Escritos de Psicología
2000, 4: 55-63

CUANTO MÁS SE SABE, ¿SE SABE MÁS CUÁNTO SE SABE?

WHEN YOU KNOW MORE, DO YOU KNOW MORE ABOUT HOW MUCH YOU KNOW?

Rosa Bersabé Morán[1]
y Rosario Martínez Arias[2]

Resumen. Se investiga la relación entre el conocimiento que los sujetos muestran sobre un tema, y la precisión de sus juicios probabilísticos. Después de contestar un examen real, se pidió a los alumnos que estimaran la probabilidad subjetiva de que cada uno de los 20 ítems, tipo verdadero-falso, fuera verdadero. El conocimiento de los alumnos se midió a través de las calificaciones obtenidas en el examen (aciertos menos errores). Los resultados indican que a mayor conocimiento, mejor precisión total en los juicios probabilísticos, esto es, menor Puntuación de Brier. La precisión total se puede desglosar en tres componentes: discriminación, calibración y ruido. La discriminación también mejoró a medida que aumentaba el conocimiento. Esto puede deberse al hecho de que la capacidad para diferenciar ítems verdaderos y falsos afecta a ambas variables. Asimismo, cuanto más se conoce una materia, mejor se calibra. Dicho de otra forma, cuanto más se sabe, más se sabe cuánto se sabe. Los datos encontrados apoyan el modelo matemático desarrollado por Björkman (1992) que predice que tanto la resolución como la calibración mejoran con el conocimiento. Se formaron cuatro grupos de conocimiento para poder dibujar gráficos de calibración y de covariación.
Palabras clave: juicios de probabilidad; calibración; riesgo; diferencias individuales.

Abstract. Investigated the relationship between knowledge and several measures of probability judgment accuracy. After answering a 20 true-false ítem exam, students were asked to estimate the subjective probability that each ítem was true. Knowledge was measured by the grades in the exam (right minus wrong answers). The overall accuracy measure improved (Brier Score decreased) with knowledge. This overall accuracy measure can be decomposed in various dimensions: discrimination, calibration and noisiness. Discrimination measures increased with knowledge. This could be explained by the fact that the ability to differentiate true and false ítems works on both variables. At the same time, the more you know a matter, the better calibration or, in other words, the better you know how much you know. These findings support the mathematical model developed by Björkman (1992). Four groups of knowledge were formed to draw calibration and covariance graphs.
Key words: probability judgments; calibration; risk; individual differences.

[1] Dpto. de Psicología Básica, Psicobiología y Metodología de las CC. del Comportamiento. Facultad de Psicología. Universidad de Málaga. 29071-Málaga. España. FAX #: 95-213 26 21. Correo electrónico: bersabe@uma.es
[2] Dpto. de Metodología de las CC. del Comportamiento. Facultad de Psicología. Universidad Complutense de Madrid. 28223-Madrid. España.

En este trabajo, se ha retomado la pregunta que plantearon Lichtenstein y Fischhoff (1977) en su ya clásico estudio sobre la calibración de probabilidades (Do those who know more also know more about how much they know?). La calibración se refiere a la medida en que las probabilidades subjetivas que estima un sujeto sobre un suceso coinciden con la proporción de veces que realmente ocurre. Imaginemos, por ejemplo, 10 exámenes en los que un alumno pronosticó una probabilidad subjetiva de aprobar del 80%. Si los juicios de ese alumno estuviesen perfectamente calibrados, aprobaría realmente 8 de esos 10 exámenes.

La cuestión, entonces, es la siguiente: ¿los que conocen más una materia calibran mejor? Lichtenstein y Fischhoff (1977), a través de cinco investigaciones, hallaron que la relación entre el conocimiento de los sujetos y la calibración no era lineal. Hasta llegar a un punto (aproximadamente el 80% de respuestas correctas), cuanto más se sabe, mejor se calibra. Sin embargo, pasado ese punto de conocimiento, la calibración empeora. Parece ser que, cuando alguien es lego en una materia, tiende a sobrevalorar lo que sabe. Esa sobreconfianza va disminuyendo a medida que se va conociendo el terreno, con lo cual, va mejorando la calibración. Esto es así hasta llegar a un punto de conocimiento en el que se comienza a infravalorar lo que se sabe, por lo que la calibración comienza a deteriorarse. Es ya casi un tópico el que, después de realizar un examen, los alumnos más brillantes piensen que van a suspender, y los menos aventajados crean poder aprobarlo. Por defecto o por exceso, los dos extremos del saber se alejan de la calibración perfecta.

Investigaciones posteriores hablan de una relación lineal entre el conocimiento y la calibración. Así, Wright, Rowe, Bolger y Gammack (1994) encontraron que los sujetos que se autoevaluaban con un mayor conocimiento del billar (snooker), en una escala de 7 puntos, no sólo acertaban más en sus pronósticos sobre quién ganaría, sino que también mostraron una menor sobreconfianza, y una mejor calibración. Estos datos no se contradicen del todo con los de Lichtenstein y Fischhoff. Ellos también hallaron una relación lineal entre el conocimiento y la calibración ($r = -0,48$; $p<0,001$). No obstante, la función cuadrática se ajustaba significativamente mejor ($R = 0,62$; $p<0,001$).

