Jornet, J.M.; Suárez, J.M.; González Such, J. y Belloch, C. (1997): Estrategias de elaboración de pruebas criteriales en Educación Superior. En C. Martínez Mediano: "Encuentros en la Facultad de Educación sobre Evaluación". Universidad Nacional de Educación a Distancia. Madrid.

ESTRATEGIAS DE ELABORACIÓN DE PRUEBAS CRITERIALES EN EDUCACIÓN SUPERIOR

JESÚS M. JORNET,
JESÚS M. SUÁREZ,
JOSÉ GONZÁLEZ SUCH
Y CONSUELO BELLOCH
DEPARTAMENTO DE MÉTODOS DE INVESTIGACIÓN
Y DIAGNÓSTICO EN EDUCACIÓN. UNIVERSITAT DE VALÈNCIA

1. INTRODUCCIÓN

En este trabajo presentamos algunas referencias para la elaboración de pruebas criteriales para la evaluación del aprendizaje en Educación Superior. En este sentido, tenemos que señalar que las pruebas estandarizadas constituyen un indicador del nivel de aprendizaje, indicador que podrá ser, en principio, más o menos válido, dependiendo del tipo de aprendizaje que se pretenda evaluar. Entiéndase, pues, que el desarrollo de pruebas no se toma necesariamente como el elemento sustancial para realizar la evaluación, sino como un instrumento que puede, si el tipo de aprendizaje a evaluar así lo permite, mejorar la evaluación a partir de un mayor grado de objetivación y estandarización de los procesos evaluativos.

En cualquier caso, la propuesta que aquí se revisa no está orientada hacia la elaboración de pruebas cerradas, dado que para la evaluación del rendimiento, aunque se disponga de varias formas paralelas (aspecto, en cualquier caso, difícil de conseguir), pensamos que el coste de desarrollo de una prueba es muy elevado, máxime cuando en una aplicación «se quema». Sólo puede tener sentido, desde una perspectiva de usos formativos de la evaluación, la creación de un marco de trabajo basado en el desarrollo de recursos evaluativos, que esté orientado a facilitar la elaboración de n-pruebas

paralelas (o al menos equivalentes), tantas como puedan necesitarse en el trabajo evaluativo que cotidianamente debe desarrollar el docente.

Para el planteamiento del marco mencionado, debe tenerse en cuenta que las características del tipo de aprendizaje a que se aspira en la Educación Superior deben guiar el énfasis en la elaboración de pruebas hacia técnicas orientadas a la evaluación de tareas que implican procesos superiores de pensamiento. Respecto a las soluciones métricas (modelos de medida, metodología de construcción de pruebas...) es muy similar al que pueden utilizarse para cualquier prueba de rendimiento y es en el ámbito de la *medición y evaluación criterial* donde encontramos propuestas más ajustadas a estas necesidades.

Teniendo en cuenta que el aspecto clave para que la/s prueba/s sea/n válida/s es la adecuación del análisis del aprendizaje que se desea evaluar, esta propuesta se basa en que la/s prueba/s debe/n ser desarrollada/s por un *comité de expertos* (profesores) de la materia. En el caso en que éstos no tengan una formación en técnicas de evaluación, será preciso integrar en dicho Comité un especialista en evaluación que actuará de asesor/dinamizador en el proceso de construcción.

Las fases de elaboración de la prueba han sido descritas en un trabajo anterior referido a las pruebas de *clase* (ver Jornet y Suárez, 1994), por lo que no incidiremos más en este aspecto aquí. Las fases se pueden sintetizar en los siguientes elementos:

- definir la finalidad de la prueba, en sí misma y en relación a los recursos evaluativos disponibles;
- especificar/definir el *dominio educativo;*
- determinar el *nivel mínimo de competencia,* a través del *estándar* y el *punto de corte,* y
- realizar aplicaciones piloto sucesivas hasta lograr un ajuste que permita la disponibilidad del instrumento y las decisiones asociadas.

En este tipo de desarrollo priman las acciones del *comité* –encaminadas a lograr un consenso intersubjetivo– en la definición del *dominio, nivel mínimo de competencia,* etc., sobre los indicadores de corte empírico, basados en modelos de medida.

La *validez y utilidad* de la prueba descansan sobre el *análisis y definición del dominio.* No obstante, llama la atención que en el ámbito criterial la mayor producción de investigación metodológica se sitúa en torno a las problemáticas de la *determinación de estándares* y *puntos de corte,* aunque se reconoce el análisis y definición del dominio educativo como el elemento central y punto de referencia de todas las acciones de elaboración de este tipo de pruebas. Es por ello que deliberadamente hemos centrado nuestro énfasis, en el presente trabajo, en el apartado destinado al análisis y especificación del dominio educativo.

No obstante, aportamos también referencias globales respecto a los restantes componentes técnicos a tener en cuenta en el desarrollo de las pruebas.

2. DOMINIO EDUCATIVO

Popham (1990) señala que el elemento clave para poder realizar una interpretación criterial de una prueba es la definición del *dominio educativo.* Ciertamente, el dominio educativo constituye el *universo de medida* y es el referente de cualquier operación de elaboración de una prueba criterial. En un trabajo anterior (Jornet y Suárez, 1989a), definíamos el *dominio educativo* como el *«conjunto de objetivos, contenidos, actividades y tareas que constituyen el objeto de la Educación, sea en general sea en un programa concreto»* (pág. 239). Asimismo debe tenerse en cuenta que en la lógica de la *evaluación criterial,* los requerimientos para la elaboración de pruebas se sustentan sobre la idea de *unidimensionalidad,* por lo que asumiendo el dominio educativo como el universo de medida, éste deberá referirse a *«la unidad mínima cuyos aprendizajes vayan a ser evaluados... cabe pensar en un dominio adecuado como el de una lección o unidad didáctica»* (Jornet y Suárez, 1994, pág. 428).

Se define el *dominio educativo* a partir del establecimiento de objetivos instruccionales y constituye la base de consecución de la *validez del test,* tanto a nivel de contenido como de constructo; o bien, si se prefiere, en el sentido de *validez curricular.* Definir correctamente el dominio educativo supone la adaptación del contenido de la prueba al planteamiento instruccional del programa, su adecuación al nivel –o niveles– requerido/s, al planteamiento metodológico-didáctico, etc... Es por ello, probablemente, la etapa más crítica en la elaboración de pruebas y se extiende hasta la *Escritura/formulación de ítems.* En este contexto, entendemos que un dominio está completamente especificado si están escritos los ítems que lo componen o bien si se describen de forma precisa las *Reglas* de generación de los mismos.

Desde un planteamiento de desarrollo de pruebas –o recursos evaluativos– de carácter estandarizado, la diferenciación del grado de conocimiento adquirido por los estudiantes, puede estar determinado en función de la graduación que se establece a través de los ítems. De esta forma, dependiendo de cómo se plantee, y escriba, el ítem se exige del estudiante un nivel diferencial (como por ejemplo, reconocimiento, comprensión o aplicación). Así, en la definición del dominio educativo, deben incluirse ya

todos los elementos interpretativos que puedan asegurar la máxima validez y utilidad de la prueba.

Es conveniente, pues, tener en cuenta que la definición del nivel alcanzado por un estudiante puede realizarse con diferentes grados de precisión. Bien se determina su mero posicionamiento global respecto de los contenidos totales del dominio educativo, bien se profundiza hasta pormenorizar su perfil en relación al tipo de habilidades que tiene adquiridas y lagunas que presenta en su aprendizaje –realizándose una descripción exhaustiva, de tipo diagnóstico, a través de la totalidad de la topografía que constituye el dominio–. No obstante, el grado de precisión requerido depende del tipo de uso que se pretenda dar a la prueba. Para usos meramente sumativos no se requiere más allá de una definición, aunque exhaustiva, de los productos globales de un aprendizaje. Sin embargo, para usos formativos es necesaria una definición pormenorizada del dominio educativo. Esto facilitará una alta precisión en la descripción no sólo del nivel global mostrado por cada sujeto, sino en concreto de la tipología del rendimiento, pudiéndose explicitar en ese caso las habilidades sobre las que se sustenta el nivel alcanzado. Es decir, se trata de dotar a la prueba de capacidad para interpretar no sólo la puntuación global, sino también las respuestas concretas a los ítems.

¿Qué elementos pueden mejorar la precisión en la definición del Dominio Educativo? Aunque son muy diversos podrían sintetizarse en los siguientes:

1. Comenzar la definición del dominio educativo con un análisis del aprendizaje a evaluar desde una taxonomía o sistema de clasificación.

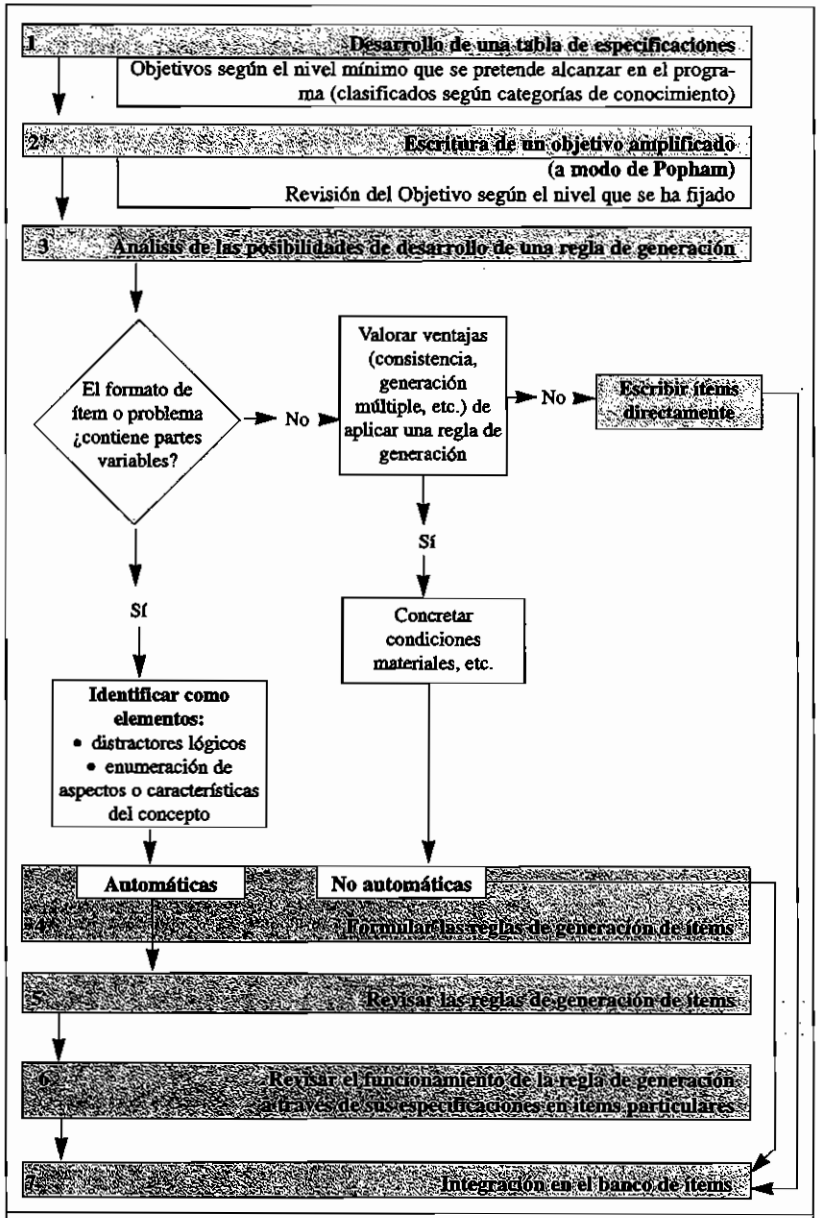
2. Graduar inicialmente el nivel de aprendizaje que se desea observar a través de objetivos.

