Algunos límites de la evaluación

Walter H. MacGinitie

La revisión que se hace aquí de toda una serie de limitaciones y de errores habituales en la evaluación educativa ayudará sin duda a los lectores a tomar conciencia del problema y, como propone el autor, a concebir la evaluación más como un método para posibilitar oportunidades que como vía para establecer categorías.

Para guiar lo que hacemos al ayudar a los estudiantes a aprender, evaluamos lo que ya saben y lo que deben saber. Por tanto, la evaluación es una parte importante de la enseñanza. Por desgracia, no es una parte de la enseñanza que podamos desempeñar especialmente bien. La naturaleza humana y los tipos de cosas que debemos evaluar ponen límites a la precisión y la eficacia de las evaluaciones escolares.

PREJUICIOS NORMALES EN LAS APRECIACIONES HUMANAS

Como nuestras evaluaciones comportan apreciaciones, pueden tener en cuenta todo tipo de observaciones significativas. El enseñante tiene una idea de lo mucho que se ha aplicado Vicky en este trabajo. El enseñante ha visto a Ichiro pegándose en el recreo justo antes del examen. Pero como nuestras evaluaciones comportan apreciaciones humanas, también sufren sus debilidades. Y no es solo que errar es humano sino que tener prejuicios también lo es: errar *sistemáticamente* es humano.

El juicio humano, sujeto a errores tanto aleatorios como sistemáticos, desempeña un papel evidente en las evaluaciones de rendimiento y de carpeta.

Comunicación, Lenguaje y Educación, 1993, *19-20*, 25-32

Sin embargo, también las puntuaciones de las pruebas están sujetas a errores de apreciación puesto que una puntuación suele tener poco significado hasta que alguien emite un juicio sobre ella. Alguien decidirá que el percentil 58 de Shelby es esperanzador, o decepcionante, o ninguna de las dos cosas.

Existe una extensa literatura sobre los errores de apreciación que cometen las personas. Aunque muchos de estos estudios se centraron originalmente en cuestiones relativas a la percepción, el razonamiento, la formación de categorías, la estimación de probabilidades y otras áreas, presentan implicaciones para los juicios en clase. Unos cuantos ejemplos pueden ilustrar lo importante que es para nosotros ser cautos en nuestras evaluaciones y estar preparados para reconsiderar nuestras decisiones.

ASIMILACION Y CONTRASTE

En un experimento típico (Jones, Rock, Shaver, Goethals y Ward, 1968), grupos de observadores adultos observaban a un varón adulto desconocido que intentaba resolver una serie de 30 problemas de razonamiento. Sin que lo supieran los observadores, la persona que intentaba resolver los problemas en realidad estaba conchabado con los experimentadores y resolvía o fallaba los problemas siguiendo una secuencia previamente determinada. El sujeto observado siempre resolvía correctamente 15 de los problemas, pero en algunos grupos los observadores le vieron resolver más de los primeros problemas y menos de los últimos, mientras que otros grupos le vieron resolver menos de los primeros y más de los últimos.

Los observadores que vieron al sujeto resolver más de lo primeros problemas lo consideraron más inteligente que los observadores que le vieron resolver más de los últimos. Por tanto, los observadores evaluaron la capacidad del observado al principio, sin modificarla después. *Asimilaron* la evidencia posterior a su primera impresión. Esta conclusión se ve reforzada por el recuerdo distorsionado de los observadores en cuanto al número de problemas resueltos por la persona observada. Aunque el sujeto siempre resolvía 15 problemas, los observadores que le habían visto resolver problemas al principio recordaban haberle visto resolver más problemas que los observadores que le habín visto resolver problemas al final. Tenemos una tendencia general a ignorar evidencias posteriores y a basar nuestras evaluaciones en las primeras evidencias siempre que presuponemos que estamos observando una característica estable (Jones y Goethals, 1972).

