anales de psicología
2011, vol. 27, nº 1 (enero), 249-255

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Sobre un método comúnmente utilizado para la puntuación de tests en la evaluación del rendimiento educativo

Miguel A. Mateo García*

Universidad Complutense de Madrid (España)

Resumen: El presente trabajo constituye una reflexión sobre la utilización de una conocida fórmula de puntuación para tests en el contexto de la evaluación del rendimiento educativo. Inicialmente, se introducen ideas básicas acerca de los tests y de los motivos para utilizar, eventualmente, una fórmula de puntuación asociada a su empleo como instrumentos para registrar información. Sin entrar en los pros y contras del uso de tests en el contexto de la evaluación del rendimiento educativo, y con base en trabajos representativos de voces autorizadas, se presentan, seguidamente, algunas observaciones críticas respecto de las posibilidades y el uso real de la fórmula de puntuación considerada en ese contexto. Dado que este uso es, aparentemente en contra de toda lógica, absolutamente mayoritario y acrítico, aquí y ahora, se sugiere la posibilidad de que tenga su raíz en la necesidad experimentada comúnmente por el ser humano de buscar procedimientos reglados, automáticos ("recetas") que guíen, aligeren la carga y liberen de la responsabilidad personal a la hora de asumir riesgos en la toma de decisiones ante las incertidumbres de la existencia. En este orden de cosas, no parece descabellado entender que esta postura ocupa el centro de la escena en el drama de buena parte de la ciencia, hoy en día, donde una notable cantidad de científicos contemporáneos está abusando de la *modelización formal* y la *cuantificación* (números a toda costa) como estrategias únicas para lograr su principal cometido (que no puede ser otro que ir dando cuenta de alguna parte de la realidad). Volviendo a lo más concreto, y para tratar de cerrar el círculo, se termina proponiendo, por enésima vez en la literatura especializada, la más sencilla y clásica de las alternativas para puntuar un test.
Palabras clave: Fórmula de puntuación; test; evaluación científica del rendimiento educativo; abuso de modelos formales; ciencia y realidad

Title: On a common formula scoring method in educational achievement testing.
Abstract: This paper tries to be a reflection on using test formula scoring for the assessment of educational achievement. To start, basic ideas about tests and putative reasons/conditions for using formula scoring are shown. Then, some criticisms, based on selected workings from reliable sources, are raised in respect of using formula scoring in the context of educational assessment. As a suggestion, it is proposed that nowadays overwhelming use of formula scoring may have its roots in human need of patterned procedures for decision making, in order to get a certain relief against uncertainty. Moreover, some reflection is suggested on formal modelling and quantification as a rather compulsory practice (numbers at any cost) by a lot of today scientists. Finally, a simple, classic, universal alternative for test scoring is proposed.
Key words: Formula scoring; tests; scientific educational assessment; formal modelling abuse; science and reality.

Introducción

De un tiempo a esta parte, estoy siendo testigo, involuntario aunque no totalmente marginal, de lo que puede considerarse como una verdadera contraposición de hecho de dos formas profundamente distintas, incluso parece que incompatibles, de entender una parte tan central del desempeño educativo como es la evaluación de conocimientos, y del rendimiento en general. Mientras que para algunos de los involucrados, voces aparentemente minoritarias, el asunto es básicamente contextual: requiere una consideración amplia de objetivos, contenidos, procedimientos, características de aquellos a quienes va dirigida la labor educativa, etc., para otros la evaluación debe ser algo objetivo, que pueda ser resumido en términos de un (¿único?) procedimiento estándar y un resultado expresado numéricamente, por medio de una fórmula universal. La impresión que uno no puede evitar es que esta contraposición no es sino un capítulo más de la difícil coexistencia, a que alude Rivière (2009), entre el posmoderno *saber fragmentado* o *experto* que, con pretextos relativos a la creciente complejidad del mundo actual y la consiguiente necesidad de especialización para desenvolverse en él, se ha constituido en paradigma contemporáneo ineludible de cualquier tipo de saber, y el ejercicio más básico de la inteligencia humana, que consiste fundamentalmente en *atar cabos* que nos permitan ir dando cuenta de lo que sucede alrededor. Ese saber experto, por su parte, aspira cada vez más a ser cuantitativo, y, en consecuencia, a formularse en términos matemáticos, o al menos numéricos.

