EVALUACIÓN Y CONTROL DE LOS CONOCIMIENTOS

Paul Black, King's College, Londres, Gran Bretaña

Introducción

Las metas

En la enseñanza, la evaluación tiene tres funciones principales. La primera consiste en orientarse en los conocimientos de los alumnos individualmente en el objetivo de la certificación. La segunda es orientarse en las adquisiciones de los grupos, de clases o escuelas para los objetivos políticos más generales. La tercera es de servir en la enseñanza y el aprendizaje.

La primera función permite establecer los procesos verbales jugando el rol de pasaportes – para mejores empleos, o para ir a la enseñanza superior cuando un alumno abandona la escuela. Para cumplir con esta función, la evaluación debe suscitar la confianza del público. Una tal evaluación tiene también el objetivo de una vista del conjunto del trabajo del alumno, eso hace que pueda ser calificado de sumativo.

La segunda función está caracterizada por la importancia de rendir cuenta al público a su vez en escuelas tomadas individualmente y del sistema educativo a nivel de un estado o de una región. Aquí el objetivo es informar para establecer una política, por la colección y el análisis de los resultados. Diferentes sistemas de controles regionales y nacionales así como los estudios comparativos internacionales rellenar este objetivo.

La tercera función proviene de lo que todo sistema de aprendizaje necesita de las informaciones de vuelta. Para esperar este objetivo, los resultados deben proveer de las informaciones sobre el aprendizaje de cada alumno en base a las cuales uno puede poner en ejecución las acciones respondiendo a las necesidades de aprendizaje de cada alumno. Una evaluación tal puede ser llamada formativa o diagnóstica.

De forma ideal, cada una de esas tres funciones demanda informaciones de tipo diferente. En la práctica, es frecuentemente necesario utilizar la misma información para llenar funciones diferentes: Un uso tal múltiple es seductor porque es económico, pero no siempre es realizable y existe siempre una tensión entre las necesidades de esas diversas funciones.

Contexto político y social

Las diferencias de historia, de tradiciones, de necesidades políticas y sociales se combinaron para producir los sistemas en los cuales la comparación entre países hace aparecer los esquemas de prácticas muy diferentes (Black 1992, Britton y Raizen 1996). Por ejemplo, en ciertos países, las certificaciones de fin de curso están enteramente bajo su responsabilidad de cada escuela, entonces los otros poseen una cierta confianza en los

profesores es impensable y las certificaciones están fundamentadas enteramente en tests externos.

Tales diferencias tienen sus orígenes históricos. Los movimientos sociales teniendo por objetivo producir una sociedad igualitaria encontró en el pasado los medios de compensar las ventajas de las escuelas privilegiadas utilizando los procesos de evaluación. En tal empresa, la naturaleza externa del sistema de evaluación, y su ausencia de vinculación orgánica con el sistema de enseñanza en las escuelas tuvo elementos esenciales. Inversamente, en las sociedades donde la igualdad de financiamiento y el status de las escuelas sería real, ello sería posible de no utilizar sistemas externos y objetivos y contar en cada escuela para realizar la evaluación.. Toda vez, un sistema de educación pública bien desarrollado y universal se asevera ser un utilizador importante de los fondos públicos. Eso explica que la justificación de su presupuesto se transforma en un problema político importante. Esta presión condujo a exigencias de cuidado y de responsabilidad.

Tales características afectan la manera como los exámenes externos imponen la elaboración de enseñanza y del aprendizaje. Las quejas pretendiendo que ellos se extenúan y deforman el aprendizaje, uno debe oponerlas al hecho que, desarrollados y aplicados de forma adecuada. Ellos pueden ser agentes de pujantes reformas.

El puesto de la física en las evaluaciones para los certificados donde los riesgos están vinculados al ingreso a la universidad, es igualmente muy variable. En los procedimientos de las universidades en la mayor parte de los países, la transformación en muchos materias, preferiblemente para la física sola, se toma en cuenta en la admisión de los cursos especializados de física. En el caso donde la prueba de ingreso a la universidad toma en cuenta una ampliag gama de asignaturas, puede tener reglas de reagrupamiento que conducen a que un alumno sea admitido para estudiar física igual si las transformaciones en física han sido débiles – como en Brasil. Desde la prueba de ingreso que se basa solamente en un pequeño número de materias – como en Gran Bretaña – es posible dar más peso a los resultados obtenidos de física, en parte gracias a la posibilidad de un ejercicio profundo, lo cual no sería realizable si los candidatos tuviesen más asignaturas.

Estructura de este capítulo

Este capítulo tratará sucesivamente de esas tres funciones, es decir la certificación, la responsabilidad frente al público, la ayuda en el aprendizaje. Inevitablemente, por vinculación a la primera, numerosas cuestiones generales se implican en un conjunto de tres deberán ser exploradas. Una sección final pasará revista a las interacciones entre esas funciones para tratar de hacer un panorama de los sistemas en los diferentes estados o naciones.

Evaluación para la certificación

Métodos – tests escritos

Los tests papel – creyón siempre han sido utilizados como medio principal para la evaluación en la enseñanza. Los tests escritos convencionales en física conforman una lista de preguntas exigiendo las definiciones, las explicaciones estándares o los cálculos, los informes de las prácticas o de las aplicaciones, y algunos problemas habituales. El ejemplo que sigue ilustra combinación:

- *Una muestra del elemento talio 207/81 Tl de 16 mg emiten radiaciones beta y es transformado en un isótopo del plomo (Pb). La duración media de vida del proceso de desintegración es de 5 minutes. Responda a las siguientes preguntas:*
- *¿Qué significan los términos : isótopos – isóbaros – isótonos?*
  Dé un ejemplo de cada uno.
- *Explique lo que quiere decir "una duración media de vida de 5 minutos para una tal desintegración?*
- *Encuentre la masa del elemento talio 207/81T1 residual después de 1/3 de hora.*
- *Calcule la constante de desintegración del elemento talio 207/81 T1*
- *Dar (sin explicar) dos métodos de stockage y de tratamiento de los desechos radioactivos.*
  (Egipto – Wassef pp. 57-68 en Black 1992)

Los tests pueden estar compuestos de preguntas más cortas, tratando separadamente los aspectos ligados a los conocimientos y a la resolución de problemas. El ejemplo siguiente es típico del estilo del problema corto:

> *Un resorte comprimido mantiene distancia de dos carretas teniendo respectivamente por masas 0,2 kg y 0,3 kg, de manera tal que las carretas que estaban inicialmente en reposo son separadas de 60 cm al extremo de 5 s. La masa del resorte los frotes son negligentes. ¿Cuál es la velocidad de esas carretas?*
> *(Hungría - Radnai pp. 84-100 en Black 1992)*

Tales preguntas en los problemas cortos necesariamente no son cuantitativas, como en este ejemplo:

> *¿Cómo un dipolo eléctrico se comporta en un campo eléctrico uniforme y en un campo eléctrico radial?*
> *(Polonia – Plazak y Mazur pp. 125-139 en Black 1992)*

Un número más importante de preguntas dentro del mismo tiempo y una fiabilidad más importante de la notación pueden ser obtenidas a partir de la utilización de tests de escogencia múltiple, que evalúan las capacidades para resolver problemas, todo como los conocimientos, como lo muestra el siguiente:

Los puntos O, P, Q, R, S y T son alineados verticalmente, en intervalos regulares.
Un objeto es soltado en caída libre de O con una velocidad inicial nula.

