UNA ESTRATEGIA DE DIAGNOSTICO DEL RENDIMIENTO ACADEMICO (LOS TESTS BASADOS EN CRITERIOS)

VÍCTOR ALVAREZ ROJO

1. INTRODUCCIÓN

1.1. *La evaluación en el proceso de enseñanza-aprendizaje*

La evaluación del rendimiento académico de los alumnos es una función específica del profesor en cualquier proceso de enseñanza-aprendizaje. Los métodos y técnicas de evaluación contribuyen directamente a la configuración y desarrollo de dicho proceso mediante la información que suministran al docente sobre la situación de partida (claves para la planificación), la adecuación del proceso a esa situación (disfunciones) y los resultados finales de una intervención didáctica específica (cambios observados).

La evaluación se configura, pues, como un proceso inseparable de cualquier secuencia de enseñanza, cuyas funciones básicas son:

a) Determinar si los alumnos poseen los conocimientos y destrezas básicas necesarias para iniciar una unidad de instrucción *(Evaluación previa).* Esta evaluación posibilita al docente:

1. Establecer las destrezas que han de ser adquiridas antes de abordar una unidad instruccional (conocimientos prerrequisitos).
2. Situar a cada alumno en un lugar respecto a una secuencia instruccional (agrupamiento).
3. Modificar la planificación de una materia o área de conocimientos en base a lo ya adquirido por los alumnos (modificación curricular).
4. Obtener información que posibilite después la evaluación de la eficacia de la enseñanza (pretest/postest).

b) Medir el progreso en la adquisición de conocimientos y destrezas durante una unidad instruccional *(Evaluación formativa).*

c) Localizar y analizar las dificultades surgidas en el aprendizaje durante una secuencia de instrucción *(Evaluación diagnóstica).*

d) Medir los resultados finales conseguidos en una secuencia de instrucción *(Evaluación sumativa).*

1.2. *Los tests como instrumentos de evaluación*

Entre los diversos instrumentos de evaluación de que dispone el docente para llevar a cabo esta tarea se encuentran los tests. La historia de los tests, que abarca ya un siglo

de existencia desde que Cattell utilizara el término por primera vez en 1890, ha sido hasta hace dos décadas la de los tests basados en normas, ya se tratase de tests mentales o de rendimiento; es decir, se construían y validaban en relación con grupos «normativos» de población. Las puntuaciones que obtienen los sujetos sometidos a una prueba basada en normas se interpretan en relación con la «norma» que ha sido establecida por el grupo de referencia al que pertenece el sujeto; a éste, por consiguiente, se le asigna una puntuación que indica su *posición relativa* en el grupo respecto a una característica de personalidad, capacidad mental o rendimiento académico.

> «La clasificación se realiza en función de normas resultantes del examen previo de un número más o menos elevado de sujetos, lo que permite situar cada una de las respuestas, totales o parciales, en una distribución estadística» (De Landsheere, 1978).

Los tests estandarizados para la evaluación del rendimiento académico, basados en normas, tuvieron desde su nacimiento una enorme aceptación entre los docentes de EE.UU. y otros países europeos, puesto que permitían realizar con rapidez y economía de medios muchas de las funciones de la evaluación educativa (clasificar, agrupar, asignar niveles de ejecución, etc.) en las más variadas áreas de conocimientos. Sin embargo, la producción y utilización en España de estos tests ha sido, por razones que no es la ocasión ahora de analizar, muy escasa.

Fue en 1963 cuando Robert Glaser, de la Universidad de Pittsburgh, establece la distinción básica entre la medición que hace referencia a una norma y la que toma como referencia un criterio, pues considera que la primera es insuficiente para evaluar determinados aspectos de una secuencia de enseñanza-aprendizaje. La diferencia fundamental entre ambas estriba en que en la primera emitimos un juicio de valor sobre un individuo comparando los resultados que obtiene con los de un grupo normativo, mientras que en *la medición referida a un criterio* tomamos como referencia un campo de conducta bien definido, lo cual nos posibilita *determinar explícitamente qué es lo que sabe y qué es lo que no sabe hacer* en ese ámbito de conducta.

La evaluación referida a un criterio exige (Thyne, 1978) «la comparación con un patrón que *no* es una norma. No se trata de una competición entre alumnos, en la que un candidato se compara con otro. En este nuevo tipo de evaluación un alumno que no alcanzara el patrón fallaría igualmente, fuese cual fuese el número de alumnos todavía peores que él. Un alumno que alcanzase el patrón no quedaría menoscabado, aunque fuese superado por otros cualesquiera».

La evaluación referida a criterios en el ámbito del rendimiento académico ha generado un tipo de instrumentos que se conocen como «pruebas de dominio» (Tenbrink, 1983) o con la denominación más conocida de *Tests Basados en Criterios* (TBC).

1.3. *Características de los tests basados en criterios*

El diagnóstico del rendimiento académico, o evaluación diagnóstica, se inscribe, como ya hemos visto con anterioridad, en el proceso de evaluación que acompaña a toda secuencia de enseñanza-aprendizaje, y persigue la detección de los errores más frecuentes que cometen los alumnos al enfrentarse con una unidad instruccional, con objeto de dirigir a los alumnos hacia programas o secuencias correctoras específicas.

