Gordon Stobart

Tiempos de pruebas: Los usos y abusos de la evaluación

GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN

Morata

GOBIERNO DE ESPAÑA | MINISTERIO DE EDUCACIÓN

INSTITUTO DE FORMACIÓN DEL PROFESORADO, INVESTIGACIÓN E INNOVACIÓN EDUCATIVA

EDICIONES MORATA, S. L.

Tema: **Evaluación educativa**

Tiempos de pruebas

Los usos y abusos de la evaluación

Por

Gordon Stobart

Traducido por

Pablo Manzano Bernárdez

Obras en coedición con el Ministerio de Educación

1. **Zimmermann, D.:** *Observación y comunicación no verbal en la escuela infantil* (3ª ed.).
2. **Oléron, P.:** *El niño: su saber y su saber hacer* (2ª ed.).
3. **Loughlin, C. y Suina, J.:** *El ambiente de aprendizaje: diseño y organización* (5ª ed.).
4. **Browne, N. y France, P.:** *Hacia una educación infantil no sexista* (2ª ed.).
5. **Selmi, L. y Turrini, A.:** *La escuela infantil a los tres años* (4ª ed.).
6. **Selmi, L. y Turrini, A:** *La escuela infantil a los cuatro años* (3ª ed.).
7. **Saunders, R. y Bingham-Newman, A. M.:** *Perspectivas piagetianas en la educación infantil* (2ª ed.).
8. **Driver, R., Guesne, E. y Tiberghien, A.:** *Ideas científicas en la infancia y la adolescencia* (4ª ed.).
9. **Harlen, W.:** *Enseñanza y aprendizaje de las ciencias* (6ª ed.).
10. **Selmi, L. y Turrini, A.:** *La escuela infantil a los cinco años* (3ª ed.).
11. **Bale, J.:** *Didáctica de la geografía en la escuela primaria* (3ª ed.).
12. **Tann, C. S.:** *Diseño y desarrollo de unidades didácticas en la escuela primaria* (3ª ed.).
13. **Willis, A. y Ricciuti, H.:** *Orientaciones para la escuela infantil de 0 a 2 años* (3ª ed.).
14. **Orton, A.:** *Didáctica de las matemáticas* (4ª ed.).*s*
15. **Pimm, D.:** *El lenguaje matemático en el aula* (3ª ed.).
16. **Moyles, J. R.:** *El juego en la educación infantil y primaria* (2ª ed.).
17. **Arnold, P. J.:** *Educación física, movimiento y curriculum* (3ª ed.).
18. **Graves, D. H.:** *Didáctica de la escritura* (3ª ed.).
19. **Egan, K.:** *La comprensión de la realidad en la educación infantil y primaria.*
20. **Hargreaves, D. J.:** *Infancia y educación artística* (3ª ed.).
21. **Lancaster, J.:** *Las artes en la educación primaria* (3ª ed.).
22. **Bazalgette, C.:** *Los medios audiovisuales en la educación primaria.*
23. **Newman, D., Griffin, P. y Cole, M.:** *La zona de construcción del conocimiento* (3ª ed.).
24. **Swanwick, K.:** *Música, pensamiento y educación* (3ª ed.).
25. **Wass, S.:** *Salidas escolares y trabajo de campo en la educación primaria.*
26. **Cairney, T. H.:** *Enseñanza de la comprensión lectora* (4ª ed.).
27. **Nobile, A.:** *Literatura infantil y juvenil* (3ª ed.).
28. **Pluckrose, H.:** *Enseñanza y aprendizaje de la historia* (4ª ed.).
29. **Hicks, D.:** *Educación para la paz* (2ª ed.).
30. **Egan, K.:** *Fantasía e imaginación: su poder en la enseñanza* (3ª ed.).
31. **Escuelas infantiles de Reggio Emilia:** *La inteligencia se construye usándola* (4ª ed.).
32. **Secada, W. G., Fennema, E. y Adajian, L. B.:** *Equidad y enseñanza de las matemáticas: nuevas tendencias.*
33. **Crook, Ch.:** *Ordenadores y aprendizaje colaborativo.*
34. **Gardner, H., Feldman, D. H. y Krechevsky, M. (Comps.):** *El Proyecto Spectrum. Tomo I: Construir sobre las capacidades infantiles.*
35. **Gardner, H., Feldman, D. H. y Krechevsky, M. (Comps.):** *El Proyecto Spectrum. Tomo II: Actividades de aprendizaje en la educación infantil.*
36. **Gardner, H., Feldman, D. H. y Krechevsky, M. (Comps.):** *El Proyecto Spectrum. Tomo III: Manual de evaluación para la educación infantil* (2ª ed.).
37. **Cooper, H.:** *Didáctica de la historia en la educación infantil y primaria.*
38. **Cummins, J.:** *Lenguaje, poder y pedagogía.*
39. **Haydon, G.:** *Enseñar valores. Un nuevo enfoque.*
40. **Gross, J.:** *Necesidades educativas especiales en educación primaria.*
41. **Beane, J. A.:** *La integración del currículum* (2ª ed.).
42. **Defrance, B.:** *Disciplina en la escuela.*
43. **Siraj-Blatchford, J. (Comp.):** *Nuevas tecnologías para la educación infantil y primaria.*
44. **Peacock, A.:** *Alfabetización ecológica en educación primaria.*
45. **Abdelilah-Bauer, B.:** *El desafío del bilingüismo.*
46. **Hargreaves, A. y Fink, D.:** *El liderazgo sostenible.*
47. **Lankshear, C. y Knobel, M.:** *Nuevos alfabetismos. Su práctica cotidiana y el aprendizaje en el aula* (2ª ed.).
48. **Arnot, M.:** *Coeducando para una ciudadanía en igualdad.*
49. **Jarman, R. y McClune, B.:** *El desarrollo del alfabetismo científico.*
50. **Stobart, G.:** *Tiempos de pruebas. Los usos y abusos de la evaluación.*

— Colección *Proyectos curriculares*

Aitken, J. y Mills, G.: *Tecnología creativa* (6ª ed.).
Dadzie, S.: *Herramientas contra el racismo en las aulas.*
Suckling, A. y Temple, C.: *Herramientas contra el acoso escolar. Un enfoque integral.*
Barkley, E. F. y cols.: *Técnicas de aprendizaje colaborativo.*

Gordon STOBART

Tiempos de pruebas

Los usos y abusos de la evaluación

GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN

EDICIONES MORATA, S. L.

Título original de la obra:
TESTING TIMES:
The uses and abuses of assessment

Coeditan:

MINISTERIO DE EDUCACIÓN
Secretaría de Estado de Educación y Formación Profesional
Instituto de Formación del Profesorado, Investigación e Innovación Educativa
Secretaría General Técnica
Catálogo de publicaciones del Ministerio: educacion.es
Catálogo general de publicaciones oficiales: 060.es

Y

EDICIONES MORATA, S. L.
Mejía Lequerica, 12. 28004 Madrid
www.edmorata.es **-** morata@edmorata.es

Derechos reservados
Depósito Legal: M-44825-2010
ISBN: 978-84-7112-629-0

NIPO: 820-10-202-6
Compuesto por: Ángel Gallardo Serv. Gráficos, S. L.
Printed in Spain - Impreso en España
Imprime: ELECE Industrias Gráficas, S. L. Algete (Madrid)
Fotografía de la cubierta: *Notas de Otoño* (2010) por Iñigo Cosín Fernández. Reproducida con autorización.

Contenido

Agradecimientos

Esta obra sostiene que la evaluación, aunque pueda referirse a una persona, es esencialmente una actividad social. Lo mismo cabe decir de la redacción de este libro. He recibido estímulos e ideas tanto de las personas que me rodean como de los que forman parte de una comunidad más amplia y de aquellos que se dedican a la evaluación.

El lugar de honor le corresponde a mi esposa, Marie Adams, que no solo me ha brindado apoyo y estímulo durante toda la redacción, sino que ha utilizado también su experiencia periodística para comentar los borradores de los capítulos. Le estoy profundamente agradecido.

Otras personas también han participado en el comentario de determinados capítulos y el libro se ha beneficiado en gran medida de ello. Les agradezco también la información reflexiva que me han facilitado, a Richard DAUGHERTY, Kathryn ECCLESTONE, Steve EDWARDS, Harvey GOLDSTEIN, Eleanore HARGREAVES, Tina ISAACS, Mary JAMES, Angela LITTLE, Andrew MACALPINE, John WHITE y Alison WOLF.

Muchas ideas que aparecen aquí son reflejo de extensos diálogos con colegas y amigos. Soy consciente de la deuda intelectual contraída por mí tanto con las personas a las que les pedí que comentaran capítulos del libro como con quienes he trabajado. Entre ellas están otros miembros del *Assessment Reform Group*, que han influido en mi pensamiento durante los diez últimos años: Paul BLACK, Patricia BROADFOOT, John GARDNER, Caroline GIPPS, Wynne HARLEN, Paul NEWTON y Dylan WILIAM.

Gracias también a Michelle COTTLE, que realizó amablemente la nada envidiable tarea de preparar la bibliografía y los índices.

Introducción

> *Cómo se llaman las cosas* tiene una importancia descomunalmente mayor que lo que son... basta crear nuevos nombres, apreciaciones y verdades aparentes para crear "cosas" nuevas.
>
> (Friedrich NIETZSCHE, 1887.)

> En la sociedad contemporánea, los tests no describen al individuo, sino que, más bien, lo construyen.
>
> (Allan HANSON, 1994.)

La evaluación, en forma de tests y exámenes, es una poderosa actividad que configura la manera que tienen las sociedades, los grupos y los individuos de entenderse a sí mismos. Aquí, desarrollaremos tres argumentos específicos:

- La evaluación es una actividad social marcada por valores y no existe nada que se parezca a una evaluación independiente de las culturas.
- La evaluación no mide objetivamente lo que hay, sino que crea y configura lo que se mide: es capaz de "componer personas".
- La evaluación influye directamente en lo que aprendemos y en cómo lo aprendemos y puede limitar o promover el aprendizaje efectivo.

Estas características otorgan a la evaluación una considerable autoridad y llevan a consecuencias constructivas o destructivas: los *usos y abusos* del título del libro.

La creación de aprendices: Los casos de Hannah y Ruth

Para dar cuerpo a estas afirmaciones acerca de que la evaluación configura nuestra forma de vernos y de aprender, presentaré a dos estudiantes que las personifiquen.

Hannah la inútil

Hannah es el nombre de una alumna inglesa de 11 años, de una clase estudiada por Diane REAY y Dylan WILIAM que se estaba preparando para los tests nacionales (*SAT**), que se realizan en Inglaterra a los niños y niñas del último curso de educación primaria. Estos tests no tienen consecuencias selectivas importantes para los alumnos, dado que ya habrán escogido sus correspondientes institutos de secundaria, pero los resultados revisten una importancia crítica para sus escuelas y maestros, pues se los juzga públicamente según esos resultados. En consecuencia, se hace gran hincapié en la preparación de los exámenes (véase el Capítulo VI), porque, para los maestros, la meta es conseguir que la mayor cantidad posible de alumnos alcance el nivel 4 o superior [1], dado que los objetivos de la escuela y los nacionales se basan en ello. Gracias a los tests y los ejercicios, los niños se conciencian del nivel que se prevé que alcancen. En este contexto, se desarrolló la conversación siguiente:

HANNAH: Tengo verdadero miedo a los SATS. La Sra. O'Brien [una maestra de la escuela] vino y nos habló de nuestra ortografía y yo no voy muy bien en ortografía y David [el maestro de la clase] está poniéndonos todas las mañanas exámenes de las tablas de multiplicar y yo soy una negada con las tablas, por eso tengo miedo a hacer los SATS y quedar como una inútil.

DIANE: No lo entiendo, Hannah. Tú no puedes ser una inútil.

HANNAH: Sí, sí que puedo, porque tienes que conseguir un nivel 4 o un nivel 5 y si no vas bien en ortografía y con las tablas de multiplicar, no llegas a esos niveles y, entonces, quedas como una inútil.

DIANE: Estoy segura de que eso no es así.

HANNAH: Sí lo es, porque eso es lo que dice la Sra. O'Brien.

(REAY y WILIAM, 1999, pág. 345.)

Para hacer más sangrante esta declaración de inutilidad, los autores señalan que Hannah era "una escritora consumada, una bailarina y artista muy dotada y muy buena resolviendo problemas, aunque ninguna de estas habilidades tiene valor ante sus propios ojos. En cambio, se considera un fracaso, una nulidad académica" (pág. 346). No se trataba de un ejemplo aislado. En la época de los SATS, los niños presentaban a los demás por sus niveles, y estos habían empezado a afectar sus relaciones sociales, de manera que Stuart, un alumno "nivel 6", se estaba convirtiendo en objetivo de acoso en el patio de recreo. Cuando se les preguntó por las consecuencias de los resultados obtenidos en los SATS, se desarrolló esta conversación:

---

* Aunque son conocidos como SATS, debido a que la intención original era introducir unas *Standard Assessment Tasks* ("tareas normalizadas para evaluación"), se trata de las pruebas de evaluación del *National Curriculum* en Inglaterra. (*N. del T.*)

[1] El nivel 4 es el nivel de rendimiento esperado de la mayoría de los niños de 11 años en la escala de 1 (el nivel más bajo) a 8 del currículum nacional. El Gobierno ha establecido el objetivo de que el 85% de los alumnos alcancen este nivel en Lenguaje y Matemáticas en 6.º curso (un objetivo que todavía no se ha conseguido). Las escuelas de primaria son evaluadas según el porcentaje de sus alumnos que alcanzan este nivel. Véase el Capítulo VI.

SHARON: Me parece que sacaré un 2; solo Stuart conseguirá un 6.
DIANE: Y, si Stuart consigue un 6, ¿qué supondrá eso para él?
SHARON: Acabará teniendo un buen trabajo y una buena vida, y eso indica que no vivirá en las calles y cosas así.
DIANE: Y si tú consigues un nivel 2, ¿qué significará eso para ti?
SHARON: Umm. Que no tengo ante mí una buena vida y que creceré y me dedicaré a hacer algo feo o algo así.

(Pág. 347.)

No hace mucho, Tamara Bibby ha encontrado actitudes semejantes en su investigación sobre clases de primaria:

> Los niños empiezan a pensar en sí mismos como niveles. Y eso mezclado con la moralidad y la bondad. Las buenas personas trabajan mucho y escuchan en clase. Si de repente resulta obvio que tu compañero o compañera alcanza niveles más bajos que tú, ¿es una buena persona? Puede poner verdaderamente a prueba la amistad[2].

Ruth la pragmática

El caso de Ruth Borland desató un debate en los media de Irlanda después de que, en una entrevista en el *Irish Times* (20 de septiembre de 2005), revelara cómo había obtenido las notas máximas en el *Leaving Certificate* nacional, el pasaporte para la universidad[3]. Ruth se presentaba como una persona decidida y competitiva ("No soporto la idea de que alguien haga algo mejor que yo"). Esta determinación se había puesto de manifiesto durante varios años en sus estudios y, tras comprobar en unas prácticas en un bufete de abogados que lo suyo no era el derecho, cambió de asignaturas y facultad antes de matricularse en ciencias empresariales.

Cuando le preguntaron por su forma de abordar los estudios, Ruth reveló que, reconociendo que "las asignaturas de empresariales eran mi fuerte, sabía que la mejor manera de obtener los puntos que necesitaba era matricularme en todas las optativas del *Leaving Certificate*" (éstas se consideran como uno de los conjuntos optativos más fáciles, cuyos resultados contribuyen a la puntuación final). En las otras materias se desenvolvió bien porque consistían en:

> Aprender la fórmula de cada examen y practicarla incesantemente. Saqué un A1 en Lenguaje porque sabía exactamente lo que se pedía en cada pregunta. Lo descubrí a partir de las respuestas de ejemplo dadas por los examinadores y sabía cuánta información hacía falta y en qué formato en cada sección de la hoja. Así es como puedes desenvolverte bien en estos exámenes... No tiene sentido saber todas esas cosas que no entran en los exámenes. Los profesores que decían: "Esto no tenéis que saberlo para los exámenes, pero os lo voy a decir de todos modos" me frustraban siempre. Yo quería mi A1; ¿qué sentido tiene estudiar materiales que no aparecen en los exámenes?

[2] *Times Educational Supplement*, 9 de febrero de 2007, pág. 13.
[3] Agradezco a Anne Looney su advertencia al respecto.

Este enfoque profundamente instrumental generó una rica correspondencia. Las cartas iban desde el horror ante un sistema educativo que hacía posible que una de las personas que habían obtenido "mejores resultados en los exámenes se mostrara tan brutalmente despreciativa de una visión amplia del aprendizaje" hasta los partidarios de ella, que elogiaban su sentido común al estudiar el sistema y su determinación y esfuerzo. A Ruth se le concedió el derecho de réplica y optó por sugerir que dirigieran su cólera hacia un sistema educativo fracasado y no a ella:

> Opté por no enfrentarme con el sistema, sino por jugar con él. Hice lo que tenía que hacer para alcanzar mis metas. Jugué, si quieren. Yo no llamaría "utilitaria" a esta actitud, sino realista. Ingresé en la universidad a estudiar las asignaturas que me gustan. Ya tendré "el placer del descubrimiento" en las asignaturas de empresariales y de economía.
>
> (*Irish Times*, 27 de septiembre de 2005.)

Y a pesar de críticas como: "Gracias a Dios, Ruth Borland va a ser actuaria. Me horrorizaría ver a alguien con su actitud ingresando en Medicina", quizá sea Ruth quien haya reído la última: acabó contratada por el *Irish Times*, en donde tenía una columna en la que daba consejos para la preparación de los exámenes.

El desarrollo del argumento

Este libro trata de cómo las evaluaciones configuran nuestra forma de vernos a nosotros mismos, igual que a Hannah y a Ruth: nuestra capacidad de aprender y el tipo de aprendices que somos. Pero no solo trata de tests educativos y exámenes. El argumento que sostengo es que estamos siendo configurados por otras formas de evaluación. Las más insidiosas son las que prometen revelar nuestras habilidades y aptitudes subyacentes. El ejemplo supremo de éstas es el test de CI, con sus presunciones históricas de revelar unas capacidades intelectuales innatas que poco pueden cambiar (Capítulo II). Ciertos movimientos, como el de las *inteligencias múltiples*, de Howard GARDNER, y el de la *inteligencia emocional*, de Daniel GOLEMAN, también definen quiénes somos, aunque puedan parecer más benignos. Pero estos también nos evalúan y nos clasifican como un tipo de aprendiz o de persona. En el Capítulo III, cuestiono algunas de las suposiciones que utilizan estos enfoques, como hago, en el Capítulo IV, con respecto a la evaluación de los *estilos de aprendizaje*.

De qué modo puede configurar la evaluación lo que aprendamos y cómo lo aprendamos es el tema de los Capítulos V al VII. Ruth Borland ejemplifica cómo pueden convertirse las evaluaciones decisivas en un fin en sí mismas: lo importante son las calificaciones, no lo que se haya aprendido. Es lo que Ronald DORE llamaba *Diploma Disease**, que, según él, es virulento en el desarrollo de economías en las que, para conseguir un trabajo, son necesarios unos niveles de cua-

* Literalmente: "enfermedad de títulos", que hemos traducido como *titulitis*. (*N. del T.*)

lificación cada vez más elevados. Examino el impacto de este tipo de rutina de exámenes en el Capítulo V y busco formas de mejorar la calidad del aprendizaje mediante evaluaciones de mayor calidad. Hannah tipifica el modo en que una cultura de rendición de cuentas orientada a objetivos configura también lo que se enseña y se aprende. En el Capítulo VI, reviso los efectos reductores y deformantes de los objetivos basados en los resultados de los exámenes. Sostengo que, aunque estos objetivos puedan tener ciertas ventajas a corto plazo, rápidamente se degradan a un "juego con el sistema" y debilitan el aprendizaje efectivo. Para limitar la influencia dañina de los tests orientados a una estricta rendición de cuentas, presento una visión diferente de lo que podría constituir una *rendición de cuentas inteligente*.

Cómo puede utilizarse la evaluación para fomentar la enseñanza y el aprendizaje eficaces es el tema del Capítulo VII: Razones para alegrarse: La Evaluación para el Aprendizaje. Este enfoque incorpora la evaluación al proceso de enseñanza y aprendizaje, en vez de centrarse en lo que se haya aprendido al final del proceso (evaluación *del* aprendizaje). En el centro de la evaluación para el aprendizaje está la calidad de las interacciones en clase y examinaré algunos aspectos de su complejidad.

El capítulo final recoge lo que hace falta para que la evaluación desempeñe un papel más positivo como ayuda para comprendernos mejor a nosotros mismos y para promover un aprendizaje más profundo. Esto implica otorgar un papel más modesto a la evaluación; una interpretación más cauta de los resultados, y un mayor reconocimiento de los elementos sociales e interactivos que intervienen, lo que nos lleva a ver cómo podemos ayudar a los aprendices a elaborar enfoques *autorreguladores* de la evaluación en los que ellos tomen sus propias decisiones sobre sí mismos en cuanto aprendices y en cuanto personas. Esto cobra cada vez más importancia dado que se preparan para un futuro desconocido en el que esas destrezas serán esenciales.

¿A quién se dirige este argumento? Los lectores en los que pienso se interesan personal o profesionalmente por la evaluación y es posible que les preocupen determinados temas, por ejemplo, el debate en torno a las inteligencias o con respecto a la influencia de los tests. Para quienes padecen la imposición acrítica de ciertas iniciativas, por ejemplo, la "inteligencia emocional" o los "estilos de aprendizaje", ofrece un punto de vista alternativo sobre cómo interpretarlas.

Esta obra toma partido, aunque muchos de los argumentos sean conocidos. No obstante, es oportuno replantear estas interpretaciones alternativas con el fin de cuestionar las afirmaciones de la psicología popular, de los planificadores y de los gurús que tienen la "Respuesta", sobre todo cuando las afirmaciones se enuncian pero no se prueban. Espero dar un sentido más interrogativo a afirmaciones como: "tenemos que organizar por capacidad", "los tests elevan los niveles" y "eres un aprendiz cinestésico".

No se garantiza de ninguna manera que los argumentos de este libro sean de "fácil lectura". No obstante, he procurado hacerlos accesibles evitando la práctica académica de insertar copiosas citas en el texto, en apoyo de cada afirmación. En cambio, para quienes quieran estudiar alguna más a fondo, incluyo en el texto suficientes pistas para localizar las fuentes en la bibliografía. He utilizado también notas al final del capítulo tanto para desarrollar algunos de los puntos más difíci-

les de conocer como para mencionar las fuentes de ciertos pensamientos o evidencias. Acepto que esto resulte irritante para los colegas universitarios, que puedan reconocer sus ideas sin ver sus nombres en el texto.

El poder de la evaluación

La evaluación, en el amplio sentido de recabar pruebas con el fin de hacer un juicio, forma parte de la trama de la vida. Nuestros antepasados tenían que decidir por dónde cruzar ríos y montañas y cuándo iniciar los cultivos. El hecho de escoger el sitio para Stonehenge y la alineación astronómica de las rocas sigue siendo a día de hoy un impresionante ejercicio de evaluación.

Sin embargo, lo que me interesa es la recogida deliberada de pruebas para hacer juicios específicos sobre personas o grupos. Allan HANSON define una prueba como una "técnica representacional aplicada por un organismo a una persona con la intención de recabar información" (pág. 19). Esta definición puede aplicarse más en general a las formas estructuradas de evaluación. Su valor como definición es que señala el carácter *representacional* de las pruebas; a menudo, una prueba sustituye y actúa como una metáfora de lo que una persona puede hacer. La adecuación de la metáfora (por ejemplo, hasta qué punto representa un test de personalidad el carácter de una persona) está en el centro de los argumentos acerca de la validez de la evaluación. Esta definición enfatiza también la dimensión social de la evaluación, incluyendo el poder que tienen los administradores de los tests sobre quienes los cumplimentan. Con frecuencia, esta recogida de información se basa en el supuesto de que los tests revelan verdades objetivas ocultas a la observación directa. HANSON lo discute:

> Estas suposiciones son erróneas. A causa de su calidad representacional, los tests miden la verdad tal como se ha estructurado culturalmente y no como algo que existe independientemente... Por su misma existencia, los tests modifican e incluso crean lo que pretenden medir.
>
> (Pág. 47.)

Esta observación refleja uno de los temas principales de este libro, como hace la cita inicial de NIETZSCHE: *que la evaluación configura quiénes y qué somos y no puede considerarse como una medida neutra de habilidades o destrezas, independiente de la sociedad*. La evaluación del individuo es, paradójicamente, una actividad intrínsecamente social.

Inventando personas

El filósofo de la ciencia Ian HACKING ha elaborado un razonamiento más amplio acerca de cómo "a veces, nuestras ciencias crean tipos de personas que, en cierto sentido, no existían antes" (pág. 2) [4]. He optado por desarrollar aquí su

[4] Alison WOLF señala que, "en cierto sentido", no es más que una cadena de palabras engañosas, cuando aceptamos que las personas difieren, por ejemplo, en la proporción altura/peso y en

razonamiento porque constituye un útil marco de referencia para comprender cómo puede la evaluación clasificar a las personas de tal manera que pueda considerarse después que esa clasificación representa alguna realidad objetiva. Evidentemente, las personas existen con independencia de las medidas y difieren de muchas maneras; lo que configura las identidades es la *elección social* de la forma de evaluarlas, clasificarlas y ordenarlas. Así, ciertas denominaciones, como "disléxico", "THDA" y "síndrome de Asperger", han pasado recientemente a ser de uso común, y hacemos presuposiciones acerca de las personas así denominadas.

Uno de los ejemplos de HACKING es el "descubrimiento" de la "personalidad múltiple" en la década de 1970. Esto condujo a un rápido incremento del número de personas que mostraban el síndrome y del número de personalidades manifestadas (la primera persona tenía 2 o 3; al final de la década, el número medio era 17). Una serie de procesos sociales formaron parte de este desarrollo, cuyo resultado final fue la aparición de una nueva persona reconocible, *la múltiple*, con una identidad reconocible. Surgieron incluso "bares" en los que los múltiples se socializaban (uno podía encontrarse allí con montones de personalidades). HACKING propone un marco de referencia de cinco elementos interactivos que ocasionan este fenómeno:

1. *Clasificación.* Esta conducta se asoció rápidamente con un "trastorno", por ejemplo, el "trastorno de personalidad múltiple" (en la actualidad, "trastorno de identidad disociativo", en el que ya no se prevé que los pacientes muestren unas personalidades completamente diferentes).
2. *Las personas.* Son los individuos infelices/ineptos que expresarán esta identidad (o individuos afortunados, en el caso del "genio").
3. *Las instituciones.* Hay clínicas, programas de formación y congresos internacionales que se ocupan del trastorno (como hizo Oprah Winfrey con el "trastorno de personalidad múltiple" —TPM—, convirtiéndose así también en una institución mediática).
4. *Saber.* Tanto el de las instituciones como el saber popular, por ejemplo, la percepción pública de que el TPM está causado por unos abusos sexuales precoces y que lo padece el 5% de la población.
5. *Expertos.* Estos generan el saber, juzgan su validez y lo usan en la práctica. Trabajan en instituciones que garantizan su estatus y ellos asesoran acerca de cómo tratar a las personas que clasifican como pacientes del trastorno.

HACKING también presentó el *efecto bucle*, que alude al modo en que los clasificados responden a sus nuevas identidades. Esto puede adoptar, en algún momento, la forma de oposición; así, los "derechos de los gays" tratan de recuperar el control de las clasificaciones legales que afectan a los homosexuales.

Los mecanismos por los que entran en liza estas clasificaciones socialmente creadas son especialmente relevantes para mis argumentos acerca de los tests

---

capacidades cognitivas. La cuestión es si hay una base sustancial de diferencia (puede "existir" la personalidad múltiple). Yo sostengo que el *modo* en que la sociedad clasifica y etiqueta a las personas desempeña un papel crítico en la configuración de las identidades.

de inteligencia, las inteligencias múltiples y los estilos de aprendizaje (Capítulos II a IV), pues han seguido en gran medida el mismo patrón. HACKING los describe como diez *motores de descubrimiento* que impulsan este proceso: 1) contar; 2) cuantificar; 3) crear normas; 4) correlacionar; 5) medicalizar; 6) biologizar; 7) genetizar; 8) normalizar; 9) burocratizar; 10) reclamar nuestra identidad (pág. 10).

Para dar una idea de cómo funcionan estos motores, utilizo su ejemplo de la obesidad, cuya incidencia ha aumentado espectacularmente en las dos últimas décadas. Ésta se cuantifica primero como un "índice de masa corporal" superior a 30 (*contar, cuantificar*), dándose después unas normas que identifican el peso inferior al normal, el normal, sobrepeso, y obesidad para cada edad (*crear normas*). Después, se correlaciona con la mala salud, por ejemplo, la diabetes. Esto va acompañado de los tratamientos médicos, químicos y quirúrgicos, para reducir el peso (*medicalizar*). Después buscamos causas biológicas, entre otras cosas porque libra de responsabilidad a la persona; así, la obesidad se convierte en un desequilibrio químico en vez de una opción personal. Ello conduce inevitablemente a la búsqueda de la base genética de la obesidad. Al mismo tiempo, se procura ayudar a la persona obesa a volver a la normalidad, en la medida de lo posible, mediante fármacos para reducir el apetito y programas para perder peso (*normalizar*). Con frecuencia, el motor burocrático tiene intenciones positivas, por ejemplo, la reciente introducción en la escuela de programas de control de la obesidad para descubrir a niños y niñas pequeños que ya sean obesos. La resistencia surge cuando el obeso empieza a sentirse perseguido y afirma que la gordura es buena, como el irónico *"Groupe de réflexion sur l'obésité et le surpoids"* (GROS) francés.

Esta secuencia explica algunas clasificaciones educativas clave que han sido generadas a través de procesos de evaluación y sociales similares. Por ejemplo, el desarrollo de los tests de CI siguió precisamente esta trayectoria, hasta el punto incluso de que los primeros aplicadores de tests de CI crearon nuevas técnicas estadísticas (por ejemplo, técnicas de escalamiento y correlacionales) para desarrollar los motores 1-4. Después, el CI se *biologizó* y *genetizó*, dándole un fundamento fisiológico y considerándolo en gran medida heredado. Se incluyó, entonces, en el entorno escolar (motores 8 y 9), por ejemplo, la selección 11+, en el Reino Unido. La oposición surgió con el reconocimiento social de la injusticia de esta forma de selección.

Palabras y significados

Adopto un enfoque elástico de los términos clave. *Evaluación*, *exámenes* y *tests* tienen una cualidad intercambiable, aunque, en la práctica general, se considere que la *evaluación* abarca un conjunto de enfoques para la recogida de pruebas, mientras que *exámenes* se utiliza más para referirse a pruebas de respuestas escritas abiertas en condiciones estandarizadas y *tests* se emplea para referirse a preguntas con múltiples opciones de respuesta que se marcan para corrección mecánica. No obstante, el uso cotidiano en el Reino Unido es inconsistente: tenemos tests del currículum nacional que son exámenes y exámenes

para la obtención de títulos que utilizan tests. A veces, yo uso "test" porque es estilísticamente adecuado, en vez de "evaluación" que probablemente sería más apropiado*.

Se observa una laxitud similar en torno a los usos de *capacidad*, *aptitud* e *inteligencia*. Yo considero intercambiables capacidad y aptitud, si bien "capacidad" es más general y "aptitud", más específica: "ser capaz de beneficiarse de la enseñanza de una asignatura concreta"[5]. En Inglaterra existe el problema de que la selección de hasta el 10% de los ingresos en escuelas especializadas, que son casi todas las escuelas secundarias, puede basarse en los tests de aptitudes, mientras que los tests de capacidad no se permiten a causa de su resonancia emocional del CI. Un responsable de juzgar las solicitudes de selección ha comentado: "una de las dificultades es que la ley utiliza estas dos palabras como si fuesen cosas diferentes y, en realidad, no lo son". Por si esto no fuese ya bastante complicado, las escuelas también incluyen en estos procedimientos el rendimiento antecedente, lo que le lleva a comentar que distinguir entre "aptitud" y "logro" era "la clase de ejercicio a que se dedican los lexicógrafos cuando no tienen bastante que hacer"[6]. Esto sirve para apoyar el argumento de este libro de que los tests de potencial son, en realidad, tests de logro. Pero, como soy crítico con respecto a la forma de entenderse la capacidad y la aptitud (Capítulo II) e interpreto su uso como una forma socialmente respetable de recolocar los tests de inteligencia, no estoy dispuesto a lanzar ninguna clase de rescate conceptual en este punto.

Otros términos que reciben un tratamiento igualmente laxo son *conocimientos* y *destrezas* o *competencias*. El uso general los reúne, abarcándose así todos los resultados de la educación. Sospecho que hay un solapamiento masivo en buena parte de nuestro pensamiento, de manera que "destreza" o "competencia" es el "saber hacer" algo. El uso de "destrezas" parece adecuado para las actividades más orientadas a la ejecución; así las destrezas de danza pueden distinguirse de los conocimientos sobre la danza. *Interpretación* se ha convertido en una palabra problemática, por la forma de colonizarla quienes la contrastan —incluyéndome yo mismo, a veces— con el *aprendizaje para alcanzar el dominio*[7]. Esto le da una connotación negativa (tiene que ver con calificaciones, comparaciones y la demostración de competencia, frente al aprendizaje intrínseco), mientras que el uso normal de la palabra la trata como lo que hacemos, por lo que mi actuación es el origen de la evidencia sobre mí (por ej., la ejecución de un test o la representación de un drama). Hay que prepararse para las ambigüedades.

En este libro aparecen algunos términos técnicos, aunque se utilizan lo menos posible con el fin de que el lector general pueda seguirlo con más facili-

---

* En el original el autor hace mención además a la complejidad añadida de las distintas acepciones que tienen *assessment* y *evaluation* en EE. UU. y el Reino Unido. En castellano es más simple pues ambos términos se traducen como evaluación. (*N. del T.*)

[5] Citado por un funcionario del *Department for Education and Skills* en el *Times Educational Supplement* (TES), 11 de agosto de 2006, pág. 4.

[6] *Times Educational Supplement* (TES), 11 de agosto de 2006, pág. 4.

[7] Esta distinción procede de los trabajos sobre los objetivos de rendimiento, muy citados, de investigadores como DWECK (véase el Capítulo VII).

dad. En vez de salpicar el texto con los términos *validez* y *fiabilidad*, he preferido hablar de finalidad, adecuación a la finalidad y consecuencias de la evaluación, términos que abarcan la mayor parte del mismo terreno. En los Capítulos V y VI, relacionaré éstos con algunas de las ideas más técnicas de "validez" y "fiabilidad".

Un término que quizá deba aclarar más es *aprendizaje*. Aunque, evidentemente, abarca un amplio conjunto de actividades (aprender a andar, aprender historia, aprender a pensar y aprender de memoria), necesita cierta delimitación en relación con el rendimiento educativo. En este libro, aparece frecuencia junto a *eficaz* o *de principios*. Ambos calificativos indican una visión del aprendizaje que lo considera como un proceso de dar sentido, que se incorpora después a lo que ya se conoce. Michael ERAUT ha definido esta visión del aprendizaje como "un cambio significativo de aptitud o de conocimientos", del que excluye "la adquisición de nueva información cuando no contribuye a esos cambios" (pág. 556). Esto no significa que no haya otras maneras de aprender, pero el libro se centra en las formas más profundas de aprendizaje que permiten transferirlo a nuevas situaciones y modificar ideas previas (véanse los Capítulos V y VII).

Contexto y sesgo

Del mismo modo que sostengo que no hay evaluación neutra, tengo que reconocer que no hay escrito neutro. Este libro está informado por una visión particular de lo que es el aprendizaje y de cómo se produce. Considero el aprendizaje como un proceso social y cultural en el que el individuo construye el significado (véase el Capítulo VII).

Esta insistencia en la cultura y el contexto contribuye a explicar la organización del libro, cuyo Capítulo Primero revisa parte de los antecedentes históricos que ahora damos por descontado en el uso de la evaluación. No es simplemente un caso de "empecemos por el principio...". La intención es bosquejar el contexto social en el que se han desarrollado unas formas de evaluación que se dan por descontadas y cuestionar si los supuestos originales todavía se sostienen. Este enfoque se desarrolla aun más en el Capítulo II, en el que afirmo que los tests de inteligencia, tal como los conocemos, fueron desarrollados por personas con un conjunto muy diferente de creencias sociales y políticas, que utilizaron esta forma de evaluación para sus propios fines sociales. Las maneras de cuestionar estas ideas, tanto por oposición directa a las mismas como por enfoques alternativos, como las "inteligencias múltiples" y la "inteligencia emocional" (Capítulo III) reflejan un contexto social cambiante, los contextos económicos y políticos cambiantes que han llevado a la aparición de la "titulitis" (Capítulo V) y al uso de la evaluación con fines de rendición de cuentas (Capítulo VI). Ambas cosas conllevan el riesgo de reducir el aprendizaje a la obtención de resultados y la "evaluación para el aprendizaje" (Capítulo VII) puede interpretarse como un intento de oponerse a ello.

Reconozco que mi propia experiencia me ha llevado a una perspectiva característicamente angloparlante, en la que el Reino Unido y Norteamérica proporcionan la mayor parte de los ejemplos. Es, sin duda, una limitación, aunque confío en que las cuestiones abordadas, por ejemplo, la titulitis, la rendición de cuentas y la inteligencia, tengan eco en otros contextos culturales.

He optado por trabajar con documentos históricos y políticos en vez de con los de los sociólogos que también se han ocupado de estas cuestiones, por ejemplo, Michel FOUCAULT y Jürgen HABERMAS. Esto se debe en parte al complejo mundo de lenguaje en el que se mueven que, con frecuencia, supone que ciertos conceptos útiles no puedan trasladarse con facilidad a otros marcos de referencia. Por ejemplo, FOUCAULT emplea la idea de "vigilancia" como forma de control social, con el examen de una microtecnología especial, combinando "el despliegue de la fuerza y el establecimiento de la verdad"[8]. Sin embargo, ni la "fuerza" ni la "verdad" tienen su significado habitual: "fuerza" señala las relaciones de poder y "verdad" señala (útilmente) la capacidad de la evaluación para definir y clasificar a las personas, muy en la línea de lo que aquí decimos. Mi enfoque ha consistido, por tanto, en tratar de absorber sus preocupaciones, sin tomar su lenguaje. Donde me aparto de estos teóricos es en la medida en que el individuo es capaz de contribuir al cambio, una perspectiva que, a menudo, parece estar ausente en su obra.

También llego a esto a través de mi propia historia educativa. Mi experiencia docente me hizo caer en la cuenta de la importancia del contexto, primero en la Zambia rural y después en el centro de Londres. África me enseñó cuánto damos por sentado, mientras trataba de explicar las educadas sutilezas de *Under the Greenwood Tree*, de Thomas Hardy, para el *Cambridge School Certificate* (*Macbeth* era mucho más fácil). Londres fue un brusco despertar, en el que muchos juzgaban irrelevantes la escuela y los exámenes. La necesidad de dar sentido a esto me llevó a reciclarme y a trabajar como psicólogo educativo, lo que, a su vez, me condujo a estudiar en América. A mi regreso, trabajé como "director de investigación" en uno de los tribunales nacionales de exámenes, lo que me aportó una visión de cómo operan los exámenes y me hizo desconfiar de las afirmaciones demasiado solemnes acerca de sus virtudes y fiabilidad, del mismo modo que las personas que han trabajado en pastelerías desconfían de las tartas. El paso siguiente fue a los organismos gubernativos de evaluación, en los que tuve experiencia de la lógica de los planificadores cuando trataban, con buena intención, de elevar los niveles. Lo que por regla general les faltaba era una idea real de cómo funcionan las escuelas, los maestros y profesores y los alumnos: de nuevo, la importancia crucial del contexto. Escribo ahora como universitario, otro contexto con sus propias formas extrañas de mistificar y "problematizar" algo que puede observarse en este libro.

[8] FOUCAULT, M. (1977): *Discipline and Punishment*, traducido al inglés por Alan SHERIDAN. Londres: Allen Lane (título original: *Surveiller et Punir: Naissance de la prison*; hay traducción castellana: *Vigilar y castigar: nacimiento de la prisión*. Trad: Aurelio GARZÓN DEL CAMPO. Madrid: Siglo XXI de España Editores, S. A., 2009).

CAPÍTULO PRIMERO

Evaluando la evaluación

> Es posible que unos renacuajos inteligentes se reconcilien con los inconvenientes de su situación, pensando que, aunque la mayoría de ellos vivirán y morirán tan sólo como renacuajos, algún día, los más afortunados de la especie se despojarán de sus colas, dilatarán sus bocas y estómagos, saltarán con habilidad a la tierra seca y croarán, dirigiéndose a sus antiguos amigos y hablándoles de las virtudes mediante las que los renacuajos con carácter y capacidad pueden llegar a ser ranas.
>
> (*The Tadpole Philosophy*, R. H. TAWNEY, 1951.)

Ayudar a algunos renacuajos a convertirse en ranas ha sido, desde los exámenes de selección para el funcionariado chino hace mil años, hasta la selectividad para ingreso en la universidad de nuestros días, una de las funciones clave de la evaluación y, a través de los años, quienes fueron seleccionados y llegaron a ocupar puestos relevantes croaron a pleno pulmón acerca del poder de la evaluación para descubrir la capacidad y el mérito.

Sin embargo, las evaluaciones formales han desempeñado también otras funciones históricas: establecer la autenticidad en épocas precientíficas; certificar la competencia laboral a través de los gremios y profesiones; identificar a los alumnos con necesidad de escolarización o clases especiales, y como instrumento de rendición de cuentas para juzgar la eficacia de las instituciones.

La intención de este capítulo es hacer explícitos algunos de los supuestos históricamente arraigados acerca de las evaluaciones formales. En este caso, estas ideas que se dan por descontadas son, en gran medida, producto de la cultura y de la historia británicas, que han influido en otras muchas culturas. Las cuestiones se centran en si todavía se sostienen los fundamentos originales y qué hemos aprendido desde entonces. Por ejemplo, el atractivo de los exámenes ha sido siempre su *justicia* y su promesa de *selección meritocrática*. Lo que nos recuerda la historia es que, aunque, sin duda, eran más justas que las influencias a las que reemplazaban, también reflejaban unos supuestos sociales y de clase acerca del mérito y de la capacidad. Estos supuestos sociales excluyeron automáticamente a las mujeres de los exámenes “abiertos” hasta finales del siglo XIX

y “protegieron” de los exámenes a la mayoría de los niños británicos de clase trabajadora hasta mediados del siglo xx (de manera que, por esta vía, nunca podrían salir del charco). En el Capítulo II, veremos que unos presupuestos culturales semejantes, incluidos los de la superioridad racial, desempeñaron un papel importante en el desarrollo del los tests de inteligencia en Gran Bretaña y en los Estados Unidos.

Cualquier evaluación de los usos e influencias de la evaluación tiene que empezar por su finalidad; sin conocer ésta, no podemos juzgar si la evaluación ha hecho lo que estaba previsto que hiciese. Por eso, empezaré por una clasificación de algunas finalidades clave, antes de pasar a examinar cómo se han expresado históricamente.

Finalidades

Las tres grandes preguntas que hay que hacer con respecto a una evaluación son:

1. ¿Cuál es la finalidad principal de esta evaluación?
2. ¿Se ajusta la forma de la evaluación a la finalidad de la misma?
3. ¿Consigue su objetivo?

Su sencillez es engañosa, dado que en ellas se esconden las principales cuestiones teóricas de la validez y la fiabilidad, y el espectro de las consecuencias imprevistas. La primera pregunta implica que puede haber múltiples fines, a veces opuestos. La “adecuación a la finalidad” se refiere al grado de adecuación de la forma de la evaluación. No queremos que se le otorgue a nadie el permiso de conducir sobre la base exclusiva de la prueba teórica. La tercera pregunta se refiere al impacto de la evaluación. No solo se trata de si hace lo que dice hacer, sino de las consecuencias que tenga para quienes realizan el examen y para otras personas.

Con estas tres preguntas, cuestionaremos algunas prácticas habituales de evaluación. Éstas se sitúan en un contexto histórico, dado que lo que quizá demos por descontado no siempre ha sido así y puede ser una consecuencia de la aceptación acrítica de una herencia cultural. Por ejemplo, en Inglaterra, la tradición de los exámenes escritos nació en las universidades de élite y se extendió a las profesiones y a las escuelas secundarias. ¿Cómo configuró esto lo que examinamos y cómo lo examinamos en la actualidad? ¿Por qué es diferente de la tradición de las “opciónes múltiples” * de los EE.UU. y otros países? Este enfoque histórico ilustra también el hecho de que algunas de nuestras preocupaciones contemporáneas tienen precedentes, por ejemplo, la insistencia en utilizar tests con fines de rendición de cuentas. Pretendemos presentar aquí un contexto en el que podamos comprender mejor dónde nos encontramos ahora y cómo hemos llegado hasta aquí.

---

* En la bibliografía en castellano, suele hablarse de “pruebas de opción múltiple”. No obstante, hemos preferido traducir “opciones múltiples” porque lo que realmente se presenta en estas pruebas son varias opciones de las que escoger una. (*N. del T.*)

La finalidad principal

Esta es, en esencia, una pregunta de "por qué": ¿por qué estamos buscando esta información? Por regla general, no es difícil dar con algunas respuestas, aunque unas sean más pobres que otras. Las encuestas acerca de la satisfacción del cliente o del personal son un buen ejemplo: nos piden que las cumplimentemos aunque no estemos muy seguros de que alguien les vaya a prestar atención, dado que no parece que haya ocurrido nada como consecuencia de las anteriores. Un examen serio de la finalidad puede llevarnos a concluir que tiene menos que ver con descubrir algo para mejorar que con cumplir el requisito de consultar al cliente.

Si hay una única finalidad, la pregunta es relativamente fácil de responder. "¿Cuál es la finalidad principal del examen de conducir?" no es ningún rompecabezas, aunque podríamos profundizar y encontrarle un valor sociológico. La pregunta es más reveladora cuando aparecen múltiples finalidades o alguna cambia. Cuando hay diversas finalidades, su equilibrio fluctúa a menudo. Una imagen útil a este respecto es la de la figura y el fondo. Se parte de la base de que hay varias finalidades, pero puede haber movimientos que pongan alguna en primer plano mientras otra se difumina en el fondo. Todo ello tiene lugar en un marco social.

Tomemos, por ejemplo, los resultados del examen al final de la escolarización obligatoria. La finalidad original era que fuese un medio tanto para obtener un título como para la promoción del alumno. En Inglaterra, como en otros muchos países, unas buenas calificaciones en el examen permitían la promoción al siguiente nivel de estudios o el acceso a la vida laboral. Los lectores británicos de cierta edad recordarán que los periódicos locales publicaban los resultados obtenidos por cada alumno, de manera que los vecinos podían ver cómo se había desenvuelto cada uno, pero había pocos comentarios sobre el rendimiento general de la escuela. En la década de 1990, la exigencia legal de que las escuelas publicaran sus resultados en formatos cada vez más estandarizados, comenzó a acentuar la importancia de la proporción de alumnas y alumnos con cinco calificaciones entre A y C en el GCSE[1]. Desde entonces, esta información administrativa a los padres ha cristalizado en un completo sistema de rendición de cuentas. Esto supone la elaboración de unas tablas de rendimiento de ámbito nacional para clasificar las escuelas y las administraciones educativas locales, basadas en el porcentaje de alumnos y alumnas que obtienen cinco calificaciones entre A* y C, así como el control del cumplimiento de los objetivos nacionales.

Esto conduce a mi "principio de la prepotencia administrativa": *Cuando se multiplican las finalidades de la evaluación, cuanto más administrativa sea la finalidad, más predominante será su papel*. Utilizo aquí "administrativa" para referirme

[1] El *General Certificate of Secondary Education* (GCSE) es el examen que realizan la mayoría de los alumnos y alumnas de 16 años al final de la escolarización obligatoria. Consta de exámenes independientes por asignaturas, que comprenden exámenes escritos y trabajos de clase. Por regla general, los estudiantes hacen ocho pruebas de entre 10 asignaturas. La calificación se basa en una escala que va de A a G (y una U si está suspendido), aunque siempre se ha considerado que las calificaciones A-C son los "auténticos" aprobados, y esto lo confirman las tablas de rendimiento del Gobierno, que se centran en esas calificaciones. La calificación A* ("A estrella") se introdujo para hacer unas distinciones más precisas entre los estudiantes con mejores calificaciones.

a las finalidades de control y de rendición de cuentas. Estas cuestiones son esencialmente sistémicas y predomina el control social. Por tanto, aunque el certificado individual sigue siendo importante para quienes se examinan, lo que en Inglaterra ha pasado a primer plano es el porcentaje de estudiantes que obtienen cinco calificaciones entre A* y C en el GCSE y, como veremos, las escuelas recurren a toda clase de estrategias descaradas para maximizar sus porcentajes, subordinando a veces los valores educativos al cumplimiento de los objetivos. Estos problemas se repiten en todos los sectores que se rigen por objetivos: desde las listas de espera de los hospitales (admitiendo primero a los pacientes fáciles) a las operadoras ferroviarias (disponiendo los horarios para que los trayectos duren más tiempo, de manera que mejore la puntualidad). Volveremos sobre los sistemas de rendición de cuentas en el Capítulo VI.

La finalidad administrativa no ocupa el primer lugar en todas las evaluaciones. Las de competencia ocupacional pueden considerarse primordialmente individuales, aunque la "licencia para ejercer" esté socialmente regulada: por ejemplo, la formación médica ha procurado restringir siempre el número de personas que ingresan en la carrera y se titulan, con el fin de mantener su elevado estatus. Mi certificado de natación de 20 m es esencialmente personal, al menos hasta que el Gobierno decida que todo el mundo debe ser capaz de nadar 20 m.

Algunas evaluaciones pueden tener una finalidad primordialmente profesional. Las realizadas en el aula pueden servir tanto para que el profesor o maestro determine en qué nivel se sitúa el aprendizaje de una clase como para cada alumna o alumno. Se parte de la base de que los fines profesionales redundarán en beneficio de los procesos de enseñanza y aprendizaje más que en el del control burocrático. La evaluación formativa ("evaluación *para* el aprendizaje") es un ejemplo de ello que desarrollaré en el Capítulo VII. Aquí, la premisa es que la única finalidad de las evaluaciones informales implicadas es conducir a nuevos aprendizajes y que cualquier uso administrativo de éstas sería inadecuado. La evaluación de los "estilos de aprendizaje" y la de las "inteligencias múltiples" también pueden considerarse como elementos de una evaluación profesional para contribuir al proceso de enseñanza y aprendizaje.

Para organizar los diversos usos de las evaluaciones, he optado por hacer tres grandes grupos, que reflejan las clasificaciones convencionales:

1. Selección y certificación.
2. Determinación y elevación de los niveles.
3. Evaluación formativa: evaluación *para* el aprendizaje.

Éstas se solapan masivamente, dado que las evaluaciones sirven para diversos fines. Por ejemplo, si la selección en la universidad se basa en un examen escolar, este examen regirá el currículum y la forma de impartirlo. La "determinación de los niveles" comprende tanto lo que se enseña como el nivel de rendimiento que se espera de los estudiantes.

La inclusión de la evaluación formativa como una finalidad aparte también suscita problemas de solapamiento. ¿Por qué no interviene en la elevación de los niveles? La justificación de un tratamiento separado es que el aprendizaje supone algo más que las calificaciones de los exámenes, aunque, a menudo, para los

planificadores, “elevar los niveles” consista simplemente en mejorar las calificaciones (véase el Capítulo VI). Cada uno de estos grandes grupos alberga una serie de finalidades más concretas que se exponen en la Tabla 1.1.

Tabla 1.1. *Fines de la evaluación*

Grupo	Subgrupos	Primario	Capítulos
Selección y titulación	Selección más justa para ingreso y promoción	Individual	1 y 5
	Certificación de la competencia (“licencia para el ejercicio profesional”)	Individual	1 y 5
	Evaluación diagnóstica (necesidades especiales / superdotados / estilos de aprendizaje)	Profesional/ administrativo	1-4
Fijación y elevación de niveles	Control nacional / niveles locales de logro - evaluación	Administrativo	6
	Mejora nacional / niveles locales de logro - rendición de cuentas	Administrativo / profesional	6
	Evaluación de clase (sumativa) - control del progreso / motivación	Profesional / individual	5-7
Formativa	Evaluación de clase (formativa) - mejora del aprendizaje	Individual / profesional	3, 4 y 7

Los orígenes

La intención de esta sección es mostrar que las funciones de selección y la de elevar los niveles de la evaluación tienen solera histórica. También puede ilustrar por qué aceptamos los exámenes como un aspecto natural de la educación que, por tanto, no se cuestiona. De acuerdo con el argumento del libro, las evaluaciones no solo se han utilizado para configurar la identidad individual, sino también para definir el estatus de profesiones y escuelas. Ha sido bastante común la creencia de que los exámenes son necesariamente justos, aunque la mayoría de la población estuviera excluida de ellos, y que pueden revelar la capacidad subyacente. Esta herencia victoriana parece ahora tan evidente de por sí que, con frecuencia, no se cuestiona.

El punto de partida heterodoxo de esta revisión histórica está constituido por la prueba de la autenticidad del folclore y el mito, que aquí adquiere prioridad sobre el puesto de honor que suele reservarse para los exámenes de selección para el funcionariado chino, que han estado en uso durante más de mil

años. En Gran Bretaña, las universidades instauraron los exámenes para mejorar los niveles; fueron introducidos después en las profesiones (por los titulados con buenas calificaciones) y posteriormente se "filtraron" en las escuelas de secundaria y, por último, en las de primaria. Este papel distintivo de las universidades, en comparación con los sistemas organizados por el Estado en el siglo XIX en Francia y en Prusia, explica algunas de las características del currículum y la evaluación que los lectores del Reino Unido dan por sentado (pero que otros cuestionarán). Entre ellas, el currículum escolar y la difícil relación de lo académico con lo profesional. Los debates contemporáneos acerca del impacto del uso de evaluaciones con fines de rendición de cuentas recuerdan los del siglo XIX.

Identidad e inocencia

Este improbable punto de partida es el resultado de aceptar la afirmación del antropólogo Allan HANSON de que el folclore muestra que una de las finalidades de las pruebas en las comunidades precientíficas era establecer la autenticidad y la inocencia. Sería esta una digresión innecesaria si no fuese porque esta finalidad sigue vigente en forma de detectores de mentiras, tests aleatorios de drogas y tests de personalidad. Aunque HANSON dedica a estos test la mayor parte de su obra *Testing Testing: Social Consequences of the Examined Life*, solo resumiré aquí brevemente esta línea de razonamiento.

Los tests de autenticidad incluían la confirmación de la *identidad* y la determinación del *carácter*, por ejemplo, el rey Arturo sacando la espada de la roca y el cuento de la princesa y el guisante (dada su sensibilidad real, se dio cuenta de que había un guisante que percibió a través del colchón, confirmando su verdadera identidad). La prueba del agua también establecía la identidad. Que una persona sea merecedora de confianza es la base de las historias sobre la resolución de enigmas y actos heroicos. La prueba de la verdadera maternidad, ideada por el rey Salomón y basada en la observación de las reacciones a la sugerencia de partir al niño por la mitad, es típica de éstas.

Las pruebas de *sinceridad, culpabilidad e inocencia* se utilizaron con frecuencia para ayudar a decidir en pleitos judiciales. Un enfoque consistía en el "juicio mediante duelo", en el que la victoria indicaría quién tenía razón. Se presumía que Dios estaba del lado del bueno. El relato bíblico de David y Goliat representaba una victoria tan asombrosa del desvalido que era una clara prueba de la ayuda divina (si Goliat hubiese ganado, la explicación podría haber sido muy diferente: después de todo, era mucho más grande e iba mejor armado). El otro enfoque consistía en "nivelar el terreno de juego", una aspiración de la evaluación a la que volvemos una y otra vez, de manera que los implicados no puedan prever quién ganará, considerándose la victoria, por tanto, expresión del juicio divino. Un extremo surrealista de esto era la ley alemana que "nivelaba" el combate entre un hombre y una mujer:

> Las posibilidades... se ajustaban enterrando al hombre hasta la cintura, atando su mano izquierda a su espalda y armándolo únicamente con una maza, mientras que

su hermosa oponente podía utilizar sus miembros y se le proporcionaba una pesada piedra bien atada con un trozo de tela[2].

Sin embargo, el "nivelado" entre combatientes de distintas clases sociales era muy diferente, otro tema al que volveremos. Por ejemplo, en Francia, si un noble se enfrentaba a un plebeyo en un combate judicial, el primero tenía el derecho de luchar a caballo con armas de caballero, mientras que el plebeyo tenía que hacerlo a pie, con un escudo y un bastón. Ciertamente, necesitaría tener a Dios a su lado.

Para muchos, el juicio por ordalía es el precursor de los exámenes o las entrevistas. Históricamente, algunos de los instrumentos de evaluación preferidos eran el agua, fría y caliente, y el hierro al rojo. La forma más conocida de este juicio era la "prueba del agua", utilizada en el siglo XVII, cuando actuaba en Inglaterra el famoso cazador de brujas Matthew Hopkins. Su técnica consistía en atar el dedo pulgar de la mano derecha de la sospechosa al pulgar del pie izquierdo, sumergiéndola a continuación en el agua con una cuerda atada alrededor de la cintura. La prueba se repetía tres veces y, si la persona flotaba, era prueba de brujería. La lógica conduce a un callejón sin salida: si fuese inocente, Dios lo indicaría cuando se hundiese. La culpable flota porque la naturaleza pura del agua no recibe a la embustera. Si necesitamos algún recordatorio de la construcción social de una interpretación así, en el sudoeste de Alemania, en el mismo período, era la inocente la que flotaba y la culpable la que se hundía. Dios se mueve por caminos misteriosos... y regionales.

Antes de distanciarnos demasiado rápidamente de esta lógica sospechosa, merece la pena reflexionar acerca de si, en nuestros días, tenemos equivalentes seculares más "leves". HANSON sostiene que el uso actual de los detectores de mentiras comparte muchas de estas características. En esta ocasión, es la ciencia, en vez de Dios, la que revela qué individuos culpables quieren esconderse. No lo hace directamente: la lectura del polígrafo supone una cadena causal de evasivas que conduce a la ansiedad, que lleva a una respuesta fisiológica mensurable. Hay una confianza similar en que el detector de mentiras dice la verdad.

Un buen ejemplo de ficción de estas particularidades se da en la comedia de televisión *Mujeres desesperadas*, cuando Bree insiste en que le hagan la prueba delante de sus hijos, que han sospechado de su madre después de que su marido Rex muriera repentinamente por causas naturales. Cuando le preguntan si ella mató a su marido, la línea permanece plana (inocente). Sin embargo, cuando el interrogador le pregunta después si está enamorada de otro hombre, el polígrafo oscila fuertemente, mostrando una serie de picos, a pesar de sus protestas. Bree, que no es consciente de sus sentimientos en la mayoría de los casos, acepta que debe de estarlo y, en consecuencia, acepta las insinuaciones de George. George, que mató a Rex, se somete al detector de mentiras y pasa la prueba; nos damos cuenta así de que es un psicópata que no muestra emociones ni sentimientos de culpa.

Este ejemplo no requiere demasiadas interpretaciones y solo pretende ilustrar los mismos principios que rigen las pruebas precientíficas de autenticidad:

[2] LEA, H. C. (1968 [1870], pág. 120).

podemos alcanzar la verdad yendo más allá de las afirmaciones y la conducta del individuo. El engaño nos estimula, en vez de desalentarnos: una cadena causal que supone una especial relación cuerpo-mente. Representa también una clara relación de poder, en la que el interrogador es un operador que actúa en nombre de la sociedad y el sistema de justicia.

En este libro, no volveremos a ocuparnos del uso de las pruebas de autenticidad, aunque formen parte de la vida moderna [3]. No obstante, comparten algunas características de las formas de evaluación que consideraremos:

- Implican el ejercicio del poder. Éste reside en quienes administran las pruebas, pero éstos, a su vez, representan el sistema social del que forman parte.
- Se ocupan de la realidad socialmente construida, en vez de alguna realidad existente de forma independiente (por ejemplo, la nobleza).
- Contribuyen a formar los constructos que miden. Incluso pueden crear lo que dicen que miden (por ejemplo, las brujas), argumento que desarrollaremos en relación con las pruebas de inteligencia.

Selección por el mérito

Éste es el punto de partida histórico convencional. La finalidad de las evaluaciones formales escritas y prácticas era seleccionar a personas, basándose en el mérito en vez de en la cuna. Éste sigue siendo uno de los atractivos permanentes de las pruebas. Los honores históricos les corresponden a las pruebas de selección de los funcionarios chinos [4]. Aunque ya en la época de la dinastía Chou (*c.* 1122-256 a.C.) había pruebas para descubrir a las personas de talento de entre la gente corriente, la dinastía Sung (960-1279 d.C.) abrió los exámenes a casi todos los varones y lo convirtió en el pasaporte hacia el poder y el prestigio. Estaban excluidos los esclavos, los jornaleros, los actores, los músicos y, por supuesto, las mujeres (las limitaciones acerca de quienes pudieran presentarse a unos exámenes "justos" es un tema recurrente). La dinastía Ming (1368-1662) dio a los exámenes una forma que sobrevivió hasta principios del siglo XX.

Las condiciones de aquellos exámenes hacen que los nuestros parezcan banales. Eran muy estrictas: encerraban a los candidatos en celdas individuales durante tres días, para que escribieran comentarios sobre clásicos confucianos, compusieran poesía y escribieran ensayos sobre historia, política y sucesos de actualidad. Después, se copiaban los textos de sus exámenes, para preservar el anonimato, y eran corregidos dos veces. Los que superaban estas pruebas, que eran entre el 1 y el 10%, pasaban a un nuevo examen preparatorio, que solía

[3] A este respecto, es relevante la idea de *vigilancia* de FOUCAULT, que se basa en la imagen del "panóptico" de Bentham, una estructura arquitectónica que permite una vigilancia de 360º. La vigilancia mediante cámaras de TV en circuito cerrado (CCTV) (más de 300 veces al día para quienes vivimos o trabajamos en el centro de Londres); las bases de datos de ADN, y el seguimiento mediante las tarjetas de crédito y las tarjetas de identidad con datos biométricos contribuyen a alimentar la sensación de control social (véase: HANSON, 1994, capítulos 4 y 10).

[4] Véase un tratamiento más completo en: HANSON, 1994, capítulo 7, en el que se basa esta sección.

aprobar la mitad que tenían luego que hacer otro examen, que se celebraba cada tres años, en la capital de la provincia. El éxito en este examen que, de nuevo, alcanzaban menos del 10% de los presentados, suponía el paso al examen metropolitano, que se celebraba también cada tres años, en la capital. Quienes aprobaban este examen acudían al palacio para el examen final, que era llevado a cabo por el Emperador. En este caso, el aprobado significaba un puesto en la administración o pasar a un nuevo período de formación. Los puestos menores se destinaban a quienes hubiesen aprobado los niveles provincial y metropolitano, aunque estos tenían que volver a examinarse cada tres años.

Este es un contexto ideal para presentar la idea de los *exámenes para nota*. La premisa que subyace es que se trata de una prueba que tiene unas consecuencias importantes para algunas o todas las partes implicadas, en este caso, el examinando. Como veremos en el Capítulo VI, la rendición de cuentas puede significar que las mayores consecuencias sean para el maestro o profesor, el centro y la administración educativa local, en vez de para el estudiante. Las pruebas para nota siempre conllevan el riesgo de que la persona o personas para quienes las consecuencias sean mas importantes traten de mejorar sus resultados por todos los medios. Los exámenes de los funcionarios chinos no eran una excepción. A pesar de todas las precauciones de seguridad, que incluía el cacheo de los candidatos antes de encerrarlos en sus celdas, había un comercio de "chuletas" y el paso de las mismas a las celdas no era una novedad. El énfasis puesto en verter los conocimientos y dogmas recibidos, en vez de buscar la creatividad, fomentaba aquellas prácticas ilícitas. Esto fue lo que atrofió el proceso de selección, aunque de manera muy lenta: el examen sobrevivió 500 años.

Fueran cuales fueran las limitaciones de este sistema, una característica positiva sobresale por encima de todas las demás: se estima que hasta el 60% de los candidatos que superaron el proceso provenían de familias que no formaban parte de la élite administrativa. Sería interesante comparar esto con la composición social de quienes aprueban en la actualidad las oposiciones de ingreso a los cuerpos británicos de funcionarios, que también pretenden realizar una selección por mérito.

En Gran Bretaña, fueron los victorianos quienes asumieron con pasión los exámenes escritos. En una época de reforma y expansión rápidas de la industria y el imperio, eran la solución ideal al problema del reclutamiento del personal prescindiendo de las influencias. El ingenuo entusiasmo decimonónico por los exámenes se basaba, según Gillian SUTHERLAND, en tres méritos observados:

- los exámenes formales se consideraban la antítesis de la corrupción y del egoísmo. Sus apelaciones a la neutralidad se consideraban bienes positivos, "reemplazando principalmente la búsqueda rastrera de una plaza por la confianza en sí mismo" (DALE, 1875);
- se consideraba que los exámenes eran pruebas más que logros o destrezas; se percibían como instrumentos para acceder a las capacidades básicas;
- la capacidad se equiparaba con el mérito y el talento, con la virtud (1992, pág. 3).

Estas premisas nos ayudan a dar algún sentido a la falta de "adecuación a la finalidad" de lo que se examinaba (por regla general, latín, griego y matemáticas) con respecto a los puestos para los que se seleccionaba a los candidatos, por ejemplo, el funcionariado de la India. La "adecuación a la finalidad" se decidía considerando estos exámenes como una prueba de personalidad y capacidad. Quienes los hacían bien demostraban tanto la diligencia como la capacidad convenientes para sus funciones administrativas. (La *American Civil Service Act* de 1883, que trataba de seleccionar al 10% de los funcionarios civiles mediante examen, hacía hincapié en las destrezas relacionadas con el trabajo.) Cuando el historiador Lord MACAULAY habló en la Cámara de los Comunes en 1833 a favor de introducir exámenes para seleccionar a los candidatos para el funcionariado en la India, dijo:

> Miremos todas las profesiones y condiciones sociales y veamos si no es cierto que quienes alcanzan una elevada distinción en el mundo han sido hombres que, por regla general, se han distinguido en su carrera académica... Cualesquiera que sean las lenguas, cualesquiera las ciencias que se acostumbren a enseñar en cualquier época o país, las personas que acabarán siendo más competentes en aquellas lenguas y aquellas ciencias serán, en general, la flor y nata de la juventud, las más perspicaces, las más trabajadoras, las más deseosas de distinciones honorables.
>
> (1898, págs. xi, 571-573.)

Así, lo que se estudiara carecía de importancia. MACAULAY utiliza como ejemplo el aprendizaje del *cheroki*, en vez del griego; lo importante era hacerlo mejor que otros. Y, si eras mejor en los exámenes, serías mejor en tu trabajo y mejor persona (el mismo MACAULAY fue becario por oposición del *Trinity College* de Cambridge, creo haber oído un croar de rana). Lo que no se decía ni se reconocía es que estos también eran los más privilegiados en cuanto a oportunidades y preparación, por lo que el "mérito" reflejaba el sistema de clases. La necesidad de ampliar la base profesional de clase media suponía elevar cuidadosamente a otros en la escala educativa mediante un sistema creciente de becas escolares y universitarias. Estas becas se basaban, precisamente, en oposiciones.

El científico Thomas Huxley recogió gráficamente este pensamiento en 1871:

> Ningún sistema educativo del país sería merecedor del nombre de "sistema nacional" ni cumpliría los grandes objetivos de la educación si no estableciera una gran escala educativa, cuya base estuviera en el arroyo y el peldaño superior en la universidad y por la que todo niño que tuviera la fuerza suficiente para subir alcanzara, empleando esa fuerza, el lugar por él buscado.
>
> (SUTHERLAND, 1996, pág. 16.)

Competencia y título

Históricamente, el uso de la evaluación para certificar la competencia ocupacional es anterior y más rico que el uso para la selección educativa. En la mayoría de las sociedades, la base de los gremios era la idea del aprendizaje, mediante el cual el principiante aprende las destrezas necesarias trabajando al

lado de un maestro (la raíz latina de *assessment** es “assare”, que significa “sentarse con”). Resultaba más evidente que la evaluación era adecuada a la finalidad, porque los aprendices tenían que demostrar el oficio que querían ejercer y las destrezas personales necesarias para ello.

No obstante, en Gran Bretaña, durante la era victoriana, en muchas ocupaciones estas prácticas dieron lugar a una selección más formal y a una titulación, haciéndose hincapié en las cualificaciones educativas. Las profesiones se definían por lo que el sociólogo Randall COLLINS llama “cierre del mercado con el honor de un elevado estatus ocupacional” (pág. 36). En Gran Bretaña, se adelantó la Medicina, compitiendo 21 asociaciones por la concesión de las licencias. En 1858, la legislación que estableció el *General Medical Council*** las fusionó. A la Medicina le siguieron la Contabilidad, la Ingeniería, la Arquitectura y el Derecho, que impusieron sus propios exámenes para la titulación. Con el tiempo, las cualificaciones universitarias fueron reconocidas como aptas para el ingreso en la profesión. En el decenio de 1920, los exámenes se habían convertido en una necesidad moderna para las ocupaciones que aspiraran a un estatus profesional.

Gillian SUTHERLAND nos ha dejado una descripción más detallada de este proceso. Así resume estas creencias sobre los exámenes:

> El mecanismo del examen se aprovechó pronto como medio para el cierre y el control... El examen formal era el instrumento soberano para el descubrimiento del talento. Este estaba a un pasito del evangelio de la meritocracia, de la equiparación del talento con la virtud. La virtud, a su vez, se parecía mucho a las cualidades que se adscribían al caballero inglés... Así, el uso que hacían los grupos profesionales y las asociaciones que otorgaban títulos de las estructuras en desarrollo de la educación formal en la sociedad en general y, en particular, de los exámenes, no solo les permitió asegurarse del cierre del mercado, sino también encerrarse en un discurso y una ideología de gran poder y resonancia, beneficiarse de ellos y promoverlos.
>
> (2001, pág. 62.)

Estas ideas están relacionadas con la posibilidad de que las evaluaciones creen lo que dicen medir. Una profesión puede definirse mediante sus evaluaciones para la obtención del título, sobre todo cuando los juicios sobre la competencia quedan envueltos en el secreto profesional.

Patricia BROADFOOT ha observado que una de las consecuencias que este movimiento tuvo para la educación fue que, dado que los exámenes se asociaban con profesiones de elevado estatus, el modelo de las pruebas escritas, teóricas, quedó investido de un estatus similar. Como veremos en nuestro comentario de la “titulitis” (Capítulo V), esto también llevó a la enseñanza a formar parte del proceso cualificador para la formación profesional. Las funciones de las escuelas y de la enseñanza, inevitablemente menos prácticas y más teóricas, permanecen en tensión en muchos sistemas, como ocurre con la división entre itinerarios académicos y profesionales. En países como Alemania, esta separación se refleja en el sistema escolar de secundaria, mientras que, en Australia, es probable que se considere que el aprendizaje profesional forma parte de la educación de todos los

* La palabra inglesa *assessment* significa “evaluación”, “valoración”, “estimación”. (*N. del T.*)
** “Consejo Médico General”. (*N. del T.*)

alumnos y alumnas. En Gran Bretaña, el legado de la estructura de clases implica que los títulos de formación profesional se consideren a menudo destinados a quienes "no son aptos" para los académicos. En consecuencia, las ocupaciones de elevado estatus, como la Medicina y el Derecho, siguen considerándose, en consonancia con sus historias, disciplinas académicas, en vez de profesionales, aunque la formación médica sea en gran medida práctica y ocupacional.

Al lado de esta tradición profesional estaban los títulos ocupacionales creados en Gran Bretaña en el siglo XIX por instituciones como el *Royal College of Art* y *City & Guilds*. También en ellas se utilizaron cada vez más los exámenes, aunque el elemento práctico continuó siendo fundamental para las titulaciones. La "adecuación a la finalidad" y la competencia siguen siendo importantes en la evaluación ocupacional. Los títulos clásicos pueden seguir considerándose como una buena formación para el funcionariado y los títulos de Física, para la banca de inversiones (significa que los candidatos "saben de números"), pero todos queremos que nuestros pilotos de aerolíneas sean extremadamente competentes en algunas destrezas muy específicas, en especial las de despegar y aterrizar.

Evaluación diagnóstica y educación especial

La selección descrita hasta ahora tenía como fin descubrir al más capaz. Las reformas educativas del siglo XIX en las naciones industriales incluían la extensión de la educación a proporciones cada vez mayores de niños y niñas. Desde 1880, la educación elemental era obligatoria para los hijos de clase trabajadora en Inglaterra y Gales. A consecuencia de ello, se descubrió que algunos niños no eran capaces de seguir una escolarización normal. Los comienzos del siglo XX contemplaron el desarrollo de tests diagnósticos para identificar a esos alumnos.

Al psicólogo francés Alfred BINET se le reconoce como el autor del primero de esos tests. Trabajó con las autoridades educativas de París para elaborar un medio de identificar a los alumnos que necesitaran una educación especial y creó una serie de pruebas entre 1905 y 1908. Su gran acierto, que ahora damos por descontado, fue observar el funcionamiento cognitivo que se esperaba en clase. Esto se apartaba radicalmente de las evaluaciones diagnósticas típicas de la época, que utilizaban el tamaño del cráneo, las características físicas ("stigmata") y la percepción sensorial (véase el Capítulo II). BINET y su colaborador, Theodore SIMON, habían perdido la confianza en esos enfoques y elaboraron una batería de tests que se pusieron a prueba con centenares de niños de París. Se calibraron por edad, de manera que la "edad mental" podía compararse con la edad cronológica. Los tests estaban formados por un subtest espacial, uno verbal y otro numérico, y generaban perfiles individuales. La finalidad que todo esto tenía para BINET era educativa: quería descubrir a los alumnos que necesitaran una atención especial para aprender.

Esto constituyó el fundamento de la evaluación diagnóstica individualizada, ampliamente utilizada aún por la mayoría de los psicólogos educativos. Aunque los tests se hayan hecho más sofisticados, el fundamento subyacente es el mismo. Las finalidades actuales son tanto administrativas como profesiona-

les, dado que la evaluación psicológica forma parte a menudo de un proceso de decisión para la asignación de recursos o la modificación de los centros escolares.

No obstante, como veremos en el Capítulo II, la obra de BINET también sentó las bases de un movimiento de administración de tests mentales que perseguía unos fines muy diferentes y a menudo siniestros. Fueron los psicólogos estadounidenses y británicos quienes vieron la posibilidad de crear tests para administrarlos en grupo que podrían permitir la clasificación de grandes grupos por sus niveles relativos de capacidad. Entre estos usos, estuvieron: la selección militar en la I Guerra Mundial; la restricción de la inmigración, y la selección para la escolarización. Esta tradición generó los tests de CI que, en Gran Bretaña y en otros países, se convirtió en la base de la selección para la enseñanza secundaria. La divergencia fundamental con respecto a la postura de BINET era el supuesto de la existencia de unas capacidades mentales innatas que la educación no podía cambiar. Estas suposiciones eran de carácter social (y no descubrimientos científicos) y daba continuidad al pensamiento sobre las clases y la superioridad racial que legitimaba las creencias acerca de la superioridad natural de los varones anglosajones de clase alta.

Establecimiento y elevación de los niveles

Una finalidad clave de la evaluación, sobre todo en educación, ha sido establecer y elevar los niveles de aprendizaje. Esta es ahora una creencia virtualmente universal: es difícil hallar un país que no esté utilizando la retórica de la necesidad de la evaluación para elevar los niveles en respuesta a los retos de la globalización. Sus orígenes en Inglaterra se remontan a la introducción de los exámenes escritos en Cambridge y Oxford, en el siglo XIX. Desde aquí, se extendieron a la función pública, a las profesiones y a la escuela secundaria, convirtiéndose en un artículo de fe victoriana. Esto influyó en la forma de llevar a cabo la evaluación, antes y ahora. El papel más directo desempeñado por el Gobierno central, tanto en Prusia como en Francia, produjo sistemas diferentes, como ocurrió con los desarrollos habidos en EE.UU.

La premisa es que la evaluación indicará lo que se ha aprendido y el nivel de comprensión y de destrezas necesario. Los exámenes escritos formales comenzaron en las universidades, pasando después a las escuelas de secundaria y a las de primaria. Aunque debatir oralmente había sido el principal medio de evaluación universitaria en las universidades medievales y del Renacimiento, ésta comenzó a ceder paso a los exámenes escritos en los siglos XVIII y XIX. En 1795, el *St. John's College*, de Cambridge, estaba tan preocupado por el rendimiento de sus estudiantes que introdujo unas pruebas internas que tenían que pasar todos los estudiantes dos veces al año; cuatro años antes, los estudiantes de Yale se habían negado a someterse a otros exámenes que no fueran el del momento de la graduación[5].

[5] Esta evidencia está tomada de HANSON (1994), capítulo 7; SUTHERLAND (1996) y BROADFOOT (1979).

Podemos hacernos una idea de por qué se impuso una evaluación más formal gracias a los indignados comentarios de Vicessimus Knox en 1778, sobre los exámenes finales de Oxford:

> Como ni el funcionario ni nadie más suele entrar en la sala (porque se considera muy impropio), los examinadores (normalmente, tres M.A. elegidos por el candidato) y los candidatos conversan a menudo sobre la última borrachera o sobre caballos, leen el periódico, o una novela, o se distraen todo lo que pueden y de cualquier manera hasta que el reloj da las once, momento en el que todos descienden y el máster firma el "testimonium".
>
> (BROADFOOT, 1979, pág. 29.)

BROADFOOT sigue comentando: "En realidad, la residencia durante cuatro años era la única cualificación para el grado; no la formación, inadecuada para una élite para la que las cualificaciones eran casi por completo sociales" (Oxford todavía otorga un MA* a sus graduados después de cinco años, sin ningún trabajo adicional y por una pequeña tasa). En 1852, los comisarios reales que investigaron los asuntos de la *Oxford University* comentaban: "para hacer que un sistema de exámenes sea eficaz, es indispensable que existan el riesgo de rechazo para los candidatos inferiores y honrosas distinciones y recompensas importantes para los capaces y diligentes" [6]. Esta lógica relaciona los niveles y la competición. Los exámenes pueden establecer niveles mínimos y, al mismo tiempo, fomentar la competición, que elevará los niveles. Esta competición adquirió una importancia enorme en Cambridge, particularmente para quien resultara ser el mejor estudiante, el "senior wrangler"** (incluso el personal de la casa hacía apuestas), mientras que un gran cucharón de madera, bajado por los estudiantes de la tribuna, esperaba al estudiante que hubiera tenido las calificaciones más bajas [7].

Este interés por la introducción de los exámenes en Oxford y en Cambridge es deliberado, dado que, a partir de esta experiencia, se establecieron los exámenes para el funcionariado y los de las profesiones, introducidos por quienes habían aprobado en aquellas universidades. Al extenderse fuera de las universidades, la práctica de los exámenes también llegó a las escuelas de secundaria. Estos desarrollos fueron imitados también en América, con la introducción de los exámenes en Yale y en Harvard a principios del siglo XIX, seguidos por los exámenes escritos en las escuelas secundarias de Boston ya en 1845 (uno de los pasos importantes dados en este proceso fue la comprobación de la importancia de que todos los estudiantes hicieran el examen al mismo tiempo. Con anterioridad, los administradores iban a toda velocidad de escuela en escuela, tratando de impedir que la información acerca de las preguntas llegara antes que ellos).

---

* *Master of Arts*: "Máster en Artes". (*N. del T.*)

[6] SUTHERLAND (2001), pág. 52.

** En la Universidad de Cambridge, *wrangler* (literalmente: "vaquero") era el estudiante que terminara el tercer curso de Matemáticas con *first-class honours* (matrícula de honor). Hasta 1909, cuando dejaron de publicarse las listas de alumnos por orden de calificaciones, el estudiante que obtuviera la máxima puntuación era nombrado *senior wrangler*. (*N. del T.*)

[7] Véase una descripción más completa en: STRAY (2001).

Los exámenes en las escuelas de secundaria

El contexto social de la extensión de los exámenes a las escuelas de secundaria era el punto de vista victoriano de que cada una de las tres grandes agrupaciones de clase —alta, media y baja— debía tener sus propias instituciones independientes, porque cada clase tenía sus propias necesidades educativas. La sorpresa es que se consideraba que la educación de la clase media era la más necesitada. Esto se debía en parte a que las necesidades de las clases bajas se consideraban mínimas: cierta educación elemental, simplemente, que satisfacían cada vez más la oferta del Estado y la de la Iglesia. Como las clases altas no se mezclaban con las clases medias y éstas, a su vez, no lo hacían con la baja, cada clase tenía que disponer de su propia provisión educativa. Ésta adoptó la forma típica de unas escuelas privadas de baja calidad. El modo de mejorar los niveles fue objeto de un debate político clave. Para algunos, incluyendo a Matthew ARNOLD, era preferible un sistema de inspección configurado como el prusiano; sin embargo, este se rechazó por engorroso y caro.

Por eso, se pensó que la solución para elevar los niveles eran los exámenes y, ¿qué mejor que encargárselos a Oxford y a Cambridge? Cada universidad estableció un sistema de "exámenes locales" para las escuelas que admitieran alumnos hasta la edad de 17 años. John ROACH presenta una detallada descripción de estos desarrollos, basándose en informes de los examinadores. Para nuestros fines, hay algunos temas recurrentes: las mejoras en el trabajo de las escuelas que han acarreado los exámenes; las diferencias entre escuelas; las diferencias de rendimiento según el género (los niños eran mejores en matemáticas y las niñas, en lenguas y expresión escrita), y la preocupación por "el carácter excesivamente rutinario de la enseñanza y el predominio del aprendizaje memorístico sin que importe mucho la relevancia o el pensamiento independiente" (pág. 155). No parece que los examinadores se preguntaran en qué medida era esto una consecuencia de los exámenes escritos, en especial de la previsibilidad de sus propias preguntas.

