Medición criterial Vs. Normativa

Francisco Javier TEJEDOR TEJEDOR

Universidad de Salamanca.

EVALUACIÓN NORMATIVA VS. EVALUACIÓN CRITERIAL

1. *Desarrollo curricular: papel de la evaluación y utilización de pruebas.*

Creo que no hay ninguna duda en admitir, al menos a nivel teórico, que la evaluación es un eslabón más de la cadena del proceso de enseñanza-aprendizaje y que, por tanto, no hay razón alguna para no entenderla como formativa, admitiendo su función de retroalimentación, de motivación generadora de nueva actividad, de potencial de supervisión y mejora del propio diseño curricular.

En la práctica, la evaluación desempeña un papel sancionador; en el mejor de los casos, como señala Gimeno (1981, p. 216): "Pretende ser una toma de conciencia de la distancia que separa al alumno de unos objetivos pretendidos, y que en las prácticas al uso son incluso no explícitos".

Esta importante disfunción es el motivo de la preocupación permanente por la evaluación, que se traduce en posturas antagónicas respecto al sentido de su necesidad, de su marco de referencia, de su soporte ideológico y, como consecuencia, de los condicionamientos de su realización y recursos técnicos implicados.

Las nuevas tecnologías y concepciones de la enseñanza, cada vez más identificadas con los contenidos diferenciales de aprendizaje y con las características del alumno, han puesto de manifiesto la exigencia de instrumentos específicos para controlar los niveles de rendimiento de cada tipo de enseñanza. El interés de estudio actualmente se vincula a los objetivos de enseñanza, lo que ha supuesto un cambio en el desarrollo y realización de las prácticas evaluativas, que se refleja en el cuestionamiento actual de las prácticas tradicionales de la evaluación, admitien-

do para ella nuevas funciones en el desarrollo curricular. Gimero (1981, pp. 215-225) expresa muy claramente esta idea:

> "El gran valor de la evaluación está en ser un instrumento de investigación en la didáctica: comprobar hipótesis de acción para ir acumulando recursos metodológicos que tienen una eficacia comprobada en la acción, e ir engrosando de esta manera el apartado de la técnica pedagógica fundamentada científicamente... La evaluación fecunda, los datos de evaluación de valor científico en lo psicológico y en lo didáctico, es la evaluación que compara el contenido psicopedagógico que contienen los objetivos con la calidad de los procesos y resultados de aprendizaje".

En contraposición a la evaluación tradicional se propugna la evaluación holística, integral. Las razones que apoyan esta elección serían:

a) La riqueza de objetivos y la variedad de efectos metodológicos reclaman una amplitud en la evaluación.

b) Cualquier efecto educativo es el resultado codeterminante de múltiples factores. Requiere un modelo de aprendizaje que comprenda a su vez el modelo explicativo de la enseñanza.

c) Los resultados del proceso son expresión de las interacciones de todos los componentes del modelo didáctico.

d)Entender la evaluación como base para una didáctica recuperativa.

El carácter holístico del proceso evaluativo se manifiesta igualmente al considerar que, en el contexto del proceso de enseñanza-aprendizaje, la evaluación se relaciona, en opinión de Rivas y Alcantud (1988), con los distintos elementos que constituyen dicho proceso:

—elementos estructurales (condiciones organizativas, equipos, materiales,...)

—procedimientos (actividades de enseñanza-aprendizaje, secuencialización de contenidos,...)

—productos (efectos detectados en el aprendizaje,...).

Finalizado el proceso de instrucción de una unidad, cabe valorar el grado de adecuado funcionamiento de cada uno de estos elementos intervinientes. Solo así, de forma sistemática e interactiva, es posible conocer el papel de todos ellos en el proceso. Necesitamos aportar diversos tipos de informaciones para analizar, mejorar y corregir las acciones educativas emprendidas. De esa manera, la evaluación conecta al profesor, al currículo y al aprendiz.

Aparece también como elemento novedoso en este planteamiento holístico la necesidad de recurrir a las diversas fuentes de información, como camino para aproximarnos a una deseable objetividad. Necesariamente hemos de preguntarnos por su viabilidad y por las posibilidades de compaginar la "evaluación holística" con el imprescindible presupuesto científico de la disociación de los fenómenos para posibilitar su estudio.

La gran verdad, a mi entender, es que en el tema de la reforma y desarrollo del currículo se ha podido prestar quizá excesiva atención a los estudios teóricos de

finalidades y objetivos, a los materiales y métodos y quizá no tanta como se debiera a los resultados alcanzados, lo que puede llegar a desempeñar un papel de "retracción", de excesiva precaución, hasta el punto que pueda suponer un cuestionamiento efectivo de ciertas innovaciones deseables.

Parece conveniente entonces llegar a mantener un equilibrio entre las actividades de desarrollo del currículo y la evaluación; en caso contrario, las fuerzas del conservadurismo dinámico pueden prevalecer y detenerse las actividades de desarrollo o, como máximo, ser continuadas por grupos aislados de entusiastas (Cave, 1979, p. 113).

En opinión de Fernández Pérez (1974, p. 36) serían tres los aspectos más importantes que están condicionando el cambio en el profesor al realizar la evaluación. Nuestra propia síntesis respecto a su opinión la expresaríamos, esquemáticamente, en los siguientes términos:

a) La falta de preparación de tipo técnico del profesorado, por carecer de información respecto a temas tales como: definición operacional del rendimiento escolar; condiciones de validez de las pruebas o exámenes que aplica; condiciones de fiabilidad aplicada a la medición de aspectos de la conducta humana...

b) La estructura objetiva del sistema de autoridad.

c) La resistencia al cambio en los procedimientos de evaluación.