La diferencia entre la evaluación diagnóstica y la evaluación formativa es sutil, dado que en ésta se incluyen necesariamente tareas en cuya ejecución por el alumno el docente puede detectar ciertos errores típicos. No obstante, la diferencia estriba en que un

instrumento de evaluación diagnóstica debe contener una muestra representativa de todos los errores que los alumnos cometen normalmente en la adquisición de una determinada unidad de instrucción.

La evaluación del rendimiento académico es una tarea específica del profesor; es él el responsable de todas sus fases. Este principio didáctico elemental, sin embargo, no tiene por qué entrar en contradicción con el apoyo que los servicios institucionales de asesoramiento (EPOEs, EPM, etc.) le puedan prestar precisamente en este ámbito específico que es la evaluación diagnóstica y en las actuaciones subsiguientes a la misma (pedagogía correctiva).

Hechas estas aclaraciones, vamos a exponer algunas de las características más notables de los TBC.

La finalidad que persiguen los TBC es la obtención de información sobre lo que un alumno ha aprendido en una secuencia instruccional, en relación con un ámbito específico del aprendizaje, dándole ocasión para que lo demuestre mediante la realización de tareas concretas. *Los principios* sobre los que se asienta la elaboración de TBC son los siguientes (Gronlund, 1978):

1. Es necesario definir y delimitar claramente *el ámbito de aprendizaje* que se desea evaluar. La determinación del *campo de conducta,* paso esencial para la elaboración del test, varía en dificultad según se refiera a las áreas básicas del aprendizaje (v.g. conocimiento de terminología) o a los resultados escolares más complejos (v.g. capacidad de aplicar principios científicos a nuevas situaciones).
2. La elaboración de TBC exige que *los objetivos* que persigue una unidad de instrucción estén definidos *en términos de conductas observables* y, a veces, puede ser también necesario que se especifiquen las condiciones en que debe demostrarse su adquisición.
3. Se requiere, igualmente, establecer los *niveles de ejecución,* a partir de los cuales podrá decirse que los alumnos han adquirido un aprendizaje (conocimientos, destrezas...).
4. La construcción de TBC requiere que contengan una *muestra representativa* de las realizaciones de los alumnos en la unidad instruccional que va a ser evaluada, dada la imposibilidad de evaluar todas las posibles.
5. La *selección de los temas* del test debe hacerse teniendo en cuenta en qué medida reflejan la conducta especificada en los objetivos establecidos para una unidad instruccional.
6. Los TBC deben englobar un *sistema de puntuación e interpretación* que permita describir adecuadamente las realizaciones de los alumnos en un área concreta de aprendizaje.

La elaboración de TBC presenta algunas diferencias, dependiendo de los niveles de aprendizaje a los que se dirige la evaluación. Algunos autores consideran que dichos niveles son fundamentalmente dos:

a) El aprendizaje de los principios y conocimientos instrumentales básicos («minimum essentials») en cualquier área del *curriculum.*

b) El desarrollo y aplicación de conocimientos y destrezas en situaciones nuevas, en ambientes diferentes o en otros ámbitos del aprendizaje.

Los TBC son más fáciles de aplicar en el primer nivel, puesto que los límites del mismo pueden ser definidos con claridad. No ocurre así en el nivel de desarrollo del aprendizaje, pues dicho nivel es prácticamente ilimitado y el aprendizaje no sigue unas etapas bien definidas lógica y cronológicamente.

2. Proceso de elaboración de los TBC

Los pasos a seguir para la construcción de un TBC no difieren en la práctica, ya se dirija a la evaluación del primero o del segundo nivel de aprendizaje.

Por otra parte, no es necesario recordar, por obvio, que la edad, las características del grupo de alumnos y su trayectoria escolar anterior deben ser factores a tener siempre en cuenta al iniciar la planificación de un TBC.

Vamos, por tanto, a analizar cada uno de los pasos.

2.1. La primera tarea a acometer es la de *delimitar el área de conocimientos y destrezas que va a ser evaluada* (= Descripción del campo).

El área de conocimientos a evaluar no debe ser ni demasiado grande, pues exigiría un número excesivo de items en el test para obtener una muestra representativa de las conductas a evaluar, ni demasiado pequeña, dado que, en ese caso, el número posible de items se reduciría peligrosamente y repercutiría en la fiabilidad del test.

Los criterios a seguir para ejecutar esta tarea son (Popham, 1980):

a) Una descripción de campo debe ser lo suficientemente breve como para poder ser utilizable por el profesor.

b) Debe circunscribir suficientemente la clase de conductas que estudia, de tal forma que distintos observadores puedan discernir después qué items de los elaborados armonizan con el campo y cuáles no.

c) Para su delimitación es muy útil considerar la *cantidad de tiempo de enseñanza* que tardarán los alumnos en poder demostrar lo aprendido. Algunos autores aconsejan dividir el contenido de un curso en unidades de conocimientos relativamente pequeñas (una a tres semanas) y sobre ellas efectuar la descripción de los campos.

d) Otro criterio muy útil para la delimitación del área que va a ser evaluada es que nos permita producir *elementos de estímulo* (items) *homogéneos* en cuanto a sus características de contenido y formato. «Cuando comenzamos a aumentar el tamaño de un campo..., la homogeneidad de estímulo de los elementos varía, pues tratamos a veces de comprimir dos clases de conducta inevitablemente discontinuas en un campo único».

