Rev. de Psicol. Gral. y Aplic., 1993, *46* (1), 75-87

CALIBRACIÓN EN JUICIOS CLÍNICOS

H. GAMBARA; O. G. LEÓN
Universidad Autónoma de Madrid

Resumen

En este artículo se presentan dos experimentos desarrollados dentro del paradigma de calibración. En ambos se utiliza una tarea de diagnóstico clínico.

El principal objetivo del primer experimento es analizar las incidencia de la explicitación de las consecuencias de las decisiones sobre la confianza en las mismas. Además, se analiza el patrón general de respuesta de los sujetos ante perfiles de diferente dificultad.

En el segundo experimento se analiza si la evaluación de las dos posibles alternativas de elección incide favorablemente sobre el patrón de calibración.

Se encontró un efecto significativo, tanto de la consideración de las consecuencias como de las posibles alternativas sobre la calibración de los sujetos, hallándose, en términos generales, una tendencia a evitar la sobreconfianza.

Abstract

This paper includes two experiments within the calibration paradigm. Both experiments use a clinical diagnostic task.

To analyze the effect of the decision consequences over the confidence is the main objective of the first experiment. Moreover, the general response pattern to profiles of different difficulty is analyzed.

The second experiment is devoted to study the incidence of the evaluation of the possible alternatives over calibration.

A significative effect of the decision consequences and of the evaluation of the alternatives were found. In general terms, we found a tendency to avoid overconfidence.

Introducción

Desde la primera revisión publicada en *Psychological Bulletin* sobre toma de decisiones (Edwards, 1954) hasta la realizada recientemente por Payne, Bettman y Johnson (1992), esta línea de investigación ha ampliado, modificado y matizado de manera significativa sus objetivos. Consecuencia de ello es el auge de ciertos temas de investigación dentro de este ámbito; entre otros muchos se encuentra la investigación en calibración, sobre la que se centrará el presente artículo.

La investigación sobre calibración incluye un conjunto de trabajos de carácter esencialmente empírico que trata de responder a la pregunta: ¿sabe la gente cuánto sabe? Característicos de tales trabajos son conceptos como el de sesgo, sobreconfianza y probabilidad subjetiva.

Aunque los antecedentes de esta área se remontan a finales de los años cincuenta (por ejemplo, Adams, 1957), fue en la década de los setenta y ochenta donde cobra su mayor auge —coincidiendo con la expansión en la investigación sobre heurísticos y sesgos—. Así, comienzan a aparecer un gran número de trabajos en relación a la sobreconfianza de los sujetos (término que definimos líneas más abajo), ignorándose frecuentemente el fenómeno inverso. Dentro de los trabajos de calibración destacan los realizados por Fischhoff, Lichtenstein y Slovic, quienes, conjuntamente o en colaboración, publicaron alrededor de diez trabajos experimentales y dos revisiones teóricas entre 1977 y 1986; precisamente la última de ellas se publica en *Judgment under Uncertainty: Heuristics and Biases*, Kahneman, Slovic y Tversky (1982).

Sin entrar a profundizar en el concepto de probabilidad subjetiva, en términos generales se puede decir que, desde el punto de vista subjetivista, per-

sonalista o bayesiano las probabilidades son números entre 0 y 1, que expresan el grado de conocimientos de los sujetos sobre los sucesos (De Finetti, 1974); de esta manera, es posible asociar distintos valores numéricos de la probabilidad a distintos grados de incertidumbre.

La tarea habitualmente utilizada consta de un gran número de ítems. El sujeto debe contestar cada uno de ellos y seguidamente estimar la confianza con la que ha seleccionado la respuesta, mediante una probabilidad subjetiva.

Un sujeto se considera un buen calibrador si, a largo plazo, y para todas las respuestas a las que ha asignado una determinada probabilidad subjetiva, la proporción correcta es igual a dicha probabilidad. Las desviaciones respecto a la calibración perfecta pueden producirse en sentidos opuestos. Se habla de sobreconfianza cuando las estimaciones de probabilidad subjetiva de los sujetos superan la proporción de respuestas correctas; en caso contrario, se hablará de infraconfianza. Los datos empíricos indican la existencia de sobreconfianza cuando el sujeto se enfrenta ante una tarea de dificultad media-alta. Cuando la tarea resulta fácil, se suele encontrar una infraconfianza general (véase Lichtenstein, Fischhoff y Phillips, 1982, para una revisión).

En los últimos años los trabajos de calibración han recibido diversas críticas (por ejemplo, Wagenar y Keren, 1985; May, 1986a, b; Paese y Sniezek, 1991). De entre ellas, vamos a destacar aquí dos: la primera se refiere a la tarea utilizada, la segunda al gran cúmulo de datos empíricos existentes, sin apenas apoyo teórico.

Respecto a la tarea, resalta el volumen de trabajos que incluyen un gran número de ítems de conocimiento general y que recuerdan a juegos como el del «Trivial Pursuit». Es común hallar trabajos en los que se pide al sujeto que conteste a preguntas tales como ¿de qué país es capital Guagadugu? Apenas se utilizan tareas que involucren decisiones reales o propias del quehacer profesional, donde una mala calibración podría tener serias repercusiones, por ejemplo, en el caso del diagnóstico clínico. En áreas afines a la calibración se han encontrado marcadas diferencias entre ítems de conocimiento general y otros tipos de ítems (por ejemplo, en el sesgo retrospectivo —Christensen-Szalanski y Fobian, 1991).

A pesar del carácter descriptivo de la mayor parte de los trabajos de calibración (Lichtenstein et al., 1982) y del aislamiento de estos trabajos respecto al marco general del que parten, la Teoría de la Decisión Conductual, recientemente se están realizando esfuerzos para determinar qué variables podrían incidir significativamente sobre la calibración (May, 1986a, b; Paese y Sniezek, 1991; Koriat, Lichtenstein y Fischhoff, 1980).

Con estos antecedentes, nos planteamos dos objetivos generales: 1) analizar el patrón de calibración en una tarea que, desde nuestro punto de vista, tuviera mayor relevancia que las habitualmente utilizadas, y 2) determinar bajo qué condiciones dicho patrón sería susceptible de modificación. Con otras palabras, estábamos interesados en analizar qué variables, no estudiadas previamente, podrían mejorar la calibración de los sujetos.

PRIMER EXPERIMENTO

Se diseñó un primer experimento con el objeto de: 1) analizar la calibración de los sujetos en una tarea de diagnóstico y comparar los resultados con la calibración hallada en otros ámbitos, y 2) analizar las posibles diferencias en calibración cuando se manipulan las consecuencias de las decisiones tomadas.

En cuanto al primer objetivo, la calibración manifestada por un psicólogo cuando se pide que emita juicios discriminativos entre trastornos no tiene por qué diferir, en principio, de la calibración encontrada en otros ámbitos del diagnóstico (Christensen-Szalanski y Bushyhead, 1981), la cual parece ser similar a la encontrada en otro tipo de contextos. En términos generales podemos predecir una calibración deficiente de los psicólogos. Calibración que será función de la dificultad de la tarea: ante perfiles fáciles de diagnosticar, los sujetos manifestarán infraconfianza; ante perfiles difíciles, sobreconfianza.

El segundo objetivo se relaciona con las consecuencias de las decisiones. Las consecuencias inherentes a la toma de decisiones es uno de los elementos claves, tanto para el proceso como para el resultado de la decisión. La misma tarea, tanto en contenido como en estructura, puede provocar en el sujeto una respuesta diferente a la misma si las consecuencias derivadas de su decisión se modifican. Por ejemplo, cirujanos expertos pueden subestimar las probabilidades de éxito en operaciones de alto riesgo para evitar, en caso de fracaso, posibles pleitos. Aunque la meta siempre será el éxito de la operación, la percepción del mismo se podrá ver condicionada por las consecuencias negativas de un posible fracaso, lo que podría llevar a utilizar un tipo de estrategia que, centrándonos en nuestro tema, se podría traducir en una infraconfianza. Así, predecimos una calibración diferencial ante una situación que conlleve consecuencias negativas frente a otra donde no se especifique ninguna consecuencia.

