

CONSTRUCCIÓN DE PRUEBAS CRITERIALES DE AULA

Leonor Buendía Eisman, Honorio Salmerón Pérez
Universidad de Granada

I. MARCO TEÓRICO Y LEGAL

Partimos del supuesto de que con programas educativos adaptados a las peculiaridades de los alumnos se puede conseguir que la mayor parte de ellos alcancen los objetivos de un curso o materia de enseñanza.

A partir de este supuesto, las pruebas referidas a normas resultarían incongruentes para la efectividad de la instrucción ya que si la mediación del profesor con su programa funciona, la dispersión de las puntuaciones en las pruebas se reduce dificultando así la interpretación con referencia al grupo (de la Orden, 1982).

— A modo de recuerdo, Hambleton y Swaminathan en 1985 nos informaban de las posibles utilizaciones de las Evaluaciones Referidas a Criterio en el aula y que resumimos en breves líneas:

- Guiar el progreso individual en programas basados en objetivos de aprendizaje.
- Diagnosticar deficiencias de aprendizaje.
- Evaluar programas de acción educativa.
- Orientar competencias en diferentes exámenes y certificaciones.
-

Es en el primer uso, la evaluación de alumnos, donde enmarcamos este trabajo de construcción de pruebas criteriosales, como herramientas al servicio del propio proyecto educativo y del currículum en que se concreta.

De esta forma, creemos que estamos en consonancia con el modelo evaluativo de la LOGSE para este ámbito, (Pérez Juste, 1992, p. 42) ya que, pueden servir a una función formativa de la evaluación, ayudar a determinar el logro de objetivos y al final de cada ciclo utilizarlos para obtener información válida que sirva para decisiones de promoción y/o certificación.

No es nuestro cometido hacer una reflexión teórica sobre lo que son Tests de Referencia Criterial (TRC), que puede verse en la obra de Gray (1978) quien ya aportaba un trabajo sobre 57 definiciones; ni tampoco hablar de los precursores de este término, (Glaser, 1963; Popham y Husek, 1969), sino aportar una experiencia práctica llevada a cabo con profesores de aula de E. Primaria sin que por ello desmerezca el soporte metodológico que avala la calidad técnica para el uso que van a tener.

— Utilizamos el esquema de desarrollo propuesto por Hambleton, R. 1990, hasta la validación de contenido de las pruebas acordando con Beggs y Lewis, 1979, que dado el propósito de las mismas, al

realizar evaluaciones internas en clase con función formativa, lo más importante es demostrar su validez de contenido.

II. DESARROLLO

1. Consideraciones preliminares a la construcción de las pruebas

a) Declaración de los propósitos de las pruebas. Es nuestra intención, dentro de una línea de investigación de nuestro Departamento, fomentar la colaboración de los centros de nuestro entorno, con la Universidad, para mejorar la calidad de la enseñanza. La evaluación al servicio de la educación es una herramienta fundamental si se inserta como tal en los proyectos educativos de los centros.

En esta línea hemos comenzado, en el espacio y en el tiempo de realizar un seguimiento de la implantación de la LOGSE.

Por ello hemos comenzado con el Primer curso del 1º ciclo de E. Primaria.

Las pruebas tendrán una intención de seguimiento individual del progreso de los alumnos en los objetivos de aprendizaje para orientarlos. Sólo los que preparemos para el último proyecto del 2º curso de Primaria tendrán el propósito de emitir certificación tal y como corresponde.

b) Áreas de contenido. Durante el presente curso hemos podido atender, por razones de tiempo, recursos y expertos, las siguientes áreas:

- Conocimiento del medio
- Lenguaje
- Matemáticas

c) Grupos a medir. Fueron todos los alumnos de doce centros del entorno de Granada que en situaciones ordinarias cursaban el 1º año del 1º ciclo de Primaria.

d) El proyecto requirió un grupo amplio de coordinadores de áreas (3), redactores de ítems (6), colaboradores (15) y expertos en áreas de contenido (12) que coordinados por los autores de este informe trabajamos durante el presente curso, 1992-93.

