenson, R.J., Block, J.H., and Ravitch, M.M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, - 113-137.

Shannon, G.A. Objective -referenced-test rescore decisions and item statistics: A matter of congruence. Paper presented at the annual meeting of the - American Educational Research Association, Montreal, Canada, april 1983.

Sheehan, D.S., and Davis, R.G. The development and validation of a criterion-referenced mathematics battery. School, Science and Mathematics, 1979, 125-132.

Shepard, L. Norm-referenced vs. criterion-referenced tests. Educationl Ho - rizons, 1979, 57, 26-32.

Sukoviak, M. Estimating reliability forms a single administration of a crite rion-referenced test. Journal of Educational Measurement, 1976, 13, - 266-275.

Sukoviak, M.J. Empirical investigation of procedures for estimating reliabi- lity for mastery tests. Journal of Educational Measurement, 1978, 15, 111-16.

Sukoviac, M. Decisions-making approaches. In R.A. Berk (Ed.). Criterion-re- ferenced measurement: The state of the art. Baltimore, Maryland: The John Hopkins University Press, 1980.

./.

Swamintahan, H., Hambleton, R.K., and Algina, J. A bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement. 1975, 12, 87-99.

Tiennann, P.W., and Markle, S.M. Analysing instructional content: A guide to instruction and evaluation. Champaign, III: Stipes Publishing, 1978.

Van der Linden, W.J. A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard settings. - Journal of Educational Measurement, 1982, 19, 295-308.

Van der Linden, W.J. Criterion-referenced measurement: Its main applications, problems and findings. Evaluation in Education, vol. 5, 97-118, 1982.

Wall, J., and Geppert, W.J. Initiating school change through a state wide - testing program in math. Delaware State Board of Education.

Wardrop, J.L., et al. A framework for analyzing the inference structure of - educational achievement tests. Journal of Educational Measurement, 1982, 19, 1-18.

Woodson, M.I. The issue of item and test variance for criterion-referenced - tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140.

Williams, S.S. Criterion-referenced tests. Improving College and University Teaching, vol. 27 N°1, W-79.

Zieky, M.J., and Livingston, S.A. Manual for setting standars on the Basic Skills Assessment Tests. Princeton, N.J.: Educational Testing Service, 1977.

Zwarts, M.A. On the Construction and Validation of Domain-Referenced Measurements. Evaluation in Education, vol. 5, 119-139, Great Britain, 1982.

ANEXO N°1 (a)

DISTRIBUCION DE ESTUDIANTES QUE REALIZARON LAS PRUEBAS EXPERIMENTALES DE CONOCIMIENTOS POR ASIGNATURA Y FORMULA

ANEXO N°1 (a)

DISTRIBUCION DE ESTUDIANTES QUE REALIZARON LAS PRUEBAS

EXPERIMENTALES DE CONOCIMIENTOS POR ASIGNATURA Y FORMULA

ASIGNATURA \ FORMULA	A	B	C	TOTAL
Ciencias	275	273	275	823
Matemáticas	255	266	268	789
Estudios Sociales	320	320	316	956
Español	321	320	308	949
TOTAL 3.517				

DESGLOSE POR INSTITUCION DE EDUCACION SUPERIOR

Universidad de Costa Rica:

	A	B	C	TOTAL
Ciencias	201 (48)	200 (47)	203 (47)	604 (142)*
Matemáticas	194 (48)	200 (47)	201 (46)	595 (141)
Estudios Sociales	228 (60)	224 (60)	226 (60)	678 (180)
Español	227 (59)	225 (60)	222 (60)	674 (179)
TOTAL	850 (215)	849 (214)	852 (213)	2.551 (642)

Universidad Nacional:

	A	B	C	TOTAL
Ciencias	42	43	42	127
Matemáticas	41	45	45	131
Estudios Sociales	62	66	63	191
Español	63	65	60	188
TOTAL	208	219	210	637

Instituto Tecnológico de Costa Rica:

	A	B	C	TOTAL
Ciencias	32	30	30	92
Matemáticas	20	21	22	63
Estudios Sociales	30	30	27	87
Español	31	30	26	87
TOTAL	113	111	105	329

* Los datos entre paréntesis corresponden a los Centros Regionales.

ANEXO Nº1 (b)

MUESTRA DE ESTUDIANTES DE LA EDUCACION DIVERSIFICADA

(IV NIVEL) A QUIENES SE APLICARON LAS PRUEBAS DE CONOCIMIENTOS

ANEXO N°1 (b)

MUESTRA DE ESTUDIANTES DE LA EDUCACION DIVERSIFICADA

(IV NIVEL) A QUIENES SE APLICARON LAS PRUEBAS DE CONOCIMIENTOS

TIPO DE COLEGIO	ESPAÑOL	CIENCIAS	MATEMATICA	ESTUDIOS SOCIALES	TOTAL
Académico Diurno - Oficial	47	67	59	94	267
Académico Diurno - Particular	31	31	33	38	133
Técnico Oficial	51	57	61	60	229
Académico Nocturno Oficial	26	16	24	16	82
TOTAL	155	171	177	208	711

DESGLOSE POR FORMULA

ASIGNATURA	FORMULA			TOTAL
	A	B	C	
Español	54	50	51	155
Ciencias	61	60	56	177
Matemática	60	56	55	171
Estudios Sociales	70	54	57	208
TOTAL				711

ANEXO Nº2

LISTAS DE OBJETIVOS POR ASIGNATURAS

ANEXO N°2

LISTAS DE OBJETIVOS POR ASIGNATURAS

Lista de objetivos de Biología

El estudiante será capaz de:

1.- Distinguir la fotosíntesis de la respiración celular.

