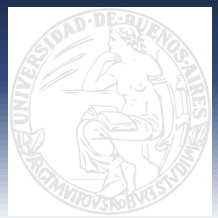
DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity

Fermín Travi **Gonzalo Ruarte**

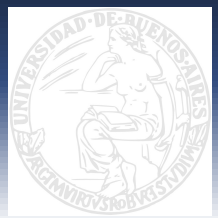
Primer Cuatrimestre de 2020

Facultad de Ciencias Exactas y Naturales, UBA



Contenido

1. Introducción
2. El Modelo
3. Experimentación
4. Resultados
5. Conclusiones



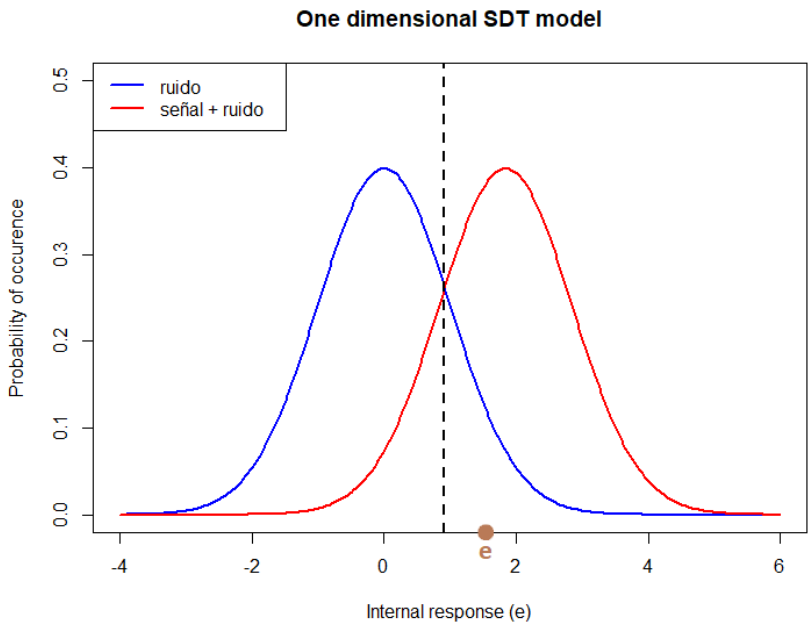
Preguntas que dieron lugar a este trabajo

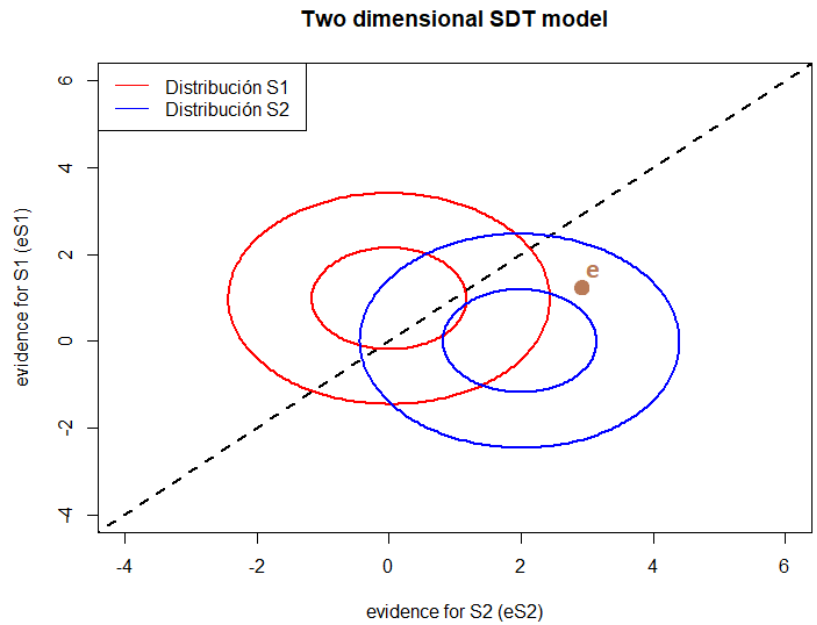
1. La puntuación de la confianza se correlaciona fuertemente con la probabilidad de que la decisión que tomamos haya sido correcta.
2. Por lo tanto, se suele asumir que puntuamos nuestra confianza ponderando la misma evidencia que utilizamos para tomar decisiones.
3. Sin embargo, al menos un estudio previo sugiere lo contrario: para evaluar la confianza, solo tenemos en cuenta la evidencia *a favor* de la decisión que tomamos.
4. ¿Es cierto que ignoramos la evidencia *en contra* de nuestra decisión al momento de evaluar nuestra confianza?
5. Si es así, ¿podríamos ser capaces de alterarla para considerar *ambas* evidencias?

