LOS EXÁMENES REFERIDOS AL CRITERIO Y AL CONCEPTO EN CIENCIAS: UN NUEVO SISTEMA DE EVALUACIÓN(1)

SATTERLY, D. y SWANN, N.
Universidad de Bristol. School of Education.

(1) Conferencia presentada en el II Congreso Internacional sobre investigación en la didáctica de las Ciencias y de las Matemáticas, Valencia 1987.

SUMMARY

Some of the processes and effects of the change from norm referenced to criterium referenced testing in science are faced.

INTRODUCCIÓN

Cada vez más los procesos de evaluación se consideran aspectos esenciales del aprendizaje y la enseñanza de las ciencias. Al mismo tiempo, los profesores se sienten insatisfechos de la evaluación por acumulación, es decir, de los exámenes de fin de curso o de ciclo formativo a fin de clasificar a los estudiantes por orden de logros. En su lugar, se concibe un papel más productivo para la evaluación: la motivación de los alumnos a través del establecimiento de objetivos claros, el diagnóstico de las dificultades de aprendizaje y la identificación de errores conceptuales, así como la evaluación del proceso de enseñanza. Hoy se reconoce ampliamente que los métodos de evaluación usados por los profesores pueden afectar profundamente no sólo a la calidad y la cantidad del aprendizaje de los estudiantes sino también a sus caracerísticas afectivas: sus intereses y actitudes hacia las ciencias y su aprendizaje. Sin embargo la mayor parte de los profesores carecen de formación en técnicas de evaluación y elaboración de exámenes y son incapaces de desarrollar pruebas que cubran las necesidades de diagnóstico y ayuden a la planificación curricular. Quizá sea más trascendente el que, incluso donde los profesores de ciencias usan pruebas diseñadas para diagnosticar las dificultades de aprendizaje y para suministrar información precisa sobre la competencia de los estudiantes (como, por ejemplo, en los exámenes referidos a criterios), su interpretación de los resultados a menudo se basa en una comprensión errónea fundamental de su utilidad (Hodson, 1986).

Existen pocas, si las hay, investigaciones a gran escala de la práctica de la evaluación en ciencias. Sin embargo los resultados de una encuesta reseñada por Hodson (1986) están apoyados por las observaciones de los autores. Básicamente existe una gran variedad en la práctica entre escuelas, dentro de las mismas e incluso entre profesores del mismo departamento, sobre el uso de la evaluación para la planificación curricular, en los fines para los que se lleva a cabo, en los tipos de evaluación que se realizan y en la interpretación de las pruebas y otros datos (Sutton, 1986). Otro descubrimiento típico ha sido el de las discrepancias entre lo que los profesores consideran deseable y lo que en realidad ocurre. Así, por ejemplo, en la encuesta de Hodson más del 80 por ciento de los profesores de ciencias que respondieron el cuestionario declararon que siempre o a menudo especificaban los objetivos, sin embargo sólo el 19 por ciento especificaron lo que se pedía a los alumnos que aprendieran. Así pues, parece que, mientras establecer objetivos para los alumnos es considerado esencial casi universalmente, estos objetivos son demasiado imprecisos para permitir la orientación de los alumnos hacia metas particulares o para permitir el diagnóstico y la identificación del siguiente paso apropiado para profesor y alumno. De un modo similar, aunque casi todos los profesores señalaron la identificación de los niños con las dificultades de aprendizaje como la finalidad más importante de la evaluación, casi el 85 por ciento de los que respondieron no estaban seguros de cómo reconocer las dificultades par-

ticulares encontradas. Y aún más, las escuelas enfatizaban considerablemente las reacciones de los niños ante el curso, mientras que menos de la mitad de los profesores de ciencias encuestados hicieron apenas intento de evaluar las actitudes y el interés en la materia (Hodson, 1986 pp. 9-11).