3. Orientar la escritura de ítems de forma que éstos:

   a) respondan a la descomposición de las diversas habilidades significativas contenidas en el aprendizaje de una tarea, y

   b) permitan la valoración de diversos niveles de aprendizaje en una tarea (por ejemplo, reconocimiento, comprensión, aplicación...).

Teniendo en cuenta los aspectos anteriormente mencionados, un esquema de actuación se recoge en el Cuadro 1.



Cuadro 1. Acciones y guía de decisiones en la definición del Dominio.

La elección del sistema de clasificación o taxonomía puede constituir en sí mismo un elemento clave para el posterior desarrollo de la prueba. No obstante, la aplicación de taxonomías sobre dominios concretos de aprendizaje, además de necesitar un análisis teórico muy costoso, conlleva enfrentarnos con los límites de las propias taxonomías: la ambigüedad para clasificar muchos aprendizajes, las lagunas que se generan en la descripción, etc... Por ello, puede resultar más parsimónico abordar tales análisis utilizando sistemas de clasificación más simples, que contengan, eso sí, la posibilidad de diferenciar entre niveles de adquisición (dado que es lo que se pretende), pero que ello lo aborden con estructuras sencillas, aplicables fácilmente a la mayor parte de dominios educativos.

En este contexto, acercamientos taxonómicos como los de Bloom (1956), Gagné (1971), Scriven (1967), etc., o bien una formulación más funcional, como el modelo de aprendizaje escolar de Carroll (1963), pueden constituir sistemas de análisis del aprendizaje desde los que derivar posteriormente las pruebas, como en otro contexto proponen Fleishman y Quaintance (1984). No obstante, como señalamos anteriormente, la dificultad de aplicación de aproximaciones taxonómicas es elevada y tampoco queda claro que aporten ventajas sustanciales respecto a otros sistemas de clasificación cuando de lo que se trata es de organizar el *dominio educativo como universo de medida.* En este sentido, aproximaciones como las de Reigeluth (1983) o Merrill (1983) pueden resultar sistemas fácilmente aplicables y muy ajustados a las necesidades de análisis en la elaboración de pruebas. Una síntesis de las categorías básicas del sistema C.D.T. de Merrill se recoge en el Cuadro 2. Merrill utiliza tres niveles para la

NIVELES DE RENDIMIENTO

RECORDAR. Es el rendimiento que requiere que el estudiante busque en la memoria para reproducir o reconocer algún ítem de información que ha sido previamente almacenado.

UTILIZAR. Es el rendimiento que requiere que el estudiante aplique alguna abstracción a un caso específico.

ENCONTRAR. Es el rendimiento que requiere que el estudiante derive o invente una nueva abstracción.

CATEGORÍAS DE CONTENIDO

HECHOS. Son piezas de información arbitrariamente asociadas como un nombre propio, una fecha, un acontecimiento.

PROCEDIMIENTOS. Son consecuencias ordenadas de pasos necesarios para conseguir algún objetivo, resolver alguna clase particular de problema u obtener algún producto.

PRINCIPIOS. Son explicaciones o predicciones de cómo suceden las cosas en el mundo. Son relaciones causa-efecto o correlacionales que se usan para interpretar hechos o circunstancias.

Cuadro 2. Sistema básico de dimensiones instruccionales de la teoría C.D.T. de Merril.

categorización del rendimiento (recordar, utilizar, encontrar), que son combinables con las categorías de clasificación de los contenidos de aprendizaje (hechos, procedimientos y principios). La ventaja de este sistema, como otros del mismo tipo, es que proporciona un referente compacto para abordar el conjunto del programa instruccional desde el desarrollo hasta la evaluación. Ejemplos de este tipo de virtualidades se pueden encontrar en el excelente compendio de Reigeluth (1987).

Un segundo elemento clave en la definición del dominio lo constituye la *formulación de objetivos.* El planteamiento que realiza Popham (1978) de *objetivos amplificados,* puede constituir un formato de análisis y formulación de objetivos muy adecuado para este propósito (ver Cuadro 3). En él se especifica no sólo el *objetivo operativo* –que contiene el nivel deseado a observar en el aprendizaje–, sino las condiciones en que se observa –evalúa– dicho aprendizaje: tipo de tarea, materiales, formato de situación en que se va a presentar al estudiante, forma –condiciones– en que el estudiante debe dar la respuesta, forma y criterios de puntuación, etc.

Esta especificación de la situación evaluativa es muy útil, dado que en su explicitación el Comité que desarrolla la prueba tiene ya un elemento de reflexión y de referencia para la formulación posterior de ítems, de forma que es un medio para asegurar la máxima asociación en la transición desde el objetivo a los ítems que lo miden.

En este proceso, la *selección del tipo de ítems* será, asimismo, otro componente clave en la definición del dominio educativo, dado que de ello depende el ajuste al nivel de aprendizaje que se desea observar.

En suma, el ajuste definitivo se producirá en la *escritura de ítems.* Para ello hay que considerar, junto a criterios de calidad técnica de ítems, qne posteriormente comentaremos, que la perspectiva que aquí se presenta tiene por objeto el desarrollo de un sistema de recursos evaluativos que permitan la generación de n-pruebas paralelas (o equivalentes). Una tecnología que puede ayudar a este propósito es el uso de *Reglas de generación de ítems.* Éstas son procedimientos de análisis de contenido que permiten automatizar la escritura de ítems[1].

En la literatura especializada se han venido presentando en los últimos años diversas propuestas de *Reglas de Generación de Elementos* (Roid y Haladyna, 1982; Roid, 1984; Oosteroff, 1994). Si bien estas propuestas no están completamente desarrolladas,

[1] Como se verá posteriormente, hay reglas de generación que responden literalmente a esta afirmación, ya que permiten formular los ítems a partir de la combinación automática de características del contenido o concepto que se desea evaluar (p. ej., el diseño de facetas). No obstante, hay otras reglas de generación que requieren un proceso de creación de cada elemento, aportando únicamente las condiciones del método o técnica para realizarlo (p. ej., el acercamiento de transformaciones lingüísticas).

Objetivo amplificado

Nivel educativo/materia/unidad didáctica

Descripción del dominio: Los estudiantes identificarán, entre un conjunto de gráficas alternativas, cuál corresponde a una distribución descrita en función de sus parámetros de simetría y curtosis.

Límites del contenido:

1. En cada ítem se describirá en términos técnicos una distribución según sus parámetros de simetría y curtosis.
   Los términos de formulación del ítem serán:
   1.ª frase: «¿Cuál de los siguientes gráficos corresponde a una distribución...».
   2.ª frase: «Descripción técnica de la distribución».
   3.ª frase: «Elige la mejor alternativa».
2. Los términos de descripción de la distribución de acuerdo con el parámetro de simetría serán: simetría, asimetría positiva, asimetría negativa.
3. Los términos de descripción de la distribución de acuerdo con el parámetro de curtosis serán: mesocúrtica, leptocúrtica, platicúrtica.

Límites de la respuesta:

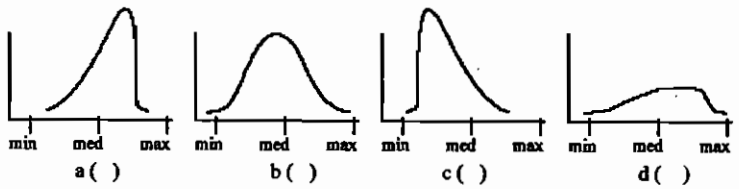
1. Los estudiantes seleccionarán una respuesta entre cuatro gráficos propuestos.
2. Los distractores se elegirán aleatoriamente entre los nuevos tipos posibles de la combinación de los diversos tipos «puros» de simetría x curtosis.
3. Se incluirá siempre una respuesta correcta.
4. No se incluirá como distractor «Ninguna de las anteriores».

Formato: Todas las cuestiones serán de elección múltiple (con 4 alternativas).

Material: Lápiz / Papel.

Instrucciones: El estudiante leerá el ítem y señalará con una «X» la alternativa correcta.

Ítem: ¿Cuál de los siguientes gráficos corresponde a una distribución asimétrica positiva y leptocúrtica? Elige la mejor alternativa.



Cuadro 3. Ejemplo de Objetivo Amplificado.

suponen un avance para el tratamiento exhaustivo de dominios educativos, en los que es compleja o difícil la escritura total de los ítems. No obstante, cada una de las tecnologías propuestas presenta ventajas diferenciales en función del material para el que se apliquen y no pueden considerarse de aplicación general a todo tipo de materiales o contenidos educativos.

No vamos a hacer una revisión exhaustiva de las tecnologías que se han propuesto para la generación de ítems[2], únicamente señalar algunas de las que han tenido mayor difusión. Entre ellas, los *Formatos de ítems* (Hively, 1966), es probablemente la más conocida y ha tenido un amplio uso, sobre todo, en relación a elementos técnicos y cuantitativos; aunque es razonablemente aplicable en todas aquellas materias que guarden una estructura fácilmente identificable. Un *formato de ítem* incluye, al menos, una *descripción general* de las características de la situación evaluativa, en donde se especifican las condiciones estimulares y de la respuesta, los materiales que se emplean para los elementos, las indicaciones para su correcta administración, el guión que se debe seguir en la misma y la forma de registro y codificación de los resultados. En este sentido, la estructura de los *formatos de ítems* garantiza una presentación estimular homogénea y adecuada para todos los elementos que se generan a partir del mismo.

