Revista Investigación Educativa - Vol. 3 - N.º 6 - 1985 (P.304-321)

Evaluación del rendimiento criterial versus normativa. Modelo de evaluación: F.C.O.

por
Mercedes Rodríguez Lajo

INTRODUCCIÓN

La presente ponencia está estructurada en dos partes.

En la primera de ellas se esboza un análisis entre comparativo los tests referidos al criterio y los tests referidos a la norma, haciendo más énfasis en los primeros. Se inicia con un breve resumen histórico de cómo surgen éstos, ante una necesidad de más riqueza en las informaciones que las proporcionadas por los instrumentos normativos. A continuación se analizan las diferencias entre ambos tipos de instrumentos para finalizar con unas anotaciones sobre la problemática que plantean los tests referidos al criterio en cuanto a la validez y fiabilidad se refiere.

Finalizada esta parte, que equivale a una síntesis teórica, un resumen esquemático del proceso de evaluación introduce a la segunda, que es la descripción de un modelo de evaluación criterial desarrollado por la ponente.

El modelo, denominado F.C.O., está construído siguiendo los pasos de un modelo trifásico de evaluación del rendimiento (preparación-implementación-interpretación) y su interés estriba en que contempla integralmente todos los aspectos del mismo, pudiendo, en consecuencia, ser una herramienta muy eficaz para el control total de un curso académico. A lo anterior coayuda el hecho de que el modelo está totalmente informatizado en un ordenador personal.

En el anexo se facilitan unos gráficos de control proporcionados por el modelo, de sumo interés en el mismo.

1. Surgimiento de la evaluación criterial frente a la normativa.

La evaluación criterial nace como consecuencia de una convergencia de hechos acaecidos en el cambio de década 1950-60: difusión de la enseñanza «pro-

gramada», el desarrollo de las teorías del aprendizaje y el auge de los objetivos conductuales y el enfoque de la enseñanza basada en objetivos. Se crea así un clima social que pone de manifiesto la necesidad de nuevos instrumentos que aporten una mayor información sobre lo que «un sujeto en capaz de hacer». Estos son denominados «tests referidos al criterio» (TRC), término propuesto por vez primera por Glaser, R. en 1963, y que define cómo: «Los que dependen del estatus absoluto de cualidades del estudiante, mientras que un test orientado a las normas dependen del estatus relativo.»

La asumición de la normalidad hacía que los instrumentos existentes se elaboraran y diseñaran de forma tal que permitieran establecer diferencias entre los sujetos comparándoles entre sí. En los tests así construídos, (tests referidos a la norma –TRN–) la puntuación directa da una información global de ejecución del sujeto y las puntuaciones se interpretan en función de la ejecución del grupo.

Ya Thorndike en 1913 señalaba cómo un sistema de calificaciones comparativas no proporcionaba una imagen clara de las habilidades del sujeto. Popham, W.J. (1978), en la misma línea, añade que «su excesiva generalidad lleva a desfases entre lo que se evalúa y lo que se enseña; el no proporcionar pistas suficientes para poner remedio a programas educativos y el emplear procedimientos técnicos en la construcción de instrumentos que tienden a eliminar los ítems más relevantes».

Desde su nacimiento, los TRC despiertan un gran interés, proliferando rápidamente los escritos sobre ellos, señalándose ya en 1978 por Hambleton, R.K. y colaboradores más de 600 referencias bibliográficas. Ello produce un lógico confusionismo terminológico. Así Gray (1978) señala 57 definiciones distintas en 40 autores, demostrando que no sólo autores distintos utilizan distintos términos sino que incluso algunos autores son inconsistentes en un mismo artículo.

Así (Glass, G. 1978), se identifican «criterio» y «estandar» o luego «criterio» y «objetivo conductual», conceptos que posteriormente aparecen claramente diferenciados. Nacen expresiones no exactamente equiparables: «tests referidos a objetivos», «de competencia mínimas», «de destrezas básicas», «de dominio», etc.

Nitko A.J. (1980) tras analizar todos los instrumentos elaborados hasta 1978 llega a la conclusión de que no hay un «prototipo» de TRC, sino que es un concepto que ha emergido con múltiples facetas pero con un proposito común: *«el emplear como referente para la interpretación de puntuaciones un dominio que debe estar previamente bien definido».*

Destacamos, finalmente, las aportaciones que hace Hambleton, R.K. (1980) en la línea de la definición anterior.

- Términos como «objetivos, «competencias» y «destrezas», pueden usarse indistintamente.
- La definición del dominio no tiene porqué incluir una referencia respecto a una puntuación de corte (cut-off) o estándar.
- Cuando se mide más de un objetivo en un test, los ítems del mismo están organizados en subtests, que se corresponden con cada objetivo.

A la síntesis señalada de las conceptualizaciones de Nitko y Hambleton, con las que estamos de acuerdo, añadiremos que: el concepto de «criterio» en los TRC se refiere a un dominio de contenido y de comportamiento hacia el cual se refiere la puntuación del test.