El conocimiento también puede adquirirse a través de la experiencia o de la práctica. Entonces, ¿cuanto mayor es la experiencia, mejor se calibra? En general, parece que sí, aunque no en todos los ámbitos ocurre lo mismo (se pueden consultar las revisiones de Lichtenstein, Fischhoff y Phillips, 1982; Chan, 1982; O'Connor, 1989; Keren, 1991; Bolger y Wright, 1994). Dentro del contexto del diagnóstico clínico, Oskamp (1962) formó tres grupos experimentales distintos: un grupo de estudiantes de Psicología y dos grupos de psicólogos clínicos con diferente grado de experiencia. Los sujetos debían clasificar 200 perfiles del MMPI como correspondientes a pacientes con o sin trastornos mentales. Los dos grupos de psicólogos presentaron sobreconfianza, aunque ésta fue algo menor que la del grupo de estudiantes. En la misma línea, Lichtenstein y Fischhoff (1980) no sólo encontraron una mejor calibración de expertos frente a novatos, sino que parecía que ésta estaba sobrecompensada con infraconfianza. También en ajedrez, Horgan (1992) observó que cuanto mejor era el ranking de los niños, mejor calibraban la probabilidad de ganar a diferentes rivales hipotéticos ($r=-0,458$; $p<0,01$). Sin embargo, Wagenaar y Keren (1985) no encontraron diferencias en la calibración de las probabilidades que estimaban tres grupos de sujetos: jugadores profesionales de blackjack, expertos en estadística, y un grupo control (miembros de una orquesta). A todos se les presentó un cuestionario con 17 preguntas en las que debían estimar la probabilidad de obtener diferentes combinaciones de naipes. Por ejemplo, una de las preguntas era: de 1000 jugadas, ¿en cuántas aparecerían tres sietes seguidos? Hay que hacer constar que los expertos en estadística no podían hacer cuentas. Si no, es de suponer que hubieran calibrado perfectamente. En cuanto a los profesionales de blackjack, no salieron mejor parados que el grupo control porque la práctica del blackjack no ayuda a realizar estimaciones de probabilidad tan complejas. Es posible que los profesionales del blackjack habrían calibrado mejor que los otros dos grupos, si la tarea hubiera consistido en estimar la probabilidad de que la banca gane, vistas las cartas de otro jugador.

La mayoría de estas investigaciones se han formulado a un nivel preteórico. No obstante, a la par han ido surgiendo modelos matemáticos que dan cuenta de la relación entre el conocimiento y la calibración (Ferrell y McGoey, 1980; Albert y Sponsler, 1989; Björkman, 1992). Björkman (1992) ha propuesto un modelo lineal que relaciona la calibración con la resolución y el conocimiento. La resolución, al igual que la calibración, es una medida de precisión de los juicios probabilísticos. Se refiere a la habilidad para

bilísticos. Se refiere a la habilidad para discriminar cuándo va a ocurrir un suceso y cuándo no. Por tanto, parece sensato pensar que la resolución esté relacionada con el conocimiento. Por ejemplo, los que tengan más conocimientos de meteorología podrán discriminar mejor cuándo va a llover y cuándo no. De hecho, el modelo predice que tanto la calibración como la resolución mejoran con el conocimiento. Es más, el conocimiento estaría más relacionado con la resolución que con la calibración. Este supuesto del modelo contrasta con los hallazgos de Lichtenstein y Fischhoff (1977) quienes encontraron que el conocimiento sí que se relacionaba con la calibración, pero no con la resolución. Tal vez, estos resultados se puedan explicar por el tipo de tarea de estimación que emplearon: se pidió a los sujetos que, después de contestar cada pregunta de dos alternativas, estimaran la probabilidad de que la respuesta dada fuera o no correcta. Esas probabilidades subjetivas estaban comprendidas entre 0,5 y 1, puesto que, en principio, no era de esperar equivocarse más que por puro azar. Este tipo de tarea de estimación difiere de la tarea en la que se pide a los sujetos que estimen la probabilidad, comprendida entre 0 y 1, de que ocurra un suceso. Por ejemplo, cuando los meteorólogos estiman la probabilidad (0-1) de que llueva al día siguiente, su conocimiento debería relacionarse directamente con su capacidad para discriminar si va a llover o no, esto es, con la resolución de sus juicios.

Con el fin de poner a prueba esta idea y, teniendo en mente el modelo matemático de Björkman, predecimos (sin estimar, esta vez, la probabilidad de acertar) que ante una tarea de estimación donde las probabilidades estén comprendidas entre 0 y 1:

1) El *conocimiento* de los sujetos estará, asimismo, relacionado con la *calibración* de sus juicios probabilísticos.
2) El *conocimiento* de los sujetos sobre un tema estará estrechamente relacionado con la *resolución* de sus juicios probabilísticos.