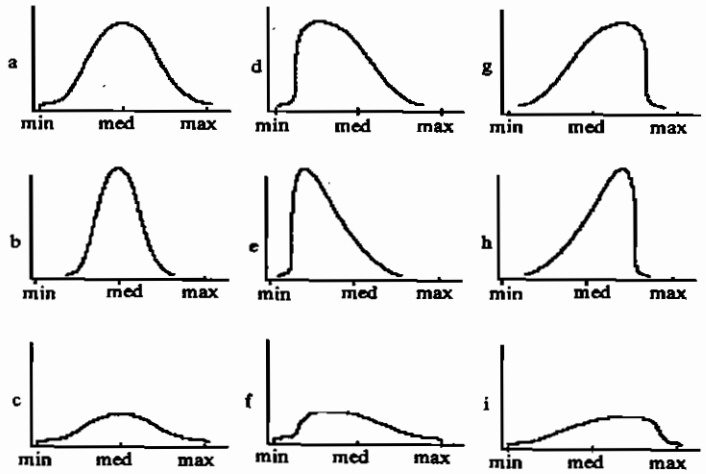
Otra tecnología para establecer *reglas de generación de ítems* es la propuesta por Bormuth (1970) y que se conoce como *transformaciones lingüísticas* (Prose Transformations). Su utilidad reside en que constituye un conjunto de procedimientos para asegurar la conexión lógica entre los ítems y los materiales de texto a partir de los que se generan. Las transformaciones lingüísticas se desarrollan de acuerdo con las siguientes fases: 1) búsqueda de frases claves relevantes en el proceso educativo –reflejadas en los materiales textuales–; 2) selección de las frases más importantes; 3) transformación de las frases[3], y 4) construcción de distractores para los formatos de elección múltiple. Con la utilización de esta tecnología pueden desarrollarse ítems en cuatro niveles de complejidad en cuanto al procesamiento cognitivo requerido: reconocimiento, comprensión, aplicación y análisis.

Por otra parte, el *diseño de facetas* (o Teoría estructural de facetas) propuesto por Guttman (1959, 1965, 1969) se basa en la traslación directa del objetivo operativo a la

[2] La exposición de los numerosos procedimientos propuestos precisa un soporte mucho más amplio que los límites que el presente trabajo imponen. Por ello, remitimos a la consulta de diversas obras que recogen un tratamiento suficiente del tema: Conoley y O'Neil, 1979; Roid y Haladyna, 1982; Jornet y Suárez, 1989a, 1994. En todos ellos, se pueden encontrar además ejemplos que permiten comprender y valorar mejor la utilidad de estos procedimientos.

[3] Las transformaciones pueden realizarse ajustándose al nivel de ítem que se desee producir. Así, tenemos para el nivel de reconocimiento (transformaciones de frases literales, de verbo y sujeto, de sujeto-objeto, o de verbo-objeto), para el nivel de comprensión (basadas en paráfrasis y/o basadas en anáfora, basada en sintaxis interfrases), etc.

Diseño de facetas
Objetivo: Los estudiantes identificarán entre un conjunto de gráficas alternativas cuál corresponde a una distribución descrita en función de sus parámetros de simetría y curtosis.
Formato de presentación: Problema presentado por escrito.
Aplicaciones del diseño de facetas
1. Sentencia directriz
«¿Cuál de las siguientes gráficas se corresponde a una distribución {faceta A}? (Elige la mejor alternativa)»
Alternativas gráficas: {faceta B}
2. Descripción de las facetas
Faceta A: Descripciones de forma de la distribución combinando simetría y curtosis: a: Simétrica-mesocúrtica; b: Simétrica leptocúrtica; c: Simetría platicúrtica; d: Asimétrica positiva-mesocúrtica; e: Asimétrica positiva-leptocúrtica; f: Asimétrica positiva-platicúrtica; g: Asimétrica negativa-mesocúrtica; h: Asimétrica negativa-leptocúrtica, i: Asimétrica negativa-platicúrtica.
Faceta B: Representaciones gráficas (cuatro alternativas: 1 correcta y 3 distractores).



3. Reglas de generación de ítems
– Composición de la formulación del ítem.
Aleatorizar faceta A: selección de una descripción.
– Selección de alternativas:
1. Asignar al azar a una alternativa de respuesta el elemento de faceta B correspondiente al extraído en faceta A.
2. Extraer aleatoriamente entre las restantes unidades de faceta B para cada una de las tres alternativas restantes.
4. Ejemplo de ítems (en cuadro 3).
5. Número total de ítems que se pueden extraer: 81.

Cuadro 4. Diseño de facetas derivado a partir del objetivo amplificado que se describe en el Cuadro 3.

forma de una *sentencia directriz* que incluye todas las condiciones definidas en el objetivo operativo. Esta sentencia recoge el planteamiento general de un ítem que integra dos tipos de elementos: a) la formulación fija del ítem, y b) las partes variables –elementos de las facetas–, por cuya combinación se generan los ítems concretos. Un ejemplo de esta tecnología se recoge en el Cuadro 4, en el que, como puede observarse, por las combinaciones entre los elementos variables se generan hasta un total de 81 ítems que presumiblemente poseen las mismas características métricas[4]. El *diseño de facetas* es aplicable, fundamentalmente, a materias muy estructuradas, por lo que los desarrollos que lo han utilizado se refieren primordialmente a materias técnicas y científicas; si bien es asimismo útil para la generación de ítems de reconocimiento y/o asociación en materias de humanidades y ciencias sociales.

Finalmente, el *Análisis de conceptos* de Tiemann y Markle (1978a, 1978b, 1985) puede entenderse como un método de generación de ítems que se sustenta en la combinación de ejemplos y contraejemplos derivados del análisis previo que se ha realizado sobre un determinado concepto. En este análisis se debe incluir la identificación de los *atributos críticos* (características universales del concepto) y *variables* (características particulares de algunas variantes o aplicaciones del concepto). Por la combinación de ambos tipos de atributos se generan los ejemplos y contraejemplos tanto para su utilización en el programa educativo como para la formulación de ítems. Con ello, entre otras cosas se garantiza una máxima homogeneización entre el proceso educativo y el de evaluación. Este tipo de análisis es especialmente útil para el desarrollo de ítems en situaciones evaluativas orientadas a la comprensión de conceptos.

Como se recoge en el Cuadro 1, la definición de un dominio educativo mediante la utilización de diversas reglas de generación induce a la producción de un conjunto de ítems susceptible de ser organizado en un *banco de ítems*. Normalmente, un banco de ítems que incluya las reglas de generación, ítems concretos, sistemas de construcción de pruebas, análisis de características, etc., es conveniente basarlo en una aplicación informática que permita optimizar su gestión.

La elección de una regla de generación depende esencialmente de su aplicabilidad sobre el tipo de aprendizaje a evaluar, el nivel requerido, etc. Por ello, el análisis y definición de un dominio educativo podrá sustentarse sobre diversas reglas y opciones de escritura de ítems, de manera que lo primordial es el ajuste del ítem resultante a los requerimientos planteados en el programa educativo.

Con independencia de las reglas de generación de elementos, las *técnicas de escri-*

[4] Obsérvese que el ítem se deriva del objetivo amplificado (Cuadro 3) que, una vez descrito, enmarca las posibilidades de desarrollo de un diseño de facetas o cualquier otra regla de generación, de acuerdo con el esquema de actuación propuesto en el Cuadro 1.

tura de ítems permiten, atendiendo a criterios técnicos de calidad, una mejora sustancial sobre las pruebas objetivas clásicas. Así, pueden establecerse diversos criterios de calidad en la formulación de ítems que ofrecen como resultado elementos bien ajustados al material didáctico utilizado, el programa, y el nivel requerido (Jornet y Suárez, 1993). Estas normas son recomendaciones orientadas a asegurar que las propiedades métricas de los ítems se ajustan a los requerimientos del contenido de los mismos.

Cuando nos referimos a la *calidad técnica* entendemos que el énfasis debe ponerse en los criterios de análisis del contenido, formulación y escritura del ítem, además de la necesidad de que el ítem se ajuste a los requerimientos del modelo de medida en que se sustenta la prueba.

Es preciso comenzar por una elección razonada del tipo de ítem a utilizar en cada caso. Esta elección deberá realizarse atendiendo a las características del contenido, del nivel cognitivo, de las adquisiciones a realizar en las diversas destrezas descritas en los objetivos del programa. Conviene, pues, vencer las inercias demasiado frecuentes en la utilización de los tipos de ítems por cuestiones relativas a la familiaridad, hábito, facilidad, prejuicio, etc., llevando a cabo una elección sólida y razonada en cada caso. En el Cuadro 5 aparece una síntesis, para los tipos de ítems más comunes, de orientaciones sobre las tipologías de habilidad a que puedan atender –en función de las estructuras de medida que condiciona cada acercamiento– y algunas recomendaciones genéricas respecto a los ámbitos de contenido que mejor se corresponden con cada caso.

Otro aspecto crucial dentro de esta problemática lo constituye el ajuste razonado, tanto en la definición del objetivo como en la construcción del ítem, del nivel cognitivo que se aspira alcanzar en cada una de las habilidades. Así, en el ejemplo de objetivos e ítem que aparece descrito en el Cuadro 4, se observa que para la resolución correcta del ítem los estudiantes pueden basarse exclusivamente en el reconocimiento. Ahora bien, si el objetivo implicara evaluar el grado de comprensión de las características de las formas de las distribuciones y su aplicación dentro del ámbito educativo sería necesario aplicar una transformación a la propuesta. Se podría, por ejemplo, plantear una situación en la que el sujeto, en lugar de describir técnicamente las distribuciones, fuera capaz de describirlas en términos de características interpretativas. Un posible acercamiento consistiría en plantear la situación para que el sujeto identificara la distribución que mejor corresponde a los resultados de aplicar a un conjunto de sujetos un test muy exigente, difícil y que no discrimina adecuadamente entre los sujetos porque presenta «efecto suelo» (ello correspondería a una distribución con unas marcadas asimetrías positiva y leptocurtosis; alternativa «e» en el ejemplo del Cuadro 4). De esta forma, el nivel cognitivo de habilidad requerido sería superior y precisaría que el sujeto fuera capaz de aplicar los conceptos –generalizando– a una situación de