2. Diferencias entre los T.R.C. y los T.R.N.

La mayoría de autores que describen las diferencias entre estos dos tipos de instrumentos (Block, J.H. 1971; Glaser, R. y Nitko, A.J. 1971; Popham, W.J. y Husek, T.R. 1969) coinciden en afirmar que no se distinguen a simple vista.

Las diferencias se sitúan en torno a estas tres áreas:

A. Especificación del contenido de dominio medido por el test.

En los T.R.C. el dominio tiene que ser descrito de forma mucho más exhaustiva. Se pretende llegar a una descripción del mismo que sea lo suficientemente breve para resultar útil y por otra, que circunscriba y delimite suficientemente la clase de conductas a estudiar, para reducir ambiguedades. Para ello se elaboran una serie de estrategias, que a su vez facilitarán, posteriormente, el muestreo y la selección de ítems, como: «los objetivos de comportamiento, «las formas de ítems» (Hively, W.B.; Pattersson H.L. y Page, S.A. 1968), «los objetivos ampliados» (Millman, J. 1974), «las especificaciones del test IOX (Popham, W.J. 1983) etc.

B. Desarrollo, análisis y selección de ítems.

En los *T.R.N.* los ítems:
* Se *contruyen* en función de objetivos.
* Se *analizan y seleccionan* no sólo en función de su adecuación al objetivo sino también de sus características estadísticas, eliminando los no discriminativos. Predominan procedimientos empíricos «a posteriori».

En los *T.R.C.*,
* *Se construyen* a partir del universo de dominio «bien definido».
* *Se analizan y seleccionan* las agrupaciones de ítem/objetivo («clusters») buscando que los mismos sean representativos. Predominan los procedimientos «a priori» de análisis y se utilizan métodos de muestreo en la selección, según las características del dominio.

C. Interpretación de las puntuaciones.

Si, como afirma Tenbrink, T.D. (1981), un referente es «aquello a lo que uno se refiere cuando formula un juicio» en los T.R.N. el referente es relati-

vo (la ejecución del grupo) y en los T.R.C. el referente es absoluto («el dominio»).

Otra diferencia, señalada por Mehrens, W. (1982) es que en los T.R.N. el número de objetivos muestreados suele ser más amplio que en los T.R.C. Estas pruebas cubren menor ámbito de generalidad.

Las diferenicas señaladas parecen indicar que los T.R.N. sirven mejor a propósitos sumativos mientras que los T.R.C. sirven mejor a propósitos formativos. Paralelismo que no tiene porque darse. Si bien es cierto que para propósitos formativos es necesaria la referencia criterial, la información normativa también será útil y, viceversa, la referencia criterial será de sumo interés en los propósitos sumativos. Toda prueba se construye de acuerdo con los objetivos y la riqueza informativa que de ella se obtenga dependerá de la rigurosidad con que se ha construido, en cuanto a su representatividad. Sin representatividad no podrá realizarse la interpretación criterial, pero sí la normativa.

3. La validez y fiabilidad en las pruebas de referencia criterial: problemas que plantean.

Dadas las características de estos instrumentos se presenta una serie de problemas que afectan directamente a su valoración.

La Fiabilidad no es una propiedad intrínseca de los instrumentos, sino de éstos cuando se aplican, se refiere a los resultados obtenidos. Será una condición necesaria para la validez, pero no suficiente.

En torno a la fiabilidad se presentan dos problemas:

1) La reducción de la variabilidad de las puntuaciones.

En la Psicometría clásica los índices para determinarla se basan, en general, en el cálculo del coeficiente de correlación de Pearson. Como éste depende de la variabilidad de las puntuaciones, si ésta es baja, también lo será el coeficiente de fiabilidad. Este hecho, constatado por muchos autores, ha llevado a decidir que el método correlacional puede ser engañoso e inadecuado. Habrá que interpretarlos con cautela y si la varianza es reducida, descartarlos.

Prefieren por eso hablar de «índices de acuerdo», más que de «coeficientes de fiabilidad» pues dichos enfoques no están definidos ni interpretados en componentes de varianza.

2) El contexto de la decisión.

La naturaleza de las alternativas de decisión influye sobre el tipo de procedimientos de fiabilidad que se empleen.

Surgen distintas conceptualizaciones de fiabilidad y procedimientos alternativos para su estimación. Hambleton R.K. y colaboradores (1978) distinguen *tres conceptualizaciones de fiabilidad* en torno a las cuales se pueden clasificar los métodos.

a) Fiabilidad como medida de acuerdo en las decisiones de clasificación.

Se enfatiza en la consistencia de las decisiones cuando se repite un instrumento en un mismo grupo a través de formas paralelas o aleatoriamente paralelas.

Estarían, entre otros, los procedimientos de Hambleton, R.K. y Novick, M.R. (1973); Huynh, H. (1976) o Subkoviack, M.J. (1976). Se denominarán «índices de pérdida de umbral». Implicarán puntuaciones, «cut-off», en función de la cual se tomarán las decisiones y las pérdidas asociadas a los errores de decisión, se consideran de igual importancia.

b) Fiabilidad de las puntuaciones cut-off.