Y somos conscientes de que muchos logros educativos son muy estables. Por ejemplo, sabemos que un buen lector o escritor no suele convertirse en un lector o escritor malo de una semana a otra y ni siquiera de un año a otro. Pero la aptitud para la lectura o la escritura *puede* cambiar. Además, el rendimiento en lectura o escritura de cualquier estudiante puede variar considerablemente de una tarea a otra o de un momento a otro. Hasta un buen futbolista comete errores y, de vez en cuando, un futbolista mediocre hace un gran partido. Sin embar-

go, muchas de nuestras evaluaciones de la capacidad para la lectura o la escritura suelen estar sesgadas en favor de las primeras impresiones.

Esta tendencia a asimilar -a hacer que nuestros juicios posteriores encajen con los iniciales- es solo una de muchas tendencias que pueden desvirtuar nuestras evaluaciones. También tenemos la tendencia a *exagerar la diferencia* entre las primeras evidencias y las evidencias posteriores. Es probable que este efecto de *contraste* se produzca cuando evaluamos una serie de productos que consideramos totalmente independientes entre sí, como cuando son hechos por distintos estudiantes (Jones y Goethals, 1972). Así, es probable que un trabajo malo parezca peor de lo que es si lo vemos después de haber visto un trabajo bueno de otro estudiante. Y es probable que un buen trabajo parezca mejor de lo que es, si lo vemos después de haber examinado un trabajo malo de otro estudiante.

SESGO DE NEGATIVIDAD

Nuestros juicios humanos también están *sometidos a sesgos de negatividad*, primeramente demostrados de una manera formal por Asch (1946). Asch pidió a unas personas que se formaran una impresión de una persona a la que se describía mediante una lista de rasgos. Por ejemplo, para un grupo de personas el sujeto fue descrito como "inteligente, hábil, aplicado, cálido, determinado, práctico y prudente". Para otro grupo la misma persona fue descrita como "inteligente, hábil, aplicada, fría, determinada, práctica y prudente". La simple sustitución de la palabra *cálida* por *fría* creó grandes diferencias en la impresión que las personas consultadas tenían del sujeto descrito. Asch encontró que determinados adjetivos negativos tenían unos efectos desproporcionados en las impresiones que se formaban.

Posteriores experimentos demostraron que la información negativa suele influir en las evaluaciones con mucha más fuerza que la información positiva (Kanouse y Hanson, 1972). Evidentemente, debemos procurar que informaciones ocasionales decepcionantes no distorsionen nuestras evaluaciones.

SESGO DE CATEGORIA

Los sesgos de asimilación y de negatividad suelen combinarse en los efectos de las categorías que utilizamos. En cuanto hemos colocado a una persona en *cualquier* categoría -joven o vieja, hombre o mujer, brillante o lento, buen lector o mal lector- es probable que el contenido estereotípico de la categoría influya en nuestra evaluación (Fiske y Taylor, 1984). La ocupación del padre de un estudiante, por ejemplo, puede sesgar nuestras creencias sobre el entorno familiar del estudiante. En un estudio, unas personas vieron una cinta de vídeo de una mujer cenando con su esposo. Las personas a las que se les dijo que la mujer trabajaba como camarera tendieron más a recordar que bebía cerveza y que tenía un aparato de televisión que las personas a las que se dijo que trabajaba como bibliotecaria. Pero las personas a las que se dijo que trabajaba

como bibliotecaria tendieron más que las otras a recordar que llevaba gafas y que tenía discos de música clásica (Cohen, 1981).

Por tanto, cuando colocamos a una persona en una categoría tendemos a privarla de parte de su individualidad. Y es incuestionable que colocamos a los niños en categorías: disléxicos, retrasados, buenos lectores, emocionalmente perturbados, ... la lista podría seguir mucho más. Si de un niño solo supiéramos que pertenecía a una de estas categorías -por ejemplo, emocionalmente perturbado- ¿no tendríamos a veces la tentación de pensar que también sabríamos *otras* cosas de ese niño?