2.2. El paso siguiente consiste en especificar los *resultados del aprendizaje* que se espera que los alumnos puedan demostrar una vez finalizada la unidad instruccional a la que se dirige el test.

La forma de llevar a cabo esta tarea es sencilla. En primer lugar, se deberán formular los objetivos generales que se persiguen con una unidad instruccional. De dos a cuatro objetivos suelen ser suficientes. A continuación confeccionaremos una lista de tareas concretas que los alumnos deberán realizar para demostrar que han adquirido dichos objetivos.

Veamos un ejemplo.

Supongamos que uno de los objetivos de una unidad instruccional fuera el siguiente:

— «Conocer el significado de la terminología básica de dicha unidad».

La posible lista de tareas para poder demostrar su adquisición sería como sigue:

— Escribir una definición de cada término.
— Redactar otro significado diferente de cada término.
— Hacer distinciones entre términos de significados semejantes.

— Identificar sinónimos de un término.
— Identificar antónimos de un término.
— Identificar el significado de un término en una frase.
— Confrontar términos.

En realidad, esta segunda etapa viene a ser algo similar a lo que en una programación curricular se conoce como la elaboración de objetivos generales y objetivos operativos para una unidad de instrucción. En la medida en que estos últimos son verdaderos *objetivos conductuales,* es decir, están redactados en forma de conductas observables, más fácil resulta después la evaluación del aprendizaje de conocimientos y destrezas. Evidentemente, para la construcción de los items de un TBC esta tarea es esencial.

2.3. A continuación es necesario *confeccionar un esquema del contenido* de la unidad instruccional con objeto de saber sobre qué aspectos concretos del contenido ha versado la enseñanza y poder así disponer de referencias claras para poder redactar los items del test.

Veamos un ejemplo.

Supóngase que se ha delimitado la siguiente área de conocimientos para ser evaluada:

Los mapas de tiempo atmosférico

Los alumnos deberán demostrar un conocimiento de la terminología y de los fenómenos atmosféricos elementales que puedan ser representados en un mapa. Asimismo, han de ser capaces de interpretar mapas de tiempo atmosférico similares a los utilizados por los medios de comunicación (TV y Prensa).

Los objetivos generales que se establecieron para esta unidad instruccional fueron cuatro:

1. Conocimiento de terminología básica.
2. Conocimiento de los símbolos de representación gráfica de los elementos atmosféricos.
3. Conocimiento de hechos específicos.
4. Interpretación de mapas del tiempo.

Con el desglose de estos objetivos en tareas tendríamos una lista de conductas que especifican los resultados del aprendizaje que esperamos que los alumnos adquieran/exhiban mediante la unidad instruccional «Los mapas de tiempo atmosférico». Sin embargo, si no especificáramos más, dos profesores incluirían, con toda probabilidad, contenidos ligera o sensiblemente diferentes en esta unidad y elaborarían dos instrumentos de evaluación criterial lógicamente diferentes. De aquí que sea necesaria la descripción de los contenidos de esa unidad, que se plasma en *un esquema de contenido.* Veamos el correspondiente al de la unidad que nos ocupa (Cuadro I).

CUADRO I

ESQUEMA DE CONTENIDO PARA LA UNIDAD INSTRUCCIONAL «MAPAS DE TIEMPO ATMOSFERICO»

A. PRESIÓN ATMOSFÉRICA

1. Medida y representación de la presión atmosférica.
2. Factores desencadenantes de la presión atmosférica.
3. Relación entre presión atmosférica y cambios del tiempo.

B. TEMPERATURAS

1. Medida y representación de las temperaturas.
2. Factores reguladores de las temperaturas.
3. Regímenes térmicos.

C. HUMEDAD ATMOSFÉRICA Y PRECIPITACIONES

1. Medida y representación de la humedad atmosférica.
2. Factores generadores de la humedad atmosférica.
3. Tipos de precipitaciones.
4. Medida y representación de las precipitaciones.

D. VIENTOS

1. Medición de la velocidad y dirección.
2. Factores determinantes de la velocidad y dirección del viento.
3. Símbolos para la representación de la velocidad y dirección del viento.

E. NUBES

1. Clases de nubes.
2. Características de los diferentes tipos de nubes.
3. Causas de la formación de las nubes.
4. Interrelación nubes/tiempo atmosférico.
5. Símbolos para la representación de los diferentes tipos de nubes.

F. FRENTES

1. Clases de frentes.
2. Formación de los frentes.
3. Interrelación frentes/tiempo atmosférico.
4. Símbolos de representación de los frentes.

