



RESULTADOS DE UNA EVALUACIÓN CRITERIAL EN EL ÁREA DE CIENCIAS SOCIALES, GEOGRAFÍA E HISTORIA DE LA EDUCACIÓN SECUNDARIA OBLIGATORIA ¹

JUAN L. CASTEJÓN COSTA (*)
LEANDRO NAVAS MARTÍNEZ (*)
GONZALO SAMPASCUAL MAICAS (**)

RESUMEN

En este artículo ofrecemos los resultados de una evaluación referida al criterio partiendo de los contenidos establecidos en el currículo de la Educación Secundaria Obligatoria para el área de Ciencias Sociales, Geografía e Historia. Se establece un punto de corte empírico siguiendo el método de Berk para cada criterio, para cada bloque de contenido y para toda el área, y un punto de corte racional considerando las estimaciones de los profesores. La muestra la componen 932 alumnos-as de 3º y 4º de la Educación Secundaria Obligatoria y 26 profesores. Nuestros datos indican que el porcentaje mínimo de dominio para el final del segundo ciclo se sitúa alrededor de un tercio del total. Además la estimación de los puntos de corte realizada por los profesores no discrepa mucho de los puntos de corte obtenidos de forma empírica. No obstante, existe una ligera

discrepancia entre las calificaciones de los profesores y los resultados de la evaluación criterial.

INTRODUCCIÓN

El presente trabajo se inscribe en un doble marco teórico, por una parte, la evaluación criterial de los objetivos educativos entendidos como desarrollo de capacidades, y por otra, la teoría de la consistencia en la toma de decisiones.

La evaluación en la Educación Secundaria Obligatoria supone valorar el tipo y el grado de aprendizaje que los alumnos han alcanzado con respecto a las capacidades indicadas en los objetivos generales, tal y como se señala en el Real Decreto 1345/1991. El nivel de cumplimiento de tales objetivos no debe medirse de modo inflexible, hay que tener en cuenta la situación del alumno, sus características y posibilidades.

(*) Universidad de Alicante.

(**) Universidad Nacional de Educación a Distancia.

(1) Este trabajo es parte del proyecto de investigación «Procesos de evaluación y decisiones de promoción en el área de Ciencias Sociales, Geografía e Historia de la Educación Secundaria Obligatoria: Análisis de consistencia desde la Evaluación Criterial», subvencionado por la Secretaría de Estado de Educación en el Concurso Nacional de Proyectos de Investigación Educativa para 1993.

Además, si la evaluación ha de cumplir una finalidad formativa debe realizarse teniendo como criterio de referencia, no la media estándar del grupo de clase sino la evolución personal de cada alumno con relación a sí mismo, si bien, para poder hacer esto, es necesario disponer de pautas mínimas y escalas graduales de los aprendizajes, que permitan trazar la evolución de cada alumno en función de los niveles que esas escalas señalan. En este sentido, la evaluación tiene tres referentes claves: los objetivos generales, los contenidos y los criterios de evaluación, que deben concretarse en una serie de tareas que permitan poner en práctica los aprendizajes logrados durante el proceso educativo. En consecuencia parece necesario definir, de manera clara y operativa, los aprendizajes sobre los que se va a evaluar a los alumnos. Esto es todavía más prioritario en relación con los aprendizajes mínimos que un alumno deberá haber alcanzado a fin de lograr la promoción de curso o ciclo. Esta promoción debe de estar en función de la madurez del alumno y de sus posibilidades de progreso en los estudios posteriores, como se indica en la Resolución de 28 de mayo de 1993 de la Secretaría de Estado de Educación.

Ello nos lleva directamente al primero de los marcos teóricos a los que anteriormente aludíamos: el campo de la evaluación criterial, que se constituye en una estrategia adecuada para ayudar a los profesores en la toma de decisiones con respecto a la promoción o no de los alumnos, y de utilidad para la adopción de medidas educativas acordes con el principio de atención a la diversidad.

La finalidad de los test referidos al criterio, desde las primeras aproximaciones al tema realizadas por Glaser (1963), Glaser y Nitko (1971), es poder definir con exactitud la posición de un sujeto respecto al dominio de una conducta bien definida. El criterio en el cual se basa una prueba criterial no está condicionado por el nivel de

realización del grupo de alumnos; el límite en el que se basa la toma de decisiones se establece especificando uno o varios puntos de corte o niveles mínimos de dominio de las destrezas básicas. La característica clave de esta evaluación es que los criterios para fijar los niveles mínimos de dominio, aunque puedan ser diversos (Shepard, 1980; Hambleton y Zaal, 1991) han de estar respaldados por la experimentación.

Este tipo de evaluación permite, además, alternativas y acciones instruccionales que posibilitan la intervención inmediata en el proceso de enseñanza-aprendizaje ajustándolo a las necesidades educativas individuales de cada estudiante.

El desarrollo de las aplicaciones prácticas de la evaluación criterial tiene actualmente una gran difusión en diversos ámbitos geográficos, siendo así notables las realizaciones llevadas a cabo en Holanda, Escocia y Francia, por citar algunos países de nuestro entorno cercano. Los aspectos teóricos de la evaluación criterial han sido tratados extensamente en nuestro país (Rivas et al., 1985; Rivas y Alcantud, 1989; Jornet y Suarez, 1989a, b y c; Rivas, Jornet y Suarez, 1995). Los planteamientos de la evaluación criterial han mostrado su validez en diversos trabajos llevados a cabo en nuestro contexto educativo, aplicados fundamentalmente a las áreas de Matemáticas y Lenguaje dentro de la Enseñanza Primaria (Rivas et al., 1985; Alcantud, Vidal y Gilabert, 1985; Rivas y Alcantud, 1989; Gómez Arbeo, 1990).

Es conveniente aplicar tales planteamientos en el marco de la Educación Secundaria que propugna la Reforma Educativa y en áreas distintas a las que tradicionalmente han sido objeto de evaluación criterial.