Otro argumento utilizado a menudo para justificar el "abandono" del estudio de la adecuación entre el desarrollo del currículo y la evaluación, reside en manifestar que no pocos de los objetivos propuestos son tan complejos que no son susceptibles de evaluación... Esta postura, radicalizada, imposibilita realmente toda acción evaluativa y no significa más que, a mi entender, una actitud negativa ante el propio proceso de racionalización del currículo.

Yo prefiero pensar, de acuerdo con Rivas y Alcantud (1988, p. 23), que el rendimiento escolar, conceptualizado como el producto final del proceso de enseñanza-aprendizaje que un individuo obtiene a lo largo de dicho proceso en un curso, nivel o materia, es la base para caracterizar el resultado de la acción educativa individual y de ahí, trasladable para analizar y mejorar el propio proceso instructivo. El rendimiento escolar se operacionaliza en las notas y calificaciones que pasan de esa forma a tener valor y reconocimiento social. Toda preocupación por cómo llegar a establecerlas adecuadamente será por tanto especialmente relevante.

Una vez asumida la conveniencia de profundizar en el sentido de la evaluación, vamos a comentar algunos aspectos relacionados con la utilización de pruebas, asumiendo desde este mismo momento que la aplicación de pruebas de evaluación del rendimiento no es, en modo alguno, la única posibilidad de obtener información para realizar la evaluación; eso sí, creemos que es una forma muy extendida de realizar la evaluación del rendimiento y que supone, en principio, un deseo de objetivar el proceso de medición que necesita de planteamientos más rigurosos a la hora de interpretar los resultados. Planteamientos rigurosos que suponen dos niveles de acción:

a) Adecuación de las pruebas a los objetivos.

b) Análisis técnico de las pruebas y de los items que las conforman.

Comenta Wheeler (1976, p. 300):

> "Cuando las escuelas, los profesores y los alumnos se valoran hasta cierto punto por los resultados de las puebas, la índole de la prueba determina en gran medida la calidad del proceso escolar... lo que los profesores enseñan y como lo enseñan, lo que los alumnos aprendan y como lo aprenden..."

Lafourcade (1977, p. 213) refuerza nuestro punto de vista cuando opina:

> "El análisis de los ítems que integran un test de rendimiento preparado por el maestro o profesor representa una tarea que indefectiblemente debe ser emprendida si se desea conocer la real efectividad de la enseñanza y de los instrumentos utilizados en la evaluación. El análisis deberá programar:
>
> a) Una buena información acerca del logro de los objetivos discriminados en la unidad e indicados en la correspondiente tabla de especificaciones.
>
> b) Una estimación de las características psicométricas de los ítems.
>
> c) Una eficaz y permanente preocupación por mejorar los instrumentos de evaluación".

2. *Aplicaciones psicométricas tradicionales.*

En el modelo normativo, el evaluador toma explícita o implícitamente como referencia la ejecución de los sujetos que pertenecen a un mismo grupo llamado "normativo". Las notas características de este grupo y su representatividad afectarán al alcance de la evaluación. Se trata de un modelo basado en la medición de diferencias individuales, asumiendo la distribución normal del rendimiento. La norma del grupo (la media, por ejemplo), utilizada como criterio o punto de corte, será más alta o más baja en función de las capacidades de los sujetos que componen dicho grupo.

Las etapas que habitualmente se siguen en la construcción de una prueba de referencia normativa, en opinión de Mateo (1987), son las siguientes:

1ª) Establecimiento del marco referencial: identificación de conductas o contenidos que se quieren medir; búsqueda de fuentes teóricas, metodológicas o instrumentales relacionadas con la prueba; determinación de los objetivos de la prueba,...

2ª) Etapa de identificación,: identificación de las propiedades que queremos medir a través del test, identificación del objeto que va a ser medido e identificación de la regla de asignación numérica que nos permitirá asignar números a cada propiedad medida.

3ª) Etapa de construcción propiamente dicha: selección de ítems; estandarización de las condiciones internas y externas de la aplicación; especificación de los criterios de puntuación; selección de la unidad de análisis; aplicación de la escala

de ensayo; puntuación, valoración y tabulación de las respuestas; análisis de ítems en la prueba de ensayo; selección de los ítems que constituirán la escala de medida; aplicación de la escala de medida a una unidad de análisis.

4ª) Etapa de objetivización de la prueba: obtención de indicadores de fiabilidad y validez. La fiabilidad puede ser entendida bien como medida de estabilidad, bien como medida de consistencia interna. Por su parte, la validez podrá ser definida como validez de criterio (estimación a partir de las puntuaciones individuales en un test de la puntuación en alguna otra variable llamada criterio); como validez de contenido (estimación del rendimiento de un sujeto en el universo de situaciones que el test intenta representar); o como validez de constructo (se evalúa un test a la luz de un constructo especificado).

5ª) Etapa de obtención de puntuaciones: determinación del tipo de puntuación requerida y de la utilización que de ella va a hacerse.

Respecto al análisis de items, presentamos el esquema que tradicionalmente se sigue en el marco de la construcción normativa de pruebas, al amparo de la teoría clásica de test. Proponemos un nivel de análisis que denominamos "de explicación", con los siguientes considerandos a determinar:

1) Media y varianza de cada ítem. Su relación con la media y la varianza de la prueba.