Dado que uno, como individuo de su tiempo, se ve afectado por el espíritu de la época, lo quiera o no y en muchas más circunstancias de lo que desearía, este trabajo trata de ser básicamente una reflexión, apoyada en fuentes autorizadas, sobre algunos aspectos del papel de las fórmulas para la puntuación del rendimiento educativo, tanto desde el punto de vista del experto como, sobre todo, desde la perspectiva de alguien que intenta, desde hace mucho y con resultados irregulares, conseguir que vayan *encajando* de la mejor manera posible *piezas del puzzle*. Comoquiera que, en la práctica, es extremadamente frecuente la utilización para la evaluación del rendimiento en la educación de instrumentos del tipo conocido como de prueba objetiva o test, parece oportuno dar comienzo al desarrollo del trabajo con una breve introducción a algunos aspectos relevantes de los tests

* **Dirección para correspondencia [Correspondence address]:** Miguel A. Mateo García. Dpto. de Metodología de las Ciencias del Comportamiento. Despacho D 2123-O. Facultad de Psicología, Universidad Complutense de Madrid. Campus de Somosaguas, 28223 Pozuelo de Alarcón, Madrid (España). E-mail: mmateo@psi.ucm.es

Los tests como instrumentos para la valoración de lo no observable

Los tests son instrumentos construidos para la recogida sistemática de información sobre un aspecto de la realidad que no es directamente observable, originados y desarrollados principalmente en el seno de la Psicología. El uso de un test descansa sobre varios supuestos: (1) en un determinado momento, para cada sujeto (o *unidad de observación*), el aspecto de la realidad que se pretende registrar se manifiesta de un modo supuestamente característico y estable (al menos temporalmente), "verdadero"; (2) este *nivel característico* no es directamente observable; se expresa en términos de un comportamiento prototípico (indicio), de manera que entre nivel (latente) y comportamiento (observable) se da una correspondencia *uno a uno*; (3) es posible establecer también una correspondencia uno a uno entre cada comportamiento y un valor numérico, y, por tanto, entre cada nivel característico y un valor numérico: esto sería lo que significaría *medir*. Un test se propone ser, en principio, un instrumento para la medición de *variables* o *constructos latentes* de interés para la Psicología, que se emplea frecuentemente también en ámbitos más o menos afines.

Un test típico, de los que se utilizan, por ejemplo, en la valoración de la inteligencia o las presuntas aptitudes cognoscitivas en los seres humanos, está formado por un cierto número de elementos o ítems, cada uno de los cuales constituye un *estímulo* para tratar de poner de manifiesto un comportamiento que refleje un matiz característico del aspecto de la realidad que se pretende registrar. En un test típico tal, cada ítem suele constar de un enunciado (o una representación gráfica, etc.) presentado por escrito, que lleva asociadas varias propuestas, expresadas en el mismo formato o similar, y que se presentan como opciones alternativas de respuesta posibles (test de papel y lápiz); únicamente una de estas tiene por definición, por razones *teóricas*, la consideración de correcta o *acierto*, mientras que todas las restantes posibilidades tienen un mismo estatus de *error* (*ítem de elección múltiple*). En general, lo que se requiere que haga un sujeto ante un test tal es tratar de elegir la alternativa de respuesta que considera correcta (tarea de reconocimiento) para todos y cada uno de los ítems que lo forman.

Una teoría psicométrica de tests constituye un intento de fundamentar la medición de variables psicológicas mediante este tipo de instrumentos específicos. La Teoría Clásica de los Tests, cronológicamente la primera que surge, encuentra su base en el modelo de Spearman, un modelo de medida originado en el seno de uno de los intentos primeros de medir una supuesta capacidad individual: la inteligencia (innata, para muchos de los pioneros, y, a partir de cierta edad, prácticamente fija durante gran parte del ciclo vital –hasta la progresiva declinación debida al envejecimiento-). La Teoría Clásica de los Tests aporta a la Psicología un conjunto de conceptos y técnicas que le permiten estimar la calidad de sus datos (fiabilidad, validez…). En este marco, surge un procedimiento extremadamente simple e intuitivo para atribuir a un sujeto una puntuación (X) que represente su rendimiento ante un test formado por ítems de elección múltiple: consiste en contar (sumar) el número de aciertos (A) que ha logrado (X=A). Gulliksen (1950), propone utilizar este procedimiento o esquema de puntuación siempre que se verifique una doble condición: (1) las instrucciones que se den al administrar el test han de poner énfasis en que los sujetos deben responder a los ítems de forma consecutiva, en el orden en que aparecen presentados, lo que implica que el número de ítems que *vayan quedando* sin responder (*omisiones intermedias,* O_i) debe ser cero o insignificantemente pequeño, y (2) se debe permitir el tiempo suficiente para que casi todos los ítems puedan ser respondidos por casi todo el mundo, lo que implica que el número de ítems que queden sin *intentar* responder (*omisiones finales*: O_f) sea también pequeño (Gulliksen, 1950, p. 246)[1].