De un modo bastante claro, estas y otras observaciones similares señalan la incertidumbre generalizada de los profesores sobre sus prácticas de evaluación e indican las discrepancias entre lo que idealmente se pide a un programa de exámenes y lo que en realidad ocurre en los departamentos de ciencias. Abrumadoramente, la evaluación consiste en conceder, generalmente al completar un trabajo, notas o grados que sirven entonces para comparar unos estudiantes con otros, o para reflejar los diferentes niveles con que los profesores juzgan que los alumnos han alcanzado criterios subjetivos de éxito o de una actuación aceptable. Estas notas suministran poca información sobre lo que los estudiantes en realidad saben y lo que pueden hacer, tampoco identifican la naturaleza de las dificultades de aprendizaje y aportan una base pobre para comunicar la competencia de los estudiantes a los interesados. Muchos autores mantienen que estas deficiencias son típicas de la evaluación referida a normas que, cuando se hace mal, permite hacer poco más que una comparación con los otros estudiantes. En este sentido indican quien lo ha hecho «mejor», el «peor», el del «término medio», etcétera, pero no pueden detallar con precisión los puntos fuertes y débiles de los estudiantes ni ayudar materialmente a decidir qué enseñar a continuación. Incluso cuando los profesores declaran estar usando exámenes referidos a criterios, el resultado con frecuencia se expresa como una puntuación, un porcentaje o una calificación y, lejos de comparar la actuación de los alumnos con un nivel medio explícito y cuidadosamente definido, lo oscurece todo excepto los juicios comparativos sobre los méritos relativos de los alumnos examinados. Los efectos de tales comparaciones sobre los alumnos más desaventajados son casi con seguridad nocivos.

Aunque los defectos de los métodos de evaluación referidos a normas son bien conocidos, los exámenes construidos correctamente con procedimientos de puntuación objetivos tienen ciertas ventajas sobre aquéllos en los que los profesores emplean criterios subjetivos de un modo idiosincrático. Por ejemplo, Cavemi (1987) ha demostrado que las calificaciones son influenciadas en gran medida por muchos factores extraños a la actuación que en realidad se está evaluando, incluyendo el efecto de halo, consistente en la influencia sobre la calificación de las notas previas del estudiante y del conocimiento que tiene el evaluador del alumno. Además, si los exámenes se construyen intentando medir las habilidades subyacentes que se supone son las responsables de las diferencias individuales de actuación, la teoría clásica de los exámenes permite calcular el error de medida que adjudica explícitamente un cierto margen probable de error a la puntuación o calificación de un estudiante, adoptando así una aproximación científica a lo que vagamente se llama «precisión» de la evaluación. A pesar de la abundancia y el fácil acceso a textos que tratan de la validez de la evaluación, pocos profesores tienen tiempo o pericia para calcular estas estimaciones sobre los exámenes que han diseñado. De todos modos, la opinión generalizada en el Reino Unido se ha vuelto en contra de la actuación referida a normas y es partidaria del potencial tanto práctico como teórico de la actuación referida a criterios.

Las ventajas y la conveniencia de la evaluación referida a criterios sobre la referida a normas han sido descritas claramente por Black y Dockrell (1980) entre muchos otros. Las ventajas que se señalan son:

1) El establecer criterios de actuación específicos y explícitos que se usen para evaluar el trabajo de un estudiante proporciona una fijación de objetivos mucho más clara para los que aprenden y para los profesores.

2) El comparar el trabajo de los estudiantes con criterios específicos permite un diagnóstico de las dificultades del que aprende y el diseño de un perfil más detallado de su competencia y actuación.

3) El aprender a dominar un criterio estimula el aprendizaje por sí mismo y anima a los estudiantes a «competir» contra sus propias actuaciones anteriores, lo que es preferible a competir contra otros estudiantes en un proceso en el que la mayoría está destinada al «fracaso».

4) La evaluación referida a criterios intensifica la motivación del estudiante para «actuar bien» y reconoce los progresos de todos, no sólo los de unos pocos estudiantes afortunados.