La velocidad media entre O y T es igual a la velocidad instantánea que se sitúa en el intervalo siguiente:
a) entre O y P *b) entre P y Q* *c) entre Q y R*
d) entre R y S *e) entre S y T*
(Japón – Ryu pp. 101.123 en Black 1992)

De hecho del número importante de preguntas en las cuales los alumnos deben ensayar de atacarse en un tiempo determinado, de tales cuestiones pueden llegar a cubrir un campo amplio y ser más fiables. Su inconveniente es que ellos no muestran las razones de la escogencia de los alumnos y que los estudios han mostrado que cerca de un tercio de los alumnos que seleccionaron la respuesta correcta lo hicieron por razones erróneas (Tamir 1990). La utilización exclusiva de esos tests puede conducir en negligencia, en la enseñanza, la discusión y la argumentación que son dos aspectos preciosos y válidos de la investigación científica. En las pruebas nacionales, ellos han sido durante mucho tiempo preponderadamente en las evaluaciones en USA, ellos fueron componentes minoritarios en ciertos países (por ejemplo en el reino Unido, en Suecia) y no son del todo utilizados por otros (por ejemplo en Francia).

Todos los ejemplos previamente citados ponen en juego una combinación de conocimientos de física y de capacidades para seleccionar y aplicar esos conocimientos. Desde que las preguntas largas tienen una estructura – como en el primer ejemplo, esta estructura ayuda a aquellos que responden, a su vez presentando la selección del saber necesario para el problema, y guiado frente a la estrategia que permite abordarlo.

No es necesario limitarse a los conocimientos memorizados. Ciertos tests escritos presentan al alumno informaciones bajo la forma de un pequeño artículo sobre un tema de física, y evaluando enseguida las capacidades para comprender y aplicar haciendo preguntas sobre ese texto – para las preguntas abiertas cortas o las preguntas de selección múltiple (Ver Black 1992).

Evaluación de los "saber-hacer procedimentales"

Hubo tentativas para evaluar los "saber – hacer" separadamente del contenido físico asociado. Los problemas que aquello conlleva pueden ser ilustrados por las tareas prácticas teniendo por objeto evaluar los saber - hacer asociados a la medida. Uno de los problemas es relativo al contexto - de numerosos alumnos que parecen ser capaces de utilizar los instrumentos de medida y los procedimientos desde que ellos están directamente solicitados en los contextos artificiales y aislados, no se sirven de esa misma incapacidad desde que él les ha solicitado de tomar a cargo una investigación, igual desde que los instrumentos de medida le han sido provistos; ellos utilizan solamente una comparación cualitativa, igual desde que ellos han mostrado, en un contexto diferente, su capacidad de utilizar los instrumentos. Es esencial tener una visión más completa de los saber - hacer en juego. La capacidad de servir de manera precisa de una escala y de disponer y ordenar los instrumentos no es suficiente. Un científico debe ser preciso sobre aquello que él desea medir – por ejemplo, medir un "débito" necesita comprender que es necesario tomar las parejas de medidas vinculadas entre ellas. Es más, el investigador debe juzgar cuando

medir, esto requiere un juicio sobre el hecho que el impulso útil de cuantificación puede y debe ser aplicado al problema (Black 1990).

Así, una evaluación de los saber – hacer relativos a los instrumentos y a las escalas no tiene mucho valor en si, del hecho que tales saber – hacer no se revelan útiles que a la luz de un modelo conceptual del sistema estudiado y de las variables en juego. Eso es lo mismo para los "saber - hacer" – por ejemplo la observación que no es una recepción pasiva pero que es esencialmente una actividad selectiva guiada por las hipótesis sobre aquello que debe ser seleccionado. Resulta que las preguntas que miden el "saber - hacer" específicos de manera aislada no pueden aportar informaciones útiles sobre la capacidad de un alumno para utilizar los saber - hacer de un trabajo científico.

Otros métodos - orales y prácticos

Los tests orales son los componentes importantes de las pruebas en muchos países europeos del Este, pero son raramente utilizados fuera. En la evaluación de los trabajos prácticos, la logística y los otros problemas de utilización de los equipamientos materiales para los ejercicios de evaluación han conducido a tratar de utilizar reemplazando a los tests de papel y creyón.

Tal aproximación puede tener un retorno de los efectos indeseables sobre la enseñanza y las correlaciones entre las transformaciones de los alumnos con el material real y su respuesta con su "equivalente" en los tests escritos son débiles. Toda vez, muy pocos países, notablemente el Reino Unido e Israel, emplean los tests con el material a título de procedimiento de los exámenes (Britton y Raizen 1996).

Una forma habitual de test práctico fue de especificar un conjunto de procedimientos con los equipamientos dados y de solicitar ciertos tipos de respuestas en términos de medidas efectuadas y de análisis de los resultados. Las obligaciones precisas impuestas entonces mejoran la fiabilidad, pero de los saber - hacer tales como la concepción de las experiencias y de las escogencias de los aparatos son negligentes. La "experiencia" es de hecho un medio para verificar ciertos saber - hacer específico teniendo vinculación en la ciencia experimental. Un punto de vista alternativo consistió en organizar ejercicios muy cortos, cada uno probando un saber – hacer específico en un contexto dado. Los saber - hacer evaluados pueden ser por ejemplo tomar medidas precisas con los instrumentos pre reglados, hacer tomar en cuenta observaciones cualitativas sobre un fenómeno inhabitual.