TIPOS DE ÍTEM/PRUEBA	HABILIDADES MEDIDAS / CAMPOS DE APLICACIÓN
RESPUESTA BREVE	Conocimiento de terminología. Conocimiento de hechos específicos Conocimiento de principios. Interpretación simple de datos (algo más compleja cuando se utiliza material figurativo). Habilidad para resolver problemas numéricos. Habilidad para completar e igualar ecuaciones químicas. Son particularmente útiles en matemáticas y ciencias donde se requiere una respuesta computacional o se debe escribir una fórmula o ecuación. También con idiomas extranjeros donde se busca medir partes específicas de información. Adecuados para medir el conocimiento de definiciones y términos teóricos.
ASOCIACIÓN	Ámbitos de aprendizaje en los que se pretende la aplicación de una base de asociación homogénea en un conjunto de pares.
VERDADERO-FALSO	Habilidad para identificar la adecuación de las afirmaciones de hechos, definición de términos, frases de principios y similares. Habilidad para reconocer relaciones causa-efecto. Aspectos simples de lógica. Son particularmente útiles para medir las creencias en concepciones incorrectas populares y en supersticiones. Si se construyen cuidadosamente, pueden medir procesos mentales superiores (comprensión, aplicación, interpretación).
ELECCIÓN MÚLTIPLE	**Objetivos de aprendizaje:** Conocimiento de terminología. Conocimiento de hechos específicos. Conocimiento de principios. Conocimiento de métodos y procedimientos. **Resultados en los niveles de comprensión y aplicación:** Habilidad de identificar hechos y principios. Habilidad para interpretar relaciones de causa-efecto. Habilidad para justificar métodos y procedimientos.
INTERPRETATIVOS	Habilidad para aplicar un principio. Habilidad para interpretar relaciones. Habilidad para reconocer y establecer inferencias Habilidad para reconocer la relevancia de la información. Habilidad para desarrollar y reconocer hipótesis posibles. Habilidad para identificar la relevancia de argumentos y juzgarlos como erróneos, en su caso. Habilidad para identificar la adecuación de procedimientos. Habilidad para formular y reconocer conclusiones válidas. Habilidad para reconocer asunciones que subyazcan a las conclusiones. Habilidad para reconocer las limitaciones de los datos Habilidad para reconocer y establecer problemas significativos. Habilidad para diseñar procedimientos experimentales. Todos los productos similares basados en la habilidad de los sujetos para seleccionar una respuesta.
DE DESARROLLO RESPUESTA BREVE	Habilidad para explicar las relaciones. Habilidad para describir las aplicaciones de principios. Habilidad para presentar argumentos relevantes. Habilidad para formular hipótesis. Habilidad para formular conclusiones válidas. Habilidad para establecer asunciones necesarias. Habilidad para describir las limitaciones de los datos. Habilidad para explicar métodos y procedimientos. Todos los productos similares basados en la habilidad de los sujetos para emitir una respuesta.
DE DESARROLLO RESPUESTA EXTENSA	Habilidad para producir, organizar y expresar ideas. Habilidad para integrar aprendizajes de diferentes áreas. Habilidad para crear formas originales (p.e., diseñar un experimento). Habilidad para evaluar el valor de las ideas.

Cuadro 5. Síntesis de tipos de ítems y habilidades o campos de aplicación a que pueden dirigirse.

utilización psicopedagógica del análisis de datos. Esta alternativa se podría solucionar adecuadamente desde diversos tipos de formato de ítem, aunque, como se recoge en el Cuadro 5, los acercamientos más globales y «cualitativos» permiten elevar el nivel cognitivo de la medida con mayor «naturalidad».

En este contexto de utilización es conveniente atender a las ventajas e inconvenientes que cada tipo de ítem puede presentar en cada caso para ajustarse al máximo a lo que requiere el objetivo –la unidad del dominio educativo definido–. Dentro de este proceso pueden ser útiles las reflexiones que se recogen de forma sistematizada en los Cuadros 6 y 7, y que sintetizan las propuestas de diversos autores a este respecto, procurando atender a las que han logrado mayor consenso.

Asimismo, se deben seguir unas normas respecto a la escritura de los ítems. Algunas de estas normas son recomendaciones «racionales» derivadas de la experiencia y otras han sido estudiadas empíricamente y están avaladas por resultados de investigación. En el Cuadro 8, a modo indicativo, se recogen sugerencias para la escritura de ítems en general. Así, estas recomendaciones debe entenderse que se ajustan, con pequeñas adaptaciones, a reglas que se deben seguir, sin importar el tipo de ítem. No obstante, existen otras muchas recomendaciones específicas para cada tipo de ítem (ver Jornet y Suárez, 1993), que se deben tomar en consideración además de las que aquí se recogen y que no se incluyen por razones obvias de espacio. Como puede apreciarse, las recomendaciones atienden tanto a cuestiones procedimentales como del propio contenido de los ítems a desarrollar. Dentro de las primeras se plantean reflexiones sobre su formulación lingüística y la organización de la dimensión temporal necesaria en toda medida. Por otra parte, desde la perspectiva de los contenidos a que atiende el ítem, se ofrecen recomendaciones sobre la significación, ajuste y relación entre los ítems, así como cuestiones relativas al nivel de dificultad «a priori» que se fija en cada caso a través de su formulación.

Finalmente, en la escritura de ítems debe tenerse en cuenta para la formulación de distractores y, en general, para el proceso de evaluación en su conjunto, las tipologías de error en función de su significación para el proceso de aprendizaje. Una referencia a este acercamiento y su uso en un contexto de evaluación orientada a la mejora se recoge en Rivas, Jornet y Suárez (1995, pág. 542). No obstante, éste es un ámbito en el que es necesario seguir profundizando y llevar a cabo experiencias específicas variadas que permitan establecer reglas más sólidas y eficaces al respecto.

El planteamiento criterial de construcción de pruebas supone la estandarización de todo el proceso, desde la especificación del dominio hasta la interpretación de la puntuacion final. Esta estandarización en ocasiones no es posible o bien constituye un proceso gradual al que puede aspirarse a partir de la experiencia continuada. De este

TIPO DE ÍTEM	VENTAJAS	INCONVENIENTES
Respuesta breve (Lagunas)	Facilidad de construcción y administración. Reducción de la posibilidad de respuesta por conocimiento parcial. Aportación de información diagnóstica.	Sólo se aplican a cuestiones que pueden responderse mediante una palabra o frase muy breve. No se ajustan a medir situaciones que requieran síntesis e interpretación, en los que sólo haya una respuesta correcta. Orientan hacia un aprendizaje excesivamente memorístico, empobreciendo los hábitos de estudio. La puntuación no es tan rápida y precisa por la variedad de respuestas aceptables.
Asociación	Su forma compacta permite incluir más ítems en un examen. Requieren poco tiempo de lectura. Se ajustan a una corrección mecanizada y objetiva.	Si no se tiene cuidado en su preparación, las listas de asociación pueden orientar-se más a la memoria serial que a la asociación. Es difícil encontrar cuestiones que permitan formular este tipo de ítems.
Verdadero/Falso	Son buenos para niños pequeños y personas con dificultades lectoras. Su tiempo de lectura-respuesta es menor al de otros tipos de ítems, por lo que pueden incluirse más por unidad de tiempo. Se ajustan a una corrección mecanizada y objetiva. Son muy flexibles, se pueden adaptar a la mayor parte de áreas de contenido.	Las puntuaciones están muy influenciadas por la adivinación. Son bastante susceptibles a la ambigüedad y mala interpretación, lo que posiblemente incide negativamente en un menor nivel de fiabilidad. Es fácil copiar en este tipo de ítems. Tienden a ser menos discriminativos que los de elección múltiple. Son susceptibles a la tendencia de respuesta por aquiescencia. No deben ser utilizados en aquellas situaciones en las que la respuesta no es totalmente verdadera o falsa. Son susceptibles a la inclusión de determinantes específicos, para forzar que la respuesta sea totalmente verdadera o falsa.
Elección múltiple	Se da un mayor muestreo de contenido por lo que, generalmente, conducen a una mayor validez de contenido. La fiabilidad de las puntuaciones de los tests puede ser muy elevada con un número suficiente de ítems de alta calidad. Se ajustan a una corrección mecanizada y objetiva. Se pueden obtener subpuntuaciones diagnósticas basadas en un análisis de distractores. Las teorías de los tests (IRT, generalizabilidad, clásica...) se acomodan fácilmente a respuestas binarias. Están relativamente menos afectados por los conjuntos de respuesta que otros tipos de ítems objetivos.	Son relativamente difíciles de construir, en ocasiones resulta complicado encontrar un número suficiente de alternativas. Hay tendencia a construir ítems de RM que demandan solamente recuperar información de hechos concretos (aunque sucede menos que con otros tipos de ítems objetivos). Entre los ítems objetivos, es el que más tiempo se tarda en responder, especialmente cuando se piden discriminaciones precisas. Están sesgados a favor de sujetos: – con habilidades para los tests objetivos y que asumen más riesgo en las respuestas; – más hábiles para detectar la ambigüedad. No se adaptan bien para medir la habilidad para organizar y presentar ideas.

Cuadro 6. Síntesis de ventajas e inconvenientes de ítems objetivos.

TIPO DE ÍTEM	VENTAJAS	INCONVENIENTES
Ítems interpretativos	El material introductorio hace posible medir la habilidad para interpretar materiales escritos, diagramas, mapas, dibujos y otros medios de comunicación que podemos encontrar en situaciones cotidianas. Permiten medir resultados más complejos de aprendizaje que con un ítem objetivo simple. Dada su estructura más amplia, minimiza la influencia de información no relevante sobre el comportamiento de objetivos de aprendizaje complejo. Las series de ítems objetivos fuerzan a utilizar sólo los procesos mentales que requieren, lo que hace posible también medir aspectos separados de la habilidad para resolver problemas y para utilizar procedimientos objetivos de puntuación.	Dificultad de construcción. Especialmente cuando el material introductorio es escrito, es fuerte el requerimiento de habilidad lectora. Aunque es eficiente para medir aspectos específicos del proceso de resolución de problemas, no puede medir la capacidad de resolución de problemas global del sujeto. Está orientado a objetivos de aprendizaje a un nivel de reconocimiento.
Ítems de desarrollo o ensayo	Es relativamente fácil de preparar, sobre todo en términos comparativos con otras opciones. Es la única forma de evaluar significativamente una habilidad para componer una respuesta y presentarla de forma textual propia. Permite medir aprendizajes complejos que no pueden ser medidos con otros procedimientos. Produce un «buen efecto» en el aprendizaje de los estudiantes. Los estudiantes lo prefieren frente a otras opciones (p.e., elección múltiple). Posee validez ecológica, al enfrentar al sujeto con una situación más real y compleja. En los niveles inferiores se puede utilizar también para mejorar habilidades de escritura.	Su pobre o limitado muestreo de contenido, especialmente en las cuestiones amplias. La baja fiabilidad en los sistemas de puntuación, especialmente si se utilizan sistemas globales. La gran cantidad de tiempo requerido en la corrección.

Cuadro 7. Síntesis de ventajas e inconvenientes de ítems de formato complejo y/o subjetivos.