De otra parte, tras las decisiones que el profesorado ha de adoptar para realizar el Proyecto Curricular, en orden de importancia se sitúan las decisiones acerca de la evaluación y promoción de los alumnos. La relevancia de tales decisiones es indis-

cutible por la trascendencia que tienen para el alumnado, y para las instituciones educativas; y buena muestra de ello es que han sido contempladas por la Administración educativa de nuestro país (Resolución de 28 de mayo de 1993 de la Secretaría de Estado de Educación sobre criterios y procedimientos para decidir la promoción y titulación del alumnado en la Educación Secundaria Obligatoria).

De acuerdo con esta normativa, la promoción o no de los alumnos de un curso o ciclo al siguiente está en función de la estimación del equipo docente en general, y de cada profesor en particular, acerca de la madurez y del grado de desarrollo de las capacidades del alumno. El conjunto de profesores de cada grupo de alumnos ha de decidir la promoción o no, desde el primer ciclo al segundo, o desde tercer curso a cuarto, de los alumnos que aún siendo evaluados negativamente en algunas de las áreas alcancen en conjunto los objetivos educativos del curso o ciclo respectivo. En coherencia con el carácter continuo e integrador que ha de tener la evaluación, se establece que al decidir la promoción al término del tercer ciclo o al final del tercer curso, el conjunto de profesores deberá valorar el progreso de los alumnos en las diferentes áreas, en relación al desarrollo de las capacidades generales contenidas en los objetivos educativos del currículum.

Se señala, además, que la decisión de promoción se adoptará siempre que el alumno haya desarrollado las capacidades que le permitan proseguir con aprovechamiento los estudios del ciclo o curso siguiente, aún en el caso de que haya sido evaluado negativamente en algunas de las áreas o materias.

El criterio para decidir la promoción es la estimación por parte del profesorado del grupo, de las posibilidades del alumno para proseguir con aprovechamiento sus estudios a la vista de las capacidades generales desarrolladas. Indicándose además que, junto a la determinación del dominio

de los contenidos específicos de cada una de las áreas, la decisión de promoción implica un juicio sobre el nivel de desarrollo y de adquisición alcanzado por el alumno, en lo que concierne a las capacidades generales necesarias para continuar su proceso de aprendizaje en el ciclo o curso siguiente.

Tal estimación tiene cierto carácter subjetivo y requiere ser validada a través de su consistencia en relación a otras estimaciones más objetivas, como las derivadas de la evaluación criterial. Toda vez, además, que las decisiones de promoción han de estar en función de las «posibilidades de progreso en estudios posteriores», el análisis de la consistencia entre la estimación del profesor y la evaluación criterial en la decisión de promoción, enlaza directamente con el punto teórico anterior, en lo relativo al establecimiento de criterios mínimos de superación.

El análisis de la consistencia en la toma de decisiones tiene su fundamento en la Teoría de la Decisión (Cronbach y Gleser, 1965; Wiggins, 1973; Silva, 1980) que permite analizar la validez predictiva, y la utilidad y eficacia de los distintos tipos de pruebas de evaluación. Estos criterios de validez y utilidad se convierten en instrumentos muy adecuados a la hora de la toma de decisiones que afectan a personas e instituciones. Algunas de las propuestas realizadas dentro de la evaluación criterial enlazan directamente con los procedimientos de toma de decisiones (Berk, 1976, 1980, 1984).

De los principales acercamientos para la toma de decisiones, distribución binomial (Kriewall, 1972), teoría de la decisión bayesiana (Novick, Lewis y Jackson, 1973), prueba de la razón de la probabilidad secuencial (Nitko, 1974), y el de Berk (1976), se opta por este último, una vez que supera las deficiencias fundamentales de los anteriores: fallos al definir la operacionalización del dominio en términos de la ejecución observada, la medida de los objetivos,

y el aumento de probabilidad de error asociado con la estimación del criterio. Este método no se corresponde con ningún modelo general (modelo de estado, modelo continuo basado en la teoría de la decisión o con el denominado «modelo de compromiso»), pero en opinión de Jornet y Suárez (1989b) tiene transcendencia a nivel metodológico.

Para Rivas, Jornet y Suárez (1995) las operaciones básicas de construcción de un test referido al criterio incluyen: 1. Especificación del dominio educativo, 2. Análisis de ítems, 3. Determinación de estándares y puntos de corte, 4. Comprobación de la Fiabilidad y, 5. Validez. Aunque de acuerdo con estos autores estas operaciones no se realizan en etapas independientes, sino que se producen de forma paralela o transversal, estando presentes en varios puntos del proceso.

La especificación del dominio educativo es uno de los elementos centrales en la construcción de pruebas referidas al criterio. El dominio educativo es «el conjunto de *Objetivos, Contenidos, Actividades y Tareas que constituyen el objeto de la Educación—en general— o de un programa o una unidad educativa —en particular—*» (Jornet y Suárez, 1989a, p. 238). Se trata del conjunto exhaustivo de tareas que manifiesten, a nivel de conducta, los aprendizajes objeto de una determinada unidad educativa (Rivas, Jornet y Suárez, 1995). Este dominio educativo forma el universo de medida desde el punto de vista práctico de la construcción del test. Mientras que desde el punto de vista teórico, la definición del dominio educativo al que se refiere el test está a su vez estrechamente relacionada con la especificación del contenido educativo, lo cual permite interpretar los resultados obtenidos en referencia a criterios internos, contenidos de la materia y unidad o tema bajo estudio.

El establecimiento de los puntos de corte es, junto a la definición del dominio, el aspecto más importante de la evaluación

criterial y la construcción de pruebas referidas al criterio.

La cuestión que se plantea en el ámbito educativo, a la cual es muy sensible la evaluación criterial, es establecer qué estándares de competencia mínima son necesarios para ser considerado apto en un dominio. Sin embargo, en este campo los criterios de realización no suelen ser tan evidentes como en el ámbito profesional.

Una solución a la cuestión anterior se plantea de acuerdo con Rivas y Alcantud (1989) en términos de la capacidad del estudiante para continuar o no dentro del proceso educativo. El estándar mínimo de competencia suele establecerse en el punto en que se garantiza con suficiente probabilidad la continuación en el curso, ciclo o nivel siguiente. Esto, desde luego, requiere asumir la existencia de una jerarquía de habilidades y conocimientos, graduados en orden de dificultad. «*La constatación a través de mediciones educativas relevantes de que unos individuos pueden progresar adecuadamente en el nivel inmediatamente posterior y otros, con desigual resultado, fallan en su progresión, es una información importante para tener acotado el término de competencia educativa*» (Rivas y Alcantud, 1989, p. 55).