Este enfoque se convirtió en el modelo para otros tribunales de exámenes de universidades, que permitieron acceder a los mismos a un conjunto más amplio de estudiantes. Los exámenes continuaban a finales del siglo XX, y muchos ingleses tendrán un montón de certificados de la *University of London* o del *Joint Matriculation Board* (que representaba a las universidades del norte). Sus fusiones y control posteriores, dirigidos por el Gobierno, constituyen una historia aparte de un cambio orientado a la rendición de cuentas.

El motivo de esta incursión histórica selectiva es recordarnos que las evaluaciones que ahora predominan en la escuela son descendientes directas de los intentos victorianos de mejorar la escuela utilizando los exámenes para controlar tanto la enseñanza como el currículum. John WHITE (2004) sostiene que, durante los últimos 100 años, ha habido pocos cambios fundamentales en cuanto a la influencia de la universidad en el currículum. Aunque hoy día el lenguaje pueda haber cambiado, la finalidad subyacente no.

Lo que ha cambiado es la escala de los exámenes. Un elemento menos visible de la difusión de los exámenes de secundaria era que la inmensa mayoría de la población escolar era excluida. Si una de las afirmaciones básicas de este libro

es que los exámenes configuran la forma de vernos a nosotros mismos, esta situación podría considerarse como algo bueno. Sin embargo, igual que suspender el 11+* u otros tests de selección de secundaria, el hecho de quedar excluido de los exámenes no es algo neutro. El mensaje que encierra es que esos estudiantes no son lo bastante buenos para examinarse, una forma ideal de prolongar la estratificación social. Un ejemplo revelador de lo que decimos fue la elevación en 1947 de la edad de escolarización obligatoria en Inglaterra a los 15 años. Esto implicaba que los antiguos alumnos de las *secondary-modern schools*** —creadas para los estudiantes que no hubiesen conseguido ingresar en las selectivas *grammar* y *technical schools*— que hubieran abandonado la escuela a los 14 años podían volver como estudiantes para obtener el *School Certificate*. Esta oportunidad fue rápidamente eliminada mediante la *Circular 103* del Gobierno, que prohibía a las escuelas que no fuesen *grammar schools* admitir al examen para el *School Certificate* a alumnos que no tuvieran 17 años. No hace falta ser un sociólogo del "capital cultural" profundamente convencido para sospechar que esto tenía menos que ver con la bondad para con los menos afortunados que con preservar las ventajas de los ya aventajados. La mayoría de los renacuajos no abandonan la charca.

Rendición de cuentas: Pago por resultados

En la actualidad, la rendición de cuentas es quizá la principal finalidad de los tests y exámenes externos. La lógica es que, si algo está funcionando, será mensurable mediante indicadores que registren la mejora. Los resultados de la evaluación son una medida obvia: si mejoran, los niveles estarán mejorando. Evaluaremos esta lógica en el Capítulo VI; esta sección solo pretende mostrar que este enfoque no es nuevo.

Un ejemplo clásico es el plan de *pago por resultados* incluido por Robert LOWE en su "código revisado" (ley parlamentaria) de 1862. Era una época de creciente demanda de escolarización primaria y creciente gasto del Gobierno en ella. ¿Cómo podría garantizarse el valor por el dinero, sobre todo ante informes de maestros que hacían caso omiso de los alumnos más flojos para concentrarse en los más brillantes? La solución de Lowe consistió en introducir un sistema de ayudas para las escuelas elementales, de manera que la mayor parte del dine-

* En las escuelas inglesas, el examen *11+* o *Eleven plus* era una prueba de acceso a secundaria que se administraba a algunos alumnos del último curso de primaria. En la actualidad solo sigue vigente en algunos condados y municipios ingleses. (*N. del T.*)

** En 1944, se promulgó en el Reino Unido la *Butler Education Act* que implantaba en Inglaterra y Gales un sistema tripartito de educación secundaria, que quedaría profundamente modificado en la década de 1970. De acuerdo con esa ley, los centros de secundaria serían de tres tipos: el primero, la *grammar school*, se destinaba al 25% de los alumnos presentados al examen 11+ que obtuvieran las mejores calificaciones, a los que se prepararía fundamentalmente para seguir una formación académica. El segundo tipo, la *secondary modern school*, se destinaba a la mayoría de los alumnos de secundaria, cuyas puntuaciones en el 11+ quedaran por debajo de las del primer 25% de alumnos. Por último, el tercer tipo, la *technical school* se pensó para ofrecer una formación técnica, orientada a la vida laboral. En realidad, se construyeron muy pocas, por lo que tuvieron poco relieve en la educación inglesa. (*N. del T.*)

ro se distribuía de acuerdo con el resultado de los niños en los exámenes de lectura, escritura y aritmética. Los "niveles" del examen estaban relacionados con la edad y ningún alumno podía volver a presentarse al mismo nivel. Los exámenes fueron dirigidos por la inspección del Gobierno que, en general, se oponía a lo que se consideraba una extralimitación de sus funciones.

La defensa que hizo Lowe de este enfoque, que sobrevivió durante 30 años y, en principio, redujo el gasto del Gobierno, tiene cierto sabor contemporáneo:

> ¿Cuál es el objeto de la inspección? ¿Consiste simplemente en hacer las cosas agradables, dar a las escuelas tanto como se pueda sacar del erario público, con independencia de su eficiencia, o quieren decir que nuestras subvenciones no solo deben ser ayudas, subsidios y regalos, sino que deben dar buenos frutos?... ¿Están ustedes a favor de la eficiencia o del subsidio? ¿Hay que beneficiar a una escuela porque es mala y, por tanto, pobre, o porque es una buena escuela y, en consecuencia, eficiente y en buenas circunstancias?
>
> (*Hansard's Parliamentary Debates*, 13 de febrero de 1862, pág. 205.)

Verdaderamente, el sistema era de "pago por resultados", porque parte del salario del maestro o profesor se basaba a menudo en lo que obtenían las escuelas, mediante financiación gubernativa, del éxito de los exámenes anuales. Cuando Edward HOLMES, ex inspector principal, reflexionaba sobre esto en 1911, hacía una crítica mordaz del impacto del plan sobre la calidad del aprendizaje:

> Los niños... eran instruídos en los contenidos de esos libros hasta que los sabían casi de memoria. En aritmética, hacían sumas abstractas, obedeciendo reglas formales, día tras día y mes tras mes; y les sugerían diversos trucos y medios que se esperaba que les permitieran saber mediante qué reglas concretas tenían que responder a las diversas preguntas de la ficha de aritmética... No se prestaba ninguna atención, excepto en una pequeña minoría de escuelas, a la preparación real del niño, a promover su crecimiento mental (y de otro tipo). La única preocupación del maestro era hacer que, de un modo u otro, aprobara el examen anual... Dejar que un niño descubriera algo por sí mismo, resolviera algo por su cuenta, se habría considerado como una prueba de incapacidad, por no decir locura, del maestro y hubiera llevado a unos resultados que, desde el punto de vista del "porcentaje", probablemente hubiesen sido desastrosos.
>
> (Págs. 107-108.)

Volveremos sobre estas declaraciones en el Capítulo VI, cuando examinemos el programa de tests *"No Child Left Behind"* en los EE.UU. y la evaluación del currículum nacional en Inglaterra. Este plan de pago por resultados del siglo XIX nos recuerda que las controversias actuales sobre la rendición de cuentas tienen precedentes.

Conclusión

La evaluación no se produce de forma accidental, sino que es una actividad social que tiene una finalidad. Las formas de evaluación que utilizamos también están socialmente determinadas y reflejan unas estructuras sociales. El desarro-

llo de los sistemas de evaluación ha sido invariablemente bienintencionado, pues su finalidad era conseguir una selección más justa y unos niveles mejores de enseñanza y de aprendizaje, tanto en las universidades como en las escuelas. En la práctica, las evaluaciones formales, como los exámenes, han conducido a una selección más justa que el sistema de influencias al que reemplazó. La elaboración de pruebas para identificar a quienes quizá no fueran capaces de beneficiarse de la escolarización regular fue también un avance positivo, aunque veremos que el desarrollo de esta cuestión distó mucho de ser apacible (Capítulo II).

Históricamente no se reconocieron en medida suficiente las limitaciones de los exámenes, en particular su forma de reflejar y respaldar las divisiones de clase de la época y la cantidad de personas que fueron excluidas. La evolución histórica en Inglaterra constituye un ejemplo específico, en el que las universidades de élite desarrollaron unas formas de evaluación que después se extendieron a las profesiones y más tarde a las escuelas de "clase media". En otros lugares, los sistemas de evaluación se han desarrollado de manera diferente y han reflejado otros esquemas sociales [8].

La influencia de los exámenes en lo que se enseñaba y en el rendimiento de los estudiantes también ha sido con frecuencia positiva pero, de nuevo, no se han reconocido suficientemente las consecuencias negativas. El efecto de la restricción de la enseñanza a unos exámenes previsibles y su efecto embrutecedor con el paso del tiempo fueron algunas de ellas. La política del pago por resultados en la Inglaterra del siglo XIX es un ejemplo revelador de las distorsiones provocadas por el uso de la evaluación con fines de rendición de cuentas en cuestiones importantes.

¿Han cambiado mucho las cosas desde el enamoramiento victoriano de los exámenes? Nos hemos apegado aún más a ellos en cuanto a la escala y la regularidad de la evaluación, estimándose que un estudiante típico escolarizado en Inglaterra hasta los 18 años habrá realizado más de un centenar de evaluaciones externas [9]. Estas evaluaciones se han hecho más inclusivas, participando en ellas la inmensa mayoría de los alumnos y alumnas, aunque siguen vigentes ciertas cuestiones relativas a la justicia, en particular los supuestos culturales y de clase (Capítulo V). Lo que ha ocurrido en muchos países es que la finalidad de *rendición de cuentas* ha saltado a primer plano y esto ha influido, a menudo negativamente, en la escuela (Capítulo VI). Si examinamos los orígenes y fines históricos de lo que quizá demos por sentado, podremos cuestionar más eficazmente nuestras prácticas al uso.

[8] Véase el clásico *Secondary School Examinations*, de ECKSTEIN y NOAH (1993).

[9] El grueso de éstas está constituido por las 10 asignaturas, más o menos, examinadas en el GCSE, a los 16 años, con dos o más pruebas, y de los *A-levels* modulares realizados a los 17-18 años, con 6 módulos por cada asignatura, de los que, por regla general, los estudiantes hacían tres. Los módulos pueden repetirse para subir nota, por lo que la mayoría de los estudiantes repiten algunos. El número de módulos ha de reducirse, aunque, en 2007, el Gobierno anunció un plan piloto de pruebas bianuales adicionales del currículum nacional (*"progress tests"*: "pruebas de progreso") para los alumnos de 7 y de 14 años, que aumentarán espectacularmente el número de exámenes.

CAPÍTULO II

Los tests de inteligencia: Cómo crear un monstruo

> Siempre ha habido una fuerte tendencia a creer que lo que ha recibido un nombre debe tener una entidad o ser, tener una existencia independiente por sí mismo. Y, si no puede encontrarse ninguna entidad real que responda al nombre, los hombres no suponían por esa razón que no existiera, sino que imaginaban que era algo particularmente abstruso y misterioso.
>
> (John STUART MILL, 1806-1873.)

Este capítulo examina la capacidad de la evaluación para crear y no solo para medir. He escogido las pruebas de inteligencia porque ilustran convincentemente el argumento de que una evaluación puede crear aquello que dice medir. La evaluación puede tomar una idea o especulación y hacer que parezca que existe realmente mediante el uso de la medida y dándole nombres y clasificaciones. De este modo, *cosificamos* nuestras ideas: les damos una existencia independiente.

Nuestra idea de *inteligencia* es el producto de este tipo de procesos. Esto no quiere decir que no exista la inteligencia, sino que es posible que algunas de las creencias anglófonas predominantes acerca de ella la hayan cosificado de este modo. Estas creencias transformaron la inteligencia en una entidad biológica (una causa de nuestro comportamiento), con la que nacemos y que no cambia. Michael HOWE, que rechaza esta interpretación biológica, señala que la inteligencia, como el éxito o la felicidad, es una consecuencia. Igual que el éxito no es la razón por la que alguien triunfa, "la inteligencia es el nombre abstracto que denota el estado de ser inteligente, pero no es la explicación del mismo" (pág. ix).

Los *motores de descubrimiento* de HACKING, que describimos en la Introducción, recogen perfectamente el proceso de convertir una consecuencia descriptiva en una causa biológica. Los cuatro primeros (contar, cuantificar, crear normas, correlacionar) implican, por regla general, técnicas estadísticas, algunas de las cuales fueron ideadas por los mismos creadores de los tests de inteligencia[1].

[1] Tanto GALTON (1822-1911) como SPEARMAN (1863-1945) hicieron aportaciones originales a la estadística, aportaciones que siguen vigentes en la actualidad. GALTON se interesó por desarrollar for-

Después, estas avanzaron para *biologizar*, *genetizar* y *normalizar*, algo que los promotores británicos y norteamericanos del CI (cociente de inteligencia) estaban más que dispuestos a hacer, un paso que Alfred BINET, el creador francés de los tests de inteligencia, se negó a dar. Para él, la inteligencia siguió siendo un resultado que podía modificarse.

La historia de los tests de inteligencia refuerza el argumento de que la evaluación es una actividad social, aunque sus defensores la presenten como una medida científica imparcial. Las figuras más destacadas actuaban movidas en gran medida por sus creencias ideológicas, que se basaban en supuestos hereditarios, raciales y de clase. Este bagaje histórico ha encontrado un lugar en nuestras actitudes y nuestro vocabulario actuales. Desde una crítica de los tests de CI podemos cuestionar este legado y encontrar modos más constructivos de enfocar la cuestión.

Los tests de capacidades: El nuevo "Clismo"[2]

Una respuesta inmediata podría ser que las pruebas de CI pertenecen al pasado y que, desde entonces, hemos avanzado. Esto es en parte cierto: en la actualidad, la práctica educativa de la evaluación se centra mucho más en las pruebas de rendimiento y, por regla general, los tests de CI ya no son fundamentales en la selección escolar a los 11 años[3]. A mi juicio, la tradición pervive en forma de tests de *capacidad* y de *aptitud* —que *están* muy extendidos. Por ejemplo, en Inglaterra, para entrar en secundaria, más de dos tercios de los niños y niñas de 11 años realizan el *Cognitive Ability Test* (CAT) *, editado comercialmente. Esta prueba es poco más que un test de inteligencia con una nueva presentación, con apartados verbal, no verbal y numérico. Otros niños y niñas de 11 años tendrán que ser seleccionados para escuelas especiales mediante tests de aptitudes, que difícilmente se distinguen de los tests de capacidad (véase la Introducción).

La paradoja es que los maestros y profesores que rechazarían el uso de tests de CI parecen muy satisfechos cuando se utilizan tests de capacidad que predicen el rendimiento posterior. Lo que me preocupa es que, aunque la capacidad podría no ser más que una forma diferente de aludir al "logro" o al "rendimiento, en realidad comparte los supuestos de los tests de inteligencia: se considera que la capacidad es la *causa del rendimiento, en vez de un aspecto del mismo*. Y, como veremos, hay una tendencia a interpretar esto como un atributo con el que los niños llegan a la escuela. Así, una puntuación de capacidad tiene la misma fuerza

---

mas de escalar las puntuaciones en una dimensión, por ejemplo, la distribución de la altura, que utilizaba la conocida "curva en forma de campana", y el concepto de "desviación típica". SPEARMAN desarrolló el análisis factorial, un avance radical también para identificar *rasgos latentes* que no podían medirse directamente, pero sí descubrirse a partir de técnicas correlacionales. Harvey GOLDSTEIN señala que fueron unas aportaciones impresionantes, dado que estos cálculos tenían que hacerse en gran parte a mano, y estas limitaciones técnicas implicaban que solo se podía trabajar con números limitados de factores, frente a la multidimensionalidad que permiten los ordenadores en la actualidad.

[2] Título del capítulo sobre la capacidad de GILLBORN y YOUDELL (2001).

[3] Aún hay tests de CI en el *11+* que siguen utilizándose en diversas administraciones educativas de Inglaterra y de Irlanda del Norte, aunque estos van a suprimirse; véase la nota 7 de este capítulo.

* *NfER*: *National Foundation for Educational Research*: "Fundación nacional de investigación educativa". (*N. del T.*)

que una puntuación CI para configurar la identidad del alumno o alumna (por ej., "capacidad baja") y determinar las expectativas de los maestros o profesores.

Por eso, aunque me centre en el desarrollo de los tests de CI, hay que tener en cuenta que esto ha invadido también nuestro pensamiento actual acerca de la capacidad y la aptitud. La investigación de David GILLBORN y Deborah YOUDELL (2001) en centros de secundaria indica que eso es lo que ocurre. Afirman los autores que la capacidad "ha llegado a interpretarse (tanto por los planificadores como por los profesionales) como una representación de la idea de sentido común de "inteligencia" (pág. 65). Como dice uno de sus profesores, "uno no puede *dar* la capacidad de otro" (pág. 78). El mayor peligro del uso actual de la palabra "capacidad" es, según ellos, que:

> actúa como una versión *no reconocida* de "inteligencia" y del "CI". Si reemplazáramos "capacidad" por "CI", saltarían muchas alarmas que permanecen en silencio porque "capacidad" actúa como una reconstrucción incontaminada, aunque poderosa, de todas las creencias que antes se plasmaban en términos como "inteligencia".
>
> (Pág. 81.)

Este es un tema que han recogido Susan HART y sus colaboradores en su proyecto de investigación *Learning Without Limits*. Les interesaba el hecho de que los calificativos de capacidad "ejercen una fuerza activa y poderosa en los procesos de la escuela y del aula, contribuyendo a crear las mismas disparidades de rendimiento que pretenden explicar" (pág. 21). En respuesta, se propusieron crear entornos de aprendizaje que asumieran la "transformabilidad", en vez de una capacidad fija.

Por eso, aunque centre mi atención en los tests tradicionales de CI, mi intención es suscitar cuestiones acerca de nuestros supuestos vigentes sobre la capacidad, uno de los discursos predominantes en las escuelas y en la planificación, y la herencia de las creencias acerca del CI.

La creación de la inteligencia

Los tests de inteligencia comenzaron como una evaluación diagnóstica pragmática para determinar a quiénes convendría escolarizar en clases especiales. Mediante una serie de desarrollos técnicos de la evaluación, su importancia creció hasta asumir un papel social mucho más amplio, que incluía su poder para prever quiénes podrían beneficiarse de determinadas formas privilegiadas de educación. Fomentaban la creencia de que nacemos con una cantidad de inteligencia heredada, que difiere significativamente entre individuos y grupos. Como esto respaldaba las creencias predominantes acerca de la estratificación social y concordaba con las perspectivas sociales de sus principales exponentes "científicos", arraigó firmemente en la psicología popular de las culturas angloparlantes. Lo que intenta hacer este capítulo es mostrar cómo se produjo este proceso de "intencionalidad de ser" en las pruebas de inteligencia.

Como veremos, no tenía que haber sido así. Alfred BINET, que elaboró el primer test de inteligencia, adoptó un enfoque muy diferente, como hizo la tradición, desde THURSTONE y GUILFORD en adelante, que se opuso a la idea de una inteli-

gencia unitaria. *Inteligencias múltiples*, de Howard GARDNER, hace esto mismo hoy día (Capítulo III). A mi juicio, el enfoque "fijo y unitario" ganó porque se ajustaba al clima de la época y ofrecía una justificación fácil para entender las condiciones y las políticas sociales favorables a quienes ostentaban el poder. Este *Zeitgeist* incluía también ideas del positivismo científico, con su llamamiento a la selección científica eficiente y a la meritocracia, con su preocupación por la selección de los más capaces.

Aunque se proclamaba que todo esto era consecuencia del estudio científico, la realidad era muy diferente: unas creencias firmes acerca de la herencia y la superioridad racial dictaron la evaluación de la inteligencia, no nacieron de ella.

La visión de BINET

El subtítulo de este libro es: "los usos y abusos de la evaluación". En un capítulo que versa en gran parte sobre los abusos, el francés Alfred BINET (1857-1911) presenta una norma mediante la que juzgar a quienes le siguieron. La narración de BINET solo presenta una descripción gráfica de cómo pueden cambiar los fines de la evaluación y cómo lo benigno se convierte en maligno.

La contribución de Alfred BINET consistió en elaborar los primeros tests de inteligencia. Psicólogo experimental y teórico, trabajó con las autoridades educativas de París para identificar a los niños que no fuesen capaces de desenvolverse en la enseñanza general. El enfoque de BINET era pragmático y se centró en lo que era necesario para el aprendizaje escolar. La primera versión estaba formada por una batería de tests basados en actividades y conocimientos que la mayoría de los niños tenía oportunidad de aprender antes de empezar a ir a la escuela. Los 30 ítemes de su test preescolar de 1905 incluían atender instrucciones sencillas, comparar longitudes y pesos, distinguir semejanzas y diferencias entre objetos, decir palabras que rimaran entre sí y formular preguntas para situaciones diversas. *Estos tests no pretendían medir facultades específicas de la mente, sino dar una visión general del funcionamiento del niño*. Su propio resumen decía: "Casi podríamos decir: 'Importa muy poco en qué consistan las pruebas, con tal de que sean numerosas'" (1911, pág. 329). A partir de las respuestas, podía determinarse una "edad mental" y compararse con la edad cronológica del niño. La diferencia determinaría la necesidad o no de escolarización especial. Las tareas se revisaron y extendieron a los niños mayores antes de la temprana muerte de BINET, en 1911.

El trascendental avance de BINET: desde el exterior al interior del cráneo [4]

Aunque el enfoque de BINET pueda parecernos muy obvio, en su época fue un desarrollo radical. La evaluación de las competencias cognitivas de los niños

[4] En esta sección, me he basado en los capítulos 2-4 de *The Mismeasure of Man*, de Stephen Jay GOULD (1996; hay traducción castellana: *La falsa medida del hombre*. Trad: Ricardo POCHTAR BROFMAN y Antonio DESMONTS. Barcelona: Crítica, 2004).

representaba una ruptura clara con respecto a las prácticas de evaluación de su tiempo. La ciencia normal de la época utilizaba diversas medidas externas y físicas para decidir acerca de la "anormalidad". Una evaluación científica de la inteligencia que gozaba de popularidad era la craneometría: la medida del tamaño del cráneo, que había sido elevado a categoría científica por su compatriota Paul BROCA. Este había concluido que:

> En general, el cerebro es mayor en los adultos maduros que en los ancianos, en los hombres que en las mujeres, en los hombres eminentes que en los hombres de talento mediocre, en las razas superiores que en las razas inferiores.
>
> (1861, pág. 304.)

El mismo BINET había publicado varios artículos utilizando este método, antes de reconocer que sus ideas preconcebidas habían influido en sus medidas. Esto ocurrió después de que su ayudante, Théodore SIMON (que no tenía ninguna postura arraigada que defender), le presentara diferentes interpretaciones de las mismas cabezas de "idiotas e imbéciles". Fue entonces cuando pasó de este enfoque "médico" improductivo a otro más psicológico.

En otros lugares, Francis GALTON, en Inglaterra, y James CATTELL, en los EE.UU., utilizaron una serie de medidas físicas, como los tiempos de reacción, para determinar la inteligencia relativa. GALTON, que lo medía todo, estudió la fuerza, la agudeza visual, la velocidad de las reacciones y la capacidad de distinguir colores, entre otras variables. En 1890, CATTELL propuso diez medidas: el apretón de mano más fuerte posible; el movimiento más rápido posible de la mano y el brazo; la cantidad de presión a la que empieza a sentirse como dolor, y la precisión al juzgar diez segundos de tiempo [5]. La justificación de CATTELL de estas medidas nos retrotraen al razonamiento victoriano sobre la capacidad que comentamos en el Capítulo Primero: "es, no obstante, imposible separar la energía corporal de la mental" (pág. 374). Aunque podamos cuestionar inmediatamente este aserto (¿qué habrían hecho con Stephen Hawking?), las expresiones como "de mirada clara" y "de mente viva" indican que, a menudo, mantenemos de alguna manera esa relación.

BINET adoptó una postura muy diferente de los que trataban de clasificar grupos sobre la base de unos rasgos fijos o de tamaños relativos. A él le preocupaba que incluso su propio constructo de la edad mental pudiera utilizarse erróneamente y se considerara que representaba una entidad real. También previó algunas consecuencias no deseadas de su enfoque: que los maestros utilizaran las puntuaciones como base para deshacerse de los niños revoltosos y los que no mostraran interés. Criticaba a los maestros que, en presencia del alumno, utilizaban expresiones como: "Este niño nunca llegará a nada... tiene poco talento... no es en absoluto inteligente" (1909, pág. 100). BINET tenía muy claro que lo que trataba de hacer era facilitar el ambiente que ayudara a aprender a los niños. Para él, la tarea consistía en incrementar la inteligencia de los alumnos, que describía como *la capacidad de aprender y asimilar la enseñanza* (pág. 104).

[5] HANSON (1994), pág. 206.

Dado lo que sucedió, debemos considerar que BINET presentaba una visión humanitaria que fue rápidamente oscurecida por quienes le siguieron y por quienes abrazaron un conjunto muy diferente de valores sociales.

El proceso de cableado: Biologizar, genetizar

Podemos medir y generar puntuaciones y categorías de cualquier cosa: altura, felicidad, productividad y destrezas de conducción. El paso crítico es lo que *inferamos* de aquellas. Aunque pueda ser fácil ponerse de acuerdo en la altura de una persona, que la consideremos "alta" dependerá de dónde estemos: puede que sea alta en Hong Kong, pero no en los Países Bajos. Podemos tener más dificultades para ponernos de acuerdo en cómo medir la felicidad y, si podemos hacerlo, quizá no queramos ir más allá que considerar esto como el producto de una compleja red de procesos: decir que alguien es feliz porque posee felicidad no parece muy explicativo. Lo mismo cabe decir con respecto a la productividad: un valor de productividad describe, pero no explica, un resultado. En una fábrica, no buscamos un componente de productividad que sea independiente de todos los demás procesos. Algo análogo ocurre con la inteligencia; el problema es que, históricamente, se ha hecho una inferencia que va más allá: la inteligencia *explica* tanto como describe; por tanto, su conducta es inteligente porque usted es inteligente. HOWE señala que, si aplicamos esta lógica a las destrezas de conducción, de la actuación de una persona en un test de conducir, habría que inferir si tiene o no una capacidad innata para conducir, "condenando a quienes suspendieran en el primer intento a toda una vida de dependencia del transporte público" (pág. 15).

Por tanto, el problema de la inteligencia (y de la capacidad) está en esta "intencionalidad de ser" causa biológica. Los tests están construidos para puntuar nuestra inteligencia subyacente, que difiere entre individuos. Como estos individuos proceden de diferentes grupos sociales y raciales, es muy fácil comparar puntuaciones medias entre grupos (aunque ésta no sea una comparación válida; véase más adelante) y hacer juicios relativos sobre éstos. Como los resultados se ajustan muy bien a las creencias y estructuras sociales: nuestros dirigentes tienen un CI elevado y nuestros trabajadores tienen un CI bajo, se convierten en una verdad. La medida se presenta como científica y los psicómetras facilitan un sofisticado apoyo estadístico. Esto permite extender la prueba diagnóstica de quienes tienen necesidades especiales a todos los niños, así como a los adultos, dado que todos tenemos CI.

Al considerar la inteligencia como una entidad (como, por ejemplo, la altura), tenemos que ubicarla: necesita un hogar físico. El siguiente paso racional consiste, por tanto, en "biologizar", situando fisiológicamente la inteligencia, por ejemplo, como una fuente de energía o una capacidad de procesado. Después, esto se "genetiza": si hemos nacido con ella, debemos haberla recibido de nuestros padres. Esta operación causal, en gran medida especulativa, redondea el proceso: la inteligencia es real, heredada, puede medirse con precisión y causa diferencias de rendimiento tanto entre individuos como entre grupos.

Históricamente, este proceso no es tan caricaturesco como pudiera parecer a primera vista. Lo que trata de demostrar el apartado siguiente es que quienes

continuaron con el trabajo de BINET tenían unas creencias sociales arraigadas para cuya promoción se utilizaron los tests de CI, igual que se había hecho antes con las medidas del cuerpo y del cráneo. Aunque fueran líderes en la medida científica, fueron sus creencias preexistentes y no las pruebas las que los llevaron a sus ideas acerca de la inteligencia. La evaluación se utilizó entonces para respaldar esas creencias. Como la evaluación es, fundamentalmente, una actividad social, se basaba en unas premisas sociales y tiene consecuencias sociales. Lo que exponemos a continuación no es solo un resumen histórico arcano, sino que su legado permanece vigente en nuestros pensamientos actuales.

Creencias hereditarias y administración de tests en masa en Estados Unidos

El desarrollo de los tests de inteligencia en el mundo angloparlante presenta dos vías diferentes. Ambas comparten unos supuestos comunes acerca de la herencia y las diferencias individuales y raciales de inteligencia. Fueron los estadounidenses quienes recogieron el test de BINET y lo adaptaron para una administración más general. Aunque Henry GODDARD lo aplicó en un primer momento en el contexto de la escolarización especial, fue Lewis TERMAN, un psicólogo de Stanford, quien lo adaptó para un público más amplio, de manera que el test de inteligencia Stanford-Binet ha permanecido hasta nuestros días como una de las pruebas predominantes de inteligencia. Lo que hicieron TERMAN, Robert YERKES y otros fue aprovechar un desarrollo reciente de la evaluación estadounidense: el formato de opciones múltiples. Este simplificaba la administración de los tests en masa. Utilizando ítemes que se correlacionaran con el test Stanford- BINET, que se administra por separado a cada persona, desarrollaron los tests *Army Alpha* y *Army Beta* (el "Beta" era para quienes no sabían leer). Se administraron estos tests a 1,75 millones de reclutas.

Lo interesante de este repaso histórico es que, aunque parezca que los tests tuvieron una influencia limitada en los resultados de guerra, sus consecuencias fueron de un alcance enorme, porque condicionaron la sociedad a aceptar que esas pruebas eran una forma meritocrática de selección basada en la medida científica. En 1922, E. L. THORNDIKE escribió en relación con la educación:

> Sin duda, es imprudente impartir enseñanza a los estudiantes sin tener en cuenta sus capacidades para aprovecharla, si, gracias a un ingenio y una experimentación suficientes, podemos contar con unos tests que midan de antemano sus capacidades. (Pág. 7.)

Muy pronto, los tests del ejército pusieron de manifiesto las limitadas capacidades mentales de quienes los hicieron. Los estadounidenses blancos tenían una edad mental de 13 años, justo por encima del nivel de imbecilidad en el sistema de clasificación; los inmigrantes rusos, italianos y polacos se desenvolvieron peor (a lo que no ayudaron unas estadísticas incorrectas cuando combinaron ambos tests), mientras que los estadounidenses negros fueron los peores de todos. En este punto, las arraigadas creencias hereditarias del grupo configuraron su inter-

pretación. Estas diferencias se consideraron como el resultado de una dotación genética inferior, en vez de serlo de diferentes ambientes y oportunidades. Así lo interpretaba TERMAN: "Los hijos de padres triunfadores y cultos puntúan más alto que los hijos de familias desgraciadas e ignorantes, por la sencilla razón de que su herencia es mejor" (pág. 115). Ya en 1912, GODDARD había visitado con regularidad la isla Ellis, donde desembarcaban primero la mayoría de los inmigrantes europeos. Él y sus ayudantes femeninas "descubrían" a los deficientes, que eran sometidos a tests (utilizando a un traductor) para hallar su nivel de inteligencia. La mayoría eran declarados "débiles mentales".

Para GODDARD, el modo de mejorar los niveles nacionales de inteligencia era restringir la inmigración de los grupos inferiores, a favor de lo cual hizo campaña, obteniendo los resultados apetecidos. También defendió las colonias en las que vivieran los imbéciles y en las que se les impidiera reproducirse. TERMAN también hizo campañas sociales a favor de la reproducción selectiva. Quería que se propagasen los "capaces y buenos" y que se restringiera la reproducción de los "inferiores y viciosos":

> Todos los débiles mentales son, por lo menos, criminales en potencia. Nadie puede discutir que toda mujer débil mental es una prostituta potencial. El juicio moral, como el juicio empresarial, el juicio social o cualquier otro proceso superior de pensamiento es una función de la inteligencia.
>
> (1916, pág. 11.)

Por tanto, una puntuación elevada en un test de inteligencia no solo indica las competencias intelectuales, sino también el valor moral y social, el eco de la actitud victoriana ante los exámenes que vimos en el Capítulo Primero. Así puede crecer un constructo.

La aportación hereditaria británica: Estadística y eugenesia

Como su homóloga estadounidense, la rama británica contempló a estadísticos y psicólogos de primera fila, como Francis GALTON (1822-1911), Charles SPEARMAN (1863-1945) y Cyril BURT (1883-1971) que utilizaron sus competencias psicométricas al servicio de unos puntos de vista claramente hereditarios acerca de la inteligencia.

GALTON y la distribución de la inteligencia

A GALTON le interesaba cartografiar y escalar medidas psicológicas en una sola dimensión, la conocida curva de la distribución normal en forma de campana en la que la mayoría de las puntuaciones se agrupan en torno a la media. Esta técnica de escalado implica que las puntuaciones de los tests de inteligencia pueden normalizarse, de manera, por ejemplo, que tengan una media de 100 y una desviación típica de 15. Esto significa que algo más de dos tercios de la población (68%) tendrá un CI entre 85 y 115, y solo alrededor del 2% tendrá puntuaciones superiores a 130 (dos desviaciones típicas por encima de la media) y otro 2%, por

debajo de 70. El interés, entonces, se centra en quienes tienen una inteligencia superior (el libro clave de GALTON, de 1869, fue: *Hereditary Genius*) y, de modo especial, en quienes tienen una inteligencia baja. Le preocupaba que estos últimos estuvieran reproduciéndose más deprisa que los primeros y que la inteligencia se heredase. GALTON estaba convencido de que estaba ocurriendo esto y, habiendo acuñado el término *eugenesia* en 1883, defendió la regulación del matrimonio y del tamaño de la familia según las capacidades innatas de los padres. También tenía unas ideas muy arraigadas acerca de la superioridad racial y, una vez más, estaba deseando ordenar grupos. Su medida era la tasa a la que una raza produce genios; absurdamente, "calculó" que uno de cada seis antiguos atenienses entraba en esa categoría, muy por encima del uno de cada 64 anglosajones y del uno de cada 4.300 negros.

La "g" de SPEARMAN

La aportación matemática de Charles SPEARMAN al desarrollo de los tests de inteligencia sentó las bases teóricas de la idea de inteligencia como una única escala que podría representarse por un número. Su desarrollo de técnicas de análisis factorial justificó la idea de una "inteligencia general" (*g*) [6], que publicó por primera vez en 1904. La técnica utilizaba correlaciones entre puntuaciones de tests para identificar este factor común *g* subyacente; algunos tests tenían un carga más grande de *g*, es decir, medían de modo más directo la inteligencia general (esto se relaciona también con los supuestos de adecuación a la finalidad de los exámenes victorianos —tanto la gramática latina como la administración colonial pueden acercarse a *g*— por lo que una puede servir para predecir el desarrollo de la otra). GOULD argumenta convincentemente que:

> prácticamente todos sus procedimientos surgen como justificaciones de determinadas teorías de la inteligencia. El análisis factorial, a pesar de su estatus como pura matemática deductiva, se inventó en un contexto social y se utilizó por unas razones concretas. Y, aunque su base matemática sea indiscutible, su uso persistente como instrumento para el aprendizaje acerca de la estructura física del intelecto ha estado envuelto en profundos errores conceptuales desde el principio.
>
> (Pág. 268.)

Estos errores giran en torno a la creencia de que "ese concepto nebuloso, socialmente definido como inteligencia puede identificarse como una 'cosa' con un lugar en el cerebro y un grado definido de heredabilidad" (pág. 269). SPEARMAN acometió su labor con arraigadas creencias hereditarias. En contraste con la amplia clasificación de BINET, SPEARMAN presenta una clasificación estricta de individuos y razas en relación con la cantidad de inteligencia heredada. Estos

[6] Harvey GOLDSTEIN (2003) considera que este trabajo precoz es, en muchos sentidos, revolucionario: "la idea de que se podrían haber observado indicadores de un constructo latente subyacente que permitiría estimar un modelo adecuadamente especificado. El debate versa sobre los supuestos necesarios para hacer que el modelo funcione y ahí está el problema" (comunicación personal).

mismos resultados podrían haber sido interpretados de un modo puramente ambiental: las personas se desenvuelven bien en una serie de tareas porque sus familias y las escuelas les han proporcionado diversas destrezas que los ayudan en su actuación, haciendo posible la mejora de la inteligencia. Pero esa interpretación no cabía en la idea de SPEARMAN, por lo que también él, como miembro de la *Eugenics Society*, consideró que la restricción de la reproducción (y del voto) era una forma de proteger el nivel de inteligencia de la nación.

La radicalización de la perspectiva hereditaria de BURT

El sucesor de SPEARMAN en el *University College London* fue Cyril BURT, psicólogo matemático que dejó una profunda huella en el sistema educativo inglés. BURT asumió la postura de SPEARMAN acerca de la inteligencia general y endureció la base hereditaria de ésta:

> Este factor intelectual general, central y omnipresente, presenta otra característica, también desvelada por los tests y la estadística. Parece que es heredada o, al menos, innata. Ni el saber ni la práctica, ni el interés ni la aplicación sirven para aumentarlo.
>
> (1937, págs. 10-11.)

BURT fue el psicólogo oficial del *London County Council*, por lo que el contraste con el enfoque de BINET en París no podía ser más marcado. Aunque su trabajo estaba directamente relacionado con la escolarización especial, su idea del retraso consistía en que, en la mayoría de los casos, "se debía a factores mentales intrínsecos; en consecuencia, primordialmente es innato y, en tanto en cuanto lo sea, está más allá de toda esperanza de curación" (BURT, 1937, pág. 10). BURT pudo apoyar ciertas intervenciones educativas cuando los problemas eran achacables a factores específicos (*s*), que son susceptibles de mejora.

Por tanto, no se trata aquí de excentricidades inocuas de psicólogos y estadísticos de primera fila. Tampoco podemos descartarlas como una historia irrelevante de otra época. Las afirmaciones de este grupo han calado profundamente en la psique de las culturas anglófonas, conduciendo a la selección educativa sobre la base de las puntuaciones de CI; a la aceptación generalizada de las diferencias de clase y raciales en cuanto a la inteligencia, y a una visión de la inteligencia como algo innato y fijo. Sus actividades respaldan las más disparatadas afirmaciones sociológicas acerca del uso de la evaluación para el control social, la disciplina y el mantenimiento del capital social.

Sostenemos que estos supuestos generales han dejado un residuo cultural que tenemos que reconocer, de manera que, en muchas de nuestras relaciones cotidianas, utilizamos la palabra "inteligencia" de manera simplista y sentenciosa. Incluso nuestros insultos: *cretino, idiota, imbécil*, eran clasificaciones técnicas de los niveles de CI. (Entre los discos antiguos de vinilo que me han dado familia y amigos, tengo uno de 1978 de la banda *punk* Jilted John, con la canción "Gordon is a Moron", que me dieron la familia y los amigos; a propósito, *morón* era la clase que estaba por encima de "idiota" e "imbécil", pero por debajo de "torpe"). Las diferencias raciales de CI son una cuestión discutida en los EE.UU., avivada por

trabajos como *The Bell Curve*, de HERRNSTEIN y MURRAY (1994). Así también los debates políticos acerca de reducir el presupuesto de los programas compensatorios, de limitado valor si la capacidad está fijada y los pobres carecen de ella.

La ubicación de la inteligencia

En 1923, el psicólogo Edwin BORING presentó una definición operacional de inteligencia como "lo que prueban los tests". Siempre he considerado esto como una perogrullada, pero, en el contexto de la presente argumentación, entiendo ahora su significado más profundo: *nuestra forma de entender la inteligencia es, en gran medida, el resultado de nuestra forma de probarla*. Allan HANSON ha señalado tres formas de configurar el concepto de inteligencia debidas a los tests:

1. La idea de que la inteligencia es una única cosa se enraíza en el hecho de que el resultado de los tests de inteligencia se expresa a menudo en una única escala, como un CI, aunque el test mismo conste de varias partes diferentes. Cuando hay una única escala, se supone en general que hay una única cosa a la que se refieren las puntuaciones.
2. El segundo atributo: la inteligencia es cuantitativa y unas personas tienen más que otras, se deriva de la práctica de informar de los resultados de los tests de inteligencia mediante puntuaciones de una escala numérica. Solo los fenómenos cuantitativos pueden expresarse con números. Y cuando los números varían de una persona a otra, lo mismo debe ocurrir con la cantidad de inteligencia que representan los números.
3. La idea de que la cantidad de inteligencia poseída por cada individuo está fijada de por vida se deriva de la creencia de que los tests de inteligencia no solo miden lo que ya sabe una persona, sino su *capacidad* de aprender... [ésta] está impresa físicamente en la persona. Por tanto, se considera que la inteligencia de cada individuo está fijada por la herencia. (Págs. 255-256).

A éstas, yo añadiría otras tres de mi cosecha:

1. El desarrollo de los tests de inteligencia de opciones múltiples permitió una eficiente administración de tests en masa, ofreciendo una medida y una clasificación social a gran escala.
2. El uso del mismo test con diferentes grupos llevó a que fueran clasificados en relación con la inteligencia.
3. El uso de las puntuaciones normalizadas fomentó la creencia de que las diferencias de puntuaciones indicaban unas diferencias cualitativas precisas[7].

[7] No es así. Por ejemplo, las puntuaciones límite para seleccionar a alumnos para las *grammar schools* en el Reino Unido solían basarse en el número de plazas disponible, en vez de en una cualidad representada por quienes conseguían la nota de corte. Lo mismo ocurría para la selección de alumnos para educación especial. La nota de corte, en torno a 70, era esencialmente pragmática: en una escala normalizada, esto significaría alrededor del 2% de la población escolar. John GARDNER y Pamela COWAN (2005) han demostrado que las actuales pruebas de selección del *11+* en Irlanda del

¿Cuáles son las pruebas?[8]

Lo que consiguieron SPEARMAN y BURT en Inglaterra y TERMAN, GODDARD y JENSEN en los EE.UU., fue hacer que sus convicciones sociales formaran parte de la psicología popular de la cultura anglófona: la inteligencia es heredada, fija y diferencialmente distribuida entre clases y razas. De vez en cuando, estas creencias se "renuevan científicamente", como hicieron de manera muy notable el libro *Bias in Mental Testing*, de Arthur JENSEN (1980), que rehabilitó la *g* de SPEARMAN, y el superventas de 800 páginas *The Bell Curve: Intelligence and Class Structure in American Life*, de Richard HERRNSTEIN (1994).

Este apartado cuestiona estas afirmaciones. Lo haremos a través de tres fuentes principales de pruebas. La primera es que el CI no es fijo, sino *maleable*, y ha cambiado rápidamente con el tiempo. La segunda discute la perspectiva genética simplista de los proponentes de la inteligencia heredada, en la que la inteligencia es el resultado de una transmisión genética directa que implica solo unos pocos genes. La tercera contempla la afirmaciones relativas a las diferencias raciales intrínsecas y cuestiona el punto de vista de que los tests de CI son independientes de las culturas (o "reciben escasas influencias culturales", como viene diciéndose últimamente, en una expresión más prudente), además de examinar la legitimidad de la comparación de grupos en tests que pretenden medir diferencias individuales.

El efecto FLYNN

Uno de los hechos menos conocidos acerca de los tests de CI es que tienen que ser recalibrados periódicamente para volver a situar la media en 100 y ajustarlos porque las niñas obtienen puntuaciones superiores en promedio a las de los niños. James FLYNN, un especialista en Ciencias Políticas de la *Otago University* de Nueva Zelanda, se dio cuenta de que las puntuaciones han ido subiendo progresivamente en el transcurso del siglo XX. Así, en el momento de la renormalización, la puntuación media del CI es siempre mayor que 100 y ha de devolverse a ese punto introduciendo ítemes más difíciles. FLYNN estudió las pruebas internacionales de este fenómeno y resume así sus hallazgos:

> Tenemos a nuestra disposición datos de veinte naciones y no hay una sola excepción al descubrimiento de incrementos masivos de CI con el tiempo.
>
> (1998, pág. 26.)

Estima que se produce una mejora de tres puntos cada década en los tests estándar de CI (Stanford-Binet y Weschler). Puede que no parezca mucho, pero

---

Norte son muy poco fiables, de manera que hasta el 30% de los alumnos seleccionados pueden haber sido erróneamente clasificados por la forma de agruparse las puntuaciones alrededor de la puntuación crítica de corte.

[8] Este comentario se basa en gran medida en aportaciones a *The Rising Curve* (dirigido por Ulrich NEISSER, 1996), que examina las pruebas acerca de los cambios de las puntuaciones de inteligencia con el paso del tiempo.

significa que alguien que estuviera en el punto medio en la distribución de 1990 (CI de 100), habría obtenido una puntuación que lo incluiría en el 18% superior (CI 115) en la normalización de 1932.

Las implicaciones

El incremento en las puntuaciones de CI con el paso del tiempo, una vez permitidas las renormalizaciones, plantea un problema crítico para quienes mantienen una idea de la inteligencia como algo "innato y fijo". Esos rápidos incrementos no pueden ser el resultado de cambios genéticos, por lo que hay que buscar la manera de explicar esto. *The Bell Curve* pasa por alto en gran medida el argumento, limitándose a reconocer a regañadientes que los negros norteamericanos han mostrado un mayor incremento del CI con el tiempo (aunque aún van por detrás). No obstante, Richard LYNN, de Irlanda del Norte, defensor contemporáneo de la rama hereditaria-eugenésica de la inteligencia (que puede seguirse hasta GALTON), trata de explicar estos descubrimientos. Y, para él, la explicación está en la *nutrición* y el *elemento educativo en los tests de CI*. Para quienes no están lastrados por unas ideas arraigadas acerca de la inteligencia "innata y fija", éstas parecen unas líneas de investigación prometedoras. LYNN tiene que hacer algunas contorsiones, sobre todo en cuanto a la línea educativa. Revisaremos brevemente sus argumentos, por ser ilustrativos de los problemas que el efecto FLYNN plantea a la postura hereditaria.

La nutrición

Esta afirmación se adentra en lo que parece terreno común para los enfoques innato y ambiental. Cuando mejora la nutrición, puede producirse un desarrollo más sano que, a su vez, mejora la capacidad mental. La analogía más corriente, sobre la que volveremos, es la altura. Cuando la nutrición ha mejorado, también lo hace la altura media; así, como las puntuaciones de CI, también ésta ha aumentado a ritmo constante. Para LYNN, el paralelismo es irresistible: es el mismo proceso de realización del potencial genético en ambos casos. En realidad, evocando la craneometría del siglo XIX, señala que "el tamaño de la cabeza y el tamaño del cerebro también han aumentado durante el último medio siglo... La significación de este hecho es que el tamaño del cerebro es un determinante de la inteligencia" (1998, pág. 211). Aunque en ningún sitio indiquen las pruebas tal cosa —en muchas naciones industrializadas, el incremento de la altura fue más lento en la década de 1970, pero no así el de CI[9]—, la idea de que una nutrición mejor conduce a un mejor funcionamiento cognitivo es verosímil. Es el desencadenante de los programas de "vitaminas" y de "aceite de pescado" para la mejora cognitiva, con respecto a los cuales las pruebas arrojan resultados diversos y las afirmaciones hechas son exageradas[10]. En su revisión de las pruebas acerca del

[9] Véase: MARTORELL (1998).

[10] Véase, por ejemplo: *The Fish Oil Files* ("Los archivos del aceite de pescado") en la columna *Bad Science*, del *Guardian* de 16 de septiembre de 2006.

papel de la nutrición en el desarrollo de la inteligencia, Marian SIGMAN y Shannon WHALEY indican que la nutrición no es un factor aislado:

> La nutrición raramente opera sola en la configuración de las competencias intelectuales; unos individuos mejor alimentados solo pueden aprender y rendir mejor si tienen acceso a experiencias que configuren adecuadamente su desarrollo para las exigencias de su cultura. Además, las mejoras nutricionales pueden ser las responsables del ascenso del CI en ciertos momentos de la historia de un país y no en otros, dependiendo de los cambios históricos de las disponibilidades alimenticias, las demandas de pensamiento abstracto y el contacto con los tipos de competencias requeridos por los tests de inteligencia.
>
> (1998, págs. 175-176.)

En torno a esta compleja interacción de nutrición y aprendizaje, los caminos empiezan a separarse. LYNN se da cuenta de adónde puede llevar esta lógica y afirma que los incrementos de CI con el tiempo se producen cuando los niños tienen 2 años "y, desde luego, antes de los 4-6 años". Esto descarta los efectos de estimulación de una mejor escolarización, la TV, los libros de crucigramas, "porque estos no influyen en los niños de 2 años y tienen un impacto mínimo sobre los de 4-6 años" (en el Ulster deben de educar a los niños de forma diferente [11]). La consecuencia es "que solo sigue en pie la teoría de la nutrición" (1998, pág. 212). Lo que pasa por alto es que una mejor nutrición probablemente forme parte de un conjunto de "mejores": mejores condiciones, mejor cuidado de los hijos y mejor salud; todos ellos pueden ser determinantes de la inteligencia. Un mejor ambiente puede aportar unas condiciones más conducentes al aprendizaje, incluyendo el hecho de que el hambre no distraiga, que se reflejarán en unas mejores puntuaciones de CI, a menos que consideremos que el CI es independiente del aprendizaje. Las matrices de RAVEN ofrecen un interesante estudio monográfico a este respecto.

El curioso caso de las matrices de Raven

El "Raven" puede no significar nada hasta ver la imagen (Figura 2.1), que se reconocerá como uno de los ejercicios mentales de tipo rompecabezas que todos hemos hecho en alguna ocasión. Éste no es un ítem real (no quiero estimular efectos de práctica), pero plantea las mismas exigencias. La tarea consiste en utilizar la lógica de las dos primeras columnas y las dos primeras filas para determinar qué debe ir en la celda inferior derecha [12].

[11] En enero de 2007, el periódico de clase media *The Sunday Times* regaló a sus lectores el DVD *Brainy Baby* ("Bebé inteligente") pensado para niños de entre 6 y 36 meses. Empieza con unos niños que juegan con unos juguetes; así, Jeremy, de 2 años, que jugaba con ladrillos, era presentado como "futuro arquitecto", mientras que Piper, de 9 meses, era un "futuro contable". Había ejercicios para el cerebro derecho ("inspira el pensamiento creativo") y ejercicios para el cerebro izquierdo ("inspira el pensamiento lógico"). El DVD en cuestión se limitaba a poco más que a estimular el juego con juguetes educativos y la práctica precoz de formas, letras y números. *The Sunday Times* se distribuye en el Úlster.

[12] La solución es una simple cruz (una línea vertical, una línea horizontal), sin recuadro alrededor. La regla es que cada línea (vertical, horizontal, diagonal "media") debía estar en dos, pero solo

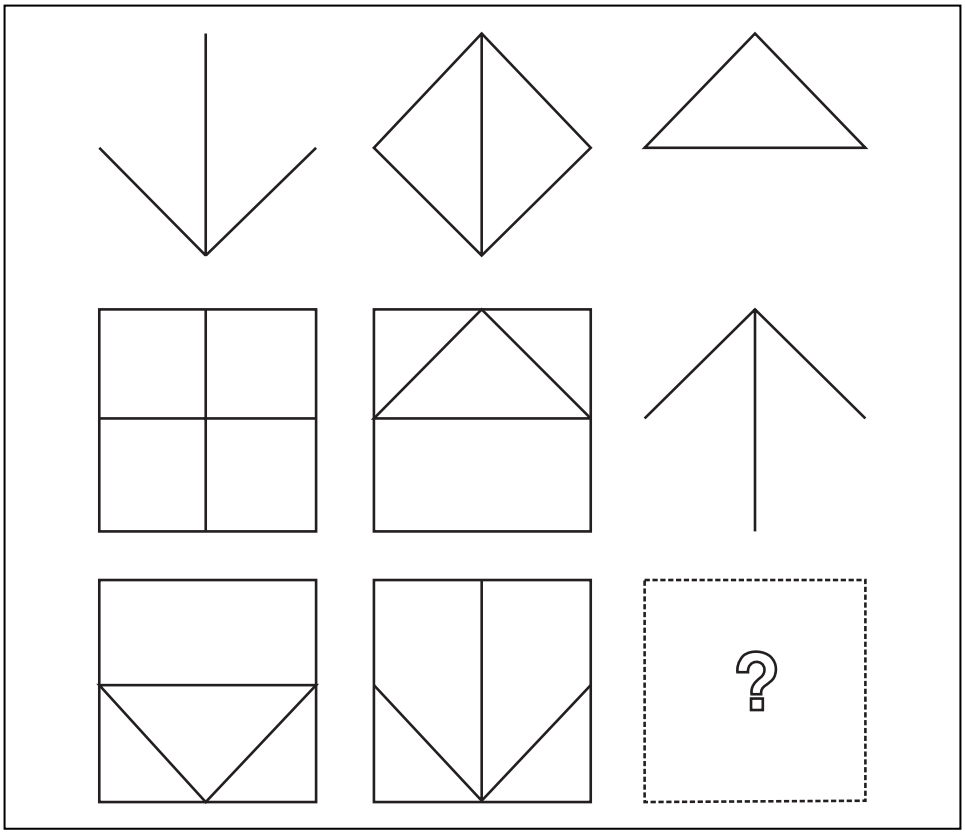


Figura 2.1. *Ilustración de un problema del tipo de las matrices de Raven.*

Lo que hace tan populares las Matrices progresivas de Raven es que parecen independientes de la cultura y de la escolarización. No hay lectura, no hay aritmética; solo puro razonamiento espacial y deductivo. Por eso, estas u otras parecidas han encontrado un hueco en la mayoría de las baterías de tests de inteligencia. El *NfER* Cognitive Ability Test* (CAT), que se utiliza mucho al principio de la educación secundaria en Inglaterra para prever el rendimiento, también las incluye. Fue Arthur JENSEN, cuyo influyente *Bias in Mental Testing* de 1980 rehabilitó la *g* de SPEARMAN, quien declaró que las matrices de Raven "miden aparentemente *g* y poco más" (1980, pág. 646) y "probablemente sea el instrumento más seguro que poseemos en la actualidad para descubrir a niños intelectualmente bien dotados de entornos deprimidos" (pág. 648).

---

en dos casillas, de las tres de cada fila y de cada columna. Esta es "la distribución de la regla de dos valores" (CARPENTER y cols., 1990, pág. 409).

* "Test de capacidades cognitivas". Las siglas "CAT" pueden inducir a error por ser las mismas que las del "Test de apercepción temática para niños", que no tiene nada que ver con éste. (*N. del T.*)

Así que aquí está el problema para la escuela de pensamiento de lo "fijo e innato": las puntuaciones en el test de Matrices progresivas de Raven han estado elevándose aún más deprisa que en los tests de CI, aunque miden la inteligencia "fluida" más que la "cristalizada", más condicionada ambientalmente. En los Países Bajos, donde se utilizaron con varones de 18 años, dentro de un programa de iniciación militar, las puntuaciones subieron de forma constante entre 1952 y 1982: el equivalente de 21 puntos de CI en 30 años (todo ello en una época en la que los neerlandeses no crecieron mucho más). FLYNN ha descubierto este efecto en otros muchos países [13].

¿Qué hace LYNN ante esta aparente maleabilidad de la inteligencia? Utiliza la táctica sorpresiva de declarar que, con respecto a las matrices de Raven, "la mejor manera de interpretarlas es como efectos de la escuela" (pág. 212), que, por tanto, hace que aquellos incrementos sean falsos:

> El Raven requiere la aplicación de los principios matemáticos de la adición, la sustracción, la progresión y la distribución de valores... En las tres décadas en las que se han producido estos incrementos en las puntuaciones, unas proporciones crecientes de chicos y chicas de 15 a 18 años han permanecido en la escuela, en la que han aprendido técnicas matemáticas que han aplicado a la solución de los problemas de las matrices.
>
> (Págs. 212-213.)

Me alegro mucho de aceptar el razonamiento de LYNN (aunque FLYNN ofrezca uno ligeramente diferente; véase más adelante) y hay pruebas que demuestran cómo puede mejorarse la ejecución. Se consigue dominando la cinco reglas que determinan lo que hay que poner en la última celda [14]. Mirado de este modo, revela lo que Patricia GREENFIELD llama "género cultural convencionalizado... La matriz es una forma de representación visual específica de una cultura. Para resolver problemas de matrices, hay que entender el complejo marco de representación en el que se presentan" (pág. 106). Puede que las hojas de cálculo, los gráficos y las pantallas ópticas hayan contribuido recientemente a fomentar estas competencias analíticas.

Sea cual fuere la causa de estas mejoras, son esencialmente ambientales. La importancia de esto reside en que, si la forma más pura de medida de *g* acaba dependiendo en gran parte de la enseñanza y del ambiente, ¿qué ocurre con los tests de vocabulario, razonamiento y aritmética, más directamente influidos ("cristalizados") por el ambiente? La solución obvia es reclasificar los tests de CI como *medidas de rendimiento general social y educativo*, volviendo adonde comenzó BINET antes de que la inteligencia se cosificase y separase del aprendizaje. Cuando hacemos esto, la correlación entre puntuaciones de CI y las de rendimiento educativo puede contemplarse de otra manera: *se produce porque ambas miden conocimientos y destrezas que se solapan, por lo que están muy correlacionadas*, no porque unas (CI) determinen las otras (rendimiento escolar).

Para los teóricos de la maleabilidad, esta es una explicación aceptable; para los del carácter fijo ("entidad") e innato, es un problema, porque ambas tienen que

[13] FLYNN, J. (1987).
[14] Véase: CARPENTER y cols. (1990).

permanecer separadas, dado que la inteligencia tiene una forma independiente y está localizada fisiológicamente (en alguna parte). Paradójicamente, esto es un problema para FLYNN, que ha tratado de explicar su propio "efecto".

Las paradojas de James FLYNN

Estamos en deuda con FLYNN por demostrar que las puntuaciones de CI se han elevado de un modo que desafía la explicación exclusivamente genética, especialmente cuando HERRNSTEIN y MURRAY, y LYNN, creen que, en todo caso, las cosas van a peor, desde el punto de vista genético (véase más adelante). Dada la sentencia de BURT: "ni el conocimiento ni la práctica, ni el interés ni la aplicación sirven para aumentarla" (1937, págs. 10-11), esto es un problema para quien adopte la postura del carácter "fijo e innato". FLYNN también ha tenido problemas para tratar de explicar su propio "efecto":

> Sigo convencido de que ni el talento (la capacidad de aprender más deprisa y dar saltos creativos) ni la inteligencia para comprender el béisbol (la capacidad de absorber las reglas usuales de la conducta social) han aumentado significativamente. Sin embargo, aunque creo que la solución mejorada de problemas en la sala de examen debe indicar alguna mejora similar en el mundo real, si bien sutil... todos mis instintos me dicen que una mejor comprensión de las competencias para el examen, más el descubrimiento de competencias del mundo real relacionadas con aquellas, producirán el conjunto de efectos necesario para identificar probables causas.
>
> (1998, págs. 59-61.)

Esta solución inicial necesita aún cierto desarrollo. Lo que FLYNN está buscando es un enlace ausente entre lo que requieren en realidad las competencias para el test, que pueden ser diferentes de las que se dice que requieren, y algunas destrezas cotidianas cada vez más utilizadas que se ajusten a los requisitos de competencias para el test. Podría considerarse un caso clásico de querer alumbrar un concepto: la inteligencia es una entidad que ha eludido parcialmente a los examinadores, por lo que los incrementos son, en gran medida, artificiosos y falsos: la inteligencia "real" no ha cambiado, aunque no estemos seguros de lo que sea.

En busca de una explicación, FLYNN rechaza una preparación mejorada de los tests, dado que los incrementos anteceden al período en el que la administración de tests se hizo habitual y persistió durante el tiempo en el que se hicieron impopulares. De igual manera, las teorías nutricionales no pueden explicar directamente los incrementos pues también tendríamos que esperar un mejor funcionamiento cognitivo en la vida cotidiana (que él no percibe) y porque la evidencia es muy contradictoria. Rechaza también otras explicaciones ambientales, como la urbanización y los cambios de estatus socioeconómico. Incluso la educación, un candidato prometedor, se considera solo como un interviniente menor, dado que los menores incrementos en las puntuaciones de CI han sido los de los tests de rendimiento, como los de aritmética, información y vocabulario. También se desecha la posibilidad de que las escuelas puedan estar enseñando mejor las destrezas de solución de problemas descontextualizados como "una hipótesis vacía" hasta

que se identifiquen las destrezas y se relacionen con las destrezas de los tests de CI. FLYNN ha seguido trabajando sobre este problema y, como veremos más adelante, ahora cree haber establecido algunas de estas relaciones.

Así, hemos cerrado el círculo: la inteligencia es lo que miden los tests de inteligencia pero, como los resultados son inestables, debemos construir una idea de "inteligencia" que trascienda los tests.

¿Qué ha causado los cambios?

El enfoque alternativo consiste en considerar los tests de inteligencia como medidas derivadas culturalmente que se adaptan cada vez más a entornos complejos. Patricia GREENFIELD pone el telón de fondo de esto:

1. Las culturas definen la inteligencia por lo que es adaptativo en su nicho ecocultural concreto.
2. Las definiciones de "inteligencia" son tanto ideales culturales como enunciados científicos (1998, pág. 83).

Considera esta autora que las ideas occidentales de inteligencia tienen más que ver con la *comprensión del mundo físico* que con el mundo social; con *ser capaz de pensar por sí mismo* que con la conformidad, y con valorar la *velocidad de reacción*. Contrasta con las ideas tradicionales africanas de "inteligencia", que implican competencias sociales, respeto a la forma de actuar de la sociedad y la deliberación, constructos todos ellos bien adaptados a una cultura estática, orientada a los parientes.

Un razonamiento similar puede hacerse con respecto a las ideas confucianas de "inteligencia"[15]. El estudio de Jin LI sobre las creencias culturales estadounidenses y chinas acerca del aprendizaje descubrió que, aunque los estudiantes estadounidenses consideraban que los "talentos" y "capacidades" constituyen una "cualidad inherente a una persona, que la capacitan para aprender", los estudiantes chinos "no consideraban la inteligencia como una cualidad inherente a una persona, sino como algo que puede aumentarse mediante el aprendizaje" (pág. 265).

Es interesante señalar que FLYNN (1991) descubrió que, aunque las puntuaciones medias de CI de los chino-norteamericanos estaban inmediatamente por debajo de 100, se estimaba que su rendimiento escolar y posteriores éxitos ocupacionales eran equivalentes a los de los europeonorteamericanos con un CI medio de 120. Puede haber muchas razones culturales que expliquen esto, incluyendo el compromiso moral con el proceso de perfeccionamiento personal y la insistencia colectiva en el rendimiento de los estudiantes. Un hallazgo relevante de la revisión transcultural de David WATKINS es que "los educadores chinos tienden a considerar tanto la creatividad como el entendimiento como procesos lentos que requieren mucho esfuerzo, mucha repetición y mucha atención, y no como procesos relativamente rápidos, marcados por la perspicacia" (pág. 161).

[15] Estoy muy agradecido a Evelyn Cao por su ayuda en este terreno.

Por tanto, es muy posible que, en las interpretaciones chinas de la inteligencia, no se otorgue la misma prioridad a la velocidad de respuesta, un elemento crítico de los tests de CI.

Los cambios sociales causan el incremento del CI

Si se considera que la inteligencia refleja un determinado nicho ecocultural, las mejoras de las puntuaciones de CI pueden interpretarse como un "paquete" cultural de factores interactivos, algunos de los cuales operan con más fuerza en algunas etapas. Aquí es adonde parece haberle llevado a FLYNN su pensamiento más reciente: ha llenado su "hipótesis vacía". Indica que las puntuaciones de CI describen una situación estática *en un punto determinado del tiempo durante el que se congela el cambio social*, por lo que la inteligencia parece un concepto unitario. Sin embargo, los incrementos del CI con el paso del tiempo "describen una situación dinámica en la que las prioridades sociales cambian de múltiples maneras... las competencias cognitivas del mundo real afirman su autonomía funcional y navegan con independencia de *g*... y la inteligencia parece múltiple. Si quieres ver *g*, para la película" (2006, pág. 6). Esto requiere una explicación, pero el esfuerzo merece la pena. Lo que FLYNN muestra a través de un análisis posterior de los incrementos con el paso del tiempo es que tales incrementos dependen mucho de determinados subtests de CI, por ejemplo, el de semejanzas y el de matrices de Raven, apreciándose pocos cambios en vocabulario, información y aritmética. Por tanto, nada ha cambiado mucho culturalmente en cuanto a las prioridades sociales en torno a la lectura y la aritmética: nuestros abuelos también aprendieron éstas.

Lo que, para FLYNN, durante este período ha "navegado con independencia de *g*" es la absorción de conceptos científicos y clasificaciones abstractas de la sociedad, y el valor cultural occidental vinculado a "la solución de problemas sobre la marcha sin un método aprendido previo" (pág. 8). Llama a estas interacciones "multiplicadores sociales" (pág. 15), porque los cambios sociales han llevado a rápidos cambios del entendimiento. Así, por ejemplo, si están de moda internacionalmente los rompecabezas *sudoku*, esto puede afectar rápidamente a estos tipos de destrezas de razonamiento lógico (que no son diferentes de las necesarias para las matrices de Raven). Su propio ejemplo es el ítem de semejanzas: "¿qué tienen en común los perros y los conejos?" (pág. 9). Nuestros abuelos podrían haber dado una respuesta funcional: "los perros se usan para cazar conejos", mientras que la respuesta requerida: "son mamíferos", les habría parecido trivial (¿a quién le importa?). Es tal la infiltración de este tipo de clasificación científica abstracta que la mayoría de los niños y niñas de 10 años reconocerían que "mamíferos" tiene sentido, aunque también ellos hubiesen dado una respuesta más concreta ("cuatro patas").

Por eso, aunque un test de CI realizado hoy día parecería dar de nuevo una puntuación relativa al mismo concepto unitario, esta instantánea prescindiría de todo lo que les hubiera ocurrido a los componentes individuales que navegan libremente con el paso del tiempo. Los ítemes similares, que priman las clasificaciones abstractas, se considerarán ahora como pensamiento normal y, sobre la

marcha, pensaríamos en hacer este tipo de enlace, mientras que antes esto se hubiese considerado una forma de razonar muy artificial.

Este razonamiento puede hacerse de un modo más general, por ejemplo, se ha demostrado que las mejoras nutricionales elevan las puntuaciones donde hay malnutrición, pero no se observan diferencias donde la dieta es suficiente. Es probable que una educación perfeccionada y prolongada mejore los resultados porque los tests de CI se han centrado en constructos que se refieren esencialmente a la inteligencia académica y que se utilizan para la selección[16]. La afirmación de LYNN de que los incrementos se producen antes de la escolarización formal pueden indicar también que los cambios en la crianza de los niños y en el tamaño de las familias están implicados en este proceso.

Esto nos deja un montón de factores ambientales con los que dar explicaciones verosímiles de un descubrimiento que no se discute: las puntuaciones de CI han ido elevándose de forma constante. Nuestra tarea aquí no es repartir causas de los incrementos, sino llevar la discusión *a considerar las puntuaciones de CI culturalmente dependientes y, por tanto, la probabilidad de que cambien al cambiar la cultura*. Las puntuaciones de CI pueden descender en el futuro si los cambios sociales marginan la inteligencia académica "lineal", por ejemplo, primando las formas multilineales de procesamiento de información que hallamos en las visualizaciones presentes en la web.

Con el fin de alejar el pensamiento de la idea de la inteligencia como una disposición innata e independiente de la cultura, presento un enfoque diferente, basado en la demostración de CECI y sus colegas de que:

> Con independencia de cómo conceptualicemos la "inteligencia" y el "rendimiento", la realidad empírica es que las tendencias de una imitan las tendencias del otro... las distinciones teóricas que hacen algunos entre "inteligencia" y "rendimiento" son irrelevantes para la realidad empírica de que una buena medida de una casi siempre está muy correlacionada con una buena medida del otro.
>
> (1998, pág. 290.)

Este enfoque *considera los tests de CI como tests de rendimiento generalizado*, que responden a la misma combinación de variables (escuela, padres, genética) que influyen en otras formas de rendimiento. Así veía BINET las de los niños pequeños: medir lo que podría esperarse que hubieran aprendido durante sus primeros años. Esto también explica los incrementos de las puntuaciones en las matrices de Raven, como una expresión del ahora socialmente valorado pensamiento sobre la marcha y de la manipulación de símbolos abstractos y de la lógica. Anne ANASTASI con sentido práctico lo resumía así:

> En la actualidad, se acepta de forma generalizada [por los psicólogos] que todos los tests cognitivos miden *capacidades desarrolladas*, que reflejan la historia de apren-

[16] Por ejemplo, CAHAN y COHEN (1989) examinaron la actuación de más de 11.000 alumnos y alumnas de 4.º a 6.º en Israel. Estudiaron a alumnos que solo diferían en edad unas pocas semanas, pero que estaban en cursos distintos en virtud de la fecha de nacimiento establecida como límite para el ingreso en la escuela. Después, compararon a éstos con alumnos con una diferencia de edad de 1 año, pero que estaban en el mismo curso. En 9 de 12 pruebas, los efectos de la escuela eran más fuertes que los de la edad.

> dizaje del individuo. Los instrumentos tradicionalmente denominados "tests de aptitud" evalúan aprendizajes aplicables en general, relativamente incontrolados y vagamente especificados. Esos aprendizajes se producen dentro y fuera de la escuela.
>
> (1985, pág. xxix.)

La consecuencia de esto es la inversión de la lógica de nuestro pensamiento. Unas puntuaciones altas o bajas de CI no son las *determinantes* del rendimiento académico, sino que *forman parte de él*. Por tanto, cuando se afirma que un test de capacidad CAT/SAT, etc., predice el rendimiento futuro, esto no tiene nada que ver con una capacidad subyacente, sino con logros educativos generalizados que son útiles para prever el rendimiento en exámenes futuros [17].

Este enfoque se presta a considerar la inteligencia como maleable, producto de nuestra experiencia. Igual que el ambiente ha llevado a la elevación de las puntuaciones de CI con el paso del tiempo, así también pueden fluctuar para los individuos en relación con su experiencia. Una vez más, la dirección de la lógica es crítica: un ambiente empobrecido generará puntuaciones de CI más bajas y quienes obtengan puntuaciones altas se habrán beneficiado de los entornos más ricos. Esto invierte la lógica tradicional de que una persona está en un ambiente pobre *porque* tiene una inteligencia baja, o que está prosperando porque su inteligencia es alta, como recoge la expresión de BURT: "las grandes desigualdades de renta personal son en gran medida, aunque no por completo, un efecto indirecto de las grandes desigualdades de inteligencia innata" (1943, pág. 141).

Genes dudosos

La precoz deformación de la idea de inteligencia de BINET perpetrada por los psicómetras británicos y estadounidenses estaba enraizada en la creencia cultural de que la capacidad se heredaba, una idea que casaba cómodamente con la estratificación social de la época. Ésta requería también especular sobre los mecanismos genéticos y la base fisiológica de esta transmisión de la inteligencia. Como productos de su tiempo, veían con frecuencia la base genética de la inteligencia de manera simplista: un único gen de cada progenitor. Este enfoque se derivaba directamente de los trabajos de MENDEL con guisantes, una combinación de genes recesivos y dominantes. Así, para GODDARD, la deficiencia mental estaba regida por un único gen recesivo: "si ambos padres son débiles mentales, todos los hijos serán débiles mentales. Es obvio que no deberían permitirse tales matrimonios" (1914, pág. 561). Añádase a eso el hecho de que los pobres que, por definición, tenían inteligencia baja, producían más hijos que los ricos y la consecuencia era una preocupación casi histérica por el descenso de los niveles de

[17] Esto ofrece una interpretación diferente, por ejemplo, del hallazgo de Steve STRAND en 2006 de que el CAT administrado al ingresar en la educación secundaria era un predictor ligeramente mejor del rendimiento en exámenes posteriores que los tests del currículum nacional administrados a los 11 años, y el predictor óptimo era una combinación de ambos. Mi interpretación sería que ambos son tests de rendimiento, de los que el CAT evalúa las destrezas más generalizadas, por lo que juntos sirven razonablemente bien para prever el rendimiento y las destrezas exigidas para el GCSE, a los 16 años.

inteligencia. HERRNSTEIN y MURRAY (1994, pág. 341) insistían más recientemente en la misma idea cuando decían: "es preocupante lo que le está ocurriendo al capital cognitivo del país (EE.UU.)".

Dada esta transmisión, ¿dónde se localiza la inteligencia? Esta pregunta dio lugar a todo un catálogo de fisiología especulativa. Para GALTON y CATTELL, tenía que ver con los niveles de energía general, que explican utilizando muchas medidas físicas (tiempos de reacción, fuerza de agarre, etc.). Más recientemente, JENSEN (1993) ha recuperado este enfoque, atendiendo a las respuestas eléctricas a los estímulos y a la velocidad de la respuesta fisiológica, por ejemplo, a inyecciones de glucosa, además de extender su trabajo a los animales. SPEARMAN identificó la *g* con los niveles de energía o fuerza mental a disposición de toda la corteza cerebral y más. Esta energía constante (parte de lo que GOULD llama la "envidia de la física" de SPEARMAN) también activaba los "motores" de los factores *s*, además de determinar el nivel de *g*. Para BURT, *g* se refería a la cantidad y complejidad del tejido cortical, mientras que los factores específicos estaban situados en áreas específicas de la corteza cerebral.

Desde entonces, hemos descubierto que la transmisión genética raramente es sencilla, sobre todo en relación con el complejo funcionamiento que implica la "inteligencia". No hay un gen *a* para la inteligencia, como no hay ninguno para la felicidad, la salud o la agresividad, pero esto no siempre resulta evidente tal como se expone (¿recuerdan el cromosoma XX de los criminales violentos y el gen "gay"?) [18]. Esto no quiere decir que no haya un componente hereditario en lo que se considera inteligencia, aunque adoptemos la definición básica de BINET: "la capacidad de aprender y asimilar la instrucción", "capacidad" conlleva una idea de predisposición. Es probable que las diferencias individuales sean el resultado de una combinación de factores genéticos sobre los que actúe el ambiente. La importancia de esto es que la inteligencia heredada no es una entidad sencilla (una cantidad fija de energía mental, etc.), sino una predisposición compleja que depende del ambiente en cuanto a su forma de expresarse. De ahí que estemos atentos al efecto *indirecto* de múltiples fuentes genéticas y sus interacciones enormemente complejas.

Un ejemplo médico útil de esta situación nos lo dan Michael RUTTER y colaboradores en su revisión de 2006 *Gene-Environment Interplay and Psychopathology*. Señalan que los genes implicados en los trastornos mentales son comunes (un tercio de la población los lleva, aunque los trastornos sean mucho más raros) y solo presentan una baja probabilidad de riesgo:

> En consecuencia, no tiene sentido presentar estos como los genes "de" un determinado trastorno mental. Están implicados en los procesos causales que conducen al trastorno mental, pero solo con otros genes y una serie de influencias ambientales. En pocas palabras, son parte de una causación multifactorial y no tienen ningún efecto genético directo.
>
> (Pág. 230.)

[18] *The Observer* (7 de agosto de 2005) informaba del trabajo realizado en la *Cambridge University* para localizar un *gen de matemáticas*. Decía también que un enfoque probablemente más productivo era el de Yulia Kovas, en el *Institute of Psychiatry* de Londres, que informa de "un fondo de entre 50 y 100 marcadores de ADN, cada uno de los cuales produce un pequeño efecto... y tienen un efecto activador, y no que un determinado gen nos haga ser mejores o peores haciendo sumas" (pág. 17).

A algunos lectores puede decepcionarles el hecho de que se dé algún crédito a la herencia genética ("naturaleza"), cuando eso parece debilitar el argumento que considera la inteligencia como un constructo artificial. Mi respuesta es que la conducta inteligente tiene una base biológica; el elemento artificial está en que se considere como algo fijo y localizable. Las pruebas indican la presencia de diferencias de aprendizaje entre niños desde el nacimiento (pregunte a cualquier padre) y de semejanzas entre los que comparten los mismos genes: los estudios de gemelos idénticos separados (que sabemos que BURT falsificó, aunque otros lo hicieron de manera más cuidadosa[19]). Podemos decir que todo es explicable recurriendo al ambiente, en relación con las experiencias prenatales, pero a menudo esto parece más una especie de justificación de un artículo de fe que una revisión de la evidencia.

Esto no significa que tenga que entrar en asignaciones especulativas naturaleza/aprendizaje (40/60, etc.). Como dice el biólogo Stephen GOULD:

> Cuando los factores causativos... interactúan de forma tan compleja, y a lo largo del crecimiento, para producir un intrincado ser humano adulto, no podemos, en principio, analizar la conducta de ese ser en porcentajes cuantitativos de causas fundamentales remotas... Las cuestiones verdaderamente sobresalientes son la maleabilidad y la flexibilidad, no los falaces análisis por porcentajes. Un rasgo puede ser heredable en un 90%, pero completamente maleable. Unas gafas de 20 dólares... pueden corregir por completo un defecto de visión 100% heredado.
>
> (Pág. 34.)

Por tanto, la clave no es si heredamos algo, sino hasta qué punto es *maleable* y *flexible* esa herencia. Esto se refiere a la *interacción* de lo biológico y lo ambiental que, a menudo, los psicómetras tratan de eliminar de la ecuación, y no a la simple aditividad (40% + 60%). Tomemos el ejemplo de la gordura. Unas personas pueden estar genéticamente predispuestas a ser más gruesas que la media cuando el nivel de nutrición es alto. Sin embargo, la misma predisposición genética puede llevar a que la persona sea más delgada que la media cuando los niveles de nutrición son bajos. En este caso, no tiene mucho sentido "analizar los porcentajes" y, cuanto más extremo es el contexto ambiental, menor es el impacto genético. Tampoco es constante. La obsesión de BURT y otros con el "porcentaje heredado" (sus correlaciones sin cambios en sucesivos estudios de gemelos idénticos permitieron detectar el fraude[20]) surge de la consideración de la inteligencia como una entidad fija que controla qué desarrollo es posible, dado que permanece constante y limita —un programa. Si el porcentaje es alto, como ellos creían que era, poco puede hacerse para remediarlo. "Heredado" significaba "*inevitable*".

Si consideramos la base biológica de la inteligencia como algo mucho más complejo y flexible, dado que tiene que ver con las respuestas de forma no pro-

---

[19] Véase: NEISSER, U. (1996): "Intelligence: Knowns and Unknowns", *American Psychologist*, 51 (2), págs. 77-101.

[20] Véase: GOULD (1996, págs. 264-269). El fraude consistió en que, entre 1943 y 1966, publicó una serie de artículos en la revista que dirigía sobre la inteligencia de gemelos idénticos que habían sido educados en hogares diferentes. No solo eran muy sospechosos los datos (y nunca vistos), sino que los dos investigadores, Margaret Howard y J. Conway, no existían. La historia completa se narra en: L. S. HEARNSHAW (1979): *Cyril Burt, Psychologist*. Londres: Hodder and Stoughton.

gramada, su maleabilidad es una consecuencia natural. Así, en la línea de BINET, la cuestión es: *"¿cómo puedo aumentar la inteligencia?"* Harán falta distintos anteojos para diferentes personas y, para estas, la fuerza puede variar en diferentes puntos de sus vida. La cuestión solo les parecerá extraña o contradictoria a quienes sean esclavos de la psicología popular anglosajona, "innata y fija".

Diferencias raciales de CI

Los padres fundadores estadounidenses y británicos de las pruebas de inteligencia estaban convencidos de que unos grupos eran intrínsecamente más inteligentes que otros. Se incluían aquí tanto grupos raciales como clases sociales. Las visitas de GODDARD a la isla Ellis, donde él y sus ayudantes buscaban a los inmigrantes que desembarcaban con apariencia de débiles mentales (¿acaso puede alguien tener un aspecto de persona inteligente después de pasar un mes encerrado bajo cubierta?) para administrarles a continuación un test, demuestran su interés al respecto. Se trataba de inmigrantes de clase trabajadora, que no hablaban inglés, procedentes del sur y del este de Europa, cuyos resultados en los tests los situaban como inferiores en inteligencia, lo que le llevó a hacer campaña para que se impusieran restricciones a la inmigración. Los tests de inteligencia del ejército de la I Guerra Mundial también se analizaron en relación con la raza, obteniendo peores resultados los norteamericanos negros.

La diferencia de medidas de CI entre estadounidenses blancos y negros se convirtieron en una cuestión explosiva a lo largo del siglo XX y hasta hoy, que se aviva periódicamente como, por ejemplo, con JENSEN en la década de 1980, y HERRNSTEIN y MURRAY, en su *The Bell Curve* de 1994. Su trabajo también se presentó como un informe científico imparcial, aunque los valores sociales que los impulsaban eran similares a los de sus predecesores. Influyó claramente en los debates acerca de si merecía la pena financiar los programas de recuperación para las minorías deprimidas, dadas sus capacidades intelectuales, limitadas y fijas. Este tipo de ciencia se hizo muy popular entre los políticos dados al recorte presupuestario de la era Reagan.