2) Dificultad (facilidad) de cada ítem y su poder de discriminación.

3) Homogeneidad de los ítems: correlación biserial, correlación tetracórica, correlación de Pearson.

4) Correlación ítem-prueba.

5) Correlación ítem-criterio.

6) Características psicométricas de la prueba: fiabilidad y validez.

7) Indices de fiabilidad y validez de cada ítem (su contribución a la fiabilidad y a la validez de la prueba).

8) Análisis factorial (interconexión de los ítems y su relación con la consecución de objetivos).

9) Interpretación psicométrica y didáctica.

Este nivel de análisis requiere el manejo de las matrices de vaciado, de varianza-covarianza y de correlaciones entre los ítems. Posibilita la realización de comprobaciones psicométricas: media y varianza de la prueba en función de la medida y varianza de los ítems; varianza de la prueba como suma de los elementos de la matriz varianza-covarianza; definición de validez como cociente entre la suma de los índices de validez y de fiabilidad; definición de desviación típica de la prueba como suma de los índices de fiabilidad de los ítems,...

Podemos preguntarnos qué elementos incorpora este esquema que puedan pensarse como novedosos en relación con propuestas anteriores. En mi opinión serían:

a) Profundizar en los aspectos estríctamente psicométricos, incluyendo en la estrategia de análisis aspectos tales como:

— los índices de fiabilidad y validez de los ítems y su repercusión en la fiabilidad y validez de la prueba.

— la factorización de los ítems de la prueba lo que nos posibilitará conocer la relación entre los ítems de la prueba y su vinculación con los diferentes dominios de objetivos evaluados.

b) Poner de manifiesto una contradicción importante que surge en las interpretaciones psicométrica y didáctica del análisis de ítems, sólo apuntada por algunos autores: "Los mejores ítems desde el punto de vista psicométrico no lo son (no deberían serlo) desde el punto de vista didáctico".

La explicación es simple: entre las características psicométricas elementales y básicas aparecen el índice de dificultad del ítem y, con él relacionado, el índice de discriminación. Desde el punto de vista psicométrico, los mejores ítems son aquellos que tienen un índice de dificultad medio (p=0,50, si son dicotómicos) y un poder de discriminación máximo (p.q.=0,25, si son dicotómicos). Esta calidad psicométrica de los ítems ha venido siendo el criterio básico para su valoración en el contexto de la evaluación educativa, como queda patente en las siguientes citas:

> "Si todos los alumnos responden bien a un estímulo o todos se equivocan, la fiabilidad se verá muy afectada. Hay mayores probabilidades de que los ítems de dificultad media otorguen una mayor consistencia a la prueba... El grado de dificultad que se asigne a los ítems deberá ser tal que solamente la mitad del curso los pueda responder correctamente..." (Lafourcades, 1977, p. 204).
>
> "Es necesario prescindir de los ítems números... ya que no discriminan entre los mejor y los peor preparados y ello debido a que han sido acertados por todos los sujetos..." (Rodríguez Diéguez, 1980, p. 342).

3. *Insuficiencia de las propuestas* psicométricas tradicionales: pruebas de referencia normativa.

La utilización de la medida del rendimiento escolar con los mismos presupuestos teóricos que sirven para la medida psicológica,tiene consecuencias indirectas que resultan ser inadecuadas a la realidad educativa; esto es, la medición psicológica de rasgos aceptados como estados estables y rasgos más o menos permanentes de los individuos, tiene la utilidad de comparar la ejecución o comportamiento habitual de una persona con la obtenida por un grupo de referencia y por ende la comparación entre distintos sujetos que pertenecen a la misma población. El supuesto descansa en la consideración de las diferencias individuales como un "hecho natural" y supone la consideración formal de todos los supuestos básicos de la teoría clásica de test, que en su proyección al ámbito evaluativo supone (Rivas y Alcantud, 1988, p.31):

- la puntuación individual se interpreta en función de los rendimientos del grupo al que el sujeto pertenece

- la puntación permite la comparación entre los distintos individuos

- la puntuación da una idea global de la realización del sujeto pero no permite establecer estrategias individuales de mejora o corrección

La "Trampa psicológica" de que los elementos más deseables en medición psicológica sean los de varianza máxima no tiene sentido en educación pues significa optar por un nivel de rendimiento "óptimo" del 50% de las acciones realizadas. Por esta razón es necesario variar los supuestos de evaluación educativa, que ahora se fundamentarán en los siguientes considerandos:

a) La ejecución individual se constatará con algún criterio estándar, fijado previamente y eceptado como valioso

b) De la ejecución individual interesa especialmente la composición analítica del contenido de medida o la especificación de los procesos implicados en toda ejecución

c) Los resultados obtenidos permitirán la puesta en marcha de estrategias individualizadas de superación o al menos la toma de decisiones adecuadas a cada caso

d) Los criterios de selección de los elementos deberán descansar sobre planteamientos válidos y no sólo en consideraciones estadísticas de distribuciones prefijadas

La aceptación de estos presupuestos supone integrarse plenamente en la perspectiva de evaluación criterial.

Un test o prueba de referencia criterial es aquel que evalúa el status absoluto de logros de un estudiante. Estos instrumentos interpretan las puntuaciones de forma individual, sin establecer comparaciones entre los individuos, tomando como referente un estándar prefijado y no la ejecución del grupo.