MÉTODO

Participantes

La muestra estaba comprendida por 224 alumnos que cursaban la asignatura de Psicometría en la Facultad de Psicología de la Universidad Complutense de Madrid. De ellos, 48 eran hombres y 176 mujeres.

Material

Examen teórico de 20 ítems de dos alternativas (V-F). Hoja de instrucciones y cuestionario con los mismos 20 ítems del examen en los que se debía estimar la probabilidad de que ese ítem fuera verdadero y falso (información redundante). De los 20 ítems presentados, 11 eran verdaderos y 9 falsos.

Procedimiento

En primer lugar, los sujetos realizaron en situación natural el examen teórico tipo test con 20 ítems de dos alternativas (V-F). Rodeaban con un círculo la opción que consideraban correcta, pudiendo dejar preguntas sin contestar. Una vez terminado el examen, se pidió a los alumnos que leyeran la hoja de instrucciones en la que se explicaba cómo debían rellenar el cuestionario. Se les presentaban las mismas 20 preguntas del examen que acababan de hacer, y tenían que estimar la probabilidad de que ese ítem fuera verdadero y falso. Esto se hacía a través de porcentajes. Por ejemplo, una respuesta de "verdadero"=0% y "falso"=100% indicaría que se estaba completamente seguro de la falsedad de ese ítem. Por el contrario, una respuesta de V=50% y F=50% significaría que eran igualmente probables las dos alternativas. En la hoja de instrucciones se aclaró, además, que la prueba era voluntaria, que no repercutiría en ningún sentido a la hora de valorar el examen, y que el propósito era el de estudiar la calidad de las pruebas tipo test. Con esto, se pretendía eliminar de algún modo el posible efecto de la "deseabilidad social" que pudiera hacer que los alumnos se quisieran mostrar más seguros en sus juicios de probabilidad de lo que realmente estaban.

El método empleado en la mayoría de los trabajos sobre calibración difiere en algunos aspectos del que se ha seguido en este estudio. En general, se suele pedir a los participantes que contesten todas y cada una de las preguntas. Se trata, pues, de una tarea de elección forzada. En segundo lugar, se les pide que indiquen (a través de porcentajes) el grado en que se estima "que la respuesta dada en ese ítem es correcta". Por tanto, esas respuestas se encuentran entre el 50%-100% porque

no es de esperar errar más que el mero azar. En nuestro trabajo, la primera tarea de contestar los ítems es de elección no forzada, y la estimación que se pide es sobre que el "ítem sea verdadero", que no correcto, con lo cual los porcentajes abarcan el rango de 0-100. Se hace hincapié en estas distinciones en el procedimiento por la repercusión que tienen a la hora de interpretar las medidas de precisión en los juicios. Los porcentajes estimados por los sujetos se transformaron en once categorías de probabilidades subjetivas: [0-0,05) [0,05-0,15) [0,15-0,25) [0,25-0,35) [0,35-0,45) [0,45-0,55] (0,55-0,65] (0,65-0,75] (0,75-0,85] (0,85-0,95] (0,95-1].

El conocimiento de la asignatura se midió mediante la nota que cada alumno obtuvo en los 20 ítems del examen (aciertos menos errores). Para dibujar los gráficos de calibración y covariación (figuras 1 y 2), se tuvieron que formar cuatro grupos de conocimiento por los cuartiles de las notas. El conocimiento que demostró cada grupo se refleja en los datos de la tabla 1.

Tabla 1
Medias y desviaciones típicas (D.T.) en los veinte items del examen. Cada grupo está formado por 56 alumnos.

	CONOCIMIENTO DE LA MATERIA			
	Bajo Media (D.T.)	Medio-bajo Media (D.T.)	Medio-alto Media (D.T.)	Alto Media (D.T.)
Aciertos	8.625 (1.244)	10.393 (1.216)	11.821 (0.993)	13.536 (1.537)
Errores	6.179 (1.377)	4.821 (1.441)	3.857 (1.167)	2.375 (1.214)
Nota	2.446 (1.220)	5.571 (0.806)	7.964 (0.738)	11.161 (1.604)

RESULTADOS

Se calcularon distintas medidas de precisión en los juicios probabilísticos para cada sujeto. Pasemos a comentar cada una de esas medidas, descritas en detalle por Yates (1990).

Precisión Total

Puntuación de Probabilidad. La medida de precisión en los juicios probabilísticos más ampliamente utilizada se atribuye a Brier (1950), y se conoce como la Puntuación de Brier, Puntuación Cuadrática, o Puntuación de Probabilidad media ($\overline{PP}$). Ésta se define como:

$$\overline{PP} = \frac{\sum (f_i - d_i)^2}{N} \qquad (1)$$

En esta investigación,

f_i es la probabilidad que estima el sujeto de que el ítem i-ésimo sea verdadero.

$d_i=1$ si el ítem i-ésimo es verdadero

$d_i=0$ si el ítem i-ésimo es falso

Por tanto, $\overline{PP}$ es una función cuadrática que mide el grado de diferencia entre cada uno de los juicios de un sujeto, y lo que ocurre realmente. $\overline{PP}$ se encuadra en el intervalo [0,1]. Cuanto más precisos son los juicios de una persona, menor será su $\overline{PP}$.