Si se pueden construir exámenes que satisfagan estas pretensiones, evidentemente nos ayudarán para proporcionar una evaluación que resulte mucho más útil en la enseñanza y el aprendizaje de lo que han demostrado ser tradicionalmente los exámenes referidos a normas. Sin embargo, la investigación de Hodson (1986) sugiere que el potencial de la evaluación referida a criterios en ciencias no ha sido comprendido ni las pretensiones de sus efectos educativos y psicológicos han sido evaluadas por una investigación empírica objetiva. En cualquier caso algún paso inicial se ha dado. En el Reino Unido las encuestas del Assessment of Performance Unit (APU) han recogido amplios datos sobre las habilidades, conocimientos y comprensión de las ciencias que posee un gran número de niños representativos (normas) de varias edades. Además, existe la tendencia hacia un nuevo sistema de exámenes (el General Certificate of Secondary Education, GCSE) en el que la referencia a criterios desempeña un gran papel. Es comprensible por tanto, que muchos examinadores de disciplinas científicas, al buscar criterios razonables para alumnos de distintos niveles de competencia, hayan recurrido a las encuestas del APU como

guía, resultando así, un poco irónicamente, que los criterios de estos nuevos exámenes tienen su origen en estudios a gran escala de las normas. Además, técnicas referidas a criterios para evaluar las habilidades prácticas en ciencias básicas han sido publicadas (TAPS, 1983; Bryce, 1983), pero la encuesta de Hodson sugiere que la formación continua es esencial si las prácticas de evaluación de los profesores se han de acercar a las necesidades ideales. Quizá resulte algo sorprendente que el potencial de diagnóstico de la actividad referida a criterios haya recibido tanto énfasis (Black y Broadfoot, 1983) puesto que identificar las dificultades de aprendizaje con que se encuentran los estudiantes es sólo el primer paso en un diagnóstico, de modo similar a la información sobre los síntomas que recoge un médico. Igual que un síntoma determinado puede asociarse con una variedad de enfermedades, existe la posbilidad de que una actuación determinada que se considere inservible para alcanzar un criterio pueda ser el resultado de más de un simple error conceptual o una habilidad deficiente por parte del estudiante. Todo lo que muchos elementos de un examen referido a criterios pueden hacer en la actualidad es suministrar información sobre si se puede afirmar o no que un alumno ha alcanzado el criterio. Esto puede ser de algún valor para describir la actuación del que aprende, pero sirve de poco para identificar la razón fundamental de esa actuación. Tomemos un ejemplo muy sencillo. Supongamos que se está examinando a un alumno para conocer su habilidad para medir, dentro de un determinado margen de error, la masa de un número de objetos. Si no se alcanza el criterio se puede atribuir a la comprensión incompleta por parte del alumno de las reglas y principios para usar una balanza de platillos, su «inexactitud» al leer la aguja, un fallo en la comprensión de la relación entre masa y peso o incluso una mala interpretación de las masas a su alcance. Evidentemente saber que un niño «comete un error» es importante, pero la actuación correctora a realizar depende de la «causa» (el diagnóstico) de la dificultad con que se enfrenta y no del error cometido.

De acuerdo con Popham (1978) y Horne (1987) se diría que cada uno de los «tipos» de error antes mencionados representa diferentes, aunque interrelacionados, dominios del aprendizaje. Aunque no existe acuerdo entre los psicólogos sobre lo que constituye un dominio del aprendizaje, cada uno de los casos anteriores es característicamente diferente y parece que se puede ubicar en una progresión de complejidad que vaya desde realizar discriminaciones simples y llevar a término procedimientos hasta los principios y los conceptos tal como han sido descritos, por ejemplo, por Gagné (1977). Aunque la distribución de tareas relacionadas de complejidad variada dentro de jerarquías de aprendizaje ha tenido cierta popularidad entre los educadores de las ciencias, comparativamente ha desempeñado poco papel en el diseño de elementos de exámenes. Sin embargo, si se adopta el principio de que un error determinado en la actuación de un estudiante se puede relacionar con un número de categorías de aprendizaje cualitativamente diferentes, cada una de las cuales requiere una manera de enseñanza correctora, entonces la posibilidad de construir exámenes que sean verdaderamente de diagnóstico puede llevarse a cabo en las materias de ciencias. Horne (1987) ha sugerido que los exámenes de diagnóstico se podrían desarrollar a través de la concentración en los errores del alumno que indiquen aprendizaje imperfecto de conceptos o aplicación incorrecta de reglas, y que tales exámenes deberían calificarse como «exámenes referidos a conceptos» para distinguirlos de los otros exámenes referidos a criterios. Para que los profesores construyan estos exámenes, sin embargo, será necesario que estén familiarizados con la teoría del aprendizaje. En ausencia de esa teoría sugiere que la aproximación inicial a la evaluación de diagnóstico sea a través de los exámenes referidos a criterios.