Se enfatiza en la consistencia de las desviaciones de las puntuaciones respecto a las puntuaciones cut-off. Estarían, entre otros, los procedimientos derivados de la teoría de la generalización, índices de Brennan, R.L. y Kane, M.T. (1977); Kane, M.T. y Brennan, R.L. (1980) y Livingston, S.A. (1972). Se donominan índices de «pérdida de varianza». Implican una puntuación de corte y las pérdidas asociadas a los errores de decisión no se consideran de igual importancia.

c) Fiabilidad del dominio del individuo.

Se enfatiza en la consistencia de las puntuaciones respecto al dominio estimado de los sujetos, sin hacer referencia a las puntuaciones de corte. Entre otros procedimientos estarían los de Lord, F.M. y Novick, M.R. (1968) y Brennan, R.L. (1980).

La Validez

Características más importante de los instrumentos que lleva a tener que determinar sobre el mismo desde su mismo objeto de medida: ¿qué mide?, hasta su finalidad: ¿para qué lo mide?, pasando por el: ¿cómo y de qué forma lo mide?; en donde no sólo deben considerarse aspectos organizativos y estructurales del mismo, sino que también dependerá de la fiabilidad de las puntuaciones.

Hay que señalar, sin embargo, que a pesar de la numerosa contribución literaria, la validez de los T.R.C. *no está resuelta de forma satisfactoria.* En ellas seguirá siendo una cuestión importante pero variará el grado de interés puesto en cada uno de los distintos tipos de validez. Se referirá también «a la interpretación de las respuestas» y a «las decisiones» que de ellos se tomen. Veamos de forma muy sintética los distintos tipos de validez en los T.R.C.

La validez de criterio o funcional.

Será útil cuando queramos saber el grado en que un instrumentos covaría con otros que pretendan medir lo mismo y cuando nos interese saber si el instrumento discrimina entre sujetos clasificados sobre criterios diferentes.

Se presentarán los siguientes problemas para su determinación: la validez de los criterios externos que sirven de apoyo, la falta de variabilidad de las puntuaciones y la influencia que sobre la validez ejerce la fiabilidad.

Muchos autores opinan que en el contexto de los T.R.C. es irrelevante, considerando que el instrumento ya, por sí mismo, es un buen criterio.

Se sugerirán distintas técnicas como: los grupos contrastados (Haynes, 1978) o la estadística discriminativa (Mehrens, W. 1982).

La Validez Descriptiva o de Contenido.

Es lo más importante, considerándose en general, que por sí sola es suficiente para valorar estos instrumentos. Según Berk, R.A. (1980, implicaría:

a) *La especificación del dominio.*
b) La validez de ítems. Existiendo distintos procedimientos: «Judiciales» y «Empíricos» para establecer la congruencia entre los ítems y las especificaciones de dominio.
c) La calidad técnica de los ítems.
d) *La representatividad* de los mismos para que permitan generalizar.

Recogemos, a modo de conclusión, las siguientes palabras de Linn, R.L. (1980): «La validez de contenido proporciona una excelente fundamentación para los T.R.C., pero para soportar la validez de un instrumento se necesita, además, la validez de las inferencias y de los usos del mismo». Es decir, es un tipo de *validez necesaria pero no suficiente.*

La Validez de Constructo.

Ha sido infravalorada en el marco de la evaluación criterial. Atenderá a la *significación de las puntuaciones* del test en relación con su estructura inicial.

Las especificaciones de los objetivos, que se realiza en estos instrumentos, es similar a lo que Kerlïnger denomina «definición operativa del constructo».

Martínez, M.R. (1981) destaca este hecho señalando como los constructos son los que determinan qué conductas han de seleccionarse para su observación y la situación en que será aplicada y como todo sistema de medida está siempre ligada al sistema de constructos dentro de cuál fue contruido.

A este tipo de validez no se le ha dado la importancia suficiente.

La Validez de Decisión o de Utilidad.

Es quizás el logro más importante, pero a su vez más controvertido, de los T.R.C. a la validez.

Un instrumento será válido si es útil a los propósitos para los cuales fue elaborado. Si de ellos deben desprenderse decisiones, será importante determinar su adecuación. Se analizará este tipo de validez en función de los distintos contextos de decisión.

Así Cronbach, L.J. (1970) diferencia entre dos niveles: las que atañen a individuos y las que atañen a grupos.

Shepard, L. (1980) diferencia, dentro de las individuales, entre: para propósitos instructivos y para «certificación».

Siempre que se tomen decisiones concernientes a la clasificación habrá que establecer puntuaciones cut-off.

La «fijación de los standars» será el problema fundamental de la metodología de los T.R.C. Su importancia nos la señala Hambleton, R.K. (1982) cuando afirma: «por muy acertado que sea técnicamente un test o por mucha validez de contenido que tenga, todos los esfuerzos serán vanos si no se presta suficiente atención al proceso de establecer standars».

Habrá distintos tipos de standars (para cada objetivo, para el conjunto de objetivos, o para la ejecución del grupo) y distintos procedimientos para establecerlos (Meskauskas, J.A. 1976; Glass, G. 1978 y Martínez, M.R. 1982).

