

Metodología del test

Cómo se construyen las preguntas, cómo se elige cada una, cómo se calcula el CI estimado y cómo leer el resultado.

1. Resumen

Cognitia es un test en línea de razonamiento que estima la capacidad cognitiva general y la expresa en la escala de CI (media 100, desviación estándar 15). Usa cuatro tipos de pregunta tomados del modelo de **ICAR** (International Cognitive Ability Resource; Condon y Revelle, 2014): matrices, series de letras y números, rotación tridimensional y razonamiento verbal. **Todas las preguntas son originales**: se generan por procedimientos con dificultad controlada o se redactaron para este proyecto; no se copió ningún reactivo de ICAR, Raven, Mensa ni de otro test publicado.

La puntuación usa **Teoría de Respuesta al Ítem** con el modelo logístico de dos parámetros (2PL). El test es **adaptativo** (CAT): empieza con preguntas de dificultad media y elige cada siguiente pregunta según las respuestas anteriores, con balanceo entre los cuatro tipos y control de exposición. La habilidad se estima por **EAP** (esperanza a posteriori con prior normal) y el resultado se reporta con su **intervalo del 90%**.

Cómo leer el resultado. (1) Es una estimación, no un diagnóstico clínico; no sirve para fines escolares, legales ni de contratación. (2) La fiabilidad esperable de un test en línea como este es $\approx 0,85$, frente a $\approx 0,95$ de una evaluación clínica individual.

2. Qué mide y qué no

Los cuatro tipos de pregunta cargan en el factor general de capacidad cognitiva (g) y en habilidades de razonamiento fluido, visoespacial y verbal. El test **no** mide conocimientos escolares, creatividad, inteligencia emocional ni rasgos de personalidad, y no está diseñado para detectar discapacidad intelectual, altas capacidades con fines formales, ni deterioro cognitivo. El desempeño puede verse afectado por cansancio, distracciones, el dispositivo, el dominio del idioma y la práctica previa con problemas parecidos.

3. Tipos de pregunta y generación original

El banco actual tiene **289 preguntas**: 72 matrices, 72 series, 66 de rotación 3D y 79 verbales (30 analogías, 21 de ordenamiento lógico, 20 silogismos y 8 condicionales). Cada pregunta se valida automáticamente al construir el banco.

3.1 Matrices (razonamiento inductivo no verbal)



Cuadrícula de 3×3 con la última celda vacía. Cada celda tiene cinco atributos (forma, cantidad, relleno, tamaño y orientación de una barra de fondo). Cada atributo sigue una regla: *constante*, *constante por fila*, *progresión* o *distribución de tres valores* (cada fila contiene los tres valores en distinto orden), la taxonomía clásica de Carpenter, Just y Shell (1990).

Dificultad controlada por: número de atributos con regla no constante (1 a 4), tipo de regla (la distribución es la más difícil) y similitud de los distractores.

Distractores de tamaño y escala de presentación (versión 2 del banco). Una diferencia *solo* de tamaño entre la respuesta y un distractor únicamente se puede juzgar si se ve a la misma escala que la matriz. Desde la versión 2: (a) el sitio dibuja las opciones exactamente a la escala de las celdas (misma medida en píxeles y mismo grosor de trazo) con la matriz siempre visible, y todas las opciones caben en pantalla sin desplazarse; (b) el validador rechaza cualquier matriz en la que la respuesta y un distractor difieran solo en tamaño con una razón menor que 1,6× (los tamaños del generador son 9, 13,5 y 18: solo 9 frente a 18 = 2× es admisible).

Opciones (8) con diseño factorial 2^3 : se eligen tres atributos y cada opción combina el valor correcto o uno alternativo en cada uno. Así, el valor correcto de cada atributo aparece en exactamente la mitad de las opciones y no se puede acertar eligiendo «la opción más parecida a las demás». **Validación:** un verificador independiente prueba, para cada atributo, todas las reglas por fila y por columna sobre las 8 celdas visibles; todas las reglas que encajan deben predecir el mismo valor, y exactamente una opción debe coincidir con la predicción en todos los atributos. Las matrices que no pasan se descartan.