Método

Los sujetos

Un total de 40 estudiantes de los últimos cursos de la Facultad de Psicología de la UAM con conocimientos en Psicodiagnóstico y Psicopatología.

La tarea

Los sujetos deben realizar el diagnóstico de un conjunto de perfiles y seguidamente estimar la confianza en su juicio mediante una probabilidad subjetiva.

Para establecer tal diagnóstico, los sujetos deben considerar un conjunto de ítems de información y no un único ítem como ocurre bajo el paradigma tradicional.

Los estímulos

Se elaboraron un conjunto de perfiles utilizando los criterios diagnósticos del DSM III.

Los perfiles «puros» corresponden a trastornos afectivos mayores y a trastornos de ansiedad. Los llamamos así porque en ellos sólo se encuentran síntomas propios de esos trastornos (siempre según el DSM III).

Para la elaboración de los perfiles «mixtos» se consideraron todos los criterios diagnósticos necesarios para determinar uno de los trastornos (por ejemplo, depresión) añadiendo algunos criterios pertenecientes al otro tipo de trastorno (por ejemplo, fobia social). De cualquier manera, sólo se podía establecer un diagnóstico según los criterios del DSM III (por ejemplo, depresión).

Se añadió información no relevante para el diagnóstico en todos los perfiles, la cual constituía siempre los primeros ítems del perfil. El motivo de incluir este tipo de información fue el de presentar los perfiles de la manera más realista posible.

A continuación presentamos un ejemplo de los dos tipos de estímulos (el conjunto de perfiles utilizados puede pedirse a los autores).

Perfil («puro»):

Economista.
Es el último de cuatro hermanos. Nació cuando el hermano más pequeño tenía doce años.
Tuvo una novia con la que dice haber fracasado; actualmente no tiene pareja.
Trabaja más de diez horas diarias.
Como excesivamente, habiendo ganado ocho kilos en algo más de dos semanas.
Estado de ánimo disfórico y sentimientos de culpa.
Agitación psicomotora.
Manifiesta indecisión, aun en las cuestiones más simples.

Perfil («mixto»):

Camarera en un restaurante de moda.
Padece reumatismo.
Divorciada.
Miedo persistente a hablar en público.
Desinterés por las actividades usuales.
Pérdida de peso.
Disminución de la motivación sexual.
Pensamientos repetitivos de muerte.
Enlentecimiento psicomotor.

Se construyeron dos grupos paralelos de 15 perfiles cada uno. Estos 30 perfiles estaban formados por: 10 perfiles correspondientes a trastornos afectivos mayores, concretamente a los episodios depresivos; 10 perfiles correspondientes a trastornos de ansiedad, y 10 perfiles mixtos, 5 correspondientes a un diagnóstico de depresión mayor (o trastorno afectivo) y 5 correspondientes a un diagnóstico de trastorno de ansiedad.

La variable dependiente fue el grado de adecuación de las estimaciones de probabilidad subjetiva; es decir, el grado de correspondencia entre las estimaciones de seguridad de los sujetos (probabilidades subjetivas) y la realidad de los sucesos. Esta variable se operativizó a través de las medidas habitualmente utilizadas bajo este paradigma:

Curvas de calibración. Estas curvas reflejan la relación existente entre el porcentaje de respuestas correctas y las estimaciones de probabilidad subjetiva. En el eje de abscisas se presentan las estimaciones de probabilidad subjetiva de los sujetos. En el eje de ordenadas, el porcentaje de respuestas correctas.

Cuando la curva coincide con la diagonal existe una calibración perfecta. Una curva por debajo de la diagonal expresa sobreconfianza; por encima, infraconfianza.

A pesar de la utilización generalizada de la curva de calibración, la interpretación de ésta no siempre es fácil porque ¿cómo de alejada debe encontrarse la curva de la diagonal para hablar de sobre o de infraconfianza? Para subsanar este problema proponemos establecer los límites confidenciales correspondientes a cada rango de probabilidad con respecto a la diagonal.

La idea es establecer, a partir de los intervalos confidenciales, un área para el número de respuestas concretas que se obtengan experimentalmente. Esta área delimitará la «buena o mala calibración» de los sujetos. Diremos, por tanto, que existe sobreconfianza, no si la curva está por debajo de la diagonal, sino si está por debajo de dicha área. Lo mismo ocurrirá con la infraconfianza; se producirá cuando la curva de calibración supere dicha zona. En definitiva, cuando se encuentren alejamientos de la diagonal no atribuibles al azar.

Concretamente, los intervalos se calcularán con un nivel de confianza del 95 por 100 en todos los experimentos.

Estimación de los límites superiores:

$$P + Z\,[(P^{*}Q)/N]^{1/2}$$

Estimación de los límites inferiores:

$$P - Z\,[(P^{*}Q)/N]^{1/2}$$

Los valores que puede tomar P son: 0,5, 0,6, 0,7, 0,8, 0,9 y 1,0. $Q = (1 - P)$.

Índice de calibración (I.ca): El cuadrado de la diferencia entre la probabilidad subjetiva (PS) y la proporción correcta (PC) dentro de cada categoría de respuesta, ponderada por el número de respuestas en cada categoría. La puntuación máxima es 1 y la mínima 0 (calibración perfecta).

Índice de confianza (I.co): La diferencia entre la probabilidad subjetiva (PS) y la proporción correcta (PC) dentro de cada categoría de respuesta, ponderada por el número de respuestas en cada categoría. Esta puntuación indica el sentido de la calibración. Una puntuación positiva (PS > PC) indica sobreconfianza; una puntuación negativa (PS < PC), infraconfianza.

Variables independientes: 1) la dificultad de la tarea, y 2) las consecuencias del diagnóstico efectuado. Ambas intrasujeto.

La primera variable presentaba tres niveles: perfiles de ansiedad, de depresión y «mixtos». Aunque los dos primeros tipos correspondían a perfiles definidos como «puros», se consideraron niveles distintos de la variable ya que, a priori, desconocíamos su nivel de dificultad.

La segunda variable estaba constituida por dos niveles, los cuales se operativizaron mediante las instrucciones que los sujetos recibían. En una condición no se mencionaba ninguna consecuencia (g. control). En la otra se implicaba al sujeto, haciendo referencia a las consecuencias que podían derivarse de su decisión (g. experimental).

Planteamos, por tanto, un diseño intrasujeto 3 × 2. Cada condición incluía el diagnóstico de 15 perfiles. Las condiciones y el orden de presentación de los estímulos se contrabalancearon.

Procedimiento

En una primera fase, los sujetos recibían un cuadernillo con 15 perfiles, una hoja de instrucciones (la cual determinaba el tipo de condición) y una hoja de respuesta. En la condición sin consecuencias, los sujetos debían indicar si cada uno de los perfiles correspondía a un diagnóstico de trastorno de ansiedad o afectivo. En caso de estar realizando la condición contraria, debían dirimir si cada uno de los supuestos pacientes debía continuar o no con un tratamiento en el que se administraban ansiolíticos o antidepresivos. Tratamientos que podían tener repercusiones serias en los sujetos si no se les administraba la medicación adecuada.

En segundo lugar, debían estimar para cada perfil una probabilidad entre 0,5 y 1, representativa de la confianza con la que decidían. Una probabilidad de 1 indicaba una seguridad absoluta sobre la respuesta, mientras que una probabilidad de 0,5 significaba que lo mismo que se había contestado de una manera se podía haber contestado de la otra. No se podía estimar una confianza por debajo de 0,5, ya que siempre se debía elegir la alternativa más probable. Pasada una hora, se volvía a presentar a los sujetos un cuadernillo de perfiles (paralelos a los de la primera fase), una hoja de instrucciones (pertenecientes a la condición no realizada previamente) y una hoja de respuesta. Se tardaba media hora aproximadamente en realizar cada una de las dos condiciones.