2.- Relacionar la estructura y la función de los componentes de la célula: membrana celular, núcleo, cromosomas, mitocondrias, cloroplastos y ribosomas.

3.- Reconocer las características funcionales de las enzimas.

4.- Distinguir los diferentes niveles de organización de los seres vivos: Ecosistema, Población, Comunidad, Bioma, Organismo y Bioesfera.

5.- Identificar cada uno de los niveles tróficos, en las cadenas alimenticias.

6.- Identificar las características del proceso de mitosis.

7.- Resolver problemas genéticos de cruces de monohíbridos.

Lista de objetivos de Física

El estudiante será capaz de:

1.- Realizar conversión de unidades, dados los factores de conversión necesarios.

2.- Aplicar la ley de conservación de la energía mecánica en problemas de movimiento de un cuerpo.

./.

3.- Resolver problemas aplicando la ecuación del movimiento uniforme.

4.- Calcular la velocidad al llegar al suelo, o el tiempo de caída, de un cuerpo que se suelta desde una altura determinada.

5.- Interpretar gráficas de velocidad en función del tiempo, para movimiento rectilíneo.

6.- Calcular la magnitud de la resultante de un sistema de 2 vectores co - planares y concurrentes, no perpendiculares ni colineales.

7.- Resolver situaciones que impliquen la aplicación de la 2da. Ley de Newton del movimiento.

Lista de objetivos de Química

El estudiante será capaz de:

1.- Relacionar los conceptos materia-energía con los cambios de estado de la materia.

2.- Relacionar los conceptos de: número atómico, ion, número de masa, con el número, tipo y carga de las partículas subatómicas.

3.- Relacionar la capacidad electrónica de cada nivel y subnivel con los - números cuánticos.

4.- Relacionar la posición de los elementos en la Tabla Periódica con sus propiedades físicas y químicas.

5.- Identificar el tipo de enlace químico que se establece entre elementos.

6.- Interpretar el significado de ecuaciones químicas.

./.

7.- Resolver problemas sobre disoluciones, tomando en cuenta expresiones - molares.

Lista de objetivos de Español

El estudiante será capaz de:

1.- Aplicar las leyes del acento y del hiato.

2.- Reconocer el acento ortográfico en palabras monosilábicas.

3.- Aplicar el uso corrido de las letras v, b, z, s, c, h, m/p,-b.

4.- Usar la letra mayúscula de acuerdo con las normas ortográficas.

5.- Emplear las normas establecidas para la concordancia nominal y verbal.

6.- Emplear las preposiciones y frases prepositivas según las reglas idiomáticas.

7.- Aplicar correctamente los signos de puntuación (coma, punto, punto y - coma y dos puntos).

8.- Determinar las ideas que conforman un párrafo.

9.- Distinguir entre las características de cada parte de un informe de - investigación: índice, introducción, desarrollo, conclusión y bibliogra - fía.

10.- Explicar el nacimiento y evolución del Español.

11.- Analizar sintácticamente oraciones simples.

12.- Detectar errores sintácticos como dequeísmo, leísmo, uso correcto de - los relativos: que, quien, cual, donde y el uso de haber como impersonal.

13.- Distinguir entre verbos conjugados y formas no personales del verbo.

14.- Discriminar entre los movimientos literarios: modernismo, realismo y -

./.

surrealismo.

15.- Reconocer la idea central en párrafos dados.

16.- Identificar errores comunes en el uso de verbos irregulares tales como: diptongación, acentuación, conjugaciones especiales.

17.- Reconocer errores idiomáticos frecuentes en el medio costarricense como supresión o adición de la de y sustitución de la l por r.

18.- Discriminar entre oraciones simples y compuestas.

19.- Identificar las características básicas de los géneros literarios.

0.- Analizar las obras Poema del Mío Cid y El Ingenioso Hidalgo don Quijote de la Mancha, en cuanto a personajes, temas, características y trascendencia histórica.

Lista de objetivos de Estudios Sociales

El estudiante será capaz de:

1.- Caracterizar el proceso de consolidación de la democracia como forma de organización política en Costa Rica.

2.- Reconocer las principales funciones y atribuciones de los poderes que conforman el Estado Costarricense.

3.- Identificar los principales derechos y deberes ciudadanos de los costarricenses contenidos en nuestra Constitución Política.

4.- Distinguir los principales campos de acción de las instituciones autónomas costarricenses.

5.- Reconocer los razgos fundamentales de la social democracia, la democra

./.

cia cristiana y el comunismo.