Los estudios relativos a los tests fuera de contexto que pueden ser mal interpretados han conducido a emprender la evaluación a partir de las tareas experimentales abiertas y completas. Si la comparación y el respeto de las condiciones de pruebas convencionales son requisadas, entonces lo que es posible hacer es muy limitado. Una contraposición suplementaria en tales condiciones es que todos los miembros de un grupo deben trabajar simultáneamente en un tiempo limitado, la evaluación debe fundamentarse en las reportes escritos de los alumnos. Sus acciones efectivas y sus razones de reaccionar no pueden ser observadas o interrogadas directamente. Igualmente, del hecho que algún sistema no puede manipular más de un pequeñísimo número (probablemente uno o dos) de tareas completas, la generalización de los resultados es un problema serio (Shavelson y otros. 1993).

Tan importantes limitaciones pueden ser sobrepasadas por la evaluación de los trabajos llevados sobre una larga duración e interpuesto en las condiciones normales de la enseñanza. Una vinculación escrita de tales trabajos puede esconder algunos de los aspectos importantes de las capacidades porque un alumno puede hacer pruebas, aspectos que pueden solamente ser estimados si los procedimientos con los cuales los alumnos han trabajado han sido observados y comprendidos. La sola solución posible frente a tales problemas es, para los profesores, de implicarse como evaluador, eso permite tener acceso a las características tales como la forma precisa donde el problema ha sido formulado, los efectos de la colaboración en el grupo, las contraposiciones con las cuales los alumnos debieron trabajar y las razones de las decisiones de los alumnos concerniendo la escogencia de una estrategia. Tal atención requiere una formación profunda de los profesores así como del tiempo y la posibilidad de poder observar atentamente el trabajo individual o el trabajo de grupo.

Bien entendido, aquí existe una paradoja. Como los objetivos de la enseñanza de la física se convierten en más reales y menos artificiales, ellos conducen a las actividades que reflejan más la complejidad y el desorden de la vida real, y esas últimas se convierten así en menos susceptibles de hacer el objeto de algún proceso reproducible de evaluación.

Tales dificultades no se aplican solamente a la evaluación de los trabajos prácticos. Algunas de las numerosas tareas escritas que fueron utilizadas están seriamente limitadas por las obligaciones de tiempos de los tests externos usuales. Si vamos al final de la discusión de las ideas y su explotación de manera de abrir las nuevas preguntas, se llega a los tests conducentes a la producción de textos en un tiempo suficiente para que el alumno pueda reflexionar y estructurar sus ideas.

Combinar los métodos

Todos los métodos de evaluación tienen sus defectos, y la mejor selección frecuentemente depende del contexto educativo o social particular. Por ejemplo, la disposición del material o la posibilidad de una vigilancia apropiada para evitar las prácticas deshonestas pueden variar enormemente de una situación a otra. Es igualmente necesario considerar que una combinación de métodos pueden ser necesaria, a la vez para reforzar la fiabilidad y para compensar lo sesgado que toda técnica introduce necesariamente.

La panoplia de los métodos empleados es muy variada en ciertos países, es muy restringida en los otros. De un lado, se encuentra la utilización exclusiva de las preguntas de selección múltiple, como en USA donde se emplean las agencias especialistas en evaluación. Por otro lado se consigue un abanico de ocho tipo de tareas para la prueba Nuffield en el Reino Unido (Black, 1992). El conjunto más largamente aprobado comprende una mezcla de preguntas de selección múltiple y los problemas cortos.

Una de las razones de tal variedad resulta del hecho que la familiaridad con ciertos métodos y la tradición de su utilización puede inhibir la posibilidad de buscar seriamente las soluciones alternativas. Por ejemplo, existe una tradición en ciertos países según la cual la cuestión teórica matemática es el test el más significativo de la competencia de un físico.

Así, esas preguntas se dan viendo la importancia más grande que, dijéramos, las competencias a abordar un problema experimentalmente o en escribir una crítica sobre un tema.

Otra razón puede ser la impresión que ciertos métodos no pueden dar resultados confiables: uno puede pensar que las preguntas de selección múltiple, con su notación objetiva, sus pre-tests de verificación y las posibilidades de análisis estadístico sobre un gran número de preguntas dan resultados que son bien fiables que los otros tipos de evaluación. Esto puede conducir a considerar el solo método a emplear.

La tercera razón puede ser el costo. Para las pruebas fuertes, la inversión para preparar buenas preguntas de escogencia múltiple está justificada por la posibilidad de obtener calificaciones confiables a un débil costo. La escogencia y la notación de muchos otros tipos de preguntas son costosas en tiempo y en experticia de los examinadores. En particular, la evaluación de los trabajos prácticos está frecuentemente excluido a causa de su costo y de su factibilidad.

Fiabilidad

La pregunta de saber si sí o no una gama de preguntas es realmente importante es además compleja. Los principales factores implicados son la fiabilidad, la validez, el feed-back, la calidad y lo sesgado. La fiabilidad es lo más simple a considerar. Un examen puede ser fiable si se está seguro que los mismos resultados fueron obtenidos a partir de un examen paralelo, es decir a partir de un conjunto de preguntas y su baremo con los mismos objetivos y los mismos métodos. Es posible obtener una medida de fiabilidad verificando la coherencia interna de las respuestas, pero eso no es posible desde que se dispone de un número razonablemente importante de preguntas y desde que uno pueda suponer que ellos deben dar resultados homogéneos. Un test más estricto es de dar los conjuntos de preguntas en paralelo a los mismos estudiantes o al menos de dar a algunos estudiantes un número tal de preguntas en el mismo campo y de determinar a partir de los resultados, el número mínimo de preguntas necesarias para reducir el error antes descrito de un cierto umbral. Es difícil de realizar de manera sistemática tales tests – la confianza de los jurados de la universidad y de la escuela en la fiabilidad de los tests externos cortos es generalmente injustificado de hecho que no está fundamentado en pruebas. La fiabilidad de la notación es igualmente una postura en los exámenes nacionales – una formación cuidadosa de los jurados y las dobles correcciones son esenciales.