FUENTE: Gronlund, p. 26.

2.4. Cualquier estrategia de evaluación presenta ciertas dificultades que hay que tener en cuenta. En primer lugar, la definición del campo de conducta (conocimientos y destrezas) a evaluar en un área cualquiera del rendimiento académico es imposible de perfilar totalmente en la mayoría de los casos. De aquí que, en realidad, lo que hacemos es valuar el grado en que la conducta de los alumnos se acerca a los objetivos conductua-

les que hemos definido previamente, dando por supuesto que ellos sintetizan el campo de conducta global. Y, en segundo lugar, hemos de contar con las limitaciones inherentes a todo instrumento de medición, en cuanto a su extensión y naturaleza, en relación a la conducta que pretende evaluar (aplicabilidad); de hecho, cualquier instrumento únicamente nos permite medir una muestra limitada de las muchas posibles manifestaciones conductuales del alumno en un área determinada del rendimiento académico.

En consecuencia, una vez establecidos los objetivos y confeccionado el esquema de contenido, se impone la tarea de *seleccionar una muestra* de las tareas/resultados/conocimientos que previamente habían sido considerados como representativos de la posesión/dominio de un campo de conducta. Para ello se construye lo que se denomina *Tabla de Especificaciones del Test,* que consiste en una tabla de doble entrada; en la entrada superior (columnas) aparecen los objetivos, consignándose a su vez las áreas básicas de contenidos en la entrada de la izquierda (filas).

Las cuadrículas resultantes en la tabla de especificaciones nos indican los diversos ámbitos de la conducta en cuestión que han de estar presentes en el test. Para que el test contenga una muestra representativa del dominio de conducta que vamos a evaluar deberemos, además, especificar en dichas cuadrículas el número de items necesarios para evaluar adecuadamente, según su importancia, cada uno de los ámbitos de esa conducta. Sin embargo, no existen normas precisas sobre el número de items a emplear; como norma general se suele indicar que deben elaborarse varios items para cada ámbito —dependiendo su mayor o menor cuantía de la importancia que un ámbito tenga en el contexto de la unidad curricular— y 10 o más para cada objetivo.

Una tabla de especificaciones para el ejemplo que estamos manejando podría ser la reflejada en el Cuadro II.

2.5. Una vez elaborada la tabla de especificaciones del test podemos abordar la siguiente tarea, que consiste en *establecer niveles de ejecución* que nos permitan realizar juicios de valor sobre lo demostrado por los alumnos en relación con las diferentes áreas del contenido de la unidad instruccional: ¿Cuánto esperamos que sean capaces de adquirir (contenidos y destrezas)? ¿A partir de qué nivel de adquisiciones diremos que un alumno domina cada uno de los objetivos establecidos para un área específica de conocimientos?

Los niveles de ejecución se establecen en forma de *porcentajes.* Generalmente se considera que un porcentaje de items correctamente resueltos que se sitúe entre el 80 y el 85 por 100 es un punto de referencia adecuado para los TBC dirigidos a unidades de instrucción para la adquisición de principios y conocimientos instrumentales básicos («minimum essentials»). En el ejemplo que nos ocupa, los niveles de ejecución quedaron fijados así:

Objetivo	*Nivel de ejecución (%)*
1. Conocimiento de terminología básica	85
2. Conocimiento de los símbolos de representación gráfica	100
3. Conocimiento de hechos específicos	85
4. Interpretación de mapas del tiempo	80

Sin embargo, cabe preguntarse si esta forma de proceder no es, en definitiva, arbitraria. ¿Por qué esos porcentajes y no otros? Efectivamente, la arbitrariedad de esos niveles

CUADRO II

TABLA DE ESPECIFICACIONES PARA UN TEST DESTINADO A LA EVALUACION DE LA UNIDAD INSTRUCCIONAL «MAPAS DE TIEMPO ATMOSFERICO»

OBJETIVOS GENERALES / *AREAS DE CONTENIDO*	1. *Conocimiento Terminología básica*	2. *Conocimiento Símbolos Representación*	3. *Conocimiento Hechos específicos*	4. *Interpretación Mapas de tiempo atmosférico*	TOTAL Items
A. Presión atmosférica	2	2	2	2	8
B. Temperaturas	2	2	2	2	8
C. Humedad atmosférica y precipitación	2	2	3	3	10
D. Vientos	2	3	3	4	12
E. Nubes	3	2	3	2	10
F. Frentes	3	3	3	3	12
TOTAL Items	14	14	16	16	60

es difícilmente soslayable. Para establecer niveles de ejecución es necesario tener en cuenta los resultados que un grupo de alumnos obtiene, es decir, apoyarnos en la información que nos puede suministrar la evaluación basada en normas. Así pues, la forma de proceder sería la siguiente:

a) Establecer inicialmente niveles de ejecución, en porcentajes, teniendo en cuenta la naturaleza de los objetivos, el tipo de estímulo/respuesta que se exige al alumno (dificultad de los items) y la importancia que esos objetivos van a tener para el rendimiento académico posterior.

b) Revisar los niveles de ejecución iniciales a la luz de los resultados obtenidos en experiencias anteriores de enseñanza de la unidad instruccional en cuestión, con objeto de establecer expectativas más realistas respecto al aprendizaje en un medio educativo concreto.