SESGO DE CONFIRMACION

Es evidente que, en cuanto se nos mete una idea en la cabeza, tendemos a ignorar otras posibilidades. Una razón es que las personas tienden a no comprobar sus creencias. Este *sesgo de confirmación* (Evans, 1989) aparece en el trabajo correctivo como, por ejemplo, cuando decidimos que el problema de un estudiante tiene una causa determinada y luego no nos ponemos a buscar evidencias en contra, evidencias de causas adicionales o diferentes.

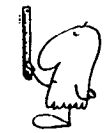
Y hasta la evaluación de nuestras evaluaciones llega a estar sesgada: Tendemos a tener una confianza excesiva en nuestras evaluaciones; sobrevaloramos lo que sabemos y las veces que tenemos razón (Evans, pp. 97-99, 103). No es de sorprender: Como raramente comprobamos nuestras creencias, rara vez descubrimos que nuestro juicio ha sido erróneo.

EL PROBLEMA DE LA SELECCION DE LA MEDIDA Y EL DILEMA DE LA VALIDEZ

Las evaluaciones también estan limitadas por la dificultad de elegir qué es lo que se debe evaluar y cómo se debe evaluar: *El problema de la selección de la medida*. Hay muchas cosas que esperamos que aprendan los estudiantes y podríamos elegir entre muchos métodos para intentar evaluar cada uno de estos aprendizajes.

Por ejemplo, ¿cuántos aspectos significativos tiene la capacidad de lectura? ¿Cincuenta? ¿Cincuenta mil? Usando una matriz de estrategias, tipos de retórica, materias, estructuras semánticas y finalidades de la lectura, fácilmente podríamos desarrollar una lista con miles de aspectos (MacGinitie, 1990). ¿Qué aspectos deberíamos optar por evaluar?

Una respuesta a este problema que ahora se recomienda con frecuencia es guardar muchas muestras del trabajo de un alumno en una carpeta. Al parecer, algunos afirman que las carpetas resuelven el problema de la selección de la medida en la evaluación de la lectura midiéndolo todo; dicen que pueden ofrecer "una imagen *completa* de la capacidad de lectoescritura de un estudiante" (Valencia, 1990, p. 340; cursiva en el original). Pero las carpetas no pueden ofrecer realmente una medida para todo; las carpetas no solucionan el problema de la selección de la medida, pero sí que ilustran el dilema de la validez.



El *dilema de la validez* se refiere a la observación de que, con frecuencia, cuando tratamos de incrementar la validez de las evaluaciones, reducimos su fiabilidad. Sin embargo, para que una evaluación sea válida -para que mida bien lo que se supone que debe medir- también debe ser fiable, es decir, debe medir *algo* de una manera razonablemente coherente.

La fiabilidad suele ser un problema en la evaluación a base de carpetas. ¿Cómo se pueden combinar las impresiones producidas por los diversos elementos contenidos en una carpeta? ¿Debe limitarse el evaluador a examinar todos los contenidos y asignar una puntuación global en base a los elementos que se recuerden al final del examen? Es probable que dos evaluadores destaquen elementos distintos y que, en consecuencia, puntúen los elementos de una manera distinta y lleguen a valoraciones globales completamente diferentes. La baja fiabilidad resultante limita la validez de las evaluaciones basadas en carpetas llevadas a cabo de esta manera.

Un método alternativo para evaluar una carpeta es asignar una puntuación a cada elemento y luego combinar estas puntuaciones de una manera sistemática. ¿Pero cómo se deberían combinar estas puntuaciones? ¿Son algunos elementos el doble de importantes que los demás? ¿Diez veces más importantes? ¿Son algunos elementos más importantes para unos estudiantes que para otros? ¿Sobre qué base se podrían justificar las respuestas a estas preguntas? Distintos conjuntos de ponderaciones podrían otorgar valoraciones muy distintas al trabajo de un mismo estudiante.