A fin de investigar algunos de los procesos y consecuencias del cambio del sistema de exámenes referido a normas, al referido a criterios en ciencias, un estudio de un año de duración se realizó entre alumnos de primer año que habían elegido ciencia combinada en una «comprehensive school» de las afueras de Bristol. Como en muchas de tales escuelas, las ciencias ocupan el 15 por ciento del tiempo y se organizan en torno a temas principales. Durante las clases se enfatiza la enseñanza de seis habilidades principales: observacionales, de recogida de datos, de medida, manipulativas, de procedimiento y de seguimiento de instrucciones. Cada clase de ciencias con alumnos de distinto nivel es impartida por dos profesores que toman nota del trabajo fuera del aula, el trabajo en clase y las notas de los exámenes para el posterior comentario del resto de profesores y la preparación de los informes escritos formales para los padres. El examen del contenido del trabajo evaluado llevó a los nueve profesores del equipo a identificar los exámenes escritos, consistentes sobre todo en elementos de conocimiento con la finalidad general de hacer distinciones entre los alumnos (referidos a normas), como lo dominante en su práctica, quedando muy lejos de lo que idealmente se requería. Puesto que se piensa que la innovación en las prácticas de evaluación es más probable que surja de abajo arriba (es decir, cuando los profesores sienten la necesidad de mejorar su práctica habitual) que desde arriba abajo (es decir, cuando se les indica que lo hagan por ajustarse a una tendencia o por presión autoritaria), el departamento, bajo la orientación del segundo autor, identificó 12 discrepancias fundamentales entre las prácticas corrientes de evaluación y lo que idealmente se requería si había que llevar a cabo los objetivos para el curso del primer año. La mayor prioridad se dió a la necesidad de diagnosticar las necesidades de aprendizaje de los alumnos y a las futuras experiencias del currículo, seguidas, en orden descendente, por: la valoración de hasta qué punto los objetivos se lograban;

la valoración de la enseñanza y los materiales del curso; la negociación de una evaluación justa con los alumnos; la orientación a los alumnos para la elección de futuros cursos y carreras; la ubicación de los alumnos en los grupos o equipos de aprendizaje apropiados; y los informes a los padres. Las prioridades de los métodos de evaluación existentes diferían considerablemente de este orden.

Se organizó una serie de reuniones de trabajo para examinar modos de producir un conjunto de técnicas de evaluación comunes, así como la instrumentación para llevar a cabo, dentro de lo posible, las prioridades antes establecidas. Dado que el conocimiento por los profesores de las teorías psicológicas del aprendizaje era presumiblemente limitado, se decidió producir una serie de exámenes referidos a criterios, a fin de evaluar las habilidades fundamentales y no intentar confeccionar exámenes referidos a conceptos que serán necesarios para evaluar la comprensión en áreas más basadas en conocimientos del currículo. Hubo gran dificultad para establecer criterios objetivos de evaluación y la descripción completa de la puntuación de los criterios, siendo necesario detallar los procedimientos de puntuación de la mayor parte de los exámenes. Como los profesores reconocieron que un simple elemento para cada criterio en el examen no proporcionaría el grado de fiabilidad necesario, se diseñaron también elementos paralelos. Dentro de lo posible los exámenes se realizaron durante las actividades normales de enseñanza, pero muchos requerían la creación de situaciones especiales de laboratorio, por lo tanto los objetivos de las nuevas estrategias de evaluación se explicaron detalladamente a los técnicos de laboratorio. Como las nuevas prácticas de evaluación requerían cambios del estilo y énfasis de la enseñanza, se concertaron reuniones periódicas de profesores para comentar y asesorar los cambios de actuación.