2.6. *Elaboración de los items del test*

La confección definitiva de un test se plasma en la construcción de los items. Es ésta una tarea delicada, pues del material que el alumno tiene que manejar y de su correcta elaboración dependerá que los resultados sean válidos o no, a pesar de que los pasos anteriores hayan sido dados adecuadamente.

Las tareas a realizar son dos: selección del tipo de items a utilizar y redacción de los items.

a) *La elección de la clase de items* que van a utilizarse para la confección del test se centra básicamente sobre dos categorías de items (Popham, 1983): items que implican selección de respuestas e items que exigen elaborar las respuestas. Las modalidades que pueden revestir cada una de estas clases de items son:

Items de selección de respuestas:

— Elección binaria: Sí/No, Verdadero/Falso...
— Elección múltiple.
— Confrontación.

Items de elaboración de respuestas:

— Ensayo.
— Respuesta breve.

Ambas clases de items presentan ventajas e inconvenientes que sintetizaremos brevemente, dado que se saldría del cometido de este trabajo su consideración en detalle y que, por otra parte, aparecen profusamente tratados en cualquier manual de técnicas de evaluación. Los *items de selección de respuestas* son adecuados para evaluar el conocimiento de información fáctica que tiene el alumno: identificar respuestas correctas, distinguir entre alternativas, confrontar términos/cosas/elementos, etc. Los *items de elaboración de respuestas* «constituyen la única manera razonable de valorar la aptitud de los estudiantes para escribir, para sintetizar ideas o para realizar ciertos tipos de operaciones intelectuales complejas que requieren originalidad». Calcular, describir, enumerar, definir, resolver problemas... serían respuestas a exigir a través de esta clase de items.

Además de estas diferencias, hay que tener en cuenta las consecuencias que de ellas se derivan para *la calificación* de las respuestas emitidas por el alumno mediante

una u otra clase de items. La diferencia fundamental estriba en que los items de selección de respuestas pueden calificarse de manera *objetiva,* mientras que en los de elaboración de respuestas son inevitables diferentes grados de *subjetividad* en la calificación.

La decisión de utilizar una u otra clase de items debe estar guiada por el tipo de «resultados del aprendizaje» que se hayan establecido para la unidad instruccional a evaluar (véase 2.2). Hay autores como Popham que insisten mucho en la conveniencia de seleccionar una *única* clase de items para confeccionar el test, eligiendo aquella que sea más congruente con el tipo de resultados esperados y más susceptible de generalización. Un test construido a base de items de diferentes tipos permite medir de forma más representativa la conducta en cuestión; sin embargo, «a menos que deseemos hacer un test interminable, sólo podremos incluir un item, o a lo sumo unos cuantos, de acuerdo con cada una de las posibles tácticas de medición... (que) no pueden proporcionarnos una estimación fiable de esa conducta... Si, en cambio, optamos por lo primero, nuestros resultados son más fáciles de interpretar... (y) sacamos lo mejor de los dos ámbitos: un reflejo adecuado del atributo que se trata de evaluar y, además, un conjunto comprensible de resultados del test» (Popham, 1983, p. 165). Si esta recomendación no pudiera ser seguida en su totalidad, deberemos tener en cuenta, al menos, que los items elaborados para evaluar los contenidos de cada objetivo deben ser necesariamente homogéneos. No obstante, las dificultades para mantener una razonable homogeneidad del estímulo (items) remiten siempre a una revisión del dominio de conducta (= descripción del campo) establecido en la primera fase (véase 2.1).

b) *Redacción de los items del test*

Redactar items es una tarea que, siendo sencilla, hay que realizar, no obstante, de forma cuidadosa, con objeto de que resulten capaces de discriminar entre el estudiante que ha adquirido unas destrezas y el que no lo ha conseguido aún. Gronlund establece las siguientes *reglas generales* para la redacción de items:

1.ª El contenido de los items (tareas, realizaciones...) debe ser *congruente* con los resultados del aprendizaje, que en su momento se describieron para la unidad instruccional que ahora se quiere evaluar. Para determinar dicha congruencia basta confrontar ambas conductas esperadas.

2.ª La tarea a ejecutar en cada item tiene que estar definida *con claridad,* de tal forma que no exista la posibilidad de que un alumno fracase en su realización simplemente porque no entendió bien qué es lo que tenía que hacer.

3.ª La redacción de los items debe ser *concisa,* es decir, ha de eliminarse todo aquel material (digresiones, añadidos...) que no sea imprescindible para la presentación del problema a resolver o de la tarea a ejecutar.

4.ª La redacción de los items no debe contener *factores distorsionantes* que impidan al alumno responder correctamente. Estos factores pueden ser el vocabulario, complejidad de las frases y demanda de varias conductas, una de las cuales es irrelevante para el propósito del item.