Validez

Un concepto suficientemente general de validez es a la vez necesario y convincente en toda tentativa de mejoramiento de la calidad de las evaluaciones. La proposición introductiva de la revista realizada por Messick (1989) da una definición que hace autoridad:

> *La validez es un juicio evaluativo global del grado en el cual los resultados empíricos y las justificaciones teóricas confirman la adecuación y la justicia de*

> *las interferencias y de las acciones fundamentadas sobre las puntuaciones de los tests o de otros modos de evaluación.*

Así, si un test supuesto muestra las competencias de un alumno por la medida, ese es el objeto del juicio del experto de saber si el test llama a la utilización de las capacidades de las medidas que son importantes en ciencias. Si un test fue concebido para ayudar a prevenir las aptitudes en las futuras adquisiciones, eso puede ser estimado a destiempo, estudiando la correlación de los resultados del test con aquello que el test debía predecir. Más de una actividad de evaluación está próxima de la actividad real para la cual los resultados del test son considerados como pertinentes, más el test tendrá tendencia de satisfacer los criterios de validez. En esta perspectiva, la evaluación en la clase tiene más oportunidad de éxito que en los tests escritos formales en tiempo limitado.

Efectos sobre el aprendizaje

Uno de los centros de interés relativos a la evaluación es su efecto sobre el aprendizaje. Las investigaciones han demostrado que preparar a los alumnos en los tests de selección múltiple puede ser desfavorable en una buena práctica del aprendizaje. Esto es explicado por Resnik y Resnik (1992) de la siguiente forma:

> *Los alumnos que practican la lectura principalmente en la manera como ella interviene en los tests – y existen pruebas que eso es lo que se produce en muchas clases – arriesgándose a ser poco confrontadas por las exigencias y a las posibilidades del razonamiento que está en el espíritu de un programa que gira en torno a la reflexión.*
>
> *Los alumnos que practican las matemáticas en la forma correspondiendo a aquella de los tests estandarizados se arriesgan a no ser nunca confrontadas al tipo del pensamiento matemático investigado por todos aquellos a quienes les concierne la reforma de la enseñanza de las matemáticas...*
>
> *Las evaluaciones deben ser concebidas porque, desde que ustedes hacen eso que es natural – es decir, preparar los alumnos para el éxito – aquellos ejercieron los tipos de competencias y desarrollaron los tipos de saber – hacer que son los verdaderos objetivos de la reforma de la enseñanza.*

Calidad

Los diferentes factores que afectan la calidad de los textos y de las evaluaciones, uno de aquellos que destacan más es el tiempo tomado por los tests externos. Si un país investido cerca de ocho horas en los tests en una materia para determinar las perspectivas de una carrera futura, como otro puede reemplazar la misma función en menos de dos horas? Todo test apunta a mostrar los diferentes campos de transformación que son importantes en la materia – el control de la A.P.U. en la edad de 13 años en 1984 necesitó 35 tests de una hora para obtener resultados fiables sobre los campos juzgados importantes para la ciencia (Johnson 1988).

Una táctica para sobrepasar esta dificultad es la de restringir la gama de campos evaluados, por tanto esto reduce de antemano los objetivos que el examen pretende. Por eso mismo, en razón de la postura que ellos representan en la mayor parte de los sistemas, ellos influyen sobre los objetivos de aprendizaje en la escuela.

Una cuestión próxima se reporta en la calidad centrada en el equilibrio entre las preguntas en las cuales es posible responder por los procedimientos de rutina utilizando los algoritmos, objetos de aprendizaje, y las preguntas que exigen una traducción reflexiva y una aplicación de principios y de procedimientos: En los sistemas de numerosos países, el análisis muestra que el equilibrio inclinado frecuentemente a favor de los primeros. Eso en parte ser debe a las limitaciones de tiempo para los tests. La tarea muy difícil para la mayoría de los evaluadores es de concebir las preguntas que exigen una reflexión más que un aprendizaje maquinal, a un nivel para el cual la media de los alumnos pueden tener una buena ocasión de éxito.

Rodeos

Todo procedimiento de evaluación es una interacción entre ciertas preguntas, ítemes y / o procedimientos, y los alumnos que son evaluados. Muchas razones pueden hacer que la interacción funcione de manera defectuosa, produciendo así los rodeos o las imperfecciones dentro de los resultados (Gipps y Murphy 1994).

Un ejemplo que ha sido bien estudiado es aquel entorno del sexo. Un problema teniendo una importancia comparable es aquella del entorno cultural o étnica. Las dificultades se sitúan en el lenguaje utilizado en la enseñanza de las materias, en la gama y la naturaleza de los ejemplos utilizados en la vida cotidiana y en la hipótesis cultural inherente a la ciencia occidental.

Existen igualmente numerosas formas de discriminar injustamente los alumnos en tanto que individuos, igual al seno de la misma categoría de sexo y de cultura, por la presentación, el contexto o el lenguaje de las evaluaciones. Se puede aseverar que los alumnos que producen respuestas extrañas y aparentemente desprovistas de calidad muestran gracia a las discusiones que se trata de respuestas sensatas ligadas a una mala interpretación de la pregunta (Gauld 1980).

Referencia a la norma y al criterio: reagrupamiento y perfil

La tradición dominante en todos los niveles de enseñanza ha sido utilizar los tests con una norma referenciada. Desde que ellos son utilizados, el énfasis porta sobre la comparación de un alumno cualquiera con el conjunto de la población verificada y evaluada. El punto de vista alternativo consiste en dar la prioridad a los criterios, de manera a lo que el resultado de una evaluación significa que un alumno ha satisfecho los criterios dados sin vinculación directa con los éxitos de los otros alumnos (Gipps 1994).

Una cuestión próxima es aquella del reagrupamiento de los resultados. Un método corriente es de tomar las notas a las preguntas tratando de diferentes temas y evaluando diferentes

capacidades, y de sumar esos resultados para dar a cada alumno un puntaje total. Este punto de vista puede ser comparado a las tentativas de contar repollos y zanahorias. Una solución es entonces reportar los resultados bajo la forma de un perfil compuesto de elementos tomados en cuenta separadamente. Entonces es posible decir que la puntuación de cada componente tiene una significación en relación con el éxito sobre los criterios dados.

Una consecuencia de tal discusión es de emitir las dudas sobre la validez de la práctica convencional de los tests conteniendo las preguntas que se establecen sobre una variedad de contextos, de los conocimientos y de saber – hacer y sumar las notas para producir un solo número con una norma referenciada. Eso vuelve a emitir las dudas sobre la identificación de los resultados y sobre la justificación de esta práctica.

Evaluaciones por los profesores

En principio, un profesor que puede describir el éxito de un alumno en el transcurso del tiempo y en los contextos diferentes y que puede discutir las respuestas particulares a manera de comprender el pensamiento que los sostiene, puede constituir una descripción de una mejor fiabilidad que algún test externo pudiera no hacerlo. Igualmente, parece que solo las evaluaciones hechas por los profesores aseguran la validez cuando se trata de evaluar los saber - hacer de los alumnos fuera de la resolución de problemas realistas y esto en contextos diversos.