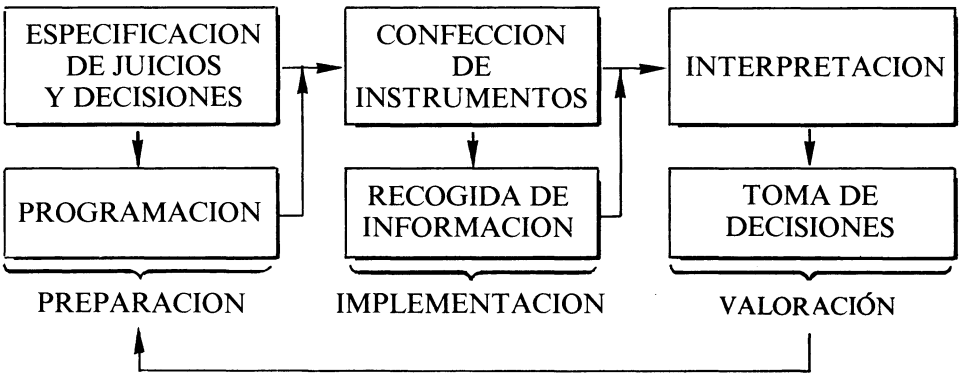
El hecho de que al aplicar distintos métodos a un mismo test que tiene un único propósito, se obtengan standars diferentes ha hecho que se ponga en tela de juicio los métodos. Será un problema que afectará directamente a la validez de la decisión. Lo importante, después, será determinar que las decisiones que se tomen son las correctas.

Los conceptos de validez y fiabilidad se funden en el marco de la Teoría de la Generalización (Cronbach, L.J. y col. 1972).

LA EVALUACIÓN DEL RENDIMIENTO: EL MODELO F.C.O.

El proceso de evaluación puede ser representado esquemáticamente por el siguiente modelo trifásico, en el cual, como puede verse, todas las fases están concatenadas y relacionadas.

PROCESO DE EVALUACIÓN. MODELO TRIFASICO



El modelo de evaluación F.C,O. que se presenta en esta ponencia, objeto de mi tesis doctoral, aún en fase elaboración, sigue el modelo trifásico expuesto y siendo eminentemente criterial posee elementos normativos complementarios.

El Modelo «F.C.O.»

A continuación se presenta un modelo de evaluación criterial del rendimiento académico mediante pruebas de elección múltiple.

A dicho modelo le hemos denominado F.C.O., iniciales, respectivamente, de «facilidad» y «cut-off», debido a que la característica más sobresaliente del mismo es, como veremos, que las distintas puntuaciones «cut-off» se establecen automáticamente a partir de las facilidades de los ítems que intervienen en las pruebas.

Una faceta fundamental del modelo que se presenta es que está totalmente informatizado en un ordenador personal. Ello permite al educador la utilización del mismo en su casa, según sus disponibilidades de tiempo.

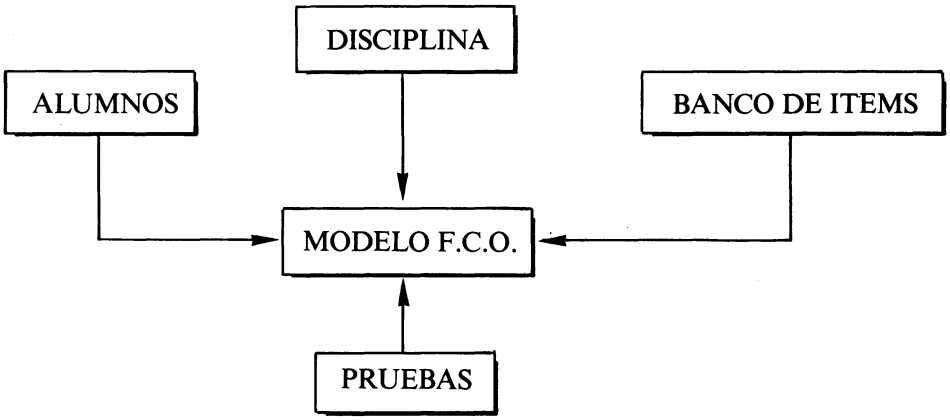
Debido a esta informatización, el modelo, integrado junto con otros procesos, permite el control y seguimiento al día del curso académico del cual es responsable el educador.

1. Componentes que intervienen en el modelo

Dado que el modelo controla el desarrollo de un curso académico veamos a continuación qué componentes son precisos para el control del mismo.

Evidentemente el primer componente es la población de sujetos a los cuales se desea controlar. Otra componente es la disciplina que se enseña, la cual debe

MODELO QUE CONTROLA UN CURSO ACADEMICO



estar «bien definida» y estructurada coherentemente «a priori» al inicio del curso. Finalmente, son precisas unas pruebas periódicas para efectuar las sucesivas evaluaciones.

Las componentes anteriores son incorporadas en el modelo como archivos, los cuales se van relacionando y actualizando convenientemente. Se incorpora, asimismo, un archivo auxiliar, pero de suma importancia, que es el banco de ítems, el cuál recoge la experiencia y resultados históricos de todas las pruebas realizadas anteriormente.