5.ª En la redacción de los items debe ponerse especial cuidado para no introducir involuntariamente *claves* que guíen al alumno hacia la respuesta correcta: claves gramaticales, asociaciones verbales, redacción diferencial de la respuesta correcta, colocación de la respuesta correcta, etcétera.

6.ª El contenido de un item *no debe dar información* que pueda ser utilizada por el alumno para responder correctamente *otros items* del test.

7.ª Los items del test deben tener un *nivel adecuado de dificultad.* Los tests dirigidos a evaluar el dominio de los principios y conocimientos instrumentales básicos («minimum essentials») deben tener un nivel relativamente bajo de dificultad, determinado solamente por la propia naturaleza de los conocimientos y destrezas que el alumno debe dominar con maestría para poder enfrentarse a las siguientes unidades instruccionales. En cambio, para los tests destinados a evaluar niveles superiores de aprendizaje (desarrollo y aplicación de conocimientos y destrezas), el planificador del test debe prever, siempre que sea posible, progresivos grados de dificultad para poder situar el grado de progreso del alumno hacia la consecución de cada uno de los objetivos fijados.

El grado de dificultad de los items depende de la complejidad de los contenidos de una unidad instruccional y del tipo de resultados que se espera que los alumnos exhiban al enfrentarse con esos contenidos. A la hora de confeccionar el test se suelen situar al comienzo del mismo unos pocos items de baja dificultad para motivar positivamente al alumno y a continuación se van colocando los items en orden creciente de dificultad: desde el 95 por 100 (95 por 100 de expectativas de respuestas correctas) hasta el 5 por 100 (5 por 100 de expectativas de respuestas correctas).

8.ª Las respuestas correctas que se prevean para los items deben ser aquellas *generalmente aceptadas* por la comunidad científica. Si entran dentro de lo opinable van a presentar serios problemas para la validez del test.

9.ª Siempre que sea posible es preferible redactar los items *en forma afirmativa,* pues la forma negativa no nos garantiza otra cosa sino que el alumno sabe lo que «no es» un hecho, concepto..., pero no nos asegura que sepa «lo que es». Por otra parte, el énfasis de la enseñanza se pone generalmente en los hechos positivos y no en sus excepciones. La evaluación, pues, debe ser congruente con esa estrategia didáctica.

10.ª Es necesario redactar un *número suficiente* de items que nos permitan muestrear adecuadamente los resultados del aprendizaje que vamos a evaluar y agruparlos, en la redacción definitiva del test, por objetivos bajo encabezamientos o títulos que claramente especifiquen dichos objetivos; esto facilita después la corrección e interpretación de los resultados.

La construcción de TBC es una tarea que puede abordarse en equipo, por departamentos o por los profesores que imparten una misma materia en uno o varios centros. Para ello, una vez fijados los objetivos, resultados esperados, etc., puede ser una estrategia adecuada la creación de *un fichero de items* por objetivos.

3. ANÁLISIS DE LOS RESULTADOS DE LOS TBC

Una vez que se ha aplicado el test a los alumnos que siguieron la unidad instruccional, el análisis de los resultados ha de ser doble.

a) *Análisis de los items*

El análisis de cada uno de los items del test proporciona información al docente sobre las dificultades que el alumno encuentra en su proceso de aprendizaje; tiene, pues, una finalidad diagnóstica. Por otra parte, nos informa sobre la adecuación y efectividad de los items y sobre el progreso de los alumnos hacia los objetivos establecidos; en definitiva, acerca de la efectividad de la labor docente.

Esta tarea no presenta ninguna dificultad y se puede llevar a cabo mediante la construcción de una tabla similar a la que aparece en el Cuadro III. Se han marcado con el

CUADRO III

ANALISIS DE LOS ITEMS DE UN TBC SOBRE «MAPAS DE TIEMPO ATMOSFERICO»

OBJETIVOS	1. *Conocimiento de terminología básica*													
CONTENIDOS	*Presión atmosférica*		*Temperaturas*		*Humedad/Prec.*		*Vientos*		*Nubes*			*Frentes*		
ITEMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Alumnos:														
Juan Hurtado	+	+	+	+	+	+	+	+	—	—	—	+	—	—
Antonio Alvarez	+	+	+	+	—	—	+	—	+	—	—	—	+	+
Ricardo Díaz	+	-	+	+	+	+	—	+	—	—	+	+	—	+
Juana Pérez	—	—	—	+	+	—	+	—	—	—	+	+	—	—
Ana Rosado	+	—	+	+	+	+	—	+	—	—	—	—	—	—
Pedro Hernández	+	+	+	+	+	+	+	+	+	—	+	+	+	—
Celia Báez	+	+	+	+	+	+	—	+	+	—	—	+	—	—
Angel Alba	+	+	+	+	+	—	+	+	+	—	—	+	+	—

signo (+) las respuestas correctas y con el signo (—) las incorrectas. Como puede apreciarse, el análisis cualitativo del rendimiento académico de cada alumno es perfectamente asequible.