En principio, cuando se le presenta un ítem de elección múltiple a un sujeto, este puede, en la práctica, (a) elegir la alternativa correcta como respuesta, (b) elegir alguna de las alternativas no correctas, o (c) decidir no pronunciarse al respecto (*omisión* o "respuesta en blanco"). El comportamiento del sujeto en cada caso concreto depende, fundamentalmente, de que verdaderamente conozca con certeza la respuesta apropiada o no la conozca. Si se da esta segunda circunstancia, van a desempeñar un papel central aspectos de la personalidad del sujeto, tales como su naturaleza conservadora o cauta, su disposición a asumir riesgos, etc., así como, en ocasiones, pueden influir condiciones o características particulares del proceso de aplicación del test. Evidentemente, cuando un sujeto se encuentra ante un ítem para el que no conoce con certeza la respuesta correcta, puede tratar de aventurarla, incluso aunque haya recibido instrucciones específicas para no hacerlo; si no dispone siquiera de ningún indicio orientativo, la elegirá *al azar* (la tratará de "adivinar"). Esto significa que el número de aciertos *podría* constituir una sobreestimación de la puntuación *verdadera* de un sujeto en un test (su "verdadero" nivel, supuestamente característico y estable –al menos temporalmente-, en aquello que se pretende medir).

Puntuación mediante fórmula

Los psicómetras han diseñado, desde los primeros tiempos, procedimientos para reducir las irregularidades que podrían acompañar a la puntuación de un sujeto en un test formado por ítems de elección múltiple debido a un proceso tal de *adivinación*. Básicamente, se trataría de *compensar* o *descontar* (estadísticamente) ese efecto de *adivinación*, mediante la utilización de una *fórmula de puntuación* (*formula scoring* o puntua-

[1] Este autor señala que también se podría aplicar el procedimiento aunque el número total de ítems que queden sin respuesta sea grande, siempre que el sujeto lea cada ítem y trate de buscar honradamente la respuesta correcta (p.246). En resumen, siempre que sea razonable confiar en que el conocimiento del sujeto de las respuestas correctas a los ítems quede bien reflejado por el número de aciertos.

ción mediante fórmula). Habitualmente, una fórmula de puntuación resta del número de aciertos una cierta proporción del número de errores. La versión más conocida y utilizada, que en lo sucesivo será denominada *fórmula de puntuación* (sin más), es

$$X = A - E/(k-1),$$

donde,
X = puntuación atribuida a un sujeto en el test,
A = número de ítems respondidos correctamente (aciertos),
E = número de ítems respondidos de un modo no correcto (errores)
K = número de alternativas de respuesta posibles para cada ítem (igual para todos ellos).

Como sucede con toda representación (modelo) matemática de un aspecto de la realidad, el modelo de puntuación que se expresa en tal fórmula sólo es aplicable propiamente (en rigor) bajo ciertas condiciones o supuestos, de naturaleza previa, que delimitan y aseguran su validez y que, por tanto, obligan al usuario. En este caso, esos supuestos son:

1) Un sujeto, o bien conoce *con certeza* la respuesta correcta a un ítem o no la conoce *en absoluto* (todo o nada).
2) En el primer caso, *necesariamente* da esa respuesta correcta (acierto). En el segundo caso, *puede* que trate de elegir al azar (adivinar) una opción de entre todas las k posibles; entonces, la probabilidad de que elija una cualquiera de estas en concreto es, evidentemente, 1/k. Es decir, la probabilidad de que elija por azar la respuesta correcta es 1/k, y la probabilidad de que cometa un error es, por tanto, [k-1]/k.
3) Cuando un sujeto no responde a un ítem, no es posible considerar ni que lo acierta ni que comete un error (omisión o respuesta en blanco).

La mayoría de los autores -y de las utilizaciones de la fórmula en la práctica- requieren que, al ser administrado el test, los sujetos reciban la instrucción de responder a un ítem únicamente si conocen con certeza la respuesta correcta; es decir, se "prohibe" la conjetura, guiada por un conocimiento parcial o presuntamente impreciso. Bajo tales circunstancias, la fórmula de puntuación pretende *eliminar* sobreestimaciones de las puntuaciones debidas a la mera adivinación afortunada. Según Gulliksen, el propósito de la utilización de la fórmula seria "…estimar el número de ítems para los que el sujeto conoce realmente la respuesta correcta" (Gulliksen, 1950, p.248). Quienes abogan por la utilización sistemática de la fórmula aducen que permite reducir la varianza error debida a esa adivinación (en particular, para los sujetos con niveles más bajos en la variable que se pretende medir), y, en consonancia, mejorar las propiedades psicométricas (fiabilidad y, por extensión, validez) de los datos.