Las dificultades evidentes son la falta de experticia de los profesores relativas a la evaluación, la falta de posibilidades de comparación de estándares entre diferentes escuelas y por supuesto entre los profesores de una misma escuela, así como los peligros de los prejuicios y deshonestidad .
Alguno de esos obstáculos no puede ser sobrepuesto rápidamente ni a un costo débil.

La comparación entre escuelas puede ser obtenida fuera de discusión en las reuniones de los grupos de escuelas donde los criterios son discutidos y donde los ejemplos de trabajo son intercambiados. De tales reuniones han sido consideradas como muy válidas para la formación de los profesores implicados, revelando frecuentemente cuántos de los profesores están aislados en lo concerniente a sus estándares y sus esperas. Existen otros métodos de homogeneización de los estándares entre escuelas, notablemente por la utilización de visitadores externos inspeccionando los procedimientos así como las muestras de trabajo (Black 1993).

El Reino Unido parece ser el único país en dar, fuera de las pruebas de certificación nacionales, la importancia de la notación, por los propios profesores de los alumnos, de trabajos efectuados fuera de las condiciones formales del examen. Para que ella sea aceptable, disponiendo de la confianza pública, son necesarias reglas precisas y atentas y un sistema que, extrayendo las muestras de las escuelas y verificando sus estándares, pueda asegurar la integridad y la comparación de los resultados. En Suecia es igualmente único de otra manera. En ese país, las evaluaciones de los profesores son los elementos principales en la determinación del resultado global. La prueba externa sirve para calibrar la repartición de las escuelas en su conjunto, pero deja al profesor libre de tomar las decisiones sobre algunos de los alumnos (Black 1992).

3 – La evaluación vista para mostrar resultados

Aquí el objetivo consiste en reseñar la toma de decisiones políticas como recurso y la utilización de informaciones provenientes de las evaluaciones. Entonces los resultados públicos de las certificaciones son utilizados como indicadores de transformación de las escuelas, los detalles de esos resultados fueron errados. Toda vez, para ese objetivo, no es necesario producir resultados fiables y completos para cada individuo. Dando a los alumnos tests diferentes, la transformación global puede ser estudiada en detalle.

Si uno desea estudiar los resultados entre la transformación y otros factores que pueden ser adoptados por la política pública o por aquella de la escuela, las informaciones sobre esos factores deben ser recogidos. Así, la tarea implica la selección y la colección de los resultados sobre los factores tales como la talla de la clase, el contexto familiar del alumno, el tiempo que invierte para aprender, las dotaciones de los laboratorios, y así seguidamente, a manera de aumentar los resultados sobre las transformaciones de los alumnos. El análisis de las posibles relaciones entre los resultados debe ser complejas en razón a las múltiples correlaciones que deben ser exploradas con atención. Cada uno de los factores (por ejemplo el tipo de escuela) puede ser correlacionada con, y aparece así como causa de las variaciones de transformación, entonces que él sea solamente el representante de un factor diferente (por ejemplo los resultados de los alumnos para entrar en una escuela) al cual está asociado. La interpretación de las relaciones es igualmente difícil del hecho de las correlaciones, si son correctamente aisladas por los medios estadísticos, porque ellas no pueden ofrecer por ellas mismas las pruebas de las causas (Woodhouse y Goldstein 1996).

Cuando existen estudios evaluativos a nivel nacional, deben reflejar una estructura de los objetivos y de los criterios. A medida dónde se producen los ítemes de tests de calidad y de los resultados detallados relativos a los objetivos y a los criterios, los trabajos llaman la atención de los profesores e influencian su trabajo: Así, en el "Nacional Science Monitoring" del Reino Unido, la estructura escogida ha sido la de insistir en ventaja sobre la observación y sobre la concepción de investigaciones prácticas (Black 1990). Así, los trabajos correspondientes pudieron haber prescrito, y se puede decir que la operación ha sido por tanto un proyecto de evaluación del hecho que ello ha permitido ir más allá de las prácticas del momento. Así, los estudios evaluativos pueden proveer la ocasión de promover la innovación, tal como ellas pueden representar una fuerza pujante de conservación si ellos reflejan solamente los objetivos y los procedimientos establecidos.

4 – Soporte para el aprendizaje – evaluación formativa

Introducción a los intercambios

La evaluación formativa supone el empleo de un conjunto de muchos tipos de resultados para un solo objetivo. Este objetivo es la modificación del trabajo de aprendizaje para adaptarla a las necesidades que han sido puestas al día por los resultados. Eso es solamente desde que la evaluación sigue esta vía transformándose en formativa. La reacción de regreso puede ir de una reacción inmediata de la clase a una cuestión, hasta una revista

detallada de una variedad de resultados permitiendo de estimar los progresos sobre el conjunto de un tema. En términos de a teoría del control, la utilización del feed –back en ese sentido puede ser considerado como una necesidad evidente.

Las características comunes de las evaluaciones formativas de numerosos países resultan de las investigaciones empleando encuestas (Black 1993). En particular está bien establecido que los programas de evaluación formativa han sido delicadamente conducidas efectivamente en el mejoramiento del aprendizaje del alumno. Una característica común es que la renovación de las prácticas de evaluación forman parte de las modificaciones más amplias de las estrategias de enseñanza y no es solamente un ajuste suplementario.

Otra característica es que, generalmente, y eso es cierto en particular para la evaluación formativa, la evaluación no es valorada por medio de las prácticas y las prioridades de los profesores. Más aún, después que ellos habían preconcebido una gran insistencia para la evaluación por los profesores, como en el programa nacional de Inglaterra y del país de Gales, esas instrucciones han sido generalmente mal interpretadas. Las investigaciones conducidas en vista de las evaluaciones han establecido que la mayoría de los profesores, y en particular los de las escuelas primarias, han restringido el sentido de la evaluación por los profesores en aquel de evaluación sumativa. De las encuestas recientes han igualmente mostrado que en los trabajos de ciencia, muy pocas evaluaciones formativas son realizadas (Russell y otros. 1994).