Se encontró que la $\overline{PP}$ iba disminuyendo a medida que se conocía más la materia (r= -0,806; p<0,001, significación bilateral). Esto indicaría que cuanto más se sabe, más precisos son los juicios probabilísticos sobre ese tema.

Componentes de la Precisión Total

La precisión en los juicios de probabilidad no es un concepto indiferenciado sino que se puede desglosar en distintos componentes: calibración; discriminación y ruido (Murphy, 1973; Yates, 1982).

Calibración. Se pueden distinguir tres índices de calibración. Analicemos cada uno de ellos:

Indice de Calibración por Categorías de juicios (ICC). Ya vimos cómo en la Puntuación de Probabilidad se comparaba cada juicio de probabilidad con lo que ocurría realmente en cada uno de los ítems (0: falso; 1: verdadero). En el ICC, en cambio, se toman todos los ítems donde se emitió el mismo juicio de probabilidad, y se compara ese juicio con la proporción de esos ítems que realmente eran verdaderos. Por ejemplo, supongamos que un alumno ha estimado una probabilidad de 0,7 en 10 ítems. Con una calibración perfecta, 7 de esos 10 ítems serían realmente verdaderos, y 3 falsos. El ICC se calcula de la siguiente manera:

$$ICC = \frac{\sum N_j (f_j - \bar{d}_j)^2}{N} \qquad (2)$$

donde,

j indica las diferentes categorías de probabilidades subjetivas (0, .1, .2, .3, .4, .5, .6, .7, .8, .9, y 1);

N_j es el número de juicios registrados en la categoría j de probabilidad subjetiva;

$\overline{d_j}$ es la proporción real de ítems verdaderos, tomando aquellos en los que se estimó la categoría j de probabilidad subjetiva.

Tal como se desprende de la fórmula, cuanto menor sea la puntuación en el ICC, mejor será la calibración. Los datos revelan que el ICC es mejor, cuanto mayor es el conocimiento de los alumnos sobre la materia que se pregunta (r= -0,473; p<0,001, significación bilateral). Por tanto, parece que cuanto más se sabe de algo, más se sabe cuánto se sabe o, lo que es lo mismo, mejor se calibra.

La mayor parte de los trabajos de calibración aportan gráficos de calibración. Para dibujarlos, se formaron cuatro grupos de conocimiento con 56 alumnos cada uno (ver tabla 1). En cada grupo, se reunieron los juicios de todos los sujetos, es decir, como si los alumnos se hubieran ido relevando para enjuiciar. Cada grupo (o "macrosujeto") obtuvo, entonces, una única puntuación en cada medida de precisión de los juicios. Esas puntuaciones son las que recoge la tabla 2.

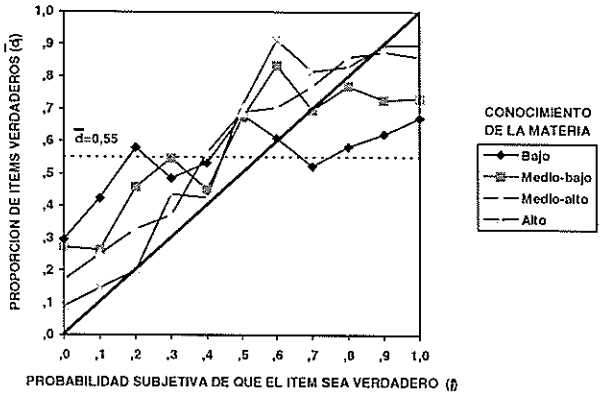
En los gráficos de calibración, el ICC se representa como las desviaciones de la función respecto a la diagonal 1:1. Si un sujeto calibra perfectamente en todas las categorías de sus juicios, su curva de calibración coincidirá con la diagonal 1:1. En la figura 1, se muestran las curvas de calibración de cada grupo de conocimiento. Así, por ejemplo, el punto más a la derecha de la curva de calibración del grupo de conocimiento "bajo" muestra que, de los ítems en los que se estimó un porcentaje del 100% de ser verdaderos, sólo el 66.94% resultaron serlo realmente. Como se puede apreciar en la figura 1, las desviaciones respecto a la diagonal 1:1 son mayores cuanto menor es el conocimiento. Podemos ver ahora lo que ya habíamos analizado a través del ICC: cuanto más se sabe, mejor se calibra por categorías de juicios.

Tabla 2
Medidas de precisión reuniendo todos los juicios de probabilidad en cada grupo de conocimiento. N indica el número de juicios de probabilidad emitidos por cada grupo ("macrosujeto").

Medida de precisión [a]	CONOCIMIENTO DE LA MATERIA			
	Bajo (N=1118)	Medio-bajo (N=1097)	Medio-alto (N=1109)	Alto (N=1107)
Total $\overline{PP}$	.288	.255	.202	.160
Calibración				
ICC ↓	.061	.046	.023	.021
Sesgo 0	-.114	-.119	-.105	-.088
ICL ↓	.013	.014	.011	.008
Discriminación				
RM ↑	.021	.038	.068	.108
Pend ↑	.143	.235	.335	.435
Ruido				
Disp ↓	.093	.096	.082	.073

[a] ↑: mejor a medida que aumenta; ↓ : mejor a medida que disminuye; 0 : cero es el mejor valor.