Lo que se requiere es establecer un estándar de competencia mínima que venga asegurado por un programa de enseñanza aprendizaje. Esta competencia mínima, si es de tipo profesional, va a venir definida por la demostración de las destrezas necesarias para la realización de tareas que se llevan a cabo en esa actividad profesional. Si la competencia es de tipo escolar, puede estar definida por la progresión futura a lo largo de cursos y ciclos.

Los procedimientos para establecer el punto de corte varían considerablemente. Existen diversas clasificaciones propuestas por distintos autores (Martínez Arias, 1981, 1995; Rivas y Alcantud, 1989; Álvaro Page, 1993; Rivas, Jornet y Suárez, 1995).

Berk (1976, 1980, 1984) recoge varias ideas presentes en tales procedimientos, como son la utilización de grupos con y sin instrucción, la teoría de la decisión, y la estimación de la utilidad de la prueba en la toma de decisiones. La alternativa propuesta por Berk es la de seleccionar de forma empírica una puntuación óptima de corte, basada en las respuestas que dan muestras de validación formadas por sujetos instruidos y no instruidos, de forma que esta puntuación haga máxima la probabilidad de decisiones correctas.

El método de Berk *«es conceptualmente útil y bastante lógico»* (Jornet y Suárez, 1989b, p. 293). Aunque algunos autores como Shepard (1980) estiman que el principal problema de este método radica en la dificultad de definir qué constituye un grupo instruido y no instruido, una vez que se supone que el grupo instruido ha recibido instrucción *«eficaz»* antes de que los sujetos sean considerados competentes (o *masters*) durante una cantidad de tiempo que no siempre puede determinarse claramente. Además el punto de corte óptimo depende del grado de conocimientos del grupo no-instruido.

En la situación escolar están claramente definidos los grupos pre-instrucción, (al inicio de un ciclo o etapa) y post-instrucción (al final de ella) y el tiempo (cursos) dedicado a la adquisición de determinado nivel de competencia o maestría, por lo que estas limitaciones se hacen mucho menos evidentes o dejan de ser tales.

Lo que se necesita, en situaciones reales, es definir puntos de corte empíricos que puedan compararse con los criterios establecidos por la Administración educativa siguiendo procedimientos lógico-rationales basados en el juicio de expertos. Estos puntos de corte empíricos van a darnos una visión clara de hasta qué punto se alcanzan aquellos estándares teóricos cuando se tiene en cuenta el rendimiento real de muestras lo suficientemente amplias de alumnos si-

tuados en etapas determinadas del sistema educativo.

Este conjunto de características hace que consideremos este procedimiento como uno de los más útiles para el establecimiento del punto de corte.

A partir de estos supuestos teóricos, nuestro trabajo trata fundamentalmente de definir de modo claro y operativo los aprendizajes mínimos sobre los que evaluar a los alumnos en el área de Ciencias Sociales, Geografía e Historia para el segundo ciclo de la Enseñanza Secundaria Obligatoria (ESO), construyendo para ello una prueba criterial en la que se establezcan, por experimentación, los puntos de corte o niveles mínimos de dominio de las destrezas básicas en el área mencionada; otros aspectos a tener en cuenta son: valorar la validez predictiva, utilidad y eficacia de esta prueba; analizar la consistencia entre las evaluaciones del profesor y la evaluación criterial y comparar los resultados de la estimación de los puntos de corte con el procedimiento empírico con los puntos de corte establecidos racionalmente por el profesorado.

MÉTODO

MUESTRA

La muestra está compuesta por 932 alumnos-as que cursan 3º y 4º de Educación Secundaria Obligatoria y 26 profesores-as de Ciencias Sociales, Geografía e Historia pertenecientes a 8 Institutos de Educación Secundaria de la Comunidad Autónoma de la Región de Murcia, que en los cursos académicos 93-94 y 94-95 anticipaban la implantación de la Reforma Educativa en la Educación Secundaria.

La unidad de muestreo para los alumnos la constituye el centro y se extrae de la población de los 26 Centros de Educación Secundaria de la Comunidad Autónoma de Murcia según listado facilitado por la Sec-

ción de Planificación y Estadística de la Dirección Provincial de Educación y Ciencia en Murcia. Se parte de un arranque aleatorio i, y se seleccionan los distintos centros por muestreo sistemático con $k = 3$.

La muestra de profesores queda determinada por aquéllos que imparten docencia en el área de Ciencias Sociales a los grupos muestreados, asumiendo que tal circunstancia es aleatoria.

INSTRUMENTOS

Los instrumentos utilizados en la presente investigación se elaboran de modo específico para la misma, y son los siguientes:

- *Prueba de evaluación criterial para las Ciencias Sociales en el 2º Ciclo de la ESO (PECCS-2ESO)*. Para la elaboración de esta prueba, en un primer paso se emplean las estrategias de análisis racional y «de acuerdo interjueces» con profesores del área en ejercicio, y se definen de manera operativa los aprendizajes mínimos sobre los que asentar la evaluación, partiendo como referente principal de los objetivos, contenidos curriculares y criterios de evaluación que para el área completa y para su afianzamiento durante toda la etapa, fueron fijados en el Anexo correspondiente al Real Decreto por el que se establecía el Currículo de la Educación Secundaria Obligatoria. Así, cada uno de los criterios de evaluación expresados en el currículo oficialmente establecido queda definido de modo operativo en función de, en primer lugar, las capacidades implicadas en los mismos y, en segundo lugar, según los contenidos de tipo conceptual a los que van referidos.

En un segundo paso, partiendo de ese universo de medida relativo a capacidades y contenidos implicados en los criterios de evaluación, una comisión de profesores del área, con amplia experiencia docente, construyó un conjunto de ítems destinados a valorar distintos niveles de logro de las capacidades, procedimientos, conceptos y actitudes a los que los criterios se referían. Por ese procedimiento fueron generados cerca de doscientos ítems.