Una razón preponderante de esta debilidad es que las evaluaciones sumativas, y notoriamente los tests externos, dominan frecuentemente la enseñanza de hecho que sus resultados son tenidos en alta estima y aquellos de otras formas de evaluación. Esto tiene numerosas consecuencias perjudiciales. Los tests externos crean modelos y falsas imágenes de las evaluaciones de los tests. Por ejemplo, una práctica corriente de enseñanza es organizar un test al final de la unidad de enseñanza o al final del trimestre que parece lo más cercano a los tests externos para entregar resultados, o igualmente para hacerlos conocer en público. De hecho que los resultados no han sido utilizados para modificar la enseñanza y el aprendizaje, no se trata de una evaluación formativa. Así una práctica más o menos frecuente de evaluación sumativa es establecida, eso que conduce a la evaluación ha sido asimilado en un test, portador de prejuzgados negativos, sea como una prueba de fuego para los alumnos, sea como un trabajo pesado e improductivo para sus profesores.

Por lo que sea, existen otras razones al débil desarrollo de las prácticas de evaluación formativa. Ellas son relativas a numerosas dificultades prácticas para recoger y registrar los resultados a través de todas las exigencias cotidianas de la enseñanza. Ellas son también un déficit en términos de modificación, de retomar o de diferenciación de la enseñanza para responder a los resultados de la evaluación.

Dos ejemplos

Dos ejemplos específicos, cada uno concerniendo al desarrollo de la práctica en una escuela inglesa, servirían para ilustrar algunos de esos problemas. En la primera escuela (Parkin y Richards 1995), los profesores de ciencias decidieron utilizar la auto – evaluación de los alumnos y las discusiones que han seguido entre profesor y alumnos como base de su

evaluación. Para cada módulo del curso, de los criterios constituidos en objetivos que fueron explicados en un lenguaje accesible a los alumnos. Para cada lección, cada uno de los alumnos disponían de una hoja dándole los criterios con un lugar al lado en la cual los alumnos debían escribir otros comentarios – por ejemplo su placer y su interés.

Más tarde, el profesor anotó cada una de las respuestas a los criterios con uno de los tres códigos siguientes: A – para una comprensión total, P para una comprensión parcial, V – desde el trabajo no ha sido más que "visitado" por el alumno.

A partir del momento donde este método fue introducido, se requirió de un año para que los alumnos utilizaran de manera productiva – al principio, muchos alumnos escribían comentarios muy breves y muy vagos, pero después de un año, esos comentarios han cambiado, ellos se han vuelto muy explícitos y perspicaces y como muy útiles. Los alumnos no estaban acostumbrados a reflexionar sobre su propio aprendizaje como blanco sobre frente a la realización de los objetivos. Ellos deberían igualmente romper con la idea de la evaluación como siendo un test formal.

Algunos alumnos, sobretodo los más débiles, no querían admitir el fracaso, y decían algunas veces haber comprendido entonces que eso no era verdad. Los profesores han insistido cerca de cada alumno para que sus hojas queden como documentos privados destinados a ayudar al profesor a ver los problemas de los alumnos y a proveer una ayuda en caso de necesidad.

En la última escuela (Fairbrother 1995), un profesor de física de una clase de alumnos de 12 – 13 años querían que ellos abordaban el curso sobre la electricidad y el magnetismo de manera más responsable. Tenía por objetivo ayudarlos a:

- situar cada lección en el contexto general del curso;
- tener un resumen de eso que habían hecho en vista de las revisiones;
- ver lo que venía justo después de cada curso.

Le dio a cada alumno una "hoja de revisión" para la secuencia que contenía 25 declaraciones de objetivos, por ejemplo:

> *Saber cómo hacer un electroimán y cómo hacer variar su fuerza.*
> *Saber que es necesario disponer de un circuito completo para todo aparato eléctrico.*
> *Saber que un hilo eléctrico ubicado en un campo magnético va a tratar de mover desde que es atravesado por una corriente.*
> *Saber cómo los interruptores, los repetidores, las resistencias, los captadores y los puertos lógicos pueden ser utilizados para resolver un problema simple, por ejemplos los sonidos de alarma, los avertisseurs de gel, los claros urbanos automáticos.*

La mayor parte de los alumnos tuvieron una pequeña idea de la forma de utilizar esta lista, por ejemplo verificando sus notas en su cuaderno a partir del contenido, o verificando si ellos conocían eso que había sido solicitado. Ciertos alumnos los menos organizados simplemente perdieron esta lista, otros han simplemente puesto de lado y no han hecho referencias a ello.

La explicación del profesor concerniendo a ese fracaso fue que el profesor llevaba mucho sobre el solo contenido y no sobre la forma de aprender. La hoja de revisión tenía por intención tratar ese problema, pero el profesor no había tomado conciencia al principio cuantos en verdad enseñaron sobre su utilización de esta hoja había sido necesaria. Por ejemplo, desde que él solicitaba a los alumnos de revisar con ellos para el test, la mayor parte entre ellos se enredan Parece tener razones principales en ello. La primera fue que los alumnos no sabían cómo extraer, de todo lo que ellos hacían, eso que ellos habían supuesto saber y comprender. Los profesores conocían la diferencia entre los *objetivos* en los cuales ellos quisieran acabar, y los *medios* con los cuales ellos tratan de alcanzarlos. Los alumnos no veían esta diferencia. Una segunda razón fue que los alumnos no sabían aquello que los profesores esperaban de ellos en lo que concierne a sus conocimientos y su comprensión. La mayor parte de ellos aprenden con la experiencia de eso que se espera de ellos, y para muchos alumnos, esta experiencia es dura y desmoralizadora. Ciertos entre ellos, sobretodo los más débiles, no aprendieron jamás.

Desarrollar una buena práctica

La preponderancia tradicional de la función sumativa se traduce por el hecho que hay una lucha para que la evaluación formativa exista y se desarrolla (Fairbrother y otros. 1995). Las tentativas para dar valor la evaluación del profesor pueden reducirse en práctica muy fácilmente en una más grande utilización de esta evaluación para los objetivos sumativos, y en una aplicación más frecuente de las evaluaciones de los profesores, con una colección y un almacenamiento de los resultados que se transforman en carga. La función sumativa puede inhibir de muchas formas el desarrollo de la función formativa para los profesores. La práctica sumativa puede inducir en error, los tests externos son modelos pobres para la evaluación sumativa porque:

- en los tests sumativos, la necesidad de tener un solo resultado global significa que los resultados relativamente diferentes (por ejemplo para la práctica y la teoría) deben ser agregados de forma frecuente arbitraria: la evaluación formativa no debe hacer la misma cosa;
- la evaluación sumativa tiene problemas particulares de referencia a los criterios, en parte del hecho de la necesidad de sumar notas, en parte porque no se puede fiar a un juicio personal fuera de la decisión de la aplicación de criterios generales al trabajo individual de los alumnos; de tales problemas son bien menos importantes en la práctica de la evaluación formativa;
- el trabajo sumativo debe insistir sobre los estándares de uniformidad y de fiabilidad en la recolección y la grabación de los resultados, lo que no es necesario en el trabajo formativo y que inhibe la libertad y la atención a las necesidades individuales que exige el trabajo formativo;
- entonces los procesos sumativos deben ser considerados como siendo equitativos, la práctica formativa con sus prioridades de identificación y de ayuda adaptar a las necesidades de aprendizaje de cada alumno puede conducir entorno a diferentes alumnos de formas muy variadas;
- los objetivos sumativos pueden exigir grandes pruebas argumentadas de los resultados – por ejemplo un audit. – y eso agrega trabajo y deforma la práctica formativa, entonces el trabajo formativo solicita más acción sobre los resultados que sobre su almacenamiento.