Resultados y discusión

Para analizar los datos efectuamos un ANOVA de medidas repetidas (3 × 2) para cada una de las cuatro VD consideradas. El supuesto de igualdad de varianzas-covarianzas fue contrastado mediante la prueba de esfericidad. Esta prueba dio un valor no significativo para todas las variables, indicando el cumplimiento de dicho supuesto.

También se procedió a la elaboración de las curvas de calibración.

Tipo de perfiles

Como se desprende de la tabla 1, en todas las variables hemos encontrado diferencias significativas entre los perfiles mixtos y puros. No las hemos encontrado, sin embargo, entre los perfiles de depresión y de ansiedad.

Tal significación en las diferencias en la *proporción de diagnósticos correctos* revela que la manipulación experimental realizada, para esta variable, fue adecuada. En los perfiles «mixtos» la proporción media de respuestas correctas fue significativamente inferior a dicha proporción en los ítems «puros»: perfiles de depresión y perfiles de ansiedad (tabla 2).

En términos del *índice de confianza*, se observa una tendencia a la sobreconfianza en la tarea más difícil y una tendencia a la infraestimación de la confianza en las tareas más fáciles.

Curvas de calibración. Las curvas de ansiedad y de depresión quedan por encima de la diagonal; la curva para perfiles mixtos queda por debajo de esta recta. Ahora bien, la curva para ansiedad queda fuera de los intervalos, lo que básicamente no ocurre en los perfiles de depresión y mixtos (figuras 1 y 2*a*, *b*, *c*).

Por tanto, parece que la calibración de los sujetos es función de la dificultad de los perfiles. Si la tarea resulta excesivamente fácil, los sujetos presentan

TABLA 1

Valores de F y sus correspondientes niveles críticos del ANOVA 3 × 2 para cada una de las variables

Medidas	Factor	F	N. crítico
PROPORCIÓN CORRECTA	Perfiles	$F_{2,78} = 40,80$	0,0000
	Condición	$F_{1,39} = 22,20$	0,0000
	Interacción	$F_{2,78} = 2,68$	0,0751
PROBABILIDAD SUBJETIVA	Perfiles	$F_{2,78} = 28,34$	0,0000
	Condición	$F_{1,39} = 0,65$	0,4234
	Interacción	$F_{2,78} = 3,33$	0,0410
ÍNDICE DE CALIBRACIÓN	Perfiles	$F_{2,78} = 14,88$	0,0000
	Condición	$F_{1,39} = 5,92$	0,0197
	Interacción	$F_{2,78} = 0,28$	0,7561
ÍNDICE DE CONFIANZA	Perfiles	$F_{2,78} = 22,43$	0,0000
	Condición	$F_{1,39} = 18,33$	0,0001
	Interacción	$F_{2,78} = 1,23$	0,2981

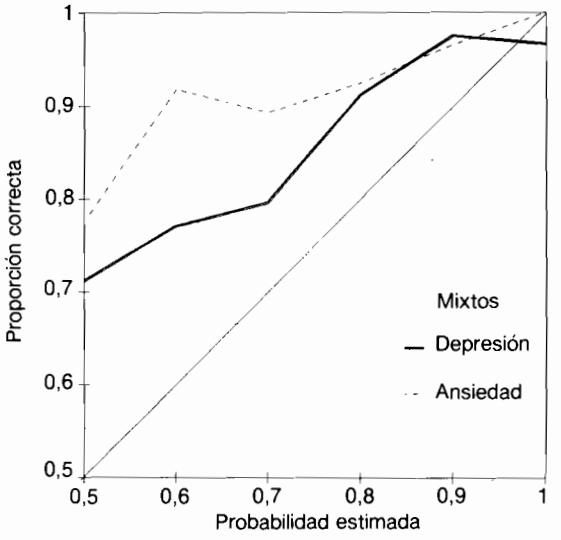

Figura 1. Curvas de calibración para los tres tipos de perfiles.

una infraestimación de su confianza. A medida que la tarea se va haciendo más difícil, se va produciendo una tendencia a la disminución de la infraconfianza.

Antes de continuar con la exposición de los resultados haremos un paréntesis para comentar un dato que nos llamó la atención. Nos referimos a la discrepancia existente entre la media del índice de calibración (vésase tabla 2) y las curvas de calibración. Por ejemplo, la curva de calibración del grupo control está más cerca de la diagonal que la curva del grupo experimental; sin embargo, la media del I.ca del gru-

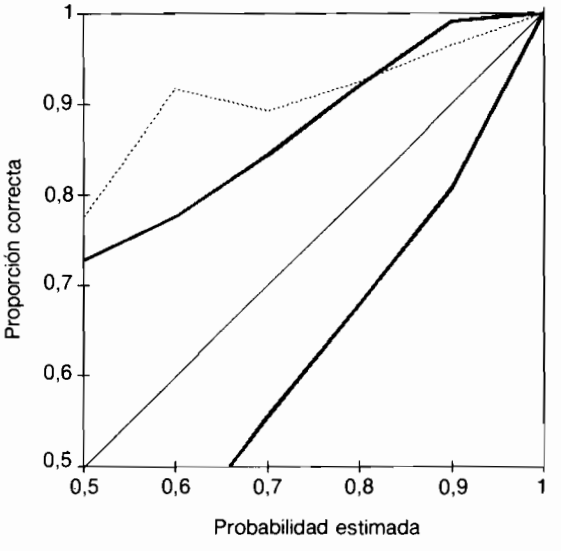

Figura 2.*a*) Curva de calibración para los perfiles de ansiedad.

TABLA 2

Medias para cada factor sobre cada una de las variables. Proporción media de respuestas correctas (PC), probabilidad subjetiva media (PS), índice de calibración (I.ca) y confianza (I.co)

	PC	PS	I.ca	I.co
Perfiles				
Ansiedad	0,92095	0,77344	0,09279	– 0,14752
Depresión	0,85803	0,75229	0,09993	– 0,10577
Mixtos	0,69505	0,70450	0,15159	+ 0,00945
Condición				
G. control	0,77036	0,74700	0,12637	– 0,02336
G. experim.	0,87899	0,73982	0,10318	– 0,13917

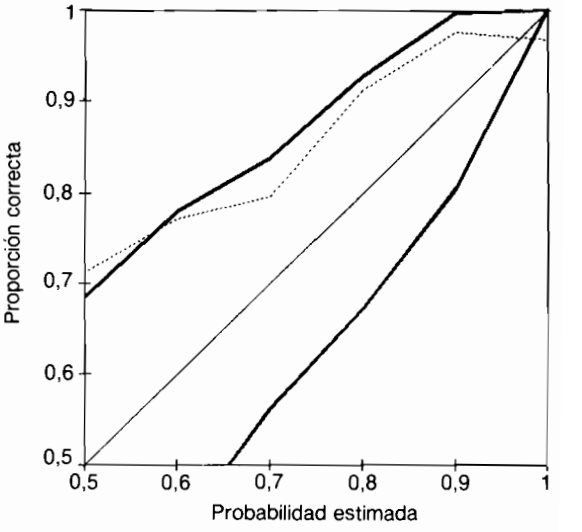

Figura 2.*b*) Curva de calibración para los perfiles de depresión.