Con un planteamiento algo diferente, Frary (1988) propone que al administrar el test, a los sujetos se les debe dar la instrucción de que no respondan a un ítem únicamente en el caso de que no tengan la más remota idea de cuál podría ser la respuesta correcta, indicándose que han de intentar responder si tienen algún indicio, aunque no estén absolutamente seguros. Se trataría de evitar la adivinación pura y dura, pero animando a los sujetos a utilizar la información parcial de que dispongan, por ejemplo para eliminar alguna opción errónea (*adivinación guiada*), de manera que se eviten subestimaciones en las puntuaciones por no haber arriesgado con cierto fundamento y en beneficio propio (Frary, Cross, & Sewell,1985, hablan de que impedir el uso de información parcial es injusto y no equitativo, puesto que limita a los sujetos al no permitirles poner en juego todas sus capacidades). En este sentido, Frary defiende que la fórmula de puntuación está dirigida meramente a ajustar las puntuaciones, "descontando" las ganancias estimadas como debidas exclusivamente al azar (como consecuencia de la ignorancia absoluta), pero no a penalizar la adivinación en general (de toda clase). Para este autor, es un error interpretar el resultado del uso de la fórmula como una estimación de la puntuación que un sujeto obtendría si no fuese posible la adivinación en modo alguno.

En general, autores como el ya citado Frary, o Rowley & Traub (1977), señalan la inconsistencia de los resultados encontrados en torno a la supuesta mejora de las propiedades psicométricas de las medidas asociada a la utilización de la fórmula de puntuación (de hecho, para reducir la varianza error, y, así, mejorar la fiabilidad, proponen como solución más favorable el aumento de la longitud del test añadiendo ítems *fáciles*, a los que sean capaces de responder correctamente muchos sujetos).

La puntuación mediante fórmula en la evaluación del rendimiento educativo

Centrándonos en un posible ámbito concreto de aplicación, la investigación empírica llevada a cabo al efecto no ha demostrado consistentemente que existan notorias ventajas o desventajas que se deriven del uso de la fórmula de puntuación en la evaluación del rendimiento educativo mediante tests del tipo aquí considerado. No es lo *normal* que un estudiante que se va a someter a examen de una materia no tenga ni idea de la misma, que ignore absolutamente las respuestas a todos o a muchos de los ítems que formen un test/examen que sea representativo (que tenga adecuada validez de contenido), y para el que se han impartido instrucciones apropiadas; más bien al contrario, lo que cabe esperar es que la mayor parte de los sujetos tengan al menos conocimientos parciales sobre bastantes de los aspectos de la materia. Esto implica, por una parte, que no parece muy plausible que, en condiciones normales, se de una circunstancia clave para el uso de la fórmula: que el número de ítems sin respuesta sea muy grande para algunos sujetos y muy pequeño para otros. Por otra parte, cuando lo razonable es esperar que los sujetos no ignoren por completo las respuestas a los ítems, no parece que el uso de la fórmula de puntuación proporcione estimaciones mejores de los resultados en el test que el simple número de aciertos. Gulliksen

(1950) descarta, incluso, el uso de la fórmula cuando la dificultad de los ítems es tal que la proporción de sujetos que los pueden responder correctamente por azar es menor que 1/k. Desde luego, este sería lo típico en la situación *estándar* deseable en el ámbito de la evaluación del rendimiento educativo. Además, parecería descabellado esperar que un profesor, con el objetivo de asignar calificaciones a estudiantes en una materia, vaya a preparar *a propósito* un examen que pueda ser respondido correctamente al azar por la mitad (k=2), o la tercera parte (k=3), o cualquier otro porcentaje del alumnado. Por otra parte, no está nada claro cómo podría asegurarse el profesor de que, para quien responda al azar (solo para ellos, naturalmente), son equiprobables todas las posibles opciones alternativas de respuesta propuestas para todos y cada uno de los ítems que conforman un examen, ni que, en la realidad, lo trate de hacer… Las diferencias entre la caracterización típica de una situación real de evaluación educativa y las condiciones asociadas a la puntuación mediante fórmula parece que no señalarían precisamente la necesidad de su utilización.

En otro orden de cosas, Frary (1982, 1988), señala que cuando los sujetos no siguen la instrucción de no tratar de adivinar la respuesta en caso de ignorancia absoluta de la misma, la calidad psicométrica de las puntuaciones obtenidas a través del uso de la fórmula tiende a ser muy similar a la de las puntuaciones en términos del número de aciertos. Desde luego, parece razonable *a priori* que, en una situación de examen, donde sienten que está en juego su trabajo durante el curso y la posibilidad de pasar la asignatura, algún estudiante trate de conjeturar o incluso adivinar abiertamente las respuestas a tantos de los ítems cuya respuesta ignora como se vea capaz. Desde esta perspectiva, pues, tampoco se aprecia la necesidad de apelar a la fórmula de puntuación.