Keywords

1. Response Congruent Evidence (RCE): Evidencia a favor de la opción elegida.
2. Response Incongruent Evidence (RIE): Evidencia en contra de la opción elegida.
3. Balance of Evidence (BE): Se tienen en cuenta ambas evidencias.
4. BE es el paradigma que predomina a la hora de representar el comportamiento humano al tomar una decisión (binaria), ¿sucede lo mismo a la hora de determinar la confianza o predomina la RCE?

One-dimensional vs. two-dimensional SDT model



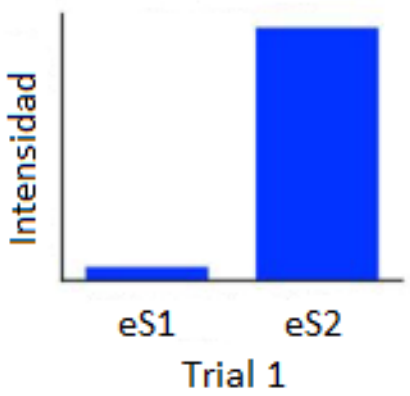


- e es un número, al igual que el criterio. Si e > criterio, respondo "si". Caso contrario, respondo "no".
- e = (eS1, eS2). El criterio es la recta eS1 - eS2 = 0. Si eS1 > eS2, respondo "S1". Caso contrario, respondo "S2".

Two-dimensional SDT

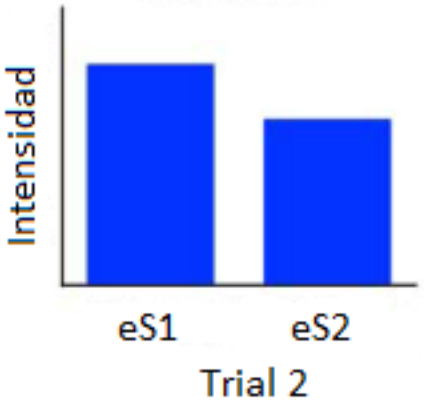
1. Las experimentaciones se realizan con dos estímulos en simultáneo, llamados S1 y S2 respectivamente.
2. Se utiliza un modelo SDT de dos dimensiones para poder analizar cuánto influye la evidencia de cada estímulo por separado al momento de evaluar la confianza.
3. Las distribuciones de los estímulos se representan con círculos centrados en $(\mu_2,0)$ y a $(0,\mu_1)$ respectivamente. El radio de cada círculo corresponde a la desviación estándar correspondiente o a un múltiplo de ella.
4. La idea es que las decisiones son representadas por la distancia de un punto a los centros de ambos círculos.
5. Llamando eS1 (eS2) a la evidencia del estímulo S1 (S2), si eS1 $>$ eS2 el punto se encontrará más cerca de $(0,\mu_1)$ que de $(\mu_2,0)$, con lo cual equivale a optar por S1.
6. Existe una curva (decision axis) que representa los casos en que eS1 $=$ eS2, de forma que debajo de ella equivale a optar por S2 y por encima equivale a optar por S1.

Midiendo la confianza: BE vs. RCE



Respondo "S2" correctamente,

- BE: Como eS2 >>> eS1, respondo con confianza alta.
- RCE: Como eS2 es alto, respondo con confianza alta.