Una fuente preponderante de dificultades para el desarrollo de la evaluación formativa es que ella no puede ser simplemente pegado sobre los procedimientos de trabajo ya existentes, pero que ello debe ser elaborado con el procedimiento mismo. Esto solamente porque su utilización para guiar el aprendizaje según las necesidades no puede ser realizada que si los proyectos de enseñanza preveen un tiempo considerable para la planificación y la organización que ello supone.

La utilización efectiva de los feed-back de la evaluación demandan un juicio del profesor, así como la confianza y suavidad en la gestión del proyecto del programa, eso que no puede producir desde que el profesor se siente responsable de la planificación del programa. Así, parece que idealmente, todo procedimiento para incorporar los aspectos formativos debe ser construido por los profesores para ellos mismos. En una tal construcción, los profesores deben generar dos innovaciones – la necesidad de aplicar nuevos métodos para la diferenciación y la suavidad del aprendizaje, y la necesidad de aprender, puede ser inventar, una nueva tecnología para tener los resultados convincentes sobre el éxito de los alumnos.

La utilización de los resultados de las evaluaciones formativas puede ser el aspecto más estimulante. Existen "macros" respuestas, en término de repartición de los grupos por niveles, pero ellos no conciernen a las necesidades inmediatas: Los profesores han respondido organizando las unidades de trabajo alrededor de un núcleo y de ramificaciones: Allí el trabajo es muy variado, después el tratamiento de nuevos temas más profundos para aquellos que desean ir más lejos hasta la repetición de las bases para aquellos siendo las necesidades más fundamentales (Black 1993). Otros indican los puntos de vista menos formales y más suaves, con las revisiones aprovechando ocasiones favorables fuera de los trabajos ulteriores. Hay una postura de suavidad o de rigidez de los programas.

La tecnología de colección de resultados sobre los progresos de los alumnos comienza justo a desarrollarse. La mayor parte de los profesores siempre han utilizado una variedad de fuentes de manera informal – es necesario afinar esta práctica en la perspectiva de obtener resultados mejor utilizados. Las hojas descritas en el primer ejemplo descrito anteriormente muestra un medio para hacerlo; los resultados se distinguen en lo que ellos producen las informaciones detalladas en relación con las formulaciones de objetivos específicos – es decir que la referencia en los criterios se aplica naturalmente; porque es una necesidad para la evaluación. Es más, porque ese producto de los resultados escritos de manera sistemática, la enseñanza es aligerar de la presión de la notación y de la grabación exhaustivas de los eventos en clase. Los resultados sobre las circunstancias pueden toda vez tener una importancia en ella - misma: ciertos profesores han encontrado particularmente útil – y eso es sorprendente – suspender su enseñanza activa durante un momento – explicando claramente en la clase lo que ellos hacían y por qué – y de concentrar sobre la observación y escuchar de un pequeño número de alumnos (ver Cavendish y otros. 1990, Connor 1991).

Cuando la actividad de evaluación es construida en estrecha relación con el programa de aprendizaje, sería insensato impedir a los alumnos de comentar sus resultados, de colocarse

en pregunta y de rehacer la evaluación si ellos desean mejorar sus transformaciones. Así, una consecuencia de su rol de soporte para el aprendizaje es que las evaluaciones formativas se vuelven a su vez informales, y conducidas por los alumnos. La importancia dada a la auto evaluación de los alumnos es una característica notable. La experiencia muestra que los alumnos no pueden asumir un rol eficaz en su propia evaluación, sólo fuera de los programas a largo término concebidos para ayudarlos a aprobar y a guardar una vista del conjunto de sus objetivos de aprendizaje, de manera de aplicar los criterios correspondientes a sus propios progresos. Como lo muestran esos dos ejemplos, se debe enseñar a los alumnos cómo evaluar sus propios progresos. Una parte importante de este trabajo es la traducción de los objetivos del programa en un lenguaje que los alumnos pueden comprender, y frente al nivel de detalles que pueden ayudarlos a poner directamente en relación sus esfuerzos de aprendizaje. Es perseguir igualmente que los objetivos deben ser a la vez accesibles a corto término y suficientemente modestos en relación con las perspectivas de éxito de los alumnos. Esas exigencias conciernen particularmente a los alumnos que reencuentran dificultades especiales de aprendizaje – pero ellas son importantes para todos.

Los profesores que han elaborado la auto – evaluación de los alumnos señalan muchas ventajas – los alumnos pueden dirigir sus propios esfuerzos más claramente y eficazmente, ellos pueden estar implicados más activamente y más motivados en relación con sus propios progresos, ellos pueden entonces sugerir su propia manera de mejorar sus adquisiciones, e igual ellos pueden volver a poner en cuestión las evaluaciones que ellos estiman son injustas.

Manifiestamente, la implicación de los alumnos puede dar la conducción de un programa de evaluación formativa más fácil a los profesores. Toda vez, esta implicación cambia igualmente el rol de los alumnos como aprendices y la naturaleza – de la relación entre el profesor y los alumnos, lo que hace reposar sobre las espaldas de aquellos la responsabilidad del aprendizaje. Igualmente fuera de la necesidad de mejorar la evaluación, la necesidad primera de mejora del aprendizaje exige de tales cambios. En efecto, se ha argumentado que la metacognición, en tanto que sensibilización y auto – dirección sobre la naturaleza del aprendizaje, es esencial al desarrollo de los alumnos por el aprendizaje de un concepto. El trabajo descrito previamente sirve claramente a este objetivo (ver Brown 1987, White y Gunstone 1989, Baird y Northfield 1992). La evaluación formativa así mejorada puede conducir a los cambios que deberían ser una ayuda pujante para el desarrollo personal de los alumnos y que deberían igualmente hacer parte de todo programa para ayudarlos a ser aprendices más eficaces.