Por si todo lo anterior pareciera poco, no puede ignorarse el hecho de que la mayoría de los tests empleados para valorar el rendimiento en ámbitos educativos son utilizados como tests *normativos*. Esto significa que las puntuaciones no pueden ser interpretadas como medidas absolutas de conocimientos o logros, sino que tan solo posibilitan establecer un cierto ordenamiento entre los sujetos, un posicionamiento de cada uno en relación con un grupo o población de referencia. En esta situación, no existe absolutamente ninguna diferencia entre la interpretación (básicamente *ordinal*) derivada de puntuaciones obtenidas mediante la fórmula y la derivada del número de aciertos. Es más, en general ambos tipos de puntuaciones proporcionan la misma información, en la mayor parte de los casos. Una vez más, el principio de simplicidad llevaría a descartar la necesidad del uso de la fórmula de puntuación en los contextos de la evaluación del rendimiento educativo por medio de tests formados por ítems de elección múltiple.

Autores como Frary (1988, Frary et al., 1985) y Rowley & Traub (1977), concluyen que la puntuación mediante fórmula generalmente no solo no es necesaria, sino que es inapropiada para ser aplicada en el uso de tests empleados en la clase. Frary, únicamente admite su utilización con tests que requieran elevada velocidad de ejecución por parte de los sujetos, o que estén formados por ítems muy difíciles para la mayor parte de la población a que están destinados, por ejemplo, en selección laboral o admisión académica[2], si bien sostiene, sin embargo, que, por ejemplo, en los exámenes para expedir licencias profesionales debería ser tenido en cuenta únicamente el número de aciertos (Frary, 1988, p. 79). Por cierto, en un trabajo citado anteriormente, Frary y sus colaboradores llegan a plantear que el apoyo al uso de la fórmula de puntuación podría estar basado en motivos que nada tienen que ver con consideraciones psicométricas. En términos muy duros, hablan de errores de concepto, tanto entre los especialistas como entre el público general, y dicen querer descartar, por inmoral, la mera posibilidad de que promoviendo su utilización se pueda tan solo pretender dar una imagen favorable al uso de los tests (Frary et al., 1985, pp. 7-8).

En todo caso, para concluir, quizás podría considerarse la puntuación mediante fórmula como un heurístico, un procedimiento tentativo, auxiliar, que puede proporcionar información útil en un contexto de evaluación mucho más amplio, en algunos casos, pero nunca como un algoritmo, una regla fija a cuya suerte fiar toda conclusión. Sin olvidar que Gulliksen manifiesta explícitamente que "…los métodos [de puntuación de tests] que traten de determinar un nivel alcanzado [en lo que se pretenda valorar], tales como los utilizados en el test de Binet, requieren un tipo diferente de enfoque teórico" (Gulliksen, 1950, p. 245); probablemente, me permito añadir, un enfoque del tipo de la medición referida a un criterio.

El beneficio de la duda o "a río revuelto…"

Antes de tratar de buscar alguna alternativa práctica, en la línea sugerida por Gulliksen, valga una breve reflexión sobre un asunto que se plantea adicionalmente en el día a día de la utilización de la fórmula de puntuación en la evaluación del rendimiento educativo. El hecho de que en ella no aparezca ningún parámetro relativo a las *omisiones*, o preguntas para las que el sujeto no elige ninguna de las opciones de respuesta

[2] Incidentalmente, el empleo de tests con estos propósitos –incluido el de la valoración del logro educativo- ha sido objeto de una gran polémica a lo largo del tiempo, y lo sigue siendo hoy en día. Por ejemplo, en los Estados Unidos de América, donde se originan y desde donde se extienden prácticamente al resto del planeta estos asuntos, se ha desencadenado recientemente un escándalo a cuenta de una sentencia del Tribunal Federal en New Haven en torno al uso, sesgado por motivos étnicos, de tests para la selección y la promoción de bomberos. En un artículo publicado el la edición *on line* del New York Times del día 11 de julio de 2009, Lani Guinier, profesora de derecho de Harvard, y Susan Sturm, profesora de derecho de Columbia, van más allá de los meros asuntos de sesgo funcional, y ponen en entredicho la "fe ciega de los estadounidenses en los tests", basada en profundos errores de concepto en relación con la capacidad de los mismos como instrumentos para apreciar las potencialidades de una persona para desempeñar un cometido.
http://www.nytimes.com/2009/07/11/opinion/11guinier.html?_r=1&th&emc=th