A continuación, una comisión de expertos formada por profesores del área en ejercicio, distintos al profesorado seleccionado para la muestra, diversos asesores de Ciencias Sociales adscritos a diferentes Centros de Profesores, así como miembros de Departamentos de Orientación y de Grupos Estables de Innovación Curricular, seleccionaron los 100 ítems que definitivamente conforman la versión inicial de la prueba, valorando su nivel de relevancia y dificultad, en sendas escalas de uno a cinco puntos en las que la mínima puntuación refleja poca relevancia (escaso ajuste entre el ítem y el criterio que pretende medir), y máxima dificultad.

El análisis empírico de los ítems supone el cálculo de los índices de dificultad (porcentaje de respuestas correctas en el grupo instruido, según la teoría clásica de test); de discriminación (definido por Cox y Vargan en 1966 como la diferencia pretest-postest, y que en nuestro caso se corresponde con la diferencia entre el porcentaje de respuestas correctas en el grupo instruido y el porcentaje de respuestas correctas en el grupo no instruido); de ganancia máxima, formulado por Brennan y Stolurow (1971); el estadístico χ^2 , como índice de discrimi-

nación calculado a partir del porcentaje de respuestas correctas en el grupo instruido y en el grupo no-instruido (el índice de significación estadística informa del poder discriminador de cada pregunta entre los grupos criterio); y el coeficiente de consistencia interna de Cronbach para el total de la prueba, si el ítem es suprimido.

Una vez realizado el análisis de ítems, se combinan los resultados obtenidos según los distintos criterios de selección de ítems (juicio de expertos sobre relevancia y dificultad, e índices empíricos). Se opta por eliminar aquellas preguntas en las que la mayor parte de los criterios anteriores sugieren su supresión. La versión final de la prueba consta de 94 ítems de diverso formato que demandan a los alumnos que identifiquen, relacionen, localicen, completen, o evalúen enunciados mediante respuestas simples de tipo objetivo. Como norma general, los alumnos-as deben escribir, entre corchetes, números, letras o aspas que se correspondan con la respuesta adecuada. En el enunciado de cada cuestión se indica el modo de respuesta en cada caso.

Cada formulario comienza recabando unos datos de carácter general para la identificación del sujeto, unas instrucciones y cuatro ejemplos resueltos que pretenden representar diversas tareas demandadas en el resto de la prueba (Castejón, Navas y Sampascual, 1996).

La fiabilidad de la versión final de la prueba es aceptable. El índice de consistencia interna de Cronbach es 0,84 y el coeficiente K^2 de fiabilidad en la clasificación de Livingston (1972), que parte para su cálculo del coeficiente KR-20, alcanza un valor igual a 0,90.

- *Escala de valoración REDI del PECCS-2ESO*. Esta escala se elabora con la finalidad de evaluar la adecuación de los ítems del cuestionario PECCS-2ESO según el criterio de los profesores que imparten clase a los alumnos que lo contestaron.

Este instrumento incluye, junto a cada cuestión, dos escalas tipo Likert de cinco puntos. La primera de ellas que varía desde a, mínima relevancia, hasta e, máxima relevancia, valora el grado de relevancia que el profesorado asigna a cada ítem atendiendo a los objetivos y criterios de evaluación establecidos en el currículo oficial para este ciclo de la ESO. Igualmente, en la segunda escala se ha de evaluar la dificultad que, a juicio del profesor, presenta el ítem. Para ello ha de indicar un número entre el 1, mínima dificultad, y el 5, dificultad máxima.

VARIABLES

Las variables consideradas son las siguientes:

- Punto de corte para cada criterio de evaluación.
- Punto de corte para cada bloque temático.
- Punto de corte para el dominio de contenido.
- Calificaciones finales en Ciencias Sociales de los alumnos de 3º de ESO.
- Calificaciones finales en Ciencias Sociales de los alumnos de 4º de ESO.
- Concordancia entre calificación y resultados de la evaluación criterial.
- Punto de corte establecido por el profesorado.

PROCEDIMIENTO

Se establece un calendario de aplicación del cuestionario PECCS-2ESO para los alumnos de 3º y 4º de la ESO, al inicio del curso académico 94 - 95, durante los meses de octubre y noviembre. Y otro, para una posterior aplicación a los mismos alumnos al finalizar el curso correspondiente. En ambos casos la aplicación del cuestionario tiene lugar en el horario lectivo, requiriéndose un tiempo para completarlo de dos horas aproximadamente.

Al finalizar el curso se recaban los datos relativos a las calificaciones escolares en el área, partiendo para ello de las actas de evaluación.

A los profesores se les invita a que completen la escala de valoración REDI del Cuestionario PECCS-2ESO, que trata de valorar la relevancia y dificultad que a su juicio presenta cada uno de los ítems de la prueba contestada por los alumnos.

DISEÑO Y ANÁLISIS DE DATOS

El diseño general básico para la construcción de la prueba criterial y el establecimiento del punto de corte consiste en un procedimiento de pre-post instrucción con una línea base situada al inicio del tercer curso de la ESO y un grupo criterio post instrucción de alumnos que han completado la etapa y cuyos datos se recaban a final de cuarto curso.

Los análisis de datos incluyen el establecimiento del punto de corte para cada criterio de evaluación, para cada bloque temático, por cada nivel y para toda la etapa. El procedimiento seguido para establecer el punto de corte es el de Berk (1976; 1980), en el que se hace uso del coeficiente de validez ϕ .

La evaluación de la consistencia entre las calificaciones de los profesores y los resultados de la prueba criterial se realizan

mediante técnicas de correlación para variables dicotómicas.

El punto de corte establecido racionalmente por el profesorado se calcula a partir de la dificultad asignada a cada ítem.

RESULTADOS

PUNTO DE CORTE PARA EL FINAL DEL SEGUNDO CICLO

Se estima el punto de corte para cada criterio de evaluación, bloque de contenido y para el total de la prueba, siguiendo el procedimiento de Berk (1976, 1984).