En el caso de que se haya realizado un pretest/postest podrían igualmente determinarse tanto los progresos y dificultades de los alumnos como la efectividad de la unidad instruccional diseñada y aplicada (Cuadro IV).

CUADRO IV

ANALISIS DE LOS ITEMS DE UN TBC EN APLICACIONES PRETEST/POSTEST

ITEMS	1		2		3		4		5		...
PRETEST (A) *POSTEST (D)*	A	D	A	D	A	D	A	D	A	D	...
Alumnos:											
Esteban Martín	—	+	+	+	—	—	+	—	—	+	
José Rojo	—	+	+	+	—	—	+	—	+	+	
Manuel Andrade	—	+	+	+	—	—	+	—	—	+	
Luis Rico	—	+	+	+	—	—	+	—	—	+	
Charo Cuevas	—	+	+	+	—	—	+	—	+	+	
Isabel Duero	—	+	+	+	—	—	+	—	—	—	
.....											
.....											

b) *Análisis de los resultados en relación con el campo de conducta evaluado*

Los resultados que obtiene el alumno en el test se expresan en porcentajes de respuestas correctas para cada uno de los objetivos instruccionales. Bastará, pues, comparar esos porcentajes con los que se había establecido cuando se elaboraron los *niveles de ejecución* del test y ver cuáles han sido alcanzados y cuáles no. Esta comparación sería suficiente para la interpretación de los resultados de un TBC dirigido a evaluar los «minimum essentials», es decir, los aspectos básicos de un área de conocimientos.

Cuando se trata de TBC destinados a medir el desarrollo del aprendizaje, la forma de interpretar los resultados es la misma; sin embargo, debe tenerse en cuenta también la posición relativa que dichos resultados otorgan al alumno dentro de su grupo en relación con cada objetivo, puesto que con ambos datos, nivel de ejecución en un ámbito específico del rendimiento y posición dentro del grupo, tenemos una idea mucho más clara del rendimiento académico de los alumnos.

4. OTRAS CUESTIONES RELACIONADAS CON LA CONSTRUCCIÓN DE TBC

La elaboración de TBC exige plantearse cuestiones como la fiabilidad y validez del test, de forma similar a lo que ocurre con los tests referidos a normas. No obstante,

hay que señalar que estas cuestiones no han recibido respuestas totalmente satisfactorias, en el estado actual de desarrollo de los TBC, por parte de los investigadores. Propondremos, pues, de forma sucinta los conceptos y procedimientos generales de estimación de ambas cualidades, remitiendo a la bibliografía especializada para una ulterior profundización.

4.1. *Fiabilidad*

Una definición ya suficientemente popularizada de la fiabilidad de un test es «el grado de constancia en sus mediciones», es decir, la consistencia con que mide aquello que se está midiendo.

Los índices de fiabilidad son: la estabilidad, la equivalencia, equivalencia y estabilidad, y consistencia interna.

Estabilidad: hace referencia a la consistencia de los resultados de un test aplicado en dos momentos diferentes a un mismo sujeto o grupo de alumnos. El método utilizado para el cálculo de la estabilidad es el de «test-retest», con un período de tiempo intermedio relativamente breve (una-dos semanas). Los resultados obtenidos en el test en ambas ocasiones se someten a análisis de correlación, y si alcanzan un índice de estabilidad determinado (entre 0,80 y 0,95 para los tests basados en normas) se considera que el test es fiable.

Equivalencia: consiste en la elaboración de dos formas del test que pretenden ser equivalentes en relación con las especificaciones del test. Para ello se crean series homogéneas de items, y mediante la selección fortuita de los mismos se construyen varias formas del test. La correlación promedio que obtenemos entre ellas, antes y después de la instrucción, reflejará la verdadera fiabilidad del test.

Equivalencia y estabilidad: supone la combinación de ambos índices de fiabilidad. Se trata de obtener dos formas equivalentes del test y aplicarlas a los mismos examinados en dos (próximos) momentos diferentes, calculando a continuación el índice promedio de relación entre las diferentes formas.

Consistencia interna: «Las estimaciones de consistencia interna... se centran, no sobre la consistencia de una prueba en situaciones diferentes o en momentos diferentes, sino sobre la de los elementos individuales dentro de ella». Es decir, hace referencia a la homogeneidad de los items. Para los TBC este índice ya se ha determinado en las etapas de construcción del test, por lo cual resulta poco útil para la estimación de la fiabilidad.

Los índices de estabilidad y equivalencia son, pues, los más indicados para la estimación de la fiabilidad de los TBC, dado que nos permitirían considerar *la consistencia de las decisiones* que se derivasen del empleo de la prueba (Popham) respecto a los alumnos individualmente considerados.

4.2. *Validez*

Como siempre se dice, un test es válido si mide lo que pretende medir. Para los TBC se manejan tres tipos de validez:

Validez descriptiva: también denominada a veces «validez de contenido». Se da cuando un TBC mide las conductas (conocimientos y destrezas) que se han especificado

en la definición del ámbito de aprendizaje (= definición de campo) que se pretendía evaluar.