alternativas disponibles, implica que estas no tienen ninguna consecuencia *directa* sobre la puntuación atribuida finalmente al sujeto en el test. Cada acierto suma un punto, cada error resta una fracción de punto, cada omisión ni suma ni resta nada. Pero cuando se utiliza la fórmula en el seno de la evaluación del rendimiento educativo, al trasladar la puntuación empírica de un sujeto en el test a la acostumbrada escala entre 0 y 10, ¿no habría que tener de alguna manera en cuenta esas omisiones?, ¿no es una pregunta no respondida una *declaración* de que se ignora con certeza la respuesta correcta (es decir, una "confesión de ignorancia"? Si es así, ¿se puede tratar con neutralidad, sin que comporte algún tipo de "penalización" (cuando sí se penalizan los errores manifiestos)? Parece que el sentido común desaconsejaría tal forma de proceder: la calificación transformada a la escala decimal debería calcularse siempre teniendo como referencia el total de respuestas posibles, es decir, la máxima calificación posible en la escala empírica, nunca exclusivamente el número total de respuestas dadas por un sujeto en particular.

Establecer unos mínimos generales, comunes para todos los sujetos de una población determinada, parece lo más sensato cuando se trata de evaluar conocimientos en el ámbito educativo. Así, la calificación que deba corresponder al nivel mínimo requerido para el aprobado, al igual que las calificaciones correspondientes a los niveles de notable y sobresaliente (y matrícula de honor, si procede), deberían ser establecidas necesariamente teniendo en cuenta el número total de preguntas/ítems que constituyen el examen/test, puesto que solo de este modo es posible fijar un único nivel de exigencia igual para todos los examinados (uso del test relativo a un *criterio*).

Si un alumno puede no responder a aquellos ítems que considere oportuno sin que ello tenga ninguna consecuencia directa sobre la evaluación de su rendimiento, ¿dónde queda la validez de contenido del examen? Calcular el nivel mínimo para el aprobado únicamente sobre el número de ítems del examen que cada alumno haya acabado contestando (es decir, "ignorar las omisiones"), supone, de hecho, calificar los conocimientos de cada examinando no en términos de un único y mismo nivel de exigencia igual para todos (no en términos de un único y mismo examen), sino sobre la base de un criterio de exigencia particular, "personalizado", *ad hoc*, y diferente de unos sujetos a otros, estableciéndose un trato desigual y no equitativo en la evaluación de los conocimientos demostrados por cada alumno. Con las calificaciones así definidas, unos alumnos quedarían beneficiados y otros perjudicados en función de variables como la asunción de riesgos, la precaución, el conservadurismo, etc., espurias en relación con el conocimiento de los contenidos que pretendidamente se trata de demostrar mediante sus respuestas al examen.

Un apunte como colofón a este asunto. Cuando no existe ninguna omisión en el conjunto de las respuestas de todos los sujetos, las puntuaciones resultantes de la utilización de la fórmula son simplemente transformaciones lineales del número de aciertos; proporcionan, por tanto, la misma información. Por tanto, si las omisiones no están permitidas, si los sujetos no deben dejar sin responder un ítem a menos que no tengan ni la menor idea de cuál es la respuesta correcta, ¿dónde quedaría, pues, la necesidad o la ventaja del uso de la fórmula de puntuación?

Epílogo

Cuando era estudiante de Psicología, mis profesores me transmitieron la idea de que la Ciencia es una forma de conocimiento superior, que se distingue fundamentalmente de otras no solo por el rigor lógico de su proceder, sino, sobre todo, porque para ella la realidad no es solo la fuente donde encuentra los asuntos por los que se interesa, sino quien proporciona también el criterio último de valoración; es la realidad quien avala o rechaza las soluciones que la Ciencia propone. Oppenheimer dice que todas las ciencias surgen como refinamientos, correcciones y adaptaciones del sentido común: partiendo del sentido común, tras duros y largos trabajos, se devuelven al sentido común nociones refinadas, originales, "extrañas" (más allá de lo accesible directamente en la vida cotidiana), que enriquecen el conocimiento de la humanidad y el modo como vive (Oppenheimer, 1956, pp. 128-129).

Hoy en día se hace, por todas partes, un uso abusivo de términos como Ciencia o científico, bajo la pretensión implícita –e inconfesada- de que convierten en casi infalible, "sagrado", aquello a lo que acompañan (a menudo, por cierto, sin más garantías de "cientificidad" que la palabra de quien lo dice, o la mera costumbre). En particular, me parece bastante atrevido que, en relación con el método habitual de puntuación mediante fórmula de los exámenes de tipo test formado por ítems de elección múltiple, empleados muy frecuentemente para la evaluación de conocimientos en el ámbito universitario, se hable de "el método/procedimiento científico de puntuación [de tests]", como equivalente a la única posibilidad admisible en rigor y garante de (una presunta) objetividad, una vez más, en un sentido casi absoluto, "sagrado" (y, por tanto, ilusoria[3]).