6.- Reconocer la influencia del sector agrícola en el desarrollo de la economía costarricense.

7.- Reconocer las principales industrias privilegiadas por las políticas fiscales y arancelarias en Costa Rica, durante la segunda mitad del Siglo XX.

8.- Identificar las principales características de los movimientos de organización cooperativa en Costa Rica.

9.- Citar las principales garantías sociales establecidas en la Constitución Política vigente, en Costa Rica.

10.- Distinguir las variaciones del sufragio a través de las Constituciones de Costa Rica.

11.- Identificar las características del subdesarrollo latinoamericano.

12.- Reconocer las principales características de la tenencia de la tierra, en América Latina.

13.- Relacionar el proceso de industrialización con el desarrollo económico, en Brasil, Argentina, México, Uruguay y Chile.

14.- Determinar las principales particularidades del crecimiento de la población latinoamericana durante el Siglo XX.

15.- Identificar las características de las diferentes etapas por las que han atravesado las relaciones entre los Estados Unidos y América Latina.

./.

16.- Reconocer las funciones políticas desempeñadas por los militares en - los países latinoamericanos.

17.- Diferenciar los rasgos específicos de los movimientos revolucionarios de México, Bolivia y Cuba en el Siglo XX.

18.- Identificar las principales características de los movimientos popu - listas en Brasil, Argentina y Colombia en el Siglo XX.

19.- Ubicar los principales recursos naturales con que cuentan los países latinoamericanos.

20.- Contrastar las principales ideas políticas del liberalismo y del comunismo.

Lista de objetivos de Matemática

El estudiante será capaz de:

1.- Transformar expresiones trigonométricas en otras equivalentes que sólo contengan ángulos agudos y positivos.

2.- Asociar el signo de discriminante de un trinomio de 2° grado, su expresión analítica y su representación gráfica, o viceversa.

3.- Traducir expresiones en lenguaje exponencial a expresiones en lengua - je logarítmico, o viceversa, de acuerdo con las propiedades de logaritmos.

4.- Obtener el valor de las funciones trigonométricas de un ángulo donde α $0 \leq \alpha \leq 360°$, dado el valor de una de ellas.

5.- Aplicar los teoremas de relaciones entre rectas de la circunferencia -

./.

en la solución de problemas.

6.- Mostrar identidades por medio de la aplicación de las propiedades de - los logaritmos.

7.- Resolver ecuaciones cuya solución requiera de las propiedades de los - logaritmos.

8.- Reconocer la expresión trigonométrica que es equivalente a la expresión dada.

9.- Reconocer si una expresión polinomial real dada es o no es factoriza - ble en $\mathbb{R}$.

10.- Aplicar el "teorema del factor" para asociar la expresión analítica de un polinomio de variable real, sus ceros y su factorización.

11.- Traducir las relaciones que se dan en el enunciado verbal de un proble_ ma a una expresión algebraica en forma de ecuación de segundo grado con una incógnita.

12.- Realizar operaciones con fracciones racionales y expresar el resulta - do en forma simplificada.

13.- Reconocer si existe función en la relación dada entre los elementos de dos conjuntos.

14.- Determinar las imágenes de elementos del dominio de una función real - de variable real.

15.- Analizar los componentes de funciones lineales de acuerdo con sus re -

./.

presentaciones gráficas.

16.- Resolver problemas en los que integre conceptos geométricos y trigonométricos.

17.- Reconocer el procedimiento para efectuar operaciones con fracciones racionales.

18.- Reconocer el conjunto de soluciones de una ecuación de 2° grado con una incógnita.

19.- Analizar el enunciado de un problema que involucra conceptos geométricos de área para determinar la suficiencia y necesidad de los datos que en él se ofrecen tendiente a encontrar su solución.

20.- Reconocer la expresión en factores irreductibles de una expresión polinomial real dada.

ANEXO N°3

ESCALA PARA JUZGAR OBJETIVOS

ANEXO Nº3

ESCALA PARA JUZGAR OBJETIVOS

Estimado profesor:

Le agradecemos de antemano su valiosa colaboración en la validación de objetivos que, tal como se explicó, es una de las partes fundamentales de la investigación que se lleva a cabo.

En este documento encontrará una lista de objetivos y una escala de valoración, para que usted juzgue la importancia que tiene, el contenido representado por cada objetivo, en los conocimientos mínimos que debe poseer un estudiante al egresar de la educación diversificada.

Lo que le pedimos específicamente es, que marque en la escala, la importancia que en su opinión tiene ese objetivo. El criterio que debe usar es:

¿ES IMPORTANTE QUE EL ESTUDIANTE SEPA ESO,
COMO MINIMO, AL FINALIZAR SECUNDARIA Y DIRIGIRSE A LA UNIVERSIDAD?

Con el siguiente ejemplo explicamos la escala y lo que significa marcar cada punto en ella:

Si marca usted en la columna 1, considera que ese objetivo carece totalmente de importancia y por lo tanto no vale la pena exigirlo como evidencia para comprobar si el estudiante está bien preparado en esa materia, al egresar de secundaria.

./.