Las informaciones que recoge cada uno de los archivos anteriores, escuetamente, son las siguientes:

A. Archivo DISCIPLINA.

Es donde se guarda lo que constituye el «Dominio» de la materia. Esta es considerada como «un todo». El «Dominio» está definido y acotado (cuya adecuación o bondad se debe validar) por los «objetivos específicos» $\{w_j\}$. Así la disciplina $\{D\}$ se equipara a un conjunto de objetivos.

Por otra parte, cada objetivo se articula en «especificaciones operativas» agrupadas en torno a competencias concretas.

Para cada objetivo debe establecerse asimismo un valor escalar denominado «peso» (π_i) el cual marca la importancia de cada uno de ellos en el dominio de la disciplina $\{D\}$. Se obliga en el modelo a que $\sum \pi_i = 100$.

Por otra parte, debe determinarse «a priori» si un objetivo es considerado «esencial» o no, ya que los que posean este atributo son evaluados por el modelo más rígidamente, como se verá posteriormente.

El procedimiento seguido para establecer todo lo anterior es a criterio del educador. Sin embargo, nosotros hemos optado por el sistema judicial, existien-

do marginalmente impresos de encuesta a jueces y algoritmos programados para su ejecución.

El archivo resultante se compone, en consecuencia, de las siguientes informaciones:

- Código del Objetivo
- Descripción
- Peso
- ¿Es esencial?
- Número de especificaciones operativas componentes
- Etc.

- Número de especificación operativa
- Descripción
- Objetivo a que pertenece
- Facilidad histórica medida de la especificación
- Etc.

B. Archivo ALUMNOS.

Corresponde a la población de control. Equivale a la ficha individual que se realiza a inicio de curso. Contiene, en consecuencia, de entrada, informaciones tales como:

- Número de alumno
- Nombre
- Dirección
- Sexo
- Experiencia previa en la disciplina
- Etc.

A estas informaciones fijas el modelo actualiza, posteriormente, unas informaciones variables que son las distintas puntuaciones que se van obteniendo a partir de las pruebas efectuadas.

Tipos de puntuaciones que se recogen, fruto de la aplicación del modelo:

q_i = puntuación relativa de cada invididuo en cada objetivo $\{0_i \div 10\}$.

$|q_i|$ = puntuación absoluta de cada individuo en cada objetivo = $q_i \times \pi_i$.

Q = Puntuación total de cada individuo en una prueba. Al valor máximo posible que se puede obtener se le denomina $Q^{max.}$

Q_{10} = puntuación en base 10 de una prueba = $\frac{Q}{Q_{max.}} \times 10$.

Q^D = puntuación total del individuo en la disciplina hasta el momento.
Es igual a la suma de las puntuaciones Q obtenidas anteriormente.
Al valor máximo posible obtenible hasta el momento se le denomina $Q^D_{max.}$

Q^D_{10} = puntuación de base 10 de la disciplina hasta el momento = $\frac{Q^D}{Q^D_{max.}} \times 10$.

Así la puntuación del sujeto se recoge totalmente desglosada en los componentes que la han formado, pudiéndose analizar con detalle como se ha llegado al resultado.

C. Banco de ítems.

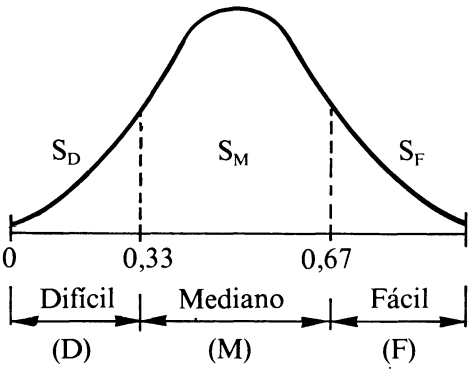
Cada ítem debe ir asociado a una especificación operativa y sólo a una. En consecuencia, queda asociado también a un objetivo específico.

Un ítem sólo se incorpora al banco cuando haya sido testado por primera vez, recogiéndose una serie de datos estadísticos asociados al mismo. De ellos es fundamental la «facilidad», la cual va actualizándose en las sucesivas pruebas en que interviene el ítem, tanto en media ($\bar{f}$) como en desviación típica (σ_f).

Para el modelo, más que el valor numérico, como se verá, es básico saber si un ítem es «fácil», «difícil», o «mediano». Por ello se define el concepto *«facilidad simbólica»* (φ), que es la proyección de la historia de la facilidad del ítem (supuesta normal) sobre un escalado de la facilidad acotado como indica el gráfico adjunto. Las áreas en que queda cortada la curva normal marcan unos «componentes de la facilidad». Así:

$\varphi = \frac{S_D}{S_D + S_M + S_F}\ D + \frac{S_M}{S_D + S_M + S_F}\ M + \frac{S_F}{S_D + S_M + S_F}\ F$. Si se tiene, por ejemplo:

$\varphi = 0{,}230^{D} + 0{,}73\ M + 0{,}04\ F$, se interpretará como que el ítem es considerado un 23 % difícil, un 73 % mediano y un 4 % fácil. Evidentemente, un ítem recien incorporado al banco, por carecer de σ_f sólo será $\varphi = D$, $\varphi = M$ o $\varphi = F$, sin componentes.