Una solución a este problema podría ser que todos los enseñantes utilizaran los mismos criterios de ponderación. Pero, de ser así, al evaluar una carpeta solo cabría incluir los elementos a los que se hubiera dado peso. En realidad, suele aconsejarse que solo se puntúe un conjunto previamente seleccionado y estandarizado de elementos de una carpeta para enjuiciar el desarrollo de la capacidad de lectura de cada estudiante: "Es importante ser *selectivo* acerca de lo que se debe incluir en la carpeta" (Valencia, 1990, p. 339).

Pero cuando se produce esta selección, una carpeta ya no puede pretender ofrecer "una imagen *completa* de la capacidad de lectoescritura de un estudiante". Por ejemplo, la explicación espontánea que da un estudiante sobre cómo ha trabajado con un ejercicio, cómo ha interpretado un pasaje o cómo ha solucionado un problema puede ser uno de los elementos más reveladores de una carpeta. Pero esta clase de información ni siquiera llegaría a ser tenida en cuenta si la evaluación se basara en un conjunto predefinido y estandarizado de materiales. Por tanto, tratar de resolver el dilema de la validez mediante el empleo de un conjunto estandarizado de ponderaciones nos hace volver directamente al problema de la selección de la medida.

Otra limitación común aunque innecesaria de la validez de las carpetas "de lectura" aparece cuando se incluyen en ellas muchas muestras de escritura o "piezas de escritura en diversas etapas de realización" (Valencia, 1990, p. 339). Naturalmente, la lectura y la escritura están estrechamente relacionadas, pero no son la misma cosa. Algunos estu-

diantes que son excelentes lectores tienen una capacidad relativamente limitada para expresar sus pensamientos, y las muestras de escritura pueden ser muy engañosas a la hora de evaluar la capacidad de lectura.

Las carpetas *pueden* ser muy útiles para la evaluación. Pueden relacionar estrechamente la evaluación con la enseñanza. Desempeñan la función, especialmente importante, de estimular a enseñantes y alumnos a registrar observaciones significativas sobre los éxitos, las dificultades y las maneras de trabajar de los estudiantes que, de no ser por esto, podrían ser olvidadas o ignoradas al planificar la enseñanza. Pero para que las carpetas se puedan emplear para llegar a una puntuación, se debe dedicar una gran cantidad de tiempo y de conocimientos a la planificación y la recopilación de colecciones, y al diseño y la realización de los procedimientos para puntuarlas.

EL SIGNIFICADO DE LA EVALUACION PARA EL ESTUDIANTE

Podemos abrigar la esperanza de influir en los estudiantes con nuestras evaluaciones, pero no podemos estar seguros del significado que nuestras evaluaciones tienen para ellos. Evaluamos el trabajo de un alumno en función de lo que nosotros buscamos como enseñantes, pero el alumno puede juzgar este trabajo en base a criterios distintos. Gordon (1990) estudió los criterios que utilizan los enseñantes y los estudiantes de sexto curso al puntuar la calidad de relatos breves escritos por otros estudiantes. Aunque el estudio está limitado por una muestra muy pequeña de enseñantes, los resultados son intrigantes. Los estudiantes acordaron entre sí perfectamente los criterios para considerar que un relato era bueno: Un buen relato debía ser imaginativo, excitante y con un argumento bien desarrollado. En cambio, los enseñantes tenían criterios que variaban mucho de uno a otro y algunos tenían criterios muy distintos a los de los estudiantes.

Así, un estudiante puede estar orgulloso de algo que a nosotros nos pasa desapercibido o estar insatisfecho con algo que encontramos aceptable. Para el estudiante, nuestra evaluación puede ser sorprendente, decepcionante o, quizá, simplemente irrelevante. "Es seguro que el juicio del espectador", dice William James (1962) "pasará por alto la raíz de la cuestión ... El sujeto juzgado conoce una parte del mundo real que el espectador que juzga no puede ver" (p. 114). Nuestras evaluaciones no pueden abarcar las alegrías y las penas personales de un estudiante ni los secretos que una mente guarda para sí.