Uno de los problemas más arduos que presenta este tipo de instrumentos es el de la conceptualización del criterio. La palabra "criterio" ha tenido varias acepciones generando un cierto confusionismo. Así, el término criterio se puede identificar: con un estándar de ejecución; con un objetivo conductual o con un dominio o logro preestablecido

3.1. *Historia*

Aunque podemos admitir, de acuerdo con Jornet (1986), que ya Thorndike en 1913 llegó a proponer la necesidad de una referencia absoluta para la interpretación de las puntuaciones de los sujetos en los test de rendimiento, diferenciándola de la referencia normativa, no será hasta 1963 con la obra de Glaser "Instructional Technology and the Measurements of Learning Outcomes" cuando de hecho se establezcan claramente las diferencias entre las expresiones "evaluación basada en normas" y "evaluación basada en criterios". Una vez establecidas las diferencias iniciales entre ambos tipos de evaluación, los primeros trabajos en el ámbito de la evaluación criterial se vinculan con el estudio de los métodos de definición del universo de medida, intentando establecer técnicas que permitan estimar la puntuación de un sujeto definida a partir de su ejecución en una muestra de elementos del universo.

En los años setenta, es sin duda el trabajo de Popham quien mejor representa los esfuerzos por sistematizar las nuevas aportaciones que van surgiendo cada vez con mayor fuerza y cuyos hitos principales serían los siguientes:

– Primera sistematización formal de la metodología propia de la evaluación referida a criterio (ERC), realizada por Millman y editada por Popham en 1974.

– Publicación de monografías sobre los problemas técnicos de la ERC.

– Publicación en 1976 de una revisión completa del tema de la determinación de estándares (Meskaukas).

– Publicación de textos que integran los dos enfoques de la construcción del test.

– Publicación del primer texto específico (Popham, 1978, edición española de 1983) en torno al tema de los estándares y punto de corte.

– Celebración del primer congreso y publicación por Berk en 1980 de los trabajos allí reunidos.

Será a partir de esa publicación cuando puede establecerse el comienzo de la última etapa del proceso de conformación de esta nueva corriente evaluadora. Los hitos que en esta etapa marcan la evolución realizada serán:

– La publicación por Hambleton (1980) de una monografía sobre toda la problemática técnico- metodológica, incluyendo ya referencias específicas a la fiabilidad y a la validez.

– Sistematización de las técnicas de formulación de items, realizada por Roid y Haladyna en 1982, retomándose el interés por los métodos de especificación del dominio como elemento básico en la construcción de un TRC.

– Publicación de manuales para la determinación de estándares de ejecución en test educativos.

– Publicación por Berk en 1984 de una revisión de su libro de 1980, actualizando el tratamiento de los temas más relevantes: formulación y análisis de items, determinación de estándares, fiabilidad, validez, análisis de sesgos,...

3.2. *Elementos básicos en la construcción de un TRC*

Son varias las definiciones que se han ofrecido sobre el concepto de criterio en este esquema evaluativo. Para Popham (1980), las medidas referidas al criterio serán aquellas que se utilizan para evaluar el status absoluto individual con respecto a algún criterio de rendimiento. Por el hecho de comparar la ejecución individual con algún criterio en lugar de con lo realizado por otros sujetos es por lo que se denominan medidas referidas al criterio.

En otros casos se ha intentado caracterizar los test orientados al criterio como aquellos que definen explícitamente las reglas que relacionan los rendimientos del test con referentes conductuales.

También pueden entenderse como pruebas construidas para obtener medidas directamente interpretables en términos de realizaciones concretas. Con una perspectiva más general, puede entenderse que un TRC será aquél que especifique la puntuación en el criterio sin hacer referencia a la distribución de puntuaciones del

grupo de sujetos examinados. Su aplicación nos ofrece información acerca del conocimiento de los examinados y produce puntuaciones interpretables en términos de tareas o logros académicos.

Aunque, como vemos, no existe total acuerdo en las definiciones dadas por los distintos autores sobre la evaluación de referencia criterial (ERC), sí que pueden extraerse algunos presupuestos implícitos en todos ellos:

- Existe un dominio bien definido, en términos de conducta; estas conductas se especifican antes de que se contruya el test.

- Se seleccionan las tareas del test de forma que constituyan un conjunto representativo del dominio que se quiere medir.

- Las puntuaciones son interpretables en términos de tareas o realizaciones.

- La valoración del sujeto es individual, no dependiente de su posición relativa en el grupo.

- La determinación del punto de corte se establece a priori.

- El estándar fijado depende de la tarea y es de carácter absoluto.

Hay dos tipos fundamentales de TRC:

- Test referidos al dominio, vinculados a programas educativos muy estructurados.

- Test de Aptitud o Mastery Test, destinados a medir la ejecución educativa de forma genérica, sin especificar tanto el dominio.

En cualquiera de estos dos tipos, los elementos básicos en la construcción de un TRC según Jornet (1986) serían:

a) Especificación y definición del dominio; la calidad de la descripción del dominio es lo que permite referir las puntuaciones de los sujetos a un criterio.

b) Análisis de elementos, concebido como sistema de control de la calidad métrica de los items.

c)Determinación del criterio de suficiencia o estándar.

d) Fiabilidad, concebida en términos distintos repecto a los modelos normativos y centrándose básicamente en el análisis de la consistencia de las clasificaciones de logro.

e) Validez, quizá la máxima aportación del esfuerzo criterial, al partir de una mejor definición del universo de medida.

Analicemos estos elementos con algún detenimiento.

3.3. *Especificación del dominio*

Es sin duda el tema central sobre el que gira la construcción de un TRC. Será la calidad de la definición del dominio lo que va a posibilitar la referencia de las puntuaciones individuales a criterios internos de calidad.