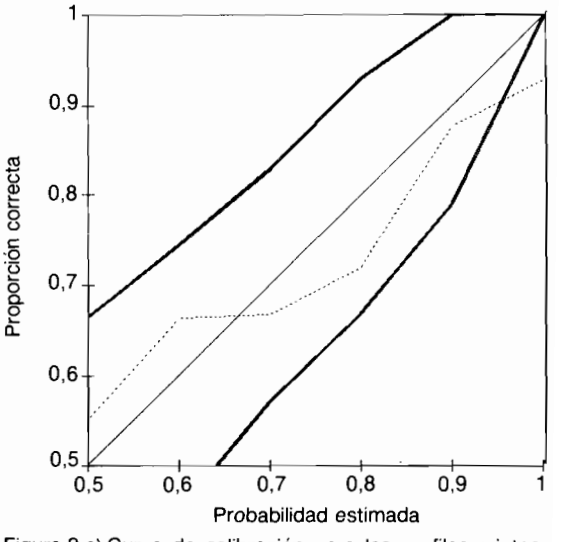

Figura 2.*c*) Curva de calibración para los perfiles mixtos. Todas con sus correspondientes intervalos confidenciales.

po control es mayor. Tal discrepancia se debe a que las curvas están realizadas sobre los datos grupales y están reflejando el índice de calibración medio de los sujetos. Es decir, existen diferencias entre la media de los índices de calibración y el índice de calibración medio debido a la naturaleza de la fórmula (algo similar a la diferencia entre la suma de cuadrados y el cuadrado de la suma). No ocurre lo mismo con el índice de confianza medio reflejado en la curva de calibración.

Esta diferencia, que nosotros sepamos, no ha sido comentada previamente en la literatura sobre el tema, a pesar de su evidente importancia. Por ello, decidimos trabajar únicamente con el I.co. Cuando hablemos de «calibración» nos referiremos a la relación entre PS y PC en términos del I.co.

Consecuencias de la decisión

El segundo factor se refiere a la influencia de las consecuencias derivadas de las decisiones tomadas sobre la actuación de los sujetos. Hemos encontrado diferencias significativas en la proporción de respuestas correctas y en la confianza (véase tabla 1).

La proporción de respuestas correctas bajo la condición con consecuencias fue mayor que bajo la condición sin consecuencias y se acentuó la infraestimación de la confianza.

Gráficamente (figura 3) se puede observar que la curva de calibración del grupo experimental queda por encima de la curva correspondiente al grupo control. La curva de calibración (figuras 4*a* y *b*) para esta última condición se incluye mayormente dentro del área establecida por los límites confidenciales, lo que indica una relativa buena calibración. Sin embargo, en la condición con consecuencias la mayoría de las categorías están fuera de los intervalos confidenciales, indicando una infraestimación de la confianza.

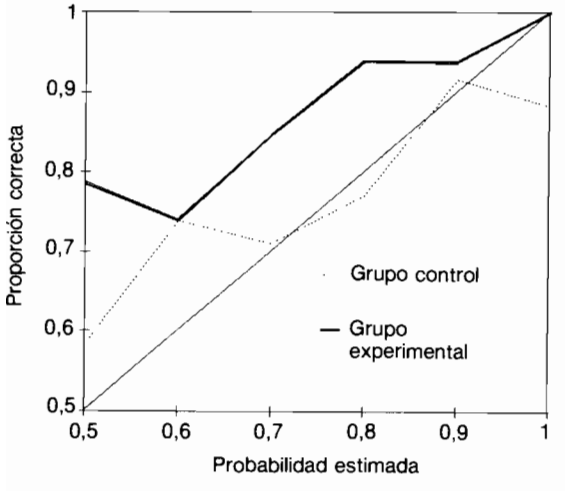

Figura 3. Curvas de calibración para el grupo control y experimental.

Por tanto, la condición en la que se hacen explícitas las consecuencias de la decisión favorece la proporción de diagnósticos correctos y favorece la infraestimación de la confianza.

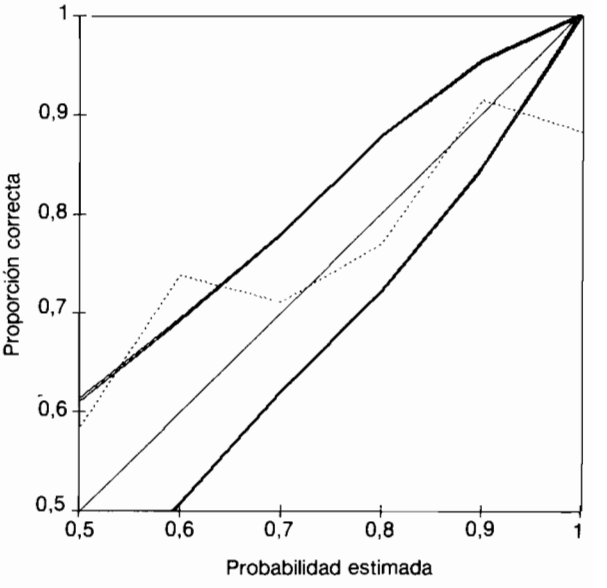

Figura 4.*a*) Curva de calibración para el grupo control.

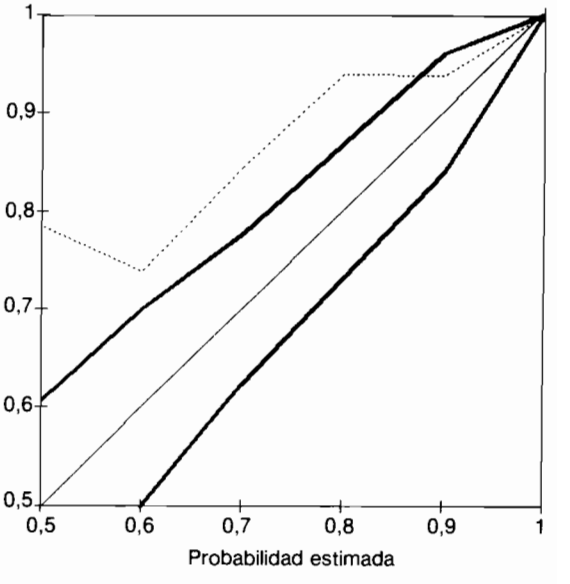

Figura 4.*b*) Curva para el grupo experimental. Ambas con sus respectivos intervalos confidenciales.

Dificultad de la tarea y consecuencias de la misma

Se encontró un efecto significativo de la interacción para las *estimaciones de probabilidad*. En la figura 5

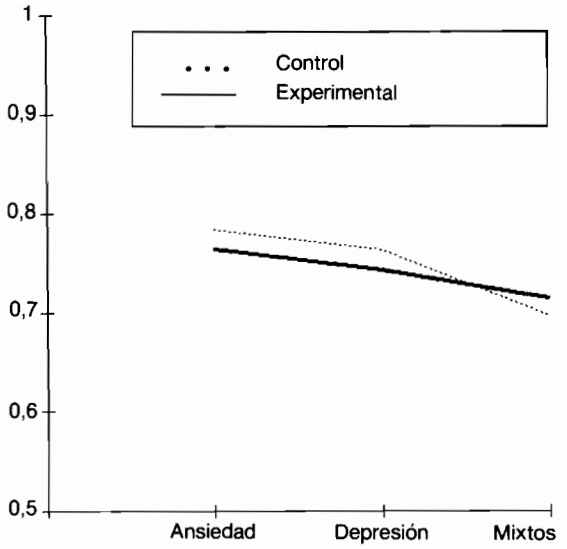

Figura 5. Relación entre los dos factores para cada una de las VD. En esta figura se representan las estimaciones de probabilidad.