La valoración del sistema fue llevada a cabo principalmente por los comentarios de los profesores, las observaciones del innovador (el segundo autor) y por las comparaciones entre las respuestas de un grupo de 50 alumnos seleccionados al azar que seguían el sistema «antiguo» de evaluación y un grupo similar que seguía el «nuevo» sistema. Los datos para el análisis fueron proporcionados por los comentarios espontáneos de los participantes y por los cuestionarios de profesores y alumnos.

1. RESULTADOS DE LA ENCUESTA DE LOS ALUMNOS

a) *Frecuencia de los exámenes.* Hubo diferencias significativas ($P < .01$; prueba de la significación de las diferencias entre proporciones independientes) entre las muestras. Los alumnos del nuevo método señalaron ser examinados mucho más frecuentemente que los del antiguo. Mientras el 100% de los alumnos del antiguo sistema afirmaron ser examinados «unas dos veces al trimestre», menos del 50% de los del nuevo sistema indicaron lo mismo. El 54% de estos últimos observaron que eran examinados «cada dos semanas o dos veces al mes aproximadamente». Evidentemente los alumnos reconocieron que la evaluación representaba un papel más importante en el currículo con el nuevo sistema.

b) *Comprensión de las finalidades de la evaluación.* Se notó una clara diferencia en la apreciación de la función de la evaluación por parte de los alumnos. Muchos más alumnos del nuevo sistema (46%) que del antiguo (12%) señalaron que la evaluación tenía lugar «para que el profesor pueda hacerte comentarios sobre tus progresos». Del mismo modo, mientras el 96% de los alumnos del nuevo sistema comprendieron que la evaluación ayudaba a los profesores «a saber cómo ayudarte mejor», sólo el 33% de los del antiguo sistema hicieron esta observación. No se advirtió ninguna diferencia, sin embargo, en el uso de la evaluación «para decir a los padres cómo progresas» ni en su uso como medio para distribuir a los alumnos en el grupo de ciencias del año siguiente.

c) *Nerviosismo y ansiedad al ser examinado.* La mayoría de los alumnos (70%) en ambas muestras indicaron sentirse nerviosos ante los exámenes, por lo menos en algunas ocasiones. Sin embargo una proporción significativamente mayor (30%) de los alumnos del nuevo sistema que del sistema antiguo (4%) señalaron sentirse «normalmente relajados». De acuerdo con esto, mientras el 34% de los alumnos del sistema antiguo observaron estar «normalmente nerviosos», sólo el 10% del nuevo sistema hicieron comentarios similares ($P < .01$ en ambos casos).

d) *Interés y progreso en las ciencias.* En general el interés en las ciencias era alto, pero se descubrieron algunas diferencias significativas. El 94% de los alumnos del nuevo sistema hablaron de las ciencias como «interesantes» o «muy interesantes», comparados con el 76% del antiguo sistema ($P < .01$) y sólo el 6% (nuevo sistema) las señalaron como «no muy interesantes» en comparación con el 24% (antiguo sistema) que las consideraban igual o «aburridas» ($P < .05$). Por lo que respecta a la percepción de los alumnos de sus propios progresos en ciencias, no se descubrieron diferencias significativas entre las dos muestras.

e) *Identificación de las dificultades de los alumnos.* Surgieron claras diferencias entre las dos muestras. El 64% de los alumnos del nuevo sistema consideraron que sus profesores sabían qué partes del trabajo encontraban más difíciles, mientras que ningún alumno del antiguo sistema hizo esta observación. Sólo el 4% de los del nuevo sistema opinaron que su profesor «no sabe realmente qué partes me resultaron difíciles», apreciación que realizó el 42% de los del antiguo sistema de evaluación.