Cuando se elabora un ítem de nueva creación se estima una facilidad simbólica sin componentes igual a la de la especificación operativa.

El archivo se compone así de los siguientes datos:

- Número de ítem
- Descripción
- Alternativa correcta
- Número de especificación operativa asociada
- Facilidad simbólica
- Etc.

D. Las pruebas

Es importante la definición de prueba en el modelo F.C.O. *«Una prueba es un conjunto de objetivos que se testan simultáneamente».* Lo anterior se recalca pues en el proceso de corrección de una prueba, el modelo puntúa cada objetivo aisladamente integrando posteriormente los resultados. En síntesis, es cada objetivo individual el que se evalúa, independientemente de cuáles otros estén presentes simultáneamente.

Dado que cada objetivo tiene un peso en la disciplina debería verificarse la condición de que *todos los objetivos tendrían que ser testados a lo largo del curso en pruebas sucesivas,* simplemente por una exigencia aritmética.

El modelo permite, sin embargo, una sutileza en la afirmación anterior transformándola en la de que *todos los objetivos han de ser presentados a test* a lo largo del curso. Entonces, si en una prueba están presentados n_1 objetivos y, en realidad, sólo se testan n_2 ($n_2 < n_1$) existe la técnica de «concentración de pesos» que incrementa los pesos de los objetivos testados, absorbiéndo los de los presentados y no testados por reparto proporcional. Se permite así que un objetivo no sea testado si las necesidades del curso lo requieren. Sin embargo, dado que la puntuación final de la disciplina se obtiene por la agregación de las puntuaciones alcanzadas en cada objetivo por sus pesos respectivos, *un objetivo sólo puede ser testado una vez a lo largo del curso.*

Todo lo anterior es válido para las pruebas que denominamos «sumativas», pero el modelo contempla otros tipos de pruebas. De ellas, las más importantes son las pruebas «formativas», las cuales no cumplen las condiciones anteriores pues un mismo objetivo puede ser testado en una formativa y en una sumativa. Sus características fundamentales son que tiene un contenido más reducido y que las puntuaciones de sus objetivos, aunque se obtienen por los mismos algoritmos que las sumativas, sólo se integran en la puntuación total de la disciplina de forma transitoria, considerándose estimaciones de las puntuaciones que obtendría el sujeto en una prueba sumativa de los objetivos involucrados.

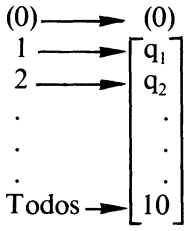
Debido a este carácter de transitoriedad las puntuaciones de los objetivos son sustituidas por las alcanzadas en una prueba sumativa posterior que los teste

nuevamente. El modelo analiza mediante unos índices de control que el nivel de puntuaciones de cada objetivo se eleva en la prueba sumativa respecto a la formativa. Ello indica así que el aprovechamiento del curso es adecuado.

Antes de pasar al contenido de los archivos de pruebas debe comentarse que el modelo posee *tres subsistemas diferentes* relativos a las mismas.

1. *Confección de la prueba.* A partir de los objetivos presentados a test, se determina el número mínimo de especificaciones operativas necesario para alcanzar la representatividad y a continuación se selecciona un ítem (del banco o de nueva creación) para cada una de ellas.

Es de destacar que la prueba queda totalmente definida una vez seleccionados los ítems de cada objetivo. El baremo de puntuaciones por objetivo es el denominado *vector cut-off* el cual especifica la puntuación alcanzada, en función del número de aciertos conseguido. Ello se efectúa posicionalmente. Así, el primer elemento señala la puntuación relativa para un acierto, el segundo la conseguida para dos aciertos, etc. El último elemento del vector (correspondiente a todos los ítems acerdatos) siempre es 10. Existe, finalmente, una línea virtual que es 0 aciertos $\rightarrow$ de puntuación relativa, la cual ya no se pone por existir siempre.



El interés del método estriba en que *conocidas las facilidades simbólicas «φ» de los ítems que componen un objetivo,* existen unos modelos matemáticos basados en curvas exponenciales, lineales y logarítmicas que *determinan automáticamente* el vector cut-off correspondiente.

Por otra parte, se determinan automáticamente unos coeficientes denominados *influencias* que *realizan* las puntuaciones alcanzadas *en el caso que el nivel de superación de los objetivos considerados esenciales* no haya superado unos mínimos preestablecidos. (Estos sólo son de aplicación en las pruebas sumativas).

Finalmente, el educador, si lo considera conveniente, construye *permutaciones* debiendo establecer tablas de correspondencia para los números de ítem entre una permutación y otra.

2. *Recogida de respuestas.* Las respuestas de los sujetos se introducen al sistema para su evaluación posterior.

3. *Evaluación.* A partir de las características de la prueba (alternativas correctas, permutaciones, vectores cut-off, coeficientes de influencia y pesos por objetivo) se evalua cada objetivo aisladamente, integrándose los resultados posteriormente.