En general, parecería que en estos tiempos, frente a la reflexión y la responsabilidad informada del científico o el técnico, se encuentra por todas partes una imparable tendencia a buscar una suerte de "seguro universal" en el uso "automático", prácticamente en todas las circunstancias, de estrategias y procedimientos que en su origen (es decir, bajo las condiciones apropiadas), y por naturaleza, son rigurosos y adecuados. Se apela a menudo a modelos, métodos y fórmulas específicos, de naturaleza cuantitativa, como si fueran universales, absolutos, como a un *mantra*. Se extiende

3 En una conferencia pronunciada en el congreso de la American Psychological Association (APA), en 1955, Oppenheimer defiende la búsqueda de la objetividad como uno de los rasgos característicos de la Ciencia, pero "…no en un sentido metafísico, sino en un sentido práctico, como la búsqueda de una seguridad de que nos entendemos unos a otros, y de que todos los practicantes cualificados hablan de lo mismo y quieren decir esencialmente la misma cosa" (Oppenheimer, 1956, p.128).

y consolida un profundo sentido de (pseudo) objetividad, que está basado en tres premisas: (a) la información que se puede expresar en números es todo lo que se necesita para alcanzar el conocimiento preciso, (b) los datos numéricos pueden sustituir a cualquier otra forma de indagación, (c) la perspicacia numérica puede sustituir al conocimiento práctico sobre la realidad (aquello de que la ciencia es cuantitativa y lo demás "coleccionismo de sellos"). Asociarle un número crea la ilusión de que una información es más sólida: se trata de una pseudo-objetividad muy extendida, que corresponde a una suerte de *alienación* de la realidad, un intento de tomar un conocimiento abstracto y supuestamente cuantitativo como sustituto del conocimiento concreto y con significado (cualitativo). Un gran número de los trabajos actuales en Psicología (y más aún en Economía –e incluso en otras Ciencias Sociales-) se elaboran tomando en consideración criterios supuestamente objetivos, así considerados única y exclusivamente por el hecho de que pueden ser expresados y manejados en forma matemática por especialistas en ello (aunque carezcan de conocimientos o experiencia en relación con los contenidos abordados -dice Niall Ferguson, un profesor de Historia de la Economía y las Finanzas en la Escuela de Negocios y la Universidad de Harvard, que "a aquellos a quienes los dioses quieren destruir, antes les enseñan matemáticas"-). Es verdad que, en consonancia con este estado de cosas, se observa asimismo una creciente valoración crítica del uso meramente "mecánico", a-teórico, y casi indiscriminado de modelos matemáticos cuantitativos en diversas áreas de aplicación del conocimiento, más allá de lo que la razón demandaría. A este respecto, y ateniéndonos como ejemplo exclusivamente a casos que han tenido repercusión pública, en los últimos meses han aparecido en foros muy diversos un gran número de reflexiones en torno a la incidencia del uso a toda costa de modelos cuantitativos en Wall Street sobre la crisis financiera y económica mundial (como muestra, pueden verse Innerarity, 2009 o Salmon, 2009). En otro contexto, también con notorias connotaciones para el público general, resulta muy ilustrativa una publicación reimpresa en el pasado mes de junio, en la que dos geólogos estadounidenses señalan verdaderas aberraciones cometidas en el estudio de temas medioambientales debido al empleo de determinados modelos matemáticos canónicos, de hecho más dirigidos al ajuste numérico que a la representación apropiada de la realidad (Pilkey & Pilkey-Jarvis, 2009). Y es que, en todo caso y recordando otra vez a Oppenheimer (1956), nunca se debería pasar por alto que, si se albergan pretensiones de "cientificidad" al respecto de un asunto determinado cualquiera de la realidad, nunca ha de ser el instrumento, formal o incluso físico, sino la realidad que se aborda, con su propia lógica, el origen de cualquier pregunta que se formula, el garante de los métodos empleados para tratar de encontrar alguna respuesta, y el referente (criterio) último de las posibles respuestas a las preguntas que ella misma ha planteado.