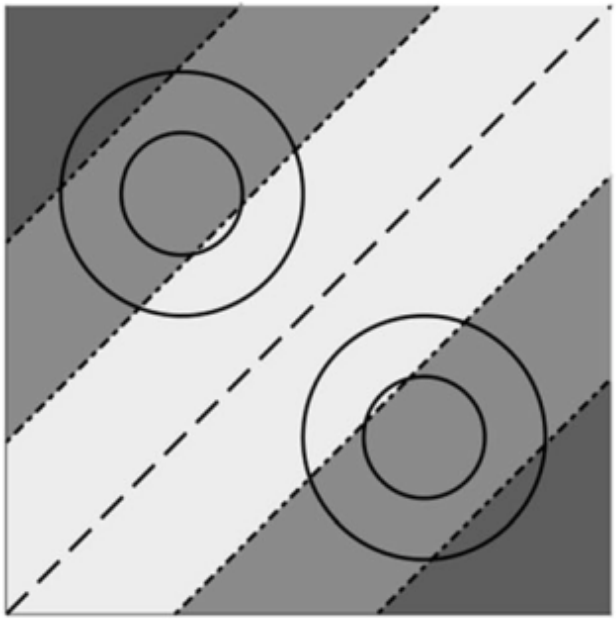


Respondo "S1" incorrectamente (el estímulo era S2),

- BE: Como eS1 es apenas mayor que eS2, respondo con confianza baja.
- RCE: Como eS1 es alto, respondo con confianza alta, a pesar de estar equivocado (!).

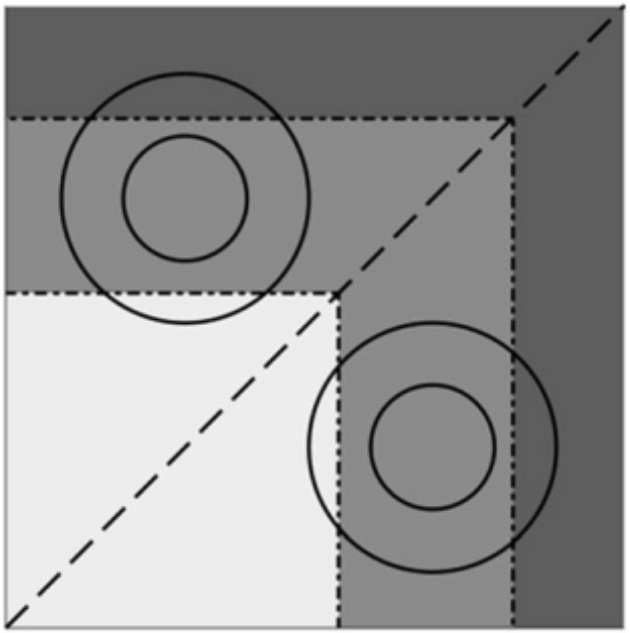
Midiendo la confianza: BE vs. RCE

BE



- Los niveles de confianza se toman a partir de $\|eS1 - eS2\|$, lo cual equivale a trazar franjas paralelas al decision axis en el gráfico.

RCE



- Los niveles de confianza se toman a partir de eS1 (ó de eS2), lo cual equivale a trazar rectángulos concéntricos dentro del gráfico partiendo de la esquina inferior izquierda.