Por "dominio" se entiende el conjunto de todos aquellos elementos (objetivos, acciones, tareas, items) que representan el propósito de la instrucción. Desde el punto de vista métrico, el dominio constituye el "universo de medida", el cual incluye todas las unidades que lo definen.

Un dominio está bien definido si sus unidades (objetivos e items) están bien especificadas y adecuadamente estructuradas, lo que puede conseguirse bien a partir de la opinión de jueces, bien por análisis empíricos.

Serán tres los niveles posibles de definición del dominio: taxonómico, por objetivos y por items. Sólo se alcanza en este último nivel el rigor preciso para fundamentar adecuadamente la evaluación.

3.4. *Análisis de elementos*

En el análisis de elementos que se realiza en una concepción criterial de la evaluación, lo que más va a interesarnos es conseguir la representatividad del elemento respecto al universo de medida, entendida en el aspecto educativo como congruencia entre el item y el objetivo, y ello en detrimento de la preocupación prioritaria y exclusiva por los parámetros de los item en el ámbito normativo (dificultad y discriminación). Estas dos preocupaciones -congruencia y eficacia en la medida- se integrarán aquí de forma mucho más armónica, respondiendo a lo que ha venido denominándose revisiones lógica y empírica, integradas en un mismo proceso de análisis de items en el ámbito de la ERC, cuyas fases serían, en opinión de Berk (1984), las siguientes:

– Revisión lógica del elemento para valorar su congruencia con el objetivo del que surge.

– Análisis de las respuestas para identificar qué partes del elemento requieren revisión o cambio.

– Análisis estadístico para evaluar la efectividad de los elementos en sus aspectos métricos.

– Revisión de los resultados estadísticos para determinar la selección de items.

La revisión lógica implica considerar las tres características básicas que afectan directamente a la validez de un item:

a) Congruencia item-objetivo, que no se determinará ya, como en las medidas de referencia normativa , mediante la correlación item-criterio externo al test, sino como adecuación del item con los diferentes objetivos prefijados, medida según la opinión de distintos jueces. Existen distintos índices para cuantificar esta congruencia.

b) Calidad técnica, referida sobre todo a la calidad de formulación y a su adecuación al programa de instrucción.

c) Análisis de sesgos, orientado hacia la consideración de aspectos de tipo ético, estético y sociológico.

La revisión empírica se lleva a cabo a partir de la denominada "matriz de vaciado de items", en la que se incorporan la puntuaciones obtenidas por cada uno de los sujetos en cada uno de los items.

Se pretende conocer el índice de dificultad y el poder de discriminación de cada uno de los items, parámetros clásicos de análisis en los planteamientos normativos, aunque ahora se tratará de redefinir su valoración. Así por ejemplo, en la

teoría normativa se prefieren items con índice de dificultad próximos a 0,50 a fin de que su poder de discriminación, medido por su varianza, sea máxima. En el contexto de la evaluación criterial se trata de conocer estos índices para analizar los efectos de la intrucción al comparar los valores pre y post-intrucción. Deseamos encontrar items con valores p próximos a cero en la medida pre-instrucción y con valores p próximos a uno en la medida post-instrucción. Esta capacidad del item para valorar los efectos de la instrucción a partir del cambio del valor del índice de dificultad del item es lo que se denomina "sensitividad instruccional".

El índice de discriminación se entenderá ahora como el índice que mide los cambios de ejecución (pre-test, post-test) o las diferencias entre los grupos en el criterio. Se basa en la idea de que un TRC maximizará las diferencias entre grupos y minimizará las diferencias entre sujetos de un mismo grupo. Las diferencias que se dan entre grupos se atribuirán al efecto del programa de instrucción.

Va a seguir manteniéndose casi en los mismos términos la interpretación del índice de homogeneidad de los items, entendida como la correlación entre el item y el total del test.

Se añade como tema importante la preocupación por la validez del item, exigiéndole capacidad para informar sobre los efectos de la instrucción. Los índices elaborados para medir esos efectos responden a muy distintos planteamientos estadísticos, que van desde los índices basados en proporciones hasta los modelos de rasgo latente, pasando por aproximaciones correlacionales y bayesianas.

3.5. *Determinación de estándares o puntos de corte*

Debe establecerse con claridad si un procedimiento se encamina a la determinación del punto de corte en la escala de puntuaciones verdaderas (establecimiento de estándares) o bien si está destinado a la determinación de un punto de corte en la escala de puntuaciones observadas (establecimiento de un punto de corte).

Cuando el objetivo es el establecimiento de estándares deseamos encontrar un nivel que garantice el adecuado aprendizaje del dominio instrucional de los individuos que lo alcancen o sobrepasen. Este nivel está expresado en puntuaciones libres de error y para su determinación se recurre a la opinión de jueces.

Los métodos destinados a la determinación de una puntuación de corte, a partir de un estándar previamente fijado, se ocupan de trasladar dicho valor a la escala observada, teniendo en cuenta las diferencias que se dan en la medición y, generalmente, optimizando las consecuencias resultantes de la decisión.

Rivas y Alcantud (1988, p.53) establecen tres grandes grupos en los métodos para el establecimiento de puntos de corte:

a) Métodos convencionales o psicométricos, donde se trata de lograr la mayor correlación entre la clasificación del rendimiento de los estudiantes respecto a un punto de superación con un criterio interno o externo

b)Métodos probalísticos, donde se parte de la teoría de la medida representacional de un universo de tal manera que el fijar el valor de un punto de corte en el continuo determine la toma de decisión subsiguiente

c) Métodos basados en la experimentación pre/post instrucción, donde se determina un punto de corte que espera a los sujetos con nivel de rendimiento alto de los de nivel bajo (exactamente sujetos pre y post instrucción).