Esta tradición partía de dos premisas clave, que han llegado a formar parte de las creencias habituales acerca de las diferencias raciales:

1. Es legítimo utilizar las puntuaciones de un test de diferencias individuales e inferir diferencias grupales.
2. Los tests de CI son independientes de las culturas, por lo que cualesquiera diferencias son el resultado de capacidades innatas.

Es preciso cuestionar ambas premisas. No decimos que las puntuaciones no difieran, sino que las inferencias extraídas de estas diferencias son erróneas. Como punto de partida, hay que aclarar qué significa *heredabilidad*. Los coeficientes de heredabilidad (por ejemplo, CI=0,8) se basan en *diferencias*, por lo que ese CI significa que el 80% de las *diferencias* de CI de las personas se deben a los genes y no el 80% de sus puntuaciones totales. Stephen CECI pone el ejemplo del número de orejas con las que nacemos. El rasgo de "orejidad" se debe a

la acción genética pero, como hay poca variación en la comunidad humana, la heredabilidad genética está en torno a 0. *Estas diferencias de heredabilidad varían directamente con la cantidad de variación ambiental*, por lo que:

> Si hubiera un drástico incremento de la pobreza en una comunidad y, a consecuencia de él, a muchos de sus niños se les negara el acceso a experiencias educativas relevantes, se reduciría el tamaño de la heredabilidad estimada de esa comunidad. Esto se debe a que los genes son más importantes en la producción de diferencias entre niños de ambientes idénticos que en la producción de diferencias entre niños de ambientes muy distintos.
>
> (1996, pág. 131).

Por tanto, cuando un test diseñado para clasificar a personas dentro de un grupo concreto se utiliza en un grupo diferente y se comparan las puntuaciones medias, no nos dice nada acerca de diferencias *innatas* entre los dos grupos. Aunque pueda haber efectos hereditarios *dentro de* cada grupo (unos puntúan más alto que otros), esto no nos permite inferir diferencias hereditarias *entre* los grupos. Otros dos ejemplos pueden ayudar a explicar mejor este razonamiento.

LEWONTON (1970) nos pide que imaginemos que se plantan dos campos de maíz con la misma variedad de semillas genéticamente modificadas, pero solo un campo se regará y se abonará suficientemente. El resultado arrojará una diferencia entre campos completamente ambiental y una varianza completamente genética intracampos.

Stephen GOULD (1996) utiliza el ejemplo de la altura, que tiene una heredabilidad mayor que cualquier valor de CI:

> Tomemos dos grupos independientes de varones. El primero, con una altura media de 1,78 m, vive en una próspera ciudad estadounidense. El segundo, con una altura media de 1,68 m, pasa hambre en una aldea del tercer mundo. La heredabilidad es del 95%, más o menos, en cada lugar, lo que tan solo significa que los padres relativamente altos tenderán a tener hijos altos y los padres relativamente bajos, hijos bajos. Esta heredabilidad de la altura intragrupo no va ni a favor ni en contra de la posibilidad de que una mejor nutrición en la generación siguiente pudiera elevar la altura media de los aldeanos del tercer mundo por encima de los prósperos estadounidenses. Del mismo modo, el CI podría ser muy heredable dentro de cada grupo y la diferencia media entre blancos y negros en Estados Unidos solo seguiría indicando las desventajas ambientales de los negros.
>
> (Págs. 186-187.)

La respuesta a estos autores puede ser un "sí... pero, ¿qué pasa si todos estamos en el mismo campo?", que nos lleva a examinar si realmente estamos o no en el mismo campo.

Tests independientes de culturas

Esta segunda premisa se utiliza para anular este argumento "ambiental" afirmando que los tests son independientes de las clases sociales y del aprendizaje, de manera que todos estamos, en realidad, en el mismo campo. Hemos visto

antes que esto no se sostiene; incluso los tests puros de CI, como el de matrices de Raven, se clasifican ahora como "de baja influencia cultural", mientras que se reconoce que los tests de inteligencia cristalizada (vocabulario, aritmética) dependen de la escolarización y de la experiencia.

Decir que los tests eran independientes de las culturas es un poco como si alguien dijese que todo el mundo, salvo su propio grupo, tiene acento. Las premisas culturales de muchos ítemes de los primeros tests se ven con facilidad; los inmigrantes europeos recientes tenían que responder a preguntas de opciones múltiples como:

> Crisco es: medicamento específico, desinfectante, pasta dentífrica, producto de alimentación*.
>
> Christy Mathewson es famoso como: escritor, artista, jugador de béisbol, humorista**.

(¿Qué tal lo has hecho?[21]). También había que responder a instrucciones verbales como:

> Cuando yo diga: "ya", haga una figura 1 en el espacio que está en el círculo, pero no en el triángulo ni en el cuadrado, y haga también una figura 2 en el espacio que está en el triángulo y el círculo, pero no en el cuadrado. Ya.
>
> (GOULD, 1996, pág. 230.)

La cuestión es: ¿Qué más sesgos culturales sutiles están incluidos en nuestros actuales tests de capacidad?

Si hacemos nuestra la idea de que no hay tests independientes de culturas y que incluso los ítemes "descontextualizados", como las fórmulas matemáticas, son expresiones culturales específicas, estamos contemplando diferencias ambientales. *Si los tests de CI y de capacidad se consideran como tests de rendimiento generalizados, la escolarización y la experiencia son críticas*. La forma de "dar de comer y de beber" a los distintos grupos se convierte en la variable explicativa clave, no lo que se hereda. La diferencia política se plantea entre una mejor compensación de la privación y la reducción del esfuerzo, puesto que poco puede hacerse por los innatamente limitados.

Vuelta a BINET: La reformulación de la inteligencia

El argumento de este capítulo es que los tests de inteligencia constituyen un caso clásico y destructivo de unos expertos que utilizan la evaluación para crear

---

* "Crisco" es una marca de productos alimenticios. (*N. del T.*)

** Christopher (Christy) Mathewson fue un famoso jugador estadounidense de béisbol que murió en 1925. (*N. del T.*)

[21] En enero de 2007, *The Sunday Times* regaló otro DVD: *Brainpower: Exercise Your Mind*. Éste tenía 200 preguntas problema creadas por *Bristish Mensa* (la sociedad que agrupa a quienes están en el 2% superior de las puntuaciones de CI). Se trata de preguntas de "cultura general", entre las que se encuentran la fecha de fundación de *Mensa* y cuántas veces ha ganado Fulham la *FA Cup* (la copa de fútbol de Inglaterra). No ha cambiado tanto.

un constructo que se cosifica y al que se otorga una existencia independiente. Este proceso empieza con la medida de una serie de actuaciones, cuyas puntuaciones se combinan en una única puntuación de una escala con la que se clasifica a las personas. Como puede puntuarse, se da por supuesto que existe. No solo se convierte en una entidad, sino que pasa a ser un poderoso instrumento social que adoptan los psicómetras con un consistente plan de acción social. Sus puntos de vista hereditarios condujeron a que la inteligencia se considerase innata y fija. La posición social se consideró como una consecuencia de la inteligencia heredada, de manera que los dirigentes de la sociedad estaban bien dotados de ella y los pobres eran pobres a causa de su limitada inteligencia. Esto también ocurrió con los grupos, de manera que los varones anglosajones ocupaban la cumbre evolutiva. La diferencia reproductiva, que se manifestaba en que las familias de los pobres eran mayores y, por tanto, transmitían su inteligencia limitada, se convirtió en una preocupación, por lo que los psicómetras trataron de restringir su reproducción —y la inmigración de otros como ellos, en un intento de impedir el declive de la inteligencia de la nación.

Aunque gran parte de esto esté ahora desacreditado (GODDARD, TERMAN y SPEARMAN acabaron retractándose de diversas maneras; BURT cavó aún más su hoyo), el daño estaba hecho. Se asume de forma generalizada que hemos nacido con cuotas fijas de inteligencia. Lo más halagador que podemos decirle a unos padres orgullosos es: "es una niña muy inteligente" (lo es y lo será para el resto de su vida). Ahora, es raro que hagamos tests de inteligencia, pero las escuelas y la selección de personal están llenas de tests de capacidad y de aptitudes, que son poco más que un cambio de rótulo de aquellos. Prevemos los resultados educativos basándonos en ellos, porque creemos que hemos dado con la capacidad subyacente, en vez de con el rendimiento actual.

La alternativa de BINET

Vuelvo a BINET porque, al iniciar las pruebas de inteligencia, tenía una visión muy diferente de su finalidad y sus consecuencias. Para BINET, la inteligencia se refería a la capacidad de aprender y de aprovechar la enseñanza. Tenía menos que ver con una capacidad subyacente de aprender que con lo ya aprendido antes y durante la escolaridad. Lo que le preocupaba era que, para algunos niños, este aprendizaje había sido insuficiente para desenvolverse adecuadamente en la enseñanza general, por lo que necesitaban ayuda extra. Lo que se pretendía era mejorar la inteligencia:

> En este sentido práctico, el único accesible para nosotros, decimos que la inteligencia de estos niños ha aumentado. Hemos incrementado lo que constituye la inteligencia de un alumno, la capacidad de aprender y asimilar la instrucción.
>
> (BINET, 1909, pág. 104.)

Criticaba severamente a esos "pensadores modernos que parecen haber dado su apoyo moral a estos deplorables veredictos afirmando que la inteligencia de un individuo es una cantidad fija, una cantidad que no puede incrementarse. Debemos protestar y reaccionar contra este pesimismo brutal; debemos tratar de demostrar que carece de fundamento" (págs. 100-101).

Esta es una visión de cómo creo que debemos contemplar la inteligencia. Las expectativas educativas se han visto perseguidas por diversas formas de este "pesimismo brutal", que ha venido muy bien a los privilegiados. "Últimamente, su inteligencia ha mejorado" parece todavía una declaración incorrecta; probablemente, preferiríamos hablar de "rendimiento" o "puntuaciones en los tests". Sin embargo, los desarrollos de la psicología cognitiva y el interés educativo por "aprender a aprender" apuntan en la dirección de que los individuos son capaces de modificar sus capacidades intelectuales (véase el Capítulo VII). Por eso, esto no son meros desiderata, sino la voluntad de arrancar la inteligencia de las manos de esos pesimistas brutales que la han convertido en una cantidad fija y heredada.

Stephen GOULD resumía así la postura y la contribución de BINET:

> **1.** Las puntuaciones son un recurso práctico; no respaldan ninguna teoría del intelecto. No definen nada innato ni permanente. No podemos denominar lo que miden como "inteligencia" o cualquier otra entidad cosificada.
> **2.** La escala es una guía aproximada, empírica, para identificar... a los niños que necesitan una ayuda especial. No es un instrumento para clasificar a los niños normales.
> **3.** Con independencia de la causa de la dificultad de los niños identificados como necesitados de ayuda, hay que hacer hincapié en la mejora mediante una formación especial. Las bajas puntuaciones no deben utilizarse para marcar a los niños como innatamente incapaces.
>
> (1996, pág. 185.)

Si volvemos a las cuestiones iniciales de la evaluación acerca de la finalidad, la adecuación a la finalidad y las consecuencias, hacen falta algunas respuestas convincentes acerca de las razones por la que cada vez utilizamos más los tests de capacidad y de aptitudes. Sospecho que todavía está en pie la creencia de que con ellos se llega a una capacidad subyacente independiente del rendimiento. En Inglaterra, hay en la actualidad un programa experimental para utilizar tests de capacidad con objeto de identificar a alumnas y alumnos que tienen talento pero reciben una enseñanza deficiente en escuelas deprimidas. Las pruebas, sin embargo, siguen demostrando que hay unas correlaciones masivas entre estas puntuaciones de "capacidad" y el rendimiento en los exámenes, lo que no sorprende en absoluto si ambos son productos del mismo ambiente de aprendizaje [22]. Paradójicamente, esta demanda de test de ingreso del estilo del estadounidense SAT llega en un momento en que el *Educational Testing Service* (ETS) ya no utiliza "aptitud", de manera que lo que antes era el *Scholastic Aptitude Test* ahora no es más que el SAT, un título de tres letras que carece de significado. ¿Por qué?

[22] El impulso se lo ha dado el filántropo Peter Lampl, cuya *Sutton Trust* ha realizado una importante labor en Inglaterra, demostrando que las oportunidades educativas y el ingreso en la universidad siguen estando sesgados a favor de los privilegiados. En parte, está financiado por el Gobierno y las universidades de élite participan activamente. La fase inicial del uso de los SATS estadounidenses modificados identificó únicamente a 29 de los 1.200 estudiantes que no habían conseguido los tres resultados necesarios en el examen de grado GCE A (interpretados como rendimiento "enseñado") que les hubieran dado acceso a Cambridge, pero cuyas puntuaciones en el SAT les habrían permitido acceder a *Harvard*, un resultado muy marginal para el coste de la evaluación. Para obtener más información sobre el proyecto *Uni Que*, véase: http://www.nfer.ac.uk.

Porque comercializar un test de capacidad cuyas puntuaciones podían mejorar mediante la actuación de los servicios de ayuda al alumnado parecía una postura difícil de defender, de ahí que: "El SAT evalúa cómo analizas y resuelves problemas, competencias que aprendiste en la escuela y que necesitarás en la universidad" (*College Board*, pág. 1).

Mi propia reformulación asume la postura de BINET y la adapta a nuestra cultura actual de administración de tests. En ella, es fundamental la idea de que consideramos los tests de inteligencia y de capacidad como *tests de rendimiento generalizado*. Su capacidad predictiva puede interpretarse como una medida del rendimiento actual, en vez de serlo de una capacidad subyacente independiente. *La "capacidad" no es la causa del rendimiento, sino una forma del mismo*. Si la inteligencia y la capacidad son consecuencias de nuestro aprendizaje y de nuestras experiencias, pueden cambiar. El interrogante: "¿hemos mejorado su inteligencia?" se convierte en una pregunta muy sensata.

BINET quería que los tests de inteligencia se limitasen a identificar a los alumnos con necesidades educativas especiales. Es ya demasiado tarde para reclamarlo; la tecnología de la evaluación significa que los caballos se han desbocado. Sin embargo, merece la pena señalar que quienes han permanecido en esta tradición, por ejemplo, los psicólogos educativos, a menudo se han mantenido próximos a su idea, y esto a pesar de la presión de las puntuaciones de corte del CI (70 para la escolarización especial, etc.). Las *British Ability Scales*, la alternativa al Stanford-Binet, hace más hincapié en el perfil de las destrezas que en la agregación de puntuaciones. La tradición de la "evaluación dinámica" asociada con Reuven Feuerstein destaca de modo especial[23]; aquí, el procedimiento diagnóstico consiste en ver cuántos progresos pueden hacerse con ayuda adulta: "la capacidad de aprender y asimilar la instrucción".

Ésta es mi ruta preferida de escape del monstruo que ha creado la evaluación: el CI innato y fijo. Otros han tomado otras rutas, haciendo hincapié en que no hay una única inteligencia general, *g*, sino múltiples inteligencias, o marginando el CI e insistiendo en la importancia de la inteligencia emocional. Relacionamos las *inteligencias múltiples* con el trabajo de Howard GARDNER, aunque forma parte de una tradición más antigua. Daniel GOLEMAN es el promotor de la *inteligencia emocional*. A estos enfoques y sus limitaciones dedicaremos seguidamente nuestra atención.

[23] Véanse referencias en: http://dynamicassessment.com.

CAPÍTULO III

El movimiento de oposición: Inteligencias múltiples e inteligencia emocional

> ... la idea impracticable de que, al clasificar a los niños según sus diversas capacidades, no necesitamos considerar su grado de capacidad general... sigue el principio de la carrera sin reglas* del País de las Maravillas, en la que todo el mundo gana y consigue algún tipo de premio.
>
> (Cyril BURT, 1955.)

Como hemos visto en el Capítulo II, para los psicómetras del CI británicos y estadounidenses, la inteligencia era una cuestión de ganadores y perdedores. La herencia genética mantenía a los privilegiados en las primeras posiciones y se preveía que los pobres se quedaran atrás. Para BURT, "debe ser una parte esencial de la educación del niño enseñarle a encarar una posible derrota en el 11+ (o en cualquier otro examen), del mismo modo que debe aprender a asumir la derrota en una carrera de media milla, en un combate con guantes de boxeo o en un partido de fútbol con una escuela rival" (1959, pág. 123). Todo esto lo decía un hombre que se entrenaba desde una edad muy temprana, mientras su padre le enseñaba las declinaciones latinas "una mañana tras otra cuando todavía estaba en la cuna"[1].

Pero siempre ha habido quien se ha resistido a la afirmación de una única inteligencia central (*g*), con su simple clasificación de niños. La oposición ha asumido diversas formas y tamaños, y he seleccionado tres ejemplos para este capítulo:

1. Enfoques psicométricos rivales que "descubren" múltiples factores.
2. Las *inteligencias múltiples*, de Howard GARDNER, que se basan en la psicología evolutiva.

---

* En esta cita, BURT alude a la *caucus-race* del Capítulo III de *Alicia en el País de las Maravillas*: *A Caucus-Race and a Long Tale* ("Una carrera de partido y un largo cuento") La expresión no tiene una traducción precisa. Como se trata de una carrera sin reglas ni duración definida, hemos optado por la traducción que se ofrece. (*N. del T.*)

[1] J. WHITE (2005, pág. 430).

3. La *inteligencia emocional*, de Daniel GOLEMAN, que minimiza la importancia de la inteligencia "académica" que representan las puntuaciones de CI.

Los tres enfoques presentan una visión más amplia de lo que es posible, escapando del "pesimismo brutal" del CI hereditario. Estimulan también unas prácticas más imaginativas de enseñanza y aprendizaje, que les han dado popularidad entre educadores y formadores. Sin embargo, aquí, mi argumento es que todos ellos caen en la misma trampa de cosificar sus formas alternativas de inteligencia, a menudo utilizando planes de evaluación de limitada validez. Lo que han hecho es ampliar el concepto de inteligencia, que lo hace más aceptable, sin cuestionar necesariamente sus características de "innata y fija". BURT cayó en la cuenta de esto en su ataque[2] contra "la carrera sin reglas del País de las Maravillas": la inteligencia se hacía aceptable afirmando que todo el mundo la tiene en una única forma, por lo que no tenemos que ser tan críticos con respecto a ella.

La primera oposición a la identificación de SPEARMAN de una única inteligencia general (*g*) llegó de otros psicómetras que partían de supuestos diferentes acerca de la naturaleza de la inteligencia. Al utilizar métodos diferentes de análisis factorial, consiguieron generar "múltiples inteligencias". Louis THURSTONE, con su *Primary Mental Abilities*, erigió la oposición contemporánea más directa a SPEARMAN y BURT. Las *inteligencias múltiples* (IM) de Howard GARDNER son una expresión muy diferente de esta tradición de las "facultades mentales". Su enfoque es el de un psicólogo evolutivo más que el de un psicómetra, y extrae sus pruebas de una amalgama de psicología académica, especulación evolutiva y neurología. Esto le permite generar aproximadamente ocho (el número no es definitivo) inteligencias independientes que las personas combinan de distintas maneras. La tercera alternativa es la de la *inteligencia emocional* (IE), de Daniel GOLEMAN, que trata de marginar el anteriormente predominante CI afirmando que el éxito depende más de las inteligencias social y emocional. Aunque la inteligencia emocional es la antítesis de los tests de CI, corre un riesgo similar de definir quiénes somos cosificando las formas social y personal de la inteligencia.

La multiplicación de las inteligencias: La tradición del análisis factorial

Comenzamos señalando que ciertos psicómetras que utilizaban otras formas de análisis factorial para demostrar múltiples formas de inteligencia se oponían al concepto de "inteligencia general" (*g*). Como los teóricos de *g*, también ellos desarrollaron métodos estadísticos para respaldar sus creencias previas acerca de las diferentes "facultades" de la mente.

Louis THURSTONE constituía la máxima amenaza contemporánea para las alegaciones de SPEARMAN. Partía de la base de que la inteligencia no puede reducir-

[2] La "carrera sin reglas" también aparece en *All Must Have Prizes*, de Melanie PHILLIPS (1996), "una virulenta crítica del modo en que han sido traicionados los niños de Gran Bretaña". No es ninguna sorpresa; la autora comparte las ideas de BURT sobre la selección, defiende las *grammar schools* y lamenta el descenso de los niveles.

se a una única medida sobre una única escala y, utilizando una forma diferente de análisis factorial[3] sobre tests muy similares, terminó seleccionando 7 capacidades mentales primarias (había comenzado con 13). Su crítica de *g* fue demoledora:

> Un factor así puede encontrarse siempre de forma rutinaria para cualquier conjunto de tests correlacionados y no significa nada más ni menos que la media de todas las capacidades requeridas por la batería [de tests] en conjunto. En consecuencia, varía de una batería a otra y no tiene una significación psicológica fundamental más allá de la arbitraria colección de tests que se le haya ocurrido a alguien juntar... No puede interesarnos un factor general que solo es la media de cualquier colección aleatoria de tests.
>
> (1940, pág. 208.)

En cambio, THURSTONE pensaba que había descubierto unas entidades mentales reales que no variaban a causa del test. No solo eso, cuando los ítemes se agrupaban en torno a estas, *g* desaparecía. Esto se debía a que no quedaba nada a lo que vincularla, dado que los datos estaban proyectados sobre capacidades específicas. Por ejemplo, si *g* es el resultado de correlacionar datos de ítemes matemáticos y de otros verbales, desaparece si hay escalas independientes matemática y verbal.

Esto lo hacía vulnerable a la misma acusación que él había hecho contra SPEARMAN de que sus capacidades mentales primarias (PMAs *) eran el producto de sus tests, en vez de ser independientes de ellos, y también ellas variaban según distintos tests. Un caso embarazoso de este tipo fue el "factor Xi", el recuento de puntos, que aparecía en tres tests y que no pudo ajustar estadísticamente a ninguna de sus PMAs. Para THURSTONE, este caso se produjo por haber prescindido previamente de otro "vector de la mente". La interpretación más obvia era que se trataba de meros artefactos de las tareas mismas, pero THURSTONE no podía aceptar esto a causa de su creencia de que las PMAs representaban entidades reales. Por eso, estaba en el mismo bando que SPEARMAN y BURT. Simplemente, había partido de supuestos iniciales diferentes acerca de la estructura de la mente (relacionada con la tradición psicológica de las facultades), desarrollando a continuación procedimientos estadísticos para demostrarlos. Como SPEARMAN y BURT, empleó la biología especulativa para cosificar estas capacidades:

> Es muy probable que las capacidades mentales primarias queden perfectamente aisladas por métodos factoriales antes de que sean verificadas por los métodos de la neurología o la genética. Al final, los resultados de los diversos métodos de investigación de los mismos fenómenos tienen que coincidir.
>
> (1938, pág. 2.)

Una vez más, nos encontramos con unas creencias fuertemente arraigadas, el desarrollo de métodos de evaluación que las respalden, la cosificación de las

---

[3] Una descripción muy legible de los distintos enfoques del análisis factorial puede encontrarse en el capítulo 6 de *La falsa medida del hombre*, de Stephen GOULD (2004 [1996]).

[*] "PMAs" es abreviatura de *primary mental abilities*: "capacidades mentales primarias". (*N. del T.*)

ideas y, después, la especulación sobre sus bases biológicas. R. D. TUDDENHAM resumió muy bien el obstáculo ocupacional del psicómetra:

> Las continuas dificultades con el análisis factorial durante el último medio siglo indican que quizá los modelos que conceptualizan la inteligencia por medio de un número finito de dimensiones lineales encierren algo fundamentalmente erróneo. A la máxima del estadístico de que todo lo que existe puede medirse, el analítico factorial ha añadido el supuesto de que todo lo que puede "medirse" tiene que existir. Sin embargo, la relación puede no ser reversible y el supuesto puede ser falso.
>
> (1962, pág. 516.)

THURSTONE representaba una línea de la psicología, que sigue muy presente entre nosotros hoy día, que considera la mente como una serie de capacidades independientes. Aunque su convicción de la existencia de sus PMAs no viniera al caso, THURSTONE nos hizo el gran servicio de cuestionar la idea de *g*, una inteligencia general que podía representarse por un único número. Para él, era preferible tener un perfil de todos los factores primarios cuya significación era conocida. Desde su postura más igualitaria, esto formaba parte de su deseo de "diferenciar nuestro tratamiento de las personas, reconociendo a cada una por sus valores mentales y físicos que la hacen única como individuo" (1946, pág. 112)[4].

Desenterrando las facultades de la mente: Las inteligencias múltiples *de Howard GARDNER*

La idea de que cada persona es una combinación única de capacidades independientes condujo a las "inteligencias múltiples" (IM) de Howard GARDNER, un enfoque que han hecho suyo rápidamente los teóricos de la educación de todo el mundo. Presenta un enfoque que rompe el estricto molde de la inteligencia académica en el que se desenvolvía SPEARMAN y, en gran medida, THURSTONE. GARDNER, como psicólogo evolutivo, representa la tradición de las "facultades de la mente", en la que la mente es producto de una serie de capacidades innatas dife-

[4] La artificialidad de la derivación de constructos mediante procedimientos estadísticos queda maravillosamente ilustrada por esta tradición analítica factorial alternativa. El extremo fue el modelo de inteligencia de "estructura del intelecto" de 150 factores de J. P. GUILFORD, que desarrolló desde la década de 1960 en adelante y que sospecho que ha sobrevivido, sobre todo, porque a los autores de libros de texto todavía les gusta repetir su diagrama cúbico de $5 \times 6 \times 5$, con sus 150 celdas. La lógica básica tiene cierto sentido: un factor es producto de *contenidos* (por ej., simbólicos, visuales), *productos* (por ej., unidades, implicaciones) y *operaciones* (por ej., memoria, transformaciones), por lo que la memoria de unidades simbólicas es uno de esos factores. Sin embargo, eso significa que acabamos con un cubo cargado de factores engañosos, por ejemplo, la transformación de implicaciones visuales. El mismo GUILFORD solo pudo demostrar la existencia de 105 de los 150 factores, la mayoría de los cuales han sido cuestionados posteriormente.

GUILFORD no estaba solo en esta generación de múltiples formas de inteligencia. En 1993, John CARROLL publicó un masivo reanálisis de datos disponibles de tests, que produjo 65-69 capacidades mentales primarias. Reconocía que habría que ordenarlas jerárquicamente en factores cognitivos más amplios, de los que escogía ocho, incluyendo los del estilo de *inteligencia fluida* (que abarcaría razonamiento, inducción, etc.), *inteligencia cristalizada* (que incluiría la comprensión verbal, la capacidad ortográfica, etc.) y la *velocidad de decisión* (por ejemplo, el tiempo de reacción).

rentes (por ejemplo, para la adquisición del lenguaje, la memoria). Después, éstas se desarrollan mediante la experiencia. Su forma de identificar estas capacidades diferentes es completamente distinta de la de la tradición psicométrica, aunque también él trata por todos los medios de determinar el número exacto de estas inteligencias. Por el momento, son ocho (y otra "a medias"):

1. *Lingüística:* sensibilidad al lenguaje escrito y hablado, para aprender idiomas y para utilizar el lenguaje para alcanzar determinados objetivos (como muestran, por ejemplo, los juristas, los escritores y los poetas).
2. *Lógico-matemática:* capacidad de analizar problemas, realizar operaciones matemáticas e investigar las cuestiones científicamente.
3. *Musical:* competencia para ejecutar, componer y apreciar la música.
4. *Corporal-cinestésica:* utilizar el propio cuerpo para resolver problemas o crear productos (actores y atletas, cirujanos y mecánicos).
5. *Espacial:* manipulación de modelos de espacios amplios (pilotos, navegantes) y de espacios más cerrados (escultores, ajedrecistas).
6. *Interpersonal:* la capacidad de comprender las intenciones, motivaciones y deseos de otras personas (vendedores, políticos, docentes).
7. *Intrapersonal:* la capacidad de comprenderse uno mismo para regular la propia vida.
8. *Naturalista:* recién llegada, es la capacidad de reconocer y clasificar las especies en su medio;

   8,5. *Existencial:* solo se acepta aún con reservas, de ahí la "mitad", pero gira en torno a la preocupación por las "cuestiones últimas".

Las *inteligencias múltiples* de GARDNER justifican una visión humana que presenta unas perspectivas mucho más ricas del aprendizaje de los niños, el currículum y cómo pueden atraer los docentes a sus alumnos. Las inteligencias mismas son idiosincrásicas y cuestionables, pero han liberado muchas escuelas y aulas de las limitaciones del estricto enseñar para el examen. También ha rechazado los tests de papel y lápiz y sus puntuaciones para determinar el perfil de una persona, prefiriendo unas formas de evaluación más auténticas, basadas en la actividad. Ésta es otra razón de su popularidad entre los profesionales que tratan de oponerse a la influencia de los regímenes insensibilizadores de tests. Su enfoque presta apoyo también a los planes políticos acerca de la inclusión y la "personalización", dado que considera que cada persona presenta un equilibrio único de competencias y necesidades.

No obstante, para hacer esto, ha tenido que justificar su enfoque acercándose a unas inteligencias "invisibles". Como son entidades innatas, que se manifiestan de forma exclusiva en cada niña o niño, tenemos el imperativo moral de ocuparnos de ellas: una sólida base a partir de la cual exigir reformas. En muchos sentidos, esto es más fácil que defender un currículum más rico y un tratamiento más respetuoso de los niños desde un punto de vista social o de aprendizaje ("¿qué clase de aprendices queremos para el siglo XXI?"), una tarea que intento realizar en este libro. Por tanto, aunque ha evitado utilizar la psicometría, ha creado unas inteligencias evaluando su idoneidad con respecto a unos criterios que él mismo ha desarrollado. Habiéndoles dado el ser al hablar de ellas, ahora se han

convertido en "reales" en el pensamiento de la gente hasta un punto que, a veces, ha alarmado al mismo GARDNER cuando se han utilizado para clasificar a los estudiantes. En su *Intelligence Reframed*, de 1999*, cuenta que se indignó al descubrir que un Estado de Australia, que había basado su programa educativo en las IM, estaba alineando ciertos grupos étnicos con determinadas inteligencias (y determinadas debilidades intelectuales). Viajó a Australia, denunció el programa en la televisión y, a continuación, el programa fue abandonado. Creo, sin embargo, que subestima las formas menos espectaculares de convertir sus inteligencias en maneras de clasificar a los niños, en vez de comprenderlos.

Así, GARDNER está ofreciendo una forma diferente del procedimiento utilizado por los analistas factoriales para inventar y, después, cosificar constructos. En un revelador aparte, el mismo GARDNER comenta:

> Me he preguntado qué hubiese ocurrido si hubiera escrito un libro con el título: *Seven Human Gifts* o *The Seven Falculties of the Human Mind* (*Siete talentos humanos* o: *Las siete facultades de la mente humana*). Sospecho que no hubiese atraído mucha atención. Resulta aleccionador pensar que el título puede influir tanto en el mundo erudito, pero tengo pocas dudas de que mi decisión de escribir sobre las "inteligencias humanas" fue profética.
>
> (Pág. 34.)

El punto de partida de GARDNER es muy distinto del utilizado por el psicómetra: "Comencé por los problemas que resuelven los seres humanos y los productos que aprecian. En cierto sentido, me remonté a las inteligencias que deben ser las responsables" (2006, pág. 21). La fuerza de su enfoque reside en que, al menos, estas inteligencias parecen "reales" en comparación con los estrictos y artificiales constructos de los psicómetras. Su vulnerabilidad estriba en decidir qué se aprecia y quién lo aprecia. Esto significa que las inteligencias están socialmente enraizadas y GARDNER tiene que demostrar que no son simplemente sus preferencias culturales, por ejemplo, su amor a la música y su respeto hacia el genio.

La definición de "inteligencia"

Lo que GARDNER quiere decir con "inteligencia" resulta escurridizo y ha evolucionado a lo largo de los 20 últimos años, a menudo como respuesta a las críticas. Su definición original, de 1983, de la inteligencia era: "la capacidad de resolver problemas o de crear productos que se valoren en uno o más ambientes culturales". En 1999, había redefinido la inteligencia como "el potencial biopsicológico para procesar información que pueda activarse en un ambiente cultural para resolver problemas o crear productos que sean de valor en una cultura" (pág. 34). El cambio es sutil y problemático: la inteligencia se ha convertido ahora en una esencia vaga, una especie de entidad neural que puede o no plasmarse en la realidad en un determinado contexto social. En palabras de John Stuart

* Hay traducción al castellano: *La inteligencia reformulada: las inteligencias múltiples en el siglo XXI* (trad.: Genís SÁNCHEZ BARBERÁN). Barcelona: Paidós Ibérica, 2003. *(N. del T.)*

MILL, se ha convertido en "algo particularmente abstruso y misterioso". Ahora existe independientemente del dominio en el que puede expresarse, mientras que, originalmente, la expresión (el "estado final") formaba parte de la inteligencia, de manera que la competencia musical demostraba la inteligencia musical.

Una razón de este cambio fue lo que, para disgusto de GARDNER, ciertos programas educativos habían empezado a identificar inteligencias con determinadas asignaturas, de manera que un profesor podía decir: "Juanito no puede aprender Geometría porque no tiene inteligencia espacial, a lo que GARDNER responde: "Sin duda, la inteligencia espacial es útil para aprender Geometría, pero hay más de una manera de dominar la Geometría" (2006, pág. 32). Una actividad concreta no puede ser la simple expresión de una inteligencia, por lo que no podemos medirla directamente. El filósofo John WHITE ha observado que esta reciente separación de la inteligencia del "dominio" (la tarea humana socialmente estructurada, por ejemplo, la geometría, la música rap o la cocina) hace ininteligible la teoría. Esto se debe a que los estados finales se han convertido en actividades (dominios) independientes de una inteligencia, mientras que los criterios para determinar una inteligencia se basan en la demostración de estos "estados finales" (2005b, pág. 9).

Todo esto puede parecer un poco "quisquilloso", pero su importancia estriba en que, como en el caso de *g*, los constructos se han cosificado y, después, se han elevado a un nivel neurológico especulativo en el que no pueden cuestionarse. Así, la selección de una inteligencia parte de un juicio social sobre lo que se valora, del que GARDNER dice que "recuerda más un juicio artístico que una evaluación científica" (1983, pág. 63), para acabar como un potencial neurológico heredado que, dadas las experiencias sociales correctas, puede expresarse como parte de un estado final.

Dejando aparte por el momento estas dificultades, ¿cómo escogemos unos candidatos adecuados a IM? GARDNER ha preparado unos prerrequisitos y unos criterios con respecto a los cuales se juzgan, lo que conduce a la pregunta: "¿y cómo escogemos los prerrequisitos y, después, en estos, los criterios?" Pero veamos primero los prerrequisitos:

1. Una competencia intelectual humana debe suponer un conjunto de destrezas de resolución de problemas... representan mi esfuerzo para centrarme en esas virtudes intelectuales que son de suma importancia en un contexto cultural.
2. Entre ellas, estas inteligencias deben recoger "una gama razonablemente completa de los tipos de capacidades valoradas por las culturas humanas" (1983, pág. 62).

Ya hay aquí cierta ambigüedad: ¿esas competencias son locales o universales? A veces una inteligencia propuesta no da la talla, por ejemplo: "la capacidad de reconocer caras", porque no parece muy valorada por algunas culturas (GARDNER no ofrece pruebas de su afirmación), mientras que otras, por ejemplo, la inteligencia musical, no se discuten.

Así, una vez superada la primera criba, ¿cuáles son los criterios para entrar en el "círculo encantado" de las inteligencias múltiples? GARDNER tiene ocho cri-

terios que aplica, aunque quizá no sea necesario satisfacer los ocho. Los extrae de cuatro disciplinas diferentes: ciencias biológicas, análisis lógicos, psicología evolutiva e investigación psicológica tradicional. Todos ellos se relacionan con su propia historia de investigación y representan una forma de evaluación muy diferente, porque no hacen uso de técnicas estadísticas; en cambio, cuentan con una base de pruebas mucho más amplia. Cada uno de estos criterios ha suscitado críticas, como también el modo de aplicarlos [5]. Resumo aquí los criterios y algunas de las objeciones para mostrar cómo las evaluaciones de GARDNER han "dado por la palabra el ser" a las inteligencias.

De las ciencias biológicas proceden:

1. *Posible aislamiento por lesión cerebral*, lo que indica que esta inteligencia tiene una localización específica en el cerebro: un producto del pensamiento sobre las "facultades". Un problema que plantea esto es que, aunque haya áreas localizadas de funciones en el cerebro (por ejemplo, de la vista), hay muchas operaciones psicológicas que no pueden vincularse a un lugar específico (por ejemplo, las destrezas interpersonales).
2. *Historia evolutiva y verosimilitud evolutiva,* por lo que hay un historia para justificar su importancia, por ejemplo, la necesidad de que los humanos primitivos tuviesen conciencia del espacio. Dado que podemos crear historias evolutivas ad hoc de cualquier cosa (ir de compras, la atracción, la religión), ¿qué utilidad tiene el criterio?

El análisis lógico genera:

3. *Una operación fundamental o conjunto fundamental de operaciones identificable.* Son las capacidades ("subinteligencias") que constituyen una inteligencia, por ejemplo, la inteligencia lingüística tiene las operaciones fundamentales de: discriminaciones fonémicas; dominio de la sintaxis; sensibilidad al uso pragmático del lenguaje, y la adquisición de los significados de las palabras. GARDNER reconoce que pueden ser capacidades muy diferentes (debilitando posiblemente el criterio **1**, pues pueden tener sus sedes en diferentes áreas del cerebro), pero mantiene que se utilizan en combinación. La ventaja es que conserva un número manejable de inteligencias; la desventaja es que una inteligencia se convierte en poco más que un nombre de un conjunto de capacidades diferentes, haciendo aún más intangible y remota la "predisposición" neural que representa.
4. *Susceptibilidad a la codificación en un sistema simbólico.* El uso humano de diversos sistemas simbólicos (lingüístico, pictórico, matemático, etc.) es producto de la evolución. GARDNER contempla nuestros principales sistemas sociales de símbolos y especula que nuestro uso eficiente de éstos significa que debe haber una "presintonía" con una inteligencia preexistente. Esta inferencia es innecesaria, sobre todo cuando el autor genera tantos, por ejemplo, la pintura, la escultura y los mapas son sistemas de

[5] Véanse, por ejemplo: la crítica de MATTEWS y cols. (2002, págs. 116-123); HOWE (1997, páginas 125-133), y WHITE (2005).

símbolos espaciales independientes: ¿por qué no considerarlos como expresiones de una evolución cultural que los individuos asimilan como un elemento de su desarrollo en una sociedad? Lo que hay que reconocer es la flexibilidad y la plasticidad del cerebro, en vez de poner en él compartimentos para diferentes símbolos. Además, él mismo se crea el problema del huevo y la gallina: ¿cómo puede proceder esta capacidad evolutiva con los símbolos a los símbolos reales?

GARDNER comenzó su carrera profesional como psicólogo evolutivo muy influido por teorías como las de Jean PIAGET. Sus criterios evolutivos lo reflejan:

5. *Una clara historia evolutiva, junto con un conjunto definible de actuaciones expertas como "estado final".* Este enfoque asume la existencia de homólogos mentales del crecimiento biológico; así, igual que la semilla se desarrolla hasta formar la planta madura, lo mismo ocurre con nuestras capacidades mentales. John WHITE ha cuestionado este símil, señalando que la mejor manera de representar nuestra evolución mental es a modo de *cambios* provocados por la socialización, por ejemplo, nuestros gustos se hacen más complejos, en vez de "desarrollarse" (2005b, pág. 4). La insistencia en los estados finales es, como ya hemos visto, problemática, dado que estos son ahora independientes de la inteligencia. Así, ser un buen matemático, un estado final social, no es lo mismo que tener inteligencia matemática, aunque presumiblemente se haya aprovechado para alcanzar esa pericia.

6. *Pruebas procedentes de individuos excepcionales*, como los niños prodigio o los *idiot savants*, que muestran que el talento en un campo puede no implicar el talento en otros, de manera que estas capacidades son independientes entre sí. A GARDNER, como a GALTON, le fascina el genio y, por tanto, necesita una inteligencia que explique los perfiles intelectuales "láser" que se centran exhaustivamente en uno o dos estados finales, en contraste con los más amplios perfiles "reflectores". La lista del reparto es previsible (Mozart, Einstein, etc.), aunque esta dependencia del genio le cause algunos problemas. Recientemente ha añadido su nueva inteligencia, la *inteligencia naturalista*, para explicar los genios de la clasificación de los objetos naturales, como Darwin y Linneo. Por su artificialidad, esto no parece muy diferente de la capacidad mental primaria de recuento de puntos de THURSTONE.

La psicología popular sobre los niños prodigio, según la cual éstos tienen talentos innatos que se expresan sin esfuerzo y precozmente, también favorece este enfoque. Michael HOWE ha cuestionado esta "explicación del talento" señalando que hay, en general, otras explicaciones que no requieren recurrir a capacidades especiales innatas. Sostiene que los niños prodigio "han recibido casi siempre una ayuda y un estímulo considerables con anterioridad al momento en que se ha considerado que tenían una capacidad sobresaliente" (pág. 132). Indica también que hay personas que han llegado muy alto en las artes y en las ciencias y no han manifestado esta eclosión precoz del talento. HOWE concluye:

> Si, como parece probable, los talentos innatos acaban siendo ficciones, una consecuencia es que se esfuma el apoyo aparente que prestan a la idea de que las diferentes inteligencias de GARDNER son auténticamente independientes, suscitando la posibilidad de que las llamadas inteligencias diferentes no sean, en realidad, sino diferentes capacidades adquiridas.
>
> (Pág. 132.)

Los dos criterios finales de GARDNER están extraídos de la investigación psicológica tradicional:

7. *Respaldo de los trabajos psicológicos experimentales.* Se basa esto en los experimentos que demuestran que las tareas no se interfieren mutuamente cuando se realizan al mismo tiempo, por ejemplo, andar y hablar, lo que sugiere que entran en acción diferentes partes del cerebro, de manera que constituyen inteligencias diferentes. Esta conclusión es, a la vez, vaga y de doble filo: ¿nunca interfiere la música en la reflexión intrapersonal ni las restricciones espaciales en la corporal-cinestésica?

8. *Respaldo de los hallazgos psicométricos.* Esto es aún más arriesgado, al utilizar GARDNER el ejemplo de la débil correlación entre las inteligencias espacial y lingüística para demostrar que, en consecuencia, son diferentes. Algunos psicómetras tienen una opinión diferente al respecto[6], por ejemplo, que hay una elevada correlación entre la inteligencia corporal-cinestésica y la visioespacial, razón por la que los entrenadores intentan que los atletas visualicen. Indican también que se ha demostrado que las capacidades lógicas y las matemáticas son mutuamente independientes, aunque GARDNER las haya combinado. ¿Y qué se sabe de la recién descubierta inteligencia naturalista en relación con la inteligencia lógica y la matemática?

He revisado estos criterios porque son los medios por los que se da existencia a las inteligencias. Superadas las pruebas, GARDNER, mediante "un acto de habla eficiente" (1999, pág. 52) las declara "inteligencia". Su confianza es semejante a la de THURSTONE: estas entidades existen; solo hay que descubrirlas, y puede haber más. Esto plantea la naturaleza subjetiva de esta selección, sobre todo cuando se considera que algunas no satisfacen los criterios, mientras que algunos criterios no se exigen para el "circulo encantado" de las inteligencias. Por ejemplo, ¿por qué no hay una *inteligencia olfativa*, dado que podría sostenerse con facilidad que cumple todos los criterios? ¿En qué compartimento diferenciado del cerebro está almacenada la inteligencia *interpersonal* (y por qué no hay *idiot savants* que solo tengan competencias sociales avanzadas, pero no las de carácter verbal o matemático)? Del mismo modo, ¿en qué es independiente la inteligencia *intrapersonal* de las competencias verbales y corporal-cinestésicas? Estas preguntas cuestionan la legitimidad de las afirmaciones acerca de unas inteligencias diferentes y neuralmente independientes.

[6] Véase MATTEWS y cols. (2002), págs. 121s.

Igualmente problemática es la relación de las actividades "fundamentales" *dentro de* una inteligencia. Como soy poeta, una cumbre de la inteligencia lingüística, ¿significa eso que me desenvolveré igual de bien aprendiendo y hablando en otras lenguas, otro componente de la inteligencia lingüística? Si soy un buen bailarín (algo que mis hijos me aseguran que no soy), esto significa que tengo inteligencia corporal-cinestésica; ¿supone esto que jugaré bien al tenis, que es también una expresión de ella? Si no, ¿supondrá esto que hay una subinteligencia de tenis y una subinteligencia de baile? Si no es cierto nada de esto, ¿qué representa, en realidad, esta misteriosa inteligencia? Me parece que la respuesta de GARDNER es que se trata de una "predisposición", que puede o no expresarse de distintas maneras, según el contexto social y estos "rasgos de capacidad como propiedades duraderas de los individuos que explican sus intereses, esfuerzos y competencias... se inserta en la tradición que se retrotrae hasta SPEARMAN, TERMAN y THORNDIKE" (2006, pág. 40).

Ahora bien, ¿por qué necesitamos un marco de referencia tan elaborado como éste, con tantas contradicciones internas para justificar este enfoque educativamente más rico del currículum, la pedagogía y la evaluación? Volveré a mi campaña de vuelta a BINET para buscar una vía alternativa hacia adelante. Si se considera que la inteligencia es "la capacidad de aprender y asimilar la instrucción" (BINET, 1909, pág. 104), estamos buscando formas de mejorar la inteligencia. Y esto es lo que hacen tantos métodos de GARDNER: la insistencia en un currículum imaginativo, estilos variados de enseñanza y atención a las virtudes y debilidades del aprendiz. La diferencia está en que GARDNER materializa esto en unas disposiciones neurales especulativas.

La otra posibilidad es considerar fundamental el contexto social y cultural, e imaginar el cerebro como flexible y maleable en nuestra respuesta a él, de manera que, por ejemplo, la inteligencia interpersonal sea una compleja integración de muchas competencias. En vez de estimular a los niños para que piensen que tienen unos puntos fuertes innatos en determinadas áreas (y puntos débiles en otras), yo reformularía esto como la tarea social de otorgar mayor valor a sus logros en áreas prácticas, creativas y físicas, junto con los que alcancen en materias más abstractas. El aprendizaje es, en esencia, un proceso social, más que un "despliegue" biológico, por lo que el centro de atención deben ser estos procesos y no unas inteligencias inaccesibles que convierten al aprendiz en problema. Como dice David OLSON:

> los maestros tienden notoriamente a explicar el éxito y el fracaso de los niños en relación con unas presuntas capacidades y unos supuestos estilos de aprendizaje, en vez de por las condiciones que hacen que el aprendizaje sea fácil o difícil.
>
> (2006, pág. 42.)

Por tanto, aunque respeto la contribución de GARDNER a unas mejores ideas y prácticas educativas, se ha basado en una teorización muy sospechosa. Las mismas prácticas pueden justificarse de forma más sencilla (véase el Capítulo VII) y sin recurrir a unas inteligencias especulativas que evaluamos y cosificamos. La visión alternativa de la inteligencia de GARDNER ha ayudado a muchos a liberarse de las restricciones de *g*. El problema es que, en realidad, no se ha despojado del determinismo genético; simplemente, lo ha extendido

de manera más liviana, permitiendo más variedad. Como ha señalado John WHITE:

> La idea... de que todos somos diferentes por nuestras capacidades innatas en las áreas de las IM puede ser tan limitadora para las percepciones de los niños acerca de sí mismos como solía ser la teoría del CI. En cierto sentido, solo es una versión pluralista de este determinismo más antiguo.
>
> (2005b, pág. 10.)

Howard GARDNER no es el único que quiere ampliar la base de lo que se entiende por "inteligencia", con el fin de alejarse de *g*. Otros psicólogos, como Robert STERNBERG y Stephen CECI[7], han elaborado teorías de la inteligencia en las que el papel del contexto social desempeña una función crítica. No obstante, me centraré a continuación a la más popular *inteligencia emocional* como otro ejemplo más de oposición al carácter central de una única inteligencia general.

La inteligencia emocional (IE)

El superventas *Emotional Intelligence: Why it Can Matter More Than IQ*, de Daniel GOLEMAN (1995*), introdujo el concepto en los planes públicos y políticos. El libro fue una explosión periodística ante la falta de reconocimiento que nuestro mundo, muy pendiente del CI, prestaba a las competencias "emocionales". El argumento básico es que nuestra inteligencia emocional pesa más en el mundo real que nuestra puntuación de CI. Los genios emocionalmente insuficientes se verán trabajando para jefes emocionalmente competentes de inteligencia media. Es un mensaje de consuelo; el CI no es lo que importa, sino nuestras competencias emocionales, y éstas pueden aprenderse. Algunas pautas arrancan del nacimiento, por ejemplo, parece que algunos niños se manifiestan desde el primer momento seguros de sí, mientras que otros se muestran

---

[7] El enfoque bioecológico de la inteligencia de Stephen CECI cuadra con el énfasis que este libro pone en el contexto (véase el Capítulo VIII). Sostiene que la forma de desarrollarse de las mentes individuales están configuradas por experiencias de toda la vida. Así, el saber y las competencias mentales son interdependientes, desempeñando el primero un papel fundamental. Por tanto, las aparentes diferencias subyacentes entre individuos y grupos pueden ser en gran parte productos de diferencias entre sociedades, en cuanto a saber cultural que infunden en el niño en desarrollo. CECI insiste también en la maleabilidad de la inteligencia y ha llevado a cabo una serie de estudios que muestran que la conducta inteligente varía según el contexto. Los tests de CI son un contexto "desencarnado" concreto, en el que tienen ventaja quienes cuentan con una extensa escolarización formal (véase: CECI: *On Intelligence. More or Less: A Bio-ecological Treatise on Intellectual Development*, 1996).

El enfoque triárquico de Robert STERNBERG reconoce también las complejas interacciones entre tres facetas diferentes de la conducta inteligente: el mundo mental del individuo y la experiencia del individuo; el mundo externo del individuo, y la forma de utilizar la gente su entorno cotidiano. En relación con las inteligencias múltiples, hay distintas formas de ser inteligente, por ejemplo, el estilo *legislativo* (crear y planear); el estilo *ejecutivo* (implementar actividades), y el estilo *judicial* (supervisar y evaluar) (véase, entre otras muchas publicaciones: STERNBERG: *Beyond IQ: a Triarchic Theory of Intelligence*, 1985).

* Hay traducción al castellano: *La inteligencia emocional* (trad.: Elsa MATEO). Barcelona: Zeta Bolsillo, 2008. (*N. del T.*)

tímidos. El libro llevó a cabo la buena tarea de afirmar la importancia de las competencias interpersonales en la conducta "inteligente". Es interesante señalar que GOLEMAN no ha definido nunca formalmente la "inteligencia emocional" y, en esa etapa, tampoco tenía plan alguno para evaluarla. En 2002, había, por lo menos, 14 tests en el mercado, uno de ellos suyo, e iban apareciendo las secuelas en forma de libros[8]. En 1998, GOLEMAN escribía que la IE era dos veces más importante que el CI, aunque no había prueba científica alguna que respaldara esa afirmación.

Lo que ofrece la inteligencia emocional es una vía positiva de escape de la clasificación de CI. Ni cuestiona realmente las premisas de los tests de CI; se limita a decir simplemente que el CI no es lo importante. Puede interpretarse como algo fijo, pero la IE no lo es, así que centrémonos en esto. A mi juicio, aunque creo que estas competencias emocionales son importantes, se basan en una teoría discutible y pueden convertirse con excesiva facilidad en una concesión para no abordar los aprendizajes difíciles. La evaluación de la IE clasifica nuestros particulares puntos fuertes y débiles en cuanto a IE, lo que, a su vez, configura nuestra forma de vernos como personas y como aprendices. Por tanto, esto "caracteriza emocionalmente a las personas" y lo hace de un modo que puede debilitar el aprendizaje.

Daniel GOLEMAN siempre ha reconocido que el concepto de "inteligencia emocional" proviene de académicos como Peter Salovey, de la *Yale University*. Lo que él hizo fue "empaquetarlo", utilizando una combinación de anécdota periodística y "cómo funciona el cerebro", para convencer a los lectores (y me incluyo) de que estábamos ignorando un elemento vital de la conducta inteligente: el emocional. No pretendo ofrecer una crítica exhaustiva de la "inteligencia emocional", pues Gerald MATTHEWS, Moshe ZEIDNER y Richard ROBERTS ya han publicado 700 páginas al respecto para quienquiera que lo necesite[9]. Mi cometido consiste en recoger las cuestiones relacionadas con la evaluación, para evaluar las consecuencias, y examinar las implicaciones para el aprendizaje.

Al releer, *La inteligencia emocional* veo que es parco tanto en definiciones como en detalles; es una señal de alarma. Y las tropas han acudido. En el nivel aplicado, se ha convertido en el marchamo de un amplio conjunto de proyectos y enfoques en diversos ambientes clínicos, educativos y ocupacionales. Sin embargo, pocos se han evaluado sistemáticamente, y MATHEWS y colaboradores se quedaron "sorprendidos y desconcertados ante el escaso contenido emocional de estos programas" (pág. 465). Su interpretación de esta situación es que muchos programas no se diseñaron específicamente como intervenciones de IE (por

[8] Entre los títulos de GOLEMAN, están: *Working with Emotional Intelligence* (1998; trad. cast.: David GONZÁLEZ RAGA, Fernando MORA ZAHONERO: *La práctica de la inteligencia emocional*. Barcelona: Kairós, 2009); *The New Leaders: Transforming the Art of Leadership into the Science of Results* (2002); *Primal Leadership: Realizing the Power of Emotional Intelligence* (2002; trad. cast.: David GONZÁLEZ RAGA, Fernando MORA ZAHONERO: *El líder resonante crea más: el poder de la inteligencia emocional*. Barcelona: Nuevas Ediciones de Bolsillo, 2003); *Social Intelligence: the New Science of Human Relationships* (2006; trad. cast.: David GONZÁLEZ RAGA: *Inteligencia social: la nueva ciencia de las relaciones humanas*. Barcelona: Kairós, 2010).

[9] Además de utilizar directamente el libro de GOLEMAN (1995): *Emotional Intelligence*, me he basado en la crítica de los desarrollos más generales del movimiento de la "inteligencia emocional" de MATTHEWS y cols. (2002).

ejemplo, programas de prevención de la delincuencia), pero se unieron bajo la bandera de la IE por el reconocimiento público del que hoy goza:

> En la actualidad, la IE cumple una función animadora, ayudando a conseguir apoyos para intervenciones potencialmente útiles (aunque no siempre en la realidad) centradas en una colección heterogénea de competencias emocionales, cognitivas y conductuales.
>
> (pág. 544.)[10]

Los creadores de tests también se han sumado a la causa, preparando inventarios y tests que tienen la misma aura científica que los tests de inteligencia. Hay un auténtico paralelismo en que, igual que en el enfoque de que la "inteligencia es lo que miden los tests de inteligencia", hay poco acuerdo acerca de lo que sea en realidad la inteligencia emocional y, por tanto, se está midiendo todo. En su libro, GOLEMAN solo tenía definiciones desechables: "capacidades... que incluyen el autocontrol, el celo y la constancia, y la capacidad de motivarse uno a sí mismo" (1995, pág. xii); "la aptitud básica para vivir... para refrenar el impulso emocional; para interpretar los sentimientos más íntimos de otras personas; para llevarse bien y fácilmente con los demás" (pág. xiii); "algunos lo llamarían personalidad" (pág. 36). Por tanto, en términos de validez de constructo, tenemos importantes problemas para decidir qué constructo evaluar.

Las confusiones conceptuales crean problemas de evaluación

La falta de acuerdo acerca de lo que sea la inteligencia emocional plantea un importante problema de validez (¿qué evaluamos?) y esto se refleja en la diversidad de tests que se han elaborado. El riesgo es que la inteligencia emocional se defina por exclusión: *es el conjunto de todas las cualidades positivas conectadas con la emoción que no sea el CI*. Para algunos, incluyendo a GOLEMAN, tiene una base esencialmente biológica: el cerebro "reciente" que aborda nuestras emociones "antiguas", o sea, el neocórtex interactuando con el sistema límbico. Para otros, es un conjunto bien definido de destrezas de procesamiento de las emociones que son más específicas y situacionales[11]. A partir de los distintos escritos sobre la IE, es difícil determinar si la inteligencia emocional es:

1. una capacidad general de los seres humanos para manejar los encuentros emocionales;
2. algo que revela diferencias individuales que pueden medirse (y clasificarse);
3. una descripción específica de la situación acerca de cómo maneja una persona la emoción, y
4. todo lo anterior.

[10] Esta es la fuerza de la crítica de ECCLESTONE y HAYES (en preparación) del modo en que muchas preocupaciones sociales están fundiéndose en torno a la emoción (IE; alfabetismo emocional; autoestima) y de cómo se ha convertido esto en un gran negocio.

[11] Por ejemplo: MAYER y cols. (2000).

El problema es que la elección del número **4** es muy popular, de manera que estos elementos se mezclan en una única escala. Volvemos a encontrarnos con el problema *g* del CI, con competencias diferentes reducidas a un único juicio que nos permite clasificar y dar un nombre a un constructo que hemos creado. MATTHEWS y sus colaboradores hacen una alusión directa a la visión pesimista de *The Bell Curve*:

> Lo más terrorífico de la destilación de esta entidad compleja en una única cualidad es que quizá algún día no muy lejano podamos leer en algún libro las ventajas de la élite emocional y el deterioro al que someten a nuestra sociedad los infradotados emocionales[12].
>
> (2002, pág. 522.)

Prosiguen estos autores señalando seis constructos posibles que miden los tests al uso: competencias emocionales básicas; conocimiento abstracto y contextualizado de la emoción (dos constructos); rasgos de personalidad; resultados de encuentros estresantes, e interacción entre la persona y el ambiente. Varios constructos de estos pueden reunirse en un único inventario o test. Sin embargo, no encajan bien; en el mejor de los casos, pueden agruparse en *tests de personalidad de autoinforme* y *tests de capacidad cognitiva basados en la actuación*. Estos dos agrupamientos están poco correlacionados, lo que indica que miden constructos muy diferentes[13].

A pesar de su falta inicial de interés, en 2001, GOLEMAN estaba redactando la segunda versión de su *Emotional Competence Inventory* (ECI), un test que incluía 20 competencias organizadas en cuatro grupos (competencia personal; competencia social; autonomía, y control de las relaciones). Se ha comercializado a través del *Hay/McBer Group*, una empresa responsable de buena parte de la formación para el liderazgo educativo en Inglaterra. La inteligencia emocional ocupa un lugar importante en los programas centrales de formación de directores escolares en Inglaterra.

El juicio de MATHEWS, ZEIDNER y ROBERTS es:

> La raíz del problema es que la IE es un constructo demasiado general para ser útil. Unas intervenciones satisfactorias requieren una comprensión relativamente minuciosa del individuo... En la psicología educativa y en la ocupacional tampoco hay pruebas de que pueda formarse a nadie en una IE genérica y descontextualizada.
>
> (2002, pág. 540.)

[12] Kathryn ECCLESTONE ha señalado que esto está ya implícito en algunas de las afirmaciones de *Antidote*, el grupo de presión del alfabetismo emocional (AE). Considera "esencial" la AE para una sana ciudadanía, y la respuesta del *Department for Education and Skills* fue que "el bienestar emocional de los niños no puede dejarse en manos de sus familias". Véase: ECCLESTONE y HAYES (en preparación).

[13] La correlación entre los dos tests más desarrollados y fiables de cada grupo: el *Bar-On Emotional Quotient Inventory* (EQ-i) (autoinforme en 15 subescalas, por ejemplo: asertividad; optimismo) y el *Multi-Factor Emotional Intelligence Scale* (MEIS) (tests que evalúan la percepción de la emoción en historias; respuesta a escenas), es solo 0,36, desesperadamente baja para tests del mismo constructo (véase: MATHEWS y cols., 2002, capítulo 13).

> La IE (si es algo en realidad) puede ser un constructo transaccional que refleje el grado de ajuste entre la competencia y destrezas de la persona y las demandas adaptativas del ambiente en el que ésta se desenvuelva.
>
> (pág. 531.)

Esta definición es útil. Por desgracia, poco tiene en común con los tests de personalidad de autoinforme, que prestan poca atención a los aspectos situacionales y que están muy correlacionados con los tests de personalidad al uso y ofrecen una información de poco "valor añadido". Otros tests de rendimiento suelen tratar la IE como un tipo de capacidad mental. De este modo, evitan la redundancia, pero plantean el problema de saber si están poniendo a prueba unas capacidades generalizadas o suscitando respuestas a demandas específicas de la situación. Esta forma de evaluación conlleva algunos problemas graves de fiabilidad con respecto al modo de puntuar unas respuestas a situaciones emocionales específicas de una cultura, por ejemplo, clasificar música según los grados de ira y felicidad[14].

La infravaloración de lo situacional

El problema de los tests es que, si son demasiado específicos para una situación, sus resultados no pueden generalizarse. Por eso se tiende a hacerlos más generalizados, robándoles parte de su validez, y dar por supuesto que esto es una "disposición". Entonces, los contextos situacional y social se debilitan. Pero las respuestas emocionales son profundamente situacionales y la IE corre el riesgo de fomentar determinados valores sociales sin reflexionar sobre su carácter situacional. Por ejemplo, ¿acaso es probable que el optimismo de un ciudadano de clase media de Edimburgo lo exprese igual un ciudadano semejante de Bagdad? ¿Los políticos y los generales "insensibles" actúan mejor, a veces, en su contexto concreto?

Muchos escritos sobre la inteligencia emocional pasan por alto las diferencias sociales y culturales. Esto es una consecuencia de considerar la IE como una disposición individual en vez de como una respuesta social. Si la inteligencia emocional tiene que ver con la adaptación a circunstancias emocionales, los detalles de éstas son importantes. La falta de conciencia social tiene menos que ver con el analfabetismo emocional que con un desconocido entorno social. Todos hemos estado en esos ambientes, bares extranjeros o ceremonias desacostumbradas, donde nos hemos sentido sin pistas que nos indiquen cómo deberíamos comportarnos. En vez de una competencia básica, la inteligencia emocional puede reflejar poco más que un estilo de vida estable en contextos en los que sabemos cómo comportarnos con personas que nos gustan y en las que confiamos.

[14] La *Multi-Factor Emotional Intelligence Scale* (MEIS) tiene una línea de identificación emocional que implica clasificar rostros, música, diseños gráficos e historias por su grado de ira, tristeza, felicidad, etc. Hay un problema acerca de lo que representa en realidad el juicio de consenso, que es el fundamento de la puntuación.

Esto nos devuelve a las finalidades (Capítulo Primero) y el problema de validez de no estar claro *qué* se evalúa (el constructo) ni la adecuación a la finalidad de los medios de evaluación. MATHEWS, ZEIDNER y ROBERTS concluyen que:

> Como estas medidas representan adiciones de auténticas capacidades, saberes culturales y contextuales, personalidad y adaptación de la persona al ambiente, las puntuaciones de los tests están abiertas a demasiadas interpretaciones para que sean útiles en la práctica. No podemos interpretar con confianza una baja puntuación como indicadora de una falta fundamental de competencia y no podemos dar por supuesto que un incremento en las puntuaciones del test represente una adquisición de competencia.
>
> (2002, pág. 540.)

Por tanto, se nos deja con la idea de que la defensa de la inteligencia emocional es una protesta útil contra el predominio de una única inteligencia general y una poderosa afirmación de la importancia de las competencias sociales. Lo que debilita su eficacia es la confusión acerca de lo que en realidad representa, complicada por las falsas certidumbres que genera su evaluación (puntuaciones, perfiles, nombres). Como con las demás evaluaciones, el riesgo es que *las categorías se cosifiquen como disposiciones y se subestime la importancia de lo situacional*. Como la IE abarca todo lo que no sea la estricta inteligencia académica, puede reducir las expectativas acerca de lo que puede y debería aprenderse. Reconozco que la IE sugiere un equilibrio para el aprendizaje: es posible que la IE lleve a un mejor clima de escuela y de aula, en el que pueda tener lugar un aprendizaje mejor. No obstante, se corre el riesgo de transmitir el mensaje de que, con independencia de lo que uno sepa, es más importante ser emocionalmente inteligente.

Configurando a las personas

En estos dos últimos capítulos hemos examinado el poder de la evaluación para crear identidades, como personas y como aprendices. La administración de tests de inteligencia es un caso clásico de clasificación y selección de niños de un modo que ha configurado sus identidades y su futuro. He aquí el inolvidable juicio de Patricia BROADFOOT:

> Los tests de inteligencia, como mecanismo de control social, no tuvieron parangón a la hora de enseñar a la mayoría condenada que su fracaso era el resultado de su propia insuficiencia innata.
>
> (1979, pág. 44.)

Los tests de CI eran el producto de las creencias culturales de un grupo de psicómetras y no de unos métodos de evaluación “neutros”. Se desarrollaron unos métodos estadísticos que apoyaron esas creencias previas, en vez de que los resultados llevaran a esas creencias. Esto apunta a la base social de la evaluación. No era esa la intención del precursor de los tests de inteligencia, Alfred BINET, que comprendió que la inteligencia era maleable y modificable mediante la educación.

También hemos visto algunas formas de oposición a estas afirmaciones de una única inteligencia innata y fija: para Louis THURSTONE, fue a través de la demostración estadística de las inteligencias múltiples; para Howard GARDNER, éstas se derivaban psicológica y culturalmente. Ambos constituyen mensajes poderosos que han supuesto beneficios educativos. No han evitado, sin embargo, unas formas similares, aunque más agradables, de cosificación que llevan a unas predisposiciones neurales invisibles.

El mensaje de Daniel GOLEMAN ha consistido en ignorar el CI y centrarse en lo emocional. En este caso, el problema es la falta de claridad con respecto a lo que sea la IE, con la consecuencia de que lo que se infiere a partir de sus diversos instrumentos de evaluación puede inducir a error. Esto conduce a una clasificación muy poco fiable, sobre todo porque no tiene suficientemente en cuenta los factores situacionales. Esto se relaciona con los *estilos de aprendizaje*, que estudiaré en el Capítulo IV.

CAPÍTULO IV

El atractivo de los estilos de aprendizaje

> En seis semanas, os prometo que los chicos que creéis que no pueden aprender estarán aprendiendo bien y con facilidad... la investigación demuestra que cada vez que utilizáis los estilos de aprendizaje, los niños aprenden mejor, rinden más, la escuela les gusta más.
>
> (Rita DUNN, 1990.)

El atractivo de evaluar los estilos de aprendizaje es sencillo e intuitivo: si sabemos cómo aprenden mejor los estudiantes, podemos utilizar este conocimiento para mejorar su rendimiento ajustando la enseñanza y los estilos de aprendizaje. Entonces, ¿por qué los revisamos aquí? Porque también ellos, a través de sus evaluaciones y categorías, corren el riesgo de crear el tipo de aprendices que somos, de "configurar aprendices". Una vez más, hay una tendencia hacia lo biológico y lo "fijo", aunque algunos enfoques procuran resistirse a ello. Como en los casos de las inteligencias múltiples y de la inteligencia emocional, a miles de maestros y formadores les han resultado útiles, por lo que parece que se trata de algo que funciona. Ahora bien, ¿funciona por las razones apuntadas? ¿Y qué confianza podemos depositar en unas evaluaciones que generan estas categorías de aprendices?

Tanto en la educación como en la formación ocupacional, los estilos de aprendizaje constituyen un gran negocio y esto ha llevado a una atmósfera comercial que no ha promovido una cultura abierta de investigación ni la coherencia teórica. Algunas afirmaciones, como la cita inicial de Rita DUNN, despiden un inconfundible tufillo de *marketing*. En su importante revisión de la base probatoria teórica y práctica de los estilos de aprendizaje, de 2004, Frank COFFIELD, David MOSELEY, Elaine HALL y Kathryn ECCLESTONE identificaron 71 instrumentos diferentes de evaluación a disposición de los profesionales, aunque muchos son productos derivados de los inventarios establecidos[1]. Yo me limitaré a revisar tres enfoques populares y dife-

[1] Este capítulo se basa en la excelente revisión y crítica sistemáticas de COFFIELD y cols.: *Learning Styles and Pedagogy in Post-16 Learning* (2004). Agradezco a Kathryn ECCLESTONE sus comentarios sobre los borradores de este capítulo.

rentes. Difieren en la medida en que consideran los estilos de aprendizaje como algo fijo y biológicamente determinado, y en el énfasis que se pone en el contexto y el contenido del aprendizaje, lo que recuerda temas de capítulos anteriores.

¿Qué son los estilos de aprendizaje?

Ésta es una de esas preguntas inocentes que derriba todo el castillo de naipes. Revela de inmediato que no hay consenso acerca de lo que son, de modo muy parecido al que vimos en relación con la inteligencia emocional. Como la mayor parte del considerable volumen de investigaciones sobre los estilos de aprendizaje se hace "dentro" de cada tendencia, con pocas referencias a otros enfoques, no es fácil decir qué tienen en común; incluso, se discuten "aprendizaje" y "estilos". Esto se debe a que algunos de ellos tienen más que ver con estructuras de la personalidad que con el aprendizaje, y los estilos abarcan disposiciones, rasgos, enfoques y preferencias, cada uno de los cuales refleja ideas muy diferentes del contexto y de la flexibilidad.

En 1983, L. CURRY organizó los distintos enfoques en un modelo de "cebolla". En el centro, estaban los estilos cognitivos de personalidad más estables; la capa siguiente abarcaba modelos que abordaban el estilo del procesamiento de información, que tiene un elemento más contextual; la capa externa hacía hincapié en las preferencias de enseñanza, sobre la que podría influirse de forma más directa. COFFIELD y colaboradores señalan que no hay pruebas longitudinales de estabilidad de los estilos cognitivos que, en gran medida, eran creaciones teóricas (según esto, de acuerdo con la metáfora de la cebolla, nada hay en el centro). Su preferencia era ordenar los instrumentos en cinco *familias* que difieren en la medida en que los "estilos de aprendizaje" se consideren basados en la constitución del individuo y relativamente fijos, o más flexibles y abiertos al cambio.

Tres instrumentos populares que representan esta gama son: el *Learning Style Inventory* (LSI), de DUNN, que considera que los estilos de aprendizaje dependen de la constitución del sujeto; el *Learning Style Inventory* (LSI), de KOLB, que considera los estilos de aprendizaje como "preferencias de aprendizaje flexibles y estables", y el *Approaches and Study Skills Inventory for Students* (ASSIST), de ENTWISTLE, que pasa de los estilos de aprendizaje a los "enfoques, estrategias, orientaciones y concepciones del aprendizaje". Las mismas cuestiones se plantean respecto a los tres: ¿se corre el riesgo de cosificar los estilos de aprendizaje?; ¿hasta qué punto es válida su evaluación?; ¿ayuda u obstaculiza el aprendizaje?

Estilos de aprendizaje visual, auditivo, táctil y cinestésico

En la actualidad, muchos docentes están familiarizados con la idea de las modalidades de aprendizaje Visual, Auditiva, Táctil y Cinestésica (incluso, para algunos, las siglas "V-A-T-C" son significativas) y a que una persona pueda ser aprendiza matutina, vespertina o nocturna. Son conceptos extraídos de los influyentes trabajos de Rita y Kenneth DUNN, que comenzaron en la década de 1960 a causa de la preocupación del *New York State Education Department* por los

alumnos y alumnas que presentaban un bajo rendimiento escolar. Se interesaron por los factores que inhibían o estimulaban el aprendizaje del individuo.

El LSI y los desarrollos posteriores[2] han tenido considerable influencia en la educación. En el nivel político, se ha plasmado en el apoyo del Gobierno de los EE.UU. a los "distritos escolares de estilos de aprendizaje", así como en el interés del *Department for Education and Skills* del Reino Unido. Hay una red internacional de profesionales y una red de apoyo[3]. Aunque el carácter comercial de esta empresa puede llevar a afirmaciones rimbombantes, es obvio que, a nivel de los profesionales, ha funcionado para muchos[4]. La cuestión es si ha funcionado por las razones aducidas o por algo más.

El Learning Style Inventory (LSI)

Los DUNN dividieron el *estilo de aprendizaje* en cuatro líneas principales y sus inventarios de autoinforme pretenden identificar las preferencias de los aprendices en cada una de estas áreas más que sus puntos fuertes en ellas. El LSI tiene 104 ítemes con una escala de 5 puntos ("completamente de acuerdo"..."completamente en desacuerdo") para personas de entre 11 y 18 años, y una escala de 3 puntos para los niños y niñas de 9 y 10 años. La línea *ambiental* incluye las preferencias acerca de los niveles de ruido, iluminación, temperatura y diseño de la habitación, por ejemplo: *me gusta escuchar música cuando estoy estudiando; estudio mejor cuando la luz es tenue*. La *sociológica* se ocupa de los grupos preferidos y del papel de los adultos, por ej.: *cuando voy bien en la escuela, las personas mayores de mi familia están orgullosas de mí*; mientras que la *emocional* atañe a la motivación, la necesidad de estructura, la responsabilidad y la constancia. La más influyente ha sido la línea *física*, con su insistencia en las preferencias de modalidad (VATC), hora del día, comida o bebida (*¿prefieres comer/beber/mascar/morder cuando te concentras o prefieres no tomar nada?*) y movilidad *(¿te mueves o te quedas sentado?)*.

Los supuestos subyacentes son que la modalidad y otras preferencias tienen una base biológica: Rita DUNN cree que "tres quintos del estilo vienen impuestos biológicamente" (1990a, pág. 15), y que estas características del estilo hacen que el mismo método de enseñanza sea "maravilloso para unos, terrible para otros" (DUNN y GRIGGS, 1988, pág. 3). Aunque haya cierta flexibilidad y sea posible algún cambio en el estilo de aprendizaje, particularmente en torno a los factores emocionales, se parte de la base de que las preferencias son relativamente estables.

---

[2] Incluyen: *Productivity Environmental Preference Survey* (PEPS), de DUNN y PRICE (1996); *Building Excellence Survey* (BES), de DUNN y RUNDLE (2002), y *Our Wonderful Learning Styles* (OWLS), de GUASTELLO y DUNN (1997).

[3] Véase: http://www.learningstyles.net.

[4] Kathryn ECCLESTONE (2002) comenta que parece que funciona de distintas maneras: para algunos profesores, no significa más que asegurarse de que utilizan diversos estilos *docentes*. La autora no ha encontrado pruebas sistemáticas en el Reino Unido de que los estilos visual, auditivo, táctil y cinestésico hayan llegado a los extremos de diagnosis y respuestas "personalizadas" adecuadas encontrados en los EE.UU., aunque sea necesario hacer una revisión en el Reino Unido.

Finalidad

El LSI y los demás inventarios elaborados por los DUNN se utilizan para identificar estas preferencias de aprendizaje, de manera que los métodos de enseñanza puedan adaptarse al estilo de aprendizaje preferido. En consecuencia, la finalidad es primordialmente diagnóstica; la intención es mejorar el ajuste entre enseñanza y aprendizaje. Cuando las preferencias no son fuertes, los estudiantes pueden adaptarse a diversos métodos de enseñanza. El modelo trata de aprovechar las preferencias, en vez de remediar los puntos débiles, y no estigmatiza distintos tipos de preferencias (aunque los niños superdotados suelen tener un estilo diferente del de los que rinden poco). Dados los orígenes de los trabajos con alumnos poco amigos de la escuela y de bajo rendimiento, intuitivamente, esto tiene sentido y concuerda con los enfoques del aprendizaje que destacan la necesidad de que los aprendices desempeñen un papel activo en su aprendizaje (véase el Capítulo VII). COFFIELD y colaboradores (2004) señalan las virtudes del modelo en cuanto anima a los docentes:

- a ver el potencial de aprendizaje de todos sus alumnos y alumnas; todos pueden beneficiarse de la educación si se atienden sus preferencias;
- a respetar las diferencias, en vez de clasificar negativamente a los estudiantes (“difícil”; “poca capacidad”);
- a ser imaginativos a la hora de adaptar su enseñanza a los estilos de aprendizaje de sus alumnos y alumnas, y a tener en cuenta sus propios estilos de aprendizaje;
- a hablar con los estudiantes acerca del aprendizaje y a facilitarles un vocabulario positivo para las conductas que antes hubiesen sido descritas de forma negativa, por ejemplo, la necesidad de moverse por el aula (“molesta”) puede exponerse en términos de aprendizaje cinestésico.

(págs. 33-34.)

Aunque estas intenciones puedan tener sentido para los docentes y el enfoque les haya resultado atractivo, sigue en pie la necesidad de evaluar la validez de este modelo de estilos de aprendizaje: ¿se basa en fundamentos sólidos? ¿Qué pruebas hay de una predisposición con base biológica a determinadas modalidades físicas y otras preferencias ambientales, sociológicas y emocionales? Ésta es una cuestión clave en relación con los temas de este libro, dado que, como los tests de inteligencia, este modelo da por supuesto un fundamento biológico sólido. ¿“Soy un aprendiz visual” implica una característica innata y, por tanto, la necesidad de adaptar la enseñanza a ella, o sus orígenes son más situacionales y probablemente varíe con el tiempo y el contexto? Es un dilema similar al de las inteligencias múltiples de GARDNER: ¿hay un imperativo biológico que exija que la enseñanza tenga que adaptarse para satisfacerlo, dado que la base biológica no puede cambiarse?

Las pruebas fisiológicas de los DUNN están sacadas de un amplio conjunto de fuentes, en especial del campo de la preferencia de modalidad, que incluye la dominancia hemisférica cerebral. Sin embargo, no se han integrado en un fundamento racional coherente, lo que llevó a SHWERY, en su revisión *Mental Measure-*

ments Yearbook, de 1994, a concluir que "el instrumento sigue lastrado por muchas cuestiones relacionadas con su validez de constructo y por la falta de un paradigma teórico a priori para su desarrollo". COFFIELD y colaboradores concluyeron que: "las referencias a la investigación cerebral, preferencias de hora del día y modalidad en el modelo de DUNN y DUNN se quedan a menudo en el nivel de las afirmaciones populares y no están apoyadas por pruebas científicas" (2004, pág. 34). Esto no quiere decir que los DUNN no hayan analizado muchas pruebas, es más Rita DUNN afirmaba que la investigación de su modelo de estilos de aprendizaje era "sin comparación, más extensa y mucho más completa que la investigación sobre cualquier otro movimiento educativo" (1990b, pág. 223). El problema es que gran parte de esta investigación de apoyo no alcanza los estándares requeridos por los investigadores independientes.

Por ejemplo, el metaanálisis de KAVALE y FORNESS del uso de los estilos de aprendizaje con estudiantes discapacitados para el aprendizaje informa que: "Cuando incluso un examen superficial ponía de manifiesto que un estudio era tan insuficiente que sus datos carecían esencialmente de sentido, se eliminaba del estudio. Ésta es la razón de que solo se incluyeran dos estudios de DUNN" (pág. 358). Como podía suponerse, Rita DUNN no lo aceptó muy bien y a esto le siguió una calurosa discusión, con afirmaciones y réplicas que solo consiguieron confundir a los profesionales. Para nuestros efectos, el mensaje es que las afirmaciones de los DUNN acerca de su modelo, en especial sobre su base biológica, son en gran medida especulativas, en vez de fundarse en pruebas firmes, evaluadas de forma independiente.

Adecuación a la finalidad

Si el objetivo del LSI es identificar las preferencias de los aprendices, ¿hasta qué punto es adecuada a la finalidad la forma de medirlas? La cuestión clave aquí es si un inventario de autoinforme de 104 ítemes puede constituir una indicación válida y fiable de un estilo de aprendizaje que incluye 22 factores (por qué sea este número y cómo se escogieron es otra área de información de validez limitada). Se sabe que las medidas de autoinforme fluctúan según los tiempos y lugares, por lo que basar las puntuaciones factoriales en un número relativamente reducido de preguntas resulta siempre problemático. El manual del LSI informa de fiabilidades test-retest superiores a 0,6 para 21 de 22 factores; sin embargo, este es un criterio laxo (el mínimo aceptable habitual es 0,7) y se corre el riesgo de obtener altos niveles de clasificaciones erróneas.

Si el LSI se está utilizando en sentido diagnóstico como base de discusión sobre las preferencias, puede ser aceptable, ya que las inferencias erróneas pueden negociarse y rectificarse. Sin embargo, a mi juicio, la evaluación no opera así en la práctica; las puntuaciones cosifican los constructos, de tal manera que se le dice al sujeto quién es: un aprendiz táctil que necesita trabajar con iluminación tenue en pequeño grupo en presencia de personas adultas.

Consecuencias

Una de las consecuencias clave es que, como se interpreta que el origen de los estilos de aprendizaje es en gran medida biológico, se consideran relativamente fijos. Esto conduce a enfatizar el *ajuste* de estilos de aprendizaje y enseñanza: la enseñanza es la que debe adaptarse, no el estilo de aprendizaje. Me preocupa que, una vez más, *estemos ante unos rasgos fijos que han sido creados mediante una medida poco fiable*. Y no son pocas las pruebas anecdóticas de personas que aceptan lo que les dicen que son y lo utilizan como excusa para aprender de un modo diferente (si uno es táctil o cinestésico, tiene poco sentido escuchar clases magistrales). Para complicar el asunto, hay escuelas que entregan distintivos para comunicar el tipo de aprendizaje ("Soy un alumno C"), de manera que los maestros lo vean y se adapten a ello.

La respuesta a esto puede ser que, como éstas son las percepciones de los aprendices acerca de cómo aprenden mejor, la evaluación no define el estilo, sino que lo determinan ellos mismos. Se pasa por alto la función de la evaluación, en la forma de los ítemes del inventario, en su configuración, es decir, el LSI está organizado para producir puntuaciones acerca de diferentes preferencias de modalidad, aunque no sea así como nos describiríamos en realidad nosotros mismos. Nos piden que nos definamos en una escala de 1 a 5 en la que a menudo querríamos decir: "sí, pero...", ante un ítem como: *En interiores, a menudo llevo un suéter o una chaqueta*, o: *Me gusta hacer cosas con los adultos*. Así es como un valor positivo —dar a los estudiantes un lenguaje para hablar sobre el aprendizaje— puede empezar a tener consecuencias negativas; los estudiantes permanecerán en su zona cómoda y esto puede generar conductas y creencias autolimitadoras, en vez de arriesgadas.