Predicción de la regla RCE

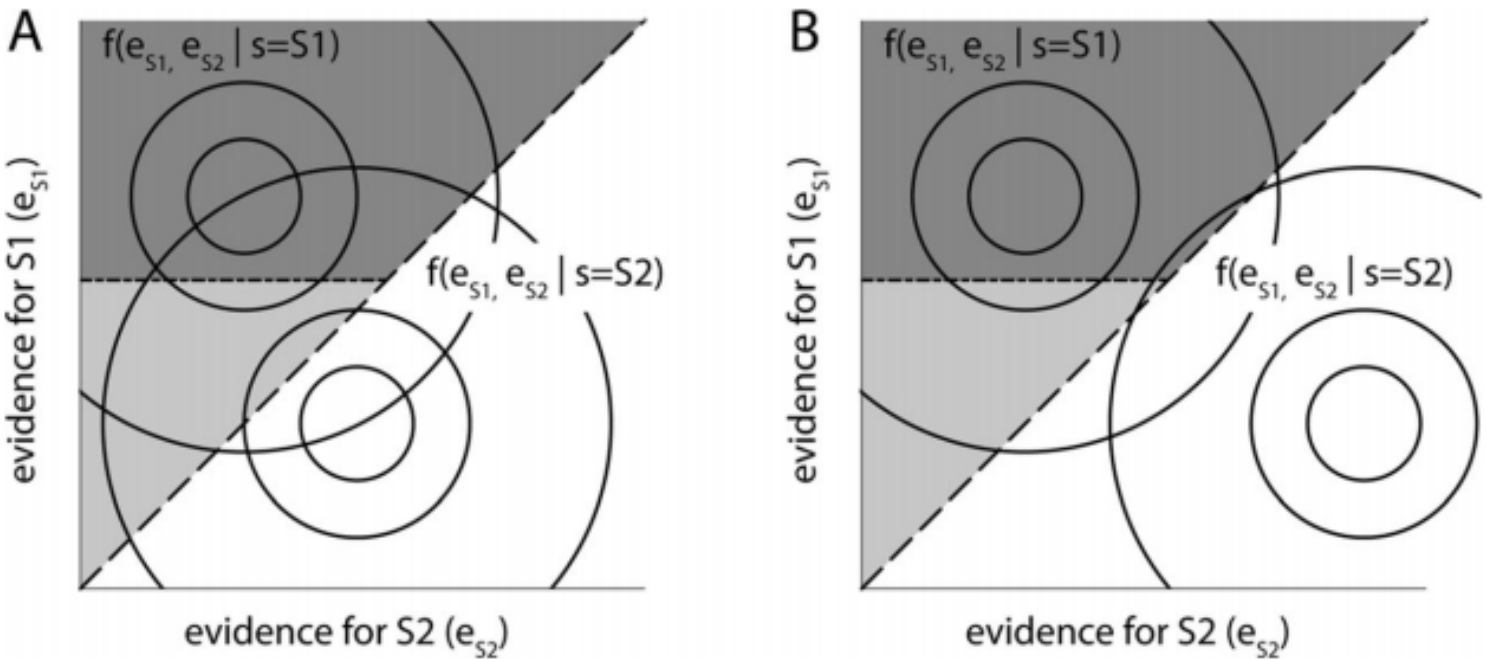


Figura: La regla RCE predice que, a medida que aumenta la *task performance* (d'), disminuye (!) la sensibilidad metacognitiva para las respuestas "S1": la fracción de respuestas "S1" *correctas* con alta confianza es la misma en ambos gráficos, pero la fracción de respuestas "S1" *incorrectas* con alta confianza es mayor en B que en A.

Predicción de la regla RCE

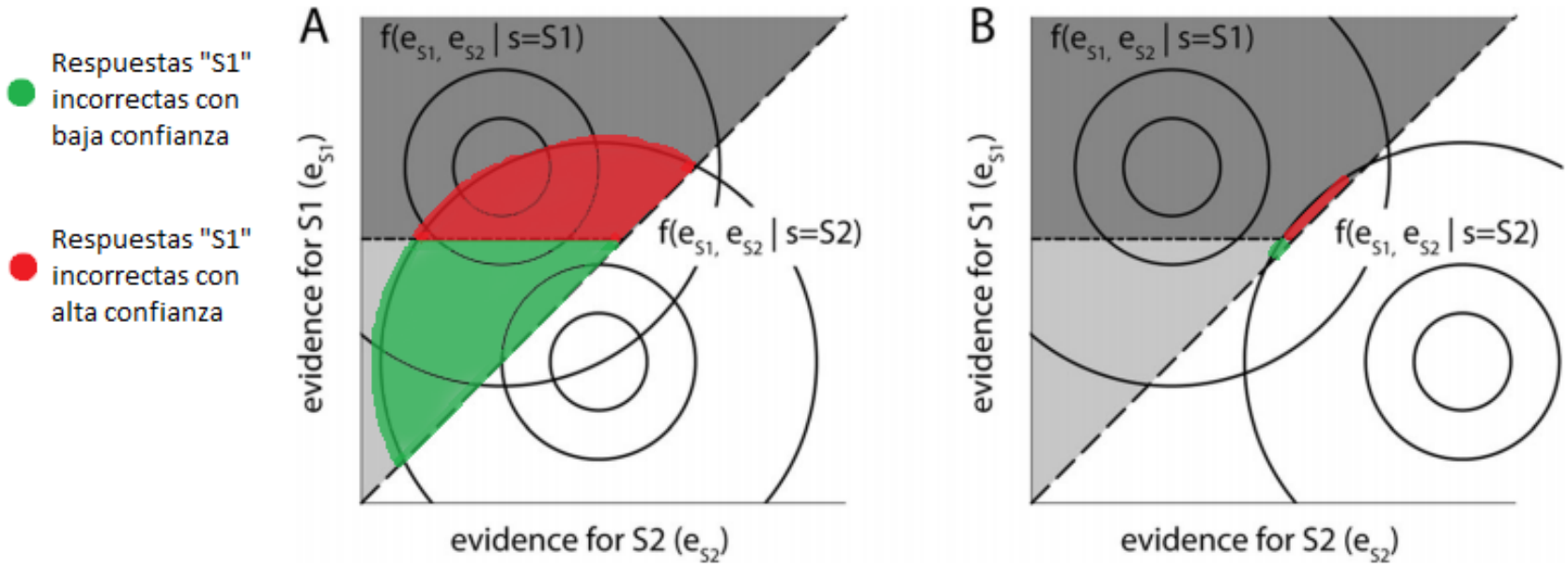
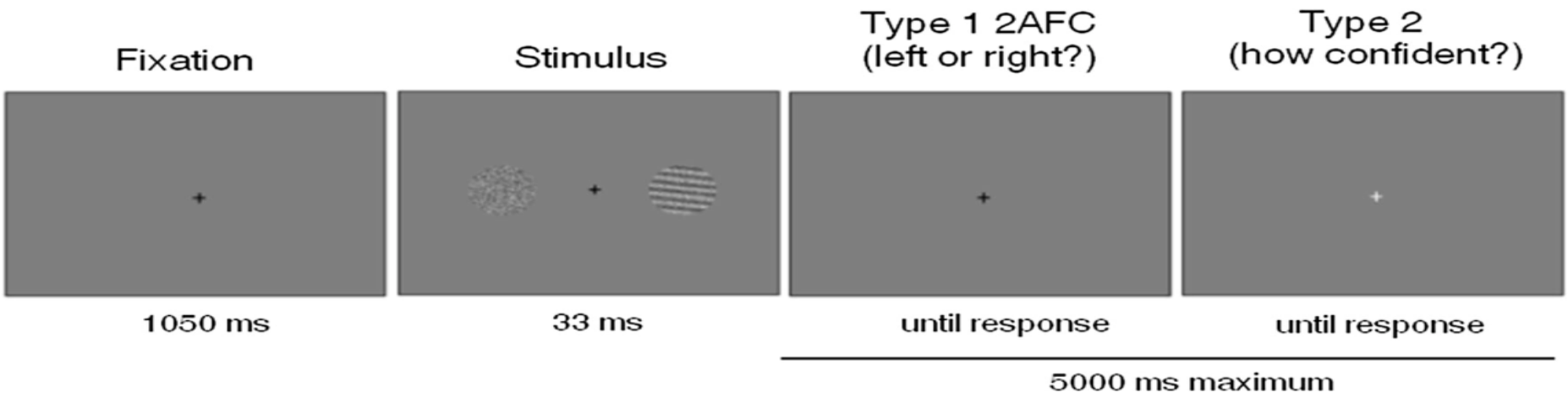


Figura: La regla RCE predice que, a medida que aumenta la *task performance* (d'), disminuye (!) la sensibilidad metacognitiva para las respuestas "S1": la fracción de respuestas "S1" *correctas* con alta confianza es la misma en ambos gráficos, pero la fracción de respuestas "S1" *incorrectas* con alta confianza es mayor en B que en A.

Diseño del experimento y sus simulaciones



1. La intensidad del estímulo S1 se mantuvo constante, mientras que la de S2 fue variando entre 5 niveles distintos, y luego se evaluó el rendimiento de la tarea (d') y la metacognición (meta-d') de manera separada para ambas reglas.
2. Para evaluar predicciones respecto de ambas reglas (BE y RCE), se hicieron simulaciones (que luego nosotros replicamos) basadas en este modelo.
3. Experimento 1: se evaluó entre BE y RCE para ver cuál representa mejor los patrones de respuesta humanos.

Experimento 1: Simulaciones

- Si *meta* $d' = d'$, entonces el observador es óptimo y usa la regla BE

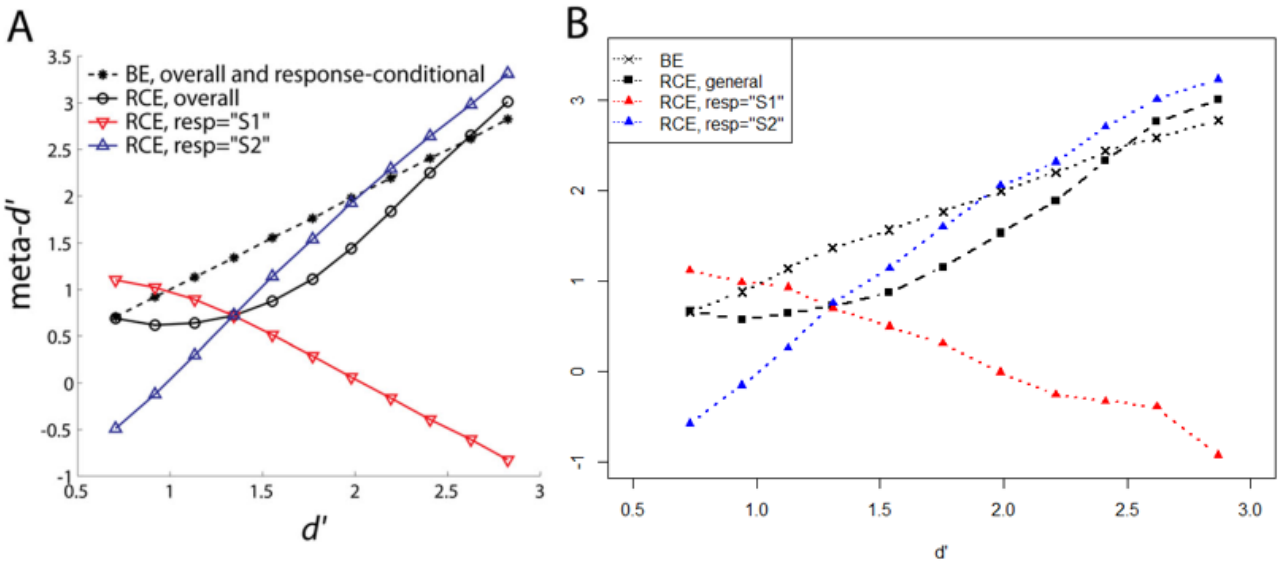


Figura: A) Predicciones hechas en el paper usando los modelos descriptos previamente. B) Simulaciones hechas por nosotros.

Experimento 1: Resultados

- Si *meta* $d' = d'$, entonces el observador es óptimo y usa la regla BE

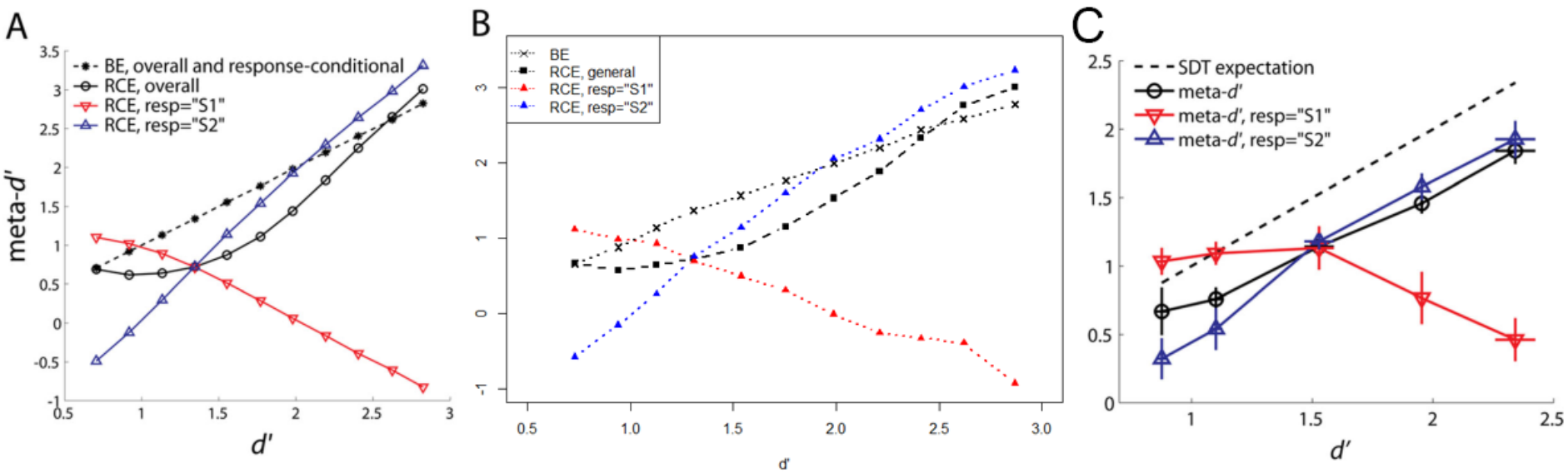


Figura: A) Predicciones hechas en el paper usando los modelos descriptos previamente. B) Simulaciones hechas por nosotros. C) Resultados del 2AFC de los sujetos de prueba.

Experimento 1: Simulaciones

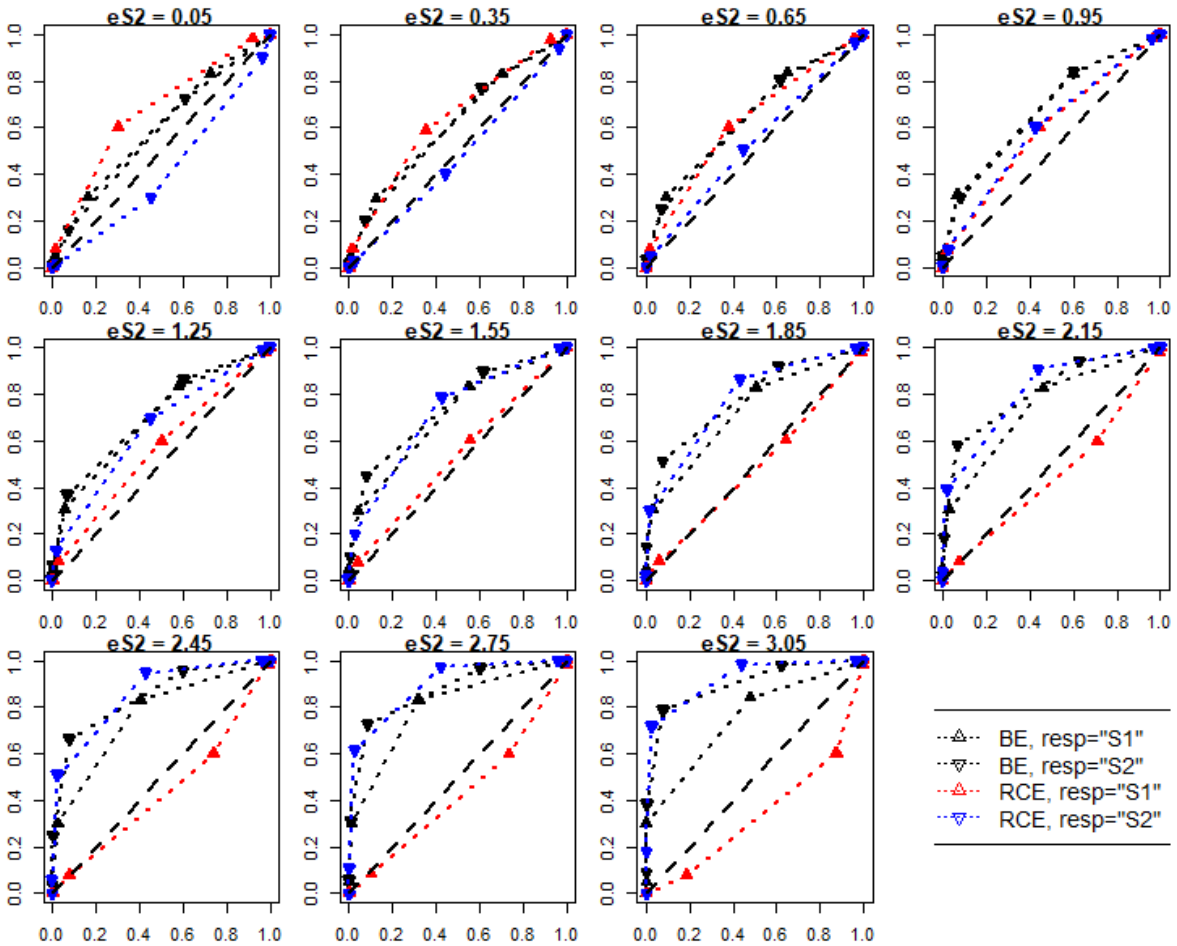
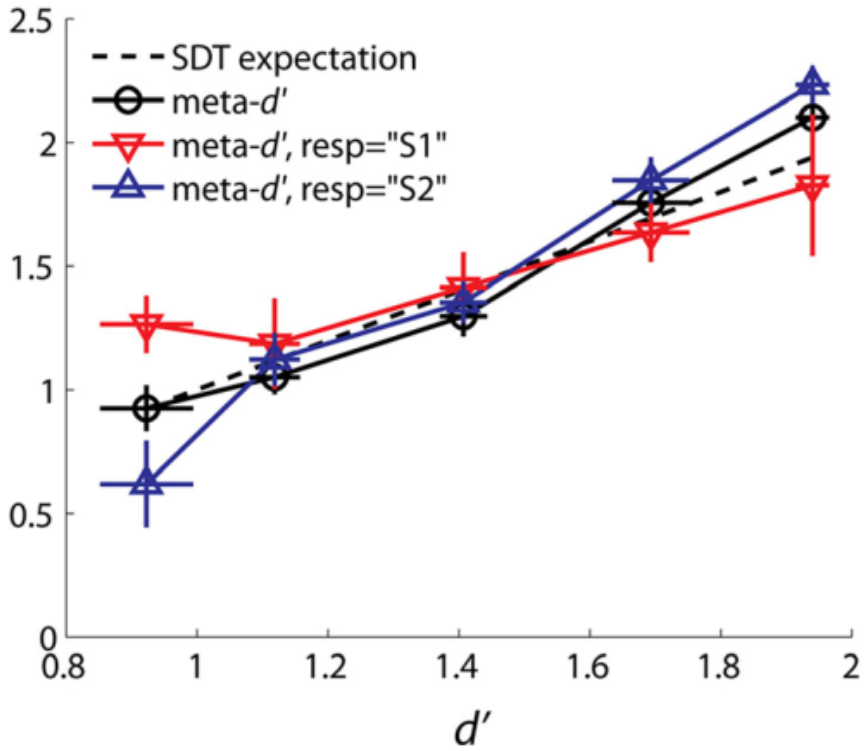