f) *Frecuencia con que los alumnos solicitaron ayuda para las dificultades.* Se realizaron preguntas sobre la

frecuencia con que los alumnos solicitaban ayuda, pero no se observaron diferencias significativas. Mientras el 94% del nuevo sistema señaló que «siempre» o «normalmente» pedían ayuda, el 84% del sistema antiguo hizo una observación similar.

g) *Cantidad de ayuda recibida.* No se observó diferencia significativa entre los alumnos de los dos grupos sobre la cantidad de ayuda recibida al encontrar dificultades. Sin embargo menos alumnos del grupo del nuevo sistema tenían la sensación de que los profesores no proporcionaban suficiente ayuda cuando se les pedía.

Si se pudiera demostrar que los alumnos aprendieron de un modo más efectivo siguiendo el nuevo sistema, se podría argumentar que la superioridad —al menos a corto plazo— de la evaluación referida a criterios sobre la referida a normas está demostrada. Sin embargo, puesto que los objetivos de la enseñanza bajo los dos sistemas no son estrictamente comparables, no se podría aplicar ninguna prueba significativa para verificar esta hipótesis. Naturalmente se debe usar la prudencia al sacar la conclusión de que las diferencias observadas entre los grupos es directamente atribuible al cambio en las prácticas de evaluación. No obstante, los grupos eran comparables, los profesores eran el mismo personal, el tamaño de las clases era equivalente y el tiempo de exposición de los alumnos a los sistemas de evaluación fue de 36 semanas en cada caso. No se controlaron los efectos de la novedad: es posible que la comparación fuera esencialmente entre los resultados de profesores trabajando con estusiasmo sobre un nuevo sistema pero desilusionados con el antiguo. Evidentemente, antes de que se pueda afirmar la «superioridad» del sistema referido a criterios, son necesarios estudios sobre períodos más largos de tiempo, en diferentes marcos y con alumnos de diferentes edades.

2. RESULTADOS DE LA ENCUESTA A LOS PROFESORES

Los resultados de la encuesta a los profesores deben interpretarse con cautela ya que se les pide que evaluen prácticas que han sido introducidas para superar aquéllas de las que estaban insatisfechos en gran medida. La posibilidad de que se inclinen a favor de una innovación en la que ellos mismos han invertido gran esfuerzo no se puede descartar. Los datos se obtuvieron de cuestionarios y comentarios libres durante reuniones y el análisis se concentró en cuatro áreas principales:

a) *Desventajas y problemas al llevar a cabo el nuevo sistema.* Los profesores que se animaron a hacer pruebas mientras los alumnos estaban ocupados en actividades normales de aprendizaje experimentaron problemas organizativos. Se mencionó el aumento del tiempo dedicado a examinar y algunos profesores comentaron que su conocimiento de las dificultades con que se enfrentaban ciertos alumnos implicaba un tratamiento reducido de parte del material. Sin embargo aquellos capaces de alcanzar resultados superiores no fueron perjudicados.

b) *Ventajas del nuevo sistema.* Los profesores comentaron que comprendían más claramente los objetivos de su enseñanza y que eran más conscientes de la naturaleza de las dificultades de aprendizaje con que se enfrentaban los alumnos. Dos profesores creyeron también que el sistema les animó a prestar más atención a los puntos fuertes y débiles de sus estilos de enseñanza y les estimuló a ser más organizados al enseñar y evaluar.

c) *Cambios necesarios.* Seis de los ocho profesores creyeron que el sistema debería seguir sin cambios un año más. Los dos disidentes estaban a favor de revisar algunos de los materiales de examen pero no especificaron los cambios que se deberían hacer.