3.2 Series de letras y números

Se muestran 6 a 10 términos (por ejemplo 12 14 16 18 20 22 24 ?) y se pide el siguiente, con 6 opciones. Familias de reglas con «profundidad» creciente: aritmética; geométrica; letras con salto constante o hacia atrás; dos series intercaladas; diferencias crecientes (segundo orden); tipo Fibonacci; operaciones alternadas (+a, ×b); diferencias geométricas; ciclos de tres pasos; pares letra-número. Los distractores son errores típicos (aplicar la regla un paso de más o de menos, continuar la otra serie intercalada, etc.). Se usa el alfabeto latino de 26 letras (sin Ñ) en los tres idiomas. **Validación:** un conjunto de resolvers (aritmética, geométrica, diferencias de orden 2 y 3, Fibonacci, diferencias periódicas, intercaladas, operaciones alternadas) se aplica a los términos visibles; si alguna regla alternativa encaja y predice un valor distinto, la serie se descarta por ambigua.

3.3 Rotación 3D



Objetos de 4 a 7 cubos (policubos tipo Shepard-Metzler, 1971) dibujados en proyección axonométrica. La tarea es reconocer cuál de 6 opciones es el mismo objeto girado. Los distractores son la imagen especular girada (no puede obtenerse por rotación) o el objeto con un cubo cambiado de lugar. **Dificultad** por número de cubos, ángulo y eje de la rotación (90° sobre el eje vertical es lo más fácil; giros compuestos de 120° o 180° sobre ejes no verticales, lo más difícil) y proporción de distractores especulares. **Validación:** el objeto debe ser quirál; se comprueba con las 24 rotaciones del cubo que exactamente una opción es equivalente al modelo y que las 6 imágenes son distintas; además se rasteriza cada dibujo y se exige que todo cubo conserve al menos el 30% del área visible de un cubo aislado (sin cubos ocultos). El orden de pintado se verificó contra un z-buffer exacto (0 discrepancias en 158 mil muestras).

3.4 Razonamiento verbal (español, inglés y portugués)

- **Ordenamiento lógico** (procedural): 3 a 5 personas y premisas del tipo «A nació antes que B»; la dificultad sube con más personas, premisas desordenadas, comparativos invertidos («después que») y preguntas por posiciones intermedias. Algunos ítems no tienen información suficiente y la respuesta correcta es «No se puede saber».
- **Silogismos y condicionales** con pseudopalabras («Todos los blicos son trefos...») para no depender de conocimiento previo: formas válidas (Barbara, Celarent, Darii, Ferio, Camestres, Baroco, modus ponens, modus tollens) e inválidas (término medio no distribuido, afirmación del consecuente, negación del antecedente, etc.). Las formas inválidas son más difíciles porque inducen conclusiones «creíbles».
- **Analogías** redactadas a mano con vocabulario común y traducidas cuidando que la relación funcione igual en los tres idiomas.

Las preguntas de una misma forma lógica están agrupadas en «familias» y el test no presenta dos de la misma familia.

4. Modelo de medición: TRI 2PL

La probabilidad de que una persona con habilidad θ responda bien la pregunta j es

$$P_j(\theta) = 1 / (1 + e^{-a_j(\theta - b_j)})$$

donde b_j es la dificultad y a_j la discriminación (Birnbbaum, 1968). La información que aporta la pregunta es $I_j(\theta) = a_j^2 P_j(\theta)(1 - P_j(\theta))$.

Parámetros iniciales. b se derivó de la complejidad del generador: la complejidad de cada pregunta (número y tipo de reglas, profundidad de la serie, ángulo de rotación, tipo de silogismo, etc.) se escaló linealmente dentro de cada tipo al rango $[-2,4; +2,4]$. a se fijó entre 1,0 y 1,5 según el tipo (matrices $\approx 1,35$, series $\approx 1,3$, rotación $\approx 1,2$, verbal $\approx 1,1$). Después se re-estiman con las respuestas reales (sección 9).

El modelo 2PL no tiene parámetro de adivinación. Con 6 a 8 opciones, quien responde al azar acierta 12–17% y obtiene un CI muy bajo; el sitio marca además las respuestas anormalmente rápidas (sección 7).