Figura 1
Curvas de calibración en cada gupo de conocimiento



Un segundo tipo de calibración es la calibración a la larga. Si la calibración fuera perfecta, entonces, la media de *todos* sus juicios sobre un suceso ($\overline{f}$) debería coincidir con la proporción de veces que ocurre realmente ($\overline{d}$). La calibración a la larga se operativiza mediante el *Sesgo*:

$$Sesgo = \overline{f} - \overline{d} \qquad (3)$$

o con su cuadrado, el *Indice de Calibración a la Larga* (ICL):

$$ICL = Sesgo^2 \qquad (4)$$

Cuanto mayor es el valor absoluto del Sesgo, peor es el ICL. Con un único sujeto, el Sesgo y el ICL proporcionan una información redundante. Sin embargo, con un grupo de sujetos, pueden revelar diferentes cuestiones. Por ejemplo, si la mitad de los sujetos de un grupo obtuvieran un Sesgo de +0,15 y la otra mitad de –0,15, el Sesgo medio resultaría nulo. En cambio, los

juicios de los sujetos individualmente aportarían un pobre ICL.

En los trabajos en los que se pide que los sujetos estimen la probabilidad (0,5-1) de que la respuesta dada sea correcta, el Sesgo es un buen indicador de la sobre/infraconfianza (Lichtenstein y Fischhoff, 1977). Un Sesgo positivo se da cuando se sobreconfía, es decir, cuando se estima haber acertado más de lo que realmente se acierta. Por el contrario, la infraconfianza se traduce en un Sesgo negativo. Sin embargo, en esta investigación, la tarea consiste en estimar la probabilidad de que el ítem, contestado o no, sea verdadero (no que la respuesta dada sea correcta). Así pues, un Sesgo negativo como el que aparece en los cuatro grupos de conocimiento (véase la tabla 2) no indica infraconfianza, sino una mayor seguridad en la falsedad que en la veracidad de los veinte ítems. Ese Sesgo negativo podría deberse a que hay más ítems verdaderos que falsos, con lo cual la tasa base ($\overline{d}$) es mayor de 0,5. También podría deberse a un sesgo de respuesta por el que se tuviera mayor confianza en las preguntas falsas que en las verdaderas. En cualquier caso, lo que sí parece claro es que el conocimiento en la materia no se relaciona significativamente con el Sesgo (r=-0,120; p=0,073, significación bilateral), aunque sí con el ICL (r=-0,198; p=0,003, significación bilateral). Tal vez, la relación con el ICL aparezca por elevar al cuadrado la medida del Sesgo. En cualquier caso, tampoco es una correlación muy elevada.

El Sesgo se representa en los gráficos de covariación (figura 2) mediante la intersección entre la línea vertical de la tasa base ($\overline{d}$) y la horizontal de la probabilidad subjetiva media ($\overline{f}$). Si esa intersección se sitúa en la misma diagonal 1:1, entonces el Sesgo será nulo, ya que la probabilidad subjetiva media coincidirá con la tasa base. Si la intersección queda por encima de la diagonal, el Sesgo será positivo. Si queda por debajo, como en los cuatro gráficos presentados en la figura 2, el Sesgo será negativo. Esto confirma lo que ya se había obtenido numéricamente.

Discriminación. La calibración compete a la habilidad para indicar apropiadamente las distintas probabilidades de que se dé un suceso. A diferencia de esto, la discriminación se refiere a la tendencia a decir algo diferente de alguna forma en las ocasiones en que ocurre un suceso que en las que no se da. La medida en la que una colección de juicios alcanza el ideal de discrimina-

Figura 2a
Gráfico de covariación para el grupo de conocimiento bajo.

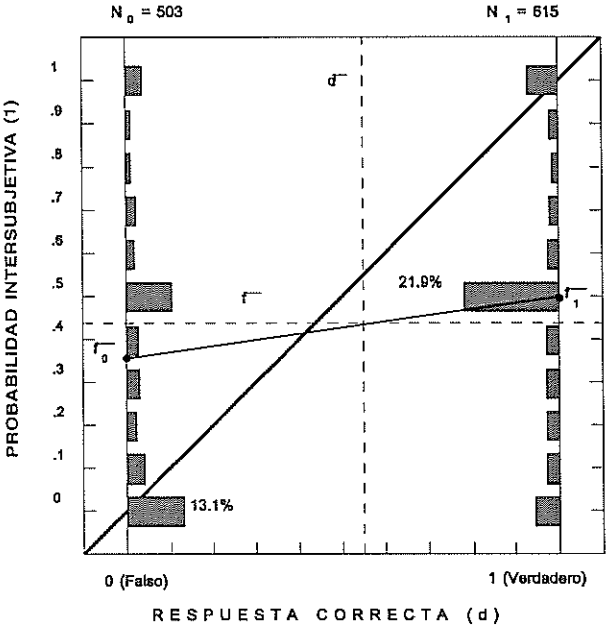


Figura 2b
Gráfico de covariación para el grupo de conocimiento medio-bajo.