2. Preparación y revisión de objetivos

Desde una metodología racional, se revisaron los objetivos propuestos por diferentes editoriales para este momento del curso académico, para este nivel y para cada área sí como los documentos legales y técnicos del MEC y publicaciones científicas.

— Las editoriales consultadas fueron: Alhambra-Logman, Anaya, Santillana, S. M. Bruño y Luis Vives.

— Mediante esta revisión se llegó a un consenso sobre un listado de objetivos para cada área. (Ver anexo I).

— A continuación y siguiendo la misma metodología, se estableció (hipotetizó) la relación de cada objetivo con los objetivos de área para el Ciclo y para la etapa así como la conexión con los objetivos generales de educación para el ciclo y etapa (ver anexo II).

3. Planificación adicional de la prueba

a) Selección de objetivos para la prueba. Junto al listado de objetivos antedichos, se describieron los contenidos y las tareas con que más podían relacionarse para presentarlo a los profesores de aula al objeto de que seleccionasen aquellos que coincidían con los programados por ellos, una vez definido su proyecto curricular.

Así se seleccionó un listado de objetivos únicos para todos los centros.

b) Redactores de ítems. Actuaron como tal los coordinadores de área: D^a M^a Teresa Pozo Morate (lenguaje); D. Mariano Machuca Aceituno (lenguaje) y D. Nicolás Morcillo Delgado (Matemáticas). Todos ellos alumnos del programa de doctorado de nuestro Departamento, licenciados en Pedagogía y expertos en cada una de las áreas.

4. Primera redacción de ítems

Una vez establecidos los diferentes conocimientos y habilidades que deberían comprender cada objetivo, éstos se transformaron en tareas (ítems) hipotetizando la relación directa entre cada pregunta y el objetivo que pretendía medir (de la Orden, 1982).

5. Validación de contenido

a) Especialistas. Actuaron como expertos en contenido 12 profesores de EGB con experiencia en Ciclo inicial superior a 5 años y más de 10 en la etapa de EGB. Como expertos en medida actuaron los autores de este informe.

b) Aspectos considerados en cada ítem. Determinar la validez de cada ítem implicaba considerar para cada uno estos tres rasgos.

COHERENCIA o grado con que la formulación del ítem se ajustaba al objetivo que dice medir.

REPRESENTATIVIDAD o grado en que la pregunta formulada es considerada como la más acertada de todas las posibles.

CALIDAD TÉCNICA o grado en que la pregunta expresa con claridad, sin pistas gramaticales falsas, la acción que debe realizar el alumno.

Para cada una de estas categorías se dieron instrucciones precisas a los jueces siguiendo las orientaciones de Hambleton, 1984b.

c) Instrumentación. A cada juez se le suministró:

* Una hoja de respuestas que comprendían:

— N^o de objetivos a medir numerados y formulados

— N^o de ítems que pretendían medir tales objetivos.

— Junto a cada ítem, se acompañaba una escala de valoración de 5 puntos. (ver anexo III).

* Un cuadernillo con la redacción de cada ítem y su numeración correspondiente.

Junto a este material, se dieron instrucciones precisas a los jueces de las que extraemos, como ejemplo, las concernientes a la valoración de COHERENCIA:

«Primero lea cuidadosamente la formulación de cada objetivo, así como, los conceptos y habilidades relacionadas con él. Después por favor, indique hasta que punto usted cree que cada pregunta refleja el dominio que pretende medir. Justifique la pregunta solamente sobre la base de ajuste entre su contenido y el contenido definido por el objetivo que la pregunta pretende medir.

Use la escala de valoración que se acompaña como se muestra a continuación»:

Pobre	Regular	Bueno	Muy Bueno	Excelente
1	2	3	4	5

La forma de expresar su juicio será rodeando con un círculo la valoración que consideres en las escalas que siguen a cada ítem.

d) Análisis de resultados:

Se expresan y a modo ejemplificador «Los obtenidos en el —área de lenguaje—; **Índices de congruencia: ítem-objetivo.**

Para este área, dado que cada objetivo pretendía ser medido por un nº de ítems de forma excluyente, o sea, ningún ítem pretendía medir aspectos de dos o más objetivos, se utilizó la formulación propuesta por Hambleton, 1984:

$$I_{ik} = \frac{(N - 1) \sum_{j=1}^R X_{ijk} + N \sum_{j=1}^R X_{ijk} - \sum_{j=1}^R X_{ijk}}{2 (N - 1) n}$$

donde

I_{ik} = índice de congruencia ítem-objetivos

N = nº de jueces

X_{ijk} = Evaluación del juez j para el ítem K en el objetivo i .