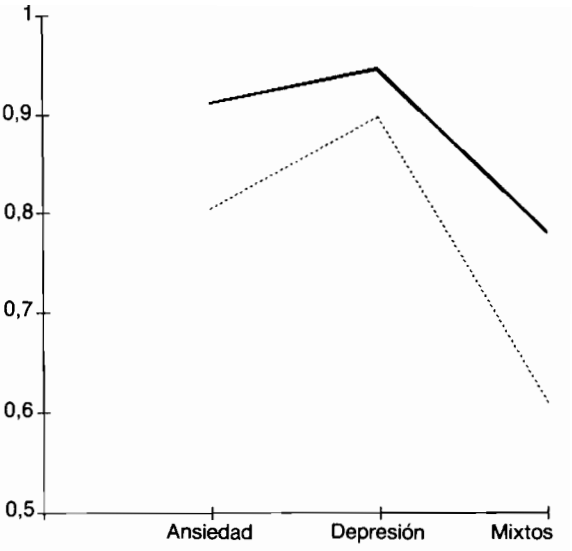

Figura 6. Relación entre los dos factores para cada una de las VD. En esta figura se representa la proporción de respuestas correctas.

se ve cómo las estimaciones de probabilidad son mayores en la condición sin consecuencias en los perfiles «puros». Sin embargo, esta relación se invierte con los perfiles «mixtos», donde las estimaciones de probabilidad de la condición con consecuencias son mayores. La figura 6 muestra cómo la *proporción de respuestas correctas* es menor en los perfiles mixtos que en los de depresión, y ésta menor que en los perfiles de ansiedad. Dicha proporción siempre es menor en el grupo control.

La figura 7 indica, por un lado, que la *confianza* en la condición con implicaciones se atenúa con respecto a la condición sin ellas; por otro, que la confianza se incrementa a medida que los perfiles se hacen más difíciles.

En suma, parece existir una tendencia general a la cautela en la estimación de probabilidades bajo una situación de demanda de compromiso, llegando incluso a producirse una acusada subvaloración de la confianza.

Conclusiones

Las conclusiones derivadas del primer experimento se pueden agrupar en tres puntos:

1. La elaboración de los perfiles según distintos niveles de dificultad resultó apropiada, hecho que se confirmó a partir de la proporción de perfiles diagnosticados correctamente (tabla 2).

2. Se obtuvo una calibración diferencial según la dificultad de la tarea, tal como se predijo en la introducción. Cuando la tarea resulta fácil, los sujetos tienden a infraestimar la confianza de sus diagnósticos (por ejemplo, I.Co = — 0,15 con PC = 0,92; a medida que la dificultad aumenta, decrece dicha infraestimación (I.Co = + 0,009 con PC = 0,70).

El hecho de no haber encontrado sobreconfianza, tal y como detalla la literatura sobre el tema, ha podido deberse a que la tarea no resultara demasiado difícil, ni aun con los perfiles mixtos. Para estos últimos, la media de proporciones de respuestas correctas fue aproximadamente 0,70. Otra posible explicación podría ser que con ítems de conocimiento general el efecto de sobreconfianza se acentua-

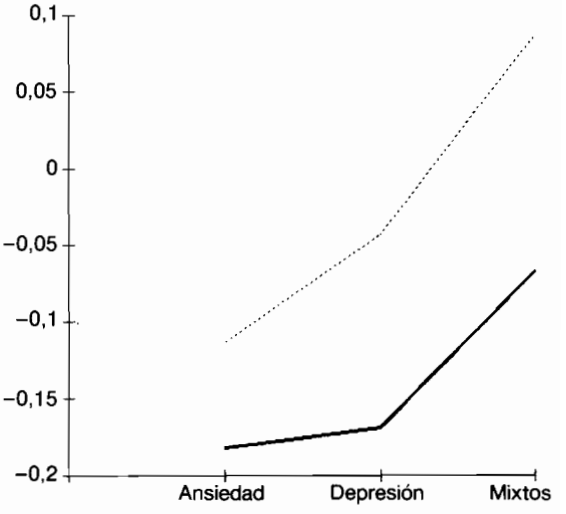

Figura 7. Relación entre los dos factores para cada una de las VD. En esta figura se representa el índice de confianza.

ría en la misma línea que comentan Christensen-Szalanski y Fobian (1991) respecto al sesgo retrospectivo.

La sobreconfianza suele encontrarse asociada a un porcentaje de respuestas correctas inferior al 65 por 100. Concretamente, Lichtenstein y Fischhoff (1977) señalan que «*la calibración va mejorando hasta que se alcanza alrededor de un 80 % de respuestas correctas, después empeora*» (pág. 180); empeora en el sentido de una infraconfianza.

3. También, como predecíamos, hemos encontrado cambios en la calibración al hacer explícitas las consecuencias de la decisión, ya que disminuye el índice de confianza (véase tabla 1).

Parece existir una propensión a la cautela en la valoración de las propias decisiones cuando la tarea puede producir consecuencias negativas. Concretamente, se encuentra que donde suele encontrarse una subestimación de la confianza, ésta se acentúa y, por el contrario, en tareas de mayor dificultad, donde es propio encontrar sobreconfianza, ésta se atenúa. Con otras palabras, existe una tendencia a ser más conservador ante la existencia de compromiso.

A continuación analizaremos las causas que subyacen a la reducción del índice de confianza cuando el sujeto debe considerar las consecuencias de sus decisiones. Por una parte, ante una decisión con repercusiones serias, los sujetos pueden sentir un cierto «miedo» a equivocarse, disminuyendo así la confianza en sus diagnósticos. En caso de estar equivocados, tal estrategia podría facilitar además una (auto)justificación, la cual podría evitar el arrepentimiento posdecisional (Janis-Mann, 1977; Hogarth, 1980).

Por otra parte, podría ocurrir que situaciones de mayor compromiso indujeran a los decisores a realizar un análisis más detallado de la tarea y de sus propios conocimientos. De hecho, varios sujetos manifestaron que atendieron más a los perfiles en la condición de consecuencias: «...Lo he pensado más a la hora de dar una respuesta», «...me he fijado más en la primera parte» (g. experimental).

Es decir, aunque los resultados muestran que los sujetos son más cautos cuando han de considerar las consecuencias de la decisión, nuestros datos no permiten ser concluyentes respecto a si ello se debe a un análisis más detenido de la tarea. Con el fin de comprobar precisamente este punto se diseñó el segundo experimento.

SEGUNDO EXPERIMENTO

El *objetivo* principal de este experimento es determinar si la consideración de las dos posibles alternativas induce a los sujetos a calibrar mejor sus decisiones. Con otras palabras, cuando la tarea demande mayor procesamiento de la información, ¿se evitará la tendencia a la sobreconfianza?

Hogarth (1980, 1987) señaló que una situación de elección no implica necesariamente la evaluación de todas las alternativas. Aunque los conceptos de juicio y elección son considerados sinónimos desde el punto de vista normativo (por ejemplo, se seleccionará la alternativa con mayor valor esperado), no lo son desde una perspectiva psicológica. Así, la elección no implica necesariamente una evaluación, aunque la evaluación o el juicio pueden representar una ayuda para la elección, parafraseando a Hogarth (1980): «*El juicio sirve para reducir la incertidumbre y el conflicto en la elección mediante un proceso de razonamiento y evaluación deliberado de la evidencia*» (pág. 131). Billings y Scherrer (1988) también son partidarios de distinguir entre juicio y elección. Estos autores comprobaron experimentalmente que las estrategias de procesamiento de la información eran diferentes dependiendo de si la tarea demandaba un juicio o una elección entre diversas alternativas; concretamente, en el caso de una evaluación previa de éstas se utilizaban más frecuentemente estrategias compensatorias, mientras que el porcentaje de estrategias no compensatorias (por ejemplo, eliminación por aspectos) aumentaba en caso de elección.

Esto nos induce a hipotetizar que, en una situación de evaluación, la mencionada reducción de incertidumbre y conflicto puede contribuir también a una reducción de la sobreconfianza, frente a una situación donde no se requiere, de manera explícita, tal evaluación. En esta misma línea, Zakay (1985) encontró que las estrategias no compensatorias venían asociadas con mayores estimaciones de seguridad en las decisiones.

Si reflexionamos sobre la tarea del primer experimento, observamos que el sujeto, para establecer los diagnósticos, no tiene por qué analizar necesariamente toda la información presentada. Así, pongamos por caso, un sujeto al leer «un intento de suicidio» puede seleccionar la alternativa de depresión, anclándose en dicho síntoma, sin considerar el resto de ellos. Una manera de evitar esto, o al menos de atenuarlo, podría ser pedir a los sujetos que evaluasen las dos posibles alternativas diagnósticas.