Sugerencias sobre la escritura de los ítems			
Procedimentales		**Contenido**	
Se debe hacer	**No se debe hacer**	**Formulación**	**Facilidad/Dificultad**
a) **Formulación lingüística** • Construcción lingüística correcta. • Redactar con la mayor claridad posible. • Formulación en forma de pregunta. • Ser lo más sintético posible. • Formulación en forma positiva. • Utilizar el formato de mejor respuesta o respuesta correcta. • Situar toda la información del ítem en la misma página, de modo que sea fácilmente abarcable con la mirada. b) **Resolución** • Minimizar el tiempo que el estudiante necesita para la lectura y comprensión del ítem. • Explicitar el tiempo para la resolución. c) **Revisión** • Disponer de tiempo para realizar la revisión de los ítems en sus diversos aspectos (formulación, contenido y puntuación). • Revisiones independientes de los ítems.	• Evitar excesos inútiles en la formulación de la cuestión. • Evitar los ítems con truco o que conducen a engaño. • Evitar porporcionar pistas indeseadas sobre la respuesta correcta (generalmente por aspectos contextuales y/o gramaticales). • No utilizar frases literales de los materiales instruccionales. • Evitar formular el planteamiento de forma negativa.	• Contenidos importantes o significativos. • Basar cada ítem en un objetivo educativo o instruccional. • Utilizar el tipo de ítem más adecuado para cada caso. • Adaptar las cuestiones a los niveles de habilidad de los sujetos y el propósito del test. • Uso de un vocabulario comprensible para el sujeto. • Ítems independientes en cuanto al contenido. • Utilizar los ejemplos de los autores como base para desarrollar los ítems. • Evitar ítems basados en opiniones.	• Evitar, al desarrollar el ítem, el conocimiento superespecífico. • El nivel de dificultad del ítem debe determinarse según lo que se precise para lograr una congruencia ítem-objetivo. • La manipulación de la dificultad del ítem (test de nivel, TRN, etc.) debe realizarse aumentando la dificultad de la respuesta requerida. Y nunca formulando la cuestión de una forma más ambigua o incomprensible. • Evitar añadir al ítem dificultad irrelevante.

Cuadro 8. Sugerencias generales para la escritura de cualquier tipo de ítem.

modo, el análisis del dominio, tal cual está aquí planteado, puede entenderse como un marco de trabajo inicial para el desarrollo de una prueba criterial, o bien, como la elaboración de recursos evaluativos que pueden utilizarse en un esquema mixto (cuantitativo/cualitativo).

Este doble enfoque es posible desde una gestión basada en la idea de *bancos de ítems,* que estarán estandarizados totalmente (tanto en su formulación, como en los criterios técnicos de selección de ítems para pruebas) en el caso de que se integre en un proceso criterial, o estarán estandarizados parcialmente (en su formulación, criterios de puntuación...) dejando los criterios de selección a la opinión del profesor, en el caso en que se integren en procesos evaluativos de corte cualitativo.

No obstante, un proceso de estas características tiene sentido globalmente considerado cuando se integra en un desarrollo estandarizado completo de la prueba, aunque claro está, ello no siempre es posible.

3. ELEMENTOS TÉCNICOS EN LA CONSTRUCCIÓN DE LA PRUEBA

El conjunto de elementos técnicos que se tienen en cuenta en la elaboración de una prueba criterial, junto a la especificación del dominio educativo, son los siguientes:

– Análisis de ítems.

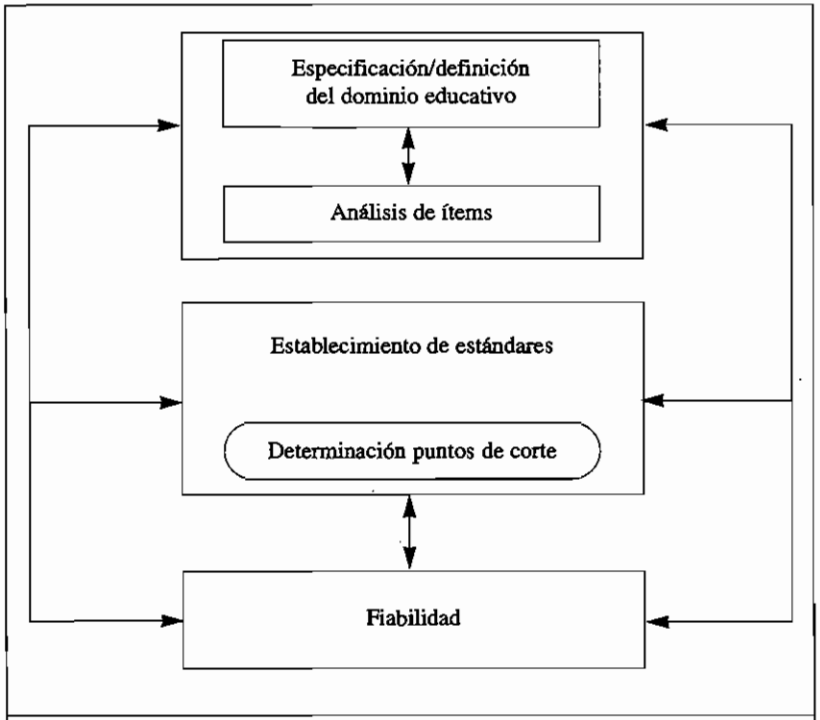
– Fiabilidad.

– Determinación de estándares y puntos de corte.

Estas operaciones tienen siempre como referente el dominio educativo –como universo de medida– y no pueden, por ello, ser desarrolladas adecuadamente sin dicha especificación. En cualquier caso, constituyen en conjunto un proceso iterativo que puede esquematizarse como se aprecia en el Cuadro 9.

Como puede observarse, todos los elementos técnicos están mutuamente implicados y sus referentes inmediatos son el dominio educativo y los estándares. En este conjunto de acciones, la validez de la prueba depende, en definitiva, de todas ellas; por ello, no le hemos dado un tratamiento diferenciado. Y la fiabilidad, como característica de la prueba, es dependiente del estándar y el punto (o puntos) de corte, qne señalan el nivel (o niveles) mínimo de competencia. El análisis de ítems se relaciona con las restantes operaciones, en especial, con la definición del dominio educativo y el establecimiento de estándares.

Una característica general de todos los acercamientos criteriales es que la base de



Cuadro 9. Operaciones de construcción de una prueba criterial y sus relaciones.

trabajo (aunque con comprobaciones empíricas posteriores) radica en el planteamiento intersubjetivo de todos los componentes implicados en la prueba, desde los objetivos (dominio educativo), hasta los niveles mínimos de competencia que deben mostrar los sujetos para considerar que dominan el contenido del programa.

Sin embargo, esto no quiere decir que las pruebas criteriales no estén sujetas a modelos de medida específicos; bien al contrario, la diversidad metodológica es rica y compleja y responde fundamentalmente al tipo de prueba que se desee desarrollar. Las opciones metodológicas se extienden desde aplicaciones del *modelo clásico,* reinterpretado para los requerimientos criteriales, pasando por aplicaciones específicas de la *teoría de la generalizabilidad,* hasta la *teoría de respuesta al ítem.* No obstante, en el contexto en que nos estamos moviendo (desarrollar recursos evaluativos para la elaboración de pruebas en Educación Superior), apostar por un modelo es prematuro, dado que pueden desarrollarse pruebas con diferentes finalidades y ámbitos de aplicación. Así, si las pruebas van a quedar restringidas a su uso en una clase específica, la propia

situación de aplicación nos lleva a optar por métodos derivados fundamentalmente del *modelo clásico* y la *teoría de la generalizabilidad;* pero si se trata de desarrollar pruebas que sean aplicables a grandes poblaciones de individuos, los modelos de construcción de referencia, necesariamente se hallan en la *teoría de respuesta al ítem* (ver Jornet y Suárez, 1994). Abordaremos brevemente una descripción de estos elementos técnicos.

Análisis de ítems

El *análisis de ítems,* encaminado a la mejora de los mismos, conlleva dos componentes mutuamente implicados, las revisiones lógica y empírica. La primera, realizada por el equipo de profesores –o Comité– que desarrollan la prueba y que actúan a modo de jueces, tiene por objeto la adecuación de los contenidos y el planteamiento de los ítems al programa (globalmente considerado). Está, pues, integrada en la definición y especificación del dominio educativo (que hemos comentado anteriormente). La segunda, de carácter empírico, tiene por objeto comprobar el funcionamiento de los ítems de acuerdo con las expectativas que los docentes tienen. Por ello, su función primordial es facilitar una retroalimentación informativa que facilite el ajuste de la prueba al programa o mejore, en su caso, el conocimiento de la funcionalidad del mismo. Su objetivo inmediato no es, pues, la selección de ítems (Berk, 1984).

La **revisión lógica** se desarrolla como un elemento de comprobación de la adecuación general de los ítems. Por ello se realiza de forma integrada con la definición del Dominio. Puede ser realizada por el mismo Comité que actúa en la especificación del Dominio, o bien, por otro que actúe como revisor. Las características básicas que se comprueban en la Revisión Lógica son las siguientes: a) Congruencia ítem-objetivo; b) Calidad técnica, y c) Sesgo de los ítems.

La *congruencia ítem-objetivo* se refiere al grado en que el ítem es una medida adecuada del objetivo del que se deriva. Su comprobación puede ser realizada por el Comité, valorando, con una única puntuación (dicotómica o escalar), el grado de adecuación. Su análisis puede ser cualitativo, o apoyarse en técnicas cuantitativas. En el caso en que la valoración se haya basado en puntuaciones dicotómicas, junto a otras opciones estadísticas, se dispone de un indicador específico desarrollado por Rovinelli y Hambleton (1977). En el caso en que la valoración haya sido escalar, se puede analizar la matriz de juicios *x* ítems con técnicas estadísticas usuales, orientando el análisis a la descripción de los niveles de adecuación de los ítems (síntesis de la opinión del conjunto del Comité), y complementándolo con el estudio de la consistencia interjueces[3].

[3] Se han realizado numerosas propuestas de indicadores cuantitativos y cualitativos que pueden ser útiles para este propósito en las revisiones de Hambleton (1980) y Crocker *et al.* (1989).

Por otra parte, el análisis de la *calidad técnica de los ítems* hace referencia al grado de corrección de los mismos, teniendo en cuenta las recomendaciones específicas para la formulación y escritura de ítems descritas anteriormente. Este aspecto se puede basar en la cumplimentación, por parte del Comité que actúa como revisor, de un registro específico para cada ítem, en el que se revisen sus características. Un ejemplo de registro de valoración de ítems de elección múltiple se recoge en el Cuadro 10.