Son de destacar dos puntos de interés:

– Se efectúa siempre corrección por azar, pero no al sistema clásico, sino uno más sofisticado que tiene en cuenta el número de errores producidos sólo en los objetivos no esenciales y efectuándose un reparto proporcional al número de ítems en estos objetivos.

– El número de aciertos (corregido al azar) final se supone que es una estimación de la puntuación verdadera que obtendría el sujeto, denominándosele, «puntuación observada».

En consecuencia, se pueden calcular unas probabilidades, para cada objetivo, de puntuaciones verdaderas de 0 aciertos, 1 acierto, 2 aciertos,... K. Ello se efectúa mediante una aproximación binomial con media la puntuación observada (φ).

$$p(k) = \binom{n}{k} \varphi^{k} (1-\varphi)^{n-k}$$

Así se obtiene un vector de probabilidades de número de aciertos. El producto escalar de dicho vector por el vector cut-off da las puntuaciones relativas de los objetivos.

$$\begin{vmatrix} p\ (1) \\ p\ (2) \\ . \\ . \\ . \\ p\ (n) \end{vmatrix} \times \begin{vmatrix} q_1 \\ q_2 \\ . \\ . \\ . \\ 10 \end{vmatrix} = p(1) \cdot q_1 + p(2) \cdot q_2 + \ldots + p(n) \cdot 10 = q$$

En conclusión de todo lo anterior, los archivos de pruebas contienen las siguientes informaciones:

- Tipo de prueba
- Items componentes (estos ya marcan las alternativas correctas, los vectores cut-off y los coeficientes de influencia).
- Número de permutación
- Tabla de correspondencia de ítems.

y finalmente:

- Número de sujeto (en el archivo de alumnos)
- Respuestas efectuadas.

2. Síntesis de funcionamiento.

Conocidos todos los elementos y archivos componentes del modelo podemos sintetizar el funcionamiento del mismo en los siguientes pasos:

1. Creación del archivo de disciplina.
2. Creación del archivo de alumnos.
3. Creación del banco de ítems.
4. Confección de una prueba.
5. Recogida de respuestas de los sujetos.
6. Evaluación.
7. A partir de los resultados de la evaluación:

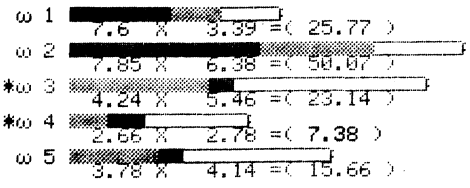
   - Confección de gráficos de seguimiento (ver anexo).
   - Actualización de puntuaciones en el archivo de ALUMNOS.
   - Actualización del banco de ítems en cuanto a facilidades y otros datos históricos se refiere.
   - Actualización del archivo de disciplina recogiendo parámetros de control para verificar la bondad de la estructuración del dominio.
   - Resultados auxiliares:

     - Evaluación normativa complementaria.
     - Análisis de normalidad de distribuciones.
     - Valoración de instrumentos.
     - Comparación entre puntuaciones sumativas y formativas, etc.

Como puede observarse, los archivos de disciplina, alumnos y banco de ítems recogen toda la información de las sucesivas pruebas, por lo que éstas, una vez evaluadas, pueden destruirse por carecer de interés posterior.

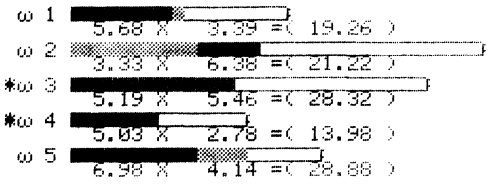
El modelo ha sido implantado en un ordenador personal Comodore-64 requiriéndose una impresora auxiliar para su funcionamiento.

SUJETO NO. 2



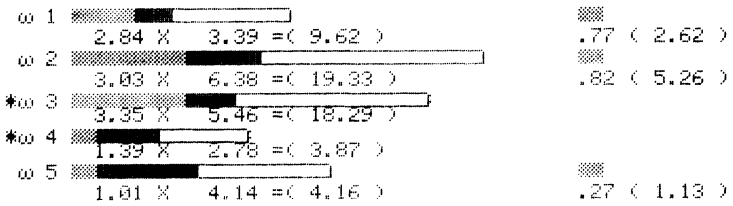
SUMA PESOS = 22.15
(TOTAL)=(122.02)
BASE 10 = 5.51

SUJETO NO. 5



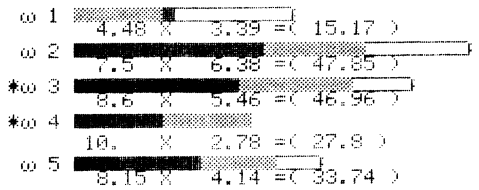
SUMA PESOS = 22.15
(TOTAL)=(111.66)
BASE 10 = 5.04

SUJETO NO. 11



SUMA PESOS = 22.15
(TOTAL)=(55.26) (9.01)
BASE 10 = 2.49 .41

SUJETO NO. 20



SUMA PESOS = 22.15
(TOTAL)=(171.53)
BASE 10 = 7.74

REFERENCIAS BIBLIOGRÁFICAS

BERK, R.A. (1980): A Framework for Methodological advances in Criterion Referenced Testing. *Applied Psychological Measurement.* 1, *4,* 563-573.