Desde el punto de vista metodológico se introduce una serie de modificaciones respecto al experimento anterior. En primer lugar, las estimaciones de la seguridad de los sujetos sobre sus diagnósticos se pedirán en porcentajes y no en probabilidades, puesto que la mayoría de los sujetos del primer experimento manifestaron que se hubiesen sentido más «cómodos» de haber estimado su confianza mediante porcentajes. Para la realización de los análisis, éstos se transformarán en PS —la utilización de porcentajes no ha provocado diferencias en el resultado principal de sobreconfianza (Arkes, Christensen, Lai y Blumer, 1987)—. Adicionalmente, consideramos conveniente aumentar el número de perfiles para obtener una medida más estable, así como utilizar únicamente perfiles «mixtos» por motivos de validez ecológica. Los psicólogos no suelen encontrarse con pacientes cuyos síntomas correspondan a un listado dentro de las categorías

de trastornos del DSM III. Por el contrario, es más frecuente encontrar personas cuyos síntomas pertenezcan a distintas categorías diagnósticas.

Método

Sujetos

Participaron un total de 30 sujetos, estudiantes del último curso de la Facultad de Psicología de la UAM, con conocimientos en Psicodiagnóstico y Psicopatología.

Estímulos

Un total de 24 perfiles elaborados a partir de los síntomas propios de trastorno de ansiedad y de depresión citados en el DSM III.

Todos los perfiles pueden ser calificados de «mixtos», aunque existe un diagnóstico principal.

El diagnóstico correcto de la mitad de dichos perfiles corresponde a trastorno de ansiedad y la otra mitad a trastorno afectivo. El procedimiento de elaboraciòn de dichos perfiles fue similar al descrito en el primer experimento.

La tarea

Los sujetos debían determinar (y evaluar, según la condición) qué tipo de trastorno psicológico predominaba en una serie de 24 perfiles, y, en segundo lugar, estimar la seguridad con la que habían dado su respuesta, en términos de porcentaje (50-100 por 100).

Diseño y procedimiento

La *variable independiente* es la demanda de la tarea: evaluación o juicio *versus* elección (v. intrasujeto).

La *variable dependiente* es la adecuación de las estimaciones de probabilidad subjetiva de los sujetos. Las medidas utilizadas son: PC, PS e I.co.

Se contrabalanceó la presentación de las condiciones y, entre la aplicación de una condición y otra, transcurrió un mínimo de dos semanas. En las dos condiciones se utilizaron los mismos perfiles, aunque se aleatorizó el orden de presentación. La duración de la tarea fue de aproximadamente media hora por condición.

La manipulación de la *demanda de la tarea* se realizó de la siguiente manera:

Condición de evaluación o juicio (Cond. J). Se presentó a cada sujeto un cuadernillo con 24 perfiles y se pidió que sobre cada uno de ellos evaluara el grado de trastorno que presentaba el supuesto paciente, tanto de ansiedad como afectivo. Esta evaluacion la realizaba el sujeto marcando una cruz sobre una escala en la que uno de los polos representaba mínimo grado de trastorno (de ansiedad o afectivo) y el otro polo, grado máximo. Esta manipulación es similar a la realizada por Billings y Scherrer (1988) para diferenciar entre juicio y elección. En segundo lugar, debían dilucidar qué tipo de trastorno predominaba y, por último, estimar la confianza de su diagnóstico, mediante un porcentaje entre 50 y 100 por 100, donde una respuesta de 100 por 100 indicaba una seguridad absoluta en el diagnóstico establecido, mientras que el 50 por 100 indicaba incertidumbre total.

Condición de elección (Cond. E). Es semejante a la condición control del primer experimento. El sujeto debía decidir si cada uno de los 24 perfiles presentados correspondía a un trastorno de ansiedad o a un trastorno de depresión, elegir una de las alternativas y manifestar el grado de confianza en su elección (también en términos de porcentaje, 50-100 por 100).

Resultados y discusión

Análisis sobre el total de perfiles

En las figuras 8*a* y *b*) se muestran las curvas de calibración correspondientes a las dos condiciones. Como se observa, en las dos condiciones se produce sobreconfianza en las categorías extremas ya que ambas curvas quedan por debajo de la diagonal. En la mayor parte de las categorías, las curvas quedan por debajo de los intervalos confidenciales.

No se encontró ninguna diferencia significativa en las tres medidas tomadas (PC, PS, I.co) cuando se realizaron las correspondientes diferencias entre medias (véase tabla 3).

El hecho de no encontrar diferencias significativas entre ambas condiciones pudo deberse a que la manipulación experimental realizada no fuese la adecuada para diferenciar entre juicio y elección. Sin embargo, como ya se dijo, se recurrió a una manipulación de la demanda de la tarea propia de trabajos de decisión donde se hace este tipo de distinción (Billings y Scherrer, 1988). De esta manera, antes de descartar nuestra hipótesis, consideramos pertinente comprobar si existían perfiles que pudieran estar distorsionando nuestros resultados. Autores como Fischhoff, Slovic y Lichtenstein (1977) y May (1986a y b) reconocen la existencia de ítems sobre los que la calibración es imposible. A estos ítems los hemos llamado ítems «equívocos» y se definen como aquellos cuya proporción de respuestas correctas es menor que 0,50 y, por tanto, provocan necesariamente sobreconfianza (recuérdese que la escala de respuesta comienza en el valor 50 por 100). Este argumento nos llevó a analizar cada uno de los ítems.

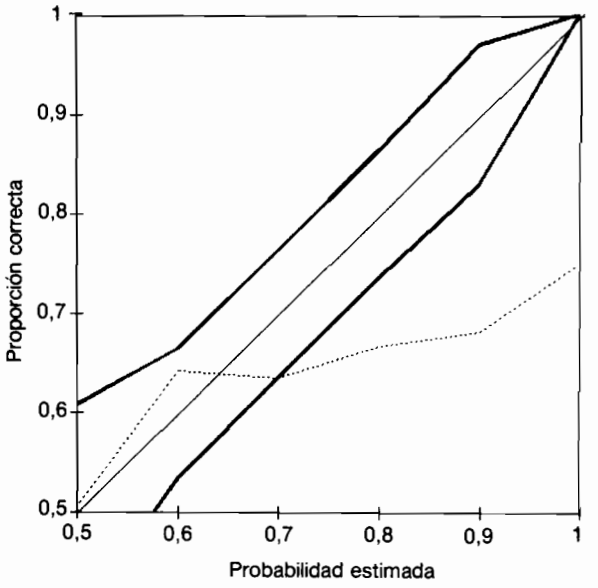

Figura 8.*a*) Curva de calibración para la condición de elección y de evaluación. Se representa la curva de calibración para la condición de elección.

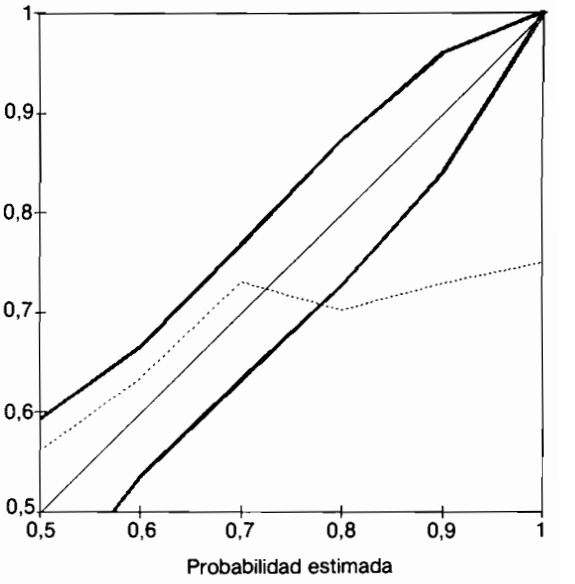

Figura 8.*b*) Para la condición de elección. Ambas con sus respectivos intervalos confidenciales.