Marcar en la 2, quiere decir que aunque tiene alguna importancia hay muchos objetivos más importantes que ese; además, cualquier estudiante se desenvolvería perfectamente sin necesidad de dominar ese objetivo.

Si marca la columna 3, significa que es un objetivo importante; es una de las cosas que usted esperaría que la mayoría de sus estudiantes dominara.

Los objetivos que califique como 4, son imprescindibles; no se puede finalizar la secundaria sin dominarlos enteramente.

Antes de iniciar la valoración lea todos los objetivos de la lista.

ESCALA

Objetivos Nº	1 No tiene importancia; no vale la pena exigirlo	2 Tiene poca importancia; no es grave si el estudiante lo ignora	3 Es importante; vale la pena exigirlo	4 Es imprescindible; todos los estudiantes deberían saberlo
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

Para finalizar, ¿Cree usted que hay objetivos del tipo 4 (imprescindi - bles) que faltan en nuestra lista?

Si es así ¿Sería tan amable de escribirlos a continuación?. (Basta con que nos indique el contenido claramente).

1.- _____

_____

_____

2.- _____

_____

_____

3.- _____

_____

_____

SU COOPERACION ES DE INAPRECIABLE VALOR.

¡MUCHAS GRACIAS!

ANEXO Nº4

OBJETIVOS AMPLIFICADOS

ANEXO N°4

OBJETIVOS AMPLIFICADOS

Los objetivos amplificados:

- Reducen la incertidumbre en cuanto a la forma y la extensión de los ítemes que se confeccionan con base en ellos.

- Ayudan a construir ítemes con las mismas características.
- Son versiones más elaboradas del objetivo conductual.
- Dan a los constructores de ítemes las reglas del juego y una idea clara del ítem que deben construir.

- Estan construidos siguiendo el siguiente esquema:

  Objetivo:

  _____

  _____

Item de muestra:

  _____

  _____

Objetivo amplificado:

- Condiciones generales del enunciado

  1.- _____

  2.- _____

./.

- Condiciones de las opciones de respuesta.

  1.- _____

  2.- _____

- Criterio de corrección.

  1.- _____

  2.- _____

Condiciones generales del enunciado

En este apartado debe incluirse:

- Estructura de la pregunta (elementos, ordenamiento, longitud).
- Si hay diferentes enunciados posibles, balance entre ellos.
- Nivel de lenguaje.
- Límites de contenido (listas de asuntos o reglas).
- Reafirmación de las condiciones básicas contenidas en el objetivo (qué elementos o situaciones contendrá el enunciado y posibles reglas de construcción de ítemes (problemas de tal tipo, oración con subrayados, un gráfico y una pregunta un gráfico y una oración, etc.)).

Condiciones de las opciones de respuesta

En este apartado debe incluirse:

- Especificación de estructura y forma de las opciones tales como: equivalencia, longitud, lenguaje, simbología, descripción de cada elemento. Es una descripción detallada de las condiciones de las opciones, para los que las escriban.
- Orden de los elementos dentro de las opciones, si se considera necesario.

./.

- Opciones no aceptables.

Posteriormente se agregarán a este apartado aquellas reglas que rigen para todos los constructores de ítemes como son:

- Deben haber 4 ó 5 opciones de respuesta.
- Deben ser de elección múltiple simple.
- Tiene que existir una única mejor respuesta, etc.

- Criterio de corrección

Detallar aquí la estructura de la opción correcta (condiciones que debe cumplir).

ANEXO N°5

NUMERO DE ITEMES QUE PASARON EN EL ANALISIS DE CALIDAD TECNICA POR OBJETIVO Y ASIGNATURA, EN LA PRIMERA REVISION

ANEXO 5

NUMERO DE ITEMES QUE PASARON EN EL ANALISIS DE CALIDAD

TECNICA POR OBJETIVO Y ASIGNATURA, EN LA PRIMERA REVISION

NUMERO OBJETIVO	ESTUDIOS SOCIALES	ESPAÑOL	MATEMATICA	QUIMICA	FISICA	BIOLOGIA
1	4	20	4	2	20	11
2	3	20	8	2	16	10
3	6	19	3	10	20	10
4	4	20	10	11	19	12
5	12	18	6			
6	12	13	14			
7	13	20	7			
8	6	20	10			
9	13	20	8			
10	8	13	8			
11	9	20	12			
12	13	20	6			

ANEXO N°6

INSTRUCCIONES PARA LOS PROFESORES

INSTRUCCIONES PARA LOS ESTUDIANTES

ANEXO Nº6

INSTRUCCIONES PARA LOS PROFESORES

Estimado profesor:

A continuación le presentamos las normas por las cuales se regirá la aplicación de las pruebas de conocimientos para los estudiantes de la Educación Diversificada (IV año). Por las razones que usted ya conoce (estandarización) las normas deben ser seguidas lo más fielmente posible.

ACERCA DEL MATERIAL

1.- El material que se deposita bajo su responsabilidad es de carácter CONFIDENCIAL; esto significa que todos los folletos de las pruebas y las hojas para respuestas deben ser devueltos al señor coordinador y los folletos no utilizados deben mantenerse sellados.

2.- Para aplicar la prueba, el coordinador de su grupo le entregará una caja con los materiales. Sin alterar el orden, usted deberá comprobar que ésta contenga:

- Los folletos correspondientes a la prueba, numerados según se indica en la tapa de la caja y ordenados en forma alterna (hay tres fórmulas, A, B y C de diferente color).

- Una hoja para sellar la caja una vez concluida la aplicación.
- Una hoja para control de asistencia. En ella cada estudiante anotará su nombre y su firma (Ver instrucción al respecto).

./.

- Un sobre que contiene las hojas para respuestas.
- Cinta adhesiva para sellar la caja una vez concluida la aplicación.
- Una hoja en blanco para anotar las impresiones sobre el desarrollo de la prueba y las preguntas o dudas que manifestaron los estudiantes en algún momento de la misma.

- Instrucciones para leer a los estudiantes antes de iniciar la prueba.
- Un comprobante de que usted recibió los folletos.

3.- Verifique que el número de folletos de la prueba anotado en la tapa, - sea el mismo que le fue entregado y escriba dichas cantidades en el compro_ bante. Como los folletos están ordenados secuencialmente, anote también - en el comprobante, el número del folleto inicial y final. Firme el compro_ bante y devuélvalo al señor coordinador.

ANTES DE INICIAR LA PRUEBA

1.- Verifique que en el aula que le corresponde haya un escritorio donde - usted pueda colocar sus documentos.

2.- Escriba en la pizarra: "NO ABRA EL FOLLETO DE PRUEBA HASTA QUE SE LE INDIQUE".

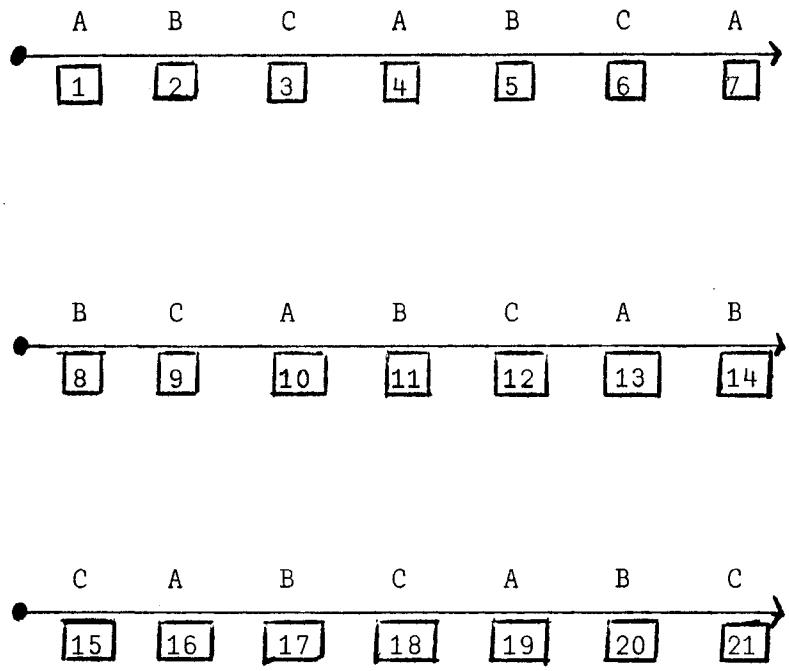
3.- Coloque los pupitres de la sala donde se va a realizar la prueba de - forma que queden 7 de frente y las filas que sean necesarias de fondo. Cerciónece de que tanto las filas como las columnas estén bien alineadas y a la mayor distancia posible. En caso de que las dimensiones de la sala - hagan más conveniente poner las 7 filas de fondo y no de frente como se in_ dica, utilice ese sistema.

./.

4.- El colocar 7 filas de frente (o de fondo, si la sala lo exige así) permite que, al repartir los folletos, éstos queden distribuidos de forma tal que ningún par de estudiantes contiguo (excepto los que están en diagonal) tenga la misma fórmula. Para lograr que esto se cumpla usted debe repartir los folletos (en el momento que corresponda): Según el siguiente esquema:

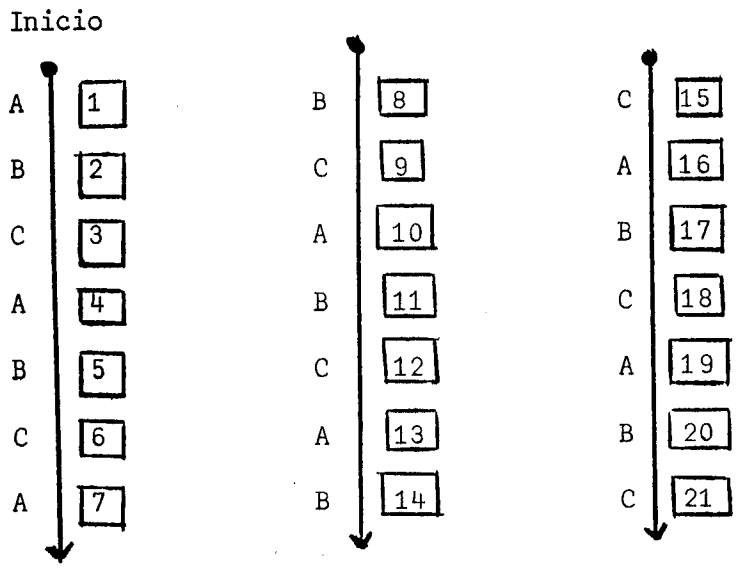
Si tiene 7 filas de frente (distribución horizontal)

Inicio



Si tiene 7 filas de fondo (distribución vertical):

./.



Observe que al finalizar cada fila, debe devolverse para iniciar la distribución de la siguiente.

5.- Pase a los estudiantes uno por uno y pídales que dejen todo lo que porten al pie de la pizarra (o en algún otro lugar adecuado para ello), excepto lápiz, lapicero, borrador y tajador.

6.- Una vez sentados todos los candidatos, preséntese y proceda a leer en voz alta las "Instrucciones para los estudiantes". Emplee la pizarra para explicar la forma en que deben marcar las contestaciones, corregirlas y codificar. Pregunte si han comprendido la explicación. Si alguno tuviera dudas, aclárelas. Para cumplir adecuadamente esta instrucción, es imprescindible que usted haya estudiado con anterioridad las "Instrucciones para los estudiantes".

DESARROLLO DE LA PRUEBA

1.- Entregue al estudiante el folleto de la prueba y la hoja para respuestas correspondiente, de tal forma que queden ordenados según se indicó en

./.

el punto 4 de la sección anterior. Asegúrece de que el color de la hoja - coincida con el folleto de prueba correspondiente.

2.- Explique la forma en que deben llenar los datos que se le solicitan en la hoja para respuestas y espere 2 ó 3 minutos para que lo hagan (el # de carnet universitario, la institución en que está matriculado, el año de - egreso del colegio y la sede o recinto no deben ser llenados. En la modalidad y la situación legal indíqueles lo correspondiente al colegio a que pertenecen).

3.- Dé la orden de inicio. A partir de este momento, la duración de la - prueba es de 1 1/2 hora.

4.- Anote en la pizarra la hora de inicio y fin de la prueba.

5.- Si se encontrara un folleto u hoja para respuestas defectuoso, debe - cambiarlo por alguno de los que usted o el coordinador tiene de repuesto, asegúrese de que sea la misma fórmula que el estudiante venía trabajando. Si por este motivo algún examinado se retrasara, deben reconocérsele esos minutos.

6.- Usted debe llamar, sólo una vez, la atención a cualquier examinado en quien observe tentativas de fraude. Si él persiste en su actitud, ordénele entregar la prueba y retirarse del aula e indíquelo en la hoja para - respuestas.

7.- Durante el desarrollo de la prueba, verifique que los examinados estén marcando correctamente en la hoja para respuestas (la equis (x) de respuesta y la cruz (+) de corrección), así como que hayan llenado correctamente

./.

los datos que se les solicitan (especialmente el nombre de la asignatura - que se está evaluando).

8.- Cualquier permiso para salir del aula debe ser autorizado por usted; - el estudiante que salga (uno cada vez) deberá dejar el folleto cerrado y - la hoja para respuestas dentro del mismo.

9.- En el momento de la aplicación de la prueba, usted no debe:

a) DEJAR AL GRUPO TRABAJANDO SOLO, NI SIQUIERA MOMENTANEAMENTE

b) HOJEAR LOS FOLLETOS DE LA PRUEBA

c) DEDICARSE A CUALQUIER OTRA ACTIVIDAD EN PERJUICIO DE LA ADECUADA - VIGILANCIA DEL PROCESO DE APLICACION DE LA PRUEBA.

d) RESPONDER A PREGUNTAS QUE SE REFIERAN AL CONTENIDO DE LA PRUEBA (PERO SI ATENDER CONSULTAS O DUDAS QUE MANIFIESTEN LOS ESTUDIANTES RESPECTO AL MISMO).

e) ABRIR FOLLETOS SELLADOS.

10.- Durante el desarrollo de la prueba anote en la hoja en blanco, entregada para este efecto, el tipo de dudas y preguntas que manifiesten los estudiantes.

FIN DE LA PRUEBA

1.- Cuando un examinado haya terminado la prueba, llámelo para que entregue el material. Alrededor de su escritorio no deben encontrarse más de un - examinado.

./.

2.- Verifique CUIDADOSAMENTE que al entregar los documentos, el examinado haya:

a) Utilizado adecuadamente las dos marcas (x y +).

b) Codificado

c) Suministrado los datos que se le solicitaron en la hoja para respuestas. "Si no lo ha hecho, devuélvale la hoja para respuestas a fin de que lo haga".

3.- Solicite al estudiante que ponga su nombre y firme en la hoja para control de asistencia y pídale que regrese cuando todos hayan finalizado la prueba (en ese momento usted sabrá exactamente a qué hora debe regresar).

4.- Coloque la hoja para respuestas, debidamente verificada, dentro del folleto de examen y éste dentro de la caja.

5.- Avise a los estudiantes cuando falten 10 minutos para concluir el período fijado.

6.- Concluido el período, llame a cada examinado a su escritorio y realice las indicaciones del "Fin de prueba".

DEVOLUCION DE LA PRUEBA

1.- Una vez que los examinados se hayan retirado:

a) Cuente los folletos

b) Ordénelos según el número de folio

c) Saque las hojas para respuestas utilizadas en el mismo orden de los folletos y colóquelas en los sobres correspondientes.

./.

d) El número total de folletos debe coincidir con el que usted recibió.

2.- Selle la caja con:

a) Los folletos debidamente ordenados

b) El sobre con las hojas para respuestas

c) La hoja para control de asistencia y materiales adicionales.

3.- Es muy importante contar con algo de retroalimentación por parte de los estudiantes, por lo tanto, si le es posible, reúnalos y solicite la opinión que tengan sobre la prueba en cuanto a dificultad, vocabulario, duración, etc. Anote los aspectos más importantes en la hoja entregada para tal fin.

4.- Una vez finalizado el período de recolección de información anterior, agradezca la colaboración que nos han brindado tanto a los estudiantes como al director del colegio.

¡MUCHAS GRACIAS POR PODER CONTAR CON SU AYUDA EN LA APLICACION DE ESTAS PRUEBAS! (NUNCA ES TARDE PARA UN AGRADECIMIENTO SINCERO)

INSTRUCCIONES PARA LOS ESTUDIANTES

1.- Buenos días (tardes, noches), mi nombre es _____,
trabajo en _____, vengo de parte de CONARE a trabajar con -
UCR, ITCR, CONARE...
uds. durante dos horas.

2.- A continuación detallaré el motivo que nos ha traido acá y en qué consiste el trabajo que van a realizar: Para las universidades es muy importante conocer los conocimientos con que ingresan sus estudiantes con el - fin de adecuar los cursos que imparte a las necesidades reales que tengan esos estudiantes.

Por lo anterior están desarrollando, coordinadas por CONARE un estudio - que les permita determinar el nivel de conocimientos básicos que un estu - diante posee al egresar de la educación diversificada.

Como parte del estudio se confeccionó una prueba que se aplicará tanto a - una muestra de estudiantes de cada una de las universidades, como a una - muestra de estudiantes que esten iniciando la educación diversificada.

Este colegio fue seleccionado dentro de esa muestra, por lo que hemos venido a que uds. nos presten su colaboración, contestando el cuestionario que les vamos a presentar.

Es importante recalcar que la calidad del estudio depende de uds., del empeño y la responsabilidad que pongan en sus respuestas y de antemano les - damos las gracias por su ayuda.

./.

Para que nosotros podamos comparar los resultados de todos los colegios - muestreados es necesario aplicar esta prueba bajo ciertas condiciones, esto es lo que en la lengua de los educadores y psicólogos se llama estandarización, por eso, se han establecido una serie de normas que rogamos cumplir al pie de la letra. A partir de este momento daré inicio al establecimiento de esas normas, ruego poner su atención y en caso de alguna duda, la podrán exponer al finalizarlas (antes de iniciar la prueba).

3.- La prueba se presenta en forma de un folleto y una hoja para respues - tas (mostrarlos) que se les entregará una vez finalizadas y aclaradas las instrucciones. Vamos a evaluar los conocimientos de _____.

4.- Deben mantener el folleto cerrado hasta que se les de la orden de iniciar la prueba.

5.- Cuentan con 1 1/2 hora para realizar la prueba. Se les avisará 10 minutos antes de que finalice el tiempo. Si terminan antes, pueden revisar sus respuesta.

6.- Durante el desarrollo de la prueba no está permitido hablar ni levan - tarse del asiento. Ai alguno tuviera necesidad de hacerlo, deberá manifestarlo levantando la mano.

7.- Al terminar el trabajo, levanten la mano para indicarles cuándo puede pasar a entregar sus documentos: el folleto de prueba y la hoja para respuestas. Recuerden que no deben levantarse de su asiento sin previa autorización.

./.

8.- No se permite el empleo de útiles tales como diccionario o calculadora.

9.- Ya que no se cuenta con la lista del grupo, rogamos que en el momento en que entreguen la prueba anoten su nombre y firmen en la hoja para control de asistencia.

10.- Los datos que se les solicitan en la hoja para respuestas deben suplirse antes de iniciar la prueba una vez entregada daré el tiempo necesario para que la llenen.

11.- Una vez leído el enunciado de cada pregunta, usted debe elegir entre las varias opciones que se le ofrecen, aquella que a su juicio responde correctamente a la pregunta o afirmación que se hace. Tomen en cuenta que algunas preguntas tienen 5 opciones de respuestas y otras tienen solo 4.

12.- Marque con una equis (x) en la hoja para respuestas el número de la opción que usted elija. Marque una sola respuesta para cada pregunta.

Por ejemplo: Si para la pregunta N°10, usted considera que la respuesta es la 3, marque así:

N°PREGUNTA	OPCIONES
10._____	(1) (2) (X) (4) (5)

13.- En el espacio que aparece junto al número de cada pregunta, anote el número de la opción que ud. eligió como correcta. En nuestro ejemplo finalmente debe aparecer así:

N° PREGUNTA	OPCIONES
10. 3	(1) (2) (X) (4) (5)

./.

14.- Si desea cambiar la respuesta, tache la equis (x) con una cruz (+), - ponga una (x) sobre la respuesta definitiva y corrija el número anotado junto al número de pregunta. Por ejemplo: Si cree que para la pregunta N°10 la respuesta correcta es la 4 y no la 3, debe aparecer así:

Nº PREGUNTA	OPCIONES
10. 3 4	(1) (2) (3) (4) (5)

15.- En caso de que honestamente en alguna pregunta, no estén seguros de - cuál es la respuesta correcta, rogamos dejarla en blanco.

16.- A la hora de marcar sus respuestas, estén atentos a que el número de - cada pregunta en el folleto coincida con el número correspondiente a esa - pregunta en la hoja de respuestas.

17.- No se dará ningún valor a lo que escriban en el folleto. Sin embargo pueden utilizar los espacios en blanco para escribir lo que deseen.

18.- Si el folleto o la hoja de respuesta de alguno de uds. tiene algún - error de impresión, levante la mano para indicarlos (en caso necesario se - le cambiará el folleto).

ANEXO N°7

CRITERIOS DE ESTRATIFICACION PARA LA SUBMUESTRA DE ESTUDIANTES UNIVERSITARIOS

ANEXO N°7

CRITERIOS DE ESTRATIFICACION PARA LA SUBMUESTRA DE ESTUDIANTES UNIVERSITARIOS

Para obtener los índices de discriminación que se describen es necesario antes de hacer comparables los grupos criterio (ambas aplicaciones) para lo cual se deberá obtener una muestra estratificada al azar de la primera aplicación con características similares a las de la segunda, en cuanto a: tamaño de la muestra, edad, tipo de colegio de procedencia, prueba realizada, sexo, año de egreso. A continuación las características de la muestra:

- Tamaño: 716 estudiantes; 179 estudiantes de cada asignatura y 58 por fórmula.

- Todos menores de 20 años (19 ó menos) y con año de egreso 1985.

- De los 58 estudiantes de cada fórmula (de cada asignatura), deben mantenerse las siguientes proporciones para las siguientes variables:

  . Sexo: 58% masculino (34 estudiantes) y 42% femenino (24 estudiantes).

  . Tipo de colegio:

  37% de colegio académico diurno oficial (22 estudiantes)

  31% de colegio técnico diurno oficial (19 estudiantes)

  18% de colegio académico diurno particular (11 estudiantes)

  12% de colegio académico nocturno oficial (6 estudiantes)

  TOTAL: 58 estudiantes

ANEXO N°8

HOJA DE COTEJO PARA EL ANALISIS DE CALIDAD TECNICA

ANEXO N°8

HOJA DE COTEJO PARA EL ANALISIS DE CALIDAD TECNICA

Objetivo número_____ Item número_____

Revisor_____ Fecha_____

Escriba una "x" bajo la columna correspondiente de acuerdo con su opinión - sobre las características del ítem; que se detallan a continuación:

	SI	CUESTIONABLE	NO
1. Está el enunciado del ítem claramente escrito para los alumnos a examinar?	—	_____	—
2. Está el enunciado libre de material irrelevante?	—	_____	—
3. Está el problema claramente definido en el enunciado?	—	_____	—
4. Están las alternativas claramente escritas para el grupo de examinados?	—	_____	—
5. Están las alternativas libres de material irrelevante?			
6. Hay una respuesta correcta o una mejor respuesta - identificable?	—	_____	—
7. Se han eliminado las palabras tales como: siempre, ninguna o todos?	—	_____	—
8. Se han empleado los posibles errores del examinando para preparar las alternativas?	—	_____	—

./.

	SI	CUESTIONABLE	NO
9. Se evita usar "todas las anteriores" o "ninguna de las anteriores" como alternativas?	—	————	—
10. Se arreglan las alternativas en orden lógico (si éste existe)?	—	————	—
11. Se colocó la clave aleatoriamente entre las alternativas?	—	————	—
12. Fueron removidas todas las palabras o expresiones repetidas de las alternativas e incluidas en el enunciado?	—	————	—
13. Son todas las alternativas aproximadamente de las misma longitud?	—	————	—
14. Se respetan las reglas gramaticales y de puntuación?	—	————	—
15. Se subrayan todas las palabras negativas?	—	————	—
16. Se eliminaron las claves gramaticales entre el enunciado y las alternativas?	—	————	—
17. Es el formato del ítem el apropiado para medir el objetivo?	—	————	—

./.

Se sugieren las siguientes revisiones

_____

_____

_____

_____

_____

Juicio final

☐	☐	☐
ACEPTARLO	ACEPTARLO (con la revisión sugerida)	RECHAZARLO

5/ Traducido y adaptado: Hambleton, R.K. Test score validity and standard-setting methods. In R.A. Beck (Ld.). Criterion-referenced measurement: The state of the art. Baltimore, Maryland: The John Hopkins University Press, 1980, p. 116-117.