BLOCK, J.H. (1971): Criterion referenced measurement: Potential. *School Review, 69,* 289-298.

BRENNAN & KANE. (1977): An Index of Dependability for Mastery tests. *Journal of Education Measurement,* 1977, *14,* (3), 277-289.

CRONBACH, L.J. (1970): Review of Bormuth's on the theory of Achievement test Items. *Psychometrika, 35,* 509-511.

CRONBACH; GLASER; NANDA Y RAJARATHAM. (1972): *The dependability of Behavioral Measurement.* John Wiley & Sons, New York.

GLASER, R. (1963): Instructional Technology and the measurement of learning autcomes. *American Psychologist, 18,* 519-521.

GLASER, R., y NITKO, A.J. «Measurement i learning and instruction» In R.L. Thorndike (Ed.) (1971): *Educational measurement.* (2nd. ed.), Washington, American Council on Education.

GLASS, G. (1978): Standars and criteria. *Journal of Educational Measurement, 13,* 4, 237-261.

GRAY, (1978): A comparasion of Piagetian theory and criterion-referenced measurement *Review of Educational Research, 48,* 223-249.

HAMBLETON, R.K. y NOVICK, M.R. (1973): Toward and integration of Theory and Method for Criterion-referenced Tests. *Journal of Educational Measurement, 10,* 159-170.

HAMBLETON; SWAMINATHAM; ALGINA Y COULSON. (1978): Criterion-referenced testing an measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1-47.

HAMBLETON, R.K. (1980): Contributions to Criterion-Referenced Testing Technology: An Instroduction: *Applied Psychological Measurement, 4,* 4, 421-424.

HIVELY, W.; PATTERSON, H.L. Y PAGE, S.A. (1968) «A «universe-defined» System of aritmetic achievement test. *Journal of Educational Measurement, 5,* 275-290.

HUYNH, H. (1976): On consistency of Decision in Criterion-Referenced Testing. *Journal of Educational Measurement, 13,* 253-264.

KANE, M.T. y BRENNAN, R.L. (1980): Agreeement Coefficients as indices of Dependability for Domain-References Test. *Applied Psychological Measurement, 4,* 1, 105-126.

LINN, R. (1980): Issues of Validity of Criterion-Referenced Measures. *Applied Psychological Measurement, 4,* 4, 574-561.

LIVINGSTON, S.A. (1972): Criterion-Referenced Applications of Classical test Theory. *Journal of Educational Measurement, 9,* 13-26.

LORD, F.M. y NOVICK, M.R. (1968): *Statistcal theories of mental test Scores.* Reading M.A. Addison, Wesley.

MARTÍNEZ, M.R. (1981): «Principios Psicométricos de la técnica en evaluación conductual» En Fernández Ballesteros, R. y Carrobles, J.A. (Eds.), *Evaluación conductual. Metodología y Aplicaciones.* Pirámide, Madrid, 157-198.

MEHERENS, W. y LEHMANN. (1979): *Measurement and evaluation in Education and Psychology,* 1973, 1978. Traducida al castellano por Compañía Editorial Continental, S.A. México, *Medición y Evaluación en la Educación y en la Psicología,* 1982.

MESKAUSKAS. J.A. (1976): Evaluation Models of Criterion-referenced testing: Views regarding Mastery and Standard-Setting. *Review of Educational Research, 46,* 1. 133-158.

MILLMAN, J. (1974): «Criterion-referenced measurement» On J. Popham, W.J. (Ed.), *Evaluation in education: Current applications.* Berkeley, Califormina: Mc. Cutcham Publishing Co..

NITKO, A.J. (1980): Distinguishing the many varieties of Criterion-Referenced tests. *Review of Educational Research, 50,* 461-485.

POPHAM, W.J. y HUSEK, T.R. (1969): Implications of Criterion-Referenced Measurement. *Journal of Educational Measurement, 6,* 1-9.

POPHAM, W.J. (1983): *Evaluación basado en criterios.* Magistrado Español, Madrid.

SHEPARD, L. (1980): Standard Setting Issues and Methods. *Applied Psychological Measurement, 4,* 447-467.

SUBKOVIACK, M.J. (1976): Estimating reability from a single administration of a mastery test. *Journal of Educational Measurement, 13,* 265-276.

SWAMINATHAN, H.; HAMBLETON, R.K. y ALGINA. (1974): Reability of Criterion-Referenced tests: A decision-theoretic formulation. *Journal of educational Measurement, 11,* 263-276.

TENBRINK, T.D. (1981): *Evaluación. Guía práctica para profesores.* Narcea, Madrid. En el original en inglés, *Evaluation: A practical guide for teachers.*

TRAUB, R.E. y ROWLEY, G.L. (1980): Reability of test Scores and Decisions. *Applied Psychological Measurement, 4,* 517-545.