TABLA 3

Media y desviación típica para cada una de las medidas tomadas en el experimento 2

Medidas	Evaluación		Elección	
	Media	D. típica	Media	D. típica
PC	0,669	0,078	0,636	0,105
PS	0,682	0,089	0,691	0,078
I.co.	+ 0,013	0,108	+ 0,055	0,129

Análisis de los perfiles

Para determinar la PC de los 24 perfiles consideramos conveniente aumentar el número de sujetos; de esta manera se presentó la tarea a 80 ss más, pertenecientes a la misma población que la muestra experimental. Se identificaron cuatro perfiles sobre los que la proporción de respuestas correctas fue inferior a 0,50.

Ahora bien, puesto que el criterio sobre el que estábamos clasificando los diagnósticos de correctos e incorrectos dependía del DSM III nos quedaba la duda de si tendríamos que haber eliminado más o menos ítems de haber utilizado otro criterio, por ejemplo, el de un conjunto de expertos. De esta manera, pedimos a un conjunto de cinco expertos (profesores del área de Psicología de la Salud de la Facultad de Psicología de la UAM) que realizaran la tarea de diagnosticar los 24 perfiles.

Con los datos obtenidos, calculamos las siguientes medidas: 1) el índice de porcentaje de acuerdo corregido, Kappa de Cohen, entre la moda de las respuestas de los jueces y la clasificación realizada según el DSM III. Obtuvimos un valor de este estadístico de 0,824, y 2) el acuerdo interjueces a partir del coeficiente Kappa generalizado para más de dos jueces —estadístico cuyo valor fue de 7,1092 (distribución normal unitaria)—. Por tanto, el acuerdo entre los jueces es significativamente mayor que el esperable por azar. Es decir, los cinco jueces tienen criterios de clasificación similares y, además, éstos son semejantes a los establecidos por el DSM III.

Observando las clasificaciones modales de los expertos para los diversos perfiles y las del DSM III, encontramos tan sólo tres desacuerdos, desacuerdos que coincidieron con tres de los cuatro perfiles con porcentaje de aciertos bajo. Apoyándonos en estos dos procedimientos, decidimos eliminar los cuatro perfiles ya mencionados.

A continuación, procedimos a analizar de nuevo los datos de los 30 sujetos de la muestra experimental, sobre el subconjunto de 20 perfiles.

Análisis sobre el subconjunto de perfiles

En la figura 9.*a*) se representa la curva de calibración de la condición de elección, en la figura 9.*b*) la curva de la condición de evaluación; ambas con sus respectivos intervalos confidenciales.

Se observa cómo las distancias entre las curvas se hacen más pronunciadas cuando se consideran sólo los 20 perfiles. Para la condición de evaluación se observa infraconfianza en las categorías de 0,6 y 0,7. La condición de elección muestra sobreconfianza en la categoría de 0,90.

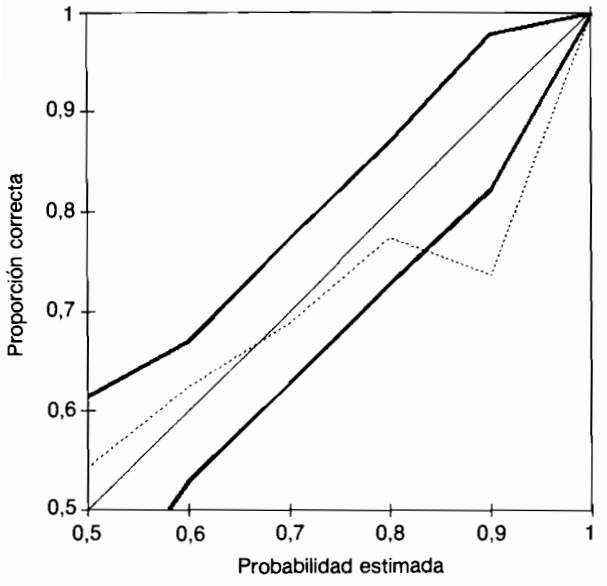

Figura 9.*a*) Se representa la curva de calibración para la condición de elección para el subgrupo de perfiles.

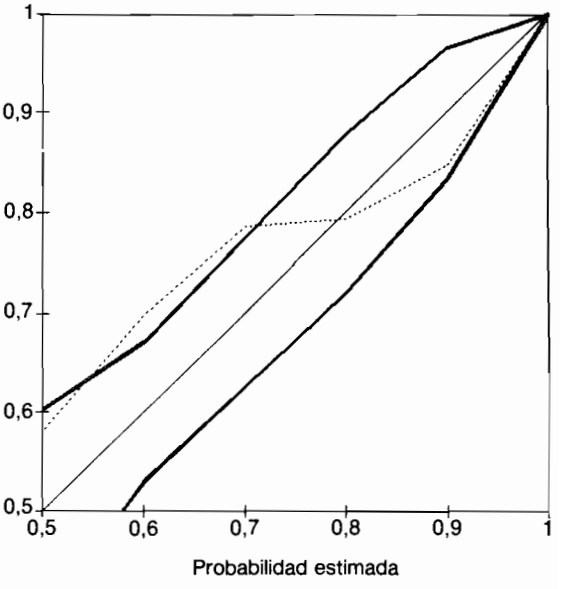

Figura 9.*b*) Para la condición de elección, con sus respectivos intervalos confidenciales.

Llama la atención haber encontrado una calibración perfecta para la categoría 1 en ambas condiciones. Esta categoría apenas es elegida, los sujetos pocas veces están completamente seguros de sus respuestas, pero cuando lo están, éstas están justificadas. Este resultado está en consonancia con el de Lichtenstein y Fischhoff (1981), quienes observaron una mejor calibración en aquellos sujetos que nunca utilizaban la categoría de certidumbre absoluta.

Como se detalla en la tabla 4, existen diferencias significativas entre las dos condiciones en la proporción de respuestas correctas y en el índice de confianza. Los sujetos en la condición de evaluación o juicio son más precisos en sus diagnósticos y menos confiados. Resultado que está en consonancia con nuestra hipótesis de partida: el hecho de considerar las dos posibles alternativas de la tarea (evaluación versus elección) hace mejorar la proporción de respuestas correctas, además de evitar la tendencia a la sobreconfianza.

TABLA 4

Media y desviación típica para cada una de las medidas tomadas en el experimento 2 para el subconjunto de perfiles. En las dos últimas columnas aparece el valor del estadístico t con su probabilidad asociada

Medidas	Evaluación		Elección		Comparaciones	
	Media	D. típ.	Media	D. típ.	t	p
PC	0,738	0,090	0,678	0,129	2,24	**0,0326**
PS	0,683	0,089	0,689	0,079	−0,38	0,7086
I.co	− 0,055	0,095	0,011	0,135	− 2,26	**0,0318**

Conclusiones del segundo experimento

En este experimento se pone de manifiesto la relevancia de hacer un análisis por ítems con el fin de contar con un material que no provoque resultados distorsionantes. Pensamos que, en muchos trabajos, el resultado de una sobreconfianza excesiva de los sujetos podría deberse, en parte, a la inclusión, en los estímulos presentados a los sujetos, de ítems «equívocos». Consecuencia de este planteamiento fue el cálculo de la PC de los perfiles utilizados en el primer experimento. En ninguno de ellos la proporción de respuestas correctas se encontraba por debajo de 0,50. Por tanto, los resultados previamente obtenidos no estaban distorsionados por el material.