d) *Influencias de los nuevos procedimientos de examen sobre el conocimiento de las necesidades de aprendizaje de los alumnos.* Los profesores hicieron comentarios favorables sobre el aumento de sus conocimientos de problemas básicos de aprendizaje de los alumnos hasta entonces no reconocidos: se señaló que esto era particularmente cierto para muchos de los alumnos que obtenían los resultados más pobres y que, no obstante, consiguieron esconder sus deficiencias en los procedimientos normales de examen. La mayor parte de los profesores razonaron que esta información sobre las dificultades de los alumnos era útil para hacer su enseñanza más apropiada. Teniendo en cuenta los efectos de Hawthorne, los posibles desequilibrios de las cuestiones planteadas por los que estaban comprometidos con el éxito del nuevo sistema y el período de prueba comparativamente corto de un solo año académico, es prematuro concluir que la nueva práctica de evaluación ha mejorado la calidad de la enseñanza de las ciencias. Sin embargo este problema es una consecuencia de la aproximación del profesor-como-evaluador y la necesidad de algún tipo de «triangulación» (Cohen y Manion, 1985, donde se emplea un evaluador externo y objetivo) se ve subrayada por conclusiones de este tipo.

CONCLUSIONES

Evidentemente es peligroso generalizar a partir de estudios a pequeña escala de las innovaciones en la práctica de exámenes en escuelas aisladas. Sin embargo se podría argumentar que las invitaciones a cambiar hacia sistemas referidos a criterios en el Reino Unido han superado en cierto modo al acceso a datos empíricos que respalden sus beneficios, y que los profesores deberían animarse cada vez más a examinar sus prácticas y a cambiarlas y valorarlas a pesar de los muchos cambios a los que se ven forzados por los intentos cada vez mayores de un control gubernamental del currículo. Las dificultades que implican los cambios des-

critos en este artículo no deberían ser minimizadas. El establecer objetivos es en sí mismo problemático; podría haber al menos 150 objetivos de enseñanza de las ciencias en el primer año y, si cada uno ha de ser evaluado objetivamente de varias maneras, un gran número de objetos a evaluar se ve implicado. En la actualidad en el Reino Unido la preocupación nacional por estos temas ocupa al menos a tres organismos (el Departamento de Educación y Ciencias, la Association of Science Education y el Secondary Examinations Council), pero seguirá siendo responsabilidad de los profesores el convertir en temas de interés nacional las cuestiones locales y prácticas dentro de cualquier orientación o presión acerca de los exámenes que pueda iniciarse «desde arriba». Existe una sutil línea de división entre el desarrollar los instrumentos de evaluación para apoyar el currículo y el permitir que esos instrumentos lo controlen, pero es más probable que ocurra lo primero cuando los profesores detectan la necesidad de nuevos procedimientos de examen que cuando esos nuevos métodos de examen proceden del exterior.

Se evidencia, a partir del estudio, que los resultados de los exámenes referidos a criterios se ven de mayor utilidad que los referidos a normas. Durante muchos años, al menos en el caso de las escuelas, los resultados de los exámenes se habían utilizado únicamente para escribir informes y «ubicar» a los alumnos. Esto puede deberse a que al tipo de información obtenida no se le ha dado otro uso, pero el «potencial de diagnóstico» de los exámenes referidos a criterios es como se ha señalado, limitado, a menos que el diagnóstico sea completo cuando los síntomas de los fracasos de aprendizaje hayan sido identificados. Un comienzo prometedor ha tenido lugar, sin embargo, ya que es fuerte la evidencia de que los nuevos procedimientos de evaluación son vistos tanto por profesores como por alumnos como de apoyo para el nuevo aprendizaje, incluso aunque no puedan por sí mismos proporcionar mucha información sobre las causas del fracaso en el aprendizaje.

La naturaleza continua del nuevo programa de evaluación permite una comunicación más regular entre profesores y alumnos sobre sus progresos en ciencias. Al ser expuestos los resultados de los exámenes como un perfil de los objetivos dominados, una imagen más clara del progreso de un alumno se puede ofrecer también a los padres. La respuesta de los padres a una reunión en la que se comunicaron los resultados de esta manera resultó ser enteramente positiva.