Para ello hubo de reducirse a tres todas las valoraciones. -1, 0 y +1, coincidiendo 0 con la posición central (3); -1 para las valoraciones 1 y 2 y +1 para las valoraciones 4 y 5.

Se recogieron en tablas las evaluaciones de los jueces y los índices de congruencia ítem-objetivo hallados para cada uno de los aspectos evaluados.

A continuación fijamos los criterios de adecuación en función de los I_{ik} obtenidos.

Como *criterio absoluto* se fijó a priori el valor de 0'75 (que se considera un índice bastante adecuado para 4 jueces). El criterio relativo se calculó, para cada uno de los aspectos evaluados hallando la media de los índices de congruencia obtenidos en cada uno de ellos por lo que es diferente para cada aspecto enjuiciado, resultando los siguientes valores:

- Coherencia 0'92
- Representatividad 0'8
- Calidad técnica 0'82

b) Experiencia de resultados. En tablas (una para cada área y aspecto evaluado) se recogen las evaluaciones de los jueces y los índices de congruencia ítem objetivo hallados para cada uno de los tres aspectos, colocando un asterisco junto los índices que sobrepasan los valores de los criterios absoluto y relativo (ver anexo IV).

Para tener una visión global de la evaluación de cada ítem recogimos en otra tabla sólo los índices de congruencia de cada uno de los tres aspectos evaluados, (ver anexo V).

Es conveniente que los ítems clasificados según los resultados obtenidos se especifiquen en un cuadro en el que aparezcan los ítems ACEPTADOS, que se admiten sin modificación salvo aquellos que van acompañados de algunos códigos, que en nuestro caso (ver anexo VI) hemos establecido con:

CT: Indica que se debe revisar la calidad técnica del mismo pues ha obtenido en este aspecto un índice excesivamente bajo.

R. CT.: Indicaría que se debe revisar el porqué no se le considera realmente representativo del objetivo con el que se asocia y también la calidad técnica:

En aquellos ítems, en que el Índice de Representatividad es bastante bajo, los hemos aceptado siempre que en Coherencia y Calidad Técnica tengan índices altos. En estos casos partimos del supuesto de que no es relevante que no se vea a este ítem como uno de los más representativos para cubrir un objetivo.

En la misma tabla (ver anexo VI) deben aparecer los ítems POSIBLES y los RECHAZADOS.

Englobamos dentro del grupo de posibles aquellos ítems que, en general, en los tres índices de congruencia no tienen una puntuación inferior a 0'25 del criterio más bajo para cada aspecto, (aquí el

criterio absoluto). En estos casos es necesario revisar estos ítems para observar donde se encuentra el motivo de una baja valoración, teniendo alguno de ellos una puntuación alta en algún índice.

Para las áreas de **matemáticas y conocimiento del medio** se aportan sólo las variaciones metodológicas en el análisis de los enjuiciamientos.

Tales variaciones son producidas por los expertos en contenido y medida al hipotetizar para cada ítem, relación de medida con varios objetivos.

En efecto los especialistas han considerado que tal y como han redactado los ítems, éstos pueden medir aspectos de varios objetivos.

Para este caso utilizamos para el cálculo del índice de congruencia ítem-objetivo la formulación propuesta por Rovinelli y Hambleton, 1977

$$I_{jk} = \frac{N}{2N-2} (\mu_k - \mu)$$

donde:

I_{jk} = índice de congruencia ítem-objetivos.

N = nº de objetivos totales evaluados por los jueces.

μ_k = media de valoración del ítem en el objetivo k.

μ = media del ítem en todos los objetivos.