Respecto a la manipulación experimental realizada. Parece que, ante una situación en la que el sujeto debe evaluar las posibles alternativas, se mejora la proporción de respuestas correctas y se confía menos en las decisiones tomadas (en términos del I.co). Lo que nos conduce a señalar las ideas de Koriat, Lichtenstein y Fischhoff (1980). Estos autores sugieren que la sobreconfianza puede deberse a un sesgo confirmatorio de la información; bien sobre una alternativa candidata antes de ser elegida, bien una vez que se ha elegido la alternativa y se revisa la evidencia para estimar la confianza. En otro contexto (toma de decisiones entre apartamentos) se ha demostrado la existencia del llamado «heurístico del candidato». Según este heurístico, los decisores

presentan desde momentos muy tempranos del proceso de decisión una alternativa candidata que sesga el proceso de recogida de información (Gambara y León, 1987).

Koriat et al. (1980) hallaron una reducción significativa de la sobreconfianza cuando pidieron a sus sujetos que especificaran razones por las que su elección podría no ser correcta. Aunque nuestra manipulación experimental no obliga a buscar razones no confirmatorias, al menos de manera explícita, el hecho de pedir que se evalúen las dos posibles alternativas puede impedir al sujeto sesgar la búsqueda de información únicamente hacia una de ellas.

Conclusiones generales

El objetivo global de la investigación ha sido analizar la confianza de los sujetos en una tarea de diagnóstico clínico, a través del paradigma de la calibración.

Un primer aspecto a resaltar se relaciona con la *tarea experimental*. Al contrario que la mayoría de los estudios sobre calibración, hemos buscado una tarea cuyo contenido fuese relevante no sólo para los sujetos sino también para el investigador, para así superar el reduccionismo de una gran parte de estos trabajos, los cuales se limitan a analizar la calibración preguntando a los sujetos, por ejemplo, ¿qué es Guadagugu?

Pensamos que, en este tipo de trabajos, un factor de trascendencia a la hora de identificar las desviaciones de los sujetos respecto a patrones normativos es la propia tarea. En este sentido, nos referiremos, de nuevo, a Christensen-Szalanski y Fobian (1991), quienes identificaron un efecto del sesgo retrospectivo tres veces mayor sobre tareas que incluían ítems de conocimiento general, que sobre otro tipo de tareas de contenido más ecológico.

El experimento primero, aparte de su obligado carácter descriptivo por la novedad de la tarea, se caracterizó por la manipulación de las *consecuencias* derivadas de los diagnósticos que los sujetos debían establecer. Esta variable no había sido considerada dentro de los estudios de calibración, pero sí dentro de un marco más amplio de toma de decisiones (por ejemplo, Billings y Scherrer, 1988). Dado que la manipulación experimental de la importancia de la decisión, cuando se hace hincapié en las repercusiones de la alternativa elegida, modifica el proceso de decisión, nos preguntamos si también influiría esta variable sobre la seguridad con la que se elige.

El diseño del segundo experimento fue consecuencia de los resultados del primero; en él también se manipuló una variable (juicio *versus* elección) que ha originado un gran cuerpo de trabajos dentro de la teoría de la decisión conductual (por ejemplo, Billings y Scherrer, 1988; Einhorn y Hogarth 1981), pero que no ha sido manipulada previamente por los investigadores en calibración. Así, se ha encontrado que existen diferencias entre *evaulación y elección* en lo que respecta a la cantidad de información analizada, a las estrategias de procesamiento (inter o intra alternativas) y a la variabilidad de información utilizada a través de alternativas. ¿Existen también diferencias respecto a la seguridad entre juicios o elecciones? El objetivo del segundo experimento fue dar respuesta a esta cuestión.

En resumen, aparte de haber encontrado sobre la tarea de diagnóstico un patrón de calibración similar al mencionado en la literatura, las variables utilizadas en nuestra investigación (*consecuencias de la decisión* y *juicio versus* elección) incidieron de manera significativa en la relación entre precisión y seguridad en los diagnósticos.

Concretamente, en una situación en la que se explicitan las consecuencias de la decisión, los sujetos mejoran la precisión de sus diagnósticos y son más prudentes en sus estimaciones de probabilidad subjetiva. Además, los resultados del experimento segundo revelan que, en una situación donde los sujetos tienen que evaluar explícitamente ambas alternativas y no sólo elegir una de ellas, la precisión de los diagnósticos aumenta y el índice de confianza se hace más pequeño, alejándose del sesgo de sobreconfianza.

Finalmente, consideramos oportuno hacer mención a algunos problemas metodológicos relativos a la variable dependiente.

Como ya señalamos, propusimos la utilización de intervalos confidenciales a la hora de interpretar las curvas de calibración; sin embargo, este mismo problema de interpretación se encuentra en relación al índice de calibración y de confianza, debido al desconocimiento de la distribución muestral de ambos estadísticos. Además, ya señalamos en el primer experimento la discrepancia de resultados entre el índice de calibración medio y la media del índice de calibración. Con el fin de solventar estos problemas, al menos el primero de ellos, se está llevando a cabo una simulación montecarlo sobre la distribución muestral de tales índices (Gambara, Ruiz y León, en preparación).

Referencias

Adams, J. K. (1957). A confidence scale defined in terms of expectes percentages. *American Journal of Psychology, 70*, 422-436.

Arkes, H. R., Christensen, C., Lai, C. y Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133-144.

Billings, R. S. y Scherrer, L. L. (1988). The effects of Response Mode and Importance on Decision-Making Strategies: Judgement versus Choice. *Organizational Behavior and Human Decision Processes, 41*, 1-19.

Christensen-Szalansky, J. J. J. y Fobian, C. (1991). The Hindsight Bias: A meta-analysis. *Organizational Behavior and Human Decision Procesess, 48*, 147-168.

De Finetti, B. (1974). *The Theory of Probability*, 2 vol. New York: Wiley.

Einhorn, H. J. y Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology, 32*, 53-88.

Gambara, H. Ruiz, M. A. y León, O. (en preparación). Una simulación de la distribución muestral del índice de calibración.

Gambara, H., y León, O. G. (1987). Uso de la información en el proceso de toma de decisiones. *Boletín de Psicología, 17*, 83-99.

Hogarth, R. M. (1980). *Judgment and Choice: The Psychology of Decision.* Chichester: Wiley (2.ª ed., 1987).

Janis, I. L. y Mann, L. (1977). *Decision Making: A Psychological Analysis of Conflict, Choice and Commitment.* New York: Free press.

Koriat, A., Lichtenstein, S. y Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107-118.

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

Lichtenstein, S. y Fischhoff, B. (1981). *The effects of gender and instructions on calibration* (Technical Report PTR-1092-81-7). Lichtenstein, S., Fischhoff, B. y Phillips, L. D. (1982). Calibration of probabilities: State of the art to 1980. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under Uncertainty: Heuristics and Biases.* New York: Cambridge University Press.

May, R. S. (1986a). Inferences, subjective probability and frequency of correct answers. A cognitive approach to the overconfidence phenomenon. En B. Brehmer, H. Jungermann, P. Lourens y B. Sevón (Eds.), *New directions in Research on Decision Making.* North-Holland: Elsevier Science Publishers B.V.

May, R. S. (1986b). Overconfidence as a result of incomplete and wrong Knowledge. En R. W. Scholz (Ed.), *Current Issues In West German Decision Research.* Bern, New York, Frankfurt: Peter Lang.

Paese, P. W. y Sniezek, J. A. (1991). Influences on the Appropiateness of Confidence in Judgment: Practice, Effort, Information and Decision-Making. *Organizational Behavior and Human Decision Processes, 48*, 100-130.

Payne, J. W., Bettman, J. R. y Johnson, E. J. (1992). Behavioral Decision Research: A Constructive Processing Perspective. *Annual Review of Psychology, 43*, 87-131.

Slovic, P., Lichtenstein, S. y Fischhoff, S. (1988). Decision Making. *Steven's Handbook of Experimental Psychology.* New York: Wiley.

Zakay, D. (1985). Post-Decisional Confidence and conflict experienced in a choice process. *Acta Psychologica, 58*, 75-80.