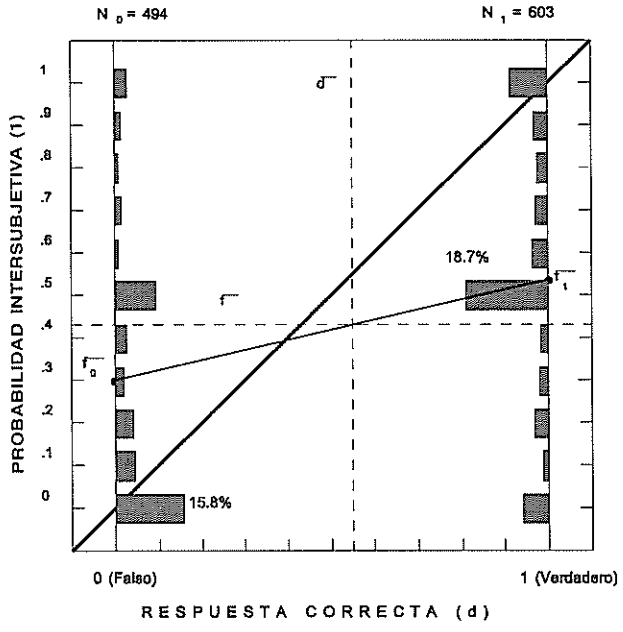


Figura 2c
Gráfico de covariación para el grupo de conocimiento medio-alto.

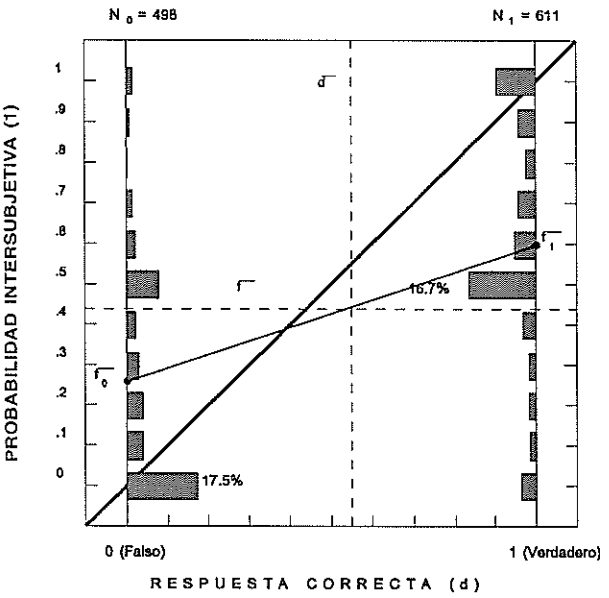
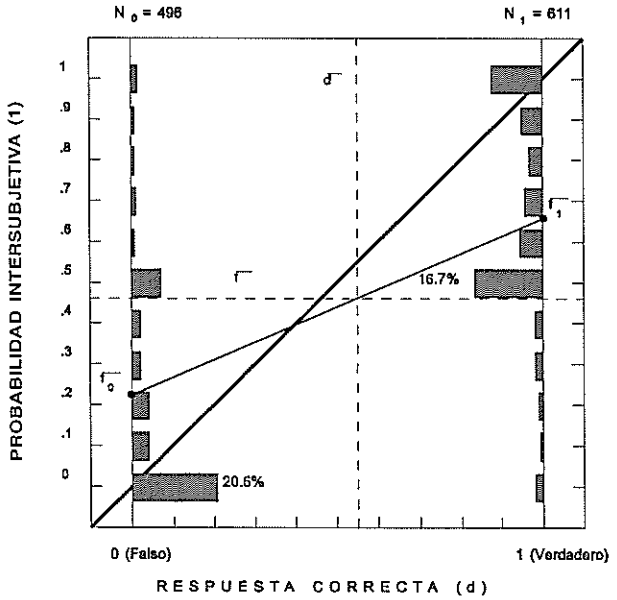


Figura 2d
Gráfico de covariación para el grupo de conocimiento alto.



ción se refleja en el estadístico de *Resolución de Murphy* (RM):

$$RM = \frac{\sum N_j(\bar{d}_j - \bar{d})^2}{N} \qquad (5)$$

El RM aumenta con el grado en que los juicios discriminan mejor (r=0,588; p<0,001, significación bilateral). En la tabla 2, se puede comprobar cómo el RM va mejorando con el conocimiento. Lógicamente, los que más saben, mejor discriminan entre los ítems verdaderos y falsos.