Volviendo al asunto concreto que es objeto del presente trabajo, como ya se planteó al principio del mismo, los métodos de puntuación que se desarrollan alrededor del procedimiento de contar el número de respuestas correctas de un sujeto a los ítems de un test tienen su origen en la Teoría Clásica de los Tests. A la propia Psicología nunca le ha parecido que, bajo cualquier circunstancia, sea posible -menos aún necesario- extrapolar *literalmente* las aportaciones establecidas en un marco teórico a otro ámbito cualquiera, sin más ni más. En suma, el marco de referencia establecido para evaluar el nivel de conocimientos adquiridos por un sujeto en un proceso de enseñanza/aprendizaje debería ser otro distinto del que fundamenta el uso de la fórmula de puntuación aquí considerada, como, entre otros muchos, señaló ya Gulliksen hace sesenta años. Seguramente en la Teoría de la Respuesta al Ítem, paradigma actual en la Psicometría, se puede encontrar ese marco mucho más apropiado. Con sus procedimientos de construcción de bancos de ítems, convenientemente calibrados, podría permitir incluso la elaboración de tests verdaderamente personalizados, en el sentido de adaptados a las características de cada sujeto, teniendo en cuenta las particularidades de cada situación y el uso al que estén destinados los resultados del proceso de aplicación del test. Y, tanto para quienes adoptasen este marco de referencia, como para los que pudieran sentirse abrumados por las dificultades y los tecnicismos propios de la Teoría de la Respuesta al Ítem, y decidieran mantenerse en una perspectiva más tradicional, a la hora de utilizar un test como instrumento para evaluar conocimientos o logros en el ámbito académico, parece fundamentado y razonable proponer, en todo caso, que se establezca, a partir de bases teóricas y empíricas, un "nivel mínimo" requerido como indicador de suficiente dominio de la materia evaluada, así como, eventualmente, otros umbrales que delimiten gradaciones posibles (uso del test *referido a un criterio*). Entonces, de acuerdo con Gulliksen una vez más, lo más sencillo sería "…establecer las condiciones que aseguren que prácticamente todos los sujetos pueden intentar responder a prácticamente todos los ítems, y…asignar a cada sujeto una puntuación igual al número de aciertos (Gulliksen, 1950, p.250)", y, en consonancia, delimitar un *número mínimo de aciertos* necesario para obtener las distintas calificaciones académicas que se contemplen (uso del test relativo a un criterio). Adicionalmente, para evitar otro tipo de problemas (entre ellos, que alguien pudiera elegir todas las respuestas posibles a todos los ítems, lo que llevaría a cumplir la condición, pero…¡de qué manera!), habría que complementar este requerimiento de un *número mínimo de aciertos* con el requisito añadido de que el número de errores cometidos no supere un determinado porcentaje del total de respuestas posibles.

Para concluir, y particularmente a beneficio de quienes, pese a todo, sigan recelosos del papel que pueda desempeñar el azar, téngase en cuenta, como orientación, que, bajo las condiciones idóneas de construcción y aplicación, la probabilidad de obtener por azar una calificación máxima (equivalente a 10) en un test de 30 elementos independientes con tres alternativas de respuesta para cada uno es inferior a .00004 (¡cuatro cienmilésimas!); incluso la probabilidad de

obtener una calificación igual a 5 es inferior a .0003 (¡tres diezmilésimas!). Y si el test contiene 50 elementos, esas probabilidades se reducen a .000008 y .000064, respectivamente. ¿Se puede considerar que es realmente demasiado riesgo?

Referencias

Frary, R. B. (1982). A simulation study of reliability and validity of multiple-choice tests under six response-scoring models. *Journal of Educational Statistics, 7*, 333-351.

Frary, R. B. (1988). Formula scoring of multiple-choice test (Correction for guessing). *Educational Measurement: Issues and Practices, 7*(2), 33-38.

Frary, R. D., Cross, L. H., & Sewell, E.H. (1985). Partial information and the "correction" for guessing. *Annual Meeting of the National Council of Measurement in Education*, Chicago.

Gulliksen, H. (1950). *Theory of mental tests*. New York:Wiley.

Innerarity, D.(2009). La política y los riesgos del futuro. *elpais.com*, jueves 27/08/2009. http://www.elpais.com... Descargado el 27/08/2009.

Oppenheimer, R.(1956). Analogy in science. *American Psychologist, 11*, 127-135.

Pilkey, O.H. & Pilkey-Jarvis, L.(2009). *Useless Arithmetic*. New York: Columbia University Press.

Rivière, M. (2009). De los 'milagros' a la mala digestión. *elpais.com*, martes 14/07/2009. http://www.elpais.com... Descargado el 14/07/2009.

Rowley, G.L. & Traub, R.E. (1977): Formula scoring, number right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*(1), 15-22.

Salmon, F. (2009). Recipe for Disaster: The Formula That Killed Wall Street. Wired *Magazine*, lunes 23 de febrero de 2009. http://www.wired.com/print/techbiz/it/magazine/17-03/wp_quant Descargado el 03/04/2009.

(Artículo recibido: 31-8-2009; aceptado: 15-7-2010)