La estrategia más adecuada para establecer la validez descriptiva de un TBC es confiarlo al juicio de varios observadores familiarizados con el campo de conducta que se desea evaluar (v.g. profesores de la misma área de conocimientos en idéntico nivel educativo) y pedirles que identifiquen aquellos items que, según su criterio, son congruentes con las especificaciones del test. Un porcentaje medio de items considerados congruentes que esté en torno al 90 por 100 conferiría una validez descriptiva satisfactoria al test.

Validez funcional: la evaluación del rendimiento académico de los alumnos no siempre persigue únicamente la descripción de lo que el alumno sabe y no sabe hacer en un momento dado. En ocasiones la evaluación pretende determinar otro tipo de realizaciones, como, por ejemplo, las posibilidades que los alumnos tienen de aplicar con éxito en otros contextos (v.g. laboratorios, actividad profesional...) los conocimientos y destrezas adquiridos en una secuencia instruccional. Se puede entonces dar el caso de que un TBC tenga una adecuada validez descriptiva pero carezca de validez funcional, es decir, no es válido para cumplir esa función asignada.

La validez descriptiva es condición previa para la validez funcional, y ésta sólo es posible determinarla en base a resultados empíricos obtenidos en el seguimiento de los sujetos que demuestren que el test cumple con eficacia la función prevista.

Validez de selección de campo (o dominio de conducta): depende de la precisión con que el que ha elaborado el test ha seleccionado un determinado campo de conducta (contenidos y destrezas) como representación del objetivo general que se pretendía que el alumno dominase.

Este tipo de validez es quizá el que más dificultades plantea para su verificación. Dos estrategias son posibles. Una, que pudiéramos denominar *indirecta,* consiste en el análisis de los procedimientos y personas que han seleccionado el dominio y que han llevado a cabo la posterior elaboración del test. ¿Son personas cualificadas en el área de conocimientos que se pretende evaluar? Si es así, cabe esperar que el test tenga una razonable validez de selección de campo.

La estrategia *directa* exige la elaboración de varias descripciones de campo con sus items correspondientes. Las pruebas así generadas se pasan a un grupo de alumnos para determinar cuál de los diversos campos discrimina mejor entre los alumnos con diferentes grados de instrucción. También puede determinarse la validez de selección de campo eligiendo a un grupo de personas competentes en el área de que se trate, aplicar a cada uno de ellos los instrumentos correspondientes a las diversas descripciones de campo efectuadas y realizar un análisis de los resultados con objeto de ver cuál de las selecciones de campo engloba a las demás, de tal forma que aquellos que obtuvieron un rendimiento óptimo en ella también lo habrían tenido aceptable en las demás.

4.3. *Presentación*

El formato definitivo en que va a ser presentado un instrumento para la evaluación/diagnóstico del rendimiento académico es muy importante para asegurarnos que no se verá afectada la validez del test. Por consiguiente, un conjunto de especificaciones o *esquema descriptivo* debe ser incluido en la edición final de cualquier TBC. Dichas especificaciones servirán para que los usuarios del test, profesores y alumnos, sepan de forma clara qué tipo de conducta pretende medir, con qué tipo de estímulos, qué clase

de respuestas van a ser exigidas y cómo se valorarán esas respuestas. Así pues, un esquema descriptivo engloba:

- Descripción general de la conducta que va a ser medida.
- Tabla de especificaciones.
- Ejemplo de pregunta.
- Características del material de estímulo.
- Características de las respuestas.
- Normas de puntuación y criterios de interpretación.

VÍCTOR ALVAREZ ROJO

BIBLIOGRAFIA

BERK, R. (Ed.) (1984): *A guide to criterion-referenced tests construction* (The John Hopkins University Press, Baltimore).

FERNÁNDEZ BALLESTEROS, R. *y otros* (1981): *Evaluación conductual. Metodología y aplicaciones* (Pirámide, Madrid).

GRONLUND, N. E. (1978): *Preparing criterion-referenced tests for classroom instruction* (Macmillan Co., New York).

HAERTEL, E. (1985): «Construct validity and criterion-referenced testing», *Review of Educational Research,* 55, pp. 23-46.

HAMBLETON, R. *y otros* (1978): «Criterion-referenced testing and measurement: A review of technical issues and developments», *Review of Educational Research,* 48, pp. 1-47.

MEHRENS, W. A. - EBEL, R. L. (1979): «Some coments on criterion-referenced and norm-referenced achievement tests», NCME, *Measurement in Education,* 10 (1).

NITKO, A. (1980): «Distinguishing the many varieties of criterion-referenced tests», *Review of Educational Research,* 50, pp. 461-485.

POPHAM, W. (Ed.) (1972): *Criterion-referenced measurement: An introduction* (Prentice-Hall, Englewood Cliffs, New Jersey).

— (1983): *Evaluación basada en criterios* (Magisterio Español, Madrid).

— (1980): *Problemas y técnicas de la evaluación educativa* (Anaya, Madrid).

POPHAM, W. y LINDHEIM, E. (1980): «The practical side of criterion-referenced tests development», NCME, *Measurement in Education,* 10 (4).