Adaptar la enseñanza al estilo de aprendizaje

Esto es *adaptar* un concepto crítico. Sigamos el argumento: como nuestras preferencias responden a un impulso biológico y son en gran medida fijas, tenemos que adaptar a ellas la enseñanza, especialmente al abordar material nuevo y difícil. Rita DUNN reivindica contundentemente la eficacia de este proceso:

> Podemos prever que los estudiantes a cuyos estilos de aprendizaje se acomoda la enseñanza rindan un 75% de una desviación típica por encima de los estudiantes a cuyos estilos de aprendizaje no se ha acomodado la enseñanza.
>
> (2003a, pág. 181.)

Aunque la investigación de los DUNN ha respaldado este tipo de reivindicación, la evidencia independiente es mucho menos convincente y presenta algunos hallazgos contradictorios[5]. Por ejemplo, en el estudio de KAVALE y FORNESS antes mencionado, se analizaron los estudios empíricos en los que se hacían corresponder la modalidad con una enseñanza especial de la lectura. Sus conclusiones

[5] Véase: COFFIELD y cols. (2004, págs. 24-30).

fueron que el diagnóstico de la preferencia de modalidad era problemático en sí mismo y que los efectos de tales enfoques eran limitados (magnitud del efecto de 0,14). Decían que, aunque la correlación "tiene un gran atractivo intuitivo,... se encontró poco apoyo empírico... Ni las pruebas adaptadas a la modalidad ni la enseñanza adaptada a la modalidad demostraron su eficacia" (pág. 237).

Con respecto a la cuestión de si, de cara a un aprendizaje eficaz, merece la pena fortalecer nuestras modalidades más débiles, en vez de fiarlo todo a las preferidas, como los DUNN han tomado la ruta "biológica", adoptan la postura más extrema. Para ellos, la cuestión es descubrir los puntos fuertes y actuar basándose en ellos. Sin embargo, observan que es muy probable que los aprendices "superdotados" tengan más de un modo preferido, a veces los cuatro, mientras que los que menos rinden prefieren un enfoque único (táctil o cinestésico). Como veremos con los modelos de KOLB y ENTWISTLE, hay otro punto de vista que dice que deberíamos desarrollar una enseñanza de modalidad mixta, dado que basarse en una modalidad única puede inhibir, en realidad, el aprendizaje: se ha demostrado que el cambio de unas modalidades a otras es creativo.

El descuido de las asignaturas

En todo esto, se descuida precisamente *lo que se está aprendiendo*. Rita DUNN sostiene que "lo que determina que los estudiantes dominen el currículum no es el *contenido*, sino *cómo se enseña ese contenido*" (2003b, pág. 270). Según esto, la pedagogía tiene que conseguir que las condiciones sean adecuadas, preocupándose poco o nada de los conocimientos específicos de la materia. En consecuencia, los materiales de apoyo están llenos de consejos para clase: "rediseña las aulas convencionales con cajas de cartón"; "apaga las luces y lee con los alumnos de bajo rendimiento o siempre que la clase esté inquieta solo con luz natural". Aunque es loable esta preocupación por el ambiente, se corre el riesgo de promover la falta de interés por las exigencias de la materia o del currículum. Si, como sostiene este libro, el aprendizaje es una actividad esencialmente social, muy dependiente de la situación, el mismo aprendiz puede tener que emplear modalidades diferentes en Educación física y en Matemáticas, que pueden diferir, a su vez, de Arte y de Música.

Me gustaría darle la vuelta al argumento de DUNN: *tanto lo que se está aprendiendo como el aprendiz determinarán el estilo de aprendizaje*. El docente tendrá que hacer todo lo posible para que lo que se esté estudiando sea accesible para el aprendiz, y eso puede requerir el uso de modalidades mixtas y una cuidadosa evaluación del punto en el que estén los estudiantes en su aprendizaje (véase el Capítulo VII). Robin ALEXANDER hace una observación similar:

> ... distintas formas de conocer y de comprender exigen diferentes maneras de aprender y de enseñar. Las comprensiones matemática, lingüística, literaria, histórica, científica, artística, tecnológica, económica, religiosa y cívica no son iguales. Unas exigen una base mucho más sólida de conocimientos prácticos y proposicionales, y todas avanzan al máximo sobre la base del compromiso con los conocimientos e ideas vigentes.
>
> (2000, pág. 561.)

Los chicos y las chicas de 15 años a quienes grabó Caroline LODGE[6] llegaron a una postura semejante, aunque en un estilo más irónico:

Linda: Cuando te preguntan sobre el estilo de aprendizaje que prefieres, si te gusta más escuchar o aprendes visualmente... no piensas en eso a diario. En clase de Lenguaje, no piensas: "¡oh!, estoy escuchando": auditivo.
John: Así es como yo aprendo.
Linda: ... y entonces se espera que le digas a la profesora: "Yo aprendo escuchando", pero no lo sabes, no piensas en ello.
Jamie: Utilizas todos. Depende de la clase. Es como en clase de Música: vas a escuchar, ¿no?
Linda: A la clase de Arte vas a mirar.
Jane: O, si te ponen una película en clase de Lenguaje, vas a verla, ¿no?
Jamie: En clase de matemáticas, no vas a escuchar los números [carcajada].
Linda: ¿Qué quieren decirme?

El descuido de lo situacional

Dado que dos de las cuatro variables principales son la *ambiental* y la *sociológica*, puede parecer un poco retorcido criticar el modelo por la falta de atención a los factores situacionales. Sin embargo, lo que estas dos variables contemplan es, esencialmente, el *clima de la clase* (o sea, el ambiente del aula y sus valores y actitudes) y no otras cuestiones más generales de carácter ambiental y sociológico. Así, lo "ambiental" abarca factores como el nivel de ruido, la iluminación, la temperatura y el diseño del aula, mientras que lo "sociológico" solo incluye los grupos de aprendizaje, la presencia de figuras de autoridad, el aprendizaje de diversas maneras y la motivación debida a los adultos.

Lo que falta aquí es el reconocimiento de factores sociológicos más generales que influyen en el aprendizaje. REYNOLDS ha lanzado una furibunda crítica contra la tradición de investigación sobre los estilos de aprendizaje por crear un concepto de aprendizaje individualizado, descontextualizado, que ignora profundas diferencias sociales:

> El mismo concepto de "estilo de aprendizaje" nubla las bases sociales de la diferencia que se expresan en las formas de enfocar el aprendizaje de las persona... la clasificación no es un procedimiento desinteresado, aunque las diferencias sociales se muestren de manera que parezcan reducibles a tecnicismos psicométricos.
>
> (1997, págs. 122, 127.)

El peligro de un modelo de estilos de aprendizaje como el de DUNN y DUNN es que los aprendices, sobre todo los que proceden de ambientes deprimidos, puedan quedar sometidos a una dieta de aprendizajes "básicos" táctiles y cinestési-

[6] LODGE, C. (2001): "An Investigation into Discourses of Learning in Schools". Tesis doctoral no publicada. Londres: *University of London*, *Institute of Education*, págs. 110-111.

cos. Esa situación recuerda mucho la oferta de un currículum (competencias funcionales) y una pedagogía (ejercicios para los tests) restringidos para los alumnos de menor rendimiento, mientras que a los demás se les ofrece un currículum más general e interesante y un amplio repertorio de enfoques de enseñanza (véase el Capítulo VI).

Entonces, ¿por qué tiene tanto éxito?

Es obvio que el modelo de los DUNN ha resultado útil para muchos profesionales de todo el mundo; para decenas de miles de docentes, "VATC" significa algo. Algo bueno tendrá, a pesar de las críticas académicas y sus limitaciones. Creo que sus virtudes cumplen algunas de las condiciones del aprendizaje eficaz (véase el Capítulo VII), aunque su explicación de lo que ocurre es, a veces, muy diferente. Este modelo:

- se centra en el aprendizaje;
- es positivo con respecto al potencial de aprendizaje de los alumnos;
- ofrece a los aprendices cierta autonomía y posibilidad de elección en cuanto a la forma de aprender;
- resalta la importancia de la relación profesor-alumno, en la que se escucha a los aprendices con respecto a su aprendizaje;
- estimula la reflexión sobre lo que ayuda a aprender;
- estimula a los docentes a ser imaginativos acerca de las condiciones y recursos necesarios para el aprendizaje en el aula.

La falta de validez del modelo recae en sus afirmaciones acerca de la base biológica de nuestras preferencias y en su dependencia de un instrumento de limitada fiabilidad para determinar el estilo de aprendizaje. Como las inteligencias múltiples, trata de ofrecer una justificación biológica para justificar sus afirmaciones. El riesgo es que los aprendices *se conviertan en* su perfil y que ello conduzca a una conducta autolimitadora. La idea de hacer corresponder el estilo docente con las preferencias individuales también es cuestionable, dado que infravalora *lo que* se esté aprendiendo al hacer hincapié en el proceso, y puede limitar el desarrollo de un conjunto de estilos de los aprendices, algo que promueve el trabajo de David KOLB, sobre el que volveremos a continuación.

Los cuadrantes del aprendizaje y el aprendizaje experiencial

Muchos lectores habrán visto el cuadrante del aprendizaje de David KOLB o algún derivado del mismo, quizá la clasificación de: *reflexivo, teórico, pragmático, activo*, de HONEY y MUMFORD (2000), o el ciclo de aprendizaje: *hacer, revisar, aprender, aplicar*, de DENNISON y KIRK (1990). Los orígenes del trabajo de David KOLB están en las técnicas de aprendizaje experiencial que introdujo en su enseñanza a estudiantes de administración en los EE.UU. Se percató de que algunos estudiantes tenían preferencias claras por determinadas actividades; por ejem-

plo, unos preferían ejercicios en grupo, mientras que otros respondían mejor a las clases magistrales. A partir de aquí, elaboró un inventario que pretendía identificar estas preferencias.

Una diferencia clave del enfoque de DUNN y DUNN era el supuesto de KOLB de que un estilo de aprendizaje no era un rasgo fijo, sino "una preferencia diferencial de aprendizaje, que cambia ligeramente de una situación a otra. Al mismo tiempo, el estilo de aprendizaje conserva cierta estabilidad a largo plazo" (KOLB, 2000, pág. 8). Esta estabilidad a largo plazo puede deberse, en parte, a que reforzamos nuestras preferencias mediante el trabajo que escogemos y nuestra forma de abordar los problemas. Éstas pueden cambiar si cambia nuestro trabajo, por ejemplo, los ingenieros que acceden a puestos directivos pueden tener que modificar sus estilos de aprendizaje.

KOLB empieza con una visión del aprendizaje como "el proceso en el que se crea el saber mediante la transformación de la experiencia. El saber se deriva de la combinación de adquirir experiencia y transformarla" (1984, pág. 41)[7]. Es un proceso dialéctico, pues tenemos que resolver el tira y afloja entre acción y reflexión; concreción y abstracción, los dos ejes independientes de su cuadrante (véase la Figura 4.1). Si soy una persona cuya forma preferida de resolver problemas

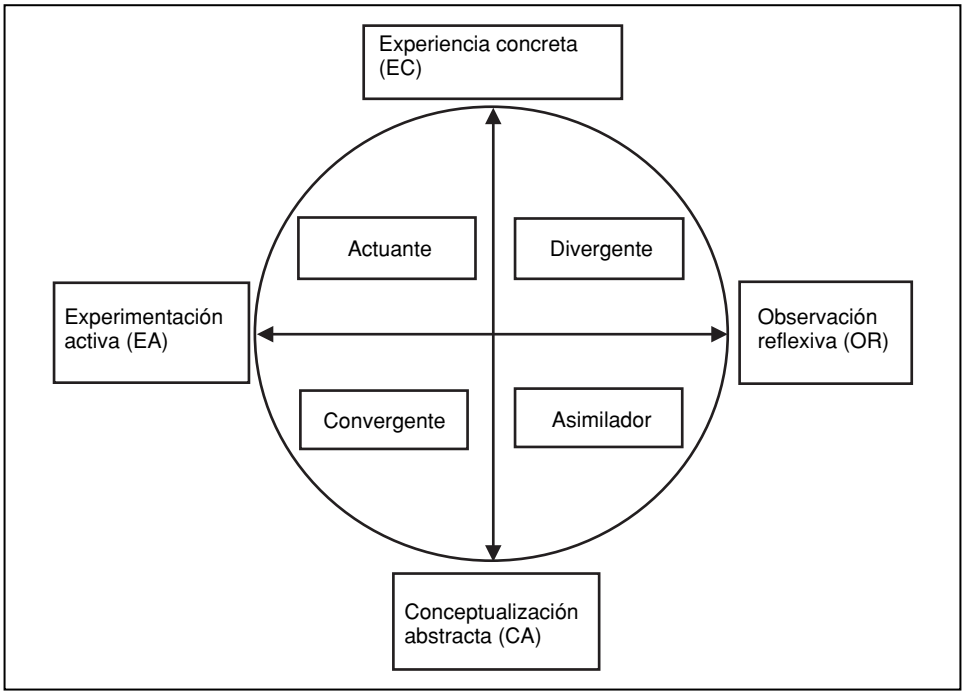


Figura 4.1. *Los cuatro estilos de aprendizaje de KOLB.*

[7] Esta definición concuerda con la definición de ERAUT, de 1997, que he estado utilizando: "Un cambio significativo de capacidad y comprensión".

se basa en la conceptualización abstracta y la experimentación activa, por ejemplo, resolver problemas técnicos, esto me dará un estilo *convergente*. En cambio, un enfoque más orientado a las sensaciones, que utilice la experiencia concreta y la observación reflexiva, se clasifica como *estilo divergente*. Para completar la información, el estilo *asimilador* se caracteriza por la preferencia por la conceptualización abstracta y la observación reflexiva (por ejemplo, el gusto por las ideas y la lógica), mientras que el estilo *actuante** se basa en la experiencia concreta y la experimentación activa (por ejemplo, las personas de "acción", a las que les gustan las experiencias nuevas y el desarrollo de planes).

A diferencia del modelo de los DUNN, estos estilos no se consideran biológicamente fijados; son el resultado de convertirse en la forma preferida de resolver conflictos acerca de cómo enfocar los conocimientos nuevos. KOLB también difiere al contemplar el aprendizaje integrado, la forma más madura de aprendizaje, como un enfoque holístico capaz de integrar los cuatro estilos básicos.

Finalidad y adecuación a la finalidad

Para KOLB, la finalidad de su *Learning Style Inventory* es facilitar "un autoexamen y un comentario interesantes que reconozcan el carácter único, la complejidad y la variabilidad de los enfoques individuales del aprendizaje". Continúa reconociendo lo que es uno de los temas centrales de este libro:

> El peligro radica en la cosificación de los estilos de aprendizaje como rasgos fijos, de manera que los estilos de aprendizaje se conviertan en estereotipos utilizados para encasillar a las personas y su conducta.
>
> (1981, págs. 290-291.)

KOLB reconoció su vulnerabilidad a esta consecuencia no deseada, al utilizar denominaciones como "convergidor"**. En su última versión, ha cambiado el sustantivo por "estilo convergente". Quizá no sea suficiente para evitar este peligro: he trabajado con una escuela en la que se decía a los alumnos que fuesen a su "cuadrante de aprendizaje" en el vestíbulo del centro, y todos sabían qué tipo de aprendices eran.

Como instrumento para centrar la discusión sobre nuestra forma de aprender, las cuestiones relativas a la validez y la fiabilidad son menos importantes que en otras formas de evaluación. No obstante, aunque la base de esto sea un simple instrumento, un inventario de 12 cuestiones, ha generado un campo de investigación masiva y una compleja superestructura teórica. El problema es que el LSI de KOLB no puede respaldar esta carga de expectativas y de actividad.

En el LSI de KOLB, se pide a los sujetos que completen 12 oraciones que describen el aprendizaje. Cada una tiene cuatro alternativas y los sujetos tienen que

---

* En el original inglés aparece *accommodating* y se ha traducido como "acomodador", pero este término no parece denotar el significado del término, por lo que hemos preferido utilizar el término "actuante". (*N. del T.*)

** En el original, aparece el neologismo inglés *converger*, como sustantivo aplicable a quien "converge". (*N. del T.*)

ordenar sus preferencias: una clasificación de elección forzosa. Así, cuando se presenta: *Aprendo mejor a partir de*, hay que clasificar por orden: *teorías racionales*; *relaciones personales*; *la oportunidad de probar y practicar*, y *observación* (dejo al lector que imagine qué cuadrante representa cada opción). A partir de aquí, se calcula la puntuación en cada una de las cuatro modalidades y en las dos dimensiones, indicando éstas la preferencia por un polo u otro.

No hace falta ser psicómetra para intuir que 48 puntuaciones de 12 cuestiones pueden plantear problemas de fiabilidad a una escala que asigna a las personas a uno de los cuatro estilos de aprendizaje y en dos dimensiones. Por tanto, la *fiabilidad* es un problema importante con respecto al modelo. Una vez más, si esto no fuese más que la base de una conversación, podría autocorregirse, pero no es así como funciona la evaluación. El riesgo está en que profesores y estudiantes tomen las puntuaciones, aunque poco fiables, como algo real. Podemos "configurar a los aprendices" sobre la base de una evidencia muy poco fiable.

Las propiedades psicométricas del LSI de KOLB se han discutido desde su presentación, a causa de sus "volátiles" fiabilidades test-retest (se plantea también una cuestión fundamental acerca de la fiabilidad de unos estilos de aprendizaje "flexibles" con el paso del tiempo). STUMPF y FREEDMAN se preguntan: "¿Cómo sabemos si la clasificación de un sujeto como asimilador se debe a características personales, factores situacionales o a un error de medida?" (1981, pág. 297). KOLB reconoce que la fiabilidad de las cuatro escalas básicas es limitada y ha promovido el uso de las puntuaciones "primordialmente para una descripción cualitativa". Cree que la fiabilidad de las dos puntuaciones combinadas puede considerarse "razonable".

Hay una bibliografía considerable y contradictoria en torno a estas fiabilidades, además de plantearse diversas cuestiones acerca de si es posible justificar sus dos ejes[8]. Si no se puede, de ello se derivan consecuencias tanto para la idea del ciclo de aprendizaje como para otros estilos de aprendizaje inspirados en KOLB: podemos estar asignando a estudiantes a cuadrantes fantasmas.

Esto nos deja con una teoría popular y conceptualmente válida del aprendizaje experiencial basada en las obras de John DEWEY, Kurt LEWIN y Jean PIAGET, acompañada de un esquema de evaluación frágil, en el mejor de los casos, y engañoso, en el peor. Este esquema de dudosa validez puede ser la base de un proceso en el que unos estilos flexibles se conviertan en patrones fijos, cosificando unas clasificaciones poco fiables y transformándolas en "el tipo de aprendiz que soy". Y, aunque se reconozcan los factores situacionales, incluyendo las destrezas del sujeto, los estilos de aprendizaje seguirán atribuyéndose a los aprendices mismos, en vez de a una respuesta situacional.

Proseguimos el argumento de que los estilos de aprendizaje dependen de lo que se aprenda y de dónde tenga lugar el aprendizaje, así como del mismo aprendiz, examinando un tercer modelo: los *Learning Approaches*, "enfo-

[8] Véase un resumen en: COFFIELD y cols. (2004), págs. 64-67. Las críticas más perjudiciales son las que cuestionan la validez de las dos dimensiones de las que dependen muchas cosas: la activa-reflexiva y la concreta-abstracta. Diversos estudios analíticos factoriales no han generado estos factores a partir de los datos, y algunos han presentado incluso diferentes combinaciones. Los análisis de WIESTRA y DE JONG solo presentan una dimensión: "aprendizaje reflexivo frente a aprendizaje a través de la actividad" (WIESTRA y JONG , 2002, pág. 439).

ques del aprendizaje", de Noel ENTWISTLE. He seleccionado este modelo menos conocido, cuyos orígenes están en la enseñanza universitaria, porque ofrece ciertas intuiciones constructivas acerca de la mejor manera de incorporar lo situacional a los estilos de aprendizaje.

Enfoques del aprendizaje profundo, superficial y estratégico

Incluso este subtítulo sugiere que vamos a hablar de algo diferente. Noel ENTWISTLE, psicólogo de la *University of Edinburgh*, ha estado trabajando durante más de 30 años sobre las estrategias que utilizan los estudiantes al abordar tareas concretas de aprendizaje. Las expresiones de este trabajo relativas al aprendizaje, que cada vez se utilizan más, son: *profundo, superficial* y *estratégico*. En este modelo, una *estrategia* es el modo de abordar los estudiantes una tarea específica de aprendizaje. Esta estrategia se basa en las exigencias percibidas. Es, por tanto, más flexible que un estilo, que trata de describir cómo prefieren enfocar, por regla general, los estudiantes las tareas de aprendizaje. Por supuesto, esto hace que la evaluación resulte más difícil, pues las estrategias son específicas de la situación, y ENTWISTLE y sus colaboradores resuelven el problema afirmando que los estudiantes muestran una regularidad suficiente "de intenciones y de procedimiento en tareas académicas similares en términos generales para justificar que se midan como una dimensión" (1979, pág. 367). Esta idea constituye un precedente peligroso sobre el que volveremos.

Los inventarios

ENTWISTLE y sus colaboradores han elaborado una serie de inventarios durante los últimos 25 años. Los dos más recientes son el *Approaches and Study Skills Inventory for Students* (ASSIST, 1997) y el *Approaches to Learning and Studying Inventory* (ALSI), que está en fase de desarrollo. El ASSIST es un inventario de 68 ítemes que requieren respuestas a manifestaciones de otros estudiantes sobre el aprendizaje (por ejemplo: *Asegurarte de que recuerdas bien las cosas*, o: *Ver las cosas de un modo diferente y más significativo*); enfoques del estudio (52 ítemes, por ejemplo: *A menudo me cuestiono cosas que oigo en clase o leo en libros*), y preferencias por distintos tipos de organización y enseñanza de la asignatura (por ejemplo: hasta qué punto *me gustan los exámenes que me permitan demostrar que he pensado por mi cuenta en el material de la asignatura*). Cada afirmación se clasifica en una escala de 5 puntos y se pide a los estudiantes que den una respuesta inmediata en relación con la asignatura concreta de la clase.

Las propiedades técnicas de los inventarios han ido fortaleciéndose progresivamente, de manera que las fiabilidades de los tres enfoques principales son satisfactorias (en torno a 0,8), aunque haya que depositar mucha menos confianza en las numerosas subescalas que generan los inventarios. De hecho, una de las cuestiones relativas a la validez de constructo es si, como en los inventarios previos de estilos de aprendizaje, las puntuaciones de estos inventarios de pre-

guntas cerradas pueden apoyar el nivel de interpretación que se hace depender de ellas. De nuevo, si el uso del inventario fuese como una "conversación sobre el aprendizaje" entre estudiantes y profesores, las malas interpretaciones pueden cuestionarse. Sin embargo, las puntuaciones y los perfiles tienen la posibilidad de convertirse en algo más que un punto de partida.

Finalidad

La finalidad pretendida de estos inventarios es la de que estudiantes y profesores puedan reflexionar críticamente sobre los enfoques del aprendizaje y el medio en el que éste tiene lugar. La intención es que ambos puedan modificarse para mejorar la calidad del aprendizaje de los estudiantes. No pretenden predecir el rendimiento; de hecho hay cierta confusión acerca de lo que pueda esperarse de las distintas estrategias: es posible que los enfoques del aprendizaje profundo no consigan los mejores resultados. Una de las consecuencias sobre las que volveremos es que la calidad de la *evaluación sumativa* en una asignatura influirá mucho en el enfoque del aprendizaje.

El modelo subyacente (véase la Tabla 4.1), en el que se basa éste, aprovecha los trabajos de Ference MARTON y Roger SÄLJÖ, que identificaron dos niveles, *superficial* y *profundo*, en el procesamiento del material de aprendizaje. Las concepciones del aprendizaje de los estudiantes influyeron en su enfoque, y estas concepciones pueden cambiar en el transcurso de una carrera, igual que su motivación en la asignatura puede ser intrínseca o extrínseca en diferentes momentos (todos hemos cursado asignaturas que "queríamos aprender" y asignaturas que "teníamos que aprender"). Estas concepciones pueden progresar del simple dualismo (es decir, hay respuestas correctas o erróneas), pasando por el relativismo y, por último, el compromiso: una postura individual coherente en una disciplina. Esto sugiere una progresión desde el aprendizaje superficial al profundo, pero hay un tercer enfoque que introduce ENTWISTLE, el del *aprendizaje estratégico*. Este enfoque está muy condicionado por las exigencias de la evaluación de la asignatura, dado que los enfoques estratégicos tratan de alcanzar las calificaciones más elevadas posibles, en contraste con los enfoques superficiales, que solo pretenden cumplir los requisitos mínimos de la asignatura. Ruth BORLAND, a quien nos referimos en la Introducción, tenía un enfoque estratégico clásico. Me detendré con cierto detalle en estos enfoques, porque se relacionan tanto con el credencialismo (Capítulo V), como con la evaluación para aprender (Capítulo VII).

La Tabla 4.1 provoca varias respuestas. La primera es la dificultad de quedarse en los *enfoques* y no tratarlos como disposiciones ("*aprendices* superficiales"), sobre todo cuando las imágenes de algunos estudiantes brillan ante determinadas características, en especial las superficiales y las estratégicas. La intención es que esto refleje los enfoques de asignaturas específicas, pero la realidad es que clasificamos a los estudiantes. MARTON y SÄLJÖ, en sus entrevistas con estudiantes, descubrieron que estos variaban sus enfoques según las exigencias de una determinada tarea; sin embargo, el problema de un inventario es que esto no lo recoge y puede acabar considerándose como un enfoque fijo.

Tabla 4.1. *Enfoques del aprendizaje y del estudio*

Grupo	Subgrupos
Profundo **Búsqueda de significado** *Intención:* desarrollar ideas por uno mismo	Relacionar ideas con los conocimientos y experiencia previos. Buscar patrones y principios subyacentes. Comprobar las pruebas y relacionarlas con las conclusiones. Examinar con cautela y críticamente la lógica y el argumento. Tener en cuenta el desarrollo de la comprensión mientras se aprende. Interesarse activamente por el contenido de la asignatura.
Superficial **Reproducir** *Intención:* cumplir los requisitos mínimos de la asignatura	Considerar la asignatura como elementos de conocimiento sin relación entre sí. Memorizar datos y desarrollar los procedimientos de forma rutinaria. Encontrar difícil dar sentido a las nuevas ideas que se presenten. Ver poco valor o significado en las asignaturas o en las tareas propuestas. Estudiar sin reflexionar en la finalidad ni en la estrategia. Sentir una presión excesiva y preocuparse por el trabajo.
Estratégico **Organización reflexiva** *Intención:* lograr las calificaciones más altas posibles	Esforzarse sistemáticamente en estudiar. Administrar el tiempo y el esfuerzo eficazmente. Buscar las condiciones y los materiales adecuados para estudiar. Controlar la eficacia de las formas de estudiar. Estar alerta ante los requisitos y criterios de evaluación. Orientar el trabajo hacia las preferencias percibidas de los profesores.

Fuente: ENTWISTLE y cols. (2001).

Entran aquí en juego algunos supuestos de valor acerca de lo que implica un buen aprendizaje y de cómo deben perfeccionarse los buenos estudiantes. Tamsin HAGGIS ha hecho una crítica enérgica de algunos de estos supuestos, en especial los que ella ve como "un conjunto de valores, actitudes y epistemologías de élite que tienen más sentido para los 'defensores' de la educación superior que para muchos estudiantes" (2003, pág. 102), sobre todo en una época en la que la educación superior es cada vez más de masas. Un ejemplo de sus supuestos culturales es "la paradoja china": se considera que la memorización forma parte de los aprendizajes rutinarios y, sin embargo, los estudiantes chinos de elevado rendimiento parecen memorizar de un modo que les lleva a una comprensión más profunda[9]. Observa esta autora:

[9] Véanse: D. WATKINS (2000) y J. LI (2003).

> Inevitablemente, sin embargo, nombrar estos elementos como ítemes independientes parece traducirse en un proceso de cosificación gradual a medida que las ideas entran en una circulación más amplia. "Los enfoques profundos del aprendizaje" se convierten en "aprendizaje profundo" y, en último término, en "procesadores profundos".
>
> (pág. 91.)

La segunda respuesta es el reconocimiento de que la cultura de la evaluación configura nuestros enfoques del aprendizaje: ¿por qué molestarse en ser más que un estudiante estratégico cuando nos van a juzgar por nuestras calificaciones y no por nuestros conocimientos y destrezas? Desarrollaré este tema en los Capítulos V y VI. Aquí, el elemento situacional es que, aunque queramos conseguir que los estudiantes utilicen los enfoques de aprendizaje profundo, ¿qué hacemos para fomentarlo cuando pensamos que nos van a juzgar por la proporción de calificaciones elevadas que obtengan nuestros estudiantes? Aunque "profundo" conlleva toda clase de sesgos de valor —es bueno ser profundo y es malo ser superficial— el mensaje del sistema es: "lo estratégico es bueno".

Aquí, el papel de la evaluación sumativa es crítico. Conseguimos el tipo de aprendizaje que merezcan nuestras evaluaciones, dado que los estudiantes deciden lo que es adecuado. Paul RAMSDEN, un colega de ENTWISTLE, dice que la capacidad de adaptar la situación de aprendizaje puede aprenderse y que:

> ... puede decirse que los estudiantes que son conscientes de sus propias estrategias de aprendizaje y de la diversidad de estrategias que tienen a su disposición, y que saben elegir de forma correcta, responden inteligentemente... o metacognitivamente en ese contexto.
>
> (1983, pág. 178.)

Puede ser más productivo considerar el enfoque estratégico como un enfoque aparte, que aprovecha los otros dos enfoques en relación con la tarea. Lo que ha demostrado el trabajo de RAMSDEN (1987) es que las exigencias de evaluación de las asignaturas eran a veces tan superficiales y basadas en el recuerdo que una respuesta estratégica "inteligente" era optar por enfoques superficiales, pues la comprensión profunda era inadecuada. ENTWISTLE y sus colaboradores reconocen que los estudiantes que tienen éxito utilizan a menudo un enfoque estratégico que aprovecha algunos enfoques profundos, mientras que, con frecuencia, los enfoques profundos por sí mismos "no se sostienen con suficiente determinación y esfuerzo para alcanzar unos niveles profundos de comprensión (2001, página 108). HAGGIS cuestiona de nuevo los supuestos subyacentes:

- los objetivos de los estudiantes son, o puede hacerse que sean, los mismos que los objetivos de los académicos (¿quieren relacionarse personal y significativamente con sus materias?);
- pueden dar sentido a los objetivos de la institución cuando se transmiten mediante la enseñanza y la evaluación;
- los estudiantes ya están en un nivel en el que pueden abordar los materiales de la asignatura tal como esperan los profesores;
- si el medio es correcto, los estudiantes tendrán la voluntad de comprometerse como se espera de ellos (o sea, en el nivel profundo deseado).

(2003, pág. 97.)

Esto refuerza aún más la importancia de la influencia de los factores situacionales y ambientales en los que tiene lugar el aprendizaje y qué tipo de aprendizaje se produce. El trabajo de ENTWISTLE da un paso adelante hacia un enfoque más situacional para evaluar nuestra forma de enfocar el aprendizaje. Aunque las expresiones *profundo*, *superficial* y *estratégico* están marcadas por los valores y transmiten el deseo académico de que los estudiantes, como sus profesores, quieran aprender por el gusto de aprender, existe al menos el reconocimiento de que *lo que* tratemos de aprender y *cómo lo evaluemos* configurará nuestra forma de aprenderlo.

Conclusión

La intención de los tres últimos capítulos ha sido reflexionar sobre el poder de la evaluación para crear constructos y clasificaciones que se tratan después como si existieran de forma independiente. Si volvemos a los "motores para configurar a las personas" de Ian HACKING, el proceso consiste en contar-cuantificar-crear normas-correlacionar-medicalizar-biologizar-genetizar-normalizar-burocratizar. El examen de la inteligencia ha visto cómo se utilizaba cada uno de estos motores con un efecto considerable e invariablemente negativo. A partir de la historia de los tests de CI, hemos visto cómo unas puntuaciones únicas, ordenadas según sus valores y que se correlacionan con el rendimiento futuro, se interpretaban en el sentido de que, en general, los pobres habían nacido con limitaciones genéticas. Esto significaba que necesitarían una provisión educativa especial (o limitada). También condujo a que pidieran que se les restringiera su reproducción. La misma lógica se aplicaba a ciertos grupos étnicos.

La oposición a lo que BINET llamaba "este pesimismo brutal" llegó a través del cuestionamiento directo de estos supuestos hereditarios y mediante la afirmación de la función central del ambiente. Un cuestionamiento indirecto surgió a través de un movimiento de oposición, que redujo la importancia del CI. Las "inteligencias múltiples" y la "inteligencia emocional" han recurrido a ideas más amplias de la inteligencia, tanto en términos de inteligencias no académicas como en los de inteligencias sociales. El problema ha sido que, aunque estas propuestas han promovido unas experiencias educativas más ricas, han corrido el mismo riesgo de "configurar a las personas", aunque se trate de personas más complejas y positivas.

Al revisar algunos enfoques de los estilos de aprendizaje, hemos visto los mismos peligros: clasificaciones benignas, creadas por unos procedimientos de evaluación endebles, que definen la clase de aprendices que somos. A menudo, falta el reconocimiento de la cualidad situacional y social de la conducta inteligente y del aprendizaje. El desinterés de los DUNN por lo que se aprende y por factores sociales más amplios es una forma extrema de lo que decimos. Aunque la teoría del aprendizaje experiencial de KOLB presta atención a estos factores, el instrumento de evaluación ha generado unas categorías poco fiables e inestables que pueden utilizarse para clasificar a los aprendices, clasificaciones que pueden quedar grabadas. Los "enfoques del aprendizaje" más situacionales de ENTWISTLE ofrecen algunas vías constructivas hacia adelante. Él también destaca la importancia de la calidad de la evaluación sumativa en la configuración del tipo de aprendizaje que tenga lugar. Examinaremos a continuación este poder de la evaluación para determinar la calidad del aprendizaje.

CAPÍTULO V

La *titulitis*: ¿Aún contiagiosa después de tanto tiempo?

> En el proceso de obtención de un título... al alumno no le preocupa el dominio de la disciplina, sino que le den el título como si la hubiese dominado. El saber que adquiere no lo adquiere por su propio valor ni para utilizarlo constantemente en una situación de la vida real, sino con el fin de reproducirlo de una vez para siempre en un examen.
>
> (Ronald DORE, 1997.)

Este capítulo y el próximo tratan el tema de cómo la evaluación configura el aprendizaje y la enseñanza, y de cómo "configura a los aprendices" cuando se dedican a trabajar por las calificaciones y los títulos. En la Introducción, hablamos de Ruth, aquella estudiante de éxito, que se las había ingeniado para "seguir el juego" con el fin de obtener las calificaciones que necesitaba para entrar en la carrera universitaria que quería. También hablamos de Hannah, cuyos desvelos para conseguir el nivel requerido le supuso que ella misma se considerase una "nulidad", aunque sus otros logros fueran impresionantes. Aquí nos ocuparemos de dos de los procesos que influyen en las identidades de los estudiantes: la *caza de títulos* (Capítulo V) y el *uso de los exámenes para rendir cuentas* (Capítulo VI). Representan el poder de la evaluación para controlar lo que sucede en la educación y la formación, y veremos cómo configura la evaluación el currículum, la enseñanza y el aprendizaje. El centro de atención de estos capítulos está constituido por las evaluaciones utilizadas con fines de selección individual y de rendición de cuentas en la escuela. En ambos casos, el título o certificado puede convertirse en un fin en sí mismo; lo que se aprenda importa menos. Es esta una visión instrumental de la evaluación que invade gran parte de la enseñanza y el aprendizaje en todo el mundo. La tarea consiste, por tanto, en ver cómo puede mitigarse esa situación y cómo podemos generar unos enfoques más productivos de la evaluación y la rendición de cuentas.

El informe más provocativo acerca de cómo afecta al aprendizaje la necesidad de conseguir cada vez más títulos es *The Diploma Desease**, de Ronald DORE, cuya primera edición es de 1976. Lo que le añade interés a este ya polémico tratamiento del "azote de la certificación" es que la segunda edición salió en 1997. Esto permitió a DORE hacer balance de lo que había predicho y retractarse cuando resultara necesario. Al mismo tiempo, Angela LITTLE publicó una revisión internacional que tituló: "The Diploma Desease Twenty Years On" (1997a), en la que ella y otros expertos examinaron qué había pasado con sus primeras predicciones. Ahora podemos revisarlo de nuevo diez años después[1].

El argumento de DORE de 1976 era que, sobre todo en las naciones en vías de desarrollo, la obtención de títulos con objeto de *conseguir un trabajo* se había convertido en la principal finalidad del aprendizaje y de los exámenes escolares, en vez de aprender por el interés intrínseco por aprender o con el fin de conseguir un trabajo mejor. La intensa competición por los puestos de trabajo "modernos" llevó a una *inflación de titulaciones* por parte de los empresarios, ya que los estudiantes trataban de llevar cierta ventaja en el proceso de selección. Aunque era racional que los individuos se comportaran así, DORE se dio cuenta de que las consecuencias serían profundamente negativas; de ahí la analogía de la enfermedad. Esas consecuencias se deberían a que se había fomentado una visión completamente instrumental del aprendizaje ("aprendizaje superficial") que provocaba un desperdicio de recursos valiosos, pues suponía que los estudiantes permanecieran más tiempo en los centros educativos para conseguir un beneficio educativo de reducido "valor añadido". Estas presiones también distorsionaron la implementación del currículum escolar, dado que solo se enseñaban, aprendían y valoraban los conocimientos y destrezas que entraban en los exámenes.

Parte del atractivo del libro de DORE se debía a que, aunque fuera un universitario que había vivido y trabajado en varios de los países sobre los que escribía, no apostó por una precaución sabia. Hizo predicciones audaces acerca de cómo afectaría la "titulitis" a países que estaban en diferentes etapas de desarrollo: desde economías asentadas, como Inglaterra, hasta otras de desarrollo más reciente, como Japón, Sri Lanka y Kenia. Cuanto más reciente fuese el desarrollo, más grave sería la fiebre. Tampoco se anduvo con rodeos sobre las consecuencias embrutecedoras de la evaluación, sobre todo el aprendizaje rutinario fomentado por la mayoría de las titulaciones.

¿Por qué dedicar un capítulo a examinar esta toma concreta de posición sobre la evaluación? La respuesta es, principalmente, porque DORE estaba interesado, desde un punto de vista comparativo, por algunas de las cuestiones clave de este libro: la finalidad, la adecuación a la finalidad y las consecuencias de la evaluación. No obstante, también tenía ideas firmes sobre la inteligencia y la capacidad, que rozan de un modo inquietante los argumentos desarrollados en el Capítulo II, por lo que conviene debatir esta cuestión. Por último, la inflación de las titulaciones y sus efectos todavía los padecemos.

---

* Hay traducción al castellano: *La fiebre de los diplomas: educación, cualificación y desarrollo*. México: Fondo de Cultura Económica, 1983. (*N. del T.*)

[1] Estoy muy agradecido a Angela LITTLE y a Alison WOLF por sus comentarios sobre un primer borrador de este capítulo.

¿Qué es la titulitis?

En 1976, el argumento de DORE era que, en una época de oferta educativa creciente pero de empleo limitado, la escuela se convierte en un "bien posicional" cuyo valor depende del número de personas que lo tengan. Como el reclutamiento laboral ha llegado a depender en gran medida de los expedientes académicos, el resultado es una *inflación de titulaciones*: el aumento constante del nivel de las titulaciones necesarias para un determinado puesto de trabajo. Así, por utilizar su ejemplo, si hay 50 candidatos para cinco puestos de conductor de autobús, el proceso de selección se simplifica escogiendo a los 10 candidatos que tengan el título de Educación Secundaria, dado que hay un motivo racional para rechazar a los otros 40. En consecuencia, hay más estudiantes que permanecen en la escuela y aspiran a obtener el título de Educación Secundaria, en un momento en el que la posesión de ese título puede dar la ventaja necesaria, una vez que los titulados han dejado de considerar degradante ese trabajo (porque, si lo hiciesen, se convertirían en "desempleados educados").

El siguiente paso del argumento era que cuanto más tarde se desarrollara económicamente un país:

- se utilizarían de forma generalizada los títulos educativos a efectos de selección ocupacional;
- más rápidamente crecería la tasa de inflación de titulaciones, y
- más se orientaría la enseñanza a los exámenes, "en detrimento de la auténtica educación".

(1997a, pág. 72.)

DORE destaca que, en el plano individual, es perfectamente racional buscar esas credenciales; el problema se plantea a nivel político.

Pero, ¿por qué es esto una enfermedad y no un grato desarrollo del capital humano? A diferencia de los políticos, que dan por supuesto que más educación significa más aprendizaje que, a su vez, significa mayor competitividad económica, DORE adopta un enfoque más "credencialista", que considera las titulaciones educativas como un dispositivo de filtro de capacidades. Es posible que el valor añadido educativo de los años de escolarización extra sea pequeño; simplemente consiguen que algunas personas tengan más probabilidades de ser seleccionadas. Se trata de un proceso esencialmente negativo, a causa de sus no buscadas y "deplorables" consecuencias. En vez de aprender por aprender o aprender para hacer un trabajo, se enfatiza el *aprender para* ***conseguir*** *un trabajo*. Este aprendizaje consiste en el "cumplimiento de unos requisitos" sin:

> ningún interés intrínseco por lo que se aprende ni convicción alguna de que sea necesario o siquiera útil para un trabajo posterior; un aprendizaje emprendido exclusivamente con la intención de aprender lo suficiente para aprobar el examen y obtener el título necesario para un trabajo.
>
> (1997b, pág. 27.)

Esto concuerda a la perfección con el enfoque "superficial" del aprendizaje de ENTWISTLE (Capítulo IV).

La segunda consecuencia es que este proceso requiere más plazas escolares y universitarias para los estudiantes que permanecen matriculados, aunque el beneficio educativo sea mínimo. Es un "derroche de recursos" (1997b, pág. 26), porque tiene menos que ver con el desarrollo del capital humano que con proporcionar unos complejos dispositivos de averiguación de antecedentes, utilizando unos recursos valiosos que podrían emplearse mejor para desarrollar una educación primaria para todos.

Lo que nos interesa directamente aquí son sus beligerantes afirmaciones sobre el papel de la evaluación en este proceso y sus consecuencias para el aprendizaje. En el centro de esto está la medida en que los exámenes, una parte esencial de los intentos por conseguir titulaciones académicas, debilitan o promueven el proceso de aprendizaje. Mi argumento es que, aunque las preocupaciones de DORE por el impacto negativo de los exámenes sobre la enseñanza y el aprendizaje son legítimas, si se sobrevaloran, su principal solución, el cambio a los tests de capacidad, es un paso atrás en vez de adelante. Esto conduce a una discusión más amplia acerca de cómo pueda reducirse este impacto negativo, si hay alternativas factibles, y la escala de la titulitis 30 años después.

¿Quién tiene la enfermedad?

En la edición original del libro de DORE, de 1976, utilizó principalmente cuatro países para elaborar su tesis: Inglaterra, Japón, Sri Lanka y Kenia, e incorporó pruebas de Cuba, Tanzania y China (durante la Revolución Cultural del presidente Mao), pues las tres constituían alternativas socialistas radicales. Los cuatro países principales representaban diferentes etapas de desarrollo: desde el desarrollo temprano de Inglaterra hasta el tardío de Kenia; se preveía que los intentos tardíos y rápidos de modernización de países como Kenia provocarían la forma más virulenta de la enfermedad, cuando los estudiantes trataran por todos los medios de acceder al elemento "moderno" de la economía, con sus trabajos mejor pagados. El cumplimiento de estas previsiones dependía en parte de otros tres factores. El primero era que el uso de los títulos dependería del tamaño y el prestigio del empleo en el sector público, que utiliza las titulaciones en la selección, y, en el caso de Japón, en las grandes empresas burocráticas. Un activo sector privado a pequeña escala suavizaría el impacto, pues los procesos de selección no son tan formales. La previsión de que se produciría una tasa más rápida de inflación de titulaciones en las economías en vías de desarrollo podrían atemperarla los gobiernos que resistieran la presión popular a favor de la expansión de la oferta laboral en los sectores secundario y terciario. El debilitamiento de la educación por la enseñanza y el aprendizaje instrumentales impulsados por los exámenes también podría mitigarse merced a la fortaleza de las tradiciones educativas premodernas que enfatizaban ciertos valores culturales como la moralidad y la preocupación social, conduciendo a que la escuela no se preocupara solo de los exámenes.

¿Hasta qué punto se cumplieron las previsiones?

Los investigadores han tenido que esperar 30 años para ver si se cumplían las atrevidas previsiones de DORE. Al revisarlas en 1997, DORE mantuvo que la idea general de su argumento sí contaba con respaldo suficiente, aunque reconocía que las cosas no se habían desarrollado tan claramente como se habían previsto.

El incumplimiento de Inglaterra

La predicción de que los "países de desarrollo más tardío" sufrirían más la titulitis no se materializó por completo. Esto se debió en parte a que Inglaterra, que en 1976 estaba relativamente libre de la enfermedad, la experimentó en niveles epidémicos durante los 20 años siguientes. Las condiciones para su expansión se crearon con el aumento de las titulaciones de formación profesional y, en especial, con la expansión de la educación superior. Alison WOLF (2002)[2] ha presentado una contundente crítica de esta evolución, en relación con la cual considera que, en Inglaterra, el Gobierno trabajó con unos enfoques simplistas del "capital humano" (una mano de obra más cualificada = una mano de obra más productiva). Dice WOLF que quedó demostrado que estos movimientos fueron un fracaso rotundo; el incremento real de títulos llegó a través de una vía mucho menos planeada por la administración central: la *educación superior*.

Esta forma de inflación de títulos surge de la exigencia de los empresarios de un primer grado o de la necesidad de ese grado para ingresar en la formación profesional. El "credencialismo" se aprecia en que a los empresarios no les preocupa demasiado la especificidad del título; lo importante es la clase (el nivel) del título y, hasta cierto punto, la institución en la que se haya obtenido. Esto se ajusta a la tesis de DORE; pero él no lo esperaba en Inglaterra. Para complicar esto, el Gobierno ha fijado el objetivo de que el 50% de los estudiantes progresen hasta la educación superior, en la que también se ha creado el título de *Foundation* de dos cursos. En general, este tiene un enfoque más aplicado, por ejemplo, un título de Educación para ayudantes de docencia no titulados. Por tanto, contra toda lógica, un título académico se convierte en la titulación profesional más adecuada, mientras que las titulaciones de carácter profesional tienen un valor real limitado.

Japón: El qué y el dónde de las credenciales

Paradójicamente, para DORE, Japón era una forma precoz del "país en vías de desarrollo". Aunque, durante siglos, habían existido escuelas para los niños samuráis (alrededor del 6 ó 7% de la población), así como escuelas locales regidas por benefactores, lo que sentó las bases de un cambio espectacular después de la II Guerra Mundial fue una serie de importantes reformas sociales y educati-

[2] Véase, por ejemplo: *Does Education Matter?*, de WOLF (2002).

vas. En 1976, Japón había puesto fin a una racha de crecimiento masivo y tenía la segunda economía en tamaño, después de la de los Estados Unidos. La influencia estadounidense se apreciaba en el paso de un complejo patrón de educación secundaria y superior (las reformas decimonónicas se habían modelado según el sistema francés) a un sistema de "itinerario único" en el que los estudiantes pasaban por nueve cursos de educación obligatoria, tres cursos de bachillerato y, después, colegios universitarios de dos cursos y universidades de cuatro cursos. Esto supuso una rápida expansión de las matrículas en los institutos de bachillerato y en el nivel terciario.

Al mismo tiempo, las estructuras sociales sufrieron un importante cambio. Se mantuvieron muchos de los cambios de la época de la guerra y esto condujo a grandes empresas —un signo de desarrollo tardío— con el reclutamiento burocrático anual de titulados. Los salarios se normalizaron de acuerdo con el nivel de titulación educativa, y surgieron presiones para reclutar a ex alumnos de las mejores universidades, sobre todo cuando el sistema educativo se expandía más deprisa que la economía.

Lo que modificó las previsiones de DORE para Japón fue que cada vez se prestaba menos atención a la obtención de titulaciones de nivel cada vez más alto y más a *dónde* se había formado cada candidato. Ikuo AMANO resume esta situación diciendo que "Japón no es tanto una sociedad de títulos de 'nivel', como una sociedad de títulos de 'institución'" (pág. 56). Esto significa que el ingreso en la escuela y en la universidad correctas es crítico, por lo que la presión evaluadora está presente desde una edad muy precoz, de manera que los niños se preparan y compiten para obtener plazas en escuelas de primaria y secundaria de prestigio, dado que éstas nutren de alumnos a las mejores universidades. En parte, esta presión ha conducido al desarrollo de un extenso sector educativo privado, que ha suavizado ligera y parcialmente esta competición para conseguir unas plazas limitadas.

No obstante, esta particular forma de competición quizá no haya debilitado la fuerza de sus afirmaciones acerca del modo en que los exámenes pueden debilitar la enseñanza y el aprendizaje. AMANO demuestra la fuerza omnipresente de la *puntuación de desviación típica* para motivar a los estudiantes a conseguir mejores puntuaciones con el fin de clasificarse por encima de sus compañeros. Lo que importa es dónde estés "en la curva" (vuelta a GALTON, Capítulo II) y no lo que sepas. Así:

> Los exámenes son el medio primordial que tiene el profesor para motivar a los alumnos para que estudien y obtengan títulos; son importantes para mantener el orden y el control en el aula. Y, en una sociedad moderna, en la que hay cada vez menos consenso acerca de lo que *sea* una buena educación, cuáles *sean* los criterios del desarrollo personal ideal, las calificaciones, el ingreso en escuelas y universidades con el nivel más alto posible en la escala de puntuación de desviación típica, se convierte... en un fin en sí mismo.
>
> (AMANO, 1997, pág. 61.)

¿Y otros países?

El impacto de "acontecimientos traumáticos" ha afectado las trayectorias previstas en varios ejemplos originales de DORE. Su simpatía por el "confucionismo romántico" de la Revolución Cultural de China cayó en el primer obstáculo, y este fue su comentario:

> ¡De qué modo tan grosero irrumpe la historia en los intentos de los sociólogos de llegar a generalizaciones acerca de las tendencias a largo plazo!... Lo que [la formulación de 1976] no tenía en cuenta era el impacto de los acontecimientos históricos traumáticos sobre las tendencias.
>
> (1997c, pág. 189.)

Entre estos acontecimientos estaban la guerra civil en Sri Lanka; el paso al socialismo de mercado en China, y el impacto del desmoronamiento de la Unión Soviética en el sistema de Cuba. Para algunos, esto exacerbó la titulitis, aunque Angela LITTLE demuestra que, en Sri Lanka, el malestar social ha hecho que las titulaciones sean aún más importantes:

> Lejos de considerarlo "un problema", en Sri Lanka, muchas personas veían los exámenes como la única forma legítima y justa de distribuir unos recursos escasos en una sociedad asolada por los conflictos... La dependencia de los exámenes en la década de 1990 no se percibe como un problema o una enfermedad, sino como un remedio paliativo para un conjunto más general de problemas políticos y étnicos cuya aparición no podía haber previsto DORE.
>
> (1997b, págs. 84-85.)

En la actualidad, este sería el caso de países como Ruanda, en el que la formación de un consejo nacional de exámenes, tras el genocidio, ha proporcionado un nivel de equidad y de incentivo en la educación nunca visto antes, un logro reconocido por la ONU[3].

La evaluación y el empobrecimiento del aprendizaje

La principal objeción de DORE a la "caza de títulos" es que debilita la educación, "un proceso de aprendizaje —sea una formación disciplinaria o métodos más libres y divertidos de experimentación— cuyo objeto es dominar algo. El saber puede buscarse por sí mismo, por el puro disfrute de utilizar la mente" (1997a, pág. 8; esto lo coloca firmemente en el "enfoque del aprendizaje profundo", del Capítulo IV). En cambio, la consecuencia de que las titulaciones sean fundamentales para la selección de personal para el trabajo es el empobrecimiento de la enseñanza y del aprendizaje. Profundamente indignado, escribió que "la recopilación de más títulos es una *mera* recopilación de títulos: ritualista, tediosa, teñida de ansiedad y aburrimiento, destructiva de la curiosidad y la imaginación; en pocas palabras: antieducativa" (1976, pág. ix).

[3] Me he basado aquí en las tesis de MA no publicadas de Peter GASINZIGWA y Alphonse KAMALI (2006).

Las "propuestas modestas" de DORE

Dado su análisis, ¿qué intervenciones podrían limitar la titulitis? DORE presenta algunas *propuestas modestas* que están lejos de ser tales: la primera sería reestructurar la educación; la segunda, cambiar lo que se evalúa. Las dos más sobresalientes son:

- Comenzar las carreras profesionales antes, en torno a los 15-17 años, hacer la selección de personal, en la mayor medida posible, en las organizaciones laborales, transformar toda la educación y la formación superior en aprendizaje en plena actividad laboral, llevado a cabo en institutos educativos especiales, bien a tiempo parcial, bien a tiempo completo.
- En todos los puntos en los que tenga que haber una selección —sobre todo con respecto a la importantísima decisión relativa a la organización laboral en la que tenga que ingresar cada persona al final del período de escolarización básica—, evitar el uso de exámenes de rendimiento del aprendizaje; con independencia de que sean tests de aptitudes, sorteos o tests especiales "encapsulados", lo esencial es que sean exámenes que no puedan ser memorizados (o no en gran medida).

(1997a, págs. 142-143.)

DORE cree firmemente en los *tests de aptitudes*, convencido de que son independientes del tipo de preparación escolar que conduce a la memorización, por lo que va en dirección opuesta a los argumentos desarrollados en los tres capítulos anteriores de este libro. Acepta que la sociedad necesita pruebas para seleccionar y ha sido ferozmente crítico de los movimientos "desescolarizadores" que trataban de abolir la escuela[4]. Su preferencia era cambiar a los *tests de aptitud*, tests de inteligencia con algunos elementos extras si fuese necesario (por ejemplo, competencias sociales), que "presumen que miden características no necesariamente innatas pero, al menos, relativamente inalterables mediante la reflexión" (1997a, pág. 155). Su razonamiento es que, si los exámenes se utilizan principalmente para investigar antecedentes, los responsables de selección de personal buscarán a estudiantes de gran capacidad en vez de a especialistas en una materia. Por eso, si podemos ordenar a los estudiantes según su capacidad relativa (dice del sistema japonés que es "un sistema de tests de CI a gran escala y muy caro") y hacerlo relativamente pronto, podemos reducir la influencia de los exámenes sobre la enseñanza e impartir una educación relevante. Como estos exámenes "no pueden ser memorizados o preparados, o no en gran medida" (pág. 154), esto daría margen para un currículum que preparara a la mayoría de los estudiantes para la vida como agricultores y como autónomos. Si esto recuerda a Cyril BURT y un sistema escolar tripartito, probablemente sea así.

[4] Considera el trabajo de Iván ILLICH, entonces en boga, como un movimiento de clase media para restringir las oportunidades de los demás, que recuerda el "ahorradles a los pobres que tengan que hacer exámenes", que comentamos en el Capítulo Primero de este libro.

Reducir lo que se somete a examen

DORE presenta también algunas estrategias alternativas. Una consiste en recortar lo que se somete a examen, por ejemplo, examinar solo de matemáticas y de una lengua: ambas son más difíciles de aprender de memoria que otras materias, como historia o geografía, y se parecen más a los tests de aptitudes. También sirven para predecir el rendimiento académico igual que lo hacen múltiples materias combinadas. Reconoce que esto puede distorsionar el currículum, pero lo rebate con tres argumentos: el primero es que podemos controlar el tiempo asignado a una materia, de manera que es posible impedir que sea desplazada de la enseñanza. Paradójicamente, el segundo es, que en absoluto es mala idea concentrarse en estas competencias básicas. El tercero es que hacen falta gran cantidad de ejercicios disciplinados para que se conviertan en reflejos automáticos, "de manera que la motivación extrínseca de la preparación del examen no sea tan dañina" (1997a, pág. 158), en especial si el resto del currículum puede ser interesante, argumento que aparece en la *Primary Strategy* de Inglaterra[5].

El test de rendimiento encapsulado

Si seguimos con "tests de rendimiento", una opción para la posterior selección ocupacional consiste en administrar tests al final de cursillos intensivos. Éstos limitarían el examen a los cursillos o a proyectos específicos de duración prefijada, lo que implica que se restringiría mucho la influencia del test en la forma de enseñar.

Cuotas y sorteos

Si todo lo demás falla, es posible reducir las presiones de los exámenes asignando plazas escuela a escuela o por sorteo. A DORE no le gusta excesivamente esta sugerencia, pues reconoce que la primera llevaría a una "competición intestina" dentro de la escuela. El sorteo es más justo, pero el problema de DORE es que pueden resultar favorecidos los menos hábiles y que "la sociedad necesita colocar a las personas más brillantes en algunas de las ocupaciones más cruciales" (1997a, pág. 161). Parece que solo tiene en cuenta lel sorteo para la selección ocupacional; no obstante, hay más de un caso de selección para escuelas de secundaria que siguen este procedimiento, como ocurre en Corea del Sur y Malta y, hace poco, en una escuela en Inglaterra[6]. En los Países Bajos, este enfoque se ha utilizado en la selección para la educación superior.

---

[5] Esta es la fuerza de *Excellence and Enjoyment*, del DfEs (2003), que puede considerarse como un esfuerzo para reparar una situación creada en primer lugar por la administración de tests de importancia decisiva y las horas reglamentadas de lectoescritura y aritmética. Véase el Capítulo VI de este libro.

[6] Una resolución judicial de 2007 sobre un centro de secundaria de Brighton (Inglaterra) confirmó el derecho a realizar un sorteo para seleccionar a sus alumnos de 7º.

Angela LITTLE, en una revisión de 1984 de estas propuestas, señaló algunas de sus limitaciones, particularmente en los contextos culturales a los que las dirigía DORE. Su investigación demostró que los padres y los niños de los países en vías de desarrollo "están muy convencidos de que el esfuerzo es el determinante primordial del éxito y el fracaso académicos... es difícil ver cómo se puede persuadir a profesores y a estudiantes para que consideren que sus capacidades y aptitudes son incontrolables, como cosas que no pueden mejorar el esfuerzo y la práctica" (1984, pág. 214). La investigadora demostró que, en estos países, la definición de "aptitud" incluye aún ideas de esfuerzo y de práctica, por ejemplo, en Sri Lanka, la palabra cingalesa que se traduce como "aptitud" significa "un reto para poner a prueba el propio nivel de erudición" (1984, pág. 209). Estas mismas actitudes significan también que la idea de utilizar un sorteo es enemiga de este enfoque y se considera que debilita profundamente el esfuerzo.

Reflexiones de DORE en 1997 sobre sus "modestas propuestas"

Al cabo de 20 años, las ideas de DORE sobre la aptitud se habían endurecido, mientras que las relativas a las repercusiones de los exámenes se habían suavizado. Define la inteligencia como "la potencialidad para la autorrealización productiva y para el logro adquisitivo" (1997a, pág. 178), lo que implica que es tanto una variable de la personalidad como una capacidad mental. Ésta es en gran medida heredada, pero el papel crucial del ambiente consiste en indicar si predomina el ideal de "autorrealización productiva" de DORE o el menos ideal "logro adquisitivo egocéntrico". El aprendizaje orientado a la titulación promueve el segundo, el enfoque estratégico y superficial, en vez del enfoque del aprendizaje profundo del Capítulo IV. En la adolescencia, nuestra disposición empieza a fijarse, de manera que "no es probable que sea muy útil... facilitar una educación universitaria que expansione y estimule la mente a unas personas condicionadas por los doce años anteriores de su escolaridad para aprender solo con el fin de ganar dinero" (1997a, pág. 179). Lo mismo vale para el trabajo: "Si un hombre ha conseguido su trabajo de funcionario merced a dieciocho o veinte años de triste conformidad con los ritos impuestos por una escolaridad orientada a la titulación, ¿quién puede culparlo si se convierte en el funcionario precavido que cumple tristemente con los ritos de la oficina?" (1997a, pág. 12). Y todo esto cuando lo que DORE cree necesario en las economías en vías de desarrollo es la iniciativa y la creatividad empresariales.

Empezamos a hacernos así una idea del mundo de DORE, modelado según la jerarquía de necesidades de MASLOW[7], en la que tenemos diversas capacidades innatas que desarrollar —unos, mediante intercambios; los más capaces, a través

[7] Véase: MASLOW (1973). Se trata de una pirámide de cinco niveles que representa cinco niveles de necesidades, de los que cuatro son necesidades por carencia (fisiológicas, de seguridad, de amor y pertenencia, de estima). Cuando estas han sido satisfechas, podemos ocuparnos de las necesidades de crecimiento (realización personal; autotrascendencia); aquí es donde encontramos la creatividad y la resolución de problemas.

del liderazgo— pero que están siendo limitadas por el "logro egoísta" de las titulaciones, fomentado por la escuela. Por eso prefiere los tests de capacidad a los de rendimiento: descubramos cuáles son las capacidades de los estudiantes y llevémoslos lo antes posible a la senda de la "autorrealización productiva", antes de que la búsqueda de títulos atrofie su crecimiento. Si consideramos la educación desde el punto de vista de la "investigación de antecedentes" y desde el del "capital humano", esto lo sitúa firmemente en el terreno de la investigación de antecedentes: los estudiantes van a la escuela con capacidades que incluso pueden reducirse estando allí. La tarea consiste en hallar su nivel de capacidad y mantener en la escuela a la mayoría solo hasta los 15-17 años, de manera que, al menos, puedan ir y beneficiarse de una educación para aprender a hacer un trabajo.

¿Acaso no son malos todos los exámenes?

Paradójicamente, cuando los puntos de vista de DORE sobre la capacidad se hacen más incisivos, hay indicios de que ha suavizado su opinión sobre los exámenes. Veinte años después, explicó que él no estaba en contra de los exámenes; atacó tanto a los desescolarizadores, que son antiexámenes, como la falta de rigor de los exámenes para la obtención de titulaciones de formación profesional en Inglaterra. También se ha convencido de que no son igual de malos para todos, "como debiera haber comprendido cualquiera que hubiera hecho tanto hincapié como yo en la importancia de las diferencias en las capacidades genéticamente determinadas, que no son en absoluto iguales para todo el mundo" (1997a, pág. xx). Los exámenes causan menos efectos en los más capaces, porque ellos pueden tomárselos con calma; a los menos capaces, pueden proporcionarles una motivación extrínseca que de otro modo no tendrían. Los exámenes siguen siendo malos para quienes "son brillantes sin ser los más brillantes, los que están pendientes de lo que se defina como premios socialmente deseables de la competición pero no estén en absoluto seguros de conseguirlos sin un angustioso y gran esfuerzo" (1997a, pág. xxii). Dadas sus ideas acerca de la capacidad, a nadie le sorprenderá que utilice distribuciones normales y desviaciones típicas para elaborar su razonamiento, de manera que, en Japón, el 0,15% superior, tres desviaciones típicas por encima de la media, es probable que "sobreviva a la feroz competitividad del examen japonés y salga relativamente indemne" (pág. xxi). Esto se debe a que, al ser muy pronto conscientes de su puntuación de desviación típica, habrán sabido que estaban en el 1% superior y, por tanto, si se habían aplicado razonablemente, ingresarían en una universidad de primera fila y podrían adoptar un enfoque más creativo de su aprendizaje. En Inglaterra, la percepción de la "confianza en la escuela pública" puede reflejar un proceso similar: la educación privilegiada conduce a una educación más rica y más segura de sí que la de quienes intentan conseguir las calificaciones más altas en las escuelas ordinarias.

Los exámenes y los menos capaces

Quizá el mayor cambio de los últimos 20 años esté en la actitud de DORE hacia los exámenes de los alumnos de menor rendimiento. Se retracta de su pedagogía centrada en el niño, que enfatiza al descubrimiento placentero como la mejor manera de aprender porque, aunque sirva para los "alumnos brillantes", es probable que su insistencia en impulsar la confianza del aprendiz lento "engañara a los aprendices lentos privándolos de la disciplina que por sí sola puede llevarlos por la dura cuesta hacia la competencia" (1997a, pág. xxii), así como formarlos en las virtudes de la puntualidad, la regularidad y la conformidad con los reglamentos.

Este enfoque se basa en parte en su reconocimiento de que la motivación para aprender es mucho más compleja de lo que se pensaba. Algunas pruebas de ello las facilitó el *Student Learning Orientations Group* (SLOG, 1987), que examinó la motivación de los estudiantes en seis países, entre los que estaban Gran Bretaña, Japón y Sri Lanka. Los resultados no respaldaron la afirmación de DORE referente a que un elevado nivel de orientación hacia la evaluación mata el interés; de hecho, se observó una correlación positiva entre evaluación e interés, aunque la de Japón fue la menos positiva. Se dio cuenta de que es posible que la motivación intrínseca no se produzca naturalmente; a veces, puede tener un principio extrínseco: lo hago porque tengo que hacerlo, pero después empiezo a disfrutarlo o, al menos, a aprender algo[8].

Creo que DORE tiene aquí un problema. Si quiere cambiar a los tests de capacidad, si ésta es innata, la capacidad no variará mucho con el trabajo duro. Si la capacidad se adquiere, ¿dónde se va a hacer el trabajo duro si no en la escuela? ¿No será esto un modo de dar ventaja a los ya privilegiados, que tienen el capital social para beneficiarse de unos exámenes más generales y abstractos? Peor aún, si me clasifican como "de poca capacidad", ¿qué puedo hacer con mi capacidad, sobre todo si me imponen un currículum "funcional", limitado, porque no "llego" a algo más exigente? El autor reconoce el problema y habla de la dificultad de:

> Institucionalizar un sistema que determine muy explícitamente el futuro de las personas según sus cualidades innatas o, al menos, en la época en la que el juicio sobre las mismas sea en gran medida inalterable. Decirle a alguien que no ha realizado bien un test de capacidad es, en cierto modo, más destructivo emocionalmente que decirle que no ha hecho bien un test de rendimiento. El fallo en un test de rendimiento le dice a la persona: tu actuación no ha sido muy buena; ponte a trabajar; prueba de nuevo el próximo año. El fallo en un test de capacidad le dice a la persona: lo siento, chico; no eres de esta clase; a cada uno lo suyo... Un fallo de la voluntad, el fallo de no trabajar lo suficiente... es algo menos intrínseco para el propio sentido del yo que "no tener la capacidad" para hacer algo.
>
> (1997a, págs. 192-193.)

Me parece que esto es una autocrítica demoledora de su postura: el suspenso en el 11+ tenía para muchos el mismo mensaje que el test de CI en el Reino Unido, pues en ambos casos las identidades de los aprendices quedaban marca-

[8] Esto recibe el apoyo de la revisión de la evaluación en el aula de Terry CROOKS (1988).

das para siempre, y recuerda el discurso del ganador de BURT acerca de saber perder (Capítulo III). Volveré sobre este tema en el Capítulo VIII.

Reconoce también que, normalmente, donde se ha implementado la administración de tests de capacidad, un grupo se lo impone a otro menos aventajado. Sus ejemplos son el insólito trío de Sudáfrica, Nueva Guinea y la selección de clase media de niños de clase trabajadora para las *grammar schools* en Inglaterra. Esta última tuvo lugar en una época en la que los niños de clase media podían ingresar pagando tasas, de manera que quienes lo hacían no tenían que competir por las plazas en tests de CI. El procedimiento fue sometido a un escrutinio mucho más minucioso cuando se abolió el pago de tasas y los alumnos de clase media tuvieron que competir con los demás.

¿Y cómo sale DORE de este notable atolladero? Tratando de escapar a unas tierras imaginarias en las que se reconoce que la capacidad es la suerte de la genética y se recompensa con la responsabilidad, pero no en relación con la *renta*. Cree que esto suavizará el golpe para quienes no la tienen, sobre todo si ganan más. Es una respuesta débil y fantástica al problema que él mismo se ha creado.

Así, su argumento sobre los tests acaba en una contradicción y en una queja. Su separación de la capacidad del rendimiento le obliga a tratar de dar sentido a la interacción del currículum, la enseñanza y la evaluación. Son también evidentes sus ingenuidades acerca de los tests de capacidad, en especial su creencia de que no es posible preparar el test, y acerca de por qué estamos motivados para aprender. Como vimos en el Capítulo II, la experiencia educativa y la preparación afectan a los tests de capacidad; DORE tenía que estar familiarizado con las tutorías de clase media que dirigió para la preparación del 11+ en Inglaterra. Cita con aprobación el SAT de EE.UU., sin embargo, como hemos visto, éste ya no se conoce como test de aptitud, sino que se presenta como prueba de competencias educativas generales y uno de los motivos de peso es la existencia de toda una industria dedicada a mejorar "tu puntuación en el SAT". Como sociólogo, DORE parece curiosamente carente de interés por los elementos de clase y de subgrupo en todo esto, un punto ciego que quizá se derive de su postura acerca de la "capacidad fija", por lo que la clase será el resultado de la capacidad y no una causa de ella.

El debate clave es si una vez eliminados los tests de capacidad, se liberarán la enseñanza y el aprendizaje, de manera que todos podamos disfrutar del aprendizaje por sí mismo (aparte de los aprendices lentos, que seguirán caminando lenta y pesadamente por las aburridas lecciones de las cuestiones básicas). Creo que esto subestima el impacto de ser clasificado: ¿quienes suspendieron el 11+ y continuaron en las *secondary-modern schools* en Inglaterra tienen una educación rica y creativa (no examinada)? Angela LITTLE hace la misma observación sobre la enseñanza y la motivación en las naciones en vías de desarrollo. DORE tiene una visión idealizada del aprendizaje: eliminemos los tests y el aprendizaje profundo se producirá de forma natural, excepto en el caso de los menos capaces. Esto no tiene suficientemente en cuenta la complejidad de la motivación en el aprendizaje. Volveré sobre estos temas mediante algunas "modestas propuestas" diferentes.

Unas modestas *propuestas diferentes*

Después de las audaces florituras de Ronald DORE, éstas son unas propuestas verdaderamente modestas. Esto se debe en gran medida a que mi enfoque es pragmático: trato de trabajar con lo que hay, en vez de inventarme un mundo ideal. Mis supuestos de trabajo son que los tests de rendimiento:

- son más válidos para la mayoría de los fines educativos que los tests de capacidad / aptitud / inteligencia;
- configuran lo que se enseña, cómo se enseña y cómo se aprende, por lo que pueden desempeñar un papel positivo en el aprendizaje si están claras la finalidad, la adecuación a la finalidad y las posibles consecuencias;
- desempeñan un papel social clave en la certificación, selección y progresión;
- deben beneficiar primordialmente a los individuos que los realizan (en vez de a otros que no los hacen).

Las consecuencias de estos supuestos nos llevan en una dirección diferente de la de DORE: hacia un enfoque más de "capital humano", que trata de añadir valor a través de la enseñanza y las titulaciones. Esto no significa que todo esté bien con respecto a las pruebas y titulaciones actuales: muchas incumplen los criterios de finalidad, adecuación a la finalidad y consecuencias (en otras palabras, carecen de validez). Muchas conducen también a un currículum limitado, una enseñanza sin imaginación y un enfoque superficial del aprendizaje. Pero, si hay un motivo válido para administrar una prueba, el remedio es mejorarla en vez de abandonarla. Justificar los exámenes puede llevar también a reducir su número, dado que, en muchos casos, será difícil justificarlos en cuanto a su finalidad y su adecuación a la finalidad.

En defensa de las pruebas de rendimiento

Un argumento central de este libro, con el que espero que, a estas alturas, estemos familiarizados, es que los tests de capacidad y aptitud son, aun por definición, engañosos. Se infiere que la capacidad previa de aprender de la persona es la *causa* del aprendizaje, que es independiente de la escolaridad y, a menudo, se considera innata y fija. Creo que esta es la idea de DORE. Mi argumento de la "vuelta a BINET" es que es preferible considerar esos tests como *tests de rendimiento generalizado*; las puntuaciones de una persona en un test nos hablan del uso que ha hecho de su experiencia, el *producto* del aprendizaje. Por eso, mi objeción se refiere primordialmente a la interpretación de los resultados, más que al contenido de las pruebas de capacidad[9]. Los tests de capacidad pueden ser

[9] Esto nos lleva a las ideas actuales acerca de la validez, que no solo implica lo que se está midiendo (el constructo) y el modo de hacer el muestreo, sino también las inferencias extraídas de los resultados y sus consecuencias. El argumento es que las inferencias están equivocadas (la capacidad como causa, en vez de como resultado), por lo que los tests de capacidad acaban siendo inválidos, aun en el caso de que el contenido pueda defenderse.

buenos predictores de la actuación futura, porque el rendimiento generalizado que se mide es un buen predictor del rendimiento fututo esperable. Cuando los empresarios se interesan por la clase de un título más que por su materia, siguen interesándose por el rendimiento general, que no es independiente del contenido. El problema es que nuestro bagaje histórico implica que sea fácil hacernos retroceder para considerar la capacidad y la inteligencia como la causa subyacente del rendimiento y no como una forma del mismo.

Mi enfoque contempla el rendimiento como un continuo, que va desde el razonamiento académico más abstracto (por ejemplo, las matrices de RAVEN, las analogías, componentes de los tests de capacidad) y las competencias de ejecuciones complejas (por ejemplo, la diagnosis médica) hasta el recuerdo más concreto de información discreta (por ejemplo, datos históricos) y tareas ocupacionales específicas (por ejemplo, conectar un enchufe). La cuestión es en qué parte de este continuo se sitúa una prueba determinada o el grueso de sus cuestiones. Utilizando este enfoque, la objeción de DORE a las pruebas de rendimiento es que plantean una clase de exigencias concretas que requiere poco más que un aprendizaje memorístico rutinario. Sin embargo, la solución no es abandonarlas, sino desplazarlas por el continuo para que puedan plantearse exigencias más complejas (véase más adelante).

Esto evita la clasificación que conllevan los tests de capacidad y sus consecuencias para el aprendizaje. Estas han sido sistemáticamente investigadas por Carol DWECK (1999), que hace la distinción entre aprendices que mantienen teorías *entitativas* de la capacidad (fija y dada) y quienes sostienen teorías *incrementales* basadas en el esfuerzo, y demuestra la diferencia que imponen estas creencias en el modo de enfocar el aprendizaje y de abordar los aprendizajes difíciles y el fracaso[10].

Las pruebas: Reglas de participación

Ésta es más una defensa a regañadientes de las pruebas de rendimiento que la imagen inversa del entusiasmo de DORE por las pruebas de aptitud. Comparto su análisis de que muchas pruebas de rendimiento dañan el aprendizaje y fomentan una enseñanza restringida y nada imaginativa. Mi apoyo es, en consecuencia, condicional, y estas condiciones nos devuelven a las tres cuestiones básicas de la validez:

- ¿Cuál es la finalidad principal de esta evaluación?
- ¿Esta evaluación es adecuada a la finalidad?
- ¿Cuáles son las consecuencias, pretendidas y no pretendidas, de esta evaluación?

Es preciso responder a estas preguntas antes de que pueda justificarse la realización de un test, y algunas respuestas pueden llevar a apoyar que se deje de utilizar.

[10] Véase el Capítulo VII y el muy legible *Self-Theories*, de DWECK (2000).

Finalidad

Un primer paso esencial, aunque a menudo ignorado, es dejar muy clara la finalidad de la evaluación. ¿Se refiere esto a establecer qué hay que aprender, a lo que se ha aprendido, al control de la clase ("si no atiendes, perderás puntos en el examen de esta semana") o al control administrativo, por ejemplo, el examen a mediados del trimestre impuesto por el centro? Una buena pregunta de seguimiento puede ser: *¿Hasta qué punto es necesaria esta evaluación?* Inevitablemente, habrá múltiples finalidades, por lo que entrará en juego el "principio de la prepotencia administrativa": cuando hay múltiples finalidades, las administrativas dominan sobre las demás. El valor educativo de la evaluación se resume en esto: ¿qué aporta a la enseñanza y al aprendizaje?

Adecuación a la finalidad

Supongamos que nuestro test propuesto tiene sentido. El siguiente conjunto de condiciones se refiere a si, en realidad, el test hace lo que pretende, lo que he llamado *test Ronseal**, por el anuncio televisivo de pinturas en el que un "manitas" declara convincentemente que "hace lo que pone en la lata" (seca en 30 minutos, etc.).

Hay varias cosas que dificultan que un test cumpla su finalidad. El test puede:

- *medir en realidad algo más*, por ejemplo, un test de matemáticas que utilice un lenguaje tan complejo que mida competencias lectoras más que competencias matemáticas, o una pregunta creativa de ensayo que sea tan previsible que mida el recuerdo de unas respuestas preparadas;
- *muestrear de forma insuficiente el contenido de la asignatura*, por ejemplo, un test de lengua que ignore las competencias de habla y escucha;
- *estar en un formato que atente contra su finalidad*, por ejemplo, un test de opciones múltiples de escritura creativa (sí, ha habido algunos) desalentará la auténtica escritura creativa.

En consecuencia, la tarea consiste en producir lo que John FREDERIKSEN y Allan COLLINS han denominado *test sistémicamente válido*:

> Un test que induce en el sistema educativo unos cambios curriculares e instructivos que fomentan el desarrollo de las competencias cognitivas que, por diseño, tiene que medir.
>
> (1989, pág. 27.)

Volveremos sobre este tema.

* *Ronseal* es una marca de barnices, pinturas y ceras, cuyo eslogan es el que aparece en el texto. (*N. del T.*)

Consecuencias

Una evaluación nunca es un hecho neutro; siempre acarrea consecuencias. Algunas de ellas estarán relacionadas con la finalidad de la evaluación; si es la selección, los resultados serán decisivos para la persona que se someta a ella. Si la finalidad es la rendición de cuentas, las consecuencias pueden ser para la escuela más que para el individuo. No obstante, en este punto, el interés está en las *repercusiones*: ¿cómo influye una evaluación en los procesos de enseñanza y aprendizaje? Para DORE, esta era la consecuencia "deplorable" de la realización de tests, que conducía a una enseñanza y un aprendizaje empobrecidos.

No pretendo minimizar las repercusiones negativas del exceso de tests. Si se percibe que el test es importante, inevitablemente se producirá una *enseñanza para el test*. La cuestión es hacer que el test sea lo bastante bueno para estimular una enseñanza y un aprendizaje efectivos. En términos de la clasificación de ENTWISTLE, buscamos, al menos, un enfoque estratégico, a ser posible con un toque de profundidad. El peligro, aun en este, es *enseñar para el test* centrándose en las exigencias específicas y previsibles del test: cómo descubrir las pistas y qué hacer entonces, un enfoque completamente superficial.

Los principios del examen

La preocupación por la finalidad, la adecuación a la finalidad y las consecuencias pueden destilarse en cinco principios principales de los exámenes. Estos principios son exigentes, tienen un doble sentido y, en la práctica, nunca se satisfarán por completo. No obstante, este es el objetivo hacia el que hay que orientar el trabajo para que el examen contribuya al proceso de aprendizaje, en vez de debilitarlo.

1. *Si los profesores van a enseñar para el examen* (y lo harán)*, éste debe reflejar las competencias y saberes que requiere el currículum.*
2. *La forma del examen influirá en la enseñanza y en el aprendizaje*: un test de opciones múltiples de "conocimientos en porciones" llevará a una "enseñanza en porciones".
3. *La previsibilidad de un examen influirá en que la enseñanza haga más hincapié en un enfoque de aprendizaje profundo o en otro superficial.* En la tradición de los exámenes antiguos, la enseñanza y el aprendizaje tienen que ver con más frecuencia con preparar y recordar cuestiones previstas que con una comprensión bien fundamentada de la materia. Para fomentar ésta, hacen falta unos elementos menos previsibles.
4. *Los exámenes deben ayudar a motivar a los examinandos, merced a la accesibilidad y a la justicia.* Esto encierra muchas cosas y algunas tensiones difíciles. Los exámenes suelen motivar a quienes los realizan bien pero, con frecuencia, se plantean pensando en quienes no los hacen bien. Los exámenes desmotivarán a no ser que se consideren cuidadosamente las necesidades de los alumnos de menos rendimiento. Por eso, tienen que ser accesibles a quienes los hacen, sin que requieran simplemente el

recuerdo de bajo nivel. La justicia tiene que ver también con el modo de abordar las diferencias culturales, sociales y de género.

5. *El modo de interpretar y utilizar los resultados tiene una importancia crítica.* Las teorías recientes de la validez de las pruebas se han centrado en *las inferencias que se hacen a partir de los resultados*, más que en las propiedades de las pruebas mismas. Por eso, una prueba bien construida puede seguir siendo inválida si los resultados se interpretan o se utilizan mal. Si utilizo un test con una finalidad no prevista o interpreto erróneamente las puntuaciones, queda comprometida la validez de la prueba.

Factores inhibidores

No quiero subestimar las dificultades prácticas de este enfoque. En cualquier sistema de evaluación, hay poderosas presiones que van en contra de crear pruebas con este nivel de adecuación a la finalidad. Una de ellas es la *presión para simplificar*, que trata de conseguir una evaluación más sencilla y más eficaz en relación con el coste. Las tareas iniciales de evaluación del currículum nacional de Inglaterra incluían algunas actividades prácticas complejas como estímulo para la evaluación, por ejemplo, los niños de 7 años experimentaban para averiguar qué objetos flotaban en el agua y por qué[11]. El a la sazón Secretario de Estado para la Educación tachó de "elaborado sinsentido" este tipo de tareas, instaurando tests de papel y lápiz. Esta presión surgirá siempre si la implementación de una forma válida de evaluación es complicada y cara.