3.6. *Fiabilidad*

En la fiabilidad en las pruebas de referencia criterial no sólo tendrá importancia la fiabilidad de las mediciones sino también otras cuestiones como la consistencia de las interpretaciones que se hagan a partir de dichas mediciones y de las decisiones que se derivan.

Hambleton y otros (1978) definieron tres grandes categorías de fiabilidad en el contexto de los TRC:

1) Fiabilidad de las decisiones de clasificación o consistencia en la toma de decisión de clasificación como aptos/no aptos. La estrategia general consistirá en aplicar un mismo test (formas paralelas) a un mismo grupo de sujetos, sucesivamente, y comprobar en qué medida existe acuerdo entre las clasificaciones derivadas de ambas aplicaciones. Si existe consistencia en la clasificación de cada individuo a través de las aplicaciones se estima que la prueba es fiable.

2) Fiabilidad de las puntuaciones del test referidas al criterio o consistencia de las desviaciones, cuadráticas o lineales, de las puntuaciones individuales desde la puntuación del punto de corte.

3) Fiabilidad de las estimaciones de la puntuación en el dominio o consistencia de las puntuaciones individuales a través de test paralelos. Destacamos el índice de Brennan como el más adecuado para medir la fiabilidad así entendida, concebido dentro de la teoría de la generalizabilidad y cuya característica más relevante es que es independiente del punto de corte y por ello su utilidad estará orientada a aquellas pruebas que no pretenden tomar decisiones.

3.7. *Validez*

Si en el modelo clásico utilizado en la evaluación normativa se asumen básicamente tres tipos de validez (de contenido, de constructo y predictiva), dentro del marco de la ERC se han producido pocas variaciones conceptuales pese al interés que se concede a este tema.

La validez de una puntuación se definirá ahora como la precisión con que las puntuaciones del test se pueden utilizar para lograr un objetivo establecido. La validez se refiere a las puntuaciones y no al test mismo.

De alguna manera se abandona el interés por la validez predictiva y se pone más énfasis en el estudio de los otros dos tipos mencionados.

En relación con la validez de contenido, se proponen dos nuevas modalidades denominadas "validez curricular" y "validez instruccional". La curricular hace referencia a la adecuación y representatividad de los diferentes tópicos en el mismo continuo: item-objetivos-curriculum. La validez instruccional se refiere al grado de ajuste entre los objetivos medidos en un test y los objetivos desarrollados en el aula. Ambas pueden considerarse sinónimas de la validez de contenido, a la que, cara a la representatividad de los items en relación con el dominio, se le va a

exigir una definición clara de dicho dominio y la especificación clara del plan de muestreo usado en la selección de elementos.

El proceso seguido en el marco de la ERC para el establecimiento de la validez de constructo es similar al procedimiento aplicado para valorar las teorías científicas, cuyas etapas serían las siguientes:

1) Estudio lógico de la prueba para determinar los constructos que supuestamente están en la base de la ejecución.

2) Analizar la relación entre variables observables, entre éstas y los constructos y entre éstos.

3) Derivar de las etapas anteriores consecuencias prácticas.

4. *Perspectivas técnicas y viabilidad de su incorporación a la evaluación en el ámbito educativo*

Estoy de acuerdo con Sawin (1970) cuando afirma que es importantísimo aclara la finalidad de aplicación de una prueba que, al menos, puede tener una doble intención:

a) Diferenciación de los sujetos (sería válido entonces el criterio psicométrico; es lo que suelen pretender los test estandarizados)

b) Evaluación de los efectos de la enseñanza, donde debe prevalecer el criterio didáctico; Popham y Baker (1970, pp. 62-75) van más allá todavía y diferencian ente " objetivos mínimos de la clase" y " objetivos mínimos del alumno".

Se aconseja una reflexión del profesor respecto al papel que va a desempeñar la prueba que aplica. Parece razonable, siempre bajo el principio fundamental de que es la naturaleza del objetivo quién determina el tipo de item, aceptar los criterios psicométricos cuando se pretenden analizar "objetivos mínimos de clase" o interesa sobremanera la diferenciación. Serán aconsejables criterios didácticos cuando se pretende analizar la consecución de "objetivos mínimos de alumnos" en el contexto de la evaluación de los efectos de la enseñanza.

Adoptar esta decisión implica situarse en el contexto de la evaluación referida a criterio, lo que exigiría el seguimiento de las pautas anteriormente señaladas en relación con las distintas fases del proceso de construcción de pruebas, con interés especial por la validez de la prueba. Entendiendo por tal en este contexto la precisión con que una prueba mide la conducta especificada (conducta, no contenido) en el objetivo sometido a comprobación. Las soluciones de tratamiento podrían ser:

En primer lugar, una potenciación de la validez de contenido, lo que puede conseguirse, y entronca directamente con la relación objetivos-evaluación, a través de las "Tablas de especificaciones" (de dominios y de objetivos), que sugieren una elección representativa de los items en función de la gama de objetivos a evaluar.

En segundo lugar, y con el mismo sustrato de fondo, es decir al adecuación entre objetivos y evaluación, una revisión factorial de la prueba que pueda ayudarnos a entender la relación existente entre lo que miden determinados bloques de items (factores) y los objetivos que deseamos evaluar mediante la prueba.