Unas cuantas limitaciones del estudio se ha señalado en lo anteriormente expuesto. El potencial de diagnóstico de los exámenes referidos a criterios no sólo ha sido exagerado sino que los efectos a largo plazo de los cambios sobre la actuación en ciencias de los alumnos no han sido investigados. Si éstos y otras actitudes positivas permanentes no se demuestran en estudios futuros, muchas de las ventajas de los exámenes referidos a criterios sólo serán meros artículos de fe en la práctica educativa.

Existen además problemas para valorar la fiabilidad de los exámenes referidos a criterios. En este estudio se han definido cuidadosamente los criterios de modo que las idiosincrasias en la apreciación del profesor se descartarán. Sin embargo sólo en las habilidades más simples examinadas se obtuvo completa objetividad. Se sabe que los exámenes que intentan valorar las adquisiciones de orden superior (como por ejemplo «la comprensión») son difíciles de formular en términos de criterios observables no sólo porque la comprensión y cuestiones similares se pueden expresar en diferentes grados y de muchas maneras (Mesick, 1984). Incluso si se puede asumir la fiabilidad, la validez no se puede dar por garantizada. Generalmente, como ha mostrado Nuttall (1987), el problema para todos los examinadores es decidir hasta qué punto los elementos del examen muestran los contextos en los que se pueden demostrar las habilidades y el conocimiento.

Aunque la idea de los exámenes referidos a conceptos descrita antes parece potencialmente de mayor ayuda para el diagnóstico, los profesores de este estudio se mostraron satisfechos de que los exámenes que habían diseñado permitían al menos una identificación más clara de los alumnos que necesitaban ayuda de recuperación, aunque no la naturaleza de la ayuda requerida.

A medida que los psicólogos han profundizado su comprensión de la actuación en los exámenes, el número de factores identificados como influyentes en la puntuación ha aumentado constantemente. También lo han hecho los efectos de los exámenes sobre los alumnos. Aunque no es parte de esta investigación examinar los efectos sobre la actitud de los alumnos hacia las ciencias o su autovaloración, los resultados preliminares favorecen que el interés de los alumnos en las ciencias parezca más positivo entre los del nuevo sistema que los del antiguo. No obstante, se mantiene la necesidad de exámenes que informen un mejor diagnóstico de las dificultades del alumno y al mismo tiempo estimulen a los más capaces.

REFERENCIAS BIBLIOGRÁFICAS

BLACK, H.D. y DOCKRELL, W.B., 1980, *Diagnostic assessment in secondary schools.* (Scottish Council for Research in Education: Edinburgh).

BLACK, H.D. y BROADFOOT, P., 1982, *Keeping track of teaching: assessment in the modern classroom.* (Routledge and Kegan Paul: London).

BRYCE, T., 1983, The diagnostic assesment of practical skills in foundation Science. *Scottish Education Review,* 15, 41-51.

CAVERNI, J.P., 1987, Knowledge acquisition assessment by experts: effects and models of the cognitive functioning of evaluators. *European Journal of Psychology of Education.* 2, 119-131.

COHEN, L. y MANION, L., 1985, *Research methods in education* 2nd Edn. (Croom Helm: London).

GAGNÉ, R.M., 1977, *Conditions of learning,* 3rd. edn. (Holt, Rinehart and Winston: New York).

HODSON, D., 1986, The role of assessment in the «curriculum cycle»: a survey of science department practice. *Research in Science and Technological Education.* 4, 7-17.

HORNE, S., 1978, Concept referenced testing. *European Journal of Psychology of Education.* 2, 143-156.

MESSICK, S., 1984, The psychology of educational measurement. *Research Report.* (ETS: New Jersey).

NUTTALL, D., 1987, The validity of assessments. *European Journal of Psychology of Education,* 2, 109-118.

POPHAM, W., 1978, *Criterion-referenced measurement.* (Prentice Hall: Englewood Cliffs, N.J.).

SUTTON, R., 1986 (Ed.), *Assessment in secondary schools: the Manchester experience.* (Longman/Schools Council: Harlow).

TAPS, 1983, *Techniques for the assessment of practical skills in foundation science.* (Heinemann: London).