Relacionada con ella está la *presión para la normalización* que, en aras de una particular interpretación de la fiabilidad (misma tarea, mismas condiciones de prueba, corrección externa o esquema de corrección), limita el alcance de lo que pueda iniciar la escuela o un alumno concreto. La presión es particularmente fuerte cuando los resultados se utilizan con fines de rendición de cuentas (véase el Capítulo VI). El argumento reza así: si van a compararse los resultados de los centros, las escuelas deben hacer las mismas pruebas y tareas. Es la misma obsesión del "café para todos" que a menudo ha perseguido los movimientos a favor de una evaluación más sistémicamente válida. Así, la "investigación histórica" se reduce para responder a un documento estímulo común en un examen.

Íntimamente ligada a la simplificación y a la normalización está la *presión a favor de la previsibilidad*. Si los tests y los exámenes son instituciones sociales, siempre habrá una presión pública para mantenerlos sin cambios en el tiempo, sobre todo si se utilizan con fines de rendición de cuentas. Esto se refleja en las tradiciones de los exámenes antiguos, en las que se esperan variaciones sobre un mismo tema ("siempre cae una sobre...") y es difícil que se impartan algunos temas porque tampoco es fácil que entren en los exámenes. En parte, una buena enseñanza para el examen tiene que ser un buen mosaico de lo que probablemente se pregunte. En el peor de los casos, esto significa, como atestiguaba

[11] Después, los alumnos tenían que explicar por qué flotan, o no, las cosas. Más elaborada aun era una tarea "integrada", que implicaba cultivar berros, medir su crecimiento en diferentes condiciones (Ciencias y Matemáticas) y, por último, saborearlos y escribir acerca de la experiencia (Lenguaje).

DORE, un enfoque del aprendizaje para los exámenes. Garrison KEILLOR presenta un irónico ejemplo en su *Lake Wobegon Days*:

> Durante años, a los estudiantes del curso superior se les exigió que leyeran ["Phileopolis"] y respondieran a preguntas sobre su significado, etc. A los profesores no se les pedía que lo hicieran, limitándose a corregir de acuerdo con las respuestas correctas facilitadas por la señorita Quoist: **1**) Extender los beneficios de la civilización y la religión a todos los pueblos, **2**) No, **3**) Platón y **4**) Un desierto no puede saciar el hambre de belleza y de aprendizaje, una vez abiertos los ojos. El examen era el mismo año tras año y, cuando los alumnos del curso superior descubrieron las respuestas y se las pasaron a los del curso siguiente, nadie volvió a leer "Phileopolis".

Crear mejores exámenes

Por tanto, la tarea consiste en elaborar evaluaciones sumativas cuyas repercusiones tengan una influencia positiva sobre la enseñanza y el aprendizaje. En algunos casos, esto puede ser tan sencillo como garantizar que se abarque el currículum acordado, una cuestión de equidad para algunos estudiantes que puedan haber sido injustamente tratados antes.

La explicación de DORE acerca de cómo podría transformarse la enseñanza en los países en vías de desarrollo si se eliminara el yugo de los exámenes, estaba algo idealizada a este respecto. Su punto de partida era que, si pudiésemos eliminar los exámenes, los profesores podrían poner manos a la obra y enseñar creativamente. Según mi propia experiencia, muchos profesores carecen del conocimiento en profundidad de la materia que se necesita para esto y los exámenes facilitan una estructura básica. La importancia de los buenos exámenes estriba en que, aun si el aprendizaje es estratégico y la enseñanza es "para el examen", todavía pueden dar como resultado una experiencia constructiva de aprendizaje.

Incorporando los cinco principios y permitiendo la influencia de los factores de presión, presento a continuación cuatro pasos prácticos para conseguir unos exámenes mejores.

Poner de manifiesto la finalidad y la exigencia del aprendizaje

Las pruebas de rendimiento son evaluaciones de algo, por regla general, un currículum, materia o especificación de competencias. El "¿por qué un examen?" inicial se refiere tanto a la finalidad como a la oportunidad. "Porque es el final del módulo/asignatura/etapa" es solo una respuesta parcial; tenemos que mirar también tanto a sus objetivos como al uso que se haga de los resultados. *Los objetivos de la asignatura, más que su contenido, deben determinar la finalidad y la forma de su evaluación*. John WHITE (2004) ha demostrado que quienes elaboraron los programas de estudio específicos de asignaturas del currículum nacional en Inglaterra prestaron poca atención a los objetivos y valores declarados del currículum. Por eso, mientras que los objetivos se refieren a fomentar la curiosidad y el trabajo colaborativo, los programas de estudio se refieren de forma abru-

madora a los contenidos. Esto resta coherencia, que todavía se debilita más por una evaluación aún más restringida, en la que, por ejemplo, los elementos aplicados de matemáticas y ciencias y los elementos de habla y de escucha en lengua no se someten a examen.

La *exigencia del aprendizaje* refleja esta preocupación por los objetivos más generales de la evaluación. ¿Qué nivel de conocimientos o competencias concuerda con estas intenciones? Algunos sistemas de evaluación utilizan la *Taxonomy of Educational Objectives*, de BLOOM (1956)*, como una jerarquía de exigencia cognitiva, con su movimiento desde el saber, a través de la comprensión, la aplicación, el análisis y la síntesis, hasta la evaluación. Esto se ha discutido, pero, por regla general, sirve como marco de referencia útil y el análisis de los exámenes utilizando ese marco identifica a menudo cuántas cuestiones corresponden al nivel más bajo, recuperando los conocimientos, en vez de demostrar la comprensión de los mismos[12]. Este primer paso práctico se refiere al modo en el que la evaluación responde a las intenciones u objetivos de la asignatura o el currículum y no a la cobertura de contenidos que predomina con frecuencia en la elaboración de los exámenes.

Hay una tendencia a definir la exigencia por el *formato* de la evaluación, de manera que, por ejemplo, se considera que las preguntas de desarrollo o de ensayo son más difíciles que las de opciones múltiples. La investigación comparativa sobre los enfoques de aprendizaje de los estudiantes australianos, japoneses y tailandeses y las exigencias de los exámenes a los que se someten, llevada a cabo por Neil BAUMGART y Christine HALSE, cuestiona estos supuestos. Estos investigadores encontraron algunas diferencias sorprendentes de actitud, de manera que los estudiantes japoneses adoptaban un enfoque del aprendizaje más activo de lo que se prevería teniendo en cuenta su estereotipo "pasivo". Cuando analizaron los distintos exámenes que habían hecho los estudiantes concluyeron que los tests de opciones múltiples realizados por los estudiantes japoneses eran a menudo más exigentes que las preguntas más abiertas a las que estaban acostumbrados los estudiantes australianos. La razón era que las preguntas japonesas implicaban un trabajo cognitivo más activo; aunque fuesen de opciones múltiples, requerían que el estudiante tuviese conocimientos previos y los utilizase para razonar con el fin de seleccionar la respuesta. "La preparación asidua para las tareas de evaluación de este tipo probablemente predispusiera a estos estudiantes asiáticos para una transferencia satisfactoria del aprendizaje y, por tanto, un elevado rendimiento en contextos de evaluación relacionados como los incluidos en los estudios comparativos internacionales de rendimiento" (1999, pág. 6). En cambio, los estudiantes australianos tenían la mayor parte de la información facilitada para las preguntas de respuesta corta, lo que les permitía "adivinar", y las preguntas abiertas eran tan previsibles que se habían convertido en gran medida en un ejercicio de recuerdo de los temas preparados.

---

* Hay traducción al castellano: *Taxonomía de los objetivos de la educación: la clasificación de las metas educacionales : manuales I y II* (traducción de Marcelo PÉREZ RIVAS). Buenos Aires: Centro Regional de Ayuda Técnica: Agencia para el Desarrollo Internacional (A.I.D), 1971. (*N. del T.*)

[12] Otros marcos de referencia para la administración de tests pueden encontrarse en la web de Cresst: http://www.cre.ucla.edu, y en informes del proyecto Bear en California: http://www.bearcenter.berkeley.edu. La taxonomía SOLO, de John BIGGS (1999) es muy útil para establecer la relación de los objetivos de aprendizaje con su evaluación.

Fomentar el conocimiento basado en principios mediante preguntas menos previsibles*

La idea de que un examen debe incluir preguntas desacostumbradas, de manera que los estudiantes tengan que utilizar sus conocimientos para poder elaborar una respuesta, puede parecer inofensiva pero, en la práctica, los factores de resistencia (manejabilidad, estandarización, previsibilidad) se activarán rápidamente. En una tradición que prima el uso de exámenes de otros años, la dependencia de la previsibilidad está muy arraigada. Gran parte de la preparación se centra en el "cuando veáis esto...", que enfatiza la búsqueda de pistas y el recuerdo de respuestas preparadas de antemano.

Pretendo pasar de la preparación *"cuando veas..."* a la preparación *"¿qué pasa si...?"*. En términos de enseñanza y aprendizaje, esta forma de preparación promueve un enfoque más orientado a los problemas, lo que, a su vez, lleva a un aprendizaje más activo. Esto se relaciona con otro trabajo comparativo, en este caso sobre la enseñanza japonesa y la estadounidense de las Matemáticas. James STIGLER y James HIEBERT examinaron lo que ocurría en las aulas japonesas con el fin de descubrir las razones del elevado rendimiento de sus alumnos en las comparaciones internacionales. Uno de los hallazgos en el terreno didáctico era que el maestro japonés planteaba con frecuencia un problema a los alumnos; éstos se reunían en grupos para resolverlo, teniendo que encontrar a veces dos métodos de resolución diferentes. Se interpretaba que la frustración y la confusión constituían una parte natural del proceso. Los maestros esperaban a que se hubiesen enfrentado al problema antes de presentar los procedimientos matemáticos que contribuyeran a la resolución; cuando sabes que tienes un problema, aprendes mejor. A un grupo que lo resolvió, se le permitió que propusiera problemas a otros grupos.

En cambio, la enseñanza estadounidense de las matemáticas se basaba en presentar por separado los distintos elementos de las técnicas que fuesen necesarios para resolver los problemas. Después, los estudiantes practicaban cada uno de ellos, planteándose a continuación el problema. Los alumnos lo resolvían, trabajando individualmente con estas técnicas, y era muy probable que el maestro rescatara de inmediato al que se quedara atascado, explicándole qué técnica aplicar. (Otra diferencia didáctica era que los maestros japoneses dedicaban más tiempo que sus colegas estadounidendeses a preparar las clases, a menudo de forma colaborativa, y menos a corregir ejercicios.)

Aunque quizá esta sea una visión idealizada de lo que ocurre en una clase en la que se practique un "aprendizaje activo", explica el deseo de disponer de pruebas que estimulen a los estudiantes a aprovechar con flexibilidad lo que saben y a abordar satisfactoriamente materiales poco habituales. El punto de partida para esto es la evaluación del maestro en el aula. El maestro sabe lo que se ha enseñado y, por tanto, puede idear unas preguntas que permitan apreciar hasta qué punto son capaces los alumnos de utilizarlo de un modo o en un contexto poco habitual. Esto no solo pone a prueba unos conocimientos "basados en principios"

* La expresión original es *principed knowledge*, tomada de Magdalene LAMPERT (1986: "Knowing, doing, and teaching multiplication", *Cognition and Instruction*, *3*, págs. 305-342), que podría definirse como una clase de conocimiento que es, a la vez, conceptual y procedimental. (*N. del T.*)

(que pueden transferirse a situaciones nuevas), sino que también facilita la retroinformación sobre las concepciones erróneas que quizá no revelaran unas respuestas de "recuerdo", más previsibles. A este respecto, un principio importante es que *las pruebas administradas en el aula no tienen que imitar continuamente las pruebas externas*, que probablemente serán siempre más restringidas. En una asignatura, el maestro o profesor puede llegar al conocimiento basado en principios de muchas maneras; la idea de hacer todo como en el examen final tendría las embrutecedoras consecuencias que detestaba DORE. Aunque los estudiantes tengan que practicar las destrezas y los formatos específicos de los exámenes, no necesitan practicarlos en todas y cada una de las pruebas de la asignatura. En el Capítulo VI, veremos que el carácter decisivo de los tests de rendición de cuentas puede forzar de este modo la enseñanza.

Al haber trabajado para tribunales de exámenes y organismos gubernamentales de tests, puedo imaginarme a mis ex colegas elevando la mirada al cielo y refunfuñando acerca de la "realidad", las "presiones" y el "está perdido". Soy muy consciente de que, en un examen, las sorpresas no son bienvenidas, por lo que cualquier cambio ha de ir precedido por una cuidadosa preparación del terreno. Tampoco digo que, en los tests y exámenes actuales, no haya ya preguntas imaginativas y estimulantes; muchos de ellos utilizan diferentes escenarios y materiales de estímulo como base de las preguntas. Lo que me gustaría ver desarrollarse es una cultura de evaluación en la que no sean suficientes la búsqueda de pistas y el recuerdo, de manera que sus consecuencias para la enseñanza y el aprendizaje consistan en estimular la resolución flexible de problemas fundada sobre el conocimiento basado en principios[13].

Hacer que las pruebas sean lo más auténticas posible

Si queremos que las consecuencias de las pruebas conduzcan a unas prácticas de enseñanza y aprendizaje que ayuden a desarrollar las competencias deseadas, cuanto más directamente evalúe un test estas competencias, más probable será que las promueva. Como vimos en nuestra definición inicial de "prueba", esta es una representación de una destreza, por lo que, cuanto más la refleje con precisión, probablemente será más válido. Por desgracia, los dos factores de la manejabilidad y la estandarización pueden causar estragos en relación con estas intenciones, reduciendo la evaluación de las destrezas a una pálida imitación de papel y lápiz.

El riesgo es que estas puntuaciones carentes de autenticidad acaben siendo más importantes que el hecho de ser capaz de poner en práctica la competencia, a causa de lo que Allan HANSON llama la *cualidad de invención de los tests*. Dice que:

> El proceso de invención opera de acuerdo con lo que podemos llamar la prioridad del potencial sobre el de la ejecución. Como los tests actúan como puerta de acceso

[13] A esta misma combinación —resolución flexible e inmediata de problemas y abstracción de nivel superior— le atribuye ahora James FLYNN (2006) el espectacular incremento de puntuaciones de CI durante los 100 últimos años (véase el Capítulo II de este libro).

> a muchos programas educativos y de formación... la probabilidad de que una persona sea capaz de hacer algo, tal como lo determinan los tests, cobra más importancia que la realización concreta de ese algo. Solo se permite que las personas accedan a estos programas, ocupaciones y actividades si demuestran primero un potencial suficiente, tal como lo miden los tests.
>
> (pág. 288.)

Aunque esto concuerda con el argumento de DORE, HANSON no ve virtud alguna en las pruebas de aptitud, pues también éstas se convierten en fines en sí mismas, en vez de ser medios para un fin. Se hace a menudo más hincapié en la titulación para hacer algo que en hacerlo realmente, por ejemplo, las calificaciones para obtener una beca de investigación, en vez de la realización concreta de la investigación.

Si operan esos procesos de invención, tenemos que hacer que las pruebas se acerquen al máximo a la "ejecución" real. Este es el fundamento racional de la evaluación auténtica, expresión utilizada para referirse a la evaluación directa de la ejecución o de las competencias. Las destrezas de ejecución, como las de la música o de la representación dramática pueden ser razonablemente sencillas de justificar: una persona que tenga un título de música tiene que haber tocado algo ante alguien. Ser capaz de hablar en la lengua extranjera que estemos estudiando parece obvio, pero no se evalúa en algunas titulaciones. Aunque la evaluación auténtica sea una práctica normal en muchas evaluaciones ocupacionales (¿quién quiere un dentista con una formación puramente teórica?), resulta mucho más difícil abrirle camino en la evaluación académica. Los tres factores inhibidores de la manejabilidad, la estandarización y la previsibilidad comienzan a surtir efecto cuando el currículum requiere competencias aplicadas, como la "investigación histórica", la "investigación geográfica" o la "presentación de una propuesta de negocio". La orientación hacia una evaluación más auténtica estimularía al estudiante a mostrar su aplicación, a ir y buscar la evidencia histórica. Con frecuencia, la preocupación por la fiabilidad ha impulsado la retracción hacia unos formatos más parecidos a los de los exámenes institucionales, en los que todo el mundo aborda la misma tarea, en vez de otras más adecuadas a la realidad del momento y el lugar.

Estas presiones son comprensibles, pero no inevitables. En Queensland (Australia), hace 30 años que la evaluación del maestro o profesor reemplazó a los exámenes institucionales como fundamento de la titulación de los alumnos de último curso y como base para la selección para la universidad. En el plano de la manejabilidad, el argumento es que la evaluación del maestro o profesor cuesta menos y genera una formación profesional continua mucho mejor para ellos. Esto se debe en parte a un programa de estandarización y moderación en el que los docentes participan con regularidad. Este enfoque permite también más libertad para escoger los temas más relevantes en el ámbito local. A causa de su carácter decisivo, para nota, la supervisión del rendimiento en el nivel de la escuela también forma parte de este sistema[14]. Esto es posible porque Queensland tiene una población relativamente pequeña y el ingreso en la universidad carece del elemento extremadamente competitivo que se aprecia en otros países.

[14] CUMMING y MAXWELL (2004).

Esto contrasta con los movimientos actuales en Inglaterra para reducir la proporción de evaluación a cargo de los docentes en los exámenes nacionales. El fundamento de los mismos es que se ha producido una pérdida de confianza en la fiabilidad de los trabajos realizados en la escuela, causada en parte por la posibilidad de descargar trabajos de Internet. Paradójicamente, el influjo de la estandarización es en gran medida el culpable de esto. El hecho de que todos los alumnos tengan que hacer las mismas tareas y que éstas se mantengan iguales año tras año hace que, comercialmente, merezca la pena colgar "respuestas modelo" en la red (se cree que hay más de 5.000 descargas de este tipo correspondientes a tareas de matemáticas para el GCSE, por lo que los trabajos de matemáticas del GCSE ocuparon el primer lugar). Así, el carácter previsible de la tarea ha reducido el trabajo a un ejercicio sobre "exámenes anteriores": "conocemos la tarea; sabemos cómo obtener los puntos; aquí tienes cómo hacerlo"[15].

De nuevo, esto no tiene por qué ser así. El Bachillerato Internacional (BI) es una titulación académica muy exigente y de elevada categoría que se utiliza como credencial de ingreso en universidades de todo el mundo. Uno de sus elementos obligatorios es una "monografía", que supone la investigación de los propios estudiantes de un tema relacionado con alguna asignatura de las que estudien. Los profesores corrigen el trabajo, que es revisado por moderadores externos. Constituye un buen ejemplo de evaluación auténtica porque promueve directamente las competencias que trata de medir. De este modo, se estimula al historiador que hace una investigación original "real", que empleará esa competencia en cualquier otro estudio de historia.

*Confiabilidad**

Necesitamos algunas herramientas analíticas para determinar cómo hacer que la evaluación sea lo más auténtica posible en una materia o titulación determinada. Un concepto útil puede ser la *confiabilidad*, pues representa la búsqueda de la interacción óptima entre dos elementos de la validez: la validez de constructo y la fiabilidad. Así, una tarea podría tener una elevada validez de constructo, ser una buena muestra de la competencia en cuestión, pero una baja fiabilidad, por no disponer de un plan acordado de evaluación. Del mismo modo, un test de opciones múltiples podría ser muy fiable en cuanto a las puntuaciones obtenidas, pero tener poca o ninguna validez de constructo para evaluar la escritura reflexiva. En ambos casos se daría una baja confiabilidad.

---

[15] En el GCSE de Ciencias, las tareas para hacer en clase son tan rutinarias (por ej.: "investigad la resistencia eléctrica de un alambre...") que una comisión parlamentaria dijo de ellas que eran "una actividad tediosa y aburrida tanto para alumnos como para profesores" (*Select Committee on Science and Technology, Third Report*, 2002).

* En el original: *dependability*. Esta palabra es, en realidad, sinónima de *reliability* ("fiabilidad") que, significa, según el diccionario Merriam-Webster, es la "cualidad o estado de ser fiable" y "*específicamente*: la medida en que un experimento, test o procedimiento de medida arroja el mismo resultado en distintos ensayos". Esta definición no coincide con la que ofrece el Diccionario de la R.A.E. de "fiabilidad", pero se acepta en estadística como uno de los criterios de valoración de pruebas y medidas. El autor, sin embargo, diferencia *dependability* de *reliability* y, con el fin de distinguir los significados de ambos términos, hemos optado por traducir el primero como *confiabilidad*. (*N. del T.*)

Buscamos un equilibrio en el que la validez de constructo de la tarea no quede excesivamente comprometida por la fiabilidad y viceversa. La manejabilidad también interviene en esto: puedo preparar evaluaciones muy confiables cuya puesta en práctica fuese demasiado cara, por ejemplo, por exigir un evaluador externo para valorar a cada estudiante. El "reloj de una sola manilla" (Figura 5.1) recoge este equilibrio; si tenemos la validez del constructo, la fiabilidad y la manejabilidad a intervalos de 20 minutos siguiendo la esfera del reloj, ¿adónde queremos que señale la manilla? Una validez del constructo y una fiabilidad elevadas (0-20 minutos) se consiguen a costa de la manejabilidad, mientras que la validez y la manejabilidad se logran a costa de la fiabilidad (40-60 minutos). Deberíamos tratar de mantener la evaluación fuera de la zona fiable y manejable (20-40 minutos), en la que se encuentra con frecuencia. La razón es que, a menudo, es un significante muy débil de las destrezas que queremos evaluar y tiene una validez del constructo limitada, pero se escoge porque es barata ("eficiente") y fiable. Es más probable que se produzcan repercusiones negativas en la zona de 20-40 minutos.

Distintas finalidades pueden llevar a diferentes posiciones de la manilla. Quiero que mi piloto esté en "y diez", con independencia del gasto, es decir, que tenga una formación impartida en simuladores realistas y en vuelo real, más una evaluación rigurosa de éstas. Me conformaré con "pasados 20 minutos" en el caso de un test nacional de matemáticas, aunque es probable que pase por alto

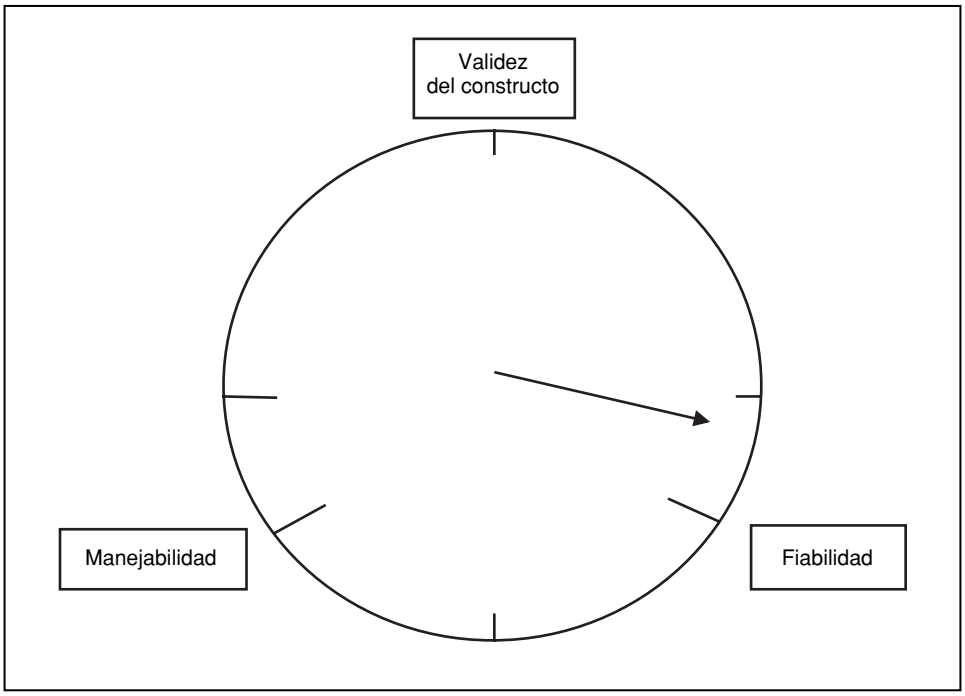


Figura 5.1. *El reloj de una sola manilla.*

las competencias aplicadas (la tentación podría ser poner pruebas de matemáticas a 30 minutos, estrictamente de papel y lápiz, pero eso saca a relucir la cuestión de por qué hay tan poca discusión acerca de la validez del currículum típico de matemáticas)[16]. Si el Lenguaje implica la escritura expresiva, podemos destacar la "autenticidad manejable" y reconocer que hay ciertos riesgos de fiabilidad ("¿pasados 45 minutos"?), por la dificultad de estandarizar la puntuación. Esto está empezando a parecerse un poco al juego "ponle la cola al burro", con los ojos tapados: ¿dónde entran el Arte, la Geografía y la Educación Física?

Una postura más útil es pensar en la evaluación sumativa en el aula como una manecilla "escondida" *que no tiene por qué apuntar simultáneamente hacia donde lo haga la manecilla del examen externo*. En un examen trimestral, la fiabilidad puede ser un motivo de preocupación menos importante que en un examen nacional. Por eso, de acuerdo con la finalidad y el argumento de imprevisibilidad, pueden realizarse evaluaciones más auténticas, en las que se haga más hincapié en la validez de constructo. No hace falta imitar las evaluaciones externas porque lo más importantes son las consecuencias para la enseñanza y el aprendizaje. Así, presentar una propuesta de negocio podría ser una actividad oral, de grupo, en vez de una actividad individual, de papel y lápiz, y la historia podría implicar una investigación local.

La consecuencia de esto puede ser un aprendizaje más basado en principios, aunque hay que preguntarse si la evaluación externa es suficientemente buena para prestar apoyo estratégico a este enfoque. Como vimos, en relación con el trabajo de RAMSDEN, en el Capítulo IV, los aprendices que utilizan un enfoque estratégico pueden decidir que un test no merece más que un aprendizaje superficial, con pistas y procurando recordar; de ahí la preocupación por mantener los tests fuera de la zona de los 20-40 minutos. Que los profesores tengan suficiente seguridad para trabajar en el nivel de los principios y evitar la imitación estricta de los exámenes externos cobra una importancia crítica. Aunque hay pruebas de que esto funciona[17] si las pruebas tienen una validez razonable, para muchos profesores que trabajan en sistemas que utilizan los resultados de los exámenes como medidas para la rendición de cuentas, esto puede encerrar un riesgo excesivo, como también los exámenes de años anteriores.

Hacerlos justos

La justicia de una prueba depende de que quienes la hacen sean capaces de ver el sentido de lo que se les pide. A menudo, esto se considera como una cues-

[16] En la comunidad de educación matemática, hay intensos debates acerca de lo que deban implicar las "matemáticas escolares" (validez de constructo), pero raramente se expresa esto en el currículum de Matemáticas. Algunos han cuestionado la importancia adjudicada a las Matemáticas, dado que, en realidad, solo utilizamos formas básicas de matemáticas, aunque los contextos sean complejos. Paul ERNEST afirma esto mismo en WHITE (2004). Joy CUMMING (2000) ha señalado que, aunque la *International Life Skills Survey* y el estudio PISA se ocupan de los "aprendizajes esenciales para la vida", tenemos currícula que consideran el álgebra y la trigonometría como "aprendizajes esenciales" (¿cuándo ha utilizado por última vez la regla del coseno?).

[17] Véase, por ejemplo: MCDONALD y BOUDO (2003).

tión relativa a la presentación, muy ligada a la legibilidad, pero tiene una profundidad mucho mayor. La cuestión de validez que está en el centro de este tema es si el examen permite a los examinandos demostrar lo que "saben, comprenden y pueden hacer" o si hay factores que lo obstaculizan. Esta preocupación no parte del examen mismo, pues hay cuestiones previas de equidad en torno a los recursos y el currículum[18]. La Tabla 5.1 resume algunos de estos problemas.

Problemas de acceso y curriculares como éstos influirán directamente en la justicia de los resultados de los exámenes. En el Capítulo Primero, vimos hasta qué punto ha sido parcial la idea histórica de justicia en relación con el acceso a los exámenes. Aunque fueron aclamados como una revolución con respecto a la injusticia de la selección mediante influencias, a menudo excluían grandes grupos de la población (por ejemplo, a las mujeres) y no se reconocían las premisas y los sesgos culturales que contenían. Esto pudo disimularse como si se les "perdonaran" los exámenes a ciertos grupos sociales, excluyéndolos así del acceso a otras oportunidades, de lo que se beneficiaban quienes los excluían (como los graduados que decían a los no graduados, peor pagados, que no merece la pena obtener un grado).

Tabla 5.1. *Cuestiones del acceso, el currículum y la evaluación en relación con la equidad*

Cuestiones del acceso	Cuestiones curriculares	Cuestiones de la evaluación
¿A quién se enseña y quién lo hace?	¿Qué conocimientos se enseñan?	¿Qué conocimientos se evalúan y se equiparan con el rendimiento?
¿Hay diferencias entre los recursos a disposición de los distintos grupos?	¿Por qué se enseña de un modo determinado a este grupo concreto?	¿La forma, el contenido y la modalidad de evaluación son adecuados para distintos grupos e individuos?
¿Qué se incluye de las culturas de los asistentes?	¿Cómo hacemos para que se enseñen de forma responsable y sensible las historias y culturas de las personas de color y de las mujeres? (APPLE, 1989).	¿Se refleja en las definiciones de rendimiento este conjunto de conocimientos culturales? ¿Cómo median los conocimientos culturales las respuestas de las personas a la evaluación de manera que alteren el constructo que se evalúa? (GIPPS y MURPHY, 1994).

Fuente: STOBART (2005).

[18] Véase: STOBART (2005).

Ahora, nos damos cuenta de que no existe la neutralidad cultural en la evaluación ni en la selección de lo que haya de evaluarse. Joy CUMMING señala que intentar presentar una evaluación como "acultural" es un error. "El saber acultural tiene unas raíces culturales definidas. Es un saber que está privilegiado en nuestras normas y procedimientos de prueba" (2000, pág. 4). Continúa la autora planteando dos cuestiones clave que se enlazan con las de la Tabla 5.1:

1. Cuando se establecen las normas y los contenidos de la prueba, ¿estamos realmente seguros de que éste es el saber que necesitamos?
2. ¿Estamos privilegiando realmente ciertos saberes para mantener una cultura dominante, asegurando, de ese modo, nuestra perpetuación como personas de éxito en la cultura educativa formal hasta la fecha?

Sostiene que lo que exigimos que debe estudiarse encierra un enorme bagaje cultural, redundante en gran parte, que tiene poca relevancia para la vida, mientras que ciertos temas importantes quedan exluídos.

Pruebas más justas

Este acceso más general y estas cuestiones curriculares más amplias forman parte del contexto de cualquier prueba. Sin embargo, los redactores de las pruebas pueden tener la sensación de que sus poderes para hacer algo al respecto son limitados. ¿Cuál es su aportación más directa a la justicia y la accesibilidad? Un primer paso es explicar cómo garantizan que su muestreo de la materia ofrezca oportunidades para los diferentes grupos que se someten a la prueba:

> Tenemos que promover una articulación más clara de los constructos de los redactores de pruebas o exámenes en los que se basa la evaluación, de manera que los examinandos y usuarios puedan evaluar la validez de constructo. Los redactores de las pruebas tienen que justificar la inclusión de contextos y tipos de modalidad de respuesta en relación con las pruebas que tenemos acerca de cómo interactúan con las diferencias de grupos y la experiencia curricular.
>
> (STOBART y GIPPS, 1998, pág. 48.)

Redacción reflexiva de las pruebas

Hemos avanzado mucho más allá de los supuestos ingenuos de los examinadores del siglo XIX acerca de que un test escrito en condiciones estandarizadas era intrínsecamente justo. *Nunca conseguiremos una evaluación justa, pero podemos hacerla más justa* y una parte de este proceso consiste en una discusión más completa y abierta. Como dice Caroline GIPPS:

> La mejor defensa contra la evaluación carente de equidad es la apertura. La apertura en relación con el diseño, los constructos y la puntuación saca a relucir los valores y los sesgos del proceso de diseño del test, da oportunidad para el debate sobre las influencias culturales y sociales y desarrolla la relación entre evaluador y aprendiz. Estos movimientos son posibles, pero requieren voluntad política.
>
> (1999, pág. 385.)

Cócteles de evaluación

Si la forma de una evaluación tiene una influencia diferencial en quienes se someten a ella, una manera de hacerla más justa consiste en ofrecer diversas formas de evaluación, de manera que quienes estén en desventaja en una evaluación tengan la oportunidad de ofrecer otra clase de pruebas de su dominio de la materia. Esta es una de las razones de la inclusión de las actividades en las titulaciones en Inglaterra; de la "evaluación del rendimiento" en los EE.UU., y de la evaluación del profesor en Suecia y Alemania. Esto no significa que estos enfoques diferentes no tengan sus propios sesgos. Eva BAKER y Harry O'NEIL han comunicado algunos hallazgos incómodos en las respuestas de minorías étnicas a la evaluación del rendimiento:

> La principal afirmación era que la reforma de la evaluación basada en la actuación es una creación de la comunidad mayoritaria que pretende retrasar el progreso de los niños desfavorecidos.
>
> (1994, págs. 13-14.)

La fuerza del enfoque del cóctel es que permite identificar cómo puede influir negativamente un formato de evaluación en algunos grupos, mientras que otros formatos facilitan el rendimiento[19].

Conclusiones

Este capítulo ha versado sobre la influencia de la evaluación en el aprendizaje y la enseñanza. *La fiebre de los diplomas: educación, cualificación y desarrollo*, de Ronald DORE, era una protesta contra el debilitamiento del aprendizaje provocado por el credencialismo, en especial en las naciones en vías de desarrollo. La escolarización acaba siendo más una forma de obtener los títulos adecuados que de aprender lo que se supone que representa el título. Así, el aprendizaje se convierte en un instrumento y la enseñanza pasa a ser una preparación fría del examen. Lo que importa es el título, no el aprendizaje. Tengo que reconocerle a DORE su incisivo análisis de cómo puede empobrecer el examen la enseñanza y el aprendizaje y su constante preocupación porque los estudiantes se dediquen a un aprendizaje profundo. Su solución a la "fiebre de los diplomas" consistía en limitar la escolaridad y utilizar las pruebas de capacidad para determinar quiénes deben ser seleccionados para una formación ulterior.

Aunque acepte gran parte de su análisis, he sostenido que sus soluciones para el problema no son útiles. Sus puntos de vista acerca de la inteligencia innata y la justicia de los tests de capacidad hacen que considere la capacidad como la causa subyacente del rendimiento, en vez de como una forma concreta de ren-

[19] Volviendo a los títulos, al arranque de este capítulo, las propuestas de TOMLISON para la reforma de los *A-levels* del GCE en Inglaterra trataban de estimular precisamente este tipo de cóctel, con exámenes, trabajos en clase y actividades fuera de la escuela combinados para la obtención del título. Las presiones políticas las archivaron, favoreciendo una acentuación aun mayor de la importancia de los exámenes.

dimiento. Sus ideas sobre el aprendizaje y la motivación no tienen suficientemente en cuenta el papel de los factores externos, incluidos los exámenes.

En consecuencia, propongo una vía diferente: mejorar la calidad de los tests de rendimiento de manera que promuevan algunos de los enfoques más profundos del aprendizaje que DORE quería, pero que suponía que los exámenes hacían imposibles[20]. Unos tests mejorados implicarían:

- presentar más claramente la *finalidad* de la evaluación (que también podría traducirse en una reducción de los tests);
- examinar la *adecuación a la finalidad* de la evaluación, haciendo hincapié en lo que se mide realmente, es decir, una cuestión de validez;
- vigilar las consecuencias, en especial las repercusiones sobre la enseñanza y el aprendizaje.

En el sistema, hay factores que operan en contra de una evaluación más válida ("auténtica"), por ejemplo, la necesidad de estandarización. Sin embargo, son posibles unas evaluaciones que promuevan una enseñanza y un aprendizaje más ricos, en especial las evaluaciones de los profesores en clase. Éstas suponen hacer *más explícita la finalidad de la evaluación* y relacionarla con los *objetivos* de la asignatura, que, con frecuencia, se refieren más a destrezas que a contenidos. Además, *estimulan un conocimiento más basado en principios* mediante el uso de preguntas y problemas menos previsibles, el paso de los enfoques de la enseñanza basados en exámenes de años anteriores, del estilo de *"¿cuándo te encuentras...?"*, a una enseñanza de *"¿qué pasa si...?"* La intención es hacer que la evaluación sea lo más auténtica posible, de manera que promueva las destrezas que pretende medir. Por último, esas evaluaciones tienen que ser todo lo *justas* que podamos, lo que suscita cuestiones relativas al acceso y al currículum, así como sobre el marco de esas evaluaciones.

Se trata de exigencias de cumplimiento difícil, pero posible. Una de las mayores amenazas contra este enfoque de un aprendizaje más eficaz surge cuando las escuelas y los docentes, por motivos de rendición de cuentas, se orientan hacia la consecución de resultados en los exámenes. Esto puede restringir en gran medida la enseñanza y el aprendizaje a la preparación de exámenes de competencias básicas. Nos ocuparemos ahora de esta amenaza.

[20] Esto podría incluir el contenido de los tests de capacidad como un elemento de unos tests de rendimiento generalizado. No es un mero juego de palabras: ese cambio modifica tanto el constructo (capacidad subyacente frente a rendimiento generalizado) como las inferencias que se hagan a partir de los resultados (talento independiente de la escuela frente a una aplicación más generalizada del aprendizaje), ambos agumentos clave de la validez.

CAPÍTULO VI

La larga sombra de la rendición de cuentas

> **Ley de Goodhart**[1]**:** Cuando una medida se convierte en objetivo, deja de ser una buena medida.
>
> Cualquier regularidad estadística tenderá a desaparecer cuando se ejerzan presiones sobre ella con fines de control.
>
> (Charles GOODHART.)

> La ley de Goodhart es el equivalente sociológico del principio de indeterminación de Heisenberg en física cuántica. Por regla general, la medida de un sistema lo perturba. Cuanto más precisa es la medida y más corta su escala temporal, mayor es la energía de la perturbación y mayor la imprevisibilidad del resultado.
>
> (Michael MCINTYRE, 2001.)

Uno de los argumentos centrales de este libro es que la evaluación puede tanto debilitar como estimular el aprendizaje. En el Capítulo V, hemos visto que, con el fin de que el individuo compita en el mercado laboral, la evaluación puede convertirse en un fin en sí misma: es el resultado lo que cuenta; la calidad del aprendizaje es irrelevante. En este capítulo, examinaremos cómo pueden afectar al aprendizaje las presiones de la rendición de cuentas para mejorar los resultados. A este respecto, es crítico el uso de medidas sencillas, como las puntuaciones de los exámenes, para juzgar si se han alcanzado los objetivos. El hecho de no alcanzarlos tiene consecuencias, tanto económicas como profesionales; en consecuencia, los exámenes adquieren una importancia decisiva y los resultados lo son todo.

---

[1] Esta es la reformulación de Marilyn STRATHERNS de la ley de GOODHART. La original, derivada de la economía por Charles GOODHART, Consejero Principal del Banco de Inglaterra, era: "en cuanto el Gobierno trata de regular un conjunto cualquiera de activos financieros, estos dejan de ser fiables como indicadores de las tendencias económicas"; esto se debe a que "las instituciones financieras pueden... idear con facilidad nuevos tipos de activos financieros" (http://www.atm.dampt.cam.ac.uk/people/mem/papers/LHCE/goodhart.html; consultada por última vez el 16 de noviembre de 2007*).

* Verificado el acceso el 8 de abril de 2010. (*N. del T.*)

Mi argumento es que, aunque el uso de los exámenes para rendir cuentas con un carácter decisivo, sus medidas estrictas y su insistencia en la mejora rápida, pueda reportar beneficios a corto plazo, rápidamente se degrada y es contraproducente. La ley de Goodhart, derivada de la economía, lo recoge muy bien: escoge un indicador estricto y observa cómo distorsiona lo que ocurre. Revisaré las presiones que la rendición de cuentas educativa, basada en los exámenes, al uso en Inglaterra y en Estados Unidos, dos de los sistemas más draconianos del mundo, ejercen sobre el aprendizaje. Sin embargo, necesitamos que se rindan cuentas, así que, ¿cómo sería una forma constructiva de hacerlo? En respuesta a esta pregunta, presento un modelo de *rendición inteligente de cuentas*.

Estamos tan acostumbrados a la rendición de cuentas en muchas esferas de la vida que resulta difícil definirla[2]. La utilizo aquí en el sentido vulgar de juzgar la eficacia de determinadas actividades, que pueden ser muy generales, como los servicios médicos, o restringidas a una iniciativa específica como por ejemplo, la reducción del absentismo escolar. Por regla general, el centro de atención se fija en la organización, por ejemplo, los hospitales, los sistemas de transporte y las escuelas, en vez de hacerlo sobre los individuos que reciben estos servicios. Esto suele implicar el uso de recursos, por lo que quienes los financian querrán saber qué se ha conseguido con sus inversiones. Cuando hay una necesidad dramática de mejorar un servicio, suele ponerse en juego una mezcla de incentivos y penalizaciones. Esta mezcla puede producir el efecto de choque buscado, pero cualquier mejora del rendimiento medido se producirá a menudo a costa de algunas consecuencias no buscadas que empiezan a distorsionar el sistema. Me limitaré a dos ejemplos: el transporte público y los tiempos de espera de los hospitales.

Objetivos de puntualidad

En el Reino Unido, muchas personas estarán acostumbradas a los objetivos de puntualidad del transporte público y a que se penalice económicamente a las compañías ferroviarias por el porcentaje de trenes que circulen con retraso. Los viajeros en ferrocarril también están acostumbrados a las modificaciones del horario de los trenes que incrementan constantemente la duración del mismo viaje, para que sea más difícil que lleguen con retraso. Es fácil que los viajeros de larga distancia hayan tenido la experiencia de viajar en un tren con mucho retraso al que detienen para que deje paso a otros trenes posteriores. Se hace esto para que los otros trenes no se retrasen y solo haya que pagar la compensación a los viajeros del tren muy retrasado. Incluso he viajado en un servicio "cancelado" que circuló por su trayecto pero sin retraso, puesto que había sido cancelado. Este artículo del *Guardian* resume perfectamente estas distorsiones:

[2] El volumen de 383 páginas *Uses and Misuses of Data for Educational Accountability and Improvement*, de HERMAN y HAERTEL (2005), no presenta, hasta donde he podido ver, una única definición formal de "rendición de cuentas"; da por supuesto que sabemos qué es.

Al haber cada vez más tráfico en las carreteras, ¿cómo van a circular los autobuses a su hora?... Una empresa de autobuses de Leeds parece haberlo resuelto... Los controladores de la ciudad de Leeds han dicho a los conductores que no recojan a viajeros cuando el tráfico sea demasiado denso. La empresa señala que "el único objeto de esta norma es recolocar el vehículo con el fin de que pueda cumplir su horario durante el resto del día".

Los tiempos de espera de los hospitales

Los tiempos de espera constituyen otro filón para obtener ejemplos. El Gobierno abordó el "escándalo" de los interminables tiempos de espera para las operaciones y estableció como objetivo reducir el tiempo que los pacientes tendrían que esperar para una operación desde que eran recibidos por el médico. Aunque esta medida produjo en parte el efecto deseado, también propició otros no deseados. Uno de éstos fue que, a menudo, pasaba más tiempo antes de ser recibido por el médico, dado que la cuenta atrás no empezaba hasta ese momento. Otro fue que los tiempos de espera se redujeran programando las operaciones más sencillas y menos necesarias por delante de las más graves y que más tiempo requerían.

Un claro ejemplo de esta clase de cinismo fue el del hospital que no salía muy bien parado en el indicador del tiempo que transcurría hasta que un paciente era atendido por el enfermero responsable de la clasificación en urgencias. El objetivo era de cinco minutos, pero muchos pacientes tenían que esperar mucho más tiempo. Al investigar la situación, la dirección del hospital descubrió que la razón era que, al estar el hospital a las afueras de la población, los pacientes llegaban con frecuencia juntos en el autobús que circulaba cada hora, por lo que los enfermeros no podían ver a todos los pacientes en cinco minutos. ¿La solución? El hospital se puso en contacto con la empresa de autobuses y acordaron trasladar la parada del autobús, retrasándola hasta la carretera, de manera que los viajeros tuvieran que ir andando hasta la clínica y se espaciaran los tiempos de llegada (los lesionados en las piernas mucho después que los lesionados en las manos, los jóvenes antes que los viejos, etc.). Los tiempos de espera se redujeron considerablemente y el hospital mejoró su comportamiento en este indicador[3].

La rendición de cuentas en la escuela

A estas mismas presiones se enfrentan escuelas y universidades cuando establecen objetivos para mejorar sus resultados en los exámenes. Los responsables políticos se han dado cuenta de que la evaluación puede utilizarse como

[3] Este ejemplo se debe a Isabel NISBET, exdirectora de Política y exdirectora de Adaptación a la Práctica del *General Medical Council* (GMC). Presentación en el congreso anual de la *Association of Educational Assessment* (AEA – Europe 2005), celebrado en Dublín.

una poderosa herramienta para la reforma de la educación. Lo que se ponga a prueba, sobre todo si conlleva consecuencias importantes, determinará lo que se enseñe y cómo se enseñe. Por tanto, esta es una vía más directa que el desarrollo paciente del currículum y la pedagogía, y produce unos resultados claros de manera relativamente barata. El modelo se ajusta también a la necesidad del economista de indicadores sencillos que puedan interpretarse para comprobar si la inversión produce beneficios.

Esto no es nada nuevo. En el Capítulo Primero, vimos que los tests se han utilizado históricamente con fines de rendición de cuentas. La introducción de los exámenes universitarios en Cambridge trataba de mejorar la calidad del estudio de los estudiantes, igual que los exámenes de las escuelas de secundaria se consideraban como una forma de mejorar la educación en las escuelas privadas de "clase media". En 1840, hubo exámenes en Boston (EE.UU.), con el fin de hacer comparaciones entre aulas y escuelas. El plan de "pago por resultados", de Robert LOWE, para promover la enseñanza de las "tres erres"* en las escuelas elementales estatales fue quizá el ejemplo supremo en Inglaterra.

Estos planes que tienen como finalidad la rendición de cuentas hacen hincapié en *cumplir los objetivos* impuestos por quienes corren con la financiación. El problema es que, a menudo, estos son más bien expresiones de aspiraciones que objetivos empíricos: se basan en la creencia social de que los niños deberían hacer las cosas mucho mejor de lo que las hacen y que la forma de conseguirlo es exigir más al sistema. La motivación procede generalmente de la impaciencia política por la aparente reticencia del sector público a cambiar, acompañada por la retórica política de los "objetivos ambiciosos" como "palancas para el cambio", junto con "el apoyo y la presión". En Inglaterra, esto ha incluido objetivos para el mismo Gobierno, por lo que la imposibilidad de alcanzar los objetivos de alfabetización y de aritmética de los niños de 11 años en 2002 condujo, en parte, a la dimisión del entonces Secretario de Estado para la Educación[4].

* Las *three Rs* o "tres erres" son las llamadas "disciplinas básicas": *reading, writing* y *arithmetic*, "lectura, escritura y aritmética". (*N. del T.*)*

[4] Estos objetivos, por ejemplo, el de 2004, de que el 85% de los niños y niñas de 11 años alcanzaran el nivel 4, todavía no se ha conseguido, aunque el plazo se amplió hasta el 2006. En 2007, el 80% había alcanzado el nivel 4 en Lenguaje y el 78%, el nivel 4 en Matemáticas. A pesar de la imposibilidad efectiva de alcanzar los objetivos, en 2007 se establecieron otros, nuevos y más exigentes, para 2009 (*Times Educational Supplement*, 6 de julio de 2007, pág. 8).

Este objetivo, que el 85% alcanzara el nivel 4, no formaba parte de la idea original de los niveles propuestos en el informe del TGAT* (1988), que introdujo la estructura de niveles en el currículum nacional. En éste, el nivel 4 se consideraba como el nivel típico de rendimiento de la mayoría de los niños y niñas de 11 años, pero se daba por supuesto que habría una proporción importante que quedaría por debajo (y por encima) de este nivel.

* *Task Group on Assessment and Testing*: "Grupo de trabajo sobre evaluación y exámenes". (*N. del T.*)

Tests para la rendición de cuentas: No Child Left Behind* *(EE.UU.) y la evaluación del currículum nacional en Inglaterra*

Los ejemplos de tests para la rendición de cuentas de Estados Unidos y de Inglaterra que utilizo aquí pueden considerarse formas extremas de lo que sucede en muchos países. Lo que los distingue son la escala y las consecuencias de los tests.

Las características clave de esta *rendición de cuentas basada en los exámenes* son:

- *metas:* se presentan como "normas" o "estándares", que representan el nivel de rendimiento deseado;
- *objetivos:* los niveles exigidos de rendimiento se especifican tanto como objetivos anuales de mejora y a largo plazo;
- *medidas:* las pruebas mediante las que se juzga el rendimiento. Pueden ser los resultados de los tests utilizados con otros fines o tests específicos de rendición de cuentas que no tengan otras finalidades importantes;
- *consecuencias:* los resultados están vinculados a penalizaciones y recompensas. Estos son los que consiguen que los tests sean tan importantes, porque el futuro de una escuela pueda estar determinado por los resultados.

Que ningún niño se quede atrás (*"No Child Left Behind"*, NCLB)

En los Estados Unidos, la política educativa es esencialmente una responsabilidad de cada Estado, celosamente guardada. En este sentido, hay poco control central y el Gobierno Federal solo aporta una pequeña proporción de los costes de la educación. Por regla general, los programas nacionales se acogen al Título 1**, una disposición que trata de promover oportunidades educativas para los más desfavorecidos. Bajo este título, la *No Child Left Behind Act* se convirtió en ley en 2002. La ley exige que las escuelas muestren un progreso regular hacia la meta de que todo el alumnado alcance un nivel elevado, con el objetivo de que *todos* sean competentes en 2014. Todos los niños se someten anualmente a exámenes de lengua y matemáticas, a los que seguirán los de ciencias entre los grados 3.º y 8.º***, así como en un curso de *high school*.

La ley aprovecha la experiencia de las pruebas estatales para calificación de los alumnos y rendición de cuentas del profesorado. De particular importancia para las escuelas de secundaria fue el "milagro de Texas", que había contemplado mejoras espectaculares de los niveles, en especial los de los alumnos pertenecientes a minorías, como consecuencia de la exigencia de someter a todos los

* "Que ningún niño se quede atrás". (*N. del T.*)

** Título 1 de la Ley de Enseñanza Primaria y Secundaria" (*N. del T.*)

*** Equivalentes, por las edades de alumnos y alumnas, a 3.º de Primaria y 2.º de E.S.O., respectivamente, en el sistema educativo español. (*N. del T.*)

estudiantes a la *"Texas Assessment of Academic Skills" (TAAS)** en 10.º grado**, y de imponer objetivos relativos al abandono de los estudios. El Gobernador de Texas era George W. Bush, y el Superintendente de las escuelas de Houston era Rodney Paige, que se convirtió en el Secretario de Educación de aquél. La cuestión es si esto fue un milagro o un espejismo***.

Las normas y niveles y las evaluaciones de la NCLB son responsabilidad de cada Estado. Todos los distritos escolares y todos los Estados deben demostrar un "progreso anual suficiente" (PAS)**** de *cada* subgrupo (origen étnico, género, inglés como segunda lengua, necesidades especiales, estatus de económicamente desfavorecido) de cara al objetivo de 2014. En consecuencia, el PAS supone alcanzar los objetivos anuales fijados por el Estado, que habrá tenido que trazar un plan para lograr las mejoras necesarias para conseguir el 100% de competencia en 2014. La arbitrariedad de estos objetivos, cuestión sobre la que volveré, complica el problema de haber establecido unos objetivos numéricos con un fundamento bastante reducido. Esto ha llevado a que los Estados escojan diferentes formas de llegar a cero en 2014, optando algunos por incrementos modestos durante los cinco años siguientes a la promulgación de la norma y unos incrementos casi milagrosos en los cursos finales. Otros han planeado un incremento anual regular, suscitando problemas de justicia con respecto a las escuelas desfavorecidas que hacen buenos progresos año tras año. Es posible que éstas no consigan cumplir esos objetivos arbitrarios a causa de las bajas proporciones de éxito de las que partieron.

La imposibilidad de cumplir los objetivos del PAS conduce a "acciones correctoras" progresivas:

1. Las escuelas que no consigan el PAS durante dos cursos consecutivos se considerará que "necesitan mejorar", y se les facilitarán asistencia técnica. A sus alumnas y alumnos se les ofrecerá la posibilidad de elegir otras escuelas públicas y de sufragar los costes del desplazamiento hasta ellas.
2. Las escuelas que no consigan el PAS durante tres cursos deberán ofrecer a sus alumnas y alumnos de familias con rentas más bajas la oportunidad de recibir enseñanza de los servicios suplementarios.
3. Tras cuatro años de incumplimiento del PAS, hay que adoptar una de estas iniciativas: reemplazar al personal docente; nombrar asesores externos; ampliar la jornada o el curso escolar; cambiar la organización interna del centro.
4. Tras cinco años de fracaso, la escuela tendrá que ser reestructurada. Para ello, deberán adoptar una de estas iniciativas: reabrir la escuela como *charter school*****; reemplazar a todo o a la mayoría del personal docen-

* "Evaluación de destrezas académicas de Texas". (*N. del T.*)

** Equivalente a 4.º de E.S.O., en el sistema educativo español. (*N. del T.*)

*** El autor juega con las palabras *miracle* ("milagro") y *mirage* ("espejismo"). (*N. del T.*)

**** *Adequate yearly progress*, al que se alude habitualmente por sus siglas: *AYP*. (*N. del T.*)

***** Las *charter schools* son centros públicos a los que se exime del cumplimiento de ciertas normas y reglamentos que obligan a la mayoría de los centros públicos aunque, a cambio, estén obligadas a conseguir unos resultados determinados de los que tienen que rendir cuentas. (*N. del T.*)

te; pasar a depender directamente del Estado, sustrayéndose al distrito escolar, o reestructurar la dirección del centro.

Las *recompensas* para las escuelas que superen el PAS previsto durante dos cursos consecutivos consisten en la concesión de los premios estatales de rendimiento académico y la designación de "escuelas distinguidas" para las que hayan obtenido los mayores progresos. También hay recompensas económicas para los maestros y profesores de esas escuelas[5].

Desde un punto de vista político, este sistema parece severo pero justo, sobre todo porque el hecho de centrarse en subgrupos implica que el bajo rendimiento de los grupos minoritarios no pueda quedar enterrado en unas estadísticas que muestren una mejora general. Esta presión para obtener resultados no tiene precedentes y, como ha calculado Robert LINN (2005), implica una tasa de mejora que quizá no se haya visto jamás.

La evaluación del currículum nacional en Inglaterra

La introducción de un currículum nacional en 1988 supuso un cambio radical para la educación en Inglaterra que, hasta entonces, era conocida por su enfoque muy local del currículum y de la evaluación. Iba acompañado de las evaluaciones nacionales, que fueron convirtiéndose progresivamente en exámenes estandarizados de Lenguaje, Matemáticas y Ciencias. A estos exámenes se someten niños y niñas de 7, 11 y 14 años, y los exámenes nacionales para el GCSE constituyen el vehículo para poner a prueba a los alumnos y alumnas de 16 años. En otros cursos, hay pruebas "opcionales" orientadas a supervisar el progreso. Las pruebas del currículum nacional son en la actualidad evaluaciones convencionales de respuestas para señalar (después de algunos intentos iniciales de crear tests referidos a criterios), a partir de las cuales los niños reciben un nivel, que representa su rendimiento en relación con el currículum nacional. A los 16 años, el rendimiento se juzga en términos de calificaciones del examen[6].

En cumplimiento del "principio de prepotencia administrativa" (pág. 15), la principal utilidad actual de los resultados de las pruebas del currículum nacional es la rendición de cuentas. El porcentaje de alumnos de 11 años que consiguen el nivel 4 o superior en cada *junior school** se publica en tablas de rendimiento, que los media transforman rápidamente en "clasificaciones de liga". En las escuelas de secundaria, el porcentaje de alumnas y alumnos de 16 años que consiguen cinco o más calificaciones A*-C en el GCSE es el indicador clave, aunque, junto a éstas, se están introduciendo medidas de "valor añadido" para indicar el progreso realizado

[5] De NICHOLS y cols. (2005), pág. 7.

[6] A los 7 años, se preveía que los alumnos se encontraran en los niveles 1-3; a los 11, estarían en los niveles 3-6, y a los 14, en los niveles 4-8. Se pretendía que un nivel representara el progreso de dos cursos, con los niveles típicos 2, 4 y 5-6 al final de las tres "etapas clave". Aunque se pensaba tener 10 niveles y llegar hasta los 16 años, la etapa clave 4 del currículum nacional (14-16 años) tenía que coordinarse con la estructura vigente de los exámenes y la información se presentaba con las calificaciones del GCSE (A*-G, siendo el objetivo clave cinco notas entre A* y C).

* En Inglaterra, escuelas de primaria. (*N. del T.*)

entre "etapas clave"*. Unos resultados deficientes suponen una mala publicidad, así como inspecciones. Si la escuela no consigue mejorarlos, se considerará "en situación de riesgo". Los equipos de inspección pueden imponer "medidas especiales", que influyen directamente en la enseñanza. El fracaso en la aplicación de estas medidas especiales, supondrá el cierre o la reorganización de la escuela.

Ambos ejemplos suponen tests de rendición de cuentas que tienen consecuencias tanto económicas como administrativas para los centros. La intención política es clara: para evitarlas, las escuelas tendrán que actuar mejor. Esto es positivo; el lado tenebroso del asunto es lo que las escuelas hacen para conseguir las mejoras exigidas, aparte de una "mejor enseñanza". De esas consecuencias de los tests de rendición de cuentas nos ocuparemos a continuación.

Consecuencias: pretendidas y no buscadas

El centro de atención de esta revisión de consecuencias es el modo en el que los tests de rendición de cuentas pueden afectar la enseñanza y el aprendizaje. La intención política es "elevar" los niveles influyendo en lo que se enseña y en cómo se enseña, y motivando a docentes, alumnas y alumnos, escuelas y administraciones educativas para que trabajen más y con mayor conciencia de su finalidad. Si los resultados mejoran, "la presión y el apoyo" habrán sido eficaces.

Las realidades de la implementación son, por supuesto, mucho más desordenadas. Dan KORETZ y sus colaboradores (2001) han identificado siete tipos de respuestas del profesorado a los exámenes nacionales para calificación y rendición de cuentas y organizaré el comentario en torno a ellos[7]. Tres son en gran parte positivos: *dedicar más tiempo de enseñanza a la materia; trabajar más para abarcar más material*, y *trabajar con más eficacia*. Uno es decididamente negativo: *hacer trampas*. Los otros tres son ambiguos y pueden inclinarse a un lado o a otro, dependiendo del contexto: *redistribuir el tiempo de enseñanza; alinear la enseñanza con las normas o estándares*, y *preparar para el examen*. No se trata de categorías claramente diferenciadas, pues, a menudo, unas se confunden con las otras. Mi punto de vista al respecto es, en general, menos positivo, a causa del modo en que influyen estas presiones en la enseñanza y el aprendizaje. He organizado las siete respuestas de KORETZ en tres categorías principales: *motivar, priorizar* y *maximizar*.

Motivar

Los docentes trabajan más y más eficazmente

Esta es la ilusión del político: una política que consiga que los funcionarios públicos trabajen más. Los supuestos implícitos son: mucho dinero, pocas mejo-

* *Key stages*: son etapas o ciclos fundamentales, cada una de las cuales consta de varios cursos. (*N. del T.*)

[7] KORETZ y cols. (2001): *Towards a Framework for Validating Gains Under High-Stakes Conditions*, CSE Technical Report 551.

ras y mucha relajación en el sistema. Robert LOWE introdujo los pagos por resultados cuando oyó que los maestros y profesores solo se molestaban en enseñar a los niños más capaces (Capítulo Primero). *No Child Left Behind* fue una respuesta a los reducidos progresos que estaban haciendo las escuelas con alumnos pertenecientes a minorías étnicas y socialmente desfavorecidas. Hay quien ha comparado las escuelas con los acorazados de la II Guerra Mundial:

> Grandes, poderosos, pesados, con tripulaciones enormes... Cuando se ordena cambiar el rumbo, lo hacen, pero hay retrasos significativos entre el momento en que se ordena el cambio de rumbo y el buque navega en una dirección diferente.
>
> (GRAHAM, 1995, pág. 3.)

Los responsables políticos se han dado cuenta de que la evaluación es una forma de tener rápidamente "la sartén por el mango", de manera que el resto del sistema tenga que seguirlos. La evidencia es que las escuelas y los docentes están trabajando mucho, aunque no siempre considerarán productivo este trabajo, sobre todo si se canaliza hacia una mayor burocracia[8]. No obstante, hay que considerar que añadir más días al curso escolar, instituir clases de recuperación fuera del horario escolar normal o dedicar más tiempo del curso escolar a la enseñanza efectiva son, en general, efectos positivos.

Desde mi punto de vista, en Inglaterra, estas presiones han producido beneficios. El beneficio clave es que se espera que las escuelas mejoren año tras año. La importancia de esto se relaciona con el Capítulo II y el contraste entre las ideas *maleables* y *fijas* de la capacidad. Una razón de las bajas expectativas de maestros y profesores con respecto a muchos alumnos y alumnas está en la psicología popular de la capacidad fija. Si un alumno suspende el 11+, poco puede esperarse de él en el plano académico. Por eso, año tras año, poco se ha conseguido, académicamente hablando, en la mayoría de las escuelas de secundaria (el 80% de los alumnos suspendió el 11+), porque "no se puede hacer mucho con esta gente", precisamente el enfoque que criticara BINET. Los tests de rendición de cuentas indican que *se espera la mejora*. Aunque la expectativa de que *todos* los niños sean muy competentes en 2014 (NCLB) es muy poco realista, al menos transmite un mensaje maleable: se prevé que las escuelas mejoren el rendimiento de todos sus alumnos: ya no puede aceptarse el "no puede hacerse mucho". Combinemos esto con unos incentivos adecuados, más palo que zanahoria en el presente, y las escuelas trabajarán más.

Trabajar con más eficacia

El esfuerzo aislado no es la solución, y "trabajar de un modo más inteligente" es el "mantra" que las escuelas se han tomado con el máximo interés. La tarea consiste en maximizar los resultados y el lado positivo de esto es que la mejora del rendimiento será un objetivo más claro. Un peligro de este enfoque es que

[8] Véanse, por ejemplo, los informes de los docentes en el proyecto VITAE, en DAY y cols. (2007): *Teachers Matter*, o el informe de HAMILTON (2003).

puede deslizarse hacia un uso cínico de las reglas en provecho propio, en el que el aprendizaje efectivo carezca de importancia. En sentido positivo, los tests de rendición de cuentas pueden llevar a que toda la escuela se esfuerce para preparar a los estudiantes; para asignar a los mejores docentes a las clases que vayan a someterse al examen, y para implicar a padres y alumnos en el esfuerzo por hacerlo bien. Esto ha ocurrido en Inglaterra, aunque las pruebas nacionales tienen relativamente poca importancia para los alumnos, dado que sus consecuencias directas son pocas. Esto se debe a que la selección de la escuela secundaria y la selección de asignaturas para el GCSE habrán tenido lugar *antes* de las pruebas, aunque puedan influir en los escenarios de enseñanza en los que se desenvuelvan los alumnos. Las consecuencias decisivas son para los maestros o profesores y para el centro, aunque las presiones sobre los docentes y las escuelas para que rindan bien se transmiten con frecuencia a los alumnos y alumnas.

El aspecto oscuro de este enfoque estratégico surge cuando los resultados mejoran por medios no relacionados con el aprendizaje. Quién aprueba y mediante qué tests y exámenes se convierten en cuestiones clave (véase: *maximizar*), como también quién obtiene ayuda extra.

Priorizar

Redistribuir el tiempo de enseñanza

KORETZ y sus colaboradores consideran esto como una de las respuestas ambiguas. La *National Literacy Strategy* y la *National Numeracy Strategy* inglesas son ejemplos espectaculares de control central de estos procesos. Estas estrategias no solo fijaban con detalle qué había que enseñar en las clases de Lenguaje y de Matemáticas de las escuelas de primaria, sino que también estipulaban cuánto tiempo había que dedicar a hacerlo, con consejos acerca del cuándo. Así, en las escuelas, teníamos la *hora del Lenguaje* y la *hora de la Aritmética*. Las evaluaciones de estas estrategias indicaban que, por regla general, tenían una influencia positiva, sobre todo la de Aritmética, si los maestros no se habían sentido muy seguros a la hora de enseñar Matemáticas, aunque el equipo canadiense que llevó a cabo estas evaluaciones tendía a tomar las puntuaciones de las pruebas en su valor nominal[9] (véase más adelante).

La redistribución implica dejar más tiempo para determinadas materias y para lo que se concentra en ellas. Tenemos ejemplos de Inglaterra de cómo se restringía el currículum de 6.º, el "curso de los exámenes" al final de la escuela primaria, con el fin de concentrarse en las materias sometidas a examen. Bill

[9] Harvey GOLDSTEIN (2003) ha criticado la aceptación indiscutida de los evaluadores canadienses (véase: EARL, 2003a) de que la elevación de las puntuaciones significa una mejora de los niveles subyacentes. Véase: "Evaluating the evaluators": http://www.cmm.bristol.ac.uk (consultada el 16 de noviembre de 2007*). Véase también la crítica de Peter TYMMS (2004).

* Verificado el acceso el 8 de abril de 2010. La dirección completa es: http://www.cmm.bristol.ac.uk/team/HG_Personal/evaluating-the-evaluators.htm (*N. del T.*)

BOYLE y Joanna BRAGG estudiaron cuánto tiempo se dedicaba a Matemáticas y a Lenguaje, descubriendo que había aumentado durante el período 1996-2004 hasta el punto de que estas dos asignaturas ocupaban más de la mitad del horario curricular, a expensas de las otras nueve asignaturas que también había que enseñar.

La respuesta política a críticas similares de este "embotar y restringir", incluso del Inspector Jefe de Escuelas, consistió en elaborar: *Excellence and Enjoyment* DfES, 2003), que animaba a las escuelas a integrar la enseñanza y el trabajo en un currículum más amplio. La evidencia del proyecto VITAE[10], que siguió a maestros de 6.º hasta 2005, era que estos maestros continuaban restringiendo gravemente lo que se enseñaba hasta que no realizaban los exámenes de mayo, momento en el que daba comienzo una enseñanza y un aprendizaje "divertidos", una vez restaurados el Arte, la Natación y las Humanidades.

El panorama es muy similar en los EE.UU. Para el maestro, solo cambia el mes:

> Ahora, estoy básicamente asustada por *no* enseñar para el examen. Sé que mi modo de enseñar está sentando unos fundamentos mejores para mis alumnos, así como llevándolos a amar el aprendizaje. No puedo esperar en cada curso a que pase marzo para poder dedicar los últimos dos meses y medio a enseñar como quiero enseñar, del modo que sé que entusiasma a mis alumnos.
>
> (DARLING-HAMMOND y RUSTIQUE-FORRESTER, 2005, pág. 299.)

Tiempo de práctica

Enseñar para el examen conlleva la práctica del mismo, que puede consumir el tiempo extra otorgado a la materia. La evidencia estadounidense demuestra que en los Estados que otorgan a los exámenes una importancia decisiva se dedica más tiempo a la práctica de los exámenes que en los Estados en los que las pruebas tienen una importancia más relativa. Los maestros que están en entornos que dan una importancia decisiva a estas pruebas, también empiezan a practicar antes y es más probable que utilicen tipos específicos de materiales que se parezcan mucho a los exámenes estatales. Se ha estimado que más del 20% del tiempo de enseñanza en Carolina del Norte y más de 100 horas por asignatura en Arizona se dedican a la práctica de los exámenes[11]. En Inglaterra, un estudio reciente financiado por el Gobierno descubrió igualmente que, en 6.º, la enseñanza estaba "dominada por períodos intensivos de preparación para el examen nacional"[12]. La preocupación está relacionada con la calidad del aprendizaje "preparatorio para el examen"; ¿hasta qué punto se centra la actividad en la técnica de realización del examen más que en el aprendizaje efectivo?

[10] DAY y cols. (2006): *Variations in Teachers' Work Lives and Effectiveness*; DAY y cols. (2007): *Teachers Matter*.

[11] HAMILTON (2003), pág. 35.

[12] BEVERTON y cols. (2005), pág. 5.

Redistribución dentro de la materia

Se reserve o no un tiempo adicional, puede darse aún una redistribución de lo que se enseña *dentro* de la asignatura. Hay gran cantidad de pruebas de que la enseñanza para el examen restringe el currículum a lo que probablemente aparezca en el examen[13]. En Inglaterra, los exámenes de Lenguaje del currículum nacional no contemplan hablar ni escuchar, a pesar de que es uno de los tres aspectos del currículum nacional, de manera que, en 6.º, se refuerza la lectura y la escritura (hay una evaluación del maestro relativa a hablar y escuchar, pero no forma parte de los resultados publicados, por lo que no se toma en serio). BEVERTON y colaboradores resolvieron así la situación: la enseñanza tendía a ser de estilo más formal y se centraba en los requisitos del sistema nacional de exámenes... era obvio que la necesidad percibida de preparar para ellos ocupaba gran parte de la práctica en el aula, al menos en gran parte de 6.º" (pág. 7). Los investigadores facilitaron una elocuente ilustración de la enseñanza que requiere la *National Literacy Strategy*. Aquí, el maestro ha entrenado a los niños hasta el caso de que, sin apuntarles, no pueden dar la respuesta "creativa" que se les pide: están demasiado bien entrenados para ello:

Maestro: ¿Qué busco cuando califico este relato?
Alumnos: La puntuación, la caligrafía.
Maestro: ¿En qué cosas importantes me fijaré?
Alumnos: El tiempo pasado, el comienzo de los párrafos, los sinónimos.
Maestro: Me fijaré en el relato completo; ¿qué es importante?
Alumnos: Un comienzo interesante.

(BEVERTON y cols., 2005, pág. 74.)

Una vez más, esto está muy en sintonía con los hallazgos estadounidenses, por ejemplo, estudios de maestros de escritura que muestran que, a consecuencia del formato del examen de escritura, habían comenzado a insistir en que los alumnos buscaran errores en los documentos, en vez de elaborar sus propios trabajos. Si el examen no pide respuestas de desarrollo, no es probable que se enseñen[14]. Está también el fenómeno del "ensayo en cinco párrafos" de Texas, en el que los estudiantes pasaron de redactar textos extensos a hacer coincidir sus escritos con las exigencias del examen, en el que un desarrollo en cinco párrafos, con cinco oraciones en cada párrafo, obtendría un aprobado. Como observa David HURSH:

> Como es probable que los estudiantes culturalmente privilegiados de clase media y de clase alta aprovechen su capital cultural para aprobar los exámenes, son los estudiantes desfavorecidos quienes tendrán que hacer los ejercicios extra. Por desgracia, aprender a escribir ensayos de cinco párrafos, de cinco oraciones cada uno no se traslada demasiado bien a la lectoescritura que sale fuera del marco del examen ni fuera de la escuela. Al esperar menos de los estudiantes desfavorecidos, se quedan muy atrás.
>
> (Pág. 614.)

[13] Véanse, por ejemplo: MADAUS (1988); HAMILTON (2003); BEVERTON (2005), y LINN (2000).
[14] Véase: HAMILTON (2003), pág. 35.

Cambio negativo

Para mí, la evidencia de los actuales tests de rendición de cuentas de Inglaterra y los EE.UU. transforma el estatus "ambiguo" de la *redistribución*, acercándola a un estatus negativo. La redistribución ha supuesto aquí una restricción tanto del currículum como de lo que se estudia en las asignaturas sometidas a examen. Esto no significa que la *redistribución* sea siempre negativa. Como vimos en el Capítulo V, puede ser positiva cuando una materia no se ha enseñado lo suficiente, cuando los docentes tienen un conocimiento muy limitado de la materia o cuando el currículum está tan poco especificado que surgen problemas relativos al derecho a enseñar la materia en cuestión. Hay importantes ejemplos positivos de ello procedentes de países como Ruanda, pero van acompañados de climas de rendición de cuentas menos punitivos.

Maximizar: Alinear y entrenar (¿y hacer trampas?)

Si la importancia es suficientemente grande y las medidas suficientemente estrechas, la ley de GOODHART predice que empezaremos a presenciar distorsiones y la corrupción del sistema. Lo que vemos, en el caso de los tests de rendición de cuentas, es un conjunto de respuestas que van desde lo ambiguo a lo inequívocamente negativo.

Alinear los niveles o estándares

Esto es una preocupación especialmente norteamericana, aunque sirve para otros países en los que no haya un currículum de alcance nacional. Como el currículum y los métodos de enseñanza son responsabilidades esencialmente locales, las instituciones financiadoras, tanto estatales como federales, han tenido históricamente un control limitado sobre ellas. En consecuencia, la evaluación ha sido una palanca clave para alinear lo que se enseña con los *niveles* o *estándares* deseados. Se interpretan éstos como niveles de asimilación de contenidos curriculares concretos: normalmente, las "materias básicas" de Matemáticas, Lenguaje y, con menor frecuencia, Ciencias. En este caso la lógica de la política es clara: si quieres que hagan bien los exámenes, tendrás que enseñar lo que éstos preguntan.

Un ejemplo precoz de esto fueron los *tests de competencia mínima* (TCM), una consecuencia del movimiento de "vuelta a lo básico" de la década de 1970. La percepción de que los niveles estaban cayendo, impulsó los TCM, que reflejaban la nueva atención prestada a los *resultados*, en vez de a las *aportaciones*, como mejores materiales o nuevos métodos de enseñanza. Se trataba, normalmente, de tests de Lectura y de Matemáticas y había que aprobarlos para obtener el título de graduado en la escuela. Estaban referidos a criterio, en el sentido de que había que alcanzar un determinado nivel de rendimiento. Como los niveles o estándares se fijaban para todo el Estado y los Estados no querían aparecer como incapaces de promover las competencias básicas, las tasas de aprobados

aumentaron rápidamente, año tras año. Esto fue acompañado de la reducción progresiva de su uso como criba para la graduación: en 1985, 33 Estados requerían que los estudiantes se sometieran a las pruebas TCM, pero solo 11 exigían aprobarlos como requisito para la graduación en la *high-school*[15]. La estrechez de este enfoque y los ejercicios de ensayo de los exámenes que lo acompañaban provocaron una violenta reacción promotora de las competencias de orden superior. El resultado fue el movimiento (de corta vida) de evaluación del rendimiento, que utilizaba evaluaciones abiertas y más complejas.

Podemos contemplar un patrón similar al TCM en el régimen de pruebas del *No Child Left Behind*. La diferencia es que el centro de atención de la rendición de cuentas son las escuelas en vez de los alumnos, aunque ya estén emergiendo con claridad patrones similares de ceñir la enseñanza al examen[16]. Aunque el Congreso conocía los riesgos de las restricciones de los tests de opciones múltiples y promoviera una combinación de formatos de evaluación, la evidencia es que la mayoría de los Estados utilizan en la actualidad tests de opciones múltiples porque son más sencillos, fiables y baratos[17]. Para mí, esto indica un cambio negativo porque las normas o estándares se convierten en lo que está en los tests ("enseñar el test"), en vez de unos campos más amplios de los que los tests son una muestra ("enseñar para el test"). Hay también pruebas de que a los estudiantes que tienen que hacer verdaderos esfuerzos para aprobar, muchos de los cuales son de minorías étnicas, se les imparte un currículum restringido y se les impone una dieta de práctica de tests, mientras que a otros se les imparte una enseñanza y se les ofrecen unas experiencias de aprendizaje más ricas[18].

Entrenamiento

El "entrenamiento" es una característica de las pruebas que cuentan para nota en todo el mundo y a menudo representa una forma del capital cultural de la clase media, cuando se trata de conseguir cierta ventaja comprando una ayuda adicional. En algunos países, por ejemplo, Brasil, en los que estas pruebas no están relacionadas con el currículum escolar, este entrenamiento adopta la forma de clases y centros de preparación intensiva que operan independientemente del sistema escolar. No obstante, en relación con la rendición de cuentas escolar, las escuelas ofrecen ese entrenamiento con el fin de ayudar a los estudiantes a obtener unos resultados que ayuden a la escuela. Esto tiene un aspecto positivo, pues los docentes dedican más tiempo y orientan más a los estudiantes, pero el aspecto negativo ocupa un lugar muy destacado, que adopta la forma de lo que David Gillborn y Deborah Youdell (2000) denominaron *selección educativa*, en la que se distribuyen unos recursos limitados a determinados grupos. A diferencia de su equivalente médico, no son los más necesitados quienes reciben la máxima ayuda, sino los estudiantes que están a punto de aprobar y a los que, por tanto, pue-

[15] Haertel y Herman (2005), págs. 13-14.
[16] Véase: Hursh (2005).
[17] Véase: Haertel y Herman (2005).
[18] Darling-Hammond y Rustique-Forrester (2005).

de promoverse con ayuda adicional. En los EE.UU., a estos alumnos se los llama "niños burbuja".

En Inglaterra, esto está aún más sistematizado, porque el Gobierno financia escuelas que cuentan con "clases de refuerzo". Éstas facilitan entrenamiento adicional previo al examen a los niños a quienes sus maestros consideran que están inmediatamente por debajo del umbral crítico del currículum nacional. En consecuencia, no se trata de distribuir recursos a los alumnos en relación con sus necesidades educativas, sino de mejorar el rendimiento con respecto a los objetivos nacionales de los que responde el Gobierno, en la medida en que los ministros sean considerados culpables si no se alcanzan aquellos. Por tanto, el Gobierno paga para obtener mejores resultados con el fin de poder declarar que ha elevado los niveles.

La cuestión ética que se plantea aquí es que estos recursos se destinan a quienes están en el límite de poder alcanzarlos; los niños que probablemente no alcancen el nivel clave quedan fuera de este proceso, como también quienes aprueben cómodamente. En este contexto, el Gobierno ha destinado considerables recursos a estas clases de refuerzo. Antes de los exámenes de 2005, se remitió un "paquete de refuerzo, de 328 páginas, que explicaba exactamente cómo debían configurar su preparación para maximizar sus resultados. Tres cuartas partes de ese paquete estaba constituido por planes detallados de clases para alumnos "límite" de nivel 4/5, que facilitaba respuestas modelo y daba consejos acerca de cómo leer por encima las preguntas. Una de las respuestas modelo era sobre la escena 1 del acto I y la escena 3 del acto II del texto de SHAKESPEARE *Mucho ruido y pocas nueces*. Se entregaron a los alumnos cuatro páginas de respuestas a la pregunta: "¿Cómo es la idea del amor examinada en estos extractos?" La pregunta de examen aparecida en la prueba nacional de 2005 fue: "¿Qué descubrimos en estos extractos acerca de la actitud de Benedick ante el amor y el matrimonio?"[19]

Esto tiene más de cínico que de educativo. Un tema central de este libro es si la evaluación debilita o promueve la enseñanza y el aprendizaje eficaces. Lo preocupante de estas formas de entrenamiento es que el objetivo es obtener un resultado determinado en vez de enriquecer el aprendizaje. Estamos volviendo a lo denunciado por DORE, que se ajusta perfectamente a esta situación.

Trampas

Para KORETZ y sus colaboradores, esta es una clara respuesta negativa. Implica manipulaciones que facilitan a los estudiantes unos rendimientos falsos. Sucede cuando los docentes señalan, facilitan o manipulan las respuestas. Esto puede ir desde elevar la ceja hasta comprobar y arreglar los exámenes después de que los hayan entregado. El *Times Educational Supplement* (1 de septiembre de 2006) informaba de que, en Inglaterra, 248 maestros fueron sometidos a investigación por "ayudar o tutelar en exceso" en las pruebas nacionales entre 2002 y

[19] Fuente: Warwick MANSELL: "Test Tips Equal Three Hundred Pages of Pressure", *Times Educational Supplement*, 1 de julio de 2005, pág. 6.

2005, habiendo sido encarcelado, al menos, un director escolar por cambiar las respuestas.

Aprovecharse del sistema

La auténtica preocupación por las trampas no se refiere tanto a estas "ayudas" aisladas como al terreno resbaladizo de *aprovecharse del sistema y hacer trampas con ello*, situaciones en las que los estudiantes no se presentan a los exámenes para mejorar los porcentajes o se presentan "estratégicamente" con el fin de elevar los resultados. David HURSH ha resumido la creciente evidencia estadounidense de que algunos de los éxitos reivindicados por los regímenes de tests de rendición de cuentas no eran todo lo que sus proponentes decían. Por ejemplo, según HURSH, "El milagro de Texas" se debía en no pequeña parte a que los administradores "reclasificaban" creativamente a los estudiantes cuando abandonaban la escolaridad o los colocaban ("repetición") en el curso inmediatamente inferior al que se sometía a las pruebas.