Listado para comprobar los ítems de elección múltiple

Preguntas	SÍ	NO
1. ¿Es el tipo de ítem más apropiado a utilizar?		
2. ¿La formulación del ítem presenta un problema completo?		
3. ¿La formulación del ítem está libre de material irrelevante?		
4. ¿La formulación del ítem está en forma afirmativa?		
5. ¿Si se utiliza, la formulación en negativo se realiza con un propósito especial?		
6. ¿El ítem está libre de pistas verbales?		
7. ¿Las alternativas son gramaticalmente consistentes con la formulación del ítem?		
8. ¿Las alternativas son breves y concisas?		
9. ¿Las alternativas son similares en longitud?		
10. ¿Existe sólo una respuesta correcta o una mejor respuesta?		
11. ¿Las alternativas incluyen determinantes específicos (nunca, siempre,...)?		
12. ¿Se utiliza la opción «todas las anteriores»?		
13. ¿Se utiliza adecuadamente la opción «ninguna de las anteriores»? Responder sólo si se emplea.		
14. ¿Las alternativas se presentan en un formato vertical?		
15. ¿Las opciones se presentan de acuerdo con un orden lógico (números o letras)?		
16. ¿Los distractores son plausibles para los que conozcan la materia?		
17. ¿Los distractores están formulados gramaticalmente igual que la alternativa correcta?		
18. ¿Los distractores representan errores habituales o posibles de los sujetos?		

Cuadro 10. Ejemplo de registro para valorar ítems de elección múltiple.

A partir de este tipo de información, con esta u otras escalas de valoración se procederá a efectuar un proceso de análisis y toma de decisiones. Este proceso puede tener un solo ciclo o varios, con retroalimentaciones sucesivas del análisis-síntesis del juicio al comité de expertos en un proceso de consenso con convergencia progresiva hasta que se alcance un nivel satisfactorio.

Por último, en esta revisión se pueden incluir valoraciones de los posibles *sesgos socio-culturales* que presenten los ítems en su formulación. No obstante, teniendo en cuenta que se trata de pruebas para la Enseñanza Superior, este aspecto puede resultar de menor relevancia que en niveles educativos inferiores, dado que el colectivo de estudiantes que acceden a la Enseñanza Superior puede considerarse más homogéneo en cuanto a niveles mínimos de formación, no estando afectado tampoco por diferencias evolutivas (como por ejemplo, se pueden observar diferentes ritmos evolutivos según el género).

La **revisión empírica,** presenta para el profesor, al menos, dos polos claros de interés:

a) Disponer de una retroalimentación empírica sobre el comportamiento de la unidad instruccional. Esto, además de una forma enriquecida, pues se aporta no sólo el resumen del test, implica la especificación a través de los ítems, e incluso, de sus partes y/u opciones.

b) Contrastar la hipótesis-definición inicial del dominio educativo con las informaciones pormenorizadas que acabamos de señalar. Esto permitirá realizar una revisión-selección de la prueba y sus elementos o incluso una modificación del dominio.

El desarrollo de una *revisión empírica de ítems* puede concretarse en las siguientes cuatro acciones:

1. *Especificar las hipótesis funcionales y requerimientos para cada uno de los ítems o grupos de ítems.* Estas hipótesis deben estar referidas, al menos, a los siguientes aspectos de interés:

   - La dificultad esperada del ítem (por ejemplo, el ítem puede estar dirigido a evaluar un concepto que el profesor entiende que es básico y que comúnmente dominan la casi totalidad de estudiantes; es decir, espera que el ítem sea fácil, respondido correctamente por la mayoría o la totalidad de los sujetos examinados);

   - La discriminación del ítem respecto al nivel de competencia en el test (por ejemplo, el ítem del ejemplo anterior debe ser acertado en cualquir caso por todos los sujetos que son dados como aptos en el conjunto del test). Estas hipótesis deben estar previamente formuladas para todos y cada uno de los ítems que componen el conjunto de la prueba, y

   - El análisis de errores y distractores. Dependiendo del tipo de ítems que integren la prueba, es conveniente formular hipótesis relativas al funcionamiento de distractores y/o las tipologías de error. Debe tenerse en cnenta que, como señalamos anteriormente, si la escritura de ítems se ha realizado de forma cuidadosa y guiada por una teoría o explicación del aprendizaje que se pretende evaluar en el ítem, el tipo de errores puede constituir un elemento adicional para la interpretación de la prueba.

2. *Recogida de información para el análisis empírico.* Dependiendo del tipo de objetivos que se planteen en la construcción de la prueba, ésta se aplica sobre un grupo de sujetos de *similares características* a los que va dirigida la misma. En el caso en que se vayan a incluir procedimientos de construcción basados en el contraste entre preinstrucción y postinstrucción es preferible una doble aplicación sobre el mismo grupo de sujetos (opción longitudinal), antes que la opción de contraste entre dos grupos diferentes de sujetos –instruidos y no instruidos– (opción transversal), que incluye mayores fuentes de invalidez.

3. *Análisis de datos*[6]. Puede orientarse basándose en la estructura básica de parámetros e indicadores clásicos incluyendo otros específicos propnestos en el ámbito de la evaluación criterial. No obstante debe tenerse en cuenta que en este contexto el uso de indicadores estadísticos de los ítems no está orientado hacia la selección de los mismos sobre sus características globales como unidades de medida, sino en todo caso el énfasis se sitúa en la posibilidad de revisión de los ítems de acuerdo con las hipótesis docentes prefijadas y comentadas anteriormente. Los componentes del análisis empírico de ítems son los siguientes:

   a) Parámetros de dificultad del ítem. Informa acerca del grado de dificultad empírica del ítem y su valor como información no reside en sí mismo (como en los tests normativos), sino en relación a la dificultad esperada por el profesor. Corresponde al indicador clásico como media del ítem.

   Adicionalmente, el análisis de los errores y/o distractores provee una información muy útil acerca del funcionamiento global de los ítems. Un simple análisis frecuencial de los mismos puede orientar adecuadamente su comprobación, siempre que se realice tomando como referencia de comparación las hipótesis previas que al respecto haya planteado el Comité que desarrolla la prueba.

[6] El análisis de datos se establecerá en función del modelo de medida elegido para el desarrollo de la prueba. En este caso, tomamos como elementos de referencia las opciones que mejor se ajustan al desarrollo de un test de clase o a la preparación de recursos evalutivos. Por el contrario, si se tratara de una prueba de certificación, el modelo de medida adecuado, necesariamente, deberá de estar basado en la Teoría de Respuesta al Ítem (IRT). En este caso, los parámetros e indicadores que se tendrían que utilizar serían los propios del modelo elegido en este contexto (ver Hambleton y Swaminathan, 1985).

b) Parámetro de discriminación del ítem. Informa acerca del grado en que el ítem diferencia entre sujetos en relación al *dominio de contenidos* o *el nivel mínimo de competencia.* La importancia de este parámetro viene determinada por su consideración con las hipótesis previas realizadas en torno al ítem. Pueden identificarse dos líneas de análisis en el mismo: *homogeneidad del ítem* y *sensitividad instruccional.*

La *homogeneidad del ítem* hace referencia a si éste actúa de forma adecuada en relación al nivel mínimo de competencia fijado para el test. Así, por ejemplo, puede plantearse como hipótesis para un ítem que será acertado aproximadamente por el 50% de los sujetos y, en todo caso, deberá «ser respondido correctamente por todos los sujetos que superen el nivel mínimo de competencia» (esto es, que demuestre un comportamiento homogéneo respecto al estándar y punto de corte). En cuanto a los indicadores, puede optarse por procedimientos basados en diferencias de proporciones (por ejemplo, diferencia entre índices de dificultad correspondientes a sujetos que superan el nivel mínimo y los que no lo hacen) o por procedimientos correlacionales (por ejemplo, correlación ítem-test). En este último caso debe tenerse en cuenta que el indicador puede verse afectado –como estadístico– en los ítems extremos en cuanto a dificultad (muy fáciles o muy difíciles).

Por otra parte, la *sensitividad instruccional* se refiere a la sensibilidad del ítem para detectar el cambio educativo. Al igual que las características anteriores, depende su inclusión de una hipótesis previa y, más específicamente, del tipo de contenidos a que vaya dirigido el test. Cuando se trata de dominios educativos muy especializados puede no ser necesaria su comprobación. Sin embargo, si se trata de dominios que hacen referencia a aspectos en los que hay una continuidad de aprendizaje evidente con otros dominios (materias muy relacionadas...) su comprobación es imprescindible. Existen múltiples indicadores de la *sensitividad instruccional* (Jornet y Suárez, 1989c), aunque suelen ser los más sencillos de cálculo los que ofrecen, asimismo, mayor claridad en su interpretación (*índice de ganancia máxima; índice de ganancia individual; índice de ganancia neta,* etc., ver Jornet y Suárez, 1994).

c) *Parámetro de Validez.* Informa acerca del grado en que el ítem mide lo que pretende medir. Si bien este aspecto constituye el núcleo central de la revisión lógica, puede ser de interés recabar información adicional empírica al menos respecto a dos componentes: la concurrencia de la medida efectuada por el ítem con la que proviene de un criterio externo al test y la ausencia de un impacto de variables indeseadas conceptualizado como sesgo educativo. En el primer caso, hace referencia a una estrategia clásica de validación criterial no deseadas, y su utilidad, como en el caso del parámetro de discriminación, vendrá determinada por su inclusión como hipótesis para la revisión del ítem. Así, por ejemplo, si el test que se está construyendo pretendemos que sea un medio más operativo para realizar la evaluación que los medios que convencionalmente venimos utilizando como docentes –que pueden resultarnos útiles pero más costosos–, la evaluación realizada por los procedimientos tradicionales puede asumirse como un criterio con el que comparar la información que provee el test y, analizar la ganancia comparativa de ésta.

Por otra parte, en la elaboración de pruebas en Educación Superior, el análisis del sesgo puede ser secundario, como comentamos con anterioridad. No obstante, su interés puede ser de carácter diagnóstico, es decir, para conocer más en profundidad el funcionamiento del dominio educativo en general y la prueba en particular (Jornet y Suárez, 1990). Esto, sobre todo, en relación a estudiantes que puedan tener diferentes formaciones colaterales a las que son objeto del programa que se evalúa. Por ejemplo, al construir una prueba para una materia troncal u obligatoria se puede analizar, mediante un estudio de sesgo, si existe un comportamiento diferencial interno a la prueba entre subgrupos de sujetos que hayan cursado diferentes optativas o grupos de optativas relacionadas con la materia a evaluar (más en un contexto, como el actual, de flexibilidad curricular). Las opciones metodológicas que probablemente se ajustan mejor a este tipo de situación son la metodología delta-plot (Angoff, 1972; Angoff y Ford, 1973) y las aplicaciones de Ji-Cuadrado (Camilli, 1979; Scheuneman, 1975).

4. *La interpretación de resultados.* Debe realizarse en relación con las hipótesis previas planteadas para cada ítem o conjunto de ítems, teniendo en cuenta que cualquier discrepancia respecto al planteamiento con qne el Comité o equipo docente haya incluido los ítems deberá conllevar un examen cuidadoso del mismo. Las consecuencias de este análisis pueden circunscribirse a la revisión de la formulación o planteamiento general del ítem o bien pueden extenderse al propio programa educativo. En este caso, debe tenerse en cuenta que una revisión del programa requiere la revisión subsiguiente del dominio.