El RM también queda plasmado en los gráficos de calibración (figura 1). En ellos, la tasa base ($\bar{d}$) se dibuja en una línea horizontal. El RM será pésimo cuando se solape esa línea horizontal con la curva de calibración. Ese caso se produciría cuando el sujeto emitiera las diferentes probabilidades subjetivas aleatoriamente, sin poder discriminar cuándo el ítem es verdadero o falso. Entonces, sería de esperar que a cualquier probabilidad subjetiva le correspondiera a la larga una proporción real de ítems verdaderos igual a la tasa base. En cambio, si alguien fuera capaz de discriminar perfectamente los ítems, estimaría probabilidades mayores de 0,5 cuando el ítem es verdadero, y menores de 0,5 cuando es falso. Por tanto, la curva de calibración para todas las probabilidades subjetivas mayores de 0,5 tendría una altura de 1 (todos esos ítems serían verdaderos), y una altura de 0 para las menores de 0,5 (todos esos ítems serían falsos). Por ello, el RM se representa como las distancias entre cada punto (altura) de la curva de calibración y la altura de la tasa base. En la figura 1, se advierte que la curva de calibración del grupo con menor conocimiento es la que más se acerca a la tasa base, esto es, la que presenta una peor discriminación entre ítems verdaderos y falsos.

La Pendiente (Pend) es otra medida que pone en evidencia la capacidad para discriminar las veces en que se da un suceso de las que no ocurre. Si los juicios tienen una buena precisión en este sentido, entonces la media de las probabilidades subjetivas que se den cuando el ítem sea verdadero ($\bar{f}_1$) tenderán a ser mayores que cuando sea falso ($\bar{f}_0$). Por esto, la Pendiente se calcula como:

$$Pend = \bar{f}_1 - \bar{f}_0 \qquad (6)$$

Naturalmente, esta medida de discriminación entre ítems verdaderos y falsos también depende del conoci-

miento de la asignatura. Efectivamente, a mayor conocimiento, mayor Pendiente ($r=0,776$; $p<0,001$, significación bilateral).

En los gráficos de covariación, queda plasmada la Pendiente en la inclinación de la recta que une $\overline{f}_0$ con $\overline{f}_1$. La discriminación máxima se corresponderá con la diagonal 1:1 (Pendiente = 1 - 0 = 1). Como se puede observar en los gráficos de covariación (figura 2), la Pendiente es más inclinada cuanto mayor es el conocimiento.

Ruido

El último aspecto de la precisión de los juicios que se analizó fue el ruido que implica una forma especial de variación. Se calcula por las varianzas de los juicios condicionales a que el ítem sea verdadero ($Var(f_1)$) y falso ($Var(f_0)$). La medida de Dispersión (Disp) proporciona una información conjunta de dicha variabilidad que, no es más que una media ponderada de dichas varianzas condicionales:

$$Disp = \frac{N_1 Var(f_1) + N_0 Var(f_0)}{N_1 + N_0} \qquad (7)$$

donde, N_1 es el número de ítems verdaderos, y
N_0 el número de ítems falsos.

También la Dispersión de los juicios apareció relacionada con el conocimiento de los sujetos ($r= -0,238$; $p<0,001$, significación bilateral). La Dispersión se muestra en los gráficos de covariación. Estos gráficos no son más que dos diagramas de barras enfrentados: el de la izquierda para las probabilidades subjetivas cuando el ítem es falso; y el de la derecha para las probabilidades subjetivas cuando el ítem es verdadero. Existirá más Dispersión cuanto mayores sean las varianzas tanto en el gráfico de barras de la izquierda como en el de la derecha. En los gráficos de covariación, se puede apreciar que, en los dos grupos con menor conocimiento (figuras 2a, y 2b), la variación de los juicios es algo mayor que en los dos grupos con mayor conocimiento (figuras 2c, y 2d). Este resultado es lógico si se piensa que los que saben más centrarán más sus probabilidades subjetivas entre 0,5 y 1, cuando el ítem sea verdadero; y entre 0 y 0,5, cuando sea falso. En cambio, los que saben menos estimarán probabilidades entre 0 y 1 tanto cuando el ítem sea negativo como cuando sea falso, es decir, tendrán mayor variabilidad a ambos lados del gráfico de covariación.

DISCUSIÓN

Vistos los resultados, parece claro que la precisión total de los juicios probabilísticos es mejor, cuanto más se conoce la materia que se enjuicia. La precisión total viene determinada por tres componentes: discriminación, calibración y ruido. Todos ellos, se relacionaron en alguna medida con el conocimiento manifestado por los sujetos.

Tanto la discriminación como el conocimiento, en este estudio, hacían referencia a la capacidad de discernir si el ítem era verdadero o falso. Por este motivo, se esperaba que ambas variables correlacionaran positivamente. De hecho, eso fue precisamente lo que se halló. Esta relación resultó incluso más fuerte que la encontrada entre el conocimiento y la calibración, tal como predice el modelo de Björkman (1992).