En tercer lugar, la verificación de hipótesis a través de constructos (validez de construtos); proceso lento, que requiere sucesivos tanteos pero sin más dificultades metodológicas que la especificación de constructos revelantes.

En cuarto lugar, el estudio de la denominada por Loevinger (1957) "validez estructural", basándose en trabajos anteriores de Coombs (1953) y Peak (1953). De forma muy general, la validez estructural plantea la posibilidad de relacionar estructuras de conducta del rasgo que está siendo medido pero que no han sido incluidas en el test o prueba.

Sin duda, el modelo que Wardrop y otros formularon en 1982 incorporaba todo el conjunto de considerandos en torno al tema que tratamos. Presentamos dicho modelo como proyecto asumible (cuadro 1)

En el modelo se incluyen cinco dimensiones a considerar respecto a la aplicación de pruebas:

A) Utilización de pruebas

B) Procedimientos para seleccionar y generar items

C) Proceso de revisión del item

D) Estudio de la precisión o fiabilidad

E) Aproximación a la validación

En A se plantea, como vemos, el continuo "diferenciación-medida" que pretende dar respuesta a las intenciones subyacentes en la aplicación de la prueba; en el continuo se delimitan varios puntos intermedios que matizan las posibles "intenciones" en la aplicación de las pruebas. En B,C,D, y E se establecen continuos referidos a cada una de las dimensiones consideradas en la forma que vemos en la figura expuesta.

Como todo modelo, es una propuesta de explicación que necesita un permanente cuestionamiento para verificar su utilidad y para verificar si efectivamente posibilita la explicación de las hipótesis subyacentes. Como aportaciones más válidas destacaríamos:

1) El planteamiento en sí del problema; es decir, el análisis necesario que debe implicar toda prueba de evaluación del rendimiento; análisis que debe realizarse en dos vertientes: la psicométrica y la didáctica (adecuación objetivos-prueba de evaluación).

2) Reconocer que el problema surge por la preocupación por el "ajuste didáctico"; es decir, las soluciones desde el punto de vista psicométrico ya estaban dadas pero se referían a pruebas de diferenciación; el problema, como digo, surge al pensar en la adecuación de esas soluciones a los requerimientos didácticos para racionalizar la aplicación de pruebas al proceso de evaluación, en el contexto de los diseños curriculares.

3) Presentar como posibles técnicas de ayuda a la solución del problema la fiabilidad intrasujetos, la validación de constructos y la validación estructural.

4) Plantear como conveniente la realización de inferencias, con base en el estudio de los items, respecto a relaciones estructurales no medidas en la prueba.

Terminamos nuestra exposición haciendo referencia, esquemáticamente, a las cautelas que consideramos han de tenerse en cuenta a la hora de planificar la actividad evaluativa:

a) Determinación previa de la finalidad de la prueba que se aplica.

b) Determinación de las características psicométricas de los items que conforman la prueba, con el fin de disponer de "información añadida" a la hora de interpretar los resultados.

c) Elaboración de las tablas de especificaciones, lo que requiere a mi entender el estudio previo de las características psicométricas de los items indicando en el punto anterior.

d) Valoración didáctica del proceso de medición, resultado de la aplicación de la prueba, lo que estará posibilitando el planteamiento de nuevas hipótesis de trabajo y en consecuencia la mejora del propio modelo didáctico seguido en el diseño del currículo, siempre con carácter de provisionalidad.

e) Integración de los conclusiones a las que hayamos podido llegar por esta vía con la procedentes de otros esquemas de actuación no necesariamente cuantitativos, a fin de aproximarnos a la consecución última de todo proceso de evaluación: su integración plena en el desarrollo del currículo, lo que la convierte inexcusablemente e formativa, en generadora de cambios educativos deseables y, en definitiva, en instrumento de mejora permanente del desarrollo del currículo, vía perfeccionamiento del propio modelo didáctico que los sustenta...

f) Someter a procesos de discusión crítica la procedencia de realizar evaluación criterial. Cuando esté claramente establecida su conveniencia desde el punto de vista didáctico, comenzar a realizar estudios sobre las características de estas pruebas en los términos aquí sugeridos.

Cuadro 1

USES OF TEST					
A.Differentiation					Measurement
maximizing job performance	fairness in job allocation	minimizing effort and disappointment in training	certifyng competence	iagnosing strength and weakness	tracking progress
ITEM GENERATION					
B. Descriptive categories					Generative Rules
Table of specificacions	list(catalog) of objectives without theory		ordered list of objectives	theretical partitioning of specified set	
ITEM AND TEST REVISION					
C. Focus on items	Focus on Rules				
	selecting and fine funing items		adjusting objetives	modifying rules theoriesgenerating or selecting items	

ASSESSMENT OF PRECISION			
D. Intersubject			Intrasubject
one time measure internal consistency,split-halves, test-retest	alternate forms	generalizability theory	repeated individual measures, time series, function fitting
VALIDATION			
E. External			Internal
correlation with external related criteria and test	content	construct, based on external and internal criteria	structural
Descriptive model for prolifing test characteristics			

REFERENCIAS BIBLIOGRAFICAS.

Bechtel, G. G. y Ofir, C. (1988): "Aggregate items response analysis". *Psychometrika*, vol. 53, 1, 93-107.

Berk, A. (1980): *Criterion-referenced measurement.*The Johns Hopkins University Press, Baltimore, Maryland.