Abandonos

Según HURSH, fue Rodney PAIGE, el entonces superintendente del *Houston School District* quien, ante el descrédito del distrito por las elevadas proporciones de abandonos, ordenó a los directores escolares que cambiaran las explicaciones de los motivos por los que los estudiantes dejaban la escuela (por ejemplo, "deja la escuela por cambio de centro"). Esto condujo a una mejora masiva de la tasa de abandonos, hasta un 1,5% en el curso 2001-2002, y a que Houston ganara un premio como distrito escolar destacado. PAIGE llegó a ser Secretario de Educación del ex gobernador de Texas George W. Bush. No obstante, una investigación estatal posterior de 16 *high schools* puso de manifiesto que, de los 5.000 estudiantes que dejaron sus centros, el 60% tenían que haber sido clasificados como abandonos, pero no lo fueron. Un director manifestó su sorpresa cuando comprobó que se acreditaba que en su centro no habían existido abandonos, cuando la matrícula inicial de 1.000 alumnos se había reducido a 300 al final del curso y que "casi todos los estudiantes que estaban siendo expulsados estaban en situación de riesgo y pertenecían a minorías" (HURSH, 2005, pág. 615).

Repetición de curso

Otra estrategia consistía en mantener como repetidores a los estudiantes en 9.º grado, el inmediato inferior al que se sometía a la *Texas Assessment of Academic Skills* (TAAS). Walter HANEY ha calculado que, en el curso 1996-1997, el 18% de todos los estudiantes estaban repitiendo 9.º y esto incluía a la cuarta parte de los estudiantes afronorteamericanos e hispanos. Solo el 58% de los estudiantes afronorteamericanos y el 52% de los hispanos estaban en 12.º, el curso de la graduación, cuatro años más tarde. Otro movimiento estratégico consistía

en clasificar a más estudiantes como alumnos con necesidades educativas especiales, de manera que, aunque todavía tuvieran que someterse al TAAS, sus resultados no se contabilizaran. En los cuatro primeros años del TAAS, el porcentaje de alumnos de educación especial ascendió en Texas del 4,5% al 7,1%. Esta estrategia cínica puede ayudar a la escuela, pero, en la medida en que la evaluación origine lo que seamos, parece un paso inhabilitador para las personas implicadas. Para HANEY, el milagro de Texas debe considerarse, más bien, como el "espejismo de Texas".

Acceso a otro nivel educativo

Los exámenes del GCSE en Inglaterra constituyen otro ejemplo de estrategias poco definidas. El objetivo de rendición de cuentas de este examen por asignaturas que hacen los alumnos al final de la escolaridad obligatoria, consiste en aprobar cinco materias con calificaciones entre A* y C (cualquier calificación entre A* y G representa un aprobado, pero, en la práctica, y en parte gracias a los objetivos, las calificaciones inferiores a C se consideran suspenso). Se elaboran unas tablas de rendimiento que permiten la clasificación de las escuelas según el porcentaje de alumnos de 16 años que obtienen calificaciones entre A* y C. Esto ha conducido a un continuo juego regulador del ratón y el gato, pues las escuelas intentaban encontrar cómo presentar a los exámenes a alumnos que les ayudasen a alcanzar sus objetivos, y el Gobierno ha procurado cerrar el paso a los fraudes.

Un ejemplo instructivo de las formas de aprovecharse del sistema surgió con el vacío legal creado cuando, con el fin de establecer la paridad de titulaciones académicas y profesionales, se equiparó el aprobado en el *Intermediate General National Vocational Qualification* (GNVQ) a cuatro calificaciones C del GCSE. Se pretendía que el *Intermediate GNVQ* fuese un curso a tiempo completo para estudiantes que hubiesen terminado la escolaridad obligatoria, es decir, por regla general, alumnas y alumnos de 17 años que asistirían a centros de educación postsecundaria. Facilitaba una amplia introducción al ámbito profesional en áreas como el mundo empresarial, la salud y la asistencia social. Sin embargo, una escuela especializada en tecnología empresarial reconoció el potencial de este GNVQ para maximizar sus resultados. Todos sus alumnos de GCSE fueron matriculados para el GNVQ en Tecnología de la Información y la Comunicación, lo que no significaba más que otra materia del GCSE que había que aprobar para que un estudiante obtuviera las cinco calificaciones requeridas entre A* y C. La *Thomas Telford School* se convirtió entonces en el centro de máximo rendimiento del país (100% de calificaciones entre A* y C) y, en su calidad de *comprehensive school**, fue agasajada y visitada por el Primer Ministro y altos funcionarios de educación. La historia no se detiene aquí: la escuela comercializó su paquete de *software* a otras escuelas que adoptaran el GNVQ en Tecnologías de la Información y la Comunicación. Llegó a hacerse tan popular que uno de los organismos de certificación educativa adoptó este plan y, gracias a ello, la escuela tiene unos ingresos de millones de libras. En la actualidad, es una escuela que, para

* Instituto inglés de secundaria que admite alumnos de cualquier nivel de aptitud. (*N. del T.*)

visitarla, hay que pagar entrada. Evidentemente, los padres quieren enviar a sus hijos a ella, por lo que su rendimiento académico general sigue siendo alto.

A medida que un mayor número de escuelas, sobre todo las desfavorecidas, fueron adoptando esta estrategia y empezaron a aparecer en las listas de "las que más han mejorado", el juego regulador del ratón y el gato comenzó de nuevo. Los críticos indicaban que estos buenos resultados podían lograrse sin aprobar las materias básicas de Lenguaje y Matemáticas. Por eso, desde 2006, entre las cinco calificaciones entre A* y C tienen que estar las de Lenguaje y Matemáticas que, en aquel momento, aprobaban menos de la mitad de los candidatos.

Aunque este sea un relato anecdótico, ilustra cómo las lagunas normativas relativas a la maximización de resultados pueden tener una cualidad sistémica y con repercusiones negativas para otros. Si las escuelas y el Gobierno, con sus objetivos nacionales, necesitan unos resultados de exámenes constantemente mejorados, hay que prever que surjan distorsiones del sistema. Creo que la razón por la que mi propia respuesta en este caso es de regocijo y no de indignación se debe a que, anteriormente, trabajé en el desarrollo del GNVQ y creo que tenía muchos méritos, como que, "accidentalmente", los estudiantes pudieran disfrutar de experiencias de aprendizaje potencialmente interesantes.

Resultados inflacionarios

La rendición de cuentas mediante los exámenes para nota implica un auténtico interés de los centros, los docentes, las administraciones educativas locales y los políticos, por mejorar los resultados de las pruebas que se utilizan, a menudo de manera bastante simplista, para medir la mejora. Hemos examinado algunas de estas respuestas y sus consecuencias. La siguiente cuestión es si esta rendición de cuentas basada en los exámenes mejora de verdad el aprendizaje que miden. La importancia de esta cuestión radica en la disposición de los políticos y los responsables de la educación, al menos de cara al público, para tomar en serio los resultados de los exámenes: si ascienden, los niveles se habrán elevado[20]. Esto se debe a que se ven atrapados por su propia lógica: la rendición de cuentas se basa en los resultados de los exámenes, por lo que éstos deben representar directamente los niveles. La respuesta —que los resultados representan directamente cómo se desenvuelven los estudiantes en el examen, pero no necesariamente los niveles subyacentes en esa materia— suele rechazarse de plano. No obstante, la examinaremos ahora con cierto detalle para ver cómo

[20] En Inglaterra, se produce una situación paradójica cuando las puntuaciones en los exámenes del currículum nacional se tratan como si representaran directamente los niveles (= resultados mejorados), mientras que los exámenes no lo son. El Gobierno (a través de la *Qualifications and Curriculum Authority*) es directamente responsable de estos exámenes. Sin embargo, la responsabilidad de los exámenes para el GCSE y para el GCE recae sobre organismos certificadores independientes y en mutua competencia. Un incremento de las proporciones de aprobados puede recibirse como una mejora, pero también es probable que se criticase a los organismos certificadores por bajar los niveles (= exigencia del examen), tratando de mejorar su cuota de mercado. En EE.UU., se observan preocupaciones similares con respecto a la calidad relativa de los exámenes estatales.

es posible que los exámenes de importancia decisiva puedan acabar siendo no representativos de lo que en realidad se esté aprendiendo.

El efecto Lake Wobegon

En la población ficticia Lake Wobegon, de Garrison KEILLOR, "los hombres son guapos, las mujeres son fuertes y todos los niños están por encima de la media". Esta mítica comunidad radiofónica ha dado nombre al *efecto Lake Wobegon*, que fue identificado por John CANNELL, un médico de West Virginia. En 1987, demostró que todos los Estados de los EE.UU. sostenían que sus estudiantes estaban por encima de la media en las pruebas estandarizadas. La inferencia pública de esos resultados de los tests era que los estudiantes de cada Estado se desenvolvían mejor que otros, por lo que su Estado lo hacía mejor que otros, algo estadísticamente imposible. Lo que ocurría, en realidad, era que los estudiantes hacían mejor las pruebas que los estudiantes que se habían utilizado para estandarizarlos cuando se elaboraron. Como señalaron Walter HANEY y sus colaboradores en una revisión de 1993, esto no puede sorprender a nadie, puesto que los estudiantes originales no tenían porqué haberse preparado para la prueba, tanto en relación con lo que estudiaran como con respecto al formato de la misma. Cuando CANNELL se hizo pasar por un superintendente escolar y visitó una organización examinadora, le indicaron que, si quería que las puntuaciones de su pobre distrito rural fueran superiores a la media, debía utilizar uno de los tests antiguos de la compañía y sus puntuaciones se elevarían año tras año[21].

Inflación de puntuaciones

Robert LINN (2000) ha señalado otro efecto de los movimientos estadounidenses a favor de los tests de rendición de cuentas de importancia decisiva durante los últimos 50 años: *invariablemente, las puntuaciones ascienden en los primeros cuatro años, más o menos, de implementación de un test y después se estabilizan*. Dan KORETZ y sus colaboradores (1991) demostraron también que si, después de este período de cuatro años de mejora de las puntuaciones, se administraba a los estudiantes el test anterior, al que había reemplazado el nuevo, sus puntuaciones caían al nivel en el que estaban en el primer año del nuevo test. En consecuencia, estas mejoras dependen esencialmente de que se haga mejor un test concreto.

Además, está la evidencia de que, mientras que las puntuaciones en las pruebas de importancia decisiva mejoran espectacularmente, la evidencia de las evaluaciones paralelas sin trascendencia para las calificaciones no acusa aquella mejora. En Inglaterra, Peter TYMMS ha demostrado que las puntuaciones en las pruebas nacionales de alumnas y alumnos de 11 años ascendieron espectacularmente entre 1995 y 2000, estabilizándose a continuación. Sin embargo, las puntuaciones en otras pruebas paralelas sin incidencia en las calificaciones

[21] HANEY y cols. (1993); WILDE (2002).

durante ese período habían mejorado mucho menos. Indicó que los niveles subyacentes solo habían mejorado modestamente y que algunas de las mejoras eran el resultado del entrenamiento para las pruebas que habían recibido los alumnos. El Gobierno rechazó con dureza las afirmaciones de TYMMS, aunque la *Independent Statistics Commission* revisó las pruebas y se manifestó, en general, a favor de TYMMS[22].

Interpretación de la inflación de puntuaciones

La inflación de puntuaciones es un fenómeno sólido en las pruebas relevantes para la rendición de cuentas. Supone una amenaza a la validez de esas pruebas, porque conduce a inferencias cuestionables que puedan extraerse de los resultados. Una figura clave de este cambio del modo de entender las cosas fue Samuel MESSICK, que señaló dos amenazas importantes para la validez: *infrarrepresentación del constructo* e *irrelevancia del constructo*, que utilizaré para organizar esta "indagación de la validez" sobre la inflación de puntuaciones en los tests.

Infrarrepresentación del constructo

En el centro de esta amenaza está la preocupación por el grado en que la prueba presenta una muestra adecuada de los conocimientos y destrezas que pretende medir. Las pruebas para nota se circunscriben con frecuencia a lo que puede medirse con facilidad y fiabilidad. Por ejemplo, las pruebas de Lenguaje del currículum nacional se centran únicamente en la lectura y la escritura, aunque el currículum nacional tiene una línea de trabajo dedicada a hablar y escuchar que, por tanto, forma parte del "constructo" de Lenguaje. Cuando se otorga una ponderación excesiva de este tipo a determinados elementos, la enseñanza se inclinará hacia esos elementos, distorsionando así el constructo. A su vez, esto llevará a inferencias incorrectas; subirán las puntuaciones de lectura y escritura, infiriéndose que se han elevado en Lenguaje, que es como se denomina la prueba, aunque es posible que las competencias de hablar y escuchar hayan descendido por haber sido marginadas.

Irrelevancia del constructo

Esta amenaza abarca las distintas formas en que las puntuaciones más altas en las pruebas pueden ser el resultado de fenómenos ajenos a un mejor aprendizaje del constructo que se evalúa. Afecta tanto la construcción de la prueba como la fiabilidad de los resultados. Por ejemplo, pueden clasificarme como deficiente en Matemáticas cuando mi baja puntuación sea, en realidad, el resultado de mis carencias de competencias de lectura: las preguntas eran demasiado difíciles de

[22] Statistics Commission Report No. 23, febrero de 2005.

leer. De modo parecido, es posible que gane puntos por razones diferentes del conocimiento del tema o del dominio de la destreza. Stephen GORDON y Marianne REESE ponen un ejemplo de esta posibilidad tomado de su investigación acerca de cómo preparan los docentes y los alumnos la *Texas Assessment of Academic Skills*. Descubrieron que la enseñanza directa para aprobar el examen puede ser muy eficaz, tanto que se podrían aprobar los exámenes

> aunque los estudiantes nunca hubiesen aprendido los conceptos de los que se examinaran. Cuando los profesores llegan a dominar este procedimiento, pueden incluso enseñarles a responder correctamente preguntas de examen *pensadas* para medir su capacidad de aplicar, analizar o sintetizar, aunque no hayan adquirido las competencias de aplicación, análisis o síntesis.
>
> (1997, pág. 364.)

Es este un problema grave en los exámenes oficiales para calificación de los alumnos, pues gran parte de la preparación consistirá en descubrir pistas y practicar la forma de responder. Es un malabarismo difícil, dado que *no* preparar a los estudiantes aumenta también esta amenaza, pues pueden no estar acostumbrados a las convenciones de los exámenes ("explica tu respuesta con ejemplos") ni al formato de los mismos y, por tanto, pueden perder puntos por no entender lo que tengan que hacer. En consecuencia, la preparación es válida hasta que reemplaza los conocimientos y destrezas que deben adquirir mediante técnicas de realización de exámenes.

Fiabilidad

En la imaginación pública, la fiabilidad se refiere esencialmente a la *corrección de pruebas*, razón por la que los tests de opciones múltiples corregidos mecánicamente se consideran automáticamente fiables. *El simple hecho de que los tests se corrijan externamente no los hace fiables*, pues hay otros muchos factores que pueden confabularse para reducir la fiabilidad de una prueba, por ejemplo: la forma de administrar la prueba, los niveles de ansiedad y de motivación de los examinandos y la proporción de puntos otorgados a determinados elementos. La forma de sumar los puntos y dónde se sitúen los límites de cada puntuación también influyen en las calificaciones. Los cambios en cualquiera de estos factores supondrán que un mismo examinando obtenga un resultado diferente si repite la prueba o si unos estudiantes idénticos la hacen al mismo tiempo. Por eso, es muy posible encontrarse con un test corregido mecánicamente que sea muy poco fiable.

La ponderación de los ítemes de un test es también una cuestión clave para la fiabilidad y un recordatorio de que la fiabilidad *es un aspecto de la validez* y no un concepto completamente independiente. Hemos visto que la infrarrepresentación del constructo es el resultado de un muestreo desequilibrado del mismo. Si un elemento está representado en la especificación de la prueba pero solo aparece débilmente contemplado en los ítemes de la misma, esta será poco fiable: la evidencia para hacer inferencias acerca de ese elemento puede ser insuficiente. Esto puede darse cuando es más fácil redactar preguntas relativas a unos ele-

mentos que a otros, por lo que las comprobaciones de calidad de las "buenas preguntas" pueden sesgar el equilibrio de la prueba. Esto se complica si a los estudiantes les resulta especialmente difícil un elemento y solo obtienen puntuaciones bajas (por ejemplo, la probabilidad en Matemáticas), de manera que las pruebas a partir de las cuales generalizar acerca de este ejercicio concreto sean muy limitadas. No obstante, el usuario puede no percatarse de ello, infiriendo, de un modo poco fiable, que una buena calificación significa que se ha comprendido el concepto de probabilidad.

La baja fiabilidad de las pruebas

A menudo, no se tiene muy en cuenta ni la escala ni los efectos de esa escasa fiabilidad. Dylan WILIAM (2001) ha calculado que los niveles de fiabilidad de los exámenes nacionales para las niñas y niños de 11 años en Inglaterra implican que alrededor del 30% de los alumnos pueden ser clasificados erróneamente, adjudicándoseles un nivel superior o inferior al correspondiente a sus logros[23]. Estas inexactitudes pueden compensarse a nivel de centro, donde las desviaciones hacia arriba de unos con respecto a su nivel real neutralizan las desviaciones de otros hacia abajo. Sin embargo, esto traslada el problema a las medidas posteriores "de valor añadido", orientadas a la rendición de cuentas, dado que es posible que, en la siguiente etapa clave, se considere que no progresan adecuadamente los alumnos a los que se haya adjudicado erróneamente un nivel más elevado del que les corresponde. Un curioso ejemplo de esta situación surgió en los exámenes nacionales de 2006, en los que algunos profesores manifestaron que las pruebas eran poco fiables porque sus alumnos "lo habían hecho demasiado bien" en relación con las evaluaciones más detalladas, llevadas a cabo por el profesor, añadiendo que "a los profesores de los centros de secundaria no les valía la pena que tuvieran que demostrar progresos con estos alumnos cuando pasaran a aquel nivel educativo" (*TES*, 7 de julio de 2006, pág. 2). Esto condujo a un raro reconocimiento del ministro de las escuelas: "Acepto que ninguna prueba es fiable al 100%. Es bien sabido que algunos alumnos tendrán una actuación mejor o peor que su auténtico nivel de rendimiento en un día concreto" (pág. 2). Obsérvese que la fiabilidad se convierte en un problema de los alumnos más que de la prueba y sus correctores, dado que la confianza pública en las pruebas no puede debilitarse.

La vida media de las pruebas utilizadas para rendir cuentas

La señal que emiten estas consecuencias negativas y la inflación de puntuaciones en los exámenes de calificación y de rendición de cuentas es que construir sistemas punitivos de rendición de cuentas sobre una base tan frágil como las puntuaciones en los exámenes acaba pronto siendo contraproducente. Esos indi-

[23] WILIAM (2001); véase también: BLACK y WILIAM (2006): "The Reliability of Assessments", en: J. GARDNER (ed.): *Assessment and Learning*. Londres: Sage, págs. 119-132.

cadores pueden encerrar cierto valor de choque a corto plazo pero, como hemos visto, esto pronto da paso a aprovecharse del sistema. Llamo a esto el *principio de la "vida media"* —una analogía del fenómeno de la degradación de los isótopos radiactivos—, en el que la medida es el tiempo que transcurre hasta que su potencia se reduce a la mitad de la original. Este principio establece que *la utilización de las pruebas destinadas a calificar a los alumnos a efectos de rendición de cuentas puede tener cierto valor a corto plazo, al centrarse en una limitación aceptada del sistema, pero tiene una breve "vida media" durante la cual se degrada y pierde su potencia*. Esto puede apreciarse en la distorsión del currículum escolar y en el debilitamiento de la integridad de las materias que se someten a prueba. Si queremos asignar un número a esta vida media, los cuatro años de LINN quizá no sean un mal punto de partida. Para entonces, su capacidad se ha aprovechado al máximo; los docentes se han acostumbrado al test, y los estudiantes saben lo que tienen que hacer gracias a exámenes anteriores, por lo que el aprendizaje se reduce cada vez más a cómo maximizar la puntuación.

Un ejemplo de "vida media" lo tenemos en los tests de Ciencias de primaria en Inglaterra. Me gustaría haber utilizado esto como un ejemplo positivo de evaluación nacional que hubiese llevado a una mejor enseñanza de las Ciencias. Sin embargo, se están suscitando interrogantes acerca de si los tests han sido contraproducentes. Michael SHAYER, un destacado educador de Ciencias, ha realizado una investigación longitudinal sobre el desarrollo de los conceptos científicos de los niños, siguiendo líneas de investigación piagetianas. Tras varios años de mejoría continuada, observa la aparición de signos recientes de un descenso del razonamiento conceptual de los niños pequeños. Considera que una posible causa de ello es la creciente carencia de juegos científicos "prácticos", el tipo de juego que ayuda a desarrollar conceptos como los de conservación de la masa y del volumen. ¿Por qué está ocurriendo esto? Porque las pruebas no exigen una experiencia práctica de ese tipo. Por tanto, las pruebas pueden haber llegado a ser contraproducentes.

La rendición de cuentas ni va a desaparecer ni debe desaparecer, por lo que hay que encontrar formas que no favorezcan las distorsiones provocadas por unos indicadores tan simplistas. Examinaremos algunas posibilidades de lo que Onora O'NEILL llamó "rendición de cuentas inteligente" en sus *Reith Lectures* de 2002.

Rendición de cuentas inteligente

La rendición de cuentas inteligente tiene dos características fundamentales. La primera es que tiene que ser más constructiva que los enfoques actuales; la segunda es que las medidas utilizadas tienen que ser más sofisticadas. O'NEILL extrajo gran parte del material de sus conferencias de los regímenes de rendición de cuentas del servicio público del Reino Unido. Examinaré aquí cómo puede resultar esto en relación con la educación.

Rendición de cuentas constructiva

Más confianza

En el análisis de O'NEILL, la actual cultura de rendición de cuentas aspira a "un control administrativo aún más perfecto la vida institucional y profesional" (O'NEILL, 2002, pág. 46). La escala de éste en la educación se refleja en el catálogo de controles del Gobierno del periodista Simon Jenkins:

> Desde que tomó posesión [1997], su departamento de Educación ha publicado 500 reglamentos, 350 objetivos políticos, 175 objetivos de eficiencia, 700 notas de orientación, 17 planes y 26 líneas de becas independientes. En 2001, *Hansard* informó de una media anual de 3.840 páginas de instrucciones enviadas a las escuelas de Inglaterra... No obstante, un objetivo rige todo: los resultados de los exámenes. Bajo el laborismo, el gasto correspondiente a exámenes ha pasado de 10 millones de libras a 600 millones.
>
> (*Sunday Times*, 24 de septiembre de 2006, pág. 18.)

Así, esta *explosión de auditorías* ha pasado del examen detallado de las finanzas a todos y cada uno de los aspectos de la vida profesional. El problema es que, aunque la meta de todo esto es aumentar nuestra confianza en los servicios públicos, las consecuencias son un *descenso* de la confianza y unas prácticas profesionales cada vez más defensivas. En teoría, estos objetivos y procedimientos publicados hacen más responsables los servicios públicos ante sus usuarios. Sin embargo, O'NEILL observa que "lo que realmente se exige es la rendición de cuentas a los reguladores, a los departamentos gubernativos, a los financiadores, ante las normas legales. Las nuevas formas de rendición de cuentas imponen formas de control central, muy a menudo un conjunto de formas diferentes y mutuamente incoherentes de control central" (2002, pág. 53). Según O'NEILL, esto lleva a que se escojan los indicadores de rendimiento por su facilidad de medida y control, en vez de porque midan con precisión la calidad de la actuación. El uso de los resultados de los exámenes se ajusta perfectamente a esta descripción.

Lo que preocupa a esta autora es que esto alimente una cultura de desconfianza y de carácter defensivo. Linda DARLING-HAMMOND describe este enfoque por su intento de "inducir el cambio mediante recompensas y sanciones extrínsecas... sobre la base de que el problema fundamental es la falta de voluntad de cambio de los educadores" (1994, pág. 23). La alternativa de O'NEILL consiste en promover:

> Más atención al buen gobierno y menos fantasías en torno al control total. El buen gobierno solo es posible si se permite a las instituciones cierto margen de autogobierno de un modo adecuado a sus tareas específicas, dentro de un marco de información económica y de otro tipo... [que] no pueda reducirse a un conjunto de indicadores de rendimiento de reserva. Quienes sean llamados a rendir cuentas deben presentar un

> informe de lo que hayan realizado... La auténtica rendición de cuentas facilita un juicio fundamental, informado e independiente del trabajo institucional o profesional.
> (2002, pág. 58.)

Hay signos precoces de esos cambios en algunos de los nuevos procedimientos de inspección escolar que se han introducido recientemente en Inglaterra. Se ha avanzado en la dirección de que las escuelas preparen su propio "formulario de autoevaluación" (FAE), en el que revisan su propia actuación, mientras que las inspecciones externas son, en teoría, una comprobación de las propias afirmaciones de la escuela. Este es un modelo que se ha utilizado en las revisiones universitarias durante unos años. La preocupación está relacionada con la posibilidad o no de que el centro de atención pueda pasar de la actuación en el examen, como indicador clave, a una visión más rica del aprendizaje y de la enseñanza. Las posibilidades de los procedimientos del nuevo sistema de inspección consisten en que prevén que las escuelas presenten su propio informe como base del juicio; en la actualidad, el proceso está regulado en exceso y predominan las puntuaciones del examen. Es un paso hacia un punto de vista alternativo de un cambio basado en la construcción del saber con el fin de mejorar, un punto de vista que asume que el problema fundamental no es la falta de voluntad, sino "la falta de conocimiento de las posibilidades de enseñanza y aprendizaje, combinada con la falta de capacidad organizativa para el cambio" (DARLING-HAMMOND, 1994, pág. 23).

Rendición de cuentas basada en valores

Una de las consecuencias de la administración de pruebas con fines de rendición de cuentas de importancia decisiva es la que presenta lo que Michael GUNZENHAUSER ha denominado *la filosofía de la educación por omisión,* que "otorga un valor desmesurado a las puntuaciones obtenidas en pruebas para calificación, en vez de al rendimiento que se supone que representan las puntuaciones" (2003, pág. 51). A causa de la fuerza de esta filosofía por omisión, en el clima actual, los docentes "pueden percatarse de que están haciendo cosas que no están a la altura de la visión que tienen de sí mismos como educadores, como hacer que sus alumnos practiquen exámenes tipo, quitar importancia o eliminar materia que no se someta a examen o enseñar para el examen" (2003, pág. 51).

Una salida constructiva de esta situación es que escuelas y docentes articulen sus propios valores y metas. Aunque su propia filosofía de la educación comparta la misma meta de buscar que todos los niños salgan airosos, pueden optar por una vía diferente, más preocupada por el aprendizaje que por las puntuaciones. El primer paso es que las escuelas pongan *por delante la rendición interna de cuentas* y elaboren respuestas a aspectos fundamentales de la rendición de cuentas: "qué esperan de los estudiantes en el plano académico, en qué consiste una buena práctica docente, quién es responsable del aprendizaje de los alumnos y cómo dan cuenta alumnos y profesores de su trabajo y su aprendizaje" (ELMORE y FUHRMAN, 2001, pág. 69).

Este enfoque lo han desarrollado en el Reino Unido, entre otros, Barbara MACGILCHRIST y sus colaboradores, en su *The Intelligent School*, y John MACBEATH,

en su *Schools Must Speak for Themselves: The Case for School Self-Evaluation*. También destacan la importancia de que las escuelas reflexionen sobre sus valores. También se ha recogido esto en el nivel político, con una retórica en torno a "liderar en el aprendizaje", y para que las responsabilidades de la dirección escolar se financien en relación con las funciones de "enseñanza y aprendizaje". Sin embargo, en Inglaterra, los objetivos de los exámenes siguen coloreando todo esto.

Mi propia postura al respecto, que recuerda mis principios de construcción de tests del Capítulo V, es que los valores en los que debe basarse la acción se refieren a la calidad y la naturaleza social del aprendizaje. El objetivo consiste en estimular el aprendizaje "basado en principios", que requiere un enfoque activo y exigente que fomente un aprendizaje más profundo y flexible. El aprendizaje tiene que hacerse más gratificante en sí mismo y no solo a través de las calificaciones obtenidas, por importantes que éstas sean.

Medidas más sofisticadas

Como hemos visto, la obsesión con los objetivos basados en las puntuaciones de los exámenes invita a aprovecharse del sistema y conduce a que se preste más atención a los resultados que a lo aprendido. La rendición inteligente de cuentas implica establecer objetivos realistas que se basen en pruebas de lo que es posible y múltiples medios de evaluación que ofrezcan un enfoque más válido de la medida del progreso. La rendición inteligente de cuentas reconoce también que el cambio lleva tiempo y necesita una evaluación eficaz de las medidas y sus consecuencias incluida en el sistema. Examinaremos ahora estas características.

Establecer objetivos realistas

El artículo citado de Simon Jenkins tenía este atractivo título: *"Fija un objetivo estúpido y conseguirás un servicio público verdaderamente disparatado"*, que encierra una verdad como un templo. Tanto en los EE.UU. como en Inglaterra, los objetivos que predominan en la enseñanza reflejan aspiraciones, en vez de derivarse empíricamente, son listas de deseos políticos promulgados mediante un proceso de reglamentación. *No Child Left Behind* tiene, en realidad, un objetivo bastante irreal: que *cada* niño que resida en los Estados Unidos haya alcanzado el nivel básico de rendimiento en 2014.

Robert Linn (2005) ha calculado que, para lograr esto en el *National Assessment of Educational Progress* (NAEP), la evaluación de 4.º grado de Matemáticas tendría que alcanzar una tasa de mejora *anual* 3,9 veces más rápida que la tasa total de incremento entre 1996 y 2003. En Matemáticas de 8.º grado, ésta tendría que ser 7,5 veces más rápida. Lo que ofrece a esto un aire aún más pronunciado de Alicia en el País de las Maravillas es que los objetivos de "progreso anual medio" (PAM) se basan en lo que debe hacerse para alcanzar este objetivo imposible (algunos Estados han determinado unos incrementos más modestos para los cursos inmediatamente siguientes y otros de carácter milagroso para los cursos inmediatamente anteriores a 2014). Así, como hemos visto, una escuela pue-

de obtener buenos progresos curso tras curso sin que, no obstante, sus alumnos de menor rendimiento dejen de mostrar un PAM insatisfactorio, mientras que una escuela privilegiada, con una elevada proporción de estudiantes que superen el nivel básico, puede progresar menos curso tras curso y, sin embargo, satisfacer los objetivos de PAM.

En Inglaterra, las pretensiones se fijan en lograr que el 85% de los alumnos de 11 años alcancen el nivel 4 en Lenguaje y Matemáticas, un objetivo que debería haberse alcanzado en 2004 y que todavía no se ha logrado porque las puntuaciones se han estabilizado. Estos porcentajes fueron el producto de la ambición política más que de la evidencia. Todos los años se crea una situación de tensión cuando se pide a las escuelas que fijen sus objetivos, que se basan en lo que creen que serán capaces de conseguir los alumnos de los cursos 2.º, 6.º y 9.º en ese año, mientras que a la administración educativa local se le fijan unos objetivos generales, meras aspiraciones sin base real, que tienen que negociar con las escuelas con el fin de equilibrar los libros ("¿podéis elevar el objetivo en torno a un 79%?"): una frustración en toda regla, cuando eso significa que hay que prescindir de los objetivos más realistas de los centros.

Visto lo que antecede, ¿cómo podrían establecerse unos objetivos realistas? Robert LINN ha propuesto un modelo basado en el principio de que "los objetivos de rendimiento deben ser ambiciosos, aunque también realistas y alcanzables con suficiente esfuerzo" (2005, pág. 3). Reclama una *prueba de existencia*, una evidencia de que el objetivo no exceda lo que se haya conseguido en las escuelas de mayor rendimiento; por ejemplo, si estas mejoraron un 3% anual durante los últimos cursos, ese podría ser un objetivo estatal realista.

Este es el tipo de enfoque que yo llamo *empírico*, en contraste con el de mera aspiración, y está en el centro de los objetivos realistas. El establecimiento de objetivos empíricos se basa en la situación en la que estamos ahora y lo que sabemos acerca de las tasas de progreso. Así, nuestras proyecciones suponen tomar una línea base, por ejemplo, el rendimiento en 2000; comprobar el progreso anual sobre ésta, y presentar unos objetivos exigentes pero alcanzables. En un sistema de rendición inteligente de cuentas, contaríamos con la inflación de puntuaciones en los primeros años si se utilizaran pruebas de calificación a efectos de rendición de cuentas (no tendría porqué ser así necesariamente). Eso se debe a que podríamos esperar unos progresos iniciales espectaculares, con una inflación de puntuaciones muy por encima del 3% de LINN, que iría reduciéndose. Si las puntuaciones son bajas, hay que evaluar qué ocurre.

Un sistema así reconocería también a quienes necesitasen hacer los mayores progresos; es probable que los alumnos que menos rindan sean quienes progresen más lentamente, cuando el progreso se mide en términos absolutos, no relativos. Por tanto, el hecho de que un grupo quede muy por detrás no significa que podamos fijar para él unas tasas de mejora aún más altas a menos que haya base para hacerlo[24]. Las intervenciones que han demostrado que mejoran el

[24] La premisa política es que los aprendices más lentos tendrán que aprender más deprisa que los demás en las nuevas iniciativas. El documento gubernativo de consulta en Inglaterra: *Making Good Progress* (2007), propone una combinación de incentivos económicos para las escuelas, exámenes bianuales regulares y orientación adicional, con el fin de conseguir que los alumnos que se

aprendizaje (por ejemplo, el programa *Reading Recovery**) permiten hacerlo, pero un currículum "básico" con ejercicios y tests, no.

Las mejoras requieren tiempo

El problema que plantean los objetivos fundamentados en meras aspiraciones es que la limitada mejora de puntuaciones curso tras curso induce una histeria política[25], que trata de tocar aun más palancas políticas en un esfuerzo desesperado de impulsar las puntuaciones hacia arriba. La manifestación más reciente de esto en Inglaterra es la imposición del Gobierno de un cambio en la forma de enseñar a leer en los primeros años. Algunos funcionarios consideran que el paso del método fonético analítico al fonético sintético es como una panacea que transformará las puntuaciones en lectura. Aunque, con el tiempo, esto pueda ayudar un poco, los políticos buscan resultados inmediatos y espectaculares.

Mi propia experiencia proviene de evaluar la *Key Stage 3 Strategy for 11-14-year-olds*** en Inglaterra, tras el "éxito" de las estrategias de Lectoescritura y Aritmética en las escuelas de primaria. En el primer curso de la experiencia piloto, las escuelas se centraron en 7.º curso, pero esto no impidió que el equipo encargado de la estrategia no se sintiera muy ansioso por los resultados de la prueba de 9.º de ese año. Cuando se cuestionó que pudiera esperarse alguna mejora, la respuesta de los responsables de la política fue que, para la financiación posterior, los ministros esperarían mejoras inmediatas (después de todo, era una estrategia 11-14). La premisa política era que "alcanzaría" la enseñanza y el aprendizaje en 9.º Esto no es una rendición inteligente de cuentas.

Uno de los mensajes clave de Michael FULLAN, el personaje canadiense internacionalmente famoso por sus trabajos para implementar una reforma educativa a gran escala, es que el cambio implica ganarse "los corazones y las mentes", es decir, la implementación de cambios efectivos requiere tiempo y un compromiso sostenido aún más largo. Louise STOLL y sus colaboradores (2003) hacen hincapié en ello en su *It's About Learning (and It's About Time)*.

Medidas múltiples

Utilizar una única medida es invitar a que surjan problemas, tanto en cuanto al modo de distorsionar el sistema como en cuanto a las consecuencias de su limitada validez. El problema es que, aunque esto sea muy conocido y se recomienden varias medidas, los titulares se basan en un único indicador. Como ya

---

encuentren por debajo de los niveles esperados hagan rápidos progresos. Se están probando como *Single Level Tests** (http://www.qca.org.uk). De este modo, pueden conseguirse los objetivos del Gobierno.

* "Exámenes de nivel único". (*N. del T.*)

[25] Expresión acuñada por STRONACH y MORRIS (1994).

* *Reading Recovery* es un programa de intervención precoz, diseñado por Marie M. CLAY, para apoyar a los niños de primer grado que tienen dificultades para aprender a leer y escribir. (*N. del T.*)

** "Estrategia de la 3.ª Etapa Clave, para alumnos de 11 a 14 años". (*N. del T.*)

hemos visto, en educación, a menudo serán las mejores puntuaciones en los exámenes. Las que no aparecen son otras medidas de la calidad de la escolarización, como la evaluación del profesorado, la satisfacción del alumnado, el absentismo y las medidas de progreso de "valor añadido".

La rendición inteligente de cuentas buscaría un uso más válido de estos datos: información conjunta de los juicios del docente y de las pruebas o, mejor aún, cierta reconciliación basada en el comentario de las pruebas en el contexto de la escuela (en Escocia, se utilizan tareas estandarizadas por encargo para validar la evaluación del maestro o profesor). Lo que ocurría en Inglaterra en el pasado es que, como en buena parte se hacía caso omiso de la evaluación del maestro o profesor, los docentes se limitaban a esperar los resultados del examen y otorgaban el mismo nivel al alumno. El sistema de rendición de cuentas ha debilitado la confianza del profesorado en que sus juicios puedan ser más fiables que los resultados de los exámenes. La rendición inteligente de cuentas supone la confianza en los docentes; ellos son compañeros, no el enemigo.

Ese enfoque no casa muy bien con la actual línea dura de rendición de cuentas en Inglaterra, pero los desarrollos de Gales y Escocia, donde la recogida de los datos de los exámenes centrales ha sido sustituida por enfoques locales de rendición de cuentas, muestran algunas de las posibilidades una vez eliminadas las tablas de clasificación de las escuelas.

Supervisión de los estándares o niveles nacionales

Hay una retórica política global acerca de la necesidad de elevar los niveles educativos de un país para que pueda llegar a tener peso o a conservarlo en el mercado global, una retórica que, como Alison WOLF ha demostrado en su libro: *Does Education Matter?**, simplifica en exceso la relación entre educación y creación de riqueza. Dado que los Gobiernos se consideran a sí mismos como responsables de la elevación de los niveles, cobra importancia supervisar su mejora. Hemos examinado los problemas que acarrea esto a través de los resultados de las pruebas de importancia decisiva.

Un modo mucho más constructivo de supervisar los estándares y niveles nacionales es tomar una muestra representativa de alumnos y utilizar evaluaciones de menor importancia (no se publican ni las puntuaciones individuales ni las de los centros por tratarse solo de una muestra) que tengan ítemes comunes curso tras curso. Esto reduce los efectos de preparación y hace más fiables las comparaciones entre distintos cursos. Esta es la lógica que subyace a la *National Assessment of Educational Progress* (NAEP), de Estados Unidos; al *National Education Monitoring Programme* (NEMP), de Nueva Zelanda, y a la *Scottish Survey of Achievement (*SSA) [26]. En Inglaterra, la *Assessment of Performance Unit* (APU) desarrolló una función similar hasta que se le puso fin tras la introducción de las pruebas del currículum nacional.

---

* "¿Importa la educación?". (*N. del T.*)

[26] Es un cambio de nombre reciente, que refleja la función potenciada de esta forma de supervisión y acompañada del fin de la recogida nacional de datos de tests; véase: HAYWARD (2007).

La finalidad que se esconde tras este desfile de siglas es mostrar que se trata de una metodología bien establecida y no una moda. El saber recibido por los expertos en evaluación indica que este enfoque da una idea mucho más precisa de la mejora de los niveles y constituye un enfoque mucho más constructivo de la rendición de cuentas al nivel del sistema nacional. La NAEP, un programa federal de supervisión que opera desde 1961, ha justificado plenamente su existencia por ser el antídoto del efecto Lake Wobegon. Cuando las pruebas estatales muestran unas mejoras espectaculares, estas siempre pueden verificarse con los datos estatales y nacionales de la NAEP. Por ejemplo, Robert LINN (2005) ha demostrado que, en 2003, las pruebas del Estado de Colorado indicaban que el 67% de los estudiantes de 8.º grado eran considerados muy competentes en Matemáticas, mientras que los datos de las pruebas del Estado de Misuri indicaban que solo lo eran el 21%. Sin embargo, los resultados paralelos de la NAEP indicaban que el nivel de Colorado era del 34% y el de Misuri, del 28%, por lo que las conclusiones extraídas de las pruebas de cada uno de estos Estados eran engañosas.

El muy respetado NEMP de Nueva Zelanda añade otra dimensión constructiva: utiliza a docentes en activo para que visiten escuelas como asesores de las actividades de carácter abierto y de grupo. La fortaleza de este enfoque, también utilizado en Escocia, estriba en que contribuye también a la formación profesional continua del profesorado, pues los docentes descubren cómo enfocan los problemas los alumnos y toman también conciencia de los niveles que deben alcanzar. Asimismo, lleva a los maestros y profesores a considerarse intervinientes activos en la supervisión nacional, en vez de receptores de la misma; un ejemplo claro de confianza profesional.

El mensaje para los responsables de la política educativa en Inglaterra es que esta forma de supervisión constituye una medida más sofisticada del progreso que, al basarse en muestras, presenta una relación costo/eficacia más favorable y es fuente de datos ricos acerca de lo que saben los niños y cómo piensan. Estos enfoques nos advierten también de la poca fiabilidad intrínseca de las "simples" preguntas de los exámenes. El siguiente ejemplo está sacado del cuestionario de Matemáticas de la APU de la década de 1980 para los alumnos de 11 años. Formulaba la misma pregunta de tres maneras y obtenía unos niveles muy diferentes de respuestas correctas*:

> Tres más 14 son ___ (97% de respuestas correctas).
> ¿Qué número es tres más que 14? (67% de respuestas correctas).
> ¿Qué número supera a 14 en tres unidades? (54% de respuestas correctas).

Una prueba nacional solo puede preguntar esto de una manera; en consecuencia, ¿qué nos diría respecto a cómo entienden la adición?

Al principio, me desconcertaba la reacción que los responsables de la política educativa en Inglaterra tenían cuando se les sugería que debíamos adoptar ese enfoque. Su respuesta típica ha sido: "no mencionen la APU". Como vimos con la

* Los enunciados originales de las preguntas están en inglés y es posible que induzcan más a error que los de la traducción que presentamos. (*N. del T.*)

reacción al trabajo de Peter TYMMS, esto se debe a que se dan cuenta de la amenaza que supone a sus proclamaciones Wobegon de mejoras espectaculares de los niveles. En el presente, solo podemos conectar datos, como hizo TYMMS, operación que es menos fiable que un muestreo sistemático, curso tras curso, de materiales relacionados con el currículum.

Incluir todo

En el Capítulo Primero, vimos que "liberar" a los grupos desfavorecidos de la presión de los exámenes también buscaba proteger el capital social de quienes se sometían a ellos y los utilizaban para progresar. La operación paralela en cuanto a rendición de cuentas es excluir a ciertos grupos de los objetivos, de manera que no cuenten de ninguna manera. Una de las características más positivas de *No Child Left Behind* es que incluye a todos estos subgrupos marginados. El problema está en la medida en que la evaluaciones sean diferenciadas, por ejemplo, si deben utilizarse distintas evaluaciones para grupos diferentes o si deben someterse a la misma prueba que los demás, pero con ciertas "concesiones" (por ejemplo, tiempo extra, un "secretario", traducción).

No obstante, cuando la evaluación se utiliza con fines de rendición de cuentas, yo abogaría por una mayor diversidad de medidas, de manera que tengamos unas más minuciosas que permitan representar gráficamente un progreso mucho más lento. La rendición inteligente de cuentas supervisaría el progreso con respecto a esta escala, en vez de reducirse a un informe que presentara poco o ningún progreso de acuerdo con una escala menos precisa. Esto puede hacer que la información sea más compleja, pero eso es lo que tenemos que esperar con medidas más sofisticadas. En Inglaterra, por ejemplo, la *Qualifications and Curriculum Authority* gubernativa ha elaborado las *P Scales* de ocho niveles para los estudiantes con discapacidades de aprendizaje que quedan por debajo del nivel 1 del currículum nacional. Estas escalas son más complejas, pero, en este contexto, estamos intentando evitar unas medidas "rudimentarias y simples".

Evaluar continuamente el sistema de rendición de cuentas

Si se cumplen la ley de Goodhart y el principio de la vida media, la supervisión de la influencia del sistema de rendición de cuentas y sus objetivos adquiere una importancia crítica. No es solo cuestión de si se alcanzan o no, sino que implica revisar cómo está modificando la rendición de cuentas la enseñanza y el aprendizaje y examinar cualquier consecuencia imprevista. Supone también apartarse sistemáticamente de unos objetivos limitados, con su corta vida media, y orientarse hacia unos cambios más sostenibles en el currículum, la enseñanza y el aprendizaje, que se reflejen en enfoques más complejos y cualitativos de la rendición de cuentas. Implica también la supervisión de la fiabilidad del sistema de rendición de cuentas.

Este enfoque exigiría también una respuesta más medida si no se alcanzaran los objetivos empíricos. En el contexto de la histeria política, las respuestas actua-

les tienden a introducir una nueva "palanca" para mejorar las puntuaciones que permita a los responsables políticos declarar que han arreglado el problema. Si nos mantenemos innovando, nunca tendremos que responsabilizarnos de lo que no haya funcionado; simplemente, seguiremos adelante y declararemos que hemos resuelto el problema. La rendición inteligente de cuentas implica hacer más hincapié en comprender por qué algo no funciona y menos en los cambios impulsados por el pánico.

Control del error de medida

Uno de los requisitos de los *Tests Standards** de la *American Education Research Association* (AERA) es:

> En los ambientes educativos, los informes de puntuaciones deben ir acompañados de un enunciado claro del grado de error de medida asociado con cada puntuación o nivel de clasificación e información acerca de cómo interpretar las puntuaciones.
>
> (1999, Standard 13.14, pág. 148.)

En la actualidad, los tests y exámenes nacionales del Reino Unido no cumplen estos criterios. Paul NEWTON ha examinado las consecuencias de informar del error de medida en un sistema que considera que las puntuaciones son precisas y exactas. Dada la fiabilidad conocida de un test, mi puntuación de 45 podría representar, por ejemplo, una "puntuación verdadera" (una inútil expresión estadística) entre 41 y 49. El problema es que mi 45 puede situarme en un nivel 4, pero no la puntuación 44. Por eso Dylan WILIAM (2001) estimaba que en los tests nacionales, al menos, el 30% de los alumnos están erróneamente clasificados. Aunque las clasificaciones erróneas puedan compensarse en el conjunto del sistema, no ocurre lo mismo en los niveles de clase e individual. Hace falta, pues, información acerca de cómo interpretar las puntuaciones y los "intervalos de confianza" que tenemos que observar a su alrededor.

Comprobar las consecuencias imprevistas

Si la rendición de cuentas ha de tener relación con la mejora de los niveles que, en educación, significa la mejora del aprendizaje, tenemos que controlar que ocurra precisamente eso. Es preciso desarrollar un concepto común del aprendizaje mejorado e idear indicadores que equiparen unas mejores puntuaciones en los exámenes con el aprendizaje mejorado. No se trata del juego del ratón y el gato, aprovechándose del sistema y cubriendo lagunas, sino de la calidad de la enseñanza y del aprendizaje que promuevan los indicadores. Una de las razones de que, en los EE.UU., el movimiento de los tests de mínima competencia diese paso al entusiasmo por el porfolio y la evaluación alternati-

* "Normas o criterios de los tests". (*N. del T.*)

va fue el daño que estaba haciendo a la enseñanza y al aprendizaje. Los *Tests Standards* exigen que:

> Cuando se recomiende el uso de un test o la interpretación de una puntuación sobre la base de que la administración de tests o el programa de tests producirá per se algún beneficio indirecto además de la utilidad de la información extraída de las puntuaciones mismas en el test, hay que manifestar explícitamente las razones de esas previsiones. Deben presentarse los argumentos lógicos o teóricos y la evidencia empírica de ese beneficio indirecto.
>
> (1999, Standard 13.14, pág. 23.)

Esta instrucción es farragosa, pero útil. Dado que hemos reiterado las distorsiones que provocan los objetivos restringidos que imponen las cuestiones importantes en juego, esto puede llevar a un rico diálogo.

Despejando la larga sombra

El argumento de este capítulo es que, aunque a veces sea necesario introducir evaluaciones de rendición de cuentas de importancia decisiva con el fin de dar al sistema o a las instituciones un centro de atención más claro, el proceso se distorsiona rápidamente. Para superar y reemplazar con rapidez esta forma de rendición de cuentas hace falta un tipo de rendición inteligente de cuentas que se centre en la calidad de la enseñanza y el aprendizaje y establezca unos objetivos de progreso exigentes pero realistas.

En este capítulo, hemos examinado lo que implica una *rendición inteligente de cuentas*. Supone depositar más confianza en los profesionales, que deben estar dispuestos a manifestar sus valores y objetivos. Todavía necesitaremos medidas de rendición de cuentas, pero éstas deben ser más sofisticadas. Para ello, es fundamental establecer objetivos realistas y basados en un conjunto de indicadores, en vez de una medida única. Con independencia de las evaluaciones que se utilicen, hay que controlar el error de medida y las consecuencias imprevistas. Las normas o estándares nacionales pueden supervisarse de un modo mucho más preciso, mediante muestreo y unas evaluaciones que tengan consecuencias menos relevantes como las que se utilizan en los Estados Unidos, Escocia y Nueva Zelanda. Todo esto ha de realizarse en un contexto social que tenga en cuenta que el cambio sostenible requiere tiempo y paciencia, factores que, en el presente, son más bien escasos.

CAPÍTULO VII

Razones para alegrarse: La evaluación para el aprendizaje

> Si enseñar fuese tan sencillo como contar, todos seríamos mucho más listos de lo que somos.
>
> (Mark TWAIN.)

> El estudiante sabe más que el maestro de lo que ha aprendido, aunque sepa menos de lo que se haya enseñado.
>
> (Peter ELBOW.)

Los dos capítulos anteriores han demostrado que la influencia de la evaluación sobre el aprendizaje ha sido, en el mejor de los casos, ambigua. Para el credencialismo y para la rendición de cuentas mediante los exámenes, la finalidad primordial de la evaluación es obtener resultados que, después, se equiparan con un aprendizaje mejorado. Hemos visto que con frecuencia no ocurre así: los resultados pueden mejorar sin que lo haga el aprendizaje. La evaluación *para* el aprendizaje (EpA) es un intento consciente de hacer de la evaluación un elemento productivo del proceso de aprendizaje. Se consigue haciendo de la evaluación en el aula una parte esencial de la enseñanza y el aprendizaje efectivos. En consecuencia, aborda directamente los principales temas del libro: cómo puede la evaluación configurar constructivamente el aprendizaje y nuestras identidades como aprendices.

En este capítulo, destacaré lo que implica la "evaluación para el aprendizaje", tanto en la práctica de clase como en la comprensión de nuestra forma de aprender. Aunque algunas estrategias docentes importantes son bien conocidas, preocupa la posibilidad de que los maestros y profesores puedan implementarlas sin comprender por qué pueden conducir a un aprendizaje eficaz. Esto lleva a los que considero principales problemas de la "evaluación para el aprendizaje": el tipo de aprendizaje que tiene lugar, los efectos de los objetivos explícitos del aprendizaje, la difícil relación con las evaluaciones sumativas a efectos de calificación y el conocimiento de lo que hace que la retroinformación sea eficaz. Se presta mucha atención a la retroinformación porque, en la EpA, se considera la clave para hacer avanzar el aprendizaje.

En este punto, tengo que declarar mis intereses al respecto, porque estoy implicado activamente en la "evaluación para el aprendizaje", tanto por mi larga permanencia como miembro del *Assessment Reform Group* como por mis escritos y mi trabajo profesional con maestros y profesores. No obstante, procuro mantener el mismo tipo de cuestionamiento crítico con el que he enfocado otros temas en este libro. Para quienes conozcan la evolución de la evaluación formativa, espero que mi identificación de algunas cuestiones clave les ayuden a clarificar el pensamiento y la práctica, dado que nuestra comprensión de estas cuestiones sigue estando en una fase inicial.

Perspectiva general

La mejor manera de considerar la "evaluación para el aprendizaje" es como un enfoque de la evaluación en el aula, más que como una teoría formulada con rigor. A este respecto, se acerca más a los "estilos de aprendizaje" y a la "inteligencia emocional" que a sistemas articulados de forma más completa, como las "inteligencias múltiples". Esto no significa que no existan bases teóricas, sino que, sencillamente, no se han organizado en una teoría independiente, y quizá no haga falta hacerlo. La EpA no es más que un elemento de un sistema más amplio que incluye el currículum, la cultura escolar y las formas de enseñar. Aunque las expresiones utilizadas son relativamente recientes ("formativo-formativa" se acuñó en 1967 y "evaluación para el aprendizaje", a mediados de la década de 1990), algunos de los temas clave cuentan con una historia mucho más larga[1].

Lo que distingue la "evaluación para el aprendizaje" de las "inteligencias" y de los "estilos de aprendizaje", que vimos en los Capítulos III y IV, es su énfasis en lo *situacional* —interacción en el aula— en vez de en las disposiciones del aprendiz concreto. Es una diferencia muy significativa; *centra la atención en lo que se esté aprendiendo y en la calidad de las interacciones y relaciones en el aula*. En este enfoque, la "evaluación" se interpreta en sentido amplio y se refiere a obtener pruebas relativas a la situación concreta de los aprendices y a facilitarles retroinformación que les ayude a avanzar. Estas pruebas pueden proceder de la observación (miradas desconcertadas, momentos de depresión) y de las interacciones en el aula, así como de productos más tangibles. Las pruebas desempeñan un papel si las respuestas se utilizan para identificar lo que se ha entendido y lo que no, y si esto lleva a una acción para mejorar el aprendizaje.

Uno de los argumentos centrales de este libro es que la evaluación configura nuestra forma de vernos como aprendices y como personas. Mi defensa de la "evaluación para el aprendizaje" se basa en su insistencia en el proceso de aprendizaje, más que en las capacidades y disposiciones de los aprendices. En términos de Carol DWECK, adopta un enfoque *incrementalista* del aprendizaje que resalta el esfuerzo y la *mejora* de la competencia. Éste contrasta con un enfoque

[1] Tanto en términos de teoría de la educación como de las prácticas encontradas en escuelas alternativas. En los EE.UU., encontramos la insistencia de John DEWEY (1938) en el aprendizaje activo y el énfasis de Ralph TYLER (1971) en unos objetivos claros. En Europa, por ejemplo, MONTESSORI enfatiza la autonomía del aprendiz y FREINET, la autoevaluación.

entitativo, que atribuye el aprendizaje a la capacidad y se centra en *demostrar* la competencia mediante calificaciones y comparaciones[2], y está relacionado con el énfasis que la EpA pone sobre la autorregulación y la autonomía de los aprendices con respecto a su aprendizaje, una competencia que se desarrolla mediante la autoevaluación y el diálogo en el aula.

¿Qué es la "evaluación para el aprendizaje"?

Se trata de una evaluación que está incluida en el proceso de aprendizaje. Sus cinco factores clave, "engañosamente sencillos" son:

- la participación activa de los alumnos en su aprendizaje;
- la retroinformación eficaz facilitada a los alumnos;
- la adaptación de la enseñanza para tener en cuenta los resultados de la evaluación;
- la necesidad de que los alumnos sean capaces de evaluarse a sí mismos;
- el reconocimiento de la profunda influencia que la evaluación tiene sobre la motivación y la autoestima de los alumnos, influencias cruciales ambas sobre el aprendizaje.

(*Assessment Reform Group*, 1999, págs. 4-5.)

Una definición muy utilizada que se deriva de estos factores es la del *Assessment Reform Group*:

> El proceso de búsqueda e interpretación de evidencia para uso de los aprendices y sus maestros para identificar en qué fase de su aprendizaje se encuentran los aprendices, adónde tienen que llegar y la mejor manera de alcanzar ese punto.
>
> (2002a, págs. 2-3.)

Este enfoque podría considerarse simplemente como un "buen ejercicio docente", sobre todo cuando la "evidencia" de la definición se interpreta como un amplio conjunto de informaciones más que como tests formales o informales. ¿Por qué insistir tanto en la evaluación? Creo que, en gran parte, es un paso deliberado dado por los miembros de la comunidad de la evaluación para reivindicar una de las finalidades clave, y en peligro, de la evaluación. En una época dominada por las pruebas sumativas para rendir cuentas, esto puede considerarse como un intento de reequilibrar los usos que se dan a las evaluaciones, haciéndolas *parte* del proceso de aprendizaje, en vez de que sean ajenas al mismo y sirvan de comprobación de lo aprendido. La evaluación es más que una instantánea de lo que se sabe en un momento dado.

A menudo, "evaluación para el aprendizaje" se usa como sinónimo de *evaluación formativa*[3]. En parte, la expresión "evaluación para el aprendizaje" se intro-

[2] Es la útil distinción que hace Chris WATKINS (2002).

[3] Suelo considerar la EpA como un acento especial *dentro* de la evaluación formativa. La EpA se ocupa primordialmente del aprendizaje interactivo del *estudiante*, con la pretensión de provocar cam-

dujo a causa de la gran cantidad de malas interpretaciones provocadas por "formativa". Una de las más problemáticas es la creencia de que las pruebas periódicas de clase, que se utilizan para supervisar el progreso, son formativas. Dada su finalidad, sería preferible considerarlas como evaluaciones minisumativas, pues la información recogida no se utiliza directamente para modificar la enseñanza y el aprendizaje. Lo mismo cabe decir de la corrección de los trabajos de clase que, una vez más, se presenta con frecuencia como formativa, cuando, en realidad, su finalidad consiste en facilitar pruebas para posteriores juicios sumativos. Más adelante, veremos que esta diferencia entre "formativa" y "sumativa" dista mucho de ser radical.

En su revisión de la bibliografía francesa sobre la evaluación formativa, Linda ALLAL y Lucie LOPEZ hacen unas distinciones útiles entre tres tipos de respuesta formativa (utilizan la expresión: *regulation*) para evaluar la información.

Interactivo

Se basa en las interacciones del aprendiz con los demás intervinientes en la actividad docente (por ejemplo, el maestro, otros estudiantes y los materiales de enseñanza). Esto impregna la actividad cotidiana en el aula. El resultado es la adaptación continuada del aprendizaje, sobre todo gracias a la retroinformación y la orientación. Este es, en gran medida, el centro de atención de quienes emplean la expresión "evaluación para el aprendizaje".

Retroactivo

Este tipo de respuesta es la evaluación formativa llevada a cabo *tras* una fase de enseñanza, a menudo, utilizando una prueba, y trata de atajar las dificultades de aprendizaje que se hayan identificado en ella. Este modelo de "prueba y recuperación" de la evaluación formativa sigue siendo el predominante en los EE.UU.

Proactivo o eficiente

Aquí, la evidencia conduce a cambios futuros en la enseñanza. En el contexto francés, con sus enfoques de "toda la clase", preocupa diferenciar actividades para satisfacer las distintas necesidades de los alumnos. Una interpretación más amplia consiste en que los docentes modifiquen su enseñanza *posterior* en respuesta a la evidencia de sus actuales alumnos. Por ejemplo, los resultados

---

bios docentes y curriculares que mejoren el aprendizaje del alumno. Paul BLACK y sus colaboradores hacen una distinción diferente. Esta supone considerar la EpA como una *finalidad*, mientras que la evaluación formativa es una *función*: "la evaluación se convierte en 'evaluación formativa' cuando la evidencia se utiliza realmente para adaptar el trabajo docente para que satisfaga las necesidades de aprendizaje" (BLACK y cols., 2002, pág. i).

detallados de un examen pueden llegar demasiado tarde para quienes se someten al mismo y pasan curso, pero pueden llevar a cambios relativos al qué y al cómo se enseñe al siguiente grupo. David CARLESS (2007) ha presentado recientemente el concepto de "evaluación formativa preventiva", en la que los docentes, basándose en su experiencia previa con estudiantes similares, se anticipan a las concepciones erróneas en vez de dejar que se desarrollen.

Aunque las tres formas pueden configurar el repertorio de evaluación formativa de un maestro o profesor, lo que conduce a diferencias de interpretación es su *ponderación relativa*. En este sentido, es fundamental si el centro de atención se sitúa en el aprendizaje de los docentes o en el de los estudiantes. Tanto en el enfoque retroactivo como en el proactivo, los docentes son los principales aprendices, puesto que adaptan ("autorregulan") su enseñanza. En el enfoque interactivo, el centro de atención es el aprendizaje de los estudiantes, mientras que el papel del maestro o profesor consiste en ir dejando progresivamente el control del aprendizaje en manos de los estudiantes para que ellos mismos se conviertan en aprendices autorregulados.

Esas diferencias de énfasis pueden provocar tensiones. Por ejemplo, los responsables de la política educativa en Inglaterra centran la atención en los docentes, dado que son los objetivos de la política; de ahí la insistencia en cómo deben modificar su enseñanza los profesores (un subproducto de esto es el supuesto de que los docentes deben dominar el proceso de aprendizaje). Sin embargo, esas políticas pueden inhibir los enfoques centrados en el aprendiz que formula la EpA y que, en términos de aprendizaje, están tratando de que los docentes hagan menos y los aprendices, más. Mientras tanto, los enfoques de orientación más comercial pueden considerar las "pruebas para la recuperación" como elementos fundamentales de la evaluación formativa; no es en absoluto sorprendente que comercialicen instrumentos diagnósticos de una forma o de otra[4].

En la actualidad, la Evaluación para el Aprendizaje se ha organizado suficientemente (de ahí las iniciales mayúsculas) para que sea considerado un movimiento diferenciado en el Reino Unido, que cada vez es más aceptado internacionalmente. Aunque la expresión tenga ya unos 15 años[5], su reciente popularidad está relacionada con el folleto *Inside the Black Box**, de Paul BLACK y Dylan WILIAM (1998b), del que se han vendido por todo el mundo más de 50.000 ejemplares. El folleto se basa en su revisión de 1998 de las pruebas de investigación sobre la evaluación en el aula: la "caja negra" es el aula y lo que sucede en su interior[6]. Desde entonces, han aparecido una serie de publicaciones, inclu-

[4] Por ejemplo, la herramienta de aprendizaje *Achieve*, de *Cambridge Assessment*, que ofrece "evaluaciones, información diagnóstica, establecimiento de objetivos y planificación a nivel individual y de aula basados en la pantalla... recoge y analiza los resultados y crea informes que perfilan los puntos fuertes y débiles de individuos, clases y grupos". http://www.cambridgeassessment.org.uk/ca/Our_Services/Assessment/Assessment_for_Learning (consultada el 16 de noviembre de 2007*).

* Verificado el acceso el 8 de abril de 2010. (*N. del T.*)

[5] La utilizaron independientemente Ruth SUTTON, Caroline GIPPS y Mary JAMES a principios de la década de 1990.

[6] La revisión de BLACK y WILIAM se publicó en un número especial de *Assessment in Education: Principles, Policy and Practice*, 1998, 5 (1): 7-74. El número también incluía una serie de respuestas a la revisión.

* "En el interior de la Caja Negra". (*N. del T.*)

yendo el folleto *Assessment for Learning: Beyond the Black Box**, del *Assessment Reform Group*, que también ha obtenido una amplia divulgación y conducido a la adopción generalizada de la "evaluación para el aprendizaje". La incorporación de algunas de sus ideas clave a las estrategias educativas del Gobierno del Reino Unido y de otros países pone de manifiesto la creciente influencia del enfoque[7].

¿Qué implica la "evaluación para el aprendizaje"?

Para quienes conozcan el trabajo de BLACK y WILIAM, orientado a los centros de secundaria, o el de Shirley CLARKE (1998, 2001), que ha influido en las escuelas de primaria, la evaluación formativa se identifica con determinadas prácticas docentes en clase[8]. Son las siguientes:

- *Intenciones del aprendizaje y criterios de éxito.* Ser más explícitos acerca de lo que se estudia y qué se requiere para una actuación satisfactoria, el "adónde tienen que llegar" de la definición. Muchos maestros y maestras de primaria del Reino Unido reconocerían de inmediato el personaje de cómic WALT (*We Are Learning To***), que manifiesta la intención del aprendizaje en un bocadillo del dibujo.
- *Hacer preguntas.* Una práctica consiste en *esperar un tiempo*, que requiere que los maestros y profesores dejen más tiempo a sus alumnos para que piensen, a menudo colaborativamente, en sus respuestas a preguntas orales. Esto anima a los docentes a hacer preguntas más interesantes, que revelen mejor "en qué fase de su aprendizaje están los aprendices", y exponer las concepciones erróneas. En fechas más recientes, se ha pasado a enfatizar más el carácter fundamental del diálogo, en vez de limitarse a hacer preguntas[9].

  Los semáforos sirven para este propósito, implican a los aprendices, a menudo como grupo que deben indicar si han entendido (verde), tienen dudas (ámbar/amarillo) o no han comprendido lo que se les ha explicado.
- *Retroinformación.* Se considera el mecanismo clave para contribuir a "la mejor manera de alcanzar ese punto" de la definición, dado que trata de reducir el espacio entre la fase en la que los aprendices están en el pre-

---

* "Evaluación para el aprendizaje: más allá de la Caja Negra". (*N. del T.*)

[7] En la actualidad, es una de las líneas de actuación para todo el centro de las estrategias tanto de primaria como de secundaria en las escuelas inglesas. En Escocia, es fundamental para la iniciativa *"Assessment is for Learning"** del *Scottish Education Department*. En Nueva Zelanda, forma parte de la *Teaching and Learning Strategy*, de ámbito nacional (http://www.tki.org.nz/r/assessment; consultada por última vez el 15 de noviembre de 2007**).

* "La evaluación es para aprender". (*N. del T.*)

** Verificado el acceso el 8 de abril de 2010. (*N. del T.*)

[8] Para los lectores internacionales, el trabajo de Rick STIGGINS, de la *ATI Foundation*, en los EE.UU. (STIGGINS, 2001); las publicaciones canadienses de Ruth SUTTON, Anne DAVIS y Lorna EARL, y el trabajo dirigido por el Gobierno en Nueva Zelanda comparten un fuerte parecido familiar.

* "Estamos aprendiendo (a)". (*N. del T.*)

[9] Este aspecto puede relacionarse con el trabajo más extenso de Neil MERCER (2000): "Talk Lessons", y con el de Robin ALEXANDER (2004), sobre la enseñanza dialógica.

sente y el lugar al que tienen que llegar. Como veremos, cada vez se reconoce más la complejidad de este proceso y que lo que a menudo se considera retroinformación no conduce a nuevos aprendizajes. En la práctica, se hace más hincapié en comentarios basados en la tarea que en las calificaciones o notas que, en expresión de Royce SADLER, están "demasiado profundamente codificadas" para dar información sobre lo que deba hacerse a continuación. Una de las afirmaciones más provocativas es que la retroinformación que se centra en el yo ("buen chico"; "me has decepcionado"), en vez de en la tarea, es inútil. Esto suscita cuestiones relativas al papel de la retroinformación relacionada con el "yo" (por ejemplo, el elogio) en la evaluación formativa.

- *Autoevaluación y evaluación a cargo de los compañeros.* Un objetivo clave de la EpA es progresar hacia una cultura de aula en la que los aprendices sean cada vez más capaces de juzgar la calidad de su propio trabajo y la de los ejecutados por los demás y comprender lo que implica un aprendizaje eficaz. El fundamento de ello es que, con el fin de evaluar su propio trabajo, los aprendices tienen que ser conscientes de lo que supone una actuación satisfactoria ("adónde tienen que llegar") y en qué fase están de su propio aprendizaje. Estas competencias sientan las bases de la autorregulación ("metacognición"), que se considera como una poderosa fuente de aprendizaje eficaz. Llevan a que el papel del docente consista en compartir deliberadamente con los aprendices su "saber gremial" acerca de la evaluación. Esto se haría mediante modelos de actuación: "¿Por qué es mejor este trabajo que ese otro?", y el docente ejemplifica como ofrecer retroinformación.

El atractivo del docente

En parte, la popularidad de la EpA entre los docentes se debe a que es práctica y se centra en lo que sucede en el aula: he aquí algunas técnicas que pueden probarse en clase. Algunas no suponen ningún trastorno, por ej., el tiempo de espera puede practicarse con facilidad en la intimidad del aula, aunque puede convertirse en una "bola de nieve" y conducir a cambios más profundos en la práctica[10]. En su centro están las *interacciones en clase* y el tipo de clima de aula que estimule el aprendizaje eficaz. En 2007, Terry CROOKS, un teórico fundamental de la evaluación formativa, destacaba los que él considera principales factores de una EpA eficaz. Todos ellos implicaban la calidad de la interacción en el aula y el desarrollo de una cultura de confianza en clase.

El peligro es que la idea de la "evaluación para el aprendizaje" no vaya más allá, convirtiéndose en una serie de "consejos prácticos" para la clase, en vez de un enfoque de la enseñanza y el aprendizaje respaldado por una teoría[11]. Mary JAMES y sus colaboradores, que llevaron a cabo en Inglaterra el extenso proyecto

[10] Véanse los informes de los profesores en: BLACK y cols. (2003): *Assessment for Learning*.

[11] El artículo de revisión de BLACK y WILIAM (1998a) se basaba en un análisis de unos 250 artículos de investigación relevantes.

*"Learning How to Learn"** sobre la mejora de la evaluación formativa, clasifican las respuestas de los docentes en términos del *espíritu* y de la *letra* de la evaluación formativa[12]. La letra es cuando las técnicas se aplican mecánicamente, sin comprender realmente lo que representan, mientras que el espíritu implica verlas como una expresión de una visión más amplia del aprendizaje que, a su vez, puede modificar las técnicas. Examinaremos a continuación esta base teórica.

Fundamentos teóricos

Dado que la "evaluación para el aprendizaje" ha sido impulsada en gran parte por universitarios, resulta un tanto sorprendente que los fundamentos teóricos hayan quedado en segundo plano. El centro de atención ha sido en gran medida pragmático, basado en lo que revela la investigación acerca de las prácticas eficaces de evaluación en el aula. Se ha producido un reconocimiento creciente de que la EpA ha partido de unos *supuestos* acerca de cómo aprendemos, sin hacerlos explícitos ni relacionarlos con una determinada teoría del aprendizaje. Esta estrategia puede defenderse si las prácticas se acomodan con facilidad a diversos enfoques; sin embargo, se corre el riesgo de que la evaluación formativa se trate como una práctica ateórica que consista simplemente en una serie de técnicas de enseñanza y de evaluación. Harry TORRANCE y John PRYOR sostienen que:

> un intento de comprender la evaluación formativa debe implicar una combinación y coordinación críticas de ideas derivadas de una serie de puntos de vista psicológicos y sociológicos, ninguno de los cuales, por sí mismo, constituye una base suficiente para el análisis.
>
> (1998, pág. 105.)

Se ha comprobado que abordar las teorías implícitas del aprendizaje de los docentes es un paso inicial importante para implementar la evaluación formativa. Chris WATKINS ha observado que el modelo dominante de aprendizaje sigue siendo el de "enseñar es hablar y aprender es escuchar" (2003, pág. 10), que hay que cambiar para "construir el conocimiento en el contexto de hacer cosas con los demás" (pág. 14). Esto cuadra bien con los enfoques actuales de la EpA.

Los supuestos clave de la EpA son que el aprendizaje es:

- un proceso social activo;
- en el que el individuo crea el significado;
- y la mejor manera de hacerlo es construir sobre lo que ya se conoce.

Cada uno de estos elementos conlleva un importante bagaje teórico que solo se despliega poco a poco. Yo me limito a presentar un breve resumen de lo ya alcanzado.

---

* "Aprender a aprender". (*N. del T.*)

[12] Véase: MARSHALL y DRUMMOND (2006).

Orígenes neoconductistas

La idea del uso formativo de la evaluación ("ayudarles a mejorar lo que quieren hacer") tiene su origen en los modelos neoconductistas del aprendizaje experto propuesto en 1971 por Benjamin BLOOM y sus colaboradores en los EE.UU. Se preveía aquí la división del aprendizaje en pequeñas unidades y, una vez enseñada una unidad, tendría lugar una evaluación formativa, normalmente en forma de un test de papel y lápiz. Es la *regulación retroactiva* de ALLAL y LOPEZ. Basándose en los resultados de éste, se adoptarían medidas correctoras con el fin de alcanzar los objetivos instructivos. La meta de la evaluación formativa era, por tanto, *la recuperación de las dificultades de aprendizaje*, resaltando la modificación de los enfoques instructivos de los docentes para conseguirlo. Este modelo todavía impregna las interpretaciones habituales norteamericanas de la evaluación formativa[13]. Por ejemplo, una práctica habitual es enseñar una unidad curricular durante unas seis semanas, seguida por un test de opciones múltiples editado comercialmente; después, se dedica una semana a trabajar sobre los temas cuyas puntuaciones hayan indicado una falta de dominio de la materia. Este período de recuperación de una semana se conoce como "evaluación formativa".

En los EE.UU., el conductismo ha dado paso a una visión *constructivista* de nuestra forma de aprender[14]. Esto implica "cómo trabaja la mente", algo que el conductismo obviaba. Se destacan aquí los procesos cognitivos mediante los que damos sentido a la información nueva. No creamos nuevas ideas de la nada: construimos sobre lo que ya sabemos y tratamos de dar sentido a la información nueva. Lorrie SHEPARD lo resume así: "El sentido hace más fácil el aprendizaje, porque el aprendiz sabe dónde poner las cosas en su marco mental de referencia, y el sentido hace útil el saber porque las probables finalidades y aplicaciones ya forman parte de la comprensión" (1992, pág. 319).

Una de las limitaciones del enfoque constructivista es que los elementos situacionales se minimizan con frecuencia. Parece que esto tiene dos causas: la primera es que el centro de atención es la creación individual del sentido. La segunda es que el interés por el modo de transferirse el aprendizaje, sobre todo en sus formas más abstractas, como las matemáticas y las ciencias, tiende a descontextualizarlo.

Constructivismo social

Probablemente, la denominación más adecuada del enfoque de la teoría del aprendizaje que subyace a las posturas actuales de la EpA, incluyendo la mía,

[13] Hay algunas excepciones notables a esto, en particular, el trabajo de Rick STIGGINS, a través de su organización ATI, y Lorrie SHEPARD, en su influyente alocución presidencial en la AERA (*Educational Researcher*, 2000).

[14] Dos libros importantes, patrocinados por la *National Academy*, que han adoptado este enfoque son: BRANSFORD y cols. (2000): *How People Learn: Brain, Mind, Experience and School*, y PELLEGRINO y cols. (2001): *Knowing What Students Know: The Science and Design of Educational Assessment*. Ambos adoptan un enfoque pronunciadamente cognitivo (constructivista), que hace poco o ningún hincapié en los aspectos socioculturales.

es la de "constructivista social". Este enfoque trata de mantener en equilibrio el aprendizaje como actividad cultural y como búsqueda de significado. Así, el aprendizaje tiene menos que ver con interpretaciones personales idiosincrásicas que con la adaptación personal de saberes y significados socialmente creados. Yo no creo mi propio sistema de matemáticas, sino que trato de dar sentido a lo que significa esta actividad social.

A nivel teórico, incluso la expresión "constructivismo social" es discutible, pues representa lo que Lorrie SHEPARD llama "teoría combinada, de punto medio", que toma elementos de "campos que, a veces, están enfrentados entre sí" (2000, pág. 6). Esto no contribuye precisamente a hacer amigos entre los puristas, tanto del campo constructivista, con su énfasis en la búsqueda individual de sentido, como del campo del aprendizaje situado, en el que se considera que el aprendizaje es el resultado de la participación en una "comunidad de práctica"[15].

No obstante, el constructivismo social tiene más historia que la de una síntesis reciente de estas teorías enfrentadas. Mary JAMES ha señalado que el énfasis en el contexto social del aprendizaje individual puede seguirse tanto hasta John DEWEY (y, antes que él, hasta William JAMES), en Estados Unidos, como hasta Lev VYGOTSKY, en la Rusia marxista. Del funcionalismo de DEWEY procede la insistencia en la interacción entre el individuo y el medio; VYGOTSKY insistía en que las relaciones sociales preceden al aprendizaje y a la interacción de acción y pensamiento. Ambos enfoques destacan la base social y cultural del aprendizaje[16].

¿Dónde nos deja esto? Paul COBB, un educador de Matemáticas, presenta una forma útil de avanzar en la comprensión de la búsqueda de significado en un contexto social. De acuerdo con el tema central de este libro, sostiene que los términos que utilizamos no "se corresponden con la realidad", sino que, simplemente, una cosa se entiende mejor con un vocabulario que con otro, por lo que el vocabulario preferido es el más útil para una finalidad determinada. Esto conduce a una postura pragmática, por lo que la adopción de uno u otro enfoque "debe justificarse en términos de su potencial para abordar cuestiones cuya resolución pueda contribuir a la mejora de la educación de los estudiantes" (1994, pág. 18). Trata el aprendizaje como una construcción (cognitiva) individual activa y como un proceso sociocultural. Utilizando las imágenes del primer plano y del fondo, lo cognitivo y lo sociocultural cambiarán sus posiciones según la finalidad:

> Creo que la perspectiva sociocultural da lugar a las teorías de las condiciones de posibilidad del aprendizaje, mientras que las teorías desarrolladas desde la perspecti-

[15] LAVE y WENGER (1991) presentan la exposición clásica de esta postura. En este enfoque, se considera que el aprendizaje supone una participación creciente en prácticas sociales, en vez tratarse de una organización mental del individuo que la transfiere a situaciones nuevas. Por tanto, la mente se ubica en el "individuo en la acción social", en vez de en la cabeza. Ejemplos de este aprendizaje "situado" son los niños brasileños de la calle que pueden sumar y restar rápidamente cantidades de dinero, que es parte de su aprendizaje cognitivo para trabajar y que, sin embargo, fracasan en las pruebas escolares convencionales de aritmética sobre el mismo material. Esto se debe a que las pruebas tienen poca relevancia contextual.

[16] Comunicación personal. Mary JAMES (2006) presenta un buen resumen de la teoría del aprendizaje en relación con la "evaluación para el aprendizaje" en su capítulo: "Assessment, Teaching and Theories of Learning", en J. GARDNER (ed.) (1983): *Assessment and Learning*. Londres: Sage.

> va constructivista se centran tanto en lo que aprenden los estudiantes como en los procesos mediante los que lo hacen.
>
> (pág. 18.)

El influyente artículo "On Two Metaphors for Learning and the Dangers of Choosing Just One" *, de Anna SFARD, expresa un pensamiento similar. Analiza la autora el uso de las metáforas de la adquisición ("hacerse con" el conocimiento) y de la participación (conocimiento mediante la acción con los demás) y cómo una actúa como un valioso control de los excesos de la otra. SFARD sostiene que "cuanto antes aceptemos el pensamiento de que nuestro trabajo está obligado a producir un mosaico de metáforas, en vez de una teoría homogénea del aprendizaje, mejor para nosotros y para aquellos cuyas vidas probablemente se vean afectadas por nuestro trabajo" (SFARD, 1998, pág. 12).

Mi lectura de la bibliografía de la EpA me indica que esa es también la postura adoptada en gran parte de ella. Si ha habido algunos cambios recientes, estos tienden a hacer más hincapié en lo sociocultural[17]. Este cambio puede ser indicativo del creciente reconocimiento de la importancia de las interacciones y relaciones en el aula en una evaluación formativa eficaz. Para nuestros fines, la postura constructivista social hace hincapié en que el aprendizaje es un proceso *social* y *activo* de búsqueda de significado. A este respecto, resulta útil aquí el concepto del *aprendizaje intencional*, en el que los aprendices tratan de aprender y el maestro trata de ayudarles[18]. Esto implica esforzarse para intentar resolver un problema determinado *y* transferir ese aprendizaje a otros problemas. Surge aquí el contraste con el conocido "hacer tareas", en el que la solución es un fin en sí misma y lo que se hace tiene poca influencia posterior en el aprendizaje.

La EpA y el aprendizaje efectivo

Las reivindicaciones que hace Evaluación para el Aprendizaje con respecto a que permite un aprendizaje más eficaz se basa en que la evaluación se utiliza para ayudar a los aprendices a:

---

* "Sobre dos metáforas del aprendizaje y los peligros de tomar solo una". (*N. del T.*)

[17] ALLAL y LOPEZ (2005) han puesto de manifiesto una tendencia reciente en la bibliografía francesa sobre la evaluación formativa, que se basa en la obra de Lev VYGOTSKY y enfatiza la actividad social, particularmente el lenguaje, como base del aprendizaje. Del mismo modo, en Inglaterra, BLACK y WILIAM (2006) presentaron una teoría de la evaluación formativa basada en la *teoría de la actividad* de ENGESTRÖM, que se deriva de la tradición sociocultural de VYGOTSKY. ENGESTRÖM, cuyo centro de atención es el lugar de trabajo, se interesa por los cambios que tienen lugar cuando se produce el aprendizaje. Utiliza el concepto de "sistema de actividad", con una compleja serie de componentes interactivos (por ejemplo, herramientas, roles, reglas y resultados). Se reconoce que el cambio es el resultado de interacciones complejas, de manera que, por ejemplo, ese aprendizaje no solo afecta al aprendiz, sino que modifica el sistema porque el aprendiz se reposiciona dentro de él. Esta teoría difiere de las del aprendizaje situado en las que el aprendizaje parece un proceso unidireccional de incorporación a una comunidad establecida de práctica.

[18] La expresión procede de BEREITER y SCARDAMAGLIA (1989). Mi explicación se basa en el valioso artículo de BLACK y cols. (2006): "Learning How to Learn and Assessment for Learning: a Theoretical Enquiry".

- tener más claro lo que hay que aprender y cómo será lo que se consiga;
- reconocer lo que comprenden y lo que no en el presente;
- percatarse de la mejor manera de avanzar.

Como ya hemos visto, los mecanismos para ello implican hacer explícitas las intenciones del aprendizaje y los criterios de éxito; utilizar preguntas y otras informaciones para descubrir lo que se comprende, y emplear la retroinformación para "salvar la distancia".