Al reflexionar sobre los límites de la evaluación descubrimos que, independientemente de lo cuidadosos que seamos, muchos de nuestros juicios estarán sesgados, que no podemos tener la esperanza de evaluar muchas cosas que son importantes, que nuestros procedimientos de evaluación, por muy realistas que queramos que sean, tendrán una validez limitada, y que nunca podemos estar seguros del significado que tienen nuestras evaluaciones para los estudiantes que son evaluados.

Puesto que nuestras evaluaciones son falibles y limitadas, las decisiones basadas en ellas deberán ser prudentes. No hay muchas decisiones sobre los estudiantes que *deban* ser *definitivas.* Casi todas las decisiones deberían ser reconsideradas periódicamente. Un programa que en su día parecía correcto debería volver a ser evaluado eventualmente. Un diagnóstico que en su día parecía correcto debería volver a ser examinado. Sobre todo, un estudiante que no pase a la primera debería tener otra oportunidad.

Deberíamos dejar de concebir la evaluación como un método para establecer categorías y empezar a concebirla como un método para establecer oportunidades. Si nuestra evaluación nos dice que un estudiante decesita ayuda en alguna tarea escolar, constituye una oportunidad para que nosotros podamos ayudar y para que el estudiante se desarrolle. Si nuestra evaluación nos dice que un estudiante ha alcanzado un nivel destacado, se trata de una oportunidad para que nosotros compartamos su alegría y para que el estudiante explore nuevas sendas para el logro. Para utilizar la evaluación con el fin de establecer oportunidades, necesitaremos modificar algunas actuaciones políticas y administrativas además de nuestra propia manera de pensar. Conocer los límites de la evaluación nos puede ayudar a conseguir estos cambios.

Referencias

ASCH, S. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258-290.

COHEN, C. E. (1981). Person categories and social perception: Testing some boundaries of the processing effects of prior knowledge. *Journal of Personality and Social Psychology, 40,* 441-452.

EVANS, J. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

FISKE, S. T. Y TAYLOR, S. E. (1984). *Social cognition.* Reading, MA: Addison-Wesley.

GORDON, C. J. (1990). Students' and teachers' criteria for quality writing: Never the twain shall meet? *Reflections on Canadian Literacy, 9,* 74-81.

JAMES, W. (1962). *Talks to teachers on psychology and to students on some of life's ideals.* Nueva York: Dover. (Obra original publicada en 1899).

JONES, E. E. Y GOETHALS, G. R. (1972). Order effects in impression formation: Attribution context and the nature of the entity. En E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins y B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 27-46). Morristown, NJ: General Learning Press.

JONES, E. E., ROCK, L., SHAVER, K. G., GOETHALS, G. R. Y WARD, L. M. (1968). Pattern of performance and ability attribution: An unexpected primary effect. *Journal of Personality and Social Psychology, 10,* 317-340.

KANOUSE, D. E. Y HANSON, L. R., JR. (1972). Negativity in evaluations. En E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins y B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47-62). Morristown, NJ: General Learning Press.

MACGINITIE, W. H. (1990, marzo). *What's the score in reading assessment?* Ponencia presentada en la Texas Testing Conference, Austin, TX.

VALENCIA, S. (1990). A portfolio approach to classroom reading assesment: The whys, whats, and hows. *The Reading Teacher,* 43, 338-340.

Algunos límites de la evaluación

Walter H. Macginitie

CL&E, 1993, 19-20, pp. 25-32

Datos sobre el autor: MacGinitie, en su día presidente de la International Reading Association, se dedica actualmente a escribir y a tareas de consultoría.

Artículo original: Some limits of assessment, en *Journal of Reading, Vol. 36*, Nº 7, pp. 556-560. Traducción de Genís Sánchez. Reproducido con autorización de Walter H. MacGinitie y de la International Reading Association (La I.R.A. no se responsabiliza de la adecuación de la traducción)

Dirección: PO Box 1789, Friday Harbor WA 98250, EEUU.