Berk, A. (1984): *A guide to criterion-referenced test construction.*The Jhons Hopkins Univ. Press, Baltimore.

Black, H. D. y Dockrell, W. B. (1984): *Criterion-referenced assessment in the classroom.* The Scottish Council for Research in Education. Macdonald Printers, Edinburgh.

Carreño Huerta, F. (1977): *Enfoques y principios teóricos de la evaluación.* Trillas, México.

Cave, R. G. (1979): *Introducción a la programación educativa.* Madrid, Anaya.

Coombs, C. H. (1953): "Theory and methods of social measurement". En L. Festinger y D. Katz (Eds.) *Research methods in the behavioral sciences.* Dryden, New York.

Dendaluce, I (1899) (coord): *Aspectos metodológicos de la investigación educativa.* II Congreso Mundial Vasco. Narcea, Madrid.

Fernández Pérez, M. (1974): *Evaluación escolar y cambio educativo.* Cincel, Madrid.

Gimeno, J. (1981): *Teoría de la enseñanza y desarrollo del currículo*, Anaya, Madrid.

Glaser, R. (1963): "Instructional technology and the measurement of learning autocomes". *American Psichologist, 18, 519-554.*

Haertel, E. (1985): "Construct validity and criterion-referenced Testing". *Review of Educational Research.* Vol. 55, 1, 23-46.

Hambleton, R. K. (1980): *Review methods for criterion-referenced test items.* Paper en Anual Meeting American Educational Research Ass., Boston.

Hambleton, R. K. (Ed.) (1983):*Applications of item response theory.* Vancouver, Educational Research Institute of British Columbia.

Hambleton, R. K. (1983): "Application of item response models to criterion referenced assessment". *Applied Psychological Measurement,* Vol. 7, 1, 33-44.

Hambleton, R. K. y De Gruijter, D. n. (1983): "Application of item response models to criterion-referenced test item selection". *Journal of Educational Measurement,* Vol. 20, 4, 1983.

Hambleton, R. K. y otros (1978): "Criterion-referenced testing and measurement: A review tchnical issues and developments". *Review of Educationtal Research,* 48, 1, 1-47.

Jornet, J. M. (1986): *Una aproximación teórico-empírica a los métodos de medición de referencia criterial.* Tesis Doctoral. Universidad de Valencia, Facultad de Filosofía y Ciencias de la Educación.

Lafourcade, P. (1977): *Evaluación de los aprendizajes.* Cincel, Madrid.

Loevinger, J. (1957): "Objetive test as instruments of psychological theory". *Psichological Reports,* 3, pp. 635-694.

Lord, F. M. (1980): *Applications of item response theory to practical testing problems.* Hillsdale, Erlbaum.

Mateo, J. (1997): *Medición y evaluación.* Documento policopiado, Universidad de Barcelona.

Meskaukas, J. A. (1976): "Evaluation models for criterion-referenced testing: Views regarding mastery and stander-setting". *Review od Educational Research,* 46, 1, 133-158.

Millman, J. (1974): "Criterion-referenced measurement". En W. J. Popham (ed) *Evaluation: Current applications.* McCutchan Publishing, Berkeley, California.

Nitko, A. J. (1980): "Distinguishing the many varieties of criterion-referenced test". *Review of Educational Research,* Vol. 50, 3, 461-485.

Peak, H. (1953): "Problems of objective observation". En L. Festinger y D. Katz (Eds) *Research methods in the behavioral sciences.* Dryden, New York.

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

Popham, W. J. (1983): *Evaluación basada en criterios.* Magisterio Español, Madrid.

Popham, W. J. y Baker, E. L. (1970): *Los objetivos de la enseñanza.* Paidos, Buenos Aires.

Rodríguez Diéguez, J. L. (1980): "*Didáctica General: Objetivos y evaluación.* Cincel-Kapelusz,Madrid.

Rodríguez Lajo, M. (1986): " Evaluación del rendimiento criterial versus normativa. Modelo de Evaluación F.C.O.". *Revista de Investigación Educativa,* Vol. 3, nº 6, 304-321.

Roid, G. H. y Haladyna, T. M. (1982): *A tecnology for test-item writing.* Academic press, New York.

Rivas, F. y Alcantud, F. (1986): *Desarrollo de instrumentos de evaluación criterial y cualitativa para la E.G.B..* Memoria de investigación, Universidad de Valencia.

Rivas, F. y Alcantud, F. (1988): *Evaluación criterial en la educación primaria.* CIDE, M.E.C., Madrid.

Sawin, E. I. (1970): *Técnicas básicas de evaluación.* Magisterio Español, Madrid.

Strandmark, N. L. y LINN, R. L. (1987): "A generalized logistic item response model parameterizing test score inappropriateness". *Applied Psychological Measurement,* 4, 355-370.

Stufflebeam, D. L. (1987): *Evaluación sistemática.* Paidos, Buenos Aires.

Van Der Linden, W. J. (1980): "Decision models for use with criterion-referenced test". *Applied Psychological Measurement,*4, 469-492.

Wardrop. J. L. y otros (1982): "A framework for analyzing the inference structure of educational achievement test". *Journal of Educational Measurement,* Vol. 19, n° 1, pp. 1-18.

Webb, N. M., Herman, J. L. y Cabello, B. (1987): "A domain-referenced approach to diagnostic testing using generalizability theory". *Journal of Educational Measurement,* 2, 119-130.

Wheeler, D. K. (1976): *El desarrollo del curriculum escolar.* Aula XXI, Educación Abierta, Santillana, Madrid.