5. Test adaptativo (CAT)

1. **Inicio:** θ inicial aleatorio en $[-0,3; +0,3]$ (cerca de la media).
2. **Balanceo de contenido:** se elige el tipo con mayor déficit respecto al 25% objetivo, de modo que los cuatro tipos quedan equilibrados.
3. **Máxima información:** dentro del tipo se ordenan las preguntas por información de Fisher en el θ actual.
4. **Control de exposición** («randomesque», Kingsbury y Zara, 1989): se sortea entre las k más informativas, con $k = \max(8, 10 - n)$. Así las primeras preguntas varían más y ninguna se vuelve demasiado conocida.
5. **Regla de paro:** el test termina cuando el error estándar es menor que 0,32 (con al menos 20 preguntas y 5 por tipo), al llegar a 30 preguntas, o al agotarse el tiempo (25 min).

6. Estimación y escala de resultados

EAP (Bock y Mislevy, 1982): $\hat{\theta}$ es la media de la distribución posterior de θ con prior $N(0, 1)$, calculada por cuadratura en 81 nodos entre -4 y 4 ; el error estándar (EE) es la desviación estándar de esa posterior. EAP es estable incluso con pocas respuestas o con todas correctas/incorrectas.

$$CI = 100 + 15 \cdot (\hat{\theta} - \mu) / \sigma \quad \cdot \quad \text{percentil} = \Phi((\hat{\theta} - \mu) / \sigma) \quad \cdot \quad \text{intervalo } 90\% = CI \pm 1,645 \cdot 15 \cdot EE / \sigma$$

donde μ y σ son la media y la desviación de θ en la población de referencia. La escala de referencia usa $\mu = 0$ y $\sigma = 1$, y el procedimiento de la sección 9 los ajusta por idioma con los datos reales. Con $EE \approx 0,32$ el intervalo del 90% mide unos ± 8 puntos, p. ej. «CI estimado 118 (rango 110–126)». Fuera de 55–145 se reporta solo el límite («< 55», «> 145»), porque en los extremos la norma no es fiable.

Sub-puntajes por tipo: se re-estima θ con EAP usando solo las preguntas de cada tipo (5 a 8 por tipo), por lo que su intervalo es mucho más ancho. Se muestran como «orientativos».

Normas por edad

Como en las escalas Wechsler, el número principal compara a la persona con su grupo de edad. La edad es obligatoria (por bandas de 5 años, de 16 a 70+; solo se guarda la banda). θ es la habilidad en una escala común; el CI por edad es

$$CI_{edad} = 100 + 15 \cdot (\hat{\theta} - \mu_{edad}) / \sigma_{edad} \quad \cdot \quad \text{intervalo } 90\% = CI_{edad} \pm 1,645 \cdot 15 \cdot EE / \sigma_{edad}$$

y es el que se usa para la clasificación descriptiva, el percentil «frente a personas de tu edad», la curva y el certificado. Como dato secundario se muestra el CI frente a la población adulta general (fórmula de arriba). Las sesiones anteriores, sin banda de edad, conservan solo la norma general.

Valores por edad. μ_{edad} se fijó a partir de las tendencias publicadas para razonamiento fluido y pruebas tipo matrices: el rendimiento sube hasta los primeros 20, se mantiene hasta cerca de los 30 y después baja de forma gradual, del orden de 0,02 desviaciones estándar por año en estudios transversales, y algo más rápido después de los 60. Modelo usado: declive = 0,02 DE/año después de los 27 años + 0,01 DE/año adicional después de los 60; la banda 16–19 queda 0,10 DE por debajo de la meseta 20–29; los valores se centran para que la media ponderada de la población adulta (pesos aproximados de la población de 16 años o más de México, Censo 2020 del INEGI) sea 0, de modo que la norma general y las de edad comparten escala. $\sigma_{edad} = 1$ en todas las bandas (no hay una fuente que justifique otra cosa para este test).

Edad	$\mu_{edad} (\theta)$	σ_{edad}	CI por edad con $\theta = 0$
16–19	0,25	1,0	96
20–24	0,35	1,0	95
25–29	0,35	1,0	95
30–34	0,25	1,0	96
35–39	0,15	1,0	98
40–44	0,05	1,0	99
45–49	-0,05	1,0	101
50–54	-0,15	1,0	102
55–59	-0,25	1,0	104
60–64	-0,37	1,0	106
65–69	-0,52	1,0	108
70+	-0,76	1,0	111

Fuentes de la forma de la curva:

- Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, 30(4), 507-514.
- Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, 16(5), 754-760.

- Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? *Psychological Science*, 26(4), 433-443.
- Schaie, K. W. (2005). *Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study*. Oxford University Press.
- Wechsler, D. (2008). *WAIS-IV Technical and Interpretive Manual* (normas de Matrices por grupo de edad). Pearson.

Clasificación descriptiva

Junto al número se muestra una categoría convencional (descriptores del tipo de las escalas Wechsler), calculada en el servidor sobre el CI estimado redondeado. Es **descriptiva**, no **diagnóstica**.

CI estimado	Español	English	Português
≥ 130	Muy superior	Very superior	Muito superior
120-129	Superior	Superior	Superior
110-119	Promedio alto	High average	Média superior
90-109	Promedio	Average	Média
80-89	Promedio bajo	Low average	Média inferior
70-79	Limítrofe	Borderline	Limítrofe
< 70	Muy bajo	Very low	Muito baixo

Regla del tramo. Una sola etiqueta sugeriría una precisión que el test no tiene. Por eso se clasifican también los dos extremos del intervalo del 90%: si caen en categorías distintas, se informa el tramo completo, de la categoría del extremo inferior a la del superior (p. ej. CI 115, rango 107-123 → «Promedio a Superior»), y se indica aparte en qué categoría cae la estimación puntual. Con un EE típico el intervalo mide unos ± 8 puntos, así que lo habitual es que el resultado abarque dos categorías. La misma clasificación aparece en la página de resultados (tabla y límites marcados en la curva) y en el certificado.

7. Tiempo y validez de las respuestas

Hay un tiempo total de 25 minutos (no por pregunta), visible y honesto: no se puede pausar, sigue corriendo si se cierra la página y el test puede reanudarse mientras quede tiempo. Cada respuesta guarda el tiempo medido en el servidor y el reportado por el navegador. Se marcan como posibles resultados no válidos: (a) $\geq 30\%$ de respuestas más rápidas que un umbral por tipo (Matrices 5 s, Series 4 s, Rotación 3D 5 s, Verbal 5 s), (b) test completo en menos de 4 minutos y (c) más del 60% de las respuestas en la misma posición. Las sesiones marcadas no entran a las normas.

8. Simulación: ¿recupera el CAT la habilidad?

Se simularon respondientes con θ verdadero conocido, $\theta \sim N(0, 1)$, que contestan según el modelo 2PL, y se pasaron por exactamente el mismo algoritmo del sitio. En las dos últimas filas los parámetros «reales» con que responden los simulados difieren de los priors que usa el test ($b_{\text{real}} = b + N(0, \sigma^2)$, $a_{\text{real}} = a \cdot e^{N(0, \sigma^2/4)}$), para medir qué pasa si los priors están equivocados.

Escenario	n	$r(\theta, \hat{\theta})$	RMSE	Sesgo	EE medio	Preguntas	Cobertura IC 90%
Parámetros correctos	5000	0,949	0,318 (4,8 pts)	0,006	0,317	24,9	89,7%
Parámetros reales $\neq$ priors ($\sigma = 0,5$)	5000	0,943	0,336 (5,0 pts)	0,031	0,317	24,8	88,3%
Parámetros reales $\neq$ priors ($\sigma = 1,0$)	5000	0,933	0,370 (5,5 pts)	0,062	0,317	24,7	84,0%

Escenario «parámetros correctos», por banda de θ verdadero:

θ verdadero	n	Sesgo	RMSE	EE medio	Preguntas
[-9, -2)	127	+0,225	0,369	0,326	28,3
[-2, -1)	686	+0,140	0,335	0,317	25,6
[-1, +0)	1671	+0,043	0,310	0,317	24,5
[+0, +1)	1715	-0,034	0,309	0,317	24,1
[+1, +2)	696	-0,125	0,330	0,317	25,6
[+2, +9)	105	-0,186	0,334	0,323	28,8

Sub-puntaje	r con θ verdadero	EE medio
Matrices	0,838	0,55 (± 14 pts)
Series	0,824	0,57 (± 14 pts)
Rotación 3D	0,820	0,58 (± 14 pts)
Verbal	0,793	0,62 (± 15 pts)

Lectura. Si el modelo es correcto, el CAT recupera θ con correlación $\approx 0,95$, error típico de unos 4,8 puntos de CI, ≈ 25 preguntas y la cobertura del intervalo del 90% es la nominal. En los extremos ($|\theta| > 2$) hay encogimiento hacia la media, propio de EAP, y menos preguntas disponibles. Si los priors están moderadamente equivocados ($\sigma = 0,5$) la precisión casi no cambia; si están muy equivocados ($\sigma = 1$) el error crece y el intervalo del 90% sólo cubre $\approx 84\%$ de las veces: por eso los parámetros se re-estiman con datos reales (sección 9). La exposición máxima de una pregunta fue 37%.

Estas cifras muestran que el algoritmo funciona **si** las preguntas se comportan como dice el modelo. No demuestran validez: que el resultado se relacione con otras medidas de capacidad cognitiva sólo puede establecerse con datos de personas reales.

9. Datos y actualización de parámetros

El sitio guarda cada sesión (sin correo ni IP): nombre de la persona (para el certificado), idioma, banda de edad, preguntas presentadas, respuesta, acierto y tiempos. Con esos datos se hace periódicamente lo siguiente:

1. Estima de nuevo a y b de cada pregunta por **máxima verosimilitud marginal con EM** (Bock y Aitkin, 1981), con priors débiles ($a \sim N(1,2; 0,5^2)$, intercepto $\sim N(\text{prior}; 1,5^2)$). Los datos de un CAT son «faltantes al azar» condicionados a lo observado, por lo que la estimación marginal sobre la matriz dispersa es válida. Preguntas con menos de 30 respuestas conservan su prior.
2. Re-estima θ de cada sesión válida con los nuevos parámetros y calcula media y desviación por idioma (normas). La escala se fija como $CI = 100 + 15 (\theta - \mu_{\text{idioma}}) / \sigma_{\text{idioma}}$.

Prueba con datos simulados (4,000 sesiones, priors con error $\sigma = 0,5$): la dificultad recuperada correlaciona 0,991 con la verdadera (RMSE 0,184), frente a 0,933 (RMSE 0,502) de los priors; la discriminación es más difícil de recuperar ($r = 0,714$), como es habitual con muestras dispersas.

10. Fiabilidad: en línea vs. clínica

Las pruebas clínicas individuales (aplicadas por un psicólogo, con normas nacionales representativas) alcanzan fiabilidades del CI total cercanas o superiores a 0,95. Un test en línea autoadministrado, más breve y con condiciones no controladas, se ubica típicamente alrededor de 0,85. En términos prácticos: con fiabilidad 0,85 el error estándar de medida es $\approx 5,8$ puntos de CI; con 0,95, $\approx 3,4$ puntos. El sitio lo explica junto al resultado y no presenta el número como exacto.

11. Nota legal y de licencias

- ICAR es un recurso de dominio público pensado para investigación; sus reactivos se distribuyen sin costo para ese fin (algunos conjuntos se solicitan a los autores). **Este producto no usa ningún reactivo de ICAR:** reutiliza el método publicado y los tipos de pregunta, con reactivos originales generados o redactados para el proyecto. Las ideas y métodos científicos no son objeto de derechos de autor, pero los reactivos concretos sí podrían serlo.
- No se usan reactivos ni marcas de Raven, Mensa, Wechsler u otras pruebas comerciales.

12. Referencias

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F. M. Lord y M. R. Novick, *Statistical theories of mental test scores*. Addison-Wesley.
- Bock, R. D., y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D., y Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.
- Carpenter, P. A., Just, M. A., y Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404–431.
- Condon, D. M., y Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, 43, 52–64.
- Embretson, S. E., y Reise, S. P. (2000). *Item response theory for psychologists*. Erlbaum.
- Kingsbury, G. G., y Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359–375.
- Shepard, R. N., y Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701–703.
- van der Linden, W. J., y Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. Springer.
- Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer* (2a ed.). Erlbaum.