El problema es que este enfoque puede reducirse con facilidad a una serie de técnicas de clase (la *letra*), sin entender claramente por qué se usan (el *espíritu*). Algunas técnicas son de por sí complejas y pueden malinterpretarse, complicando las dificultades. Por eso, como en el caso de los "estilos de aprendizaje" o en el de las "inteligencias múltiples", aunque encierren el potencial para mejorar el aprendizaje en clase, pueden distorsionarse con facilidad. En consecuencia, volveremos sobre lo que considero como las tensiones fundamentales en la implementación de la EpA. Las he agrupado en cuatro áreas principales:

1. ¿Qué se aprende?
2. Claridad frente a conformidad.
3. Lo formativo en un clima sumativo.
4. Retroinformación eficaz.

¿Aprender o aprender a aprender?

Nos encontramos con otro caso de dos conceptos y del peligro de escoger solo uno. La tarea de la EpA en facilitar el aprendizaje directo de algo ("adquisición") y el más indirecto "aprender a aprender", referido también como "metacognición", "autonomía del aprendiz" y "aprendizaje autorregulado". Uno de los riesgos de estos es que se presta tanta atención a los procesos, por ejemplo, de "aprender a aprender" que se pasa por alto *lo que* hay que aprender: se considera fundamental el proceso, no el resultado. El aprendizaje no puede desarrollarse en el vacío; hay que aprender algo. Este enfoque del proceso puede provocar también el peligro de restringir los métodos docentes. Anna SFARD advierte:

> Las prácticas educativas tienen una propensión irresistible a las recetas prácticas extremas, universales. Una mezcla de moda... que tiene mucho que ver con la *metáfora de la participación*; se traduce a menudo en la prohibición total de la "enseñanza oral", el imperativo para hacer obligatorio para todos el "aprendizaje cooperativo" y la invalidación completa de la instrucción que no se "base en problemas" o no esté situada en un contexto de la vida real. Pero esto significa poner demasiada cantidad de algo bueno en una sola olla.
>
> (1998, pág. 11.)

Dos ejemplos nos mostrarán la importancia de "utilizar ambos". El primero procede del detallado estudio de Ann WATSON de dos profesores de Matemáticas que estaban practicando la evaluación formativa en sus clases. Cuando sus alumnos no comprendían un problema determinado, la retroinformación se hacía pri-

mordialmente en términos de reflexionar sobre sus procesos de aprendizaje. De ese modo, al centrar la atención en aprender a aprender, nunca se facilita ayuda específica suficiente para resolver los problemas en los que se hayan quedado atascados. En parte, esto es el resultado de que los mismos profesores no tengan suficiente "conocimiento práctico" de matemáticas que les permita ayudar y de su reticencia a utilizar la retroinformación, la información sobre los resultados de la prueba y los informes de apoyo (de los que eran críticos). Así pues, es posible que los estudiantes se sintieran potenciados, pero seguirían sin poder resolver el problema de matemáticas.

En contraste, algunas enseñanzas pueden ser tan concretas que lo adquirido tenga poco valor para acometer otros aprendizajes. Me parece que éste es un problema mucho más general. El trabajo de recuperación basado en los tests formativos estadounidenses es con frecuencia de esta naturaleza, pues genera una "microenseñanza" basada en un ítem incorrecto de un test de opciones múltiples. Aunque se preste atención al contenido, es probable que el aprendizaje solo se generalice a ítemes similares y no se produzca un aprendizaje *intencional* que se generalice a nuevos contextos.

Para la EpA, el problema consiste en establecer un equilibrio entre estimular el aprendizaje directo y desarrollar la *autorregulación* de los aprendices que lleve a la reflexión acerca de cómo encaran su aprendizaje. No son actividades independientes: mi autorregulación informará mi aprendizaje directo, que, a su vez, desarrollará mi autorregulación. David BOUD ha introducido la valiosa idea de la *doble tarea* de la evaluación: apoyar el programa actual de aprendizaje y, al mismo tiempo, aumentar la comprensión de los aprendices de la evaluación y de su capacidad de autoevaluación futura. Lo denomina *evaluación sostenible*. Ni el simple cumplimiento de criterios ni el desarrollo de competencias de autorregulación son suficientes por separado; ambos tienen que estar presentes. Las actividades de evaluación:

> Tienen que ocuparse de la tarea inmediata y de las consecuencias para equipar a los estudiantes para un aprendizaje de por vida en un futuro desconocido... tienen que prestar atención tanto al proceso como al campo sustantivo de contenidos.
>
> (2002, pág. 9.)

Esto tiene consecuencias respecto a la forma de evaluar el impacto de las prácticas de la "evaluación para el aprendizaje". Si se afirma que ayuda a aprender, ¿cómo puede evaluarse? Uno de los atractivos de *Inside the Black Box* era que estimaba las mejoras espectaculares de los resultados de los exámenes que podían esperarse. Sin embargo, todavía hay poca evidencia empírica directa de la influencia de la EpA en el rendimiento[19]. Esto se debe en parte a que es difícil de hacer, porque la EpA puede ser solo una de las diversas iniciativas o cambios que se desarrollen en un aula. Los diseños de investigación que emplean grupos de control, que podrían explicar esto, no encajan bien con los enfoques de investigación-acción que suelen adoptarse, aunque, en los Estados Unidos, se han

[19] Una excepción ha sido el análisis de WILIAM y cols. (2004) del impacto de su proyecto de investigación-acción de evaluación formativa sobre el rendimiento de los estudiantes.

convertido, en muchos casos, en requisito para obtener financiación. No obstante, hacen falta más pruebas de la influencia sobre el rendimiento para que sean creíbles las afirmaciones acerca de la mejora del aprendizaje.

Esto no quiere decir que los proyectos de EpA no hayan sido evaluados pero, cuando lo son, suele centrarse la atención en el aspecto de la *participación*. Por regla general, se atiende a cómo han modificado su práctica los docentes, junto con los cambios de las actitudes y de la participación de los estudiantes. El *ARIA Project*, financiado por la *Nuffield Foundation*, está revisando en la actualidad la influencia de las iniciativas de evaluación en el aula y ha descubierto que la mayoría informan en estos términos [20].

¿Claridad o conformismo?

Uno de los elementos clave de la EpA es la insistencia en hacer explícito lo que se aprende y qué se considera un aprendizaje satisfactorio. El fundamento teórico es claro: es más fácil aprender cuando sabemos lo que hacemos y adónde estamos intentando llegar. El aprendizaje podría incluir las competencias de autorregulación, así como el conocimiento y la comprensión de la materia. Si el aprendizaje en clase es un proceso colaborativo, tanto docentes como alumnos tienen que saber qué se prevé. Sin embargo, conseguir claridad en este proceso es como caminar por la cuerda floja: si no está claro lo que se aprende (y por qué) y qué se considera satisfactorio, los aprendices estarán desconcertados; si se especifica de forma demasiado rigurosa, se convierte en un ejercicio de conformismo.

La falta de claridad genera el tipo de desconcierto expresado por dos de los alumnos de 15 años, que comentaron:

> No es que no haya aprendido mucho, sino que, en realidad, no entiendo lo que hago.
>
> (HARRIS y cols., 1995, pág. 253.)

> En Matemáticas especialmente (el profesor) lo explica en la pizarra y no entiendo lo que dice, pero tenemos un libro y también lo explica. Por eso, mi amigo y yo vamos leyendo este libro porque lo explica mejor y, mientras lo leemos, él ya va por el capítulo siguiente y no sabemos lo que hace. Nos perdemos y todo eso.
>
> (RUDDUCK, 196, pág. 42.)

Mi ejemplo favorito del estudiante desconcertado es el examinando que respondió a la pregunta de Geometría de la Figura 7.1.

De alguna manera, esto puede explicar por qué se observa que los estudiantes de menor rendimiento son quienes hacen los mayores progresos gracias a la evaluación formativa. Al encontrarse menos cómodos en la cultura de la escuela, es menos probable que entiendan lo que se les pide. Es muy posible que los estudiantes de alto rendimiento, que con frecuencia se encuentran a sus anchas en

[20] Véase el sitio web de ARIA: http://www.aria.qub.ac.uk.

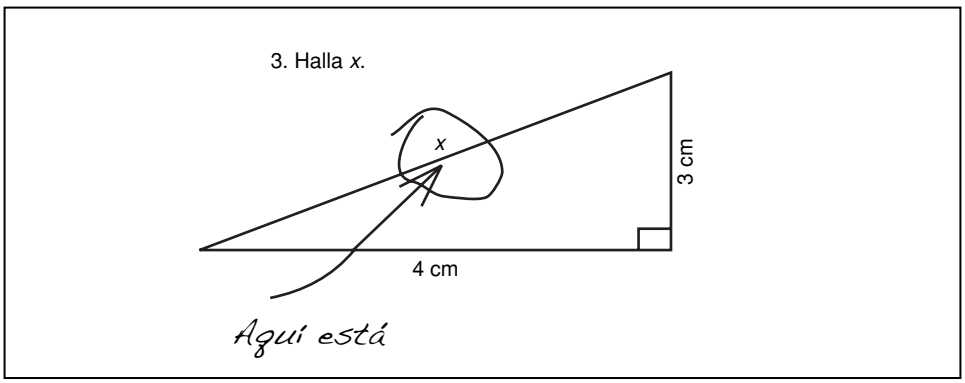


Figura 7.1. *Solución de Geometría.*

la escuela, ya hayan desarrollado las competencias de autorregulación que les permiten comprender lo que tienen que hacer, aunque no se haya manifestado con claridad.

Los peligros de ser explícitos

Sin embargo, al hacer explícito lo que hay que aprender y los criterios de éxito correspondientes se corre el riesgo de caer al otro lado de la cuerda floja. En este caso, hay una tendencia a plantear unos objetivos de aprendizaje cada vez más detallados que especifiquen el logro requerido y que se anuncien, en vez de negociarlos. Esto explica mi preferencia por la expresión *intenciones del aprendizaje*, de Shirley CLARKE, dado que transmite una sensación tanto de flexibilidad como de amplitud. Esto es importante para el espíritu de la EpA, pues puede que sea necesario adaptar o aun abandonar un plan de clase para alcanzar estos objetivos más amplios. Por ejemplo, Noel ENTWISTLE y sus colaboradores (2000) concluyeron que la mayoría de las acciones de aprendizaje en la educación superior provenían de desviaciones no planeadas. La *intención* sirve también para materias o temas en los que el aprendizaje no siempre es lineal. Mientras que, en Matemáticas, el dominio de un determinado concepto puede alcanzarse a través de una secuencia previsible de destrezas, en materias como Lenguaje o Arte, esto es mucho más variable. Una idea útil a este respecto es considerar las intenciones y metas en términos de un *horizonte*, en vez de como un punto fijo: sabemos el nivel de rendimiento que queremos alcanzar, pero distintos estudiantes lo lograrán en diferentes lugares de ese horizonte. Una narración imaginativa puede surgir en muchas formas y tamaños, por ejemplo, ¿cómo responderíamos al inquietante relato en seis palabras de Ernest Hemingway: *"For sale: baby shoes, never worn"**?

* En castellano, son siete palabras o más: "Se venden: zapatos de bebé, no usados nunca". (*N. del T.*)

Esta tendencia se intensifica cuando los objetivos se imponen al docente mediante un currículum o requisito muy especificado. Por eso, no solo se presentan a los alumnos como unos objetivos prefijados, sino también a los profesores. En tal caso, el proceso se hace cada vez más mecánico, pues se anima a los aprendices a que dominen pequeños y detallados elementos del currículum. Esta explicitud, que pretende lograr que el aprendizaje resulte claro para el alumno puede, en realidad, reducir su autonomía, en vez de promoverla, sobre todo si es el profesor quien tiene que descodificar estos requisitos [21]. Volvemos entonces al aprendizaje para obtener una calificación o la señal en una casilla (Capítulos V y VI), un enfoque que hace poco probable la transferencia del aprendizaje. Este es un ejemplo de "hacer cosas", en vez de un aprendizaje intencional. Kathryn ECCLESTONE comenta esto en el sentido de que solo se permite a los estudiantes una *autonomía procedimental*, que los deja como "cazadores y recogedores de información, sin un interés profundo por los contenidos o los procesos" (2002, pág. 36). Recogiendo el sentido de esta tendencia concreta, Harry TORRANCE ha acuñado la expresión: *docilidad a los criterios*. Concluye:

> La transparencia, no obstante, fomenta el instrumentalismo. Cuanto más clara esté una tarea para alcanzar una nota o una recompensa y más detallada sea la ayuda prestada por los tutores, supervisores y consejeros, más probable es que los candidatos tengan éxito; pero, ¿éxito en qué? La transparencia de objetivos, junto con el uso generalizado de tutorías y prácticas para ayudar a los aprendices a alcanzarlos, encierra el peligro de eliminar el reto del aprendizaje y reducir la calidad y la validez de los resultados alcanzados. Hemos identificado un movimiento, a partir de lo que caracterizamos como evaluación del aprendizaje y a través de la idea, popular en la actualidad, de la evaluación para el aprendizaje, hacia la evaluación como aprendizaje, en el que los procedimientos y prácticas de evaluación pueden llegar a dominar completamente la experiencia de aprendizaje y en el que la "docilidad a los criterios" llegue a reemplazar al "aprendizaje". Este es el reto más significativo al que se enfrenta el sector [postobligatorio] de "aprendizaje y competencias": equilibrar la explicitud de los objetivos del aprendizaje y los procesos de enseñanza frente a la validez y la conveniencia de los resultados del aprendizaje.
>
> (2005, pág. 2) [22].

Esto también constituye un reto para el sector escolar, en el que tanto las comunicaciones sobre el currículum como las especificaciones de los exámenes

---

[21] El estudio de GROVES (2002): "They can read the works but do they know what they mean?" demostró que los criterios de aprendizaje eran más exigentes que el *Tractatus* de WITTGENSTEIN, cuando los analizó con el índice de comprensibilidad Gunning Fog*.

* En castellano, se le llama a veces "índice de niebla (*fog*) Gunning". Es una medida del grado de dificultad de comprensión de un texto. Robert Gunning, hombre de negocios estadounidense, ideó esta prueba en 1952. (*N. del T.*)

[22] Este uso de la evaluación *como* aprendizaje difiere de otros como, por ejemplo, Lorna EARL (2003b); Ruth DANN (2002), y la iniciativa escocesa *"Assessment is for Learning"*, en los que el término se usa como un proceso constructivo en la "evaluación para el aprendizaje". Por tanto, se interpreta que la evaluación *refuerza* el aprendizaje, mientras que el uso que hace TORRANCE implica que la evaluación *desplaza* el aprendizaje.

presentan unas especificaciones cada vez más detalladas, basadas en los resultados[23].

Una amenaza clave para la evaluación formativa es la de que se reduzca a un mecanismo para suavizar la "impartición" de un currículum especificado y preparar para una evaluación sumativa. Una forma concreta de esta amenaza es un sutil cambio de significados obrado por los responsables de la política educativa cuando incluyen la EpA en sistemas más amplios de rendición de cuentas. En su artículo "The Trouble With Learning Outcomes"*, Trevor HUSSEY y Patrick SMITH sostienen que, en la educación superior,

> aunque los resultados del aprendizaje puedan ser útiles, si se utilizan adecuadamente, han sido usurpados y adoptados de forma generalizada en todos los niveles del sistema educativo para facilitar el proceso administrativo. Esto ha llevado a su distorsión... La interpretación adecuada de estos resultados debe desprenderse del contexto y de las actividades y experiencias preponderantes de los estudiantes.
>
> (2002, págs. 220, 232.)

Lo que empieza como un centro de atención del aprendizaje se transforma en objetivos que cumplir mediante unas puntuaciones mejoradas. Como los responsables de la política educativa no distinguen entre puntuaciones mejoradas y aprendizaje mejorado (véase el Capítulo VI), este es un paso lógico. Sin embargo, los resultados y el aprendizaje no son equivalentes, por lo que esto se convierte en una distorsión del aprendizaje. Un ejemplo elocuente de cambio tan sutil fue la descripción de la EpA hecha por el entonces *Minister of Schools*** de Inglaterra:

> La "evaluación para el aprendizaje" que se incluye en la planificación de la clase y en las estrategias de enseñanza, establece objetivos claros e identifica claramente lo que tienen que hacer los alumnos para llegar allí... y nosotros conseguiremos despegar realmente cuando todos los interesados por el progreso de los alumnos hagan el máximo uso de los datos y parámetros de referencia.
>
> (David Millband: *Observer*, 1 de junio de 2003.)

Aquí, los "objetivos" son numéricos, niveles y notas, en vez de resultados del aprendizaje, por lo que "llegar allí" puede deslizarse con facilidad a la microenseñanza de formas de obtener puntos adicionales, sobre todo cuando esto se promueve desde instancias centrales (véase el Capítulo VI).

---

[23] Por ejemplo, en Inglaterra, la QCA* ha elaborado los *Assessment Foci*** con el fin de facilitar una enseñanza más precisa de los puntos débiles del currículum nacional de Lenguaje, un modelo de prueba y recuperación de evaluación formativa; véase: SAINSBURY y cols. (2006): *Assessing Reading*, cap. 13.

* *Qualifications and Curriculum Authority*: "Administración de titulaciones y currículum". (*N. del T.*)

** "Centros de interés para la evaluación". (*N. del T.*)

* "El problema de los resultados del aprendizaje". (*N. del T.*)

** "Ministro de las Escuelas", cargo del *Department for Education and Skills*, de nivel inferior al del *Secretary of State* titular del departamento, equivalente al de un "Secretario de Estado" español. (*N. del T.*)

Neutralidad curricular

Una vulnerabilidad relacionada con esto de las versiones actuales de la EpA es que tienden a preocuparse por mejorar el aprendizaje, con independencia de lo que haya que aprender. Al tratar el currículum como un dato, esto ha llevado a una respuesta en gran parte pasiva a los contenidos y competencias implicados. En contraste, Terry CROOKS ha defendido una y otra vez la importancia de hacer que el currículo sea significativo para el aprendiz:

> Es mucho más probable que los estudiantes aprendan cosas que les preocupen que las que, para ellos, tengan poco significado o importancia. Hagan lo que hagan los profesores para evaluar a los estudiantes y guiar su aprendizaje, es mucho menos probable que tengan éxito si los estudiantes no están motivados para aprender ese material o esas competencias.
>
> (2007, pág. 1.)

En su contexto neozelandés, en el que el currículum es menos prescriptivo, esto es más fácil de realizar, pues pueden ofrecerse opciones. Sin embargo, aun en climas más prescriptivos, el teórico de la motivación Jere BROPHY señala que:

> los principios motivacionales básicos indican que todo lo incluido en un currículum debe estarlo porque merezca la pena aprenderlo por razones que puedan entender los aprendices, y estas razones deben destacarse al presentar el contenido y apoyar las actividades que se vayan a utilizar para desarrollar el aprendizaje.
>
> (1998, págs. 5-6.)

Creo que esto casa bien con la negociación de las intenciones del aprendizaje; puede no versar simplemente sobre lo que haya que aprender, sino también sobre el valor de aprenderlo. El problema surge cuando "gran parte de lo que aparece en el currículum escolar no merece la pena aprenderlo" (BROPHY, 1998, pág. 11), aunque haya que acatarlo. Si no nos preocupamos por la validez de lo que haya que aprender, es probable que la EpA se deslice hacia actitudes instrumentales cuyo objetivo sea obtener buenas notas, si hay pocas cosas cuyo aprendizaje pueda recomendarse por su valor propio. Como Michael APPLE ha señalado, esto también encierra un elemento de justicia social: es importante *lo que* enseñamos.

Lo formativo en un clima sumativo

La idea de la *doble tarea* de BOUD (2000) tiene otra aplicación: confina la evaluación formativa al aprendizaje y la evaluación sumativa a la certificación. ¿Cómo podemos asegurar el desarrollo de una "evaluación sostenible", al tiempo que cumplimos los objetivos a corto plazo de la certificación o la rendición de cuentas? Es, en esencia, una expresión práctica de las secciones anteriores sobre el aprendizaje y la claridad, grave cuando hay presiones para asignar un nivel o calificar trabajos individuales y para administrar regularmente pruebas que acarrean consecuencias importantes. Esas calificaciones enmascaran a menudo

una confusión, pues esta evaluación se presenta muchas veces como formativa (dado que informa sobre el progreso y los niveles alcanzados) cuando su función es, en realidad, sumativa (una instantánea de mi situación actual). Gran parte de este tipo de evaluaciones debería clasificarse más exactamente como *sumativa frecuente* o *minisumativa*. Lo que se haga con esta información es lo que determinará si *se convierte* en formativa: ¿lleva a aprendizajes posteriores? Por tanto, la diferencia está en la finalidad y no en el momento.

Lo que complica aún más la cuestión es que las funciones formativa y sumativa forman parte a menudo de un bucle, en vez de ser independientes una de otra. Por ejemplo, si la parte evaluada por el profesor de un ejercicio para calificación tiene especificados unos resultados detallados para cada nivel, el proceso de preparación puede ser formativo: ¿hay mucha distancia entre mi trabajo actual y los resultados buscados? y ¿qué tengo que hacer para aproximarme a ellos? Cuando alcance estos resultados, el proceso se convierte en sumativo y se me concede este nivel, para utilizar esta información en relación con el nivel siguiente, de manera que vuelve a ser formativa. El riesgo que se corre aquí es el de la docilidad a los criterios, con poca o ninguna "evaluación sostenible" o transferencia de aprendizaje.

Creo que la enmarañada relación formativa-sumativa es una de las cuestiones prácticas más difíciles para la EpA. Aunque esta cuestión ha sido abordada académicamente[24], la prueba real de la dificultad proviene de la frecuencia con la que la evaluación formativa queda en suspenso cuando lo imponen las presiones del examen. Sigue vigente la imagen de que la evaluación formativa es "una buena cosa", pero, cuando empieza la preparación de los exámenes, tenemos que enfrentarnos a "lo real". Esto significa frecuentes evaluaciones sumativas y enseñanza directa para el examen, aunque haya pruebas de que los aprendices autorregulados rinden más[25].

Esto significa que, en culturas de rendición de cuentas, con frecuentes pruebas para calificación, es más difícil hacer progresos en la evaluación formativa. En estas culturas de evaluación, solo las escuelas y los profesores más seguros de sí mismos se arriesgarán a promover el aprendizaje autorregulado, la autoevaluación y la evaluación a cargo de compañeros. Para la mayoría, la tarea consiste en cubrir todo el currículum y preparar para el examen. Hacen falta pruebas más convincentes de la contribución de la evaluación formativa a este proceso para que los profesores la mantengan en los cursos que concluyen con exámenes oficiales.

Retroinformación eficaz

Andrew MORGAN ha comparado la retroinformación formativa con un "crimen perfecto". Que sea eficaz y útil, depende de tres cosas: 1) *motivo*: el aprendiz lo necesita; 2) *oportunidad*: el aprendiz la recibe en el momento de utilizarla, y 3) *medios*: el aprendiz está dispuesto a usarla y es capaz de hacerlo. Como en algunos crímenes, a lo largo del camino se dejan muchas pistas falsas.

[24] Véase, por ejemplo: HARLEN (2006).
[25] Véanse: MCDONALD y BOUD (2003), y WILIAM y cols. (2004).

Todo el mundo facilita retroinformación; esto no es privativo de la EpA. Sin embargo, sabemos que lo que se denomina "retroinformación" a menudo no lo es, dado que, aunque la *intención* sea ayudar a aprender, esto no es la *consecuencia*; de hecho, puede haber retrasado el aprendizaje. Cuando he pedido a mis estudiantes ejemplos de retroinformación que les hayan ayudado en su aprendizaje en cualquier esfera de la vida, a la mayoría les ha resultado difícil encontrar uno. Cuando les he preguntado por retroinformaciones que hayan retrasado su aprendizaje, los ejemplos han surgido a los pocos segundos. Suele tratarse de humillaciones de los profesores. El hecho de que le digan a uno que puede estar en el coro, pero que no debe cantar, sino solo abrir y cerrar la boca, no parece que contribuya mucho a la formación musical, como tampoco parece que ayude en Matemáticas el comentario de que "no tienes nada que hacer".

Esta percepción de que la retroinformación no siempre ayuda a aprender cuenta con el apoyo de la evidencia de investigación. De su metaanálisis de investigaciones sobre la retroinformación, los psicólogos Avraham KLUGER y Angelo DENISI concluyen que:

> En más de un tercio de los casos, las intervenciones de retroinformación redujeron el rendimiento... creemos que tanto los investigadores como los profesionales confunden sus sentimientos de que la retroinformación es deseable con la cuestión de si la intervención de retroinformación beneficia el rendimiento.
>
> (1996, págs. 275, 277.)

Como la retroinformación es el mecanismo clave en la evaluación formativa para pasar de la actuación actual a la deseada, es crítico que sea bien entendida. Me parece que esto es algo que todavía no ha conseguido la EpA. Dos recientes revisiones importantes de la investigación sobre la retroinformación, una de Valerie SHUTE, en los EE.UU., y otra de John HATTIE y Helen TIMPERLEY, en Nueva Zelanda, han puesto de manifiesto la complejidad del proceso de retroinformación. Que la retroinformación tenga o no efectos positivos en el aprendizaje depende de muchos factores en interacción: la motivación, la complejidad de la tarea, la pericia del aprendiz y el nivel y la calidad de la retroinformación. Esto hace que sea muy situacional: la misma retroinformación dada a dos aprendices puede tener efectos opuestos. El simple hecho de decir a unos aprendices principiantes que están equivocados puede retrasar el aprendizaje; si se dice lo mismo a expertos interesados por lo que hacen puede ser suficiente para que aumenten sus esfuerzos y modifiquen su estrategia.

¿Qué sabemos, pues, acerca de la retroinformación eficaz, aparte de que es compleja y difícil de conseguir? Sabemos que el contexto en el que se dé y su oportunidad son importantes (la *oportunidad* de MORGAN). La forma que adopte también influye en el modo de recibirla y de actuar al respecto (*medios*), suscitando preguntas acerca de la función de las alabanzas y los puntos. Estos están íntimamente ligados a las propias actitudes de los aprendices ante su aprendizaje (*motivos*). El contexto, la forma y las actitudes están íntimamente interrelacionadas, por lo que su separación aquí solo pretende reducir la complejidad.

Contexto: El dónde y el cuándo de la retroinformación

La respuesta de Philippe PERRENOUD a la revisión de la evaluación formativa de BLACK y WILIAM de 1998 señalaba que el modelo de retroinformación allí presentado era solo una pequeña parte de las complejas interacciones de clase. Era necesario examinar más los antecedentes de la interacción de retroinformación. Pensaba en el clima del aula, pero puede ser útil extender los antecedentes a la cultura en la que tienen lugar la enseñanza y la retroinformación.

Algunas de éstas son características culturales generales, por ejemplo, la motivación de los estudiantes. En una sociedad que valora mucho la educación, por ejemplo, en los países de la costa del Pacífico, la motivación para aprender puede ser un dato social y asumirla el profesor o maestro. En muchas escuelas del Reino Unido y de los EE.UU., se da por supuesto con frecuencia que la motivación tiene que fomentarla la escuela; los estudiantes no llegan a las puertas de la escuela deseosos de aprender. La retroinformación entra entonces a formar parte de esta motivación y se utiliza con frecuencia para promover el aprendizaje futuro, en vez de abordarlo directamente. De este modo, como veremos, el elogio se convierte en un problema.

Otro ejemplo surge cuando las expectativas culturales refuerzan la consideración del aprendizaje como una actividad dirigida por el docente y de carácter didáctico, y conducen a desaprobar el uso de la autoevaluación y de la evaluación a cargo de compañeros como formas de facilitar la retroinformación[26]. En este caso, la resistencia puede proceder tanto de los estudiantes como de la sociedad en general. Ocurre, en especial, cuando existe un currículum prescriptivo y se administran pruebas que tienen consecuencias importantes, dado que, como ya hemos visto, a menudo, la retroinformación se convierte en una cuestión relativa al cumplimiento de los requisitos: "dígame que tengo que hacer para mejorar mi nota".

Al nivel del aula o del lugar de trabajo, uno de los factores clave de una retroinformación eficaz es un clima de *confianza* y *respeto*. Esto implica unas relaciones de apoyo y ayuda en el aula o el lugar de trabajo, en las que el aprendiz se sienta seguro para poder admitir dificultades y los docentes sean constructivos y estimulantes. Si el aula es un lugar competitivo en el que los estudiantes comparan sus notas y se reparten honores (el o la "estudiante n.º 1"), puede hacer más difícil el reconocimiento de las dificultades.

Oportunidad de la retroinformación

HATTIE y TIMPERLEY observan que la retroinformación es algo que siempre ocurre en segundo lugar: debe haber un contexto de aprendizaje al que se dirija. Si hay una falta básica de comprensión, "es mejor que el docente dé detalles a través de la enseñanza que facilite retroinformación sobre conceptos no muy bien comprendidos" (2007, pág. 104), algo que encaja bastante con la "evaluación formativa preventiva" de CARLESS (2007).

[26] David CARLESS (2005) ha escrito sobre estas presiones en la enseñanza en Hong Kong.

Hay una bibliografía compleja sobre la oportunidad de la retroinformación, información que no genera sencillas prescripciones. Valerie SHUTE sostiene que la retroinformación inmediata es eficaz en la corrección de errores y puede producir beneficios inmediatos; sin embargo, la retroinformación retardada se asocia con una mejor transferencia del aprendizaje, aunque la velocidad de éste pueda ser más lenta. Cuando un aprendiz aborda una tarea nueva y difícil, es mejor utilizar al principio una retroinformación inmediata (para reducir la frustración y los atascos). Sin embargo, con tareas más sencillas, es mejor retrasar la retroinformación (para prevenir sentimientos de "intrusión de la retroinformación"); lo mismo cabe decir cuando un aprendiz está activamente inmerso en una tarea. Todos hemos tenido la insatisfactoria experiencia de que nos den la solución de un rompecabezas o una pista antes de que quisiéramos que nos ayudasen. Una forma de dar sentido a esto es el concepto de *concienciación*. Se describe como la reflexión del aprendiz sobre "las pistas situacionales y los significados subyacentes relevantes para la tarea de que se trate" [27]. Si estoy atascado en una tarea compleja nueva para la que tengo un conjunto limitado de pistas y significados, la retroinformación que me aporte pistas me será útil. Si la retroinformación se proporciona demasiado pronto, el proceso no se habrá agotado y esto estimulará la *irreflexión*.

Un ejemplo instructivo de esta situación es la comparación de la enseñanza de las matemáticas en Japón y en Estados Unidos realizada por STIGLER y HIEBERT. Los estudiantes japoneses trabajaban en grupos para resolver problemas que, a menudo, implicaban dos maneras de hallar una solución. La retroinformación solo la daba el profesor cuando habían llegado tan lejos como les permitieran sus competencias, por lo que había una motivación para aprender más, es decir, concienciación. En cambio, a los estudiantes estadounidenses les enseñaban técnicas concretas antes de pedirles que trabajaran individualmente y las aplicaran a un problema. Los estudiantes que tuvieran alguna dificultad eran rescatados de inmediato por el profesor, que les decía qué técnica utilizar: todos los ingredientes de una respuesta sin reflexión con poca o ninguna transferibilidad a problemas diferentes.

Las características del aprendiz complican aun más esta complejidad. Es posible que los principiantes y de bajo rendimiento necesiten una retroinformación eficaz inmediata y explícita. En el caso de los expertos y de alto rendimiento, a quienes incluso las tareas complejas pueden resultarles relativamente fáciles, es mejor retrasar la retroinformación y facilitarla en formatos más desafiantes, por ejemplo, pistas o preguntas.

El centro de atención de la retroinformación: La tarea, no la persona

¿Por qué hay tantas retroinformaciones que obstaculizan el aprendizaje en vez de facilitarlo? Para KLUGER y DENISI [28], la respuesta está relacionada con la

[27] DEMPSEY y SALES (1993).

[28] La descripción de los niveles de HATTIE y TIMPERLEY (2007) es mucho más accesible que el tratamiento original de KLUGER y DENISI. Me he basado en gran medida en aquellos.

dirección de la retroinformación. El nivel en el que se dé es al que probablemente prestemos atención. Ese nivel puede ser uno de estos cuatro:

1. *Nivel de tarea.* Con frecuencia, se trata de una retroinformación correctora relativa a la mayor o menor precisión del trabajo, a la necesidad o no de más información y a la construcción de conocimiento más superficial. Las tareas que se benefician de este nivel de retroinformación son más las sencillas que las complejas. Alrededor del 90% de las preguntas de los docentes se dirigen a este nivel [29]. Este tipo de retroinformación tiene más fuerza cuando se refiere a interpretaciones erróneas y no a falta de información; para esto último es más eficaz ampliar la enseñanza.
2. *Nivel de proceso.* Aborda los procesos subyacentes a las tareas o relacionados con éstas y extensivos a ellas. La retroinformación a este nivel puede estar relacionada con la mejora de la detección de errores y con estrategias para abordar otras tareas más complejas. Tiene que ver con las estrategias de "aprendizaje profundo" del Capítulo IV.
3. *Nivel de regulación.* HATTIE y TIMPERLEY identifican seis factores importantes a este nivel que median la eficacia de la retroinformación. Son los siguientes: "la capacidad de crear una retroinformación interna y de autoevaluarse; la disposición a dedicar esfuerzos a buscar y abordar retroinformación; el grado de confianza o certidumbre en la corrección de la respuesta; las atribuciones de éxito o fracaso, y el nivel de competencia para buscar ayuda" (2007, pág. 94).
4. *Nivel personal.* Es una retroinformación personal que facilita juicios positivos (y, a veces, negativos) sobre el aprendiz. Se trata de expresiones como: "buena chica", "eres un fenómeno", que se oyen en muchas aulas y se utilizan en vez de los tres anteriores. El problema es que raramente es eficaz.

Atención al nivel

Si la retroinformación aborda lo que hace falta para mejorar el rendimiento en una tarea, le prestaremos atención a este nivel. Si la retroinformación se dirige a la persona, la respuesta será a nivel personal, que puede distraer del aprendizaje. La retroinformación más eficaz implica una interacción de los tres primeros niveles, que puede visualizarse como un bucle. La retroinformación que pretende empujar a los aprendices de la tarea al proceso y de éste a la regulación, es la más poderosa. Así, al haberme ayudado a hacer bien la tarea, me estimulan a relacionarlo y generalizarlo a otras tareas. Esto, a su vez, me ayuda a elaborar estrategias de autorregulación que me permitan supervisar mi progreso en una tarea (autoevaluación) y mi compromiso y esfuerzo (autocontrol). Para los aprendices más asentados y fuertes, la retroinformación puede empezar en este nivel de autorregulación: "comprueba si has respondido plenamente a tus cuestiones de investigación", y esto los llevará al nivel de tarea y al de proceso (comprobar el texto, reorganizarlo para relacionar las respuestas más directamente).

[29] AIRASIAN (1997).

La retroinformación *relacionada con la persona* queda fuera de este bucle, pues raramente contribuye a un nuevo movimiento. Esto se debe a que no contiene suficiente información sobre la tarea para encaminar al nivel de tarea o al de proceso y con frecuencia se centra más en la imagen de sí mismo que en la autorregulación. John HATTIE considera que esta retroinformación pasa por una "lente reputacional". Si me dicen que mi trabajo ha decepcionado a mi profesor, que sabe que puedo hacerlo mejor, tendré que pensar en proteger mi reputación. Puedo atribuir la calidad de mi trabajo a una falta de esfuerzo, proteger la visión que tengo de mí mismo diciéndome que tengo la capacidad de hacerlo (¿la estrategia masculina preferida?). Esta estrategia se ve favorecida si nunca realizo el máximo esfuerzo y me discapacito a mí mismo: preparando de antemano las razones por las que lo he hecho mal. Por ejemplo, el hecho de quedarse hasta tarde en el bar antes de una presentación o entrevista significa que tenemos preparada la excusa si sale mal o fracasa. BERGLAS y JONES han señalado que este tipo de comportamiento surge de una historia de retroinformación caprichosa y caótica: "no es que sus historias estén guardadas con repetidos fracasos; han sido ampliamente recompensadas, pero de formas y en ocasiones que los dejan profundamente inseguros acerca del motivo de la recompensa" (1978, pág. 407). (Yo diría que esto incluye elogios y apreciaciones banales, como: "que cuadro más bonito", cuando el niño sabe que no le ha dedicado mucho esfuerzo o reflexión). No obstante, si yo hubiese rendido al máximo y aun así decepcionara a mi profesor o a mis oyentes, ¿en qué lugar quedaría yo? Si este patrón se repitiera periódicamente, podría llevarme a un estado de "indefensión aprendida", en el que declararía: "No sirvo para esto", y lo evitaría.

El problema es que gran parte de la retroinformación que se facilita en el aula opera a nivel personal. Cuando Caroline GIPPS y sus colaboradores (2001) observaron clases impartidas por maestros de primaria, descubrieron que la mayor parte de la retroinformación incluía juicios expresados como alabanzas, estímulos o críticas hechos al nivel personal. De igual manera, los detallados estudios de BOND y sus colaboradores de 65 maestros en Australia descubrieron que la forma más común de retroinformación era el elogio.

El problema de los elogios y las recompensas

Cuando los maestros y profesores elogian o premian a los niños, es probable que los aprendices lo consideren como retroinformación positiva. La cuestión crítica es si el elogio *aparta la atención de la tarea y la dirige hacia el yo* y a actividades "reputacionales" (por ejemplo, cambiar a otras tareas más fáciles, que mantienen la imagen que tienen de sí mismos), o si *conduce a cambios en el esfuerzo, el compromiso o los sentimientos de eficacia de los aprendices* en relación con la tarea. Los docentes justificarían sus elogios aludiendo a esta fuerza motivadora, aunque la evidencia de investigación apunta en la dirección opuesta. Los metaanálisis han demostrado que las alabanzas de los docentes, como reforzadores o recompensas, tienen una influencia extremadamente limitada sobre el rendimiento de los estudiantes, aunque vayan acompañadas de retroinformacio-

nes al nivel de la tarea. KLUGER y DENISI descubrieron que la ausencia de elogios tenía una mayor influencia positiva en estas condiciones.

Podríamos esperar, entonces, que el elogio estimulara la autorregulación. Sin embargo, esto también es problemático, pues es probable que el elogio conduzca a la dependencia del aprendiz, en vez de a su autonomía. Alfie KOHN sostiene que puede

> crear una dependencia creciente de asegurarse la aprobación de otra persona. En vez de ofrecer un apoyo incondicional, el elogio hace que la respuesta positiva esté condicionada a hacer lo que pida el adulto. En vez de promover el interés por una tarea, el aprendizaje se devalúa en la medida en que se considera un prerrequisito para recibir la aprobación del maestro.
>
> (1994, pág. 3.)

Para complicar aún más las cosas, sabemos que la mayoría de los estudiantes adolescentes prefieren ser elogiados discretamente y en privado, mientras que los más jóvenes prefieren los elogios por esforzarse mucho más que por tener gran capacidad. Para algunos estudiantes, el elogio público es un castigo: afecta su reputación como "malos" estudiantes. Es interesante señalar que el elogio también puede ser contraproducente en relación con las percepciones que los aprendices tengan de su capacidad. En el caso de los estudiantes mayores, se ha demostrado que el elogio tras el éxito y la retroinformación neutra después de un fracaso se interpreta como un indicio de que el profesor considera que su capacidad es baja. En el caso de los alumnos más pequeños, ocurre lo contrario [30].

Pueden utilizarse argumentos similares con respecto a recompensas como las pegatinas de caras sonrientes y las menciones especiales. Hay aún una cuestión acerca de si deben considerarse como retroinformación, porque ofrecen muy poca información. Un metaanálisis de DECI y colaboradores descubrió una correlación negativa entre las recompensas extrínsecas y el rendimiento en la tarea, que se hacía aun más pronunciada en tareas interesantes. La postura de DECI es que tales recompensas "dificultan que las personas asuman su responsabilidad y se motiven y regulen a sí mismas" (1999, pág. 659). Solo en las tareas poco interesantes las recompensas tuvieron un efecto positivo, lo que dice mucho del currículum y de las tareas en las aulas de pegatinas y menciones. Por tanto, si no se aprende nada de interés, prestaremos atención a las recompensas. La respuesta de KOHN a la pregunta: "¿Motivan las recompensas a los estudiantes?" es: "Decididamente sí: motivan a los estudiantes para conseguir recompensas" (1994, pág. 3).

Del mismo modo que los elogios frenan la autorregulación y la autonomía del aprendiz, pueden empezar a distorsionar nuestra identidad como aprendices; un tema importante de este libro. Carol DWECK (1999) ha demostrado que quienes reciben alabanzas constantes probablemente atribuyan sus éxitos a su capacidad, en vez de a su esfuerzo. La consecuencia de esto es que los estudiantes "de sobresaliente" tienen que hacer todo lo que puedan para proteger su reputación. Esto puede incluir matricularse en asignaturas más fáciles (¿recuerdan a Ruth,

[30] Véase: HATTIE y TIMPERLEY (2007), pág. 97.

de la Introducción?) y evitar todo riesgo de fracaso. Se da entonces más importancia a las calificaciones que al aprendizaje y se evitan riesgos y contratiempos. DWECK ha demostrado también el impacto negativo que produce en el alumnado, sobre todo en ellas, el paso a centros universitarios en los que puede ser más difícil obtener éxitos. Esto les provoca dudas acerca de si tienen realmente la capacidad que pensaban que poseían por haber sido condicionados para ello. Quienes enfocaban el aprendizaje como una adquisición progresiva y basada en el esfuerzo demostraron tener mucho más aguante en estas circunstancias[31].

Este debate sobre el elogio tiene lugar en un contexto social angloparlante que hace hincapié en el "niño total" de un modo que resultaría extraño en otras culturas, sobre todo en aquellas (como Francia, Rusia, China) en las que los docentes consideran a menudo que su trabajo consiste en ayudar a los estudiantes a aprender en vez de ocuparse de su bienestar general. Esas declaraciones generales acerca de un país son potencialmente engañosas, dado que muchos maestros y profesores irán más allá de lo delimitado estrictamente por esa función docente; no obstante, la intención es, simplemente, mostrar que puede haber formas diferentes de expresar confianza en culturas educativas menos "personales". En Rusia, por ejemplo, Robin ALEXANDER ha observado que solo hay un puñado de manifestaciones de elogio, mientras que "el vocabulario de desaprobación es rico y variado" (2000, pág. 375). Sin embargo, la investigación ha demostrado que los estudiantes están dispuestos a pedir salir a la pizarra cuando no han entendido algo, por lo que profesores y alumnos pueden seguir el trabajo y corregirlo[32]. Paradójicamente, esto no ocurre tan a menudo en nuestras aulas "acogedoras". Si se percibe que los maestros y profesores quieren ayudar a los estudiantes a aprender, esto mismo puede promover un enfoque saludable del aprendizaje. La tarea del estudiante consiste en dominar la materia, ser un "buen estudiante" más que un "buen chico", y las dificultades hay que reconocerlas en la búsqueda de la comprensión.

Para los profesionales, este debate puede tener cierto aire purista. Por nuestra propia experiencia, sabemos que necesitamos cierto reconocimiento de que nuestro trabajo es aceptable (o, mejor aún, "bueno") antes de que podamos ocuparnos de lo que haya que mejorar. Podemos necesitar elogios para seguir en la brecha, por lo que presta una función autorreguladora. En la práctica, es difícil separar con facilidad la retroinformación centrada en la "tarea" y la centrada en la "persona", dado que la retroinformación basada en la tarea puede percibirse como elogio o como crítica. Como observa Michael ERAUT, "aunque quien facilite la retroinformación insista en que el objeto de ésta es la acción o el rendimiento, muchos receptores lo interpretan como un comentario sobre su persona. Así, los mensajes con intención orientadora pueden interpretarse como juicios" (2007, pág. 1). Sus pruebas tomadas de los lugares de trabajo indican que, aunque los aprendices quieran mejorar en lo que consideran como atributos importantes, sopesan la retroinformación en términos de equilibrio entre adquisición de infor-

[31] Creo que simplifica excesivamente estas distinciones, pues su trabajo se basa con frecuencia en los extremos de cada grupo, ignorando la mayoría situada en el medio, que trabaja con múltiples objetivos; véase: YOUNG (2007).

[32] Para hacerse una idea de la dinámica del aula francesa, véanse: HUFTON y ELLIOT (2001), y también RAVEAUD (2004).

mación a largo plazo y coste emocional inmediato[33]. Si la retroinformación se refiere a algo que se considera poco importante, puede ignorarse a causa de su coste emocional o prestarse atención únicamente a la retroinformación sobre las virtudes o puntos fuertes.

Por tanto, para que sea útil, el elogio debe dirigir la atención a la tarea, no a la persona. Esto queda muy bien recogido en el enfoque de la retroinformación de "las dos estrellas y un deseo", de Shirley CLARKE. Este enfoque de la corrección y la calificación implica identificar dos ejemplos que satisfagan los criterios de éxito ("dos estrellas") y seleccionar después un elemento de retroinformación ("un deseo") para mejorar el trabajo. Hay, por tanto, elogio, pero de un modo que probablemente estimula una mayor atención a la tarea. Este enfoque refuerza la insistencia de CLARKE en ofrecer solo retroinformación restringida; probablemente, todos hayamos tenido experiencia de la "retroinformación asesina", que es tan exhaustiva que nos lleva a abandonar la tarea. Insiste también la autora en dejar tiempo para que el aprendiz realice algo con la retroinformación, en vez de pasar a la siguiente unidad de trabajo y dejarlo sin tratar. Su enfoque consiste, pues, en cumplir otra *doble tarea*: ofrecer estímulo y retroinformación al nivel de la tarea o del proceso.

El problema de las notas y las calificaciones

En el clásico de culto *Zen and the Art of Motorcycle Maintenance*, de Robert PIRSIG (1974)*, se nos presenta a un profesor que se niega a poner notas a sus estudiantes y solo les hace comentarios. Pasadas unas semanas, "algunos de los estudiantes de sobresaliente comenzaron a ponerse nerviosos y su magnífico trabajo empezó a empeorar" (pág. 202), mientras que los de notable y suficiente empezaron a mejorar la calidad de sus trabajos y los de suspenso y futuros fracasados escolares empezaron a "ir a clase para ver qué ocurría". En las semanas finales del trimestre, cuando, por regla general, la mayoría saben su calificación y "se quedan cruzados de brazos, medio dormidos", los estudiantes se habían reunido en un amistoso diálogo libre "que hacía que la clase pareciese una fiesta". PIRSIG concluye:

> En realidad, las notas encubren el fracaso de la enseñanza. Un mal profesor puede pasarse todo un trimestre sin dejar nada memorable en las mentes de sus alumnos, representar la curva de las puntuaciones en una prueba irrelevante y dar la impresión de que unos han aprendido y otros no. Pero, si se eliminan las notas, los alumnos se ven obligados a preguntarse a diario qué se está estudiando. Las preguntas: ¿qué se está enseñando?, ¿cuál es el objetivo?, ¿cómo cumplen ese objetivo las clases y las tareas?, no auguran nada bueno. La eliminación de las notas expone a un enorme y espantoso vacío.
>
> (Pág. 204)[34].

[33] TROPE, Y. y cols. (2001).

* Hay traducción al castellano: *Zen y el arte del mantenimiento de la motocicleta* (traducido por: Esteban RIAMBAU SAURÍ). Barcelona: Random House Mondadori, 1999. (*N. del T.*)

[34] Agradezco a Martin FAUTLEY que descubriera y me facilitara esto.

Estas afirmaciones han resistido la prueba del tiempo, no solo con respecto a la influencia de las notas, sino al "pánico moral" que emerge si se reduce su uso. Cuando Paul BLACK y Dylan WILIAM lanzaron su folleto *Inside the Black Box*, que defiende el uso de comentarios, en vez de notas, en la evaluación en el aula, la respuesta de la prensa encerraba algo más que un tufillo de este pánico. *The Times* le dedicó incluso su principal editorial: *"TWO OUT OF TEN for educationalists who want the world to be a different place"**. Truena el artículo:

> Cualquier padre o madre sabe que los hijos progresan con los premios y, a veces, con el castigo... Desde la escuela infantil en adelante, los niños y niñas llegan entusiasmados a casa cuando han recibido una estrella dorada o una pegatina por su buen trabajo... los especialistas en educación como Paul BLACK... parecen decididos a aislar a los alumnos de la realidad. A los niños y niñas —dice—, no debe dárseles notas sobre diez ni estrellas doradas, porque después "buscarán las formas de obtener las mejores notas, en vez de lo que tienen que aprender"... Sin embargo, aprender a buscar las formas de poder ganar las mejores notas es una de las competencias más útiles que pueden adquirirse en la escuela. Gran parte de la vida tiene relación con aprovechar un sistema para obtener el máximo beneficio. Los alumnos que sepan cómo maximizar sus resultados en los exámenes estarán preparados para el mundo laboral.
>
> (*The Times*, 6 de junio de 1998, pág. 21.)

Aunque esta visión de la escuela y del trabajo no me parezca precisamente estimulante (ya sabemos lo que hace falta para aparecer en *The Times*), se encierran aquí preocupaciones justificadas que la EpA no ha tratado siempre de forma convincente. Una de ellas es el papel de las notas y de las calificaciones, dado que éstas son expresiones profundamente culturales que son la moneda de cambio de la mayoría de los sistemas educativos. Aunque la lógica sea sencilla, es decir, que las notas y calificaciones no transmiten suficiente información para hacer avanzar el aprendizaje, las consecuencias de ellas no lo son. Entonces, ¿por qué no tener notas y comentarios? Porque las pruebas indican que los comentarios son en gran medida ignorados; lo que importa son las notas.

La evidencia más directa de esto procede del trabajo experimental de Ruth BUTLER, que manipuló las condiciones de retroinformación para examinar la influencia de las notas y los comentarios sobre el aprendizaje. El descubrimiento clave fue que la condición combinada "notas y retroinformación" mostró poco más aprendizaje que la condición "solo notas", mientras que la retroinformación "solo comentario" logró un incremento del aprendizaje significativamente mayor. Intuitivamente, podríamos haber previsto que las notas con comentarios se acercaran más al rendimiento de "solo comentario", pues da una idea de hasta qué punto los alumnos están haciendo bien su trabajo y qué pueden hacer para mejorarlo. El enfoque de KLUGER y DENISI puede ayudar a explicar los hallazgos, es decir, si tra-

* "DOS SOBRE 10 para los especialistas en educación que quieren que el mundo sea un lugar diferente". (*N. del T.*)

tamos las notas como elementos esencialmente autorreferenciados. Mi 7 sobre 10 necesita una respuesta "reputacional": ¿hasta qué punto concuerda con mi idea de mí mismo y cómo queda en comparación con las puntuaciones de mis amigos y de mis rivales? La atención se dirige, pues, a la persona más que a la tarea y solo puede prestarse una atención limitada a lo que los comentarios indiquen que pueda hacerse.

Por parte del profesor, el uso de notas o calificaciones magnifica a menudo el problema de la retroinformación, cuando las notas van acompañadas de comentarios "vacíos". Como se pone una nota, se considera a menudo que la evaluación es sumativa, de manera que, con frecuencia, los comentarios adoptan la forma de juicios u observaciones generales que, de por sí, no prestan mucha ayuda a la mejora. Una investigación hecha en Inglaterra examinó los comentarios de corrección de los maestros en los trabajos de alumnos de 11 años durante un período de 7 meses, en 12 asignaturas[35]. Cuando se analizaron los comentarios dirigidos al alumno, más del 40% de los 114 comentarios escritos eran elogios sin retroinformación. Otro 25% era comentarios sobre la presentación: "no condenses tanto el texto"; "¡cuidado con la ortografía!"; "muy bien; escribe siempre con bolígrafo". El carácter muy generalizado de la retroinformación hacía imposible determinar siquiera las materias con las que ésta se relacionaba. En menos de la cuarta parte de los casos había retroinformación específica al nivel de la tarea o del proceso, por ejemplo: "¿Por qué es esto? ¿Qué crees que has aprendido realmente?"; "¿por qué haces dos pruebas diferentes?"; "¿qué aspectos de su personaje te gustan?"

El uso que hace el aprendiz de la retroinformación

Gran parte de este comentario ha tratado la retroinformación como un "regalo" del profesor. Sin embargo, el alumno tiene varias opciones relativas a lo que pueda hacer con ella. Deborah BUTLER y Philip WINNE resumen las opciones que tiene el aprendiz que la recibe y las diversas formas que puede adoptar:

> La retroinformación es la información con la que el aprendiz puede confirmar, añadir, reemplazar, perfeccionar o reestructurar la información en la memoria, con independencia de que la información se refiera a los conocimientos de un campo, conocimientos metacognitivos, creencias sobre el yo y las tareas o tácticas y estrategias cognitivas.
>
> (1995, pág. 275.)

La retroinformación puede negociarse, aceptarse y utilizarse de una o más de esas formas para ayudar a hacer avanzar el aprendizaje. No obstante, el coste emocional y de esfuerzo de actuar sobre la base de esa información puede ser excesivo, sobre todo si el compromiso con ella es bajo. Puede ocurrir que el aprendiz pueda modificar o desdibujar el objetivo, cambiarlo o rechazar la retroin-

[35] BATES, R. y MOLLER BOLLER, J. (2000): *Feedback to a Year 7 pupil*. Informe no publicado. Stockport Education Authority.

formación. Muchos de nosotros habremos cursado asignaturas que empezamos con la intención de destacarnos y después, tras alguna retroinformación (y contratiempo) inicial, decidimos que lo mejor que podíamos hacer era terminarlas y aprobarlas sin más. Esto podría llevarnos a pensar que la asignatura ya no tenía interés en relación con nuestras necesidades, antes de declarar que estaba tan mal organizada y enseñada que íbamos a dejarla.

Esto destaca la importancia de desarrollar la *autorregulación*, dado que la combinación de autovaloración y autocontrol estimulará tanto la evaluación como la perseverancia. Por eso, las intenciones del aprendizaje y los criterios de éxito son importantes en la EpA, pues permiten a los alumnos relacionar su actuación con sus objetivos y hacer ajustes en sus esfuerzos, dirección y estrategias cuando sea necesario. No obstante, esto supone un compromiso con estos objetivos de aprendizaje, un primer paso que, con frecuencia, se subestima en el aula. Esto nos devuelve a la negociación, en vez de al anuncio, de los objetivos del aprendizaje. Del mismo modo, la insistencia en la autoevaluación y en la evaluación a cargo de compañeros forma parte de este repertorio de autorregulación, pues los aprendices desarrollan sus propias destrezas de detección de errores y se hacen más partidarios de buscar y aceptar la retroinformación de los demás.

Conclusión

Evaluación para el aprendizaje: posibilidades y dificultades

El argumento de este libro es que la evaluación es una actividad esencialmente social que configura tanto la identidad del alumno como el tipo de aprendizaje que tiene lugar. La "Evaluación para el Aprendizaje" presenta una forma positiva de avanzar en ambas cosas, al destacar lo situacional y centrarse en la comprensión y la mejora del aprendizaje. En el centro de ello está la visión de unos aprendices activos y autorregulados que trabajan para dar sentido a lo que están aprendiendo y que han recibido los medios para hacerse cada vez más expertos para evaluar su propio trabajo. Resalta la base colaborativa del aprendizaje merced a unas intenciones de aprendizaje y unos criterios de éxito compartidos. La evaluación a cargo de compañeros y la autoevaluación desempeñan un papel clave en esta autorregulación, al igual que la retroinformación.

En la práctica, no es fácil alcanzar esa visión, sobre todo cuando el currículum es demasiado detallado y rígido, y la evaluación está dominada por la rendición de cuentas con consecuencias importantes. Esto puede llevar con facilidad a que la EpA se convierta en una serie de técnicas mediante las que mejorar las notas, es decir, existe el riesgo de quedarse en una simple *conformidad con los criterios*, en vez de llegar al aprendizaje productivo. Del mismo modo, la autoevaluación y la evaluación a cargo de compañeros puede quedar reducida a la "corrección de los trabajos" mientras el profesor aporta las respuestas. Una de las restricciones clave en la práctica actual es la limitada comprensión de la *retroinformación*, el medio clave para hacer progresar el aprendizaje. Aunque los profesionales dicen que "ya lo hacemos", gran parte de la práctica actual de la retroinformación no hace avanzar el aprendizaje, en especial con los elogios y las

calificaciones, dos de las monedas de cambio de gran parte de la educación angloparlante.

La idea de la *doble tarea* facilita una forma útil de salvar estas fuertes tensiones. ¿Nuestra evaluación formativa está ayudando al aprendizaje en la situación presente *y* desarrollando las destrezas de autorregulación para los aprendizajes futuros? Si solo se realiza una tarea, es insuficiente: habremos olvidado el aquí y ahora por el mañana, o quizá no hayamos dominado los conocimientos o las destrezas en primer lugar. Lo mismo cabe decir con respecto a la retroinformación: ¿nos ayuda con la tarea que tenemos entre manos *y* nos prepara para abordar las tareas futuras de modo más eficaz? La "Evaluación para el Aprendizaje" encierra un rico potencial pero, tal como es, tiene que aclarar más qué implican algunos de sus conceptos clave.

CAPÍTULO VIII

Recuperar la evaluación: Responsabilizarnos de quienes somos[1]

> En un mundo en el que los seres humanos se encuentran cada vez más carentes de normas bien definidas, de apoyo comunitario y de metas colectivas, se hace cada vez más necesario encontrar formas de ayudarles a ser capaces de definirse como individuos y a conseguir controlar su propio aprendizaje y sus carreras profesionales.
>
> (Patricia BROADFOOT y Paul BLACK, 2004.)

> La auténtica dificultad para cambiar cualquier empresa no radica en elaborar ideas nuevas, sino en escapar de las antiguas.
>
> (John Maynard KEYNES.)

> A veces, hay que retirar del idioma una expresión y mandarla a limpiar; después, puede volver a ponerse en circulación nuevamente.
>
> (Ludwig WITTGENSTEIN.)

Vivimos en tiempos de pruebas y este libro ha sido crítico con respecto a muchas prácticas actuales de evaluación. Esto no significa que podamos prescindir de la evaluación; la necesitamos para hacer juicios sobre los alumnos y para ayudarles a que aprendan. Lo que hace falta es una visión más clara del potencial y de las limitaciones de la evaluación; qué puede esperarse que haga, y cuándo se utiliza mal. Este capítulo intenta abordar esto; la intención es recuperar la evaluación limitando su poder y promoviendo los tipos de evaluación que puedan mejorar la calidad del aprendizaje. Parte de esta recuperación consiste en retirar para su limpieza algunas expresiones clave; entre las que más necesitan que las frotemos bien para sacar a la luz lo que hay debajo, están *capacidad* e *inteligencia*.

[1] La expresión *becoming responsible for who we are* ("responsabilizarnos de quienes somos") procede del título de la crítica que hace David OLSON de Howard GARDNER: "Becoming Responsible or Who We Are: the Trouble With Traits", en J. SCHALER (ed.): *Howard Gardner Under Fire* (2006). Chicago: Open Court.

Empezaba este libro con la afirmación de Allan HANSON de que "en la sociedad contemporánea, los tests no describen al individuo, sino que, más bien, lo construyen" (véase pág. 16). A lo largo de esta, he intentado demostrar el poder de la evaluación para configurar no solo cómo nos vemos a nosotros mismos, sino también cómo y por qué aprendemos. Ian HACKING, en su análisis de nuestra forma de "caracterizar a las personas", presentaba la *recuperación de nuestra identidad* como el motor último del descubrimiento: el momento en el que quienes son medidos y categorizados empiezan a oponerse a lo que otros les están haciendo. Parte del proceso de recuperación consiste en *responsabilizarnos de quienes somos*. Se trata de cuestionar las denominaciones que otros quieran adjudicarnos. Este cuestionamiento puede hacerse mediante un diálogo constructivo; también oponiéndonos a una clasificación.

Hemos visto que la administración de tests de inteligencia, las "inteligencias múltiples", la "inteligencia emocional" y los "estilos de aprendizaje" nos hablan de quiénes somos: nuestros atributos intelectuales o emocionales; cómo aprendemos; la clase de personas que somos. En este capítulo me opondré a la idea de que una instantánea de ese tipo, a menudo mal encuadrada y borrosa, pueda ofrecer una representación permanente de quienes somos y de nuestro potencial. Nunca permitiríamos que nuestro primer examen de conducir determinara nuestra categoría como conductores para el resto de nuestra vida. Sin embargo, esto es lo que siguen haciendo los tests de CI y de capacidad; las "inteligencias múltiples" y la "inteligencia emocional" presentan versiones más benévolas de un proceso similar de moldeado. Así, un aprendiz cinestésico, con un CI medio pero con un elevado CE, y todo sobre la base de un test y dos inventarios rellenados por mí, teniendo poco o nada en cuenta el contexto en el que los hice, sería una persona diferente en otro lugar.

El programa de recuperación

Este título puede generar imágenes de recuperación de la tierra tras años de contaminación industrial o de recuperación de una playa tras un repugnante vertido de petróleo. No me disgusta este tipo de imagen. Gran parte de lo que he escrito en este libro versa sobre los excesos de la evaluación, más allá de sus legítimas funciones, a menudo con consecuencias tóxicas para las identidades individuales, la enseñanza y el aprendizaje. La recuperación no es un proceso glamuroso; sugiere una *zona abandonada*, en vez de un campo verde o un cielo azul. Esa es la fuerza de la observación de KEYNES, al principio del capítulo: la dificultad está en eliminar las antiguas ideas antes de que podamos desarrollar las nuevas.

La recuperación es un proceso constructivo. Con respecto a la evaluación, supone prescindir de las afirmaciones y supuestos falsos con el fin de volver a sus fines legítimos. Este programa de recuperación se organiza en cinco elementos principales; los dos primeros reclaman un papel más modesto para la evaluación y unas interpretaciones más cautas de los resultados. El tercer elemento es más expansivo: hay que prestar más atención al contexto en el que tienen lugar las evaluaciones. Ligada a éste está la importancia de comprender las interaccio-

nes, tanto técnicas como sociales, que se producen en la evaluación. El elemento final considera cómo puede desarrollarse una *evaluación sostenible*, que dé al aprendiz las destrezas de evaluación que puedan equiparlo para un futuro ignoto.

Paso 1: Limitar las ambiciones de la evaluación; centrarse en el rendimiento

El primer paso para la recuperación de la evaluación consiste en fijar unas ambiciones más realistas para lo que podamos hacer. Esencialmente, la evaluación es una mirada atrás; ofrece un informe de lo que ya se ha aprendido. Esto es una actividad social y responderá a los valores y conceptos de esta sociedad; debemos rechazar la idea grandiosa de que puede mantenerse al margen de la sociedad y hacer juicios objetivos e independientes de las culturas. La evaluación es, simplemente, una parte de las actividades educativas y sociales más generales y comparte los mismos valores y limitaciones de éstas. La mejor manera de considerar la evaluación es como un elemento de una iniciativa más amplia; los problemas aparecen cuando domina esta iniciativa, haciendo afirmaciones que no puede respaldar.

Aunque la evaluación es una mirada atrás, esto no significa que carezca de una función predictiva. Sin embargo, el uso de la información de la evaluación para prever el rendimiento futuro ha de interpretarse con sumo cuidado. Si, hasta ahora, he aprendido bien, es posible que continúe así; si he tenido problemas, es posible que siga haciendo progresos limitados. Pero esto puede cambiar a causa de cambios en cuanto a mi compromiso, mi esfuerzo y la forma en que me ayuden a aprender.

Tener ambiciones modestas con respecto a la evaluación significa que su función principal es hacer sondeos acerca de la situación de una persona en relación con su aprendizaje. Estos sondeos acarrean consecuencias cuando se utilizan con fines de selección con implicaciones importantes. Tenemos que estar alerta ante un posible "proceso de fabricación", como lo llama HANSON, en el que "la probabilidad de que una persona sea capaz de hacer algo, como lo determina el test, sea más importante que hacerlo realmente" (pág. 288). Esto nos lleva a considerar la "adecuación a la finalidad" de esas pruebas de importancia decisiva (véase más adelante).

Replantear la inteligencia y la capacidad

Uno de los usos más destructivos de la evaluación ha sido la aseveración de que puede trascender la medida del rendimiento y facilitar una medida de la capacidad subyacente que ha conducido a ese rendimiento. Este supuesto ha otorgado su poder social a la administración de pruebas de inteligencia y de capacidad: he aquí algo que no solo mide lo que se sabe, sino que puede prever lo que será posible y lo que no. Para los primeros examinadores de CI, las puntuaciones de los tests de inteligencia predecían incluso la calidad moral de una persona, dado que vinculaban la capacidad a la calidad de los juicios morales y sociales que

pueden esperarse. El débil mental no era digno de confianza; los intelectos superiores, sí.

En el Capítulo II, intenté demostrar que estas afirmaciones carecen de justificación: los tests de capacidad son, en esencia, tests generalizados de rendimiento. Se basan en lo que ya sabemos, tanto en términos de conocimientos como en la forma de enfocar los problemas. Este fue el punto de partida de BINET y por eso creía que podíamos mejorar la inteligencia de los niños. Esta idea se perdió cuando se interpretó el CI, de acuerdo con las creencias sociales de los examinadores del CI, como una capacidad innata que llevamos con nosotros a la educación, sobre la que la escuela produce poco o ningún efecto. La evaluación desempeñó un papel clave para disimular estos supuestos: los tests de inteligencia se consideraban científicos y estadísticamente objetivos. Estas estadísticas podrían utilizarse para ordenar a las personas según su inteligencia y clasificarlas, y estas clasificaciones configurarían las identidades de millones de personas, particularmente a través de la selección en la educación y para la selección laboral. También determinaban el tipo de experiencias de aprendizaje que disfrutarían las distintas personas de capacidades diferentes, contribuyendo así a crear las mismas disparidades previstas por los tests.

Aunque se diga que el CI es, en gran medida, cosa del pasado, no puede decirse lo mismo de los tests de capacidad, que operan como una versión *no reconocida* de la "inteligencia" y del "CI". Los tests de capacidad hacen suyas muchas de las creencias que se plasmaban en los tests de CI, aunque a menudo esto se acepta de manera acrítica. Medimos la capacidad y después inferimos que esta es la causa del éxito o del fracaso, en vez de considerarlo como una medida de rendimiento que puede ayudar a prever el rendimiento futuro. La observación de WITTGENSTEIN que aparece al principio del capítulo encaja bien aquí: la *capacidad*, con las inferencias al uso acerca de su naturaleza causal y fija, infecta gravemente gran parte del pensamiento educativo y político actual. Desde luego, me encantaría ver eliminado el término del vocabulario educativo y despojado de sus connotaciones causales. No es tarea fácil, dada su asociación con las pruebas de inteligencia.

No todo el mundo ha hecho estas inferencias causales; ha existido una postura histórica desde BINET en adelante que no ha tratado la inteligencia como algo fijo y ha reconocido el papel fundamental de la experiencia en la configuración de la inteligencia. Otros han rechazado la limitada visión de la inteligencia que la restringe en gran medida a un único factor general: *g*. Entre ellos está Howard GARDNER, con sus *Inteligencias Múltiples*. El análisis del Capítulo III señalaba que también él da por supuestas unas capacidades innatas, aunque, al haber ocho, más o menos, era posible que cada persona posea un patrón característico. Aunque es una visión más positiva que el "pesimismo brutal" de los examinadores del CI, todavía arrastra un toque de determinismo biológico.

Por tanto, el primer paso del programa de recuperación consiste en ser más modesto. La evaluación depende de lo ocurrido antes; es un elemento de una iniciativa mayor, y es el producto de unos valores sociales. Decir que se sitúa fuera de esos procesos y puede hacer predicciones con independencia de lo ocurrido antes es desmesurado.

Paso 2: Interpretar los resultados con más cautela

Que una evaluación sea o no válida no solo depende de lo bien que mida lo que se somete a examen, sino también de las interpretaciones que se hagan de los resultados. Un test bien construido se invalida si no se entienden o se malinterpretan los resultados. Muchos de los abusos de la evaluación surgen de lo que se lee en los resultados y de las consecuencias de ello. En el contexto de unas ambiciones más limitadas, tenemos que hacer unas inferencias más cautas.

Nos lleva esto de nuevo a nuestras tres cuestiones básicas sobre la evaluación: ¿cuál es la finalidad principal de esta evaluación?; ¿la forma de la evaluación es adecuada a su finalidad?; ¿consigue su finalidad? Si estas preguntas se plantearan con más rigor ante las evaluaciones, tendríamos más cuidado con lo que afirmamos. Por ejemplo, si la finalidad es comprobar la lectura, tendremos que determinar qué entendemos por "lectura". ¿Nos referimos a pronunciar correctamente las palabras o a leer en silencio y comprendiendo lo que leemos?[2] Si no queda claro lo que entendemos por "leer" (o "lenguaje", "inteligencia" o "ciencia"), nuestras evaluaciones carecerán de validez y, en consecuencia, de "adecuación a la finalidad", dado que no está claro lo que medimos. Por tanto, aunque quiera restringir el *alcance* de la evaluación, también querría ser mucho más estricto con respecto a la finalidad de la evaluación y su "adecuación a la finalidad".

Aunque solucionemos las cuestiones de la finalidad y de la adecuación a la finalidad, todavía queda la del modo de interpretar los resultados. Muchos de los usos erróneos de la evaluación que hemos visto en este libro conllevan interpretaciones erróneas de los resultados. Estas pueden corresponder a inferencias excesivas, a confiar en exceso en resultados poco fiables o a hacer interpretaciones simplistas de los resultados.

Interpretaciones abusivas de los resultados

Ya hemos visto que las inferencias a parir de los resultados de los tests pueden ir mucho más allá de los datos. Los tests de CI han utilizado una única puntuación total para hacer inferencias acerca de la capacidad biológica de aprender. Aunque la evaluación de las "inteligencias múltiples" se acerca más al mundo real, todavía se utiliza para estimar disposiciones biológicas innatas. Se trata de saltos lógicos que no tenemos porqué aceptar. En la interpretación, no tenemos que ir más allá de lo que nos dice acerca de nuestro funcionamiento real. La biología y el destino son interpretaciones abusivas inadecuadas.

[2] Durante muchos años, pudo encontrarse con frecuencia en las escuelas el *Schonell Graded Words Test*. Este consistía en pronunciar correctamente cadenas de palabras aisladas de dificultad creciente (e irregularidad fonética). La "edad lectora" se determinaba por el número de palabras que se pronunciaran correctamente. En Inglaterra, la lectura se mide en la actualidad mediante un test de lectura del currículum nacional basado en la lectura silenciosa de pasajes, seguida de preguntas de comprensión e inferencia. Este es un argumento de validez que implica tanto el constructo de lectura como la mejor manera de medirla: adecuación a la finalidad. Véase: SAINSBURY y cols., 2006.

Interpretaciones poco fiables

Aunque todas las evaluaciones puedan tener consecuencias importantes, esto no significa que sean automáticamente dignas de confianza. Esta carencia de fiabilidad puede deberse a una falta de validez de constructo: es posible que no midan en realidad lo que dicen medir. Consideramos que las evaluaciones de la "inteligencia emocional" tienen una validez limitada porque no estaba claro lo que medían. Como los instrumentos de evaluación medían una mezcla de conceptos débilmente relacionados, esto generaba problemas de fiabilidad interna[3]. La conclusión de MATTHEWS y sus colaboradores con respecto a las pruebas de "inteligencia emocional" era que estaban "abiertas a demasiadas interpretaciones para tener utilidad práctica". No podemos interpretar con confianza una baja puntuación en el test como indicador de una falta fundamental de competencia y no podemos dar por supuesto que una elevación de las puntuaciones en el test represente una adquisición de competencia" (2002, pág. 540). Sin embargo, basándose en esos tests, se extraen toda clase de conclusiones y se adjudican todo tipo de calificaciones.

En relación con los "estilos de aprendizaje", vimos unos complejos marcos interpretativos aplicados a autoinventarios poco consistentes. En el caso del *Learning Style Inventory*, de DUNN y DUNN, el inventario de autoinforme generaba 22 factores de limitada fiabilidad, que aún se complicaba más con la variación de los autoinformes al cambiar tiempos y lugares de aplicación. En el caso de los cuadrantes de aprendizaje de KOLB, las inferencias se extraen de las respuestas a 12 preguntas que generan cuatro estilos de aprendizaje en dos dimensiones. Es una base frágil para concluir a partir de ella que una persona sea convergente o divergente. Sin embargo, los niños, y los adultos, saben cuál es su cuadrante y ese conocimiento los moldea.

Esto no significa necesariamente que estas evaluaciones carezcan de valor pero, en el mejor de los casos, solo pueden ser un estímulo para un diálogo relativo a nuestra forma de aprender. Esta era la intención de ENTWISTLE con sus enfoques del aprendizaje, aunque ya vimos con qué facilidad puede cristalizarse un enfoque en una disposición. Responsabilizarnos de quienes somos significa cuestionar cualquier calificación de "inteligencia emocional" o de "estilos de aprendizaje" que quieran adjudicarnos.

Interpretaciones excesivamente simplistas

Vimos que, en las culturas de rendición de cuentas, como las del Reino Unido y los EE.UU., se utilizan como objetivos unos indicadores simples y restrin-

[3] Si una prueba tiene un número de componentes que no se correlacionan mucho entre sí, surgirán problemas de fiabilidad: una buena puntuación en un componente puede quedar compensada por la baja puntuación en otro. En consecuencia, la puntuación global es difícil de interpretar, dado que no resume eficazmente mi actuación. Si me desenvuelvo bien en unos componentes y mal en otros, puedo obtener una puntuación global media sin tener una puntuación media en ningún componente. Normalmente, se prevé que los tests presenten coherencia interna y que sus componentes tengan una buena correlación mutua, es decir, si hago bien un componente, probablemente haga bien también los demás. De este modo, mi puntuación global debe reflejar mi actuación en el test.

gidos. En educación, se otorga una importancia desproporcionada a los resultados de las pruebas porque facilitan unos indicadores numéricos directos que se consideran medidas externas de rendimiento escolar. En el Capítulo VI, examinamos con cierto detalle cómo podían dar esas medidas restringidas una imagen deformada de la mejora. Se debe esto a que los objetivos, los resultados mejorados de las pruebas, se convierten en fines en sí mismos, por lo que se hacen todos los esfuerzos posibles para alcanzarlos. Esto puede motivar a las escuelas para mejorar: un efecto positivo. También puede llevar a aprovecharse del sistema de un modo que poco tiene que ver con el aprendizaje, por ejemplo, aprovechando las opciones de ingreso y escogiendo asignaturas que maximicen los resultados, con independencia de su valor educativo. La consecuencia es la *inflación de puntuaciones*, en la que las puntuaciones obtenidas en tests de importancia decisiva mejoran de forma espectacular, aunque esa mejora no se corresponde con evaluaciones de menor trascendencia. Esto indica que las mejoras son más bien el resultado del perfeccionamiento de las técnicas concretas de realización de pruebas que de la mejora en la asignatura.

Los responsables de la política educativa tienen que mantener la sencillez de los procedimientos de rendición de cuentas, por lo que tienen que oponerse a unas interpretaciones más complicadas en las que unos resultados mejores sean también consecuencia de una técnica más depurada de realización de exámenes. Para ellos, las puntuaciones de las pruebas se convierten en normas de nivel, en estándares, de manera que unos resultados mejores significan unos niveles más altos. Si los resultados no mejoran, y sabemos que, con frecuencia, se estabilizan pasados cuatro años, les asalta la histeria política. Esto supone la presentación de más iniciativas para dar un nuevo impulso a las puntuaciones; así, por ejemplo, en Inglaterra, el Gobierno ha introducido recientemente la enseñanza obligatoria de la fonética sintética como una nueva panacea para mejorar las puntuaciones en Lectura.

De esos resultados deberíamos extraer la cauta conclusión de que demuestran muy poco acerca de unos estándares subyacentes. Para evaluar éstos, necesitaríamos medidas más sofisticadas, como unas encuestas nacionales de rendimiento, basadas en una muestra representativa de estudiantes, es decir, encuestas que no provoquen las distorsiones de las pruebas de importancia decisiva. Este enfoque, junto con el uso de un conjunto más amplio de medidas de evaluación, contribuiría a una *rendición de cuentas inteligente*, un enfoque alternativo que presentamos en el Capítulo VI. La recuperación que pretendemos aquí intenta reducir la dependencia distorsionadora de unas medidas y unos objetivos de evaluación restringidos y de evitar una interpretación simplista de los resultados.

Paso 3: Reconocer el contexto

Mi enfoque hace considerable hincapié en el contexto cultural en el que operamos. En la recuperación de la evaluación, hay que tenerlo muy presente; si no se comprenden los elementos situacionales, nuestras interpretaciones siempre serán parciales.

Una limitación permanente de las evaluaciones estandarizadas es que nunca hacen del todo justicia a los factores situacionales. Esto se debe a que, por definición, tratan de presentar a todo el mundo las mismas pruebas de evaluación, el elusivo "nivel del terreno de juego". Siempre está presente el problema de que unos estudiantes habrán estado mejor preparados que otros y el de que el material puede partir de supuestos culturales compartidos por unos y no por otros. Se corre aquí el riesgo de interpretar lo "estandarizado" como "justo" (o sea, igual para todos), sin investigar si lo que se evalúa es o no injusto para algunas de las personas que se someten a la prueba. Como vimos en el Capítulo Primero, éste era el punto débil de la pasión victoriana por los exámenes: la justicia residía en someterse al mismo examen y ser juzgado por el resultado del mismo. Raramente se extendía a cuestionar a quién podía favorecer el examen por sus contenidos y requisitos. Así, los privilegiados lo hacían —y lo siguen haciendo— bien, y siguen afirmando que todo se debe al mérito; las ranas de Tawney croaban a pleno pulmón.

El Capítulo V se centró en estudiar cómo podía mitigarse esto. Rechacé la solución de Ronald DORE, consistente en sustituir los tests de CI por los de capacidad, basándome en que éstos se limitan a disfrazar el mismo problema: como tests generalizados de rendimiento, favorecen en exceso a quienes cuentan con capital cultural. En cambio, yo he buscado formas de conseguir que los procedimientos de prueba fuesen más abiertos, de manera que pudieran cuestionarse sus sesgos, unos procesos que ya estaban presentes en las evaluaciones para calificación, para nota. Tener en cuenta lo situacional significa ser mucho más consciente del contexto social de la evaluación; la justicia abarca mucho más que las oportunidades de acceder a evaluaciones estandarizadas.

El fundamento social del CI

Una de las hazañas de los primeros administradores angloparlantes de tests de CI consistió en persuadir a la gente de que el CI era *independiente de la cultura* (ahora, más modestamente, "poco dependiente de la cultura"). Como se trataba de una capacidad innata, se prescindía del contexto social. Los pobres, especialmente las minorías, heredaban unos CIs bajos y su pobreza era el resultado de ello. La explicación alternativa —ser pobre significa tener un acceso limitado a las experiencias sociales y educativas en las que se basan los tests— no concordaba con el espíritu de los tiempos. Los privilegiados eran capaces de demostrar que esto era producto de la capacidad, una defensa meritocrática completa de su posición.

En el Capítulo II, revisé los problemas de las reivindicaciones de independencia de la cultura. Una debilidad crítica era el *efecto* FLYNN, la manifestación de que las puntuaciones de CI han ido elevándose de forma constante de generación en generación. Como los cambios genéticos no pueden producirse a tal velocidad, tiene que haber una razón situacional. El análisis de FLYNN había demostrado que la mejora más rápida durante los últimos 60 años se produjo en los componentes más "independientes de la cultura" de los tests de inteligencia, por ejemplo, las "matrices progresivas" de Raven (véase la Figura 2.1), una prueba de razona-

miento abstracto. Su reciente resolución de esta paradoja pasa por el concepto de *multiplicadores sociales*. Un multiplicador social surge cuando un cambio social a gran escala lleva a rápidos cambios de comprensión, por ejemplo, la absorción de conceptos científicos más abstractos y el creciente valor social otorgado a "la resolución de problemas sobre la marcha sin un método previamente aprendido" (2006, pág. 8). Las abstracciones, por ejemplo, hallar elementos comunes en el subtest de Semejanzas y el razonamiento rápido con elementos poco habituales, son de las cosas que premian los tests de CI. Por otra parte, era más probable que nuestros abuelos establecieran vínculos funcionales (que puntúan menos) y buscaran soluciones basadas en reglas, que es lo que valoraba su sociedad.

El elemento anti-intuitivo de todo esto es que, a pesar de una mayor educación que la de nuestros abuelos, los componentes más evidentemente situacionales de los tests de CI, por ejemplo, Vocabulario y Aritmética, han mejorado menos. El argumento de FLYNN, que me parece persuasivo, es que este saber era tan importante para nuestros abuelos como para nosotros pero el efecto multiplicador ha sido menor porque donde ha faltado la educación en generaciones anteriores, ha habido unos efectos sociales multiplicadores más generalizados. Esto puede formar parte de la explicación de por qué algunos grupos minoritarios han mostrado unos incrementos aún más rápidos de las puntuaciones de CI: un mejor acceso a la educación ha llevado a mejoras en los componentes más "de aprender", así como en los más fluidos.

He revisado el efecto FLYNN del Capítulo II porque hace gran hincapié en ello: para comprender la evaluación, nunca puede prescindirse del contexto social. Debemos sospechar de cualquier evaluación que se presente como independiente de la cultura: la evaluación es una actividad cultural.

Situación más que disposición

Como en el caso de la inteligencia, una forma de infravalorar lo situacional es inferir, a partir de los resultados, unos rasgos personales independientes de la situación social. Los "estilos de aprendizaje" de DUNN y DUNN constituyen un ejemplo de lo que decimos: nuestros estilos de aprendizaje deben determinar la forma de enseñar y la tarea del docente es adaptarse para concordar con nuestras disposiciones para el aprendizaje. Yo sostengo que esto devalúa lo situacional: el aprendizaje de cada materia puede plantear exigencias diferentes y nosotros tenemos que aprender a adaptarnos a la situación. Otros teóricos de los estilos de aprendizaje, como KOLB y ENTWISTLE, se dieron cuenta de ello, pero tuvieron que esforzarse para impedir que las preferencias y los enfoques del aprendizaje se interpretaran como disposiciones para el aprendizaje. Por todo el mundo, tenemos ahora aprendices K, aprendices convergentes y profundos, cuando las respuestas a situaciones concretas se cristalizan en rasgos individuales.

