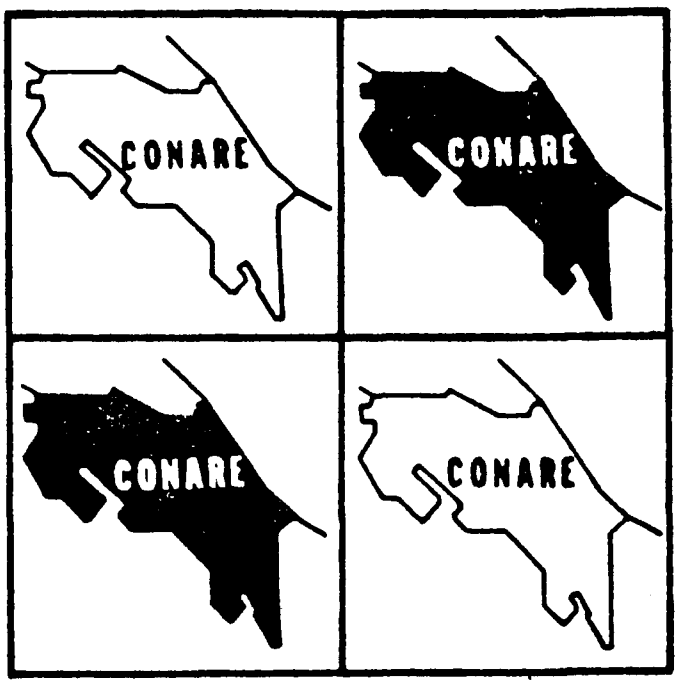
CONSEJO NACIONAL DE RECTORES
OFICINA DE PLANIFICACION DE LA EDUCACION SUPERIOR



PRUEBAS CON REFERENCIA A CRITERIOS:

UNA REVISION BIBLIOGRAFICA

DR. JUAN ML. ESQUIVEL A.

ING. MAYRA ALVARADO U.

OPES-01/87

Febrero, 1987

378
O-p
01/87

Oficina de Planificación de la Educación Superior (OPES)
Pruebas con referencia a criterios: Una revisión bibliográfica.--1.ed.--San Pedro: Sección de Publicaciones de la - OPES, 1987.
64 p.

1. Educación Superior.
I. Título

PRESENTACION

A finales de 1983 y durante los primeros meses de 1984, por iniciativa de la Comisión Interinstitucional de Estudios Relacionados con la Admisión a la Educación Superior (CIAES), se gestó la idea de realizar pruebas de conocimiento a los estudiantes que aspiraban a ingresar a las institucio - nes universitarias; esta idea contó con el apoyo de los señores Rectores y quedó plasmada en el "Proyecto Prueba Experimental de Conocimientos".

Si bien desde un principio se había pensado en la conveniencia de que, como parte del trabajo de la CIAES, se llegasen en su momento a elaborar este tipo de pruebas, las circunstancias que se dieron en ese entonces en torno a las notas de los estudiantes del Ciclo Diversificado contribuyeron en un alto grado a que se hiciera un esfuerzo especial para elaborar las pruebas y que éstas sirviesen a corto plazo como un elemento básico "para perfilar técnicamente el problema de la preparación del estudiante que egresa de la Educación Diversificada".

El paso siguiente lo constituyó la elección de la metodología que ser - viría de base a las pruebas. Se optó para ello, luego de consideradas las alternativas y los propósitos que orientaban su elaboración y utilización, la de las pruebas de conocimientos referidos a criterios, ya que ella re - presentaba, en palabras de la Comisión "el mejor medio de lograr informa - ción útil sobre los conocimientos reales que maneja el estudiante y las consecuencias que en la admisión y en las políticas académicas acarrea este hecho".

./.

De lo antes expuesto, se deriva que el interés primordial de la CIAES - fue el de poder contar con los medios adecuados para determinar cuánto saben de lo que deberían saber los estudiantes a punto de concluir sus estudios secundarios.

La especialización y relativa novedad de la metodología a ser empleada requirieron que la CIAES contase con el debido asesoramiento técnico para la elaboración definitiva del proyecto. Dicha asesoría correspondió al - Dr. Juan Manuel Esquivel, destacado científico educativo costarricense.

El Dr. Esquivel procedió a efectuar una minuciosa revisión bibliográfica, procurando obtener el material más reciente publicado sobre el tema. Para este efecto se realizaron consultas bibliográficas internacionales - con el objeto de obtener artículos publicados aun en 1984, los cuales se - clasificaron por temas y se estudiaron.

La Ing. Mayra Alvarado participó junto con el Dr. Esquivel en la revi - sión bibliográfica y en la redacción de este informe; asimismo, la Ing. - Alvarado fue la coordinadora de la ejecución del proyecto y el elemento - fundamental para que el mismo fuese concluido con todo éxito.

El encargo inicial encomendado al Dr. Esquivel se plasmó en el trabajo "Pruebas con referencia a criterios: una revisión bibliográfica", al cual estas notas sirven de presentación.

Ha sido de especial interés del Dr. Esquivel y de la Ing. Alvarado, así como de los miembros de la CIAES, el que este documento sea publicado y - que llegue a aquellos investigadores educativos vinculados con el tema, de

./.

bido a la escasa bibliografía disponible en el país y a que la mayoría no está en idioma español.

La mayor parte del material indicado en la bibliografía está disponible en la Biblioteca del Consejo Nacional de Rectores, debidamente clasificado por temas, para aquellos investigadores que deseen consultarlo y profundizar en este campo educativo.

José Andrés Masís Bermúdez
Director OPES

INDICE

./.

INTRODUCCION

En 1984, el Consejo Nacional de Rectores (CONARE), por medio de la Comisión Interinstitucional de Estudios relacionados con la Admisión a la Educación Superior (CIAES), se trazó la tarea de desarrollar pruebas de conocimientos, dentro del marco conceptual de la medición con referencia a criterios, para diagnosticar los conocimientos de los alumnos de primer ingreso a cada universidad.

Como fase inicial de ese proyecto se realizó una búsqueda de la bibliografía existente sobre el tema, tanto dentro como fuera del país. Para la búsqueda en el extranjero se utilizó el servicio que ofrece el Instituto Tecnológico de Costa Rica, sobre consultas a bases de datos especializadas.

Esta primera etapa dio como resultado una bibliografía de más de 700 títulos, una gran mayoría de ellos no existentes en el país, de los cuales se hizo una selección de 120 títulos.

La revisión exhaustiva de esos artículos que datan de 1963 a 1984, y su interpretación, dio como producto el presente trabajo (Pruebas con Referencia a Criterios: una Revisión Bibliográfica) y los lineamientos metodológicos necesarios para el desarrollo de las pruebas.

Para el desarrollo de esta revisión bibliográfica se ha dividido el tema en cinco grandes aspectos: características generales de las pruebas con referencia a criterios; validez; confiabilidad; desarrollo, selección y análisis de ítemes y otras consideraciones técnicas.

./.

Este trabajo está dirigido a aquellos investigadores en el campo de la evaluación y la medición que estén interesados en conocer la teoría que sustenta este tipo de pruebas, con la intención de contribuir al mejoramiento de la investigación que en este campo se desarrolle, en el país.

PRUEBAS CON REFERENCIA A CRITERIOS:

UNA REVISION BIBLIOGRAFICA

Pruebas con referencia a criterios

Definición

Desde que apareció en 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 cüales 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: escritura, 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 matriculados 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 tuvieron mejor rendimiento en los ítemes de palabra 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 ítemes 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 entre 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 es 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) Plazos 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=1b; (a<b) ε U; {H, V}, donde A y

./.

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 imprâctico...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 descriminen (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, estadísticos, 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). El 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 arbi - trariedad 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 clasifi cación 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 rendi - miento de una persona con capacidad de maestría mínima. Entre estos méto - dos 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 cualquiera 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 examonado 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).

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 NºED170382) (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 NºED170381) (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 Educational 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 pre- sented 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 Mea- surement, 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 indivi - dualized 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 pro - grams. 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 gruy ter, 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 N°ED 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, N°3, pp. 257-266, Fall, 1981.

Jencks, C., and Crouse, J. Aptitud vs. achievement: should we replace the SAT? The Public Interest, Vol. 67, N°21, 35, Spring 1982.

Kane, M.T., and Brennan, R.L. Agreement coefficients as indices of dependability for damain-referenced tests, ACT. Technical Bulletin N°28. Iowa City. Iowa: American College Testing Program. (ERIC Document Reproduction Service N°ED 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 N°ED 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, - NºED 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.