Los sistemas y los roles

Una buena evaluación sumativa exige la implicación de los profesores. Parece así que no existe otra solución que de vincular los profesores en los roles a la vez sumativos y formativos, utilizando para esas dos funciones algunos de sus resultados, pero no necesariamente todos. Entonces es necesario distinguir cuidadosamente los métodos y las necesidades en relación con cada uno de los dos objetivos. Asumir los dos roles de esta

manera sería muy exigente. De un lado, existen las necesidades de aprendizaje de sus alumnos, eso debería ser su primera preocupación. Por otro lado, existen las presiones y las coacciones provenientes del exterior. Los sistemas nacionales y regionales tienen fuertes posturas y crean las presiones sobre los profesores, eso que los obliga a trabajar en el seno de una estructura que guía a la vez las decisiones de sus escuelas y las expectativas de los padres. El profesor debe encontrar el medio justo entre las presiones provenientes de los dos lados.

La razón principal de la insistencia sobre este punto es que algunos de los objetivos importantes en física no pueden ser reflejados, y ser así sostenidos, por los sistemas de evaluación que reposan solamente sobre los tests externos cortos. La reforma de los controles sumativos nacionales es una necesidad seria. En 1992, al final de una revista de los tests nacionales de física en once países, yo escribí el resumen siguiente en vista de una conclusión:

> Una conclusión que se establece de este estudio es que la variedad de los métodos utilizados y la variedad de las capacidades evaluadas por esos exámenes de física son muy débiles. Ellos han tenido ciertamente un efecto restrictivo serio sobre el desarrollo de la física en la escuela y sobre el reclutamiento de físicos. Hay muchas razones en ello. La insuficiencia de los recursos, con otras coacciones del sistema, pueden explicar lo diligente de los examinadores de física a trabajar con los sistemas que ellos estimen ser, al menos, lejos del ideal y pueden ser perjudiciales para el futuro de la física. Puede ser esta situación es ella aceptada muy voluntaria para todos nosotros. (Black 1992)

Los exámenes públicos u oficiales tienen un poder particular sobre el futuro de la física. Determinando los objetivos y la estructura con los cuales los profesores de las escuelas secundarias piensan que ellos deben trabajar, ellos determinan la estructura y la imagen del sujeto a los ojos de los jóvenes. Si de tales exámenes no se suscitan o no estimulan las actividades que son importantes y atrayentes para el físico y si ellos transmiten una imagen muy restrictiva del campo, ellos atraerán demasiado poco a los especialistas, y darán a todo adulto una vista muy negativa de la física.

Referencias

Baird, J.R. et Northfield, J.R. (eds.) (1992) *Learning from the PEEL experience*. Melbourne: Monash University.

Black, H (1993) Assessment: A Scottish Model pps.91-94 in Fairbrother, R , Black, P.J. and Gill, P. (eds.) *TAPAS : Teacher Assessment of Pupils: Active Support. King's Education Papers No.3.* London: C.E.S. King's College.

Black, P.J. (1990) APU Science - the past and the future. *School Science Review* 72. 13-28

Black, P.J. (1992) *Physics Examinations for University Entrance : an International Study.* Science and Technology Education - Document No. 45. Paris : UNESCO.

Black, P.J. (1993), Formative and Summative Assessment by Teachers. *Studies in Science Education.* 21. 49 - 97.

Britton, E.D. et Raizen, S.A. (eds.) (1996) *Examining the Examinations : An International Comparison of Science and Mathematics Examinations for College-Bound Students.* Boston : Kluwer

Brown, A. (1987) Metacognition, executive control, self-regulation and other mysterious mechanisms. pps 65 - 116 in Weinert, F.E and Kluwe, R.H. (eds.) *Metacognition, Motivation, and Understanding.* Hillsdale, New Jersey: Lawrence Erlbaum.

Cavendish, S., Galton, M., Hargreaves, L. et Harlen, W. (1990) *Observing Activities.* London: Paul Chapman.

Connor, C. (1991) *Assessment and Testing in the Primary School.* London: Falmer Press.

Fairbrother, R. (1995) *Pupils as Learners.* pp.105-124 in Fairbrother et al. op.cit.

Fairbrother, R., Black, P.J., et Gill, P. (eds.) (1995) *Teachers Assessing Pupils : Lessons from Science Classrooms.* Hatfield UK : Association for Science Education.

Gipps, C.V. (1994) *Beyond Testing : Towards a Theory of Educational Assessment*, London : Falmer, .

Gipps, C.V. et Murphy, P. (1994) *A fair test ? Assessment, achievement and equity,* Milton Keynes : Open University Press.

Gauld, C.F. (1980) Subject oriented test construction, *Research in Science Education,* 10, 77--82.

Johnson S. (1988) *National Assessment : the APU Science Approach.* London : Her Majesty's Stationery Office.

Messick, S. (1989) Validity pp. 12 - 103 in Linn, R.L. (ed.), *Educational Measurement (3rd. Edition)',* London : Collier Macmillan.

Parkin, C et Richards, N (1995) *Introducing Formative Assessment at KS3 : an attempt using pupils' self-assessment.* pp 13-28 in Fairbrother et al. op.cit.

Resnick, L.B. et Resnick, D.P.(1992) Assessing the Thinking Curriculum: New Tools for Educational Reform pp. 37 - 75 in Gifford, B.R. & O'Connor, M.C.(eds.), *Changing Assessments : Alternative Views of Aptitude, Achievement and Instruction,* Boston : Kluwer.

Russell, T., Qualter, A., McGuigan, L. et Hughes, A. (1994), *Evaluation of the implementation of Science in the National Curriculum at Key Stages 1, 2 and 3.* London: School Curriculum and Assessment Authority.

Shavelson, R.J., Baxter, G.P. et Gao, X.(1993) Sampling variability of performance measurements *Journal of Educational Measurement* 30. 215--232.

Tamir, P. (1990) Justifying the selection of answers in multiple-choice questions. *International Journal of Science Education* 12. 563-573.

White, R.T. et Gunstone, R.F. (1989) Meta-learning and conceptual change. *International Journal of Science Education.* 11. 577-586.

Woodhouse, G. et Goldstein, H. (1996) The Statistical Analysis of Institution-based Data pp.135-144 in Goldstein, H. and Lewis, T. (eds.) *Assessment: Problems, Developments and Statistical*