En cuanto a la calibración, el índice de Calibración por Categorías era la medida que más nos interesaba en cuanto que podía dar respuesta al interrogante que planteábamos en el título del artículo: cuanto más se sabe, ¿se sabe más cuánto se sabe? Los resultados apoyan la respuesta afirmativa. Analizando los datos individuales, se obtuvo una relación lineal entre el conocimiento y la calibración ($r = -0,473$; $p<0,001$) muy similar a la que encontraron Lichtenstein y Fischhoff en 1977 ($r = -0,48$; $p<0,001$). No obstante, estos autores hallaron que la relación cuadrática ($R= -0,62$; $p<0,001$) se ajustaba significativamente mejor que la lineal, lo que no ocurre en nuestros datos ($R = -0,481$; $p<0,001$). Según ellos, hay un punto óptimo de calibración que se corresponde aproximadamente con un conocimiento del 80% de respuestas correctas. Con menor o mayor conocimiento, la calibración empeora. Nuestros datos, en cambio, apuntan a una relación lineal, como la encontrada por Horgan en 1992 ($r= -0,458$; $p<0,01$). Esto podría deberse a que, en nuestra muestra, sólo 3 de los 242 alumnos consiguieron una nota superior al 70%, es decir, que muy pocos alumnos llegaron a alcanzar el grado de conocimiento en el que se comienza a infravalorar lo que se sabe. Por tanto, parece ser que, a medida que se conoce más una materia (al menos hasta un cierto grado de conoci-

miento), se calibra mejor. O, dicho de otra forma, cuanto más se sabe de un tema, más se sabe cuánto se sabe. Esto se pudo visualizar en los gráficos de calibración en los que, en los grupos de mayor conocimiento, las probabilidades subjetivas se acercaban más a las reales.

Una vez encontrada esta relación entre el conocimiento y la calibración, deberíamos intentar explicar a qué es debida. Pitz (1974) sugirió que la evaluación de la incertidumbre depende de la variedad de respuestas diferentes que el sujeto puede activar. Si se activan muchas respuestas posibles, se reconocerán mejor los diferentes grados de incertidumbre. Cuanto más rica o accesible sea la base de conocimiento, mayor número de alternativas de respuesta se podrán recordar, y mejor será la calibración. El modelo matemático de Albert y Sponsler (1989) parece haberse inspirado en esta idea. Su modelo representa fenomenológicamente cómo el cerebro puede estimar probabilidades subjetivas basándose en recuerdos de experiencias pasadas similares.

En consecuencia, los datos obtenidos en nuestra investigación avalan el modelo matemático de Björkman (1992) que asume que el conocimiento se relaciona linealmente tanto con la resolución, como con la calibración. Estos resultados, a parte de sus implicaciones teóricas, pueden ser relevantes en la práctica profesional de quienes deben tomar decisiones o predecir sucesos en una situación de incertidumbre relativa: meteorólogos, médicos, psicólogos u otros profesionales. Cuanto mayor sea el conocimiento del tema que tengan estos profesionales, mayor fe podremos tener en la precisión de los juicios probabilísticos que emitan sobre sus predicciones o diagnósticos. Al menos, hasta cierto punto, ya que parece que al llegar a cierto nivel de conocimiento, la calibración se vuelve hacia atrás. Tal vez por eso, el humilde filósofo griego estimaba, con una evidente infraconfianza, aquello de que "sólo sé que no sé nada". ¿O era, acaso, un juicio perfectamente calibrado?

REFERENCIAS BIBLIOGRÁFICAS

Albert, J.M., y Sponsler, G.C. (1989). Subjective probability calibration: A mathematical model. *Journal of Mathematical Psychology,* 33(3), 298-308.

Björkman, M. (1992). Knowledge, Calibration, and Resolution: A linear Model. *Organizational Behavior and Human Decision Processes, 51,* 1-21.

Bolger, F. y Wright, G. (1994). Assessign the quality of expert judgment. *Decision Support Systems, 11,* 1-24.

Chan, S. (1982). Expert judgments under uncertainty: Some evidence and suggestions. *Social Science Quarterly, 63,* 428-444.

Ferrell, W.R. y McGoey, P.J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26,* 32-53.

Horgan, D.D. (1992). Children and chess expertise: The role of calibration. *Psychological Research, 54,* 44-50.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. Acta *Psychologica, 77,* 217-273.

Lichtenstein, S. y Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Performance, 20,* 159-183.

Lichtenstein, S. y Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149-171.

Lichtenstein, S., Fischhoff, B. y Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Martínez Arias, R. (1991). El proceso de toma de decisiones. En R. Martínez Arias y M. Yela (Eds.), *Pensamiento e Inteligencia* (pp. 411-494). Madrid: Alhambra Universidad.

Murphy, A.H. (1983). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595-600.

O'Connor, M.J. (1989). Models of human behavior and confidence in judgment: a review. *International Journal of Forecasting, 5,* 159-169.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76,* 28-547.

Pitz, G.F. (1974). Subjective probability distributions for imperfectly known quantities. En L.W. Gregg (Ed.), *Knowledge and cognition.* New York: Wiley.

Wagenaar, W. A., y Keren, G. B. (1985). *Organizational Behavior and Human Decision Processes, 36,* 406-416.

Wright, G., Rowe, G., Bolger, F. y Gammack, J. (1994). Coherence, Calibration, and Expertise in Judgmental Probability Forecasting. *Organizational Behavior and Human Decision Processes, 57,* 1-25.

Yates, J.F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132-156.

Yates, J.F. (1990). *Judgment and Decision Making.* Englewood Cliffs, New Jersey: Prentice Hall.