Figura: Las curvas ROC tipo 2 nos indican que es un tema de *sensibilidad metacognitiva* y no de *sesgo metacognitivo* (realizado a partir de nuestras simulaciones)

Experimento 2

- Vimos que los observadores utilizan la regla RCE para evaluar la confianza.
- Resta observar si es posible manipularlos de manera tal que utilicen la regla BE y así mejoren su sensibilidad metacognitiva.
- Por lo tanto, se agregó feedback según el rendimiento y por otro lado se apostaban puntos en lugar de puntuar la confianza.

Experimento 2: Resultados



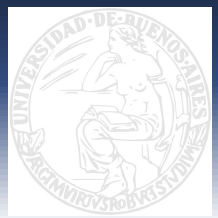
- Utilizando sistemas de feedback, se logró que los participantes adoptaran la regla BE para medir la confianza. Esto implica que no es una característica inherentemente biológica, sino una heurística adoptada por el mecanismo de toma de decisiones.

Conclusiones

1. En base a la experimentación hecha, los datos sugieren que la regla RCE refleja mejor el comportamiento humano en lo que respecta a metacognición.
2. Lo más interesante y contraintuitivo es ver cómo disminuye la sensibilidad metacognitiva en función del rendimiento de la tarea, ya que una correlación positiva entre el rendimiento de la tarea y la sensibilidad metacognitiva es un fenómeno empírico muy robusto y es una consecuencia teórica directa de SDT.
3. La regla BE sigue siendo la utilizada para elegir una respuesta.
4. Se puede alterar el comportamiento metacognitivo de los sujetos y que este se ajuste a la regla BE utilizando manipulaciones/feedback de rendimiento.

Críticas

1. Utilizaron pocos sujetos de prueba para los experimentos ($n = 3$ en el primero, $n = 4$ en el segundo). En toda experimentación con sujetos de prueba es difícil conseguir una muestra adecuada, y si hay pocos sujetos esto se agrava.
2. En el experimento 2, hay un detalle técnico que permite elegir una estrategia óptima para obtener la mayor cantidad de puntos, lo cual podría perjudicar al objetivo de la experimentación. No obstante, les dijeron a los sujetos que usen todos los niveles de confianza y los datos reflejaron que no utilizaron la estrategia óptima.



Discusión

1. Una pregunta que surge de esto es *por qué* utilizamos una estrategia que no es óptima (desde el punto de vista Bayesiano) para medir nuestra confianza.
2. Una posibilidad es que, en el contexto del mundo real, debemos decidir entre muchas alternativas posibles: una vez que decidiste que el estímulo visual es un gato (y no un perro, conejo o auto), no te interesa qué tanto se asemeja a un perro o a un auto, sino qué tanta evidencia tenés a favor de que sea un gato.
3. Más aún, en la mayoría de los casos, lo más probable es que tengamos poca información sobre qué *no* es el estímulo. Intentar mantener estimaciones sobre la calidad de evidencia de *todas* las posibilidades *no elegidas* sería muy desgastante.