Estándares, puntos de corte y fiabilidad de las pruebas

Decidir si un sujeto ha adquirido un nivel de competencia suficiente en el dominio educativo que se evalúa es una actividad básica para la promoción en el aprendizaje. Con independencia del tipo de uso de la prueba (sumativo/formativo), siempre consti-

tuye un elemento necesario. En el marco de las pruebas criteriales se ha realizado un trabajo metodológico iinportante para desarrollar técnicas que permitan interpretar las puntuaciones individuales en relación al dominio educativo, alejándose del relativismo de las interpretaciones basadas en comparaciones entre las puntuaciones de individuos.

La determinación de estándares, puntos de corte y fiabilidad, deben entenderse como un proceso iterativo[7] de identificación de la puntuación en el test más adecuada para indicar correctamente el nivel mínimo de competencia en el *dominio educativo*. Este proceso debe comenzar con la determinación del *estándar* y continúa con su ajuste empírico a través del establecimiento del *punto de corte* y la comprobación de la *fiabilidad* de la clasificación.

El establecimiento del nivel mínimo que debe mostrar el estudiante para asumir que domina/no domina (pasa/no pasa) el contenido del programa se establece a partir del *análisis de estándares* y *puntos de corte*.

La determinación del estándar se establece por métodos intersubjetivos por parte del Comité o equipo docente que desarrolla la prueba. La metodología, muy variada, recorre desde la valoración de los ítems como elemento de referencia hasta la valoración de sujetos o el establecimiento empírico de grupos de contraste (Jornet y Suárez, 1989b).

El hecho de que la base de determinación del *estándar* sea de carácter subjetivo ha suscitado una polémica que llega hasta nuestros días (Glass, 1978; Block, 1978; Hambleton, 1978; Linn, 1978; Jaeger, 1991; Cizek, 1993; Kane, 1994). En cualquier caso, debe tenerse en cuenta que se trata de determinar un *nivel mínimo de calidad de carácter absoluto* y, en este contexto, se aportan elementos de objetivación de un proceso que necesariamente debe tener su base en la subjetividad. No obstante, ello no quiere decir que la subjetividad mencionada sea la mera subjetividad individual. Se han propuesto diversos elementos que pueden orientar la objetivación en la determinación de estándares; entre ellos se pueden destacar los siguientes:

a) Orientar el proceso de determinación hacia la definición de un estándar intersubjetivo basado en el juicio de un colectivo de expertos (Comité).

b) El Comité que determina el estándar, en el tipo de pruebas que estamos considerando, es preferible que sea el mismo que realiza la definición del dominio educativo.

c) Entre los métodos propuestos, los basados en *juicios sobre los ítems* (como los

[7] Aunque la mayor parte de métodos contempla estos tres componentes de forma aislada y no en un proceso como aquí proponemos de acuerdo con los criterios expresados en otro lugar (Jornet y Suárez, 1989b). Además, estudios recientes avalan la mayor consistencia de los procedimientos que integran la interacción en su esquema de actuación (Busch y Jaeger, 1990; Jaeger, 1990b).

de Angoff o Jaeger) parecen los que mejor se ajustan al tipo de trabajo que debeu realizar los Comités[8]. Un ejemplo de aplicación que integra estos métodos modificados por los autores se recogen en Jornet y Suárez (1994, pág. 438).

d) No obstante, la determinación basada en juicios sobre los ítems suele producir estándares excesivamente exigentes, por lo que su ajuste, utilizando métodos basados eu el juicio sobre sujetos (Livingston y Zieky, 1982), puede ayudar a couseguir estándares más realistas.

Adicionalmente, otros elementos clave a tener en cuenta en este proceso son los siguientes:

1. *Entrenamiento del comité de expertos.* El establecimiento de estándares requiere un proceso minucioso de entrenamiento del Comité para asegurar que todo él emite juicios sobre una idea uniforme de exigencia –de acuerdo con los objetivos y finalidad de la prueba–. En este contexto, realizar sesiones preliminares utilizando retroalimentación acerca de las consecuencias de aplicación de sus juicios, ayuda a detectar jueces que no hayan comprendido su labor, así como mejoran el realismo del Estándar posterior (Reid, 1991).

2. *Definición de los objetivos y finalidad de la prueba.* Ayuda a realizar una definición previa del nivel de exigeucia del estándar y enmarca las preferencias del Comité en relación al *tipo de errores* (Tipo I y II) a asumir en la clasificación con el test. Por ejemplo, en una prueba de Cirugía parece más deseable eliminar por el test un sujeto apto (generar falsos negativos mediante un estándar exigente) que aceptar como apto un sujeto que no lo es (generar falsos positivos mediante un estándar poco exigente).

3. *Formatos de juicio y recogida de información.* A este respecto deben tenerse en cuenta algunas recomendaciones:

   – El formato de juicio y el sistema de puntuación utilizado deben ser sencillos y de fácil utilización por el Comité;

   – En el caso de utilizar métodos basados en el juicio sobre los ítems, es conveniente sintetizar en una única puntuación la valoración, si bien enriquece este proceso recoger las apreciaciones que cada juez realiza respecto de las valoraciones que emite (Geisinger, 1991). Asimismo, aunque las valoraciones se realizan sobre cada unidad del Dominio Educativo, es conveniente revisar ini-

[8] Los estudios comparativos de métodos, tanto teóricos como empíricos, no son totalmente concluyentes (Koffler, 1980; Brennan y Lockwood, 1980; Skakun y Kling, 1980; Shepard, 1980; Berk, 1986, Norcini *et al.*, 1987; Livingston y Zieky, 1989; Jaeger, 1990a). A este respecto Kane (1994, pág. 440) señala que hace falta una sedimentación más matizada de los resultados, porque la única realidad incuestionable es que en la práctica unos procedimientos están más extendidos que otros.

cialmente el total de unidades y seleccionar aquellas que recibirían la máxima/mínima valoración –uso de variables pívot– (Jornet, 1987; Jornet y Suárez, 1989b) con el fin de facilitar las relaciones transitivas entre las valoraciones.

– En el caso de utilizar métodos basados en el juicio sobre sujetos, el conocimiento de éstos por parte del Comité constituye un elemento esencial.

4. *Análisis de los juicios emitidos.* En estos procesos el estándar se identifica como una síntesis que plasma el acuerdo intersubjetivo. Éste se realiza a partir de procesos minuciosos –aunque pueden ser muy sencillos– de análisis de los juicios emitidos. En este caso es preciso considerar elementos de identificación de jueces extremos y, si ha lugar, la eliminación de sus valoraciones o la utilización de medidas robustas como estadísticos que sinteticen los juicios como estándar. Asimismo, la consistencia interjueces puede constituir una medida de la representatividad del estándar derivado de este proceso.

Los **puntos de corte** (dicotómicos: pasa/no pasa, o policótomos: no apto, apto, notable...) hacen referencia a la puntuación real en la prueba que deben obtener los sujetos para ser asignados a una categoría de conocimiento. Se determinan como un ajuste empírico del estándar, teniendo en cuenta las cosencuencias de su aplicación. Como en el caso del estándar, los métodos para determinar puntos de corte también son muy diversos. En cualquier caso, en este contexto se ha demostrado que el mejor ajuste se logra cuando el establecimiento se realiza considerando (a través del análisis de la aplicación sucesiva del estándar y sns variaciones) las consecuencias empíricas de la aplicación del estándar sobre sujetos que son valorados por procesos más detenidos y costosos, pero más minuciosos.

La **fiabilidad** de la prueba, en este caso, hace referencia a si la prueba es consistente al clasificar a los sujetos de acuerdo con el punto de corte establecido, por lo que, técnicamente, está implicado en este mismo proceso y, muchos autores, la identifican como estrategias de validación del punto de corte. Como en los casos anteriores, la oferta de métodos al respecto es muy variada (Jornet y Suárez, 1992) y se concreta en procedimientos orientados a valorar la consistencia en la decisión que proviene de dos administraciones[9] de la prueba.

En este contexto, obviamente, la fiabilidad depende de la localización del punto de

[9] Aunque se han propuesto diversos métodos basados en una sola administración (como el de Huyhn, 1976, y el de Subkoviak, 1976, entre otros), éstos son muy complejos y no hay seguridad acerca de la precisión de sus estimaciones.

corte[10]. La utilización de métodos sencillos y, en todo caso, que estén bien descritos como indicadores parece lo más recomendable. Éste es el caso de indicadores de proporciones como el índice p_0 (Hambleton y Novick, 1973) o el índice p* (Crocker y Algina, 1986)[11] o bien el cociente $K^{2(x,t)}$ de Livingston (1972), que está bien descrito en relación a K.R. 20 y presenta alternativas para una y dos administraciones de la prueba.

Por último, la fiabilidad proporcionará una retroalimentación muy relevante para otros procesos anteriormente mencionados. De un lado, es una base de trabajo en la construcción del estándar/punto de corte, en el proceso iterativo que hemos mencionado. Por otro, las consecuencias de la decisión y las propiedades métricas de la prueba también deben constituir un elemento de reflexión para la revisión del dominio y la estructura de los ítems.

4. NOTAS FINALES

Finalmente, si bien estos componentes son complejos, los beneficios que aportan son indudables y se centran fundamentalmente en un incremento de la objetividad en la evalnación, así como en una mayor operatividad de los procesos de evaluación debida a la estandarización.

La mejora en la evaluación puede contar, entre otros instrumentos, con la incorporación de recursos evaluativos orientados a la elaboración de pruebas. Las propuestas que aquí se han presentado se inspiran en principios de **Evaluación Criterial,** si bien se integran en un marco de trabajo alternativo, aunque no contrapuesto, a la elaboración de pruebas cerradas.

En este proceso pensamos que el énfasis debe situarse en la definición del *dominio educativo.* Cualquier prueba, aunque esté sustentada sobre un modelo de medida aplicado adecuadamente, puede ser poco válida y escasamente útil si no se ba extraído desde un dominio educativo bien definido. El modelo de medida –alternativo según el tipo de prueba a desarrollar– es esencial para ajustar la prueba resultante como instrumento, pero en sí mismo es también un instrumento, nunca un principio ni fin que justifique la utilidad de las pruebas.

[10] Por ello, la multiplicidad metodológica se relaciona con los modelos de cualificación del error (pérdida de umbral, pérdida de error al cuadrado, pérdida lineal) y en consecuencia con la consideración de la gravedad de los diversos errores de clasificación.

[11] $p_0 = p_{11} + p_{00}$; $p^*_{..} = (p_0 - 0.50) / (1 - 0.50)$. A partir de una tabla de decisión 2 x 2.

BIBLIOGRAFÍA

ANGOFF, W.H. (1972): «A technique for the investigation of cultural differences». Comunicación presentada en la Reunión Anual de la A.P.A., Honolulú, septiembre (Servicio de Reproducción de Documentos ERIC, n.° ED 069 686).

ANGOFF, W.H. y FORD, S.F. (1973): «Item-race interaction on a test of scholastic aptitude». *Journal of Educational Measurrement,* 10, 95-105.

BERK, R.A. (1976): «Determination of optimal cutting scores in criterion-referenced measurement». *Journal of Experimental Education,* vol. 45, 4-9.

BERK, R.A. (Ed.) (1980): *Criterion Referenced Measurement: The State of The Art.* Baltimore, Johns Hopkins University Press.

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

BERK, R.A. (1986): «A consumer's guide to setting performance standards on criterion-referenced tests». *Review of Educational Research,* 56, 137-172.

BLOCK, J.H. (1978): «Standards and criteria: A response». *Journal of Educational Measurement,* 15 (4), 291-295.

BLOOM, B.S. (1956): *Taxonomy of Educational Objectives. The Classification of Educational Goals. Handbook I: Cognitive Domain.* New York, MacKay.

BORMUTH, J.R. (1970): *On the theory of achievement Test Items.* Chicago, Illinois, University of Chicago Press.

BRENNAN, R.L. y LOCKWOOD, R.E. (1980): «A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory». *Applied Psychological Measurement,* 4, 219-240.

BUSCH, J.C. y JAEGER, R.M. (1990): «Influence of type of judge, normative information and discussion on standards recommended for the National Teacher Examinations». *Journal of Educational Measurement,* 27, 145.

CAMILLI, G. (1979): «A critique of the chi-square method for assessing item bias». *Laboratory of Educational Research,* Boulder, University of Colorado.

CARROLL, J.B. (1963): «A model for school learning». *Teacher College Record,* 64, 723-733.

CIZEK, G.J. (1993): «Reconsidering standards and criteria». *Journal of Educational Measurement,* 30, 93-106.

CONOLEY, J.C. y O'NEIL, H.F.Jr. (1979): «A primer for developing test items». In H. F. O'NEIL, Jr. (Ed.): *Procedures for instructional systems development.* New York, Academic Press, 95-127.

CROCKER, L.M.; MILLER, M.D., y FRANKS, E.A. (1989): «Quantitative methods for assessing the fit between tests and curriculum». *Applied Measurement in Education,* 2 (2), 179-194.

CROCKER, L. y ALGINA, J. (1986): *Introduction to Classical and Modern Test Theory.* New York, Holt, Rinehart and Winston.

FLEISHMAN, E.A., y QUAINTANCE, M.K. (1984): *Taxonomies of human performance (the description of human tasks).* London, Academic Press, Inc.

GAGNE, R.M. (1971): *Defining objectives for six types of learning.* Washington, DC, American Educational Research Association.

GEISINGER, K.F. (1991): «Using standar-setting data to establish cut-off scores». *Educational Measurement, Issues and Practice,* 10 (2), 17-22.

GLASS, G.V. (1978): «Standards and criteria». *Journal of Educational Measurement,* vol. 15, n.° 4, pp. 237-261.

GUTTMAN, L. (1959): «A structural theory for intergroup beliefs and actions». *American Sociological Review,* 24, 318-328.

GUTTMAN, L. (1965): *The structure of interrelations among intelligence tests.* Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton, New Jersey, Educational Testing Service, Princetom.

GUTTMAN, L. (1969): *Integration of test design and analysis.* Proceeding of the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey, Educational Testing Service.

HAMBLETON, R.K. (1978): «On the use of cutoff scores with criterion-referenced test in instructional settings». *Journal of Educational Measurement,* 15, 277-290.

HAMBLETON, R.K. (1980): «Contributions to criterion-referenced testing technology: An introduction». *Applied Psychological Measurement,* 4 (4), 421-424.

HAMBLETON, R.K. y NOVICK, M.R. (1973): «Toward an integration of theory and method for criterion-referenced tests». *Journal of Educational Measurement,* 10, 159-70.

HAMBLETON, R.K. y SWAMINATHAN, J. (1985): *Item Response Theory: principles and applications.* Boston, MA, Kluwer.

HIVELY, W. (1966): «Preparation of a programmed course in algebra for secondary

school teacher. A report to the National Science Foundation». Minneapolis, Minnesota National Laboratory, Minnesota State Departament of Education.

HUYNH, H. (1976): «On the reliability of decisions in domain-referenced testing». *Journal of Educational Measurement,* 13, 253-64.

JAEGER, R.M. (1990a): «Establishing standards for teacher certification tests». *Education Measurement, Issues and Practice,* 9 (4), 15-20.

JAEGER, R.M. (1990b): «Setting standards on teacher certification tests». En J. MILLMAN y L. DARLING-HAMMOND (Eds.): *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers.* Newbyry Park, CA, Sage, 295-321.

JAEGER, R.M. (1991): «Selection of judges for standard-setting». *Educational Measurement, Issues and Practice,* 10 (2), 3-6. 10, 14.

JORNET, J.M. (1987): «Una aproximación teórico-empírica a los métodos de medición de referencia criterial». Tesis doctoral no publicada. Valencia, Universitat de València.

JORNET, J.M. y SUÁREZ, J.M. (1989a): «Conceptualización del Dominio Educativo desde una perspectiva integradora en Evaluación Referida al Criterio». *Bordón,* 41, 2, 237-275.

JORNET, J.M. y SUÁREZ, J.M. (1989b): «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, 277-301.

JORNET, J.M. y SUÁREZ, J.M. (1989c): «La sensitividad instruccional: una característica métrica de los ítems de los TRC». *Bordón,* 41, 2, 303-324.

JORNET, J.M. y SUÁREZ, J.M. (1990): «Algunas notas de reflexión metodológicas acerca del estudio de distractores y el sesgo de ítems en tests educativos y psicológicos». *Revista de Investigación Educativa,* 8 (16), 551-559.

JORNET, J.M. y SUÁREZ, J.M. (1992): «La fiabilidad en los test criteriales». Documentos de Doctorado. Documento interno del Departamento M.I.D.E. Universitat de València.

JORNET, J.M. y SUÁREZ, J.M. (1993): «Recomendaciones para asegurar la Calidad Técnica de los Ítems en tests educativos». Documento interno. Departamento M.I.D.E. Universitat de València.

JORNET, J.M. y SUÁREZ, J.M. (1994): «Evaluación referida al criterio. Construcción de un test criterial de clase». En V. GARCÍA HOZ (dir.): *Problemas y métodos de investigación en educación personalizada.* Madrid, Rialp.

KANE, M. (1994): «Validating the performance standards associated with passing scores». *Review of Educational Research,* 64 (3), 425-461.

KOFFLER, S.L. (1980): «A comparison of approaches for setting proficiency standars». *Journal of Educational Measurement,* 17, 177-178.

LINN, R.L. (1978): «Demands, cautions and suggestions for setting-standards». *Journal of Educational Measurement,* 15, 301-308.

LIVINGSTON, S.A. (1972): «Criterion-referenced applications of classical test theory». *Journal of Educational Measurement,* 9: 13-26.

LIVINGSTON, S.A. y ZIEKY, M.J. (1982): *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ, Educational Testing Service.

LIVINGSTON, S.A. y ZIEKY, M. J. (1989): «A comparative study of standard-setting methods». *Applied Psychological Measurerment in Education,* 2 (2), 121-141.

MERRILL, M.D. (1983): «Componént Display Theory». En C. M. REIGELUTH (Ed.): *Instructional design: theories and models.* Hillsdale, NJ, LEA, 279-334.

NORCINI, J. LIPNER, R. LANGDEN, L. y SHECKER, C. (1987): «A comparison of three variations on a standard-setting method». *Journal of Educational Measurement,* 24, 56-64.

OOSTERHOF, A. (1994): *Classroom applications of educational measurement* (2.ª ed.). Nueva York, MacMillan.

POPHAM, W.J. (1978): *Criterion-referenced Measurement.* Englewood Cliffs, NJ, Prentice-Hall. (Traducción castellana: *Evaluación basada en criterios.* Ed. Magisterio Español, S. A., Madrid, 1983.)

POPHAM, W.J. (1990): *Modern Educational Measurement* (2.ª ed.). Boston, MA, Allyn and Bacon.

REID, J.B. (1991): «Training judges to generate standard setting data». *Educational Measurement, Issues and Practice,* 10 (2), 11-14.

REIGELUTH, C.M. (1983): «The Elaboration Theory of Instruction». En C.M. REIGELUTH (Ed.): *Instructional design: theories and models.* Hillsdale, NJ, LEA, 335-382.

REIGELUTH, C.M. (Ed.) (1987): *Instructional Theories in Action.* LEA, Hillsdale, NJ.

RIVAS, F., JORNET, J.M. y SUÁREZ, J.M. (1995): «Evaluación del aprendizaje esco-

lar: claves conceptuales y metodológicas básicas». En F. SILVA (Ed.): *Evaluación psicológica en niños y adolescentes.* Madrid, Síntesis.

ROID, G.H. (1984): «Generating the test items». En R.A. BERK (Ed.): *A guide to criterion-referenced test construction.* Baltimore, The Johns Hopkins University Press, 49-77.

ROID, G.H. y HALADYNA, T.M. (1982): *A technology for test-item writing.* New York, Academic Press.

ROVINELLI, R.J. y HAMBLETON, R.K. (1977): «On the use of content specialists in the assessment of criterion-referenced test item validity». *Dutch Journal of Educational Research,* 2, 49-60.

SCRIVEN, M. (1967): «The methodology of evaluation». En R. TYLER y cols. (Eds.): *Perspectives of Curriculum Evaluation.* Monograph Series of Curriculum Evaluation, 1, Chicago, Rand McNally.

SCHEUNEMAN, J.D. (1975): «A new method of assessing bias in test items». Comunicación presentada en la reunión anual de la AERA. Washington DC, abril (Servicio de Reproducción de Documentos ERIC, n.° 106 359).

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

SKAKUN, E.N. y KLING, S. (1980): «Comparability of methods for setting standards». *Journal of Educational Measurement,* 17, 229-235.

SUBKOVIAK, M.J. (1976): «Estimating reliability from a single administration of a criterion-referenced test». *Journal of Educational Measurement,* 13, 265-275.

TIEMANN, P.W. y MARKLE, S.M. (1978): *Analyzing Instructional Content: A guide to Instruction and Evaluation* (3.ª ed.). Champaign, Illinois, Stipes Publishing Company.

TIEMANN, P.W. y MARKLE, S.M. (1985): «Domain-referenced testing of conceptual learning». Comunicación presentada en la reunión anual de la AERA, Toronto, marzo.