En la tabla I se indica el punto de corte para cada uno de los criterios de evaluación establecidos para el área de Ciencias Sociales, Geografía e Historia en el segundo ciclo de la ESO. El punto de corte es la puntuación que hace máxima la probabilidad de decisiones correctas y el valor del coeficiente ϕ . Las probabilidades de decisiones correctas y el coeficiente de validez ϕ se obtienen a partir de la clasificación establecida entre los grupos instruidos (alumnos-as que finalizan 4º de ESO) y no instruidos (alumnos-as que comienzan 3º de ESO). Por ejemplo, en el criterio 1, que agrupa a ocho ítems, la puntuación necesaria para superarlo, correspondiente al punto de corte óptimo, es 4. Esto es, se considera que el sujeto que contesta correctamente cuatro ítems de los ocho que se utilizan para evaluar este criterio tiene un dominio mínimo del mismo. El valor correspondiente al punto de corte supone un porcentaje de dominio del 50 % del contenido para poder ser considerado apto en este criterio.

En la tabla II se muestra el punto de corte para cada uno de los bloques de contenido establecidos en el Decreto del Currículo para el área de Ciencias Sociales, Geografía e Historia, de la ESO.

TABLA I
*Punto de corte para cada criterio de evaluación obtenido para el final
del segundo ciclo de la ESO*

CRITERIO	NÚMERO DE ÍTEMS	PUNTO DE CORTE	ϕ	PORCENTAJE DE DOMINIO
1	8	4	0,37	50%
2	4	2	0,34	50%
3	4	2	0,34	50%
4	4	2	0,38	50,9%
5	8	3	0,43	37,5%
6	3	1,5	0,20	50%
7	4	2	0,26	50%
8	5	2,5	0,41	50%
9	3	1	0,45	33,3%
10	3	1	0,42	33,3%
11	5	1,5	0,48	30%
12	2	2	0,40	99%
13	2	1	0,34	50%
14	4	2	0,37	50%
15	2	1,5	0,30	75%
16	3	1	0,39	33,3%
17	3	2	0,34	66,7%
18	3	1	0,33	33,3%
19	3	1	0,34	33,3%
20	4	2	0,37	50%
21	3	1	0,36	33,3%
22*	1	—	—	52%
23	4	1	0,43	25%
24	3	1	0,41	33,3%
25*	1	—	—	38%
26	3	1	0,44	33%
27	—	—	—	—
28*	1	—	—	20%
29	2	1	0,36	50%

$n_1 = 623$; $n_2 = 517$.

* Criterios definidos mediante un sólo ítem. El porcentaje de dominio corresponde al índice de dificultad del ítem.

TABLA II
Punto de corte para cada bloque de contenido obtenido para el final del segundo ciclo de la ESO

BLOQUE	NÚMERO DE ÍTEMS	PUNTO DE CORTE	ϕ	PORCENTAJE DE DOMINIO
1	42	16	0,57	38,1%
2	26	9	0,63	35%
3	26	8	0,57	31%

$n_1 = 623; n_2 = 517.$

Al igual que en el caso anterior la puntuación mínima con la que se supera el dominio del bloque está determinada por la probabilidad de decisiones correctas y por el coeficiente de validez ϕ .

El bloque número 1 (Sociedad y territorio) requiere una puntuación mínima de 16 entre 42 respuestas posibles para que se considere que un sujeto posee el conocimiento mínimo para ser considerado apto en este bloque, lo que supone un porcentaje de dominio necesario del 38 %. El bloque número 2 (Sociedades históricas y cambio en el tiempo) necesita un porcentaje mínimo del 35 % para ser declarado apto. Finalmente, el bloque número 3 (El mundo actual) precisa un 31 % de dominio mínimo.

En la tabla III se indica el punto de corte para el total de la prueba de evaluación criterial al final de la ESO. Como podemos apreciar la puntuación mínima de dominio coincide con el valor 32. Esto nos indica que el porcentaje mínimo de dominio requerido es del 34 %. Llama especialmente la atención que siendo la pun-

tuación máxima posible de 94, la mayor puntuación obtenida por los alumnos del grupo instruido es igual a 74.

PUNTO DE CORTE PARA LA PROMOCIÓN DE 3º A 4º CURSO

En este apartado se presentan los resultados correspondientes a los puntos de corte requeridos para la promoción dentro del ciclo, es decir de 3º a 4º curso. El procedimiento empleado es el de grupos independientes instruidos y no instruidos. El grupo no instruido lo forman los alumnos que comienzan tercer curso y el grupo instruido los alumnos que inician cuarto curso.

En la tabla IV se resumen los puntos de corte obtenidos para cada bloque de contenido y para el total de la prueba, estimados a partir de los valores del coeficiente de validez ϕ . Los puntos de corte calculados de este modo constituyen la puntuación mínima necesaria para la promoción dentro del ciclo. Como podemos

TABLA III
Punto de corte para el total de la prueba al final del segundo ciclo de la ESO

NÚMERO DE ÍTEMS	PUNTO DE CORTE	ϕ	PORCENTAJE DE DOMINIO
94	32	0,68	34%

$n_1 = 623; n_2 = 517.$

TABLA IV

Puntos de corte por bloques de contenido y para el total de la prueba para la promoción dentro del ciclo (de 3° y 4° curso)

PRUEBA	PUNTO DE CORTE	PROBABILIDAD DE DECISIONES CORRECTAS	ϕ	PORCENTAJE DE DOMINIO
Bloque n.º 1	11	0,63	0,26	26%
Bloque n.º 2	6	0,64	0,28	23%
Bloque n.º 3	6	0,60	0,18	23%
Total	24	0,65	0,30	25%

$n_1 = 623; n_2 = 690.$

observar, todas las puntuaciones correspondientes al punto de corte tienen valores inferiores a los hallados para el final del ciclo (ver el punto 1 de este apartado). En este caso, el porcentaje mínimo de dominio requerido para el total de la prueba de evaluación criterial es del 25 %. De la misma forma, el porcentaje de dominio de cada uno de los bloques de contenido es inferior al obtenido al final del ciclo.

ANÁLISIS DE CONSISTENCIA

En este apartado se presentan los resultados de los análisis de consistencia entre la evaluación criterial y las calificaciones de tercero y cuarto curso.

El procedimiento consiste en establecer una doble clasificación en función del punto de corte para el total de la prueba de evaluación criterial (no supera y supera el punto de corte) y las calificaciones de fin de curso (no apto –calificación de suspenso– y apto –calificaciones de aprobado y superiores–). A partir de las correspondientes tablas de contingencia se calcula un valor de χ^2 que indica la existencia de concordancia o de discrepancia estadísticamente significativa.

La consistencia entre los resultados de la evaluación criterial y las calificaciones fi-

nales se calculan independientemente para tercer y cuarto curso.

En la tabla V se resume el análisis de contingencia obtenido para tercer curso. Como podemos observar, no existe consistencia entre los resultados de la prueba criterial y las calificaciones de los profesores: casi el 11 % de los alumnos que obtienen una calificación positiva no superan el punto de corte establecido por la prueba de evaluación criterial; mientras que el casi el 19 % de los alumnos con calificación «suspenso», sí superan el punto de corte referido. La discrepancia encontrada es estadísticamente significativa ($\chi^2 = 19,26; gl. = 1, 412; p < 0,0001$).

Como se aprecia en la tabla VI, en cuarto curso, el 18,5 % de los alumnos obtiene una calificación positiva a pesar de no superar el punto de corte en la prueba criterial; mientras que suspenden el 10,4 % que se sitúan por encima del punto de corte en la prueba. La discrepancia también resulta estadísticamente significativa ($\chi^2 = 4,96; gl. = 1, 491; p = 0,0258$) aunque en este caso el valor de la significación se encuentra muy próximo al límite de la significación convencional, indicando así que no existe tanta discrepancia como en el caso anterior. Un criterio de significación estadística más estricto nos llevaría a establecer la existencia de cierto grado de concordancia.

TABLA V
Resultados del análisis de consistencia en la clasificación entre Calificaciones y Puntuación en la prueba criterial a la mitad del ciclo (de 3° y 4° de ESO)

		CALIFICACIÓN		TOTAL
		SUSPENSO	CALIFICACIÓN POSITIVA	
Supera el criterio	No	41	45	86
		10	10,9	20,9
	Sí	77	249	326
		18,7	60,4	79,1
		118	294	412
	Total	28,6	71,4	100

TABLA VI
Resultados del análisis de consistencia en la clasificación entre Calificaciones y Puntuación en la prueba criterial al final del ciclo (4° de la ESO)

		CALIFICACIÓN		TOTAL
		SUSPENSO	CALIFICACIÓN POSITIVA	
Supera el criterio	No	26	91	117
		5,3	18,5	23,8
	Sí	51	323	374
		10,4	65,8	76,2
		77	414	491
	Total	15,7	84,3	100

El hecho de que sea mayor el porcentaje de alumnos que son calificados como aptos por el profesorado y que no superan el punto de corte de la evaluación criterial, que el porcentaje de alumnos que son calificados como no aptos por el profesorado y sí superan el punto de corte establecido, parece indicar que por tratarse del curso terminal del ciclo y de la etapa, el profesorado califica de forma menos estricta que la prueba de evaluación criterial.

ESTIMACIÓN DE LOS PUNTOS DE CORTE POR EL PROFESORADO

Para estimar los puntos de corte, partiendo de las opiniones del profesorado, se emplea una estrategia similar a la propuesta por Angoff (1971). Si bien, en lugar de emplear la probabilidad de que un examinado con una competencia mínima responda al ítem correctamente, nuestra aproximación consiste en calcular el porcentaje de aciertos esperados (1- P) a partir

de la dificultad estimada por el profesor para cada ítem (P).

En la tabla VII se muestran los puntos de corte, obtenidos siguiendo este procedimiento, expresados en términos de por-

centajes de dominio para cada criterio y los porcentajes de dominio real calculados por el procedimiento empírico de Berk (1984), a partir de las respuestas de los alumnos a la prueba criterial.

TABLA VII

Puntos de corte estimados por los profesores por criterios de evaluación. Comparación de los puntos de corte obtenidos por el procedimiento empírico de Berk (1984).

CRITERIOS	PORCENTAJE DE DOMINIO	
	PROFESORES	PROCEDIMIENTO EMPÍRICO
1	42,32	50,00
2	38,75	50,00
3	44,53	66,67
4	44,55	50,00
5	40,05	37,50
6	28,73	50,00
7	31,60	50,00
8	49,72	50,00
9	33,67	33,34
10	36,67	33,34
11	34,24	30,00
12	59,30	100
13	36,90	50,00
14	27,72	50,00
15	50,60	75,00
16	41,87	33,34
17	35,93	66,67
18	41,27	33,34
19	28,40	33,34
20	56,00	50,00
21	39,67	33,34
22*	41,20	-
23	34,55	25,00
24	32,60	33,34
25*	49,00	-
26	40,87	33,34
27	-	-
28*	51,20	-
29	52,10	50,00

* Criterios definidos mediante un sólo ítem. El porcentaje de dominio corresponde al índice de dificultad del ítem.

Los guiones en relación al criterio 27 indican la ausencia de ítems que valoran este criterio. Como se observa, no existen diferencias muy elevadas. Es de notar que en la mayoría de los criterios el porcentaje de dominio calculado por el procedimiento empírico es ligeramente superior al porcentaje de dominio estimado por el profesorado.

En la tabla VIII se ofrecen los puntos de corte estimados por los profesores, expresados mediante porcentajes de domi-

nio, para cada bloque de contenido y para el total de la prueba, así como los puntos de corte obtenidos por el procedimiento empírico de Berk (1984). En todos los casos, los profesores estiman que es necesario un tanto por ciento mayor para superar los mínimos exigibles en cada uno de los bloques y en el total de la prueba. La mayor discrepancia se produce en el tercer bloque de contenido, si bien las diferencias observadas no son muy elevadas en todos los casos, incluido el total de la prueba.

TABLA VIII

Puntos de corte estimados por los profesores por bloques de contenido y para el total de la prueba. Comparación con los puntos de corte obtenidos por el procedimiento empírico de Berk (1984)

	PORCENTAJE DE DOMINIO	
	PROFESORES	PROCEDIMIENTO EMPÍRICO
Bloque n.º 1	40,70	38,10
Bloque n.º 2	38,92	35,00
Bloque n.º 3	40,05	31,00
Total de la Prueba	40,20	34,00

DISCUSIÓN

Entre los métodos para la estimación de los distintos puntos de corte que establecen el mínimo requerido para ser considerado «*master*» o apto en el dominio de contenidos a evaluar, el procedimiento seguido se encuentra entre los acercamientos más empíricos empleados hasta ahora. Aunque las aproximaciones basadas en la distribución binomial, la teoría bayesiana de la decisión y la prueba de la razón de probabilidad secuencial son también de tipo empírico, en palabras de Berk: «*La deficiencia fundamental de todos estos métodos es su fallo para definir la maestría de forma operativa en términos de las respuestas observadas en los estudiantes, el objetivo*

o rasgo que se mide, y las características del test y de los ítems, únicamente. El nivel criterial o punto de corte se dispone generalmente sobre la base del juicio o de la experiencia...» (Berk, 1976, p. 4).

Parece probado que los procedimientos racionales para establecer el punto de corte sobreestiman o elevan de forma inadecuada la puntuación correspondiente. Como señala Shepard (1980): «*Nuestra experiencia con los comités de expertos que establecen niveles mínimos del criterio, empleando metodologías como las de Ebel o Nedelsky, es la de que establecen niveles en los que fallan más de la mitad de los candidatos*» (Shepard, 1980, p. 454). El procedimiento empleado en esta investigación selecciona de forma totalmente empírica

un punto de corte óptimo basado en las respuestas de muestras de validación formadas por estudiantes instruidos y no instruidos, lo que hace máxima la probabilidad de decisiones correctas a la hora de clasificar a los sujetos según superen o no el dominio mínimo requerido para alcanzar una evaluación positiva; lo que constituye una estrategia adecuada cuando están claramente definidos los grupos instruido y no instruido, al menos de manera formal, como ocurre en la enseñanza obligatoria.

Aunque en nuestro trabajo se calculan los puntos de corte correspondientes a la mitad del ciclo, el mayor valor informativo se encuentra en el punto de corte establecido para el final de la etapa, lo que nos indica, a su vez, el porcentaje de dominio mínimo necesario para ser considerado apto en el área. Cuando se consideran los resultados obtenidos, tanto para los criterios de evaluación, los bloques de contenido, como para el total de la prueba, se aprecia que hay cierta distancia entre el dominio real mostrado por los alumnos-as y los criterios de evaluación establecidos en el currículo. Así, la puntuación máxima alcanzada por los alumnos indica un logro del 78 % de los contenidos a los que hacen referencia los criterios. Este desfase nos lleva a cuestionarnos si los criterios de evaluación establecidos por la Administración educativa para esta etapa son un poco elevados o, estando éstos adecuadamente definidos, el proceso instruccional no se lleva a cabo adecuadamente, en su planificación o desarrollo.

El porcentaje de dominio mínimo para el total de la prueba y para el final de la etapa se sitúa en torno a un tercio (34 %) del total. Esto supone que el punto de corte establecido es moderadamente bajo (32 respuestas correctas de 94 posibles).

En cuanto a la estimación de los puntos de corte por el profesorado, resulta notable que no se produzcan grandes diferencias entre el porcentaje de dominio estimado por los profesores y el porcentaje

de dominio real mostrado por los alumnos. De nuevo, la mayor discrepancia se produce en el tercer bloque, mientras que en el total de la prueba la diferencia entre la estimación del profesorado y los resultados reales de los alumnos sólo es de seis puntos porcentuales (más alta en el caso de los profesores). Esta diferencia es mucho menor que la encontrada en otros estudios (Shepard, 1980; Jornet y Suárez, 1989b).

Los resultados relativos al punto de corte, tomados en su conjunto, parecen indicar que los criterios de evaluación son relativamente elevados, ya que el porcentaje de dominio estimado por los profesores y el observado en los alumnos se sitúa entre el 34 y el 40 %.

No se producen diferencias notables entre el porcentaje de dominio estimado por los profesores y el porcentaje de dominio real mostrado por los alumnos. Esto puede estar indicando cierto grado de validez convergente de nuestros resultados.

Por otra parte, el análisis de consistencia entre las calificaciones que otorgan los profesores a los alumnos y las puntuaciones alcanzadas por éstos en la prueba de evaluación criterial, así como entre las puntuaciones en dicha prueba y las decisiones de promoción distan, en general, de ser consistentes o coincidentes.

Las calificaciones de los profesores a final de ciclo y la puntuación alcanzada en la prueba criterial muestran una discrepancia que llega a ser estadísticamente significativa. Mientras el 15,7 % de los alumnos obtienen una calificación «No apto» de sus profesores, al finalizar el ciclo, el porcentaje de alumnos que no supera el criterio en la prueba se eleva al 23,8 %. Lo que indica el empleo de un criterio de evaluación menos estricto por parte del profesorado, teniendo en cuenta además lo bajo que se sitúa el punto de corte en la evaluación criterial.

Todo lo anterior apunta, por tanto, a la necesidad de continuar trabajando en el desarrollo de pruebas de evaluación crite-

rial y en el análisis de la consistencia entre las calificaciones y decisiones de promoción, por un lado, y los resultados de la evaluación mediante pruebas que ofrecen mayores garantías de objetividad. El desarrollo de pruebas que sigan una metodología referida al criterio puede asegurar, en mayor medida, una evaluación objetiva de los conocimientos del alumno en distintas fases, etapas o momentos del ciclo formativo. La información que aportan estas pruebas lleva directamente a la toma de decisiones instruccionales relativas al alumno individual, al grupo en general, y de un modo más amplio, al proceso instruccional del profesor, y extendiéndose incluso a la adopción de medidas normativas por parte de las distintas administraciones educativas.

El diseño y la construcción de este tipo de pruebas pueden ser abordadas tanto por el propio profesor, en el contexto de la clase (Jornet y Suárez, 1994), como por la propia Administración educativa con un ámbito de aplicación mucho mayor.

Puesto que las pruebas que siguen una metodología referida al criterio pueden asegurar una evaluación más objetiva de los conocimientos del alumno en distintas fases, etapas o momentos del ciclo formativo, se hace necesario el diseño y la construcción de este tipo de pruebas, con el objetivo de mejorar y optimizar la calidad de la enseñanza.

BIBLIOGRAFÍA

- ALCANTUD, F.; VIDAL, E. y GILBERT, R.: *Desarrollo de un sistema de instrucción y evaluación de la comprensión lectora*. Memoria de Investigación del Primer Plan de Innovación Educativa. Valencia, Consellería de Cultura Educación y Ciencia, 1985.
- ÁLVARO PAGE, M.: *Elementos de Psicometría*. Madrid, EUDEMA, 1993.
- ANGOFF, W.: «Scales, norms and equivalent scores», en R. Thorndike (ed.): *Educational Measurement*, Washington, American Cours Leducat (ERIC n.º ED 069686), 1971.
- BERK, R. A.: «Determination of optional cutting scores in criterion-referenced measurement», *Journal of Experimental Education*, 45, 1976, pp. 4-9.
- *Criterion Referenced Measurement*. Baltimore, John's Hopkins University Press, 1980.
- *A guide to criterion-referenced test construction*. Baltimore, The John's Hopkins University Press, 1984.
- «A consumer's guide to setting performance standards on criterion referenced test», *Review of Educational Research*, 56, 1986, pp. 137-172.
- BRENNAN, R. L. y STOLUROW, L. M.: *An empirical decision process for formative evaluation*. Research Memorandum n.º 4, Cambridge, Mass, Harvard CAI Laboratory, 1971.
- CAMILLI, G. y SHEPARD, L. A.: *Methods for identifying biased test items*. Thousand Oaks, CA, Sage Publications, 1994.
- CASTEJÓN, J. L.; NAVAS, L. y SAMPASCUAL, G.: *Procesos de evaluación y decisiones de promoción en el área de Ciencias Sociales, Geografía e Historia de la Educación Secundaria Obligatoria: Análisis de consistencia desde la evaluación criterial*. Informe Final. Madrid, CIDE, 1996.
- COX, R. C. y VARGAS, J. S.: *A comparison of item selection techniques for norm referenced and criterion-referenced test*. Trabajo presentado en la Reunión Anual del National Council on Measurement in Education, Chicago, 1966.
- CRONBACH, J. L. y GLESER, G. C.: *Psychological test and personal decisions*. Urbana, University of Illinois Press, 1965.
- GLASER, R.: «Instructional technology and the measurement of learning outcomes: some questions», *American Psychologist*, 18, 1963, pp. 519-521.

- GLASER, R. y NITKO, A. J.: «Measurement in learning and instruction» en Thorndike, R. L. (ed.): *Educational Measurement*. Washington, American Council of Education, 1971.
- GLASS, G. V.: «Standards and criteria», *Journal of Educational Measurement*, 15 (4), 1978, pp. 237-261.
- GÓMEZ ARBELO, B.: *Evaluación criterial*. Madrid, Instituto de Estudios Pedagógicos, Somosaguas, 1990.
- HAMBLETON, R. K. y ZALL, J. N. (eds.): *Advances in educational and psychological testing: Theory and applications*. Boston, Kluwer, 1991.
- JORNET, J. M. y SUÁREZ, J. M.: «Conceptualización del Dominio educativo desde una perspectiva integradora en Evaluación referida al criterio», *Bordón*, 41, 2, 1989, pp.237-275.
- «Revisión de modelos y métodos en la determinación de estándares y en el establecimiento de un punto de corte en evaluación referida al criterio (ERC)», *Bordón*, 41 (2), 1989b, pp. 277-301.
- «La sensibilidad instruccional: una característica métrica de los ítems de los TRC», *Bordón*, 41 (2), 1989c, pp. 303-324.
- «Evaluación referida al criterio. Construcción de un test criterial de clase» en Pérez Juste, R. (ed.): *Investigación Educativa*. Madrid, Narcea, 1994.
- KRIEWEALL, T. E.: *Aspects and applications of criterion-referenced tests*. Downers Grove, Ill: Institute for Educational Research, 1972.
- LIVINGSTON, S. A. «A criterion-referenced applications of classical test theory», *Journal of Educational Measurement*, 9, 1972, pp. 13-26.
- MARTÍNEZ ARIAS, R.: «Principios psicométricos de las técnicas de evaluación conductual» en FERNÁNDEZ BALLESTEROS, R. y Carrobes, J. A. (eds.): *Evaluación Conductual. Metodología y Aplicaciones*. Madrid, Pirámide, 1981, pp. 157-198.
- *Psicometría: teoría de los test psicológicos y educativos*. Madrid, Síntesis, 1995.
- NITKO, A. J.: «Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience» en C.W. Harris, M.C. Alkin y W.J. Popham (eds.): *Problems in criterion-referenced measurement*. Los Angeles, Universidad de California, Centro para el estudio de la Evaluación, 1974, pp. 59-82.
- NOVICK, M. R.; LEWIS, C. y JACKSON, P. H.: The estimation of proportions in M groups. *Psychometrika*, 38, 1973, pp. 19-46.
- RIVAS, F. et al.: *Proyecto Valencia. Objetivos básicos de aprendizaje en los ciclos y áreas de Lenguaje y Matemáticas en la EGB: Una aproximación de evaluación referida al criterio*. Valencia, Centro de Publicaciones Universitarias, 1985.
- RIVAS, F. y ALCANTUD, F.: *La evaluación criterial en la educación primaria*. Madrid, CIDE, 1989.
- RIVAS, F.; JORNET, J. M. y SUÁREZ, J. M.: «Evaluación del aprendizaje escolar: Claves conceptuales y metodológicas básicas» en SILVA, F. (coord.): *Evaluación psicológica en la infancia y la adolescencia*. Madrid, Síntesis, 1995, pp. 531-570.
- SHEPARD, L.: «Standard setting issues and methods», *Applied Psychological Measurement*, 4 (4), 1980, pp. 447-467.
- SILVA, F.: *Predicción y utilidad. Apuntes mimeografiados*. Valencia, Pau Llivres, 1980.
- WIGGINS, J. S.: *Personality and prediction: principles of personality assessment*. Reading, MA, Addison-Wesley Co, 1973.