Esta formulación está pensada para aquellas situaciones en que se presenta a los jueces una lista de objetivos y otra de ítems sin hipotetizar relaciones entre ellos «a priori». Como en nuestro caso se les han presentado ya relacionadas, hemos tenido que partir del supuesto de que todos los jueces evaluarían como -1 aquellos ítem de cuya formulación no se desprende relación alguna con los objetivos que pretenden medir.

6. REVISIÓN DE ÍTEMS

Los datos procedentes de la anterior validación fueron remitidos a los coordinadores de área que a su vez fueron los escritores de ítems para que, tras tener en cuenta el análisis, procedieron a una segunda redacción.

7. EXAMEN DE CAMPO DE LOS ÍTEMS Y ENSAMBLAJE DE LA PRUEBA

Dado el propósito de la prueba, se hizo coincidir con la administración de la misma a los alumnos con el examen de campo.

Ambas situaciones se están desarrollando en el momento de escribir esta comunicación por lo que no nos queda sino especificar el resto de acciones que quedan por realizar para dar por terminada la prueba.

Respecto al examen de campo, será utilizado para detectar preguntas defectuosas, comprobar o chequear sesgos, determinar índices de dificultad y consistencia de puntuaciones o decisiones.

Respecto al ensamblaje de la prueba, determinaremos la longitud total y el nº de preguntas de la prueba por objetivo. Prepararemos la composición, los cuadernos de respuestas de los alumnos y las claves de puntuación.

Esperamos, si es aceptada nuestra comunicación, aportar en su momento toda la información complementaria al proceso llevado hasta el momento.

REFERENCIAS BIBLIOGRÁFICAS

- ARTACHO LÓPEZ, R. (1992): *Elaboración del Proyecto Curricular E.P.* Zaragoza. Luis Vives.
- DEPARTAMENTO DE PROYECTOS DEL GRUPO ANAYA (1992): *Proyecto Curricular para la Educación Primaria-Andalucía*. Madrid. Anaya.
- EQUIPO ANAYA (1992): *Proyecto Las Estaciones*. Madrid. Santillana.
- GLASER, R. (1963): Instructional Technology and the measurement of Learning outcomes. *American Psychologist*, 18, 519-521.
- GRAY, W. M. (1978): A comparison of Piagetian theory and criterion-referenced measurement. *Review of Educational Research*, 48, 223-249.
- GRUPO S. M. (1992): *Proyecto Curricular y Programación*. Madrid. S.M.
- HAMBLETON, R. (1990): Criterion referenced testing. Methods and practices. En GUTKIN, T. y REYNOLDS, C. (Eds.): *The handbook of School Psychology*. New York. Wiley.
- HAMBLETON, R. K. (1984): Validating the test scores. En R. A. BERK (Ed.): *A Guide to Criterion-referenced Test Construction*, 2ª Ed. Baltimore, The Johns Hopkins University Press.
- HAMBLETON, R. K. y SWAMINATHAN, H. (1985): *Item Respose Theory. Principles and Applications*. Boston. Kluwer-Nijhoff.
- JIMÉNEZ MORENO, E. (1992): *Proyecto Albanta*. Madrid. Alhambra Logman.
- JUNTA DE ANDALUCÍA. BOJA. 20-6-92: *Decreto 105/1992 del 9 de Junio*. Sevilla.
- MEC. BOE 26-6-91: *Decreto 1007/1991 del 14 de junio*. Madrid.
- ORDEN HOZ, A. (1982): *La evaluación educativa*. Docencia. Buenos Aires.
- PÉREZ JUSTE, R. (1992): *El modelo evaluativo de la Reforma Educativa*. Madrid. UNED.
- POPHAM, W. J. y HUSEK, T. R. (1969): Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6, 1-9.
- RIVAS, F. y ALLANTUD (1989): *Evaluación criterial en Enseñanza Primaria*. Madrid. CIDE.
- ROVINELLI, R. J. y HAMBLETON, R. K. (1977): On the use content specialist in the assessment of criterion-referenced test ítem validity. *Dutch Journal of Educational Research*, 2, 49-60.