Lo mismo ocurre también con las "inteligencias múltiples", donde, una vez más, nos encontramos con la interpretación del aprendizaje como una expresión de unas disposiciones individuales innatas, en vez de serlo de la situación. La crítica que David OLSON hace de las inteligencias múltiples, que puede aplicarse de

manera más generalizada, es que despojan nuestro aprendizaje de toda responsabilidad personal, dado que no somos responsables de nuestras disposiciones o capacidades:

> Adquirimos competencias cuando nos responsabilizamos de determinados estándares... No somos responsables de simples disposiciones; las disposiciones son causas de acciones y no razones para actuar. Una teoría de la educación tiene que explicar con detalle cómo asumen los niños las responsabilidades del aprendizaje y cómo una persona, sea docente o aprendiz, llega a juzgar si se cumple con esas responsabilidades... Los docentes están notoriamente dispuestos a explicar el éxito de los niños en relación con unos supuestos estilos de aprendizaje y capacidades, en vez de con las condiciones que hacen que el aprendizaje sea fácil o difícil.
>
> (2006, pág. 42.)

La recuperación de la evaluación supone un reconocimiento más completo de los factores situacionales. Que los estudiantes aprendan y cómo lo hagan son en gran medida productos del contexto y no de los genes. En parte, ser un aprendiz autorregulado supone aceptar las responsabilidad del aprendizaje, del mismo modo que los maestros y profesores tienen que responsabilizarse de crear un contexto que ayude a aprender. Hay que tener esto en cuenta tanto "en las alturas", al diseñar el currículum y la evaluación, como en los valores y actitudes de la escuela y del aula. Las etiquetas reducen la responsabilidad cuando dejamos que configuren quiénes somos como aprendices. Recuperar la evaluación significa asumir más responsabilidad con respecto a nuestro aprendizaje: "acción, intencionalidad y responsabilidad podrían convertirse en las características centrales de una psicología que tiene especial relevancia para la educación. Podemos dejar que las capacidades, los rasgos y las disposiciones encuentren un nuevo lugar en las ciencias naturales o relegarlos al cubo de la basura de la historia" (OLSON, 2006, pág. 43).

Paso 4: Reconocer la importancia de la interacción

Parte del carácter situacional de la evaluación se debe a la importancia de la interacción. En este libro, el término se ha utilizado de dos formas diferentes: técnica y pedagógicamente. Ambas comparten el reconocimiento de que hay que tener en cuenta el modo de combinarse los elementos. No juzgamos una receta culinaria basándonos solo en los ingredientes; es crucial la manera de combinarlos. Utilizamos técnicamente la interacción para señalar que la inteligencia no puede considerarse simplemente como un fenómeno aditivo, como la combinación fija de herencia y ambiente. Los coeficientes de heredabilidad son inestables porque varían cuando las predisposiciones genéticas interactúan con los cambios habidos en el ambiente. Cuando el ambiente es extremo, la herencia importa mucho menos. Por ejemplo, quienes están genéticamente predispuestos a la obesidad en épocas de abundancia, adelgazarán más que el promedio cuando haya escasez de alimentos.

El uso pedagógico se basaba en el papel central de la interacción social en la evaluación formativa eficaz. No es solo cuestión de lo que docente y aprendiz

aporten a la clase, sino de cómo interactúen. Por ejemplo, la retroinformación implica un complejo conjunto de interacciones; la misma retroinformación dada a dos alumnos diferentes tendrá efectos opuestos. Que un experto comprometido diga que una solución es errónea puede ser suficiente para desencadenar cambios de razonamiento a nivel de proceso y una aplicación intensificada; esta misma retroinformación facilitada a un principiante poco comprometido puede llevar a modificar o evitar la tarea, con el fin de reducir los costes emocionales para el aprendiz.

Recuperar la evaluación implica prestar más atención a estas interacciones. Esto puede parecer obvio, pero tanto los psicómetras como los docentes están tentados de minimizar la fuerza de la interacción. Los primeros desean unas ecuaciones más estables[4]; a los segundos puede incomodarlos que su aportación se considere dependiente del modo de recibirla, prefiriendo que sea juzgada independientemente de ello. Presento aquí dos ejemplos de la fuerza de la interacción, uno técnico y el otro pedagógico.

Multiplicadores sociales e individuales[5]

En la sección anterior, vimos que la interacción del cambio social con componentes del test de CI había llevado a un incremento constante de las puntuaciones de CI de una generación a la siguiente. Estos cambios sociales fueron considerados como "multiplicadores", pues diseminaban rápidamente nuevas competencias e ideas que, a su vez, conducían a mejores puntuaciones, en especial acerca de unos tests que se habían considerado tests inenseñables de razonamiento abstracto.

James FLYNN ha considerado también el hecho de que unas personas progresen más que otras en determinadas competencias. En este terreno, descubrió los *multiplicadores individuales*, en los que una ligera ventaja provoca una serie de interacciones que multiplican esa ventaja. El éxito deportivo le brinda ejemplos: de niño, una ligera ventaja en altura te lleva a introducirte en un grupo de baloncesto y, gracias a eso, juegas más y lo haces cada vez mejor. Esto te conduce a jugar en el equipo, por lo que puedes contar con entrenamientos regulares que te hacen que seas aún mejor. Esto te lleva a... y muy pronto eres un destacado jugador de baloncesto, un "talento innato". Yo, por mi parte, era 5 cm más bajo, no tuve multiplicadores y ahora participo ocasionalmente en el juego social.

Puede que sea un relato simplista, pero la lógica tiene sentido. Está en sintonía con la "explicación del talento" de Michael HOWE. Su argumento en contra de la llamada de Howard GARDNER a los "individuos excepcionales", incluyendo a niños prodigio, era que "han recibido casi siempre una ayuda y un estímulo con-

[4] Para el psicómetra, el objetivo consiste a menudo en reducir los efectos de las interacciones, pues estos corresponden al margen de error de una ecuación, parte de los elementos aleatorios que no pueden controlarse, reduciendo así la contribución de los efectos controlados.

[5] Tanto Stephen CECI como Robert STERNBERG han reconocido, aunque con expresiones diferentes, la importancia de estas interacciones sociales en el desarrollo de la inteligencia.

siderables antes de la época en la que su capacidad se ha considerado sobresaliente" (pág. 132). Continuando con ejemplos deportivos, Tiger Woods apareció en TV cuando tenía 3 años, para mostrar la fuerza de su golpe. Aunque, evidentemente, tenía una coordinación precoz, no la desarrolló sin que alguien pusiera en sus manos un palo de golf de tamaño infantil, enseñándole lo que debía hacer y llevándolo a la TV. Los multiplicadores aparecieron pronto y las interacciones fueron positivas, aprovechando tanto la motivación como las actitudes, con unos resultados espectaculares[6]. Me he detenido en el deporte porque parece que otros campos generan fuertes respuestas emocionales, por ejemplo, si se emplea esta lógica con respecto al niño prodigio Mozart (cuyo padre era músico) o a Picasso (cuyo padre fue director de una escuela de arte).

El impacto más significativo de esta lógica se observa en lo más prosaico. Si, a causa de las circunstancias de su hogar, los niños llegan a la escuela con vocabularios muy diferentes y si el vocabulario escolar es similar al de los hogares de clase media, los multiplicadores entrarán en acción para algunas personas desde el primer día: "lo entenderán". Si yo no pudiera ingresar en una *grammar school** por un punto de CI y usted ingresara por un punto, gracias a ese punto, entran en acción poderosos multiplicadores individuales: usted es claramente académico, yo no.

Es probable que alguien responda a esto diciendo que a otros chicos les dan palos de golf, raquetas de tenis y amplios vocabularios y, sin embargo, no destacan. Esto puede deberse a que, ante lo que el niño da de sí —unas reacciones físicas lentas serán un problema en el tenis, pero también lo será la calidad de las interacciones—, los padres agresivos no siempre proporcionan multiplicadores positivos.

Al recuperar la evaluación, hay que reconocer de forma más completa la fuerza de la interacción. Por ejemplo, si se evalúa a un niño como "superdotado", esa evaluación ha de interpretarse cuidadosamente. Ésta se refiere a unas capacidades desarrolladas, pero tiende a interpretarse como una capacidad natural, con su connotación de talento innato. Si se ofrecen a algunas personas otros multiplicadores añadidos mediante clases especiales, ¿cómo influirán en quienes sean considerados más torpes? ¿Qué multiplicadores encierra esto para el 90% incluido en esta categoría deficitaria? Es posible que estemos ante otro caso de expresiones que deben ser retiradas y depuradas. "¿Dónde está aquí el multiplicador?" podría ser una cantinela rara pero útil.

---

[6] Del mismo modo, Lewis Hamilton, el *niño prodigio* de la Fórmula 1 en 2007, apareció en T a los 5 años para demostrar sus proezas conduciendo un coche por radiocontrol. Andy Murray, la última promesa británica del tenis, es otro ejemplo. En un perfil suyo aparecido en el *Guardian* (*"Boy on the Brink"*), se decía que "Murray cogió por primera vez una raqueta de tenis cuando tenía dos años. Cuando cumplió tres, él y Jamie [su hermano] estrellaban pelotas contra la casa, hasta el punto de que las ventanas y el papel de las paredes estaban permanentemente manchados con señales de pelotazos... La familia de Murray estaba loca por los deportes. Ella (su madre, que ahora es entrenadora de tenis) pensaba que era inevitable que sus hijos acabaran siendo deportistas, quizá para toda la vida" (*Guardian*, 7 de junio de 2007, pág. 34). Ambos lo fueron.

* Tradicionalmente, en Inglaterra, las *grammar schools* eran centros de secundaria en los que se ingresaba tras unas pruebas selectivas de aptitud. En la actualidad, siguen existiendo algunas de carácter privado. (*N. del T.*)

Interacción en el aula

Hablamos de la importancia de la retroinformación en el aprendizaje en el Capítulo VII. La retroinformación eficaz tiene esencialmente que ver con la interacción eficaz, un logro difícil, pues hay muchas cosas que pueden obstaculizarla. Esta es solo una forma de interacción en el aula; planteé también la importancia de la negociación en torno a lo que haya que aprender y a lo que se considere como una actuación satisfactoria. El espíritu de la clase y la manera de promover la asunción de riesgos en el aprendizaje también forman parte de esto, así como las preguntas y el diálogo interesantes. El ejemplo siguiente de Herbert GINSBURG ilustra la importancia de trascender los métodos estandarizados y de estimular la interacción. A Becky, de 6 años, le preguntaron: "¿Cuántas son siete menos cuatro?" Su respuesta fue: "Dos". La interacción podría haberse interrumpido en este punto y el juicio correspondiente sería que la niña tenía una memoria limitada para los datos numéricos. Cuando le preguntaron: "¿Cómo has llegado a esa respuesta?", ella replicó: "Supe que siete menos cuatro son dos porque sé que cuatro más dos son siete y, si cuatro más dos son siete, siete menos dos deben ser cuatro". Otra vez, esto podría haberse interpretado como una nueva confusión. GINSBURG señala que, aunque hay un error en los datos numéricos, "el segundo ingrediente del guiso cognitivo era mucho más interesante que... el recuerdo erróneo. La niña presentó la idea correcta de que, si 4 + 2 = 7, *debe ser cierto* que 7 – 4 = 2... un silogismo clásico" (1997, págs. 14-15). Aquí pueden llevarnos las interacciones más ricas; Becky comprende más de lo que hubiéramos supuesto.

Paso 5: Crear una evaluación sostenible

El concepto de *evaluación sostenible*, de David BOUD, presentado en el Capítulo VII, concuerda con la imagen de la recuperación. Hecho el trabajo de limpieza, ahora tenemos que trasplantar de un modo que conduzca a un crecimiento sostenible. La idea es que "cualquier acto de evaluación debe contribuir también de alguna manera al aprendizaje, trascendiendo la tarea inmediata... una evaluación que satisfaga las necesidades del presente y prepare a los estudiantes para satisfacer sus necesidades futuras" (2000, págs. 8-9). Esta es la *doble tarea* de la evaluación: una evaluación en el momento presente que deje a los estudiantes mejor preparados para la tarea siguiente.

La consecuencia de esto es que la evaluación nunca tiene una única función. Podemos decir que la evaluación está relacionada con el juicio sobre los resultados del aprendizaje en una determinada asignatura, pero nunca es tan sencillo como eso. También transmite nuestras ideas acerca de lo que es importante para nuestra materia y envía mensajes a quienes son evaluados que influirán en su aprendizaje futuro. Por eso he destacado la importancia de la calidad de las pruebas: ¿miden realmente lo que dicen que miden?, ¿qué impacto producen en quienes se someten a las pruebas? Parte de mi crítica de la evaluación sumativa al uso a base de pruebas se debe a que lleva a una "microenseñanza" muy centrada en mejorar las puntuaciones en tests específicos y previsibles que poco o nada sirve para el aprendizaje posterior. Ha fracasado en su doble cometido. Por

eso necesitamos evaluaciones efectuadas en situaciones normales, que requieran un aprendizaje más profundo y más basado en principios, parte del cual pueda transferirse a períodos posteriores. Los ejercicios de preparación de tests que buscan pistas concretas y utilizan después el recuerdo de respuestas practicadas aportan poco a la resolución de problemas excasamente habituales en el futuro.

Terry CROOKS pone un ejemplo sencillo de esto tomado de su trabajo de supervisión del aprendizaje de la aritmética en Nueva Zelanda. Descubrió que los niños a los que se había preparado para que adoptaran enfoques estándar del cálculo tenían dificultades para adoptar estrategias más eficientes:

> Al pedirles que sumaran 97 y 52, lo más probable era que estos niños lo escribieran como una suma en su forma estándar en vez de tomar 3 de 52 y resolver el problema sumando 49 y 100. Aunque se les hubieran explicado una y otra vez otras estrategias diferentes y las hubieran practicado, cuando no se les recordaba que utilizaran las estrategias nuevas, volvían al enfoque estándar. Es más probable que los estudiantes que aprenden desde los primeros años las estrategias nuevas comprendan más profundamente el número y que su enfoque sea más flexible.
>
> (2002, pág. 10.)

En mi propia experiencia reciente de observación de la enseñanza de la Aritmética en Nueva Zelanda, pude ver a niños pequeños que tenían que decidir y comentar qué estrategias podrían utilizar en un determinado problema (a menudo, les preguntaban por dos) antes de efectuar el cálculo. Esto es una evaluación sostenible.

Administración de tests que promueve un aprendizaje eficaz

En los Capítulos V, VI y VII, hemos visto cómo puede promover la evaluación un aprendizaje más rico y sostenible. La fuerza de la evaluación, sobre todo la de importancia decisiva, está en que configura lo que aprendemos y cómo lo aprendemos. Mi enfoque consistió en buscar mejoras de la calidad de los tests de rendimiento, de manera que la inevitable enseñanza para el examen condujera a una enseñanza y un aprendizaje mejores.

Para ello, es fundamental insistir en el *aprendizaje fundado en principios*, en el que se hace hincapié en la comprensión flexible y en las destrezas. La forma de promoverlo consiste en utilizar preguntas de examen menos previsibles, de manera que la preparación implique una enseñanza basada en preguntas del tipo: "¿qué pasa si...?" y no en enunciados como: "cuando veas esto...". En este contexto, la adecuación a la finalidad se comprobaría examinando si la evaluación aborda efectivamente los *objetivos* de la asignatura y no solo el contenido. Si el objetivo es: "promover la curiosidad por..." o "desarrollar puntos de vista personales sobre" una materia, ¿cómo lo estimula la forma de la evaluación? La consecuencia es que necesitamos un cóctel de evaluación más imaginativo del que exhiben muchos sistemas actuales.

Es más fácil hacer realidad esos enfoques en la evaluación en el aula. Aquí, el maestro o profesor tiene la oportunidad de utilizar evaluaciones más imaginativas y auténticas. El problema es si las escuelas y los docentes tienen la confian-

za suficiente para ofrecer una amplia combinación de evaluaciones o si van a lo seguro e imitan las pruebas externas; el enfoque que se basa en los exámenes ya realizados.

Para recuperar la evaluación, hay que resistir las presiones para reducir la evaluación a tests y ejercitar constantemente a los alumnos en las técnicas de examen. Supone unos tests mejores que promuevan un aprendizaje más profundo. Significa también estimular a los docentes para que utilicen un conjunto de evaluaciones más imaginativo y hacer mucho más hincapié en los usos formativos de la evaluación. De este modo, la evaluación utilizará su poder de un modo más constructivo al servicio del aprendizaje.

Aprendices autorregulados

El *doble cometido* clave de la evaluación implica tanto centrar la atención en la tarea inmediata como equipar a los estudiantes para un futuro desconocido. En el Capítulo VII, examiné los riesgos de resaltar demasiado las exigencias inmediatas tanto de las pruebas de importancia decisiva como de los procedimientos del aprendiz, por ejemplo, "aprender a aprender", sin prestar suficiente atención a lo *que* se aprende. Esta cuestión se halla tras el argumento de Anna SFARD (1998) de que necesitamos dos metáforas del aprendizaje: *adquisición* y *participación*, en vez de confiar solo en una. En el clima actual de rendición de cuentas, es probable que se pase por alto el aprendizaje para este futuro desconocido, en la búsqueda de unos resultados cada vez mejores en el momento presente.

Terry CROOKS ha definido el aprendizaje autorregulado en términos de control y administración del propio aprendizaje, lo que incluye "supervisar el progreso del aprendizaje, reconociendo cuándo hace falta una acción de recuperación o preventiva para mantener o aumentar la calidad y reuniendo la fuerza de voluntad precisa para continuar trabajando para alcanzar elevados niveles de trabajo" (2002, pág. 4). Este es el tipo de autorregulación que está en el centro de la profesionalidad: supervisar nuestro trabajo teniendo presentes unos estándares elevados y asumiendo la responsabilidad personal de mejorarlo.

Recuperar la evaluación supone, por tanto, conseguir un mejor equilibrio entre la evaluación sumativa del saber y las destrezas actuales y la evaluación sostenible, que estimule a los aprendices a seguir aprendiendo. La clave de este cambio está en la modificación de la función del docente. Las pruebas de importancia decisiva promueven la dependencia del profesor, tanto para interpretar el currículum como para saber lo que hace falta para el examen. Los maestros y profesores pueden también empezar a considerarse responsables del aprendizaje de sus alumnos, porque lo que tengan que aprender habrá sido especificado con todo detalle. Lo mismo puede ocurrir con la evaluación: al maestro o profesor le corresponde facilitar niveles o calificaciones "indicativos", mientras que, en esto, el estudiante es en gran medida pasivo[7]. Esa dependencia ahuyenta la sostenibi-

[7] Puede que el o la estudiante sea una Ruth BORLAND, que sepa más que el profesor acerca de lo que hace falta, aunque se trate más aquí de "autonomía procedimental" que de "autonomía personal". Esta es la útil distinción de Kathryn ECCLESTONE. Señala la autora que este puede ser un punto

lidad, cuya esencia es que los aprendices puedan evaluar y regular su propio aprendizaje. Al crear a unos aprendices más autónomos, las escuelas y los docentes tendrán que ceder parte del poder que la evaluación les confiere, compartiendo su "conocimiento gremial" con sus alumnos. Como descubrieron BLACK y sus colaboradores (2003) en su proyecto de investigación-acción, ésta fue una de las transiciones más duras para los profesores y los maestros, dado que suponía renunciar a parte del control y la autoridad en el aula que la evaluación dirigida por el docente les había otorgado.

La "evaluación para el aprendizaje" da oportunidades para reequilibrar, sobre todo aprendiendo a negociar las intenciones del aprendizaje y los criterios de éxito y desarrollando las destrezas de autoevaluación y de evaluación a cargo de los compañeros. En el Capítulo VII, afirmé que "compartiendo los objetivos del aprendizaje" se corre el riesgo de caer en la *conformidad con los criterios* si no hay una auténtica negociación entre el docente y los aprendices. La evaluación sostenible hace imperativo que el aprendiz participe activamente tanto en las intenciones del aprendizaje como en comprender qué supone cumplirlas. No basta con tener conciencia de ellas; los aprendices tienen que elaborar estrategias para supervisar ellos mismos su progreso hacia estos objetivos. Esto puede suponer fijar objetivos intermedios y comprobar sus progresos a intervalos regulares. El razonamiento es el siguiente: al enfrentarse a exigencias y estándares nuevos, el alumno tendrá los medios para examinarlos, hacerse una idea de lo que hace falta y elaborar estrategias para satisfacerlos, es decir, una evaluación sostenible.

Como vimos también en el Capítulo VII, la autoevaluación y la evaluación a cargo de compañeros desempeñan aquí un papel clave. Solo cuando comprenda lo que trato de hacer y lo que implica el éxito, seré capaz de juzgar cómo lo hago o cómo lo hacen mis compañeros. Cuanto más complejo sea el aprendizaje, menos probable será que podamos realizarlo solos. En vez de volver a caer en una dependencia del docente, la evaluación sostenible nos estimula para utilizar a quienes estén a nuestro alrededor para que nos faciliten fuentes de información, respondan a nuestras ideas y nos muestren otros puntos de vista. Es poco probable que el maestro o profesor esté presente cuando progresemos hacia otro aprendizaje, pero seguro que estarán otros aprendices y tendremos que saber cómo aprovechar su ayuda. Cuando los aprendices entran en un mundo de tareas complejas, es poco probable que puedan abordarlas en solitario. Necesitamos el apoyo, la habilidad y la retroinformación de los demás para ser aprendices eficaces.

Motivación

Al hablar de la retroinformación en el Capítulo VII, presenté la idea de que la autorregulación comprende la *autovaloración* y el *autocontrol*. La disposición a seguir tratando de dominar una tarea se basa en nuestras destrezas de autocontrol.

---

de partida necesario, pero debe haber un progreso hacia la autonomía personal y crítica para que la autonomía procedimental no acabe siendo poco más que una "tecnología de la autovigilancia" (2002, pág. 36).

Estas tienen relación con nuestra forma de obtener y utilizar la retroinformación y con nuestro compromiso con el éxito: cuánto esfuerzo estamos dispuestos a invertir; nuestra confianza en que alcanzaremos el éxito, y nuestras atribuciones acerca de éxitos y fracasos:

> Los principales impedimentos para el aprendizaje no son cognitivos. No se trata de que los estudiantes no puedan aprender, sino de que no quieren aprender. Si los educadores invirtieran una fracción de la energía que ahora gastan intentando transmitir información en tratar de estimular el gozo de aprender de los estudiantes, obtendríamos unos resultados mucho mejores.
>
> (CSIKSZENTMIHALYI, 1990, pág. 118.)

La evaluación desempeña un papel clave tanto en la autovaloración como en el autocontrol. David BOUD señala que "las actividades de evaluación deben dejar mejor provistos a los estudiantes para hacer frente a su siguiente reto o, como mínimo, no peor de lo que estarían en otro caso... parte de esto es tener suficiente confianza en que pueden abordarlo con ciertas probabilidades de éxito" (2000, pág. 8). Critica el vocabulario judicial y marcado por los valores de gran parte de las evaluaciones y el hecho de que puedan obstaculizar aprendizajes posteriores: "la evaluación daña, es incómoda y en la mayoría de nosotros ha dejado una profunda huella" (2002, pág. 2).

Conclusión: La recuperación del territorio

El Mago de Oz, el personaje de L. Frank Baum, es el venerado y poderoso dirigente de la Tierra de Oz, a quien Dorothy y sus variopintos compañeros se dirigen por la calle de las baldosas amarillas. Dorothy soporta los muchos disfraces aterradores con los que se presenta el Mago hasta descubrir que, en realidad, es Oscar Diggs, un estadounidense corriente que llegó a Oz en un globo de aire caliente. Una vez allí, utilizó montones de trucos y accesorios muy elaborados para parecer "grande y poderoso", con el resultado de que terminaron adorándolo.

Este libro ha examinado cómo la evaluación puede crear impresiones engañosas y hacer afirmaciones "grandes y poderosas". He tratado de definir el papel correcto y más humilde de la evaluación y de mostrar cómo puede contribuir al aprendizaje. Para ello, es fundamental que las personas cuestionen las etiquetas que otros quieran asignarles y que se responsabilicen de sus propias identidades como aprendices. Vivimos en tiempos de tests, pero no tenemos que estar a su merced.

Bibliografía

AIRASIAN, P. W. (1997) *Classroom Assessment*. Nueva York. McGraw-Hill.

ALEXANDER, R. (2000) *Culture and Pedagogy: International Comparisons in Primary Education*. Oxford. Blackwell.

— (2004) *Towards Dialogic Teaching: Rethinking Classroom Talk*. Nueva York. Diálogos.

ALLAL, L. y LÓPEZ, L. (2005) "Formative Assessment of Learning: A Review of Publications in French", en OECD staff, *Formative Assessment: Improving Learning in Secondary Classrooms*. París. OECD, págs. 241-264.

AMANO, I. (1997) "Education in a More Affluent Japan", *Assessment in Education: Principles, Policy and Practice*, 4, págs. 51-66.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION/APA/NCME (1999) *Standards for Educational and Psychological Testing*. Washington DC. American Educational Research Association.

ANASTASI, A. (1985) "Mental Measurement: Some Emerging Trends", en J. V. MITCHELL (ed.) *The Ninth Mental Measurements Yearbook*. Lincoln. NE. Buros Institute of Mental Measurement.

APPLE, M. W. (1989) "How Equality Has Been Redefined in the Conservative Restoration", en W. G. SECADA (ed.) *Equity in Education*. Nueva York. Falmer Press, págs. 7-35.

ASSESSMENT REFORM GROUP (1999) *Assessment for Learning: Beyond the Black Box*. University of Cambridge, UK. Assessment Reform Group.

— (2002a) *Assessment for Learning: 10 Principles*, University of Cambridge, R.U. Assessment Reform Group.

— (2002b) *Testing, Motivation and Learning*. University of Cambridge, R.U. Assessment Reform Group.

BAKER, E. y O'NEIL, H. F. (1994) "Performance Assessment and Equity: A View from the USA", *Assessment in Education: Principles, Policy and Practice*, 1, págs. 11-26.

BAUMGART, N. y HALSE, C. (1999) "Approaches to Learning Across Cultures: The Role of Assessment", *Assessment in Education: Principles, Policy and Practice*, 6, págs. 321-340.

BEREITER, C. y SCARDAMALIA, M. (1989) "Intentional Learning as a Goal of Instruction", en L. RESNICK y R. GLASER (eds.) *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*. Hillsdale, NJ. Londres. Laurence Erlbaum Associates, págs. 361-392.

BERGLAS, S. y JONES, E. (1978) "Drug Choice as a Self-Handicapping Strategy in Response to Noncontingent Success", *Journal of Personality and Social Psychology*, 36, págs. 405-417.
BEVERTON, S., HARRIS, T., GALLANNAUGH, F. y GALLOWAY, D. (2005) "Teaching Approaches to Promote Consistent Level 4 Performance in Key Stage 2 English and Mathematics". *DFES Research Brief*, Nº. 699.
BIGGS, J. (1996) "Enhancing Teaching Through Constructive Alignment". *Higher Education*, 32, págs. 347-364.
— (1999) *Teaching for Quality Learning at University*. Buckingham. SRHE and Open University Press.
BINET, A. (1909) *Les Idées Modernes Sur Les Enfants*. París. Flammarion.
— y SIMON, T. (1911) "La mesure du développement de l'intelligence chez les enfants", *Bulletin de la Société libre pour l'étude psychologique de l'enfant*, págs. 70-71.
BLACK, P. y WILIAM, D. (1998a) "Assessment and Classroom Learning", *Assessment in Education*, 5, págs. 7-71.
— (1998b) *Inside the Black Box: Raising Standards Through Classroom Assessment*. Londres. King's College (ver también *Phi Delta Kappan*, 80, págs. 139-148).
— (2006) "Developing a Theory of Formative Assessment", en J. GARDNER (ed.) *Assessment and Learning*. Londres. Sage, págs. 81-100.
BLACK, P., HARRISON, C., LEE, C., MARSHALL, B. y WILIAM, D. (2002) *Working Inside the Black Box: Assessment for Learning in the Classroom*. Londres. NFER Nelson.
— (2003) *Assessment for Learning: Putting it Into Practice*. Buckingham. Open University Press.
BLACK, P., MCCORMICK, R., JAMES, M. y PEDDER, D. (2006) "Learning How to Learn and Assessment for Learning: A Theoretical Inquiry". *Research Papers in Education*, 21, págs. 119-132.
BLOOM, B. S. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Londres. Longman Group.
—, HASTINGS, J. T. y MADAUS, G. F. (1971) *Handbook on Formative and Summative Evaluation of Student Learning*. Nueva York. McGraw-Hill.
BOND, L., SMITH, R., BAKER, W. K. y HATTIE, J. A. (2000) *Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study*. Washington, DC. National Board for Teaching Standards.
BORING, E. G. (1923) "Intelligence as the Tests Test It". *New Republic*, 6 Junio, págs. 35-37.
BOUD, D. (2000) "Sustainable Assessment: Rethinking Assessment for the Learning Society", *Studies in Continuing Education*, 22, págs. 151-167.
— (2002) "The Unexamined Life is Not the Life for Learning: Rethinking Assessment for Lifelong Learning", Professorial Lecture given at Trent Park, Middlesex.
BOYLE, B. y BRAGG, J. (2006) "A Curriculum Without Foundation". *British Educational Research Journal*, 32, págs. 569-582.
BRANSFORD, J. D., BROWN, A. L. y COCKING, R. R. (2000) *How People Learn: Brain, Mind, Experience and School*. Washington, DC. National Academies Press.
BROADFOOT, P. (1979) *Assessment, Schools and Society*. Londres. Methuen.
— y BLACK, P. (2004) "Redefining Assessment? The First Ten Years of Assessment in Education". *Assessment in Education: Principles, Policy and Practice*, 11, págs. 7-26.
BROCA, P. (1861) "Sur Le Volume et la Forme du Cerveau Suivant les Individus et Suivant les Races", *Bulletin Société D'Anthropologie Paris*, 2, págs. 139-207, 301-321, 441-446.
BROPHY, J. (1998) "Towards a Model of the Value Aspects of Motivation in Education: Developing Appreciation for Particular Learning Domains and Activities", trabajo presentado en la American Educational Research Association Conference, San Diego.

BURT, C. L. S. (1937) *The Backward Child*. Londres. University of London Press.
— (1943) "Ability and Income", *British Journal of Educational Psychology*, 13, págs. 83-98.
— (1955) "The Evidence for the Concept of Intelligence", *British Journal of Educational Psychology*, 25, págs. 158-177.
— "The Examination at Eleven Plus", *British Journal of Educational Studies*, 7, págs. 99-117.
BUTLER, R. (1988) "Enhancing and Undermining Intrinsic Motivation: The Effect of Task-Involving and Ego-Involving Evaluation on Interest and Performance", *British Journal of Educational Psychology*, 58, págs. 1-14.
BUTLER, D. L. y WINNE, P. H. (1995) "Feedback and Self-Regulated Learning: A Theoretical Synthesis", *Review of Educational Research*, 65 (3), págs. 245-281.
CAHAN, S. y COHEN, N. (1989) "Age Versus Schooling Effects on Intelligence Development", *Child Development*, 60, págs. 1239-1249.
CANNELL, J. (1987) *Nationally Normed Elementary Achievement Testing in America's Public Schools: How All Fifty States Are Above the National Average*. Daniels, WV. Friends for Education.
— (1989) *How Public Educators Cheat on Standardised Achievement Tests*. Albuquerque, NM. Friends for Education.
CARLESS, D. (2005) "Prospects for the Implementation of Assessment for Learning", *Assessment in Education*, 12, págs. 39-54.
— (2007) "Conceptualising Pre-Emptive Formative Assessment", *Assessment in Education: Principles, Policy and Practice*, 14 (2), págs. 171-184.
CARPENTER, P., JUST, M. y SHELL, P. (1990) "What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test", *Psychological Review*, 97, págs. 404-431.
CARROLL, J. B. (1993) *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge, UK. Cambridge University Press.
CATTELL, J. M. (1890) "Mental Tests and Measurement", *Mind*, 15, págs. 373-381.
CECI, S. J. (1996) *On Intelligence - More or Less: A Bio-ecological Treatise on Intellectual Development*. Englewood Cliffs, NJ. Prentice Hall.
—, Rosenblum, T. B. y Kumpf, M. (1998) "The Shrinking Gap Between High-and Low-Scoring Groups: Current Trends and Possible Causes", en U. NEISSER (ed.) *The Rising Curve: Long-term Gains in IQ and Related Measures*, 1ª ed., Washington, DC. American Psychological Association, págs. 287-302.
CLARKE, S. (1998) *Targeting Assessment in the Primary School*. Londres. Hodder and Stoughton.
— (2001) *Unlocking Formative Assessment*. Londres. Hodder and Stoughton.
COBB, P. (1994) "Where is Mind? Constructivist and Sociocultural Perspectives on Mathematical Development", *Educational Researcher*, 23 (7), págs. 13-20.
COFIELD, F., MOSELEY, D., HALL, E. y ECCLESTONE, K. (2004) *Learning Styles and Pedagogy in Post-16 Learning: A Systematic and Critical Review*. Londres. Learning and Skills Research Centre.
COLLEGEBOARD (2006) *SAT Reasoning Test*, http://www.collegeboard.com/student/ testing/sat (consultado el 16 Noviembre 2007).
COLLINS, R. (1990) "Market Closure and the Conflict Theory of the Professions", en M. BURRAGE y R. TORSTENDAHL (eds.) *Professions in Theory and History: Rethinking the Study of Professions*. Londres. Sage, págs. 24-43.
CROOKS, T. (1988) "The Impact of Classroom Evaluation Practices on Students", *Review of Educational Research*, 58, págs. 438-481.
— (2002) "Assessment, Accountability and Achievement - Principles, Possibilities and Pitfalls", trabajo presentado en la 24 Annual Conference of the New Zealand Association for Research in Education, Palmerston North, New Zealand.

CROOKS, T. (2007) *Key Factors in the Effectiveness of Assessment for Learning*, AERA Annual Meeting. Chicago.
CSIKSZENTMIHALYI, M. (1990) "Literacy and Intrinsic Motivation", *Daedalus*, 19 (2), páginas 115-140.
CUMMING, J. (2000) "After DIF, What Culture Remains?" *26th IAEA Conference*. Jerusalen.
— y MAXWELL, G. (2004) "Assessment in Australian Schools: Current Practice and Trends", *Assessment in Education: Principles, Policy and Practice*, 11 (1), págs. 89-108.
CURRY, L. (1983) "An Organisation of Learning Styles Theory and Constructs", *Annual Meeting of the American Educational Research Association*. Montreal. Quebec.
DALE, W. (1875) *The State of the Medical Profession in Great Britain and Ireland*. Dublin. J. Atkinson & Co.
DANN, R. (2002) *Promoting Assessment as Learning: Improving the Learning Process*. Londres. RoutledgeFalmer.
DARLING-HAMMOND, L. (1994) "Performance-Based Assessment and Educational Equity", *Harvard Educational Review*, 64, págs. 5-30.
— y Rustique-Forrester, E. (2005) "The Consequences of Student Testing for Teaching and Teacher Quality", en J. L. HERMAN y E. H. HAERTEL (eds.) *Uses and Misuses of Data from Educational Accountability and Improvement.* Chicago, IL. National Society for the Study of Education, págs. 289-319.
DAY, C., STOBART, G., SAMMONS, P., KINGTON, A., GU, Q., SMEES, R. y MUJTABA, T. (2006) *Variations in Teachers' Work, Lives and Effectiveness*. Londres. DfES.
—, SAMMONS, P., STOBART, G., KINGTON, A. y GU, Q. (2007) *Teachers Matter: Connecting Work, Lives and Effectiveness*. Maidenhead, UK. Open University Press/McGraw-Hill.
DECI, E. L., KOESTNER, R. y RYAN, M. R. (1999) "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation", *Psychological Bulletin*, 125, págs. 627-668.
DEMPSEY, J. V. y SALES, G. C. (1993) *Interactive Instruction and Feedback*. Englewood Cliffs, NJ. Educational Technology Publications.
DENNISON, W. F. y KIRK, R. (1990) *Do, Review, Learn, Apply: A Simple Guide to Experiential Learning*. Oxford. Blackwell Education.
DES/WO (1988) *Task Group on Assessment and Testing: A Report*. Londres. Department of Education and Science and the Welsh Office.
DEWEY, J. (1938) *Experience and Education*. Nueva York y Londres. Collier-Macmillan.
DfES (2003) *Excellence and Enjoyment: A Strategy for Primary Schools*. England. DfES 0377/2003.
DORE, R. (1976) *The Diploma Disease: Education, Qualification and Development*. Londres. Allen and Unwin.
— (1997a) *The Diploma Disease: Education, Qualification and Development*. Londres. Institute of Education, University of London.
— (1997b) "The Argument of the Diploma Disease: a Summary", *Assessment in Education: Principles, Policy and Practice*, 4, págs. 23-32.
— (1997c) "Reflections on the Diploma Disease Twenty Years Later", *Assessment in Education: Principles, Policy and Practice*, 4, págs. 189-206.
DUNN, R. (1990a) "Rita Dunn Answers Questions on Learning Styles", *Educational Leadership*, 48, págs. 15-19.
— (1990b) "Understanding Dunn and Dunn Learning Styles Model and the Need for Individual Diagnosis and Prescription", *Reading, Writing and Learning Disabilities*, 6, páginas 223-247.
— (2003a) "The Dunn and Dunn Learning Style Model and Its Theoretical Cornerstone", en S. ARMSTRONG, M. GRAFF, C. LASHLEY, E. PETERSON, S. RAYNOR, E. SADLER-SMITH, M. SCHIERING y D. SPICER (eds.) *Bridging Theory and Practice*. Proceedings of the Eighth Annual European Learning Styles Information Network Conference. Hull. University of Hull.

DUNN, R. (2003b) "Epilogue: So What?" en R. DUNN y S. GRIGGS (eds.) *Synthesis of the Dunn and Dunn Learning Styles Model Research: What, When, Where and So What - the Dunn and Dunn Learning Styles Model and Its Theoretical Cornerstone,* 7-10, Nueva York. St John's University.

— y GRIGGS, S. (1988) *Learning Styles: A Quiet Revolution in American Secondary Schools*. Reston, VA. National Association of Secondary School Principals.

— y GRIGGS, S. (1990) "Research on the Learning Style Characteristics of Selected Racial and Ethnic Groups", *Journal of Reading, Writing and Learning Disabilities*, 6, páginas 261-280.

— y GRIGGS, S. (2003) *Synthesis of the Dunn and Dunn Learning Styles Model Research: Who, What, When and Where and So What - the Dunn and Dunn Learning Styles Model and Its Theoretical Cornerstone*, Nueva York: St John's University.

—, DUNN, K. y PRICE, G. E. (1975) *Learning Style Inventory: An Inventory for the Identification of How Individuals in Grades 3 Through 12 Prefer to Learn*. Lawrence, KS. Price Systems.

—, GRIGGS, S., GORMAN, B., OLSON, J. y Beasley, M. (1995) "A Meta-Analytic Validation of the Dunn and Dunn Model of Learning Style Preferences", *Journal of Educational Research*, 88, págs. 353-363.

DWECK, C. S. (2000) *Self-Theories: Their Role in Motivation, Personality and Development*. Philadelphia. Psychology Press.

EARL, L. M. (2003a) *Watching and Learning 3: Final Report of the External Evaluation of England's National Literacy and Numeracy Strategies*. Nottingham. DfES.

— (2003b) *Assessment as Learning: Using Classroom Assessment to Maximise Student Learning*. Thousand Oaks, CA. Corwin Press.

ECCLESTONE, K. (2002) *Learning Autonomy in Post-16 Education*. Londres. Routledge-Falmer.

— y HAYES, D. (2008) *The Dangerous Rise of Therapeutic Education*. Londres. Routledge-Falmer.

ECKSTEIN, M. A. y NOAH, H. J. (1993) *Secondary School Examinations: International Perspectives on Policies and Practice*. New Haven. Yale University Press.

EDWARDS, A. (2005) "Let's Get Beyond Community and Practice: The Many Meanings of Learning by Participating", *The Curriculum Journal*, 16 (1), págs. 49-65.

ELMORE, R. y FUHRMAN, S. (2001) "Holding Schools Accountable: Is It Working?", *Phi Delta Kappan*, 83, págs. 67-72.

ENGESTRÖM, Y. (1987) *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Helsinki, Finland. Orienta-Konsultit Oy.

— (1999) "Activity Theory and Individual and Social Transformation", en Y. ENGESTRÖM, R. MIETTINEN y R.-L. PUNAMÄKI (eds.) *Perspectives on Activity Theory.* Cambridge, R.U. Cambridge University Press, págs. 19-38.

ENTWISTLE, N., HANEY, M. y HOUNSELL, D. (1979) "Identifying Distinctive Approaches to Studying", *Higher Education*, 8, págs. 365-380.

—, SKINNER, D., ENTWISTLE, D. y ORR, S. (2000) "Conceptions and Beliefs About 'Good Teaching': An Integration of Contrasting Research Areas", *Higher Education Research and Development*, 19, págs. 5-26.

—, MCCUNE, V. y WALKER, P. (2001) "Conceptions, Styles and Approaches within Higher Education: Analytic Abstractions and Everyday Experience", en R. STERNBERG y L. ZHANG (eds.) *Perspectives on Cognitive, Learning, and Thinking Styles*. Mahwah, NJ. Lawrence Erlbaum.

ERAUT, M. (1997) "Perspectives on Defining 'The Learning Society'", *Journal of Education Policy*, 12, págs. 551-558.

— (2007) "Assessment of Significant Learning Outcomes": 3º Seminario. Feedback and Formative Assessment in the Workplace, *Assessment of Significant Learning Outcomes Seminar*. Institute of Education, Londres. University of Sussex.

ERNEST, P. (2000) "Why Teach Mathematics?" en S. BRAMALL y J. WHITE (eds.) *Why Learn Maths?* Londres. University of London, Institute of Education.
FLYNN, J. (1987) "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure", *Psychological Bulletin*, 101, págs. 171-191.
— (1991) *Asian Americans: Achievement Beyond IQ*. Hillsdale, NJ. Lawrence Erlbaum Associates.
— (1998) "IQ Gains over Time: Towards Finding the Causes", en U. NEISSER (ed.) *The Rising Curve: Long-Term Gains in IQ and Related Measures*, 1ª ed., Washington, DC. American Psychological Association, págs. 25-66.
— (2006) "Beyond the Flynn Effect: Solution to All Outstanding Problems - Except Enhancing Wisdom", Paper given at a presentation for The Psychometrics Centre: Cambridge Assessment, University of Cambridge, UK.
FOUCAULT, M. (1977) *Discipline and Punishment*, traducido por Alan SHERIDAN. Londres. Allen Lane.
FREDERIKSEN, J. R. y COLLINS, A. (1989) "A Systems Approach to Educational Testing", *Educational Researcher*, 18 (9), págs. 27-32.
FULLAN, M. (2001) *The New Meaning of Educational Change*, Londres. Routledge-Falmer.
GALTON, F. (1869) *Hereditary Genius*. Londres. Macmillan.
GARDNER, H. (1983) *Frames of Mind: The Theory of Multiple Intelligences*. Londres. Heinemann.
— (1993) *Multiple Intelligences: The Theory in Practice: A Reader*. Nueva York. Basic Books.
— (1999) *Intelligence Reframed: Multiple Intelligences for the 21st Century*. Nueva York. Basic Books.
— (2006) *Multiple Intelligences: New Horizons*. Nueva York; Londres. Basic Books, Perseus Running.
— y COWAN, P. (2005) "The Fallibility of High Stakes '11-Plus' Testing in Northern Ireland", *Assessment in Education: Principles, Policy and Practice*, 12, págs. 145-165.
GASINZIGWA, P. G. (2006) "The Role of Education, Particularly Curriculum and Examination, in Social Reconstruction and National Reconciliation of Post-Genocide Rwanda", Inédito MA Report. Institute of Education, University of London.
GILLBORN, D. y YOUDELL, D. (2000) *Rationing Education: Policy, Practice, Reform, and Equity*. Buckingham, R.U. Open University Press.
— y YOUDELL, D. (2001) "The New IQism: Intelligence, 'Ability' and the Rationing of Education", en J. DEMAINE (ed.) *Sociology of Education Today*. Basingstoke: Palgrave, páginas 65-99.
GINSBURG, H. P. (1997) *Entering the Child's Mind*, Cambridge, R.U., Cambridge University Press.
GIPPS, C. (1999) "Sociocultural Aspects of Assessment", *Review of Research in Education*, 24, págs. 357-392.
— y MURPHY, P. (1994) *A Fair Test? Assessment, Achievement and Equity*. Buckingham, R.U. Open University Press.
—, HARGREAVES, E., MCCALLUM, B. y EBRARY, I. (2001) *What Makes a Good Primary School Teacher?: Expert Classroom Strategies*. Londres, Nueva York. RoutledgeFalmer.
GODDARD, H. H. (1914) *Feeble-Mindedness: Its Causes and Consequences*. Nueva York. Macmillan.
GOLDSTEIN, H. (2003) "Evaluating the Evaluators: A Critical Commentary on the Final Evaluation of the English National Literacy and Numeracy Strategies" http://www.cmm.bristol.ac.uk/team/HG_Personal/commentaries.htm (consultado el 16 Noviembre 2007).
GOLEMAN, D. (1995) *Emotional Intelligence*. Nueva York, Londres. Bantam Books.
— (1998) *Working With Emotional Intelligence*. Nueva York. Bantam Books.

GOLEMAN, D. (2006) *Social Intelligence: The New Science of Human Relationships*. Londres. Hutchinson.
GORDON, S. y REESE, M. (1997) "High Stakes Testing: Worth the Price?", *Journal of School Leadership*, 7, págs. 345-368.
GOULD, S. J. (1996) *The Mismeasure of Man*. Nueva York. Norton.
GRAHAM, P. A. (1995) "Assimilation, Adjustment and Access: An Antiquarian View of American Education", en D. RAVITCH y M. A. VINOVSKIS (eds.) *Learning from the Past: What History Teaches Us About School*. Baltimore, MD. Johns Hopkins University Press.
GREENfIELD, P. M. (1998) "The Cultural Evolution of IQ", en U. NEISSER (ed.) *The Rising Curve: Long-Term Gains in IQ and Related Measures*, 1ª ed., Washington, DC. American Psychological Association.
GROVES, B. (2002) "They Can Read the Words But What Do They Mean?" *Adults Learning*, 13, págs. 18-20.
GUILFORD, J. P. (1967) *The Nature of Human Intelligence*, Nueva York. Londres. McGraw-Hill.
GUNZENHAUSER, M. (2003) "High-Stakes Testing and the Default Philosophy of Education", *Theory into Practice*, 42, págs. 51-58.
HACKING, I. (2006) *Kinds of People: Moving Targets*, The Tenth British Academy Lecture. http://www.britac.ac.uk (consultado el 16 Noviembre 2007).
HAERTEL, E. H. y HERMAN, J. L. (2005) "A Historical Perspective on Validity Arguments of Accountability Testing", en J. L. HERMAN y E. H. HAERTEL (eds.) *Uses and Misuses of Data from Educational Accountability and Improvement*. Chicago, IL. National society for the Study of Education, págs. 1-34.
HAGGIS, T. (2003) "Constructing Images of Ourselves? A Critical Investigation Into 'Approaches to Learning' Research in Higher Education", *British Educational Research Journal*, 29, págs. 89-204.
HAMILTON, L. (2003) "Assessment as a Policy Tool", en R. FLODEN (ed.) *Review of Research in Education*, 27, Washington DC: AERA.
HANEY, W. (2000) "The Myth of the Texas Miracle in Education", *Education Analysis Policy Archives*, 8 (41). http://epaa.asu.edu.epaa/v9n2.html (consultado el 16 Noviembre 2007).
—, MADAUS, G. y LYONS, R. (1993) *The Fractured Marketplace for Standardised Testing*, Boston, MA: Kluwer. *Hansard* (13 Febrero 1862) *Hansard's Parliamentary Debates*.
HANSON, F. A. (1994) *Testing Testing: Social Consequences of the Examined Life*. Berkeley, CA. University of California Press.
HARLEN, W. (2006) "On the Relationship between Assessment for Formative and Summative Purposes", en J. GARDNER (ed.) *Assessment and Learning*. Londres. Sage, págs. 103-117.
HARRIS, S., WALLACE, G. y RUDDUCK, J. (1995) "'It's Not That I Haven't Learnt Much. It's Just That I Don't Really Understand What I'm Doing': Metacognition and Secondary School Students", *Research Papers in Education*, 10, págs. 253-271.
HART, S., DIXON, A., DRUMMOND, M. J. y MCINTYRE, D. (2004) *Learning Without Limits*. Maidenhead. Open University Press.
HATTIE, J. y TIMPERLEY, H. (2007) "The Power of Feedback", *Review of Educational Research*, 77, págs. 81-112.
HAYWARD, L. (2007) *Assessment in Education: Principles, Policy and Practice*, 14 (2), páginas 251-268.
HERMAN, J. L. y HAERTEL, E. H. (eds.) (2005) *Uses and Misuses of Data from Educational Accountability and Improvement*. Chicago, IL. National Society for the Study of Education.
HERRNSTEIN, R. J. y MURRAY, C. A. (1994) *The Bell Curve: Intelligence and Class Structure in American Life*. Nueva York. Free Press.

HOLMES, E. G. A. (1911) *What Is and What Might Be: A Study of Education in General and Elementary Education in Particular*. Londres. Constable and Co., Ltd.
HONEY, P. y MUMFORD, A. (2000) *The Manual of Learning Styles*. Maidenhead. Peter Honey.
HOWE, M. J. A. (1997) *IQ in Question: The Truth About Intelligence*. Londres. Sage.
HUFTON, N. y ELLIOTT, J. (2001) "Achievement Motivation: Cross-cultural Puzzles and Paradoxes". Trabajo presentado en la British Educational Research Association Conference, Leeds.
HURSH, D. (2005) "The Growth of High-Stakes Testing in the USA: Accountability, Markets and the Decline of Educational Equality", *British Educational Research Journal*, 31: 605-622.
HUSSEY, T. y SMITH, P. (2002) "The Trouble with Learning Outcomes", *Active Learning in Higher Education*, 3, págs. 220-233.
JAMES, M. (2006) "Assessment, Teaching and Theories of Learning", en J. GARDNER (ed.) *Assessment and Learning*. Londres. Sage, págs. 47-60.
— y PEDDER, D. (2006) "Beyond Method: Assessment and Learning Practices and Values", *Curriculum Journal*, 17, págs. 109-138.
JENSEN, A. R. (1980) *Bias in Mental Testing*. Londres. Methuen.
— (1993) "Why is Reaction Time Correlated With Psychometric *g*?" *Current Directions in Psychological Science*, 2, págs. 53-56.
— (1998) *The g Factor: The Science of Mental Ability*. Westport, CT. Praeger.
KAMALI, A. (2006) "National Examinations in Rwanda: On the Right Track?": Inédito MA Report. Institute of Education, University of London.
KAVALE, K. y FORNESS, S. (1987) "Substance Over Style: Assessing the Efficacy of Modality Testing and Teaching", *Exceptional Children*, 54, págs. 228-239.
— y FORNESS, S. (1990) "Substance Over Style: A Rejoinder to Dunn's Animadversions", *Exceptional Children*, 54, págs. 357-361.
KEILLOR, G. (1985) *Lake Wobegon Days*. Nueva York. Viking.
KLUGER, A. y DENISI, A. (1996) "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis and a Preliminary Feedback Intervention Theory", *Psychological Bulletin*, 119 (2), págs. 254-284.
KOHN, A. (1993) *Punished by Rewards: The Trouble With Gold Stars, Incentive Plans, A's, Praise, and Other Bribes*. Boston, MA. Houghton Mifflin.
— (1994) "The Risks of Rewards", *ERIC Digest*, http://www.ericdigests.org/1995-2/rewards.htm (consultado el 16 Noviembre, 2007).
KOLB, D. (1976) *Learning Style Inventory Technical Manual*. Boston, MA. McBer and Company.
— (1981) "Experiential Learning Theory and the Learning Styles Inventory: A Reply to Freedman and Stumpf", *Academy of Management Review*, 6, págs. 289-296.
— (1984) *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ. Londres. Prentice-Hall.
— (1999) *The Kolb Learning Style Inventory*. Boston, MA. Hay Resources Direct.
— (2000) *Facilitator's Guide to Learning*. Boston. Hay/McBer.
KORETZ, D. (2005) "Alignment, High Stakes, and the Inflation of Test Scores", en J. L. HERMAN y E. H. HAERTEL (eds.) *Uses and Misuses of Data for Educational Accountability and Improvement: 104th Yearbook of the National Society for the Study of Education, Part II*. Malden, MA. Blackwell Publishing.
—, LINN, R., DUNBAR, S. y SHEPARD, L. (1991) "The Effects of High Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests", trabajo presentado en la Annual Meeting of the American Educational Research Association, Chicago, IL.
—, MCCAFFREY, D. y HAMILTON, L. (2001) "Towards a Framework for Validating Gains Under High-Stakes Conditions", *CSE Technical Report*, 551.

LAVE, J. y WENGER, E. (1991) *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK. Cambridge University Press.
LEA, H. (1968 [1870]) *Superstition and Force: Essays on the Wager of Law, the Wager of Battle, the Ordeal, Torture*, 2ª ed. revisada, Nueva York. Greenwood Press.
LEWONTON, R. (1970) "Race and Intelligence", *Bulletin of the Atomic Scientists*, 26, págs. 2-8.
LI, J. (2003) "U.S. and Chinese Cultural Beliefs About Learning", *Journal of Educational Psychology*, 95, págs. 258-267.
LINN, R. L. (2000) "Assessment and Accountability", *Educational Researcher*, 29(2), págs. 4-16.
— (2005) "Issues in the Design of Accountability Systems", en J. L. HERMAN y E. H. HAERTEL (eds.) *Uses and Misuses of Data from Educational Accountability and Improvement.* Chicago, IL. National Society for the Study of Education, págs. 78-98.
LITTLE, A. (1984) "Combating the Diploma Disease", en J. OXENHAM (ed.) *Education Versus Qualifications: A Study of Relationships between Education, Selection for Employment and the Productivity of Labour*, Londres. George Allen & Unwin, págs. 197-228.
— (1997a) "The Diploma Disease Twenty Years On: An Introduction", *Assessment in Education: Principles, Policy and Practice*, 4, págs. 5-22.
— (1997b) "The Value of Examination Success in Sri Lanka 1971-1996: The Effects of Ethnicity, Political Patronage and Youth Insurgency", *Assessment in Education: Principles, Policy and Practice*, 4, págs. 67-86.
LODGE, C. (2001) "An Investigation Into Discourses of Learning in Schools", Londres. Tesis Doctoral inédita. Institute of Education, University of London.
LYNN, R. (1998) "In Support of the Nutrition Theory", en U. NEISSER (ed.) *The Rising Curve: Long-Term Gains in IQ and Related Measures*, 1ª ed., Washington, DC: American Psychological Association, págs. 207-215.
MACAULAY, Lord T. B. (1898) *The Works of Lord Macaulay*, Londres. Longmans. Green & Co.
MACBEATH, J. (1999) *Schools Must Speak for Themselves: The Case for School Self-Evaluation*, Londres. Routledge; National Union of Teachers.
MCDONALD, B. y BOUD, D. (2003) "The Impact of Self-Assessment on Achievement: The Effects of Self-Assessment Training on Performance in External Examinations." *Assessment in Education: Principles, Policy and Practice*, 10, págs. 209-220.
MACGILCHRIST, B. A., REED, J. y MYERS, K. (2004) *The Intelligent School*, Londres. Sage.
MADAUS, G. F. (1988) "The Influence of Testing on the Curriculum", en L. N. TANNER y K. J. REHAGE (eds.) *Critical Issues in Curriculum.* Chicago: 87th Yearbook of the National Society for the Study of Education, University of Chicago Press.
MARSHALL, B. y DRUMMOND, M. (2006) "How Teachers Engage with Assessment for Learning: Lessons from the Classroom", *Research Papers in Education*, 21, págs. 133-149.
MARTON, F. y SÄLJÖ, R. (1976) "On Qualitative Differences in Learning: 1 - Outcome and Process", *British Journal of Educational Psychology*, 46, págs. 4-11.
MARTORELL, R. (1998) "Nutrition and the Worldwide Rise in IQ Scores", en U. NEISSER (ed.) *The Rising Curve: Long-Term Gains in IQ and Related Measures*, 1ª ed., Washington, DC. American Psychological Association, págs. 183-206.
MASLOW, A. (1973) "Deficiency Motivation and Growth Motivation", en D. C. MCCLELLAND y R. S. STEELE (eds.) *Human Motivation: A Book of Readings*. Morristown, NJ. General Learning Press, págs. 126-146.
MATTHEWS, G., ZEIDNER, M. y ROBERTS, R. (2002) *Emotional Intelligence: Science and Myth*. Cambridge, MA, y Londres. MIT Press.
MAYER, J., SALOVEY, P. y CARUSO, D. (2000) "Emotional Intelligence as Zeitgeist", en R. J. STERNBERG (ed.) *Handbook of Intelligence.* Cambridge, R.U. Cambridge University Press, págs. 396-420.

McIntyre, M. (2006) Goodhart's Law, http://www.atm.damtp.cam.ac.uk/people/mem/papers/LHCE/goodhart.html (consultado el 15 Septiembre 2007).
Mercer, N. (2000) *Words and Minds*, Londres. Routledge.
Messick, S. (1989) "Validity", en R. L. Linn (ed.) *Educational Measurement*, 3ª ed., Nueva York, NY. American Council on Education and Macmillan, págs. 13-103.
Morgan, A. (2006) "Feedback: Assessment for Rather Than of Learning", http://bangor.ac.uk/the/documents/FEEDBACKJanuary06.ppt (consultado el 11 Julio 2007).
Neisser, U. (1996) "Intelligence: Knowns and Unknowns", *American Psychologist*, 51, págs. 77-101.
— (1998) "Introduction: Rising Test Scores and What They Mean", en U. Neisser (ed.) *The Rising Curve: Long-Term Gains in IQ and Related Measures.* 1ª ed., Washington, DC. American Psychological Association, págs. 3-22.
Newton, P. (2005) "The Public Understanding of Measurement Inaccuracy", *British Educational Research Journal*, 31, págs. 419-442.
Nichols, S., Glass, G. y Berliner, D. (2005) "High Stakes Testing and Student Achievement: Problems for the No Child Left Behind Act": Education Policy Research Unit, http:edpolicylab.org (consultado el 16 Noviembre 2007).
Nietzsche, F. (1887) *The Gay Science: With a Prelude in Rhymes and an Appendix of Songs*, traducido por Walter Kaufmann. Nueva York. Vintage Books.
Olson, D. (2006) "Becoming Responsible for Who We Are: The Trouble with Traits", en Schaler, J. A. (ed.) *Howard Gardner Under Fire*. Chicago, IL. Open Court.
O'Neill, O. (2002) *A Question of Trust*. Cambridge, R.U. Cambridge University Press.
Pellegrino, P., Chudowsky, N. y Glaser, R. (2001) *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC. National Academies Press.
Perrenoud, P. (1998) "From Formative Evaluation to a Controlled Regulation of Learning Processes. Towards a Wider Conceptual Field", *Assessment in Education: Principles, Policy and Practice*, 5, págs. 85-102.
Phillips, M. (1996) *All Must Have Prizes*. Londres. Little, Brown.
Pirsig, R. M. (1974) *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values*. Nueva York. Morrow.
Price, G. E., Dunn, R. y Dunn, K. J. (1991) *Productivity Environmental Preference Survey: An Inventory for the Identification of Individual Adult Preferences in a Working or Learning Environment, PEPS Manual*. Lawrence, KS. Price Systems.
Ramsden, P. (1983) "Context and Strategy: Situational Influences on Learning", en N. Entwistle y P. Ramsden (eds) *Understanding Student Learning*. Londres. Croom Helm.
—, Beswick, D. y Bowden, J. (1987) "Learning Processes and Learning Skills", en J. T. E. Richardson, M. W. Eysenck y D. W. Piper (eds.) *Student Learning: Research in Education and Cognitive Psychology*. Milton Keynes. Open University Press, págs. 168-176.
Raveaud, M. (2004) "Assessment in French and English infant Schools: Assessing the Work, the Child or the Culture?". *Assessment in Education*, 11 (2), págs. 193-211.
Reay, D. y Wiliam, D. (1999) "'I'll be a nothing': structure, agency and the construction of identity through assessment", *British Educational Research Journal*, 25, págs. 343-354.
Reynolds, M. (1997) "Learning Styles: A Critique", *Management Learning*, 28, págs. 115-133.
Roach, J. P. C. (1971) *Public Examinations in England 1850-1900*. Londres. Cambridge University Press.
Rudduck, J. (1996) "Lessons, Subjects and the Curriculum Issues of 'Understanding' and 'Coherence'", en J. Rudduck, R. Chaplain y G. Wallace (eds.) *School Improvement: What Can Pupils Tell Us?* Londres. Fulton.
Rutter, M., Moffitt, T. E. y Caspi, A. (2006) "Gene-Environment Interplay and Psychopathology: Multiple Varieties But Real Effects", *Journal of Child Psychology and Psychiatry*, 47, págs. 226-261.

SADLER, D. R. (1989) "Formative Assessment and the Design of Instructional Systems", *Instructional Science*, 18, págs. 119-144.

SAINSBURY, M., HARRISON, C. y WATTS, A. (2006) *Assessing Reading from Theories to Classrooms: An International Multi-Disciplinary Investigation of the Theory of Reading Assessment and Its Practical Implications at the Beginning of the 21st Century*. Slough: National Foundation for Educational Research in England and Wales (NFER). Cambridge Assessment.

SELECT COMMITTEE ON SCIENCE AND TECHNOLOGY (2002) "Science Education from 14 to 19: Third Report of Session 2001-2 (Hs-508-I)". Londres. Commons Publications, http://www.publications.parliament.uk/pa/cm200102/cmselect/cmsctech/ 508/508.pdf (consultado el 16 Noviembre 2007).

SFARD, A. (1998) "On Two Metaphors for Learning and the Dangers of Choosing Just One", *Educational Researcher*, 27 (2), págs. 4-13.

SHAYER, M. (2007) "Thirty Years on - a Large Anti-Flynn Effect? The Piaget Test Volume and Heaviness Norms 1975-2003", *British Journal of Educational Psychology*, 77, págs. 1-25.

SHEPARD, L. (1991) "Psychometricians' Beliefs About Learning", *Educational Researcher*, 20, págs. 2-16.

— (1992) "What policy makers who mandate tests should know about the new psychology of intellectual ability and learning", en B. GIFFORD y M. O'CONNOR (eds.) *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction.* Londres. Kluwer Academic Publishers, págs. 301-328.

— (1993) "Evaluating Test Validity", en L. DARLING-HAMMOND (ed.) *Review of Research in Education*, Vol. 19. Washington, DC. American Educational Research Association, págs. 405-450.

— (2000) "The Role of Assessment in a Learning Culture", *Educational Researcher*, 29, págs. 4-14.

SHUTE, V. (2007) "Focus on Formative Feedback", *ETS Research Reports*. Princeton, NJ. Educational Testing Service, http://www.ets.org/research/researcher/RR-07-11.html (consultado el 16 Noviembre, 2007).

SHWERY, C. (1994) "Review of the Learning Styles Inventory", en J. IMPARA y B. BLAKE (eds.) *The Thirteenth Mental Measurements Yearbook*. Lincoln, NE. Buros Institute of Mental Measurements.

SIGMAN, M. y WHALEY, S. (1998) "The Role of Nutrition in the Development of Intelligence", en U. NEISSER (ed.) *The Rising Curve: Long-Term Gains in IQ and Related Measures*, 1ª ed., Washington, DC. American Psychological Association, págs. 155-182.

SLOG (1987) "Why Do Students Learn: A Six Country Study of Student Motivations", *IDS Research Reports*. Brighton. Institute of Development Studies, University of Sussex.

SPEARMAN, C. E. (1904) "'General Intelligence' Objectively Determined and Measured", *American Journal of Psychology*, 15, págs. 201-293.

— (1923) *The Nature of 'Intelligence' and the Principles of Cognition*. Londres. Macmillan.

STATISTICS COMMISSION REPORT Nº. 23 (2005) *Measuring Standards in English Primary Schools*. Londres. Statistics Commission.

STERNBERG, R. J. (1985) *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge, R.U. Cambridge University Press.

STIGGINS, R. J. (2001) *Student-Involved Classroom Assessment*. Upper Saddle River, NJ. Merrill Prentice Hall.

STIGLER, J. W. y HIEBERT, J. (1999) *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom*. Nueva York. Free Press.

STOBART, G. (2005) "Fairness in Multicultural Assessment Systems", *Assessment in Education: Principles, Policy and Practice*, 12 (3), págs. 275-287.

STOBART, G. y GIPPS, C. (1997) *Assessment: A Teachers' Guide to the Issues*. Londres. Hodder and Stoughton.
— y GIPPS, C. (1998) "The Underachievement Debate: Fairness and Equity in Assessment", *British Journal of Curriculum and Assessment*, 8, págs. 43-49.
STOLL, L., FINK, D. y EARL, L. M. (2003) *It's About Learning (and It's About Time): What's in it for Schools?*, Londres. RoutledgeFalmer.
STRAND, S. (2006) "Comparing the Predictive Validity of Reasoning Tests and National End of Key Stage 2 Tests: Which Tests Are the "Best"?" *British Educational Research Journal*, 32, págs. 209-225.
STRAY, C. (2001) "The Shift from Oral to Written Examination: Cambridge and Oxford 1700-1900", *Assessment in Education: Principles, Policy and Practice*, 8, págs. 33-50.
STRONACH, I. y MORRIS, B. (1994) "Polemical Notes on Educational Evaluation and the Age of 'Political Hysteria'", *Evaluation and Research in Education*, 8, págs. 5-19.
STUMPF, S. A. y FREEDMAN, R. D. (1981) "The Learning Style Inventory: Still Less Than Meets the Eye", *Academy of Management Review*, 6 (2), págs. 297-299.
SUTHERLAND, G. (1992) "Examinations, Formal Qualifications and the Construction of Professional Identities: A British Case-Study 1880-1940", trabajo presentado en la IREX-Hungarian Academy of Sciences Conference on Constructing the Middle Class. Budapest.
— (1996) "Assessment: Some historical perspectives", en H. GOLDSTEIN y T. LEWIS (eds.) *Assessment: Problems, Developments and Statistical Issues*. Chichester. John Wiley.
— (2001) "Examinations and the Construction of Personal Identity: A Case Study of England 1800-1950", *Assessment in Education; Principles, Policy and Practice*, 8 (1), págs. 51-64.
TERMAN, L. M. (1916) *The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale*. Boston. Houghton Mifflin Company.
THORNDIKE, E. L. (1922) "Measurement in Education", en G. M. WHIPPLE (ed.) *Intelligence Tests and Their Uses*. Bloomington, IL. Public School Publishing Company.
THURSTONE, L. L. (1938) *Primary Mental Abilities*. Chicago and Londres. University of Chicago Press.
— (1940) "Current Issues in Factor Analysis", *Psychological Bulletin*, 37, págs. 189-236.
— (1946) "Theories of Intelligence", *Scientific Monthly*, 62, Supplement 5, págs. 101-112.
TORRANCE, H. (2005) *The Impact of Different Modes of Assessment on Achievement and Progress in the Learning and Skills Sector*. Londres. Learning and Skills Research Centre.
— y PRYOR, J. (1998) *Investigating Formative Assessment: Teaching, Learning and Assessment in the Classroom*. Buckingham. Open University Press.
TROPE, Y., FERGUSON, M. y RAGHUNATHAN, R. (2001) "Mood as a Resource in Processing Self-relevant Information", en J. P. FORGAS (ed.) *Handbook of Affect and Social Cognition*. Lawrence Erlbaum. New Jersey, págs. 256-274.
TUDDENHAM, R. (1962) "The Nature and Measurement of Intelligence", en L. POSTMAN (ed.) *Psychology in the Making*. Nueva York. Alfred A. Knopf, págs. 469-525.
TYLER, R. W. (1971) *Basic Principles of Curriculum and Instruction*. Chicago y Londres. University of Chicago Press.
TYMMS, P. (2004) "Are Standards Rising in English Primary Schools?" *British Educational Research Journal*, 30 (4), págs. 477-494.
VYGOTSKY, L. S. (1986) *Thought and Language*. Cambridge, MA. Harvard University Press.
WATKINS, C. (2003) *Learning: A Sense-Maker's Guide*. Londres. ATL.
—, CARNELL, E., LODGE, C., WAGNER, P. y WHALLEY, C. (2000) *Learning About Learning*. Londres. Routledge.

WATKINS, C., CARNELL, E., LODGE, C., WAGNER, P. y WHALLEY, C. (2001) "Learning About Learning Enhances Performance", *NSIN Research Matters, Institute of Education. London*, 13 (pamfleto).

— (2000) "Learning and Teaching: A Cross-Cultural Perspective", *School Leadership and Management*, 20 (2), págs. 161-173.

—, CARNELL, E., LODGE, C., WAGNER, P. y WHALLEY, C (2002) "Effective Learning", *NSIN Research Matters*. Institute of Education. Londres.

WATSON, A. (2006) "Some Difficulties in Informal Assessment in Mathematics", *Assessment in Education: Principles, Policy and Practice*, 133, págs. 289-303.

WHITE, J. (2004) *Rethinking the School Curriculum: Values, Aims and Purposes*. Londres. RoutledgeFalmer.

— (2005a) "Puritan Intelligence: The Ideological Background to IQ", *Oxford Review of Education*, 31, págs. 423-442.

— (2005b) *Howard Gardner: The Myth of Multiple Intelligences*. Londres. Institute of Education, University of London.

WIESTRA, R. y JONG, J. D. (2002) "A Scaling Theoretical Evaluation of Kolb's Learning Styles Inventory-2", en M. VALCKE y D. GOMBEIR (eds.) *Learning Styles: Reliability and Validity: Proceedings of the 7th Elsin Conference.* European Learning Styles Information Network Conference, 26-28 Junio. Gante, Bélgica. Ghent University, Department of Education, págs. 431-440.

WILDE, S. (2002) *Testing and Standards: A Brief Encyclopedia*. Portsmouth, NH. Heinemann.

WILIAM, D. (2001) "Reliability, Validity and All That Jazz", *Education*, 29 (3), págs. 3-13; 17-21.

—, LEE, C., HARRISON, C. y BLACK, P. (2004) "Teachers Developing Assessment for Learning Impact on Student Achievement", *Assessment in Education: Principles, Policy and Practice*, 11 (1), págs. 49-66.

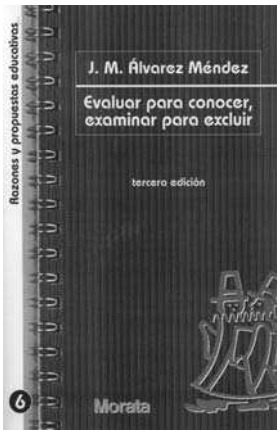
WOLF, A. (2002) *Does Education Matter? Myths About Education and Economic Growth*. Londres. Penguin.

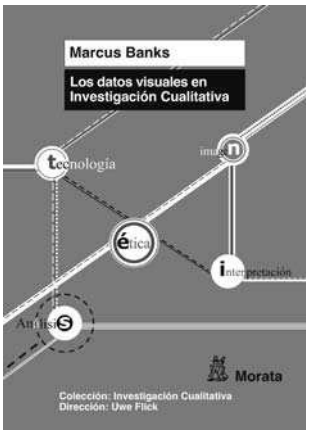
YOUNG, J. (2007) "Predicting College Grades: The Value of Achievement Goals in Supplementing Ability Measures", *Assessment in Education: Principles, Policy and Practice*, 14 (2), págs. 233-249.

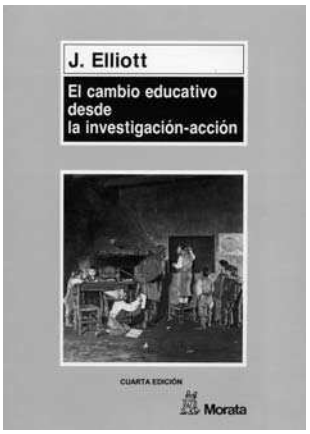
Índice de autores

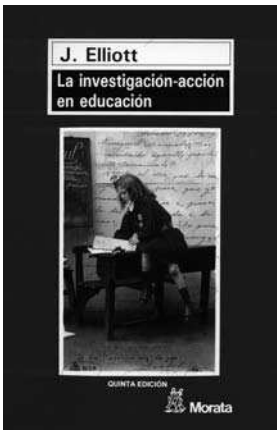
Índice de materias

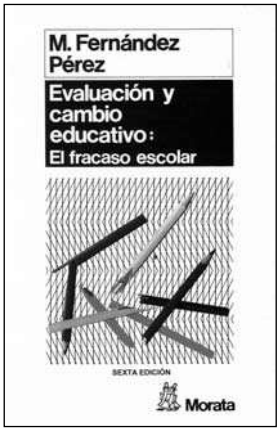
Obras de Ediciones Morata de Investigación y Evaluación

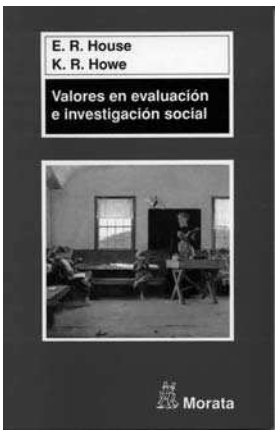


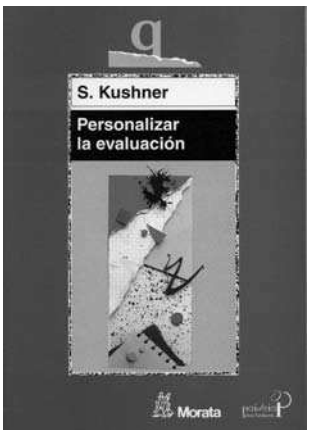


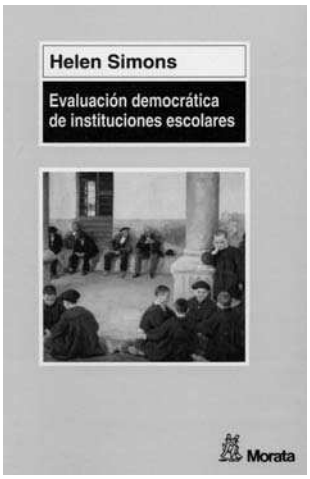


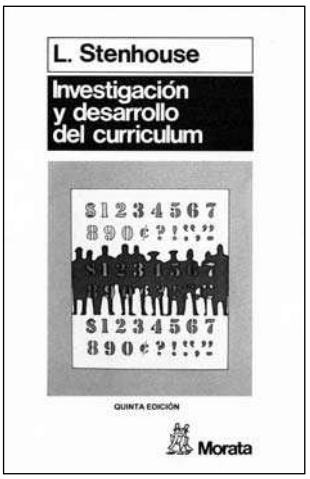


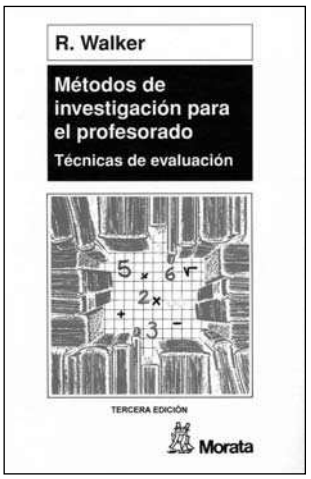








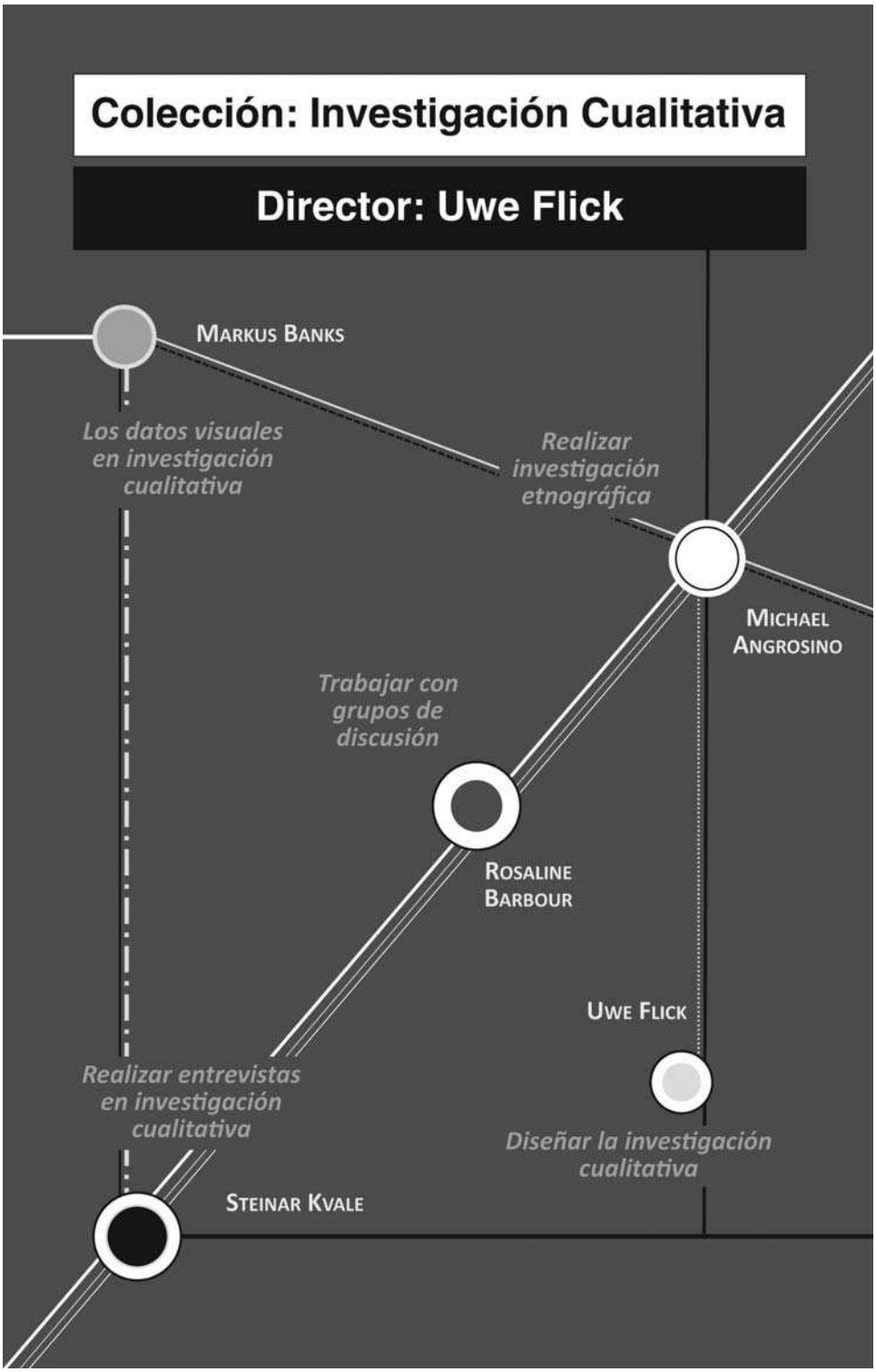




Próxima aparición:

SIMONS, H.: *Estudio de casos. Investigación y práctica.*

KVALE, S.: *Las entrevistas en Investigación Cualitativa.*



¿Qué aportan realmente los tests de inteligencia y de capacidades?
¿Las pruebas objetivas utilizadas en las evaluaciones minan los aprendizajes?

El autor sostiene que la evaluación determina el modo de vernos a nosotros mismos y nuestra forma de aprender. En un momento en el que se clasifica a las personas con arreglo a sus capacidades, estilos de aprendizaje o por su rendimiento, examina con minuciosidad los fundamentos de esas clasificaciones. Muestra también que, en nuestra cultura, donde los tests y pruebas objetivas gozan de una importante aceptación, la evaluación puede socavar a menudo los aprendizajes relevantes, fomentando una instrucción superficial "para el examen" y considerando los resultados de las pruebas como fines en sí mismos.

También investiga cómo podemos preparar evaluaciones que generen aprendizajes más significativos y profundos y que puedan desempeñar un papel constructivo en la creación de nuestras identidades como personas y como aprendices.

Tiempos de pruebas revisa los propósitos y consecuencias de las distintas modalidades de evaluar y examina críticamente los usos más destacados de muchas de ellas, tales como:

- Tests de CI y de capacidades.
- Inteligencias múltiples, inteligencia emocional.
- Estilos de aprendizaje.
- Pruebas de rendición de cuentas y para obtención de certificados y títulos.
- Evaluación formativa.

Gordon STOBART no se contenta con sacar a la luz las dimensiones perversas de ciertas modalidades de evaluación y de medición muy de actualidad, sino que también expone alternativas rigurosas y prácticas sobre cómo debe plantearse una evaluación que sirva para enriquecer los procesos de enseñanza y aprendizaje.

Este libro, accesible y provocativo, es de gran interés para profesionales de la enseñanza, investigadores, profesorado y estudiantes universitarios de educación, psicología y sociología.

* * *

Gordon STOBART, catedrático de Educación en el Instituto de Educación de la Universidad de Londres, es director de la revista internacional *Assessment in Education* y miembro del *Assessment Reform Group*.

ISBN 978-84-7112-629-0

Tema: **Evaluación educativa**

GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN