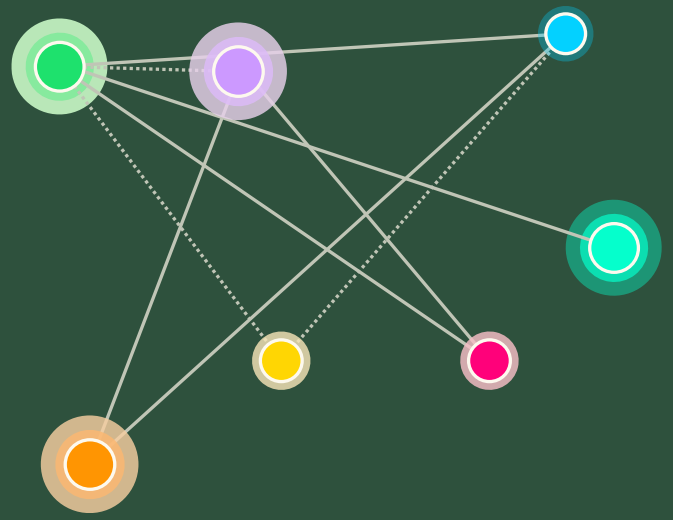
Universidad Señor de Sipán

VII Congreso Internacional de Ciencia, Tecnología e Innovación USS 2026

"Inteligencia artificial para la investigación y la innovación responsable"



SCImago

CORPORACIÓN UNIVERSITARIA REFORMADA

VII Congreso Internacional
de Ciencia, Tecnología e
Innovación USS 2026

"Inteligencia artificial para la investigación y
la innovación responsable"

Llévate la presentación

Apunta tu cámara al código

Ahí están estas mismas láminas en PDF, para seguirlas desde tu celular, descargarlas y consultar las referencias.

joseventuraleon.com/diapositivas/ia-psicometria-desafios-eticos

USS Vicerrectorado de Investigación

VII Congreso Internacional de Ciencia, Tecnología e Innovación USS 2026

"Inteligencia artificial para la investigación y la innovación responsable"

Inteligencia artificial y psicometría: desafíos éticos para la validez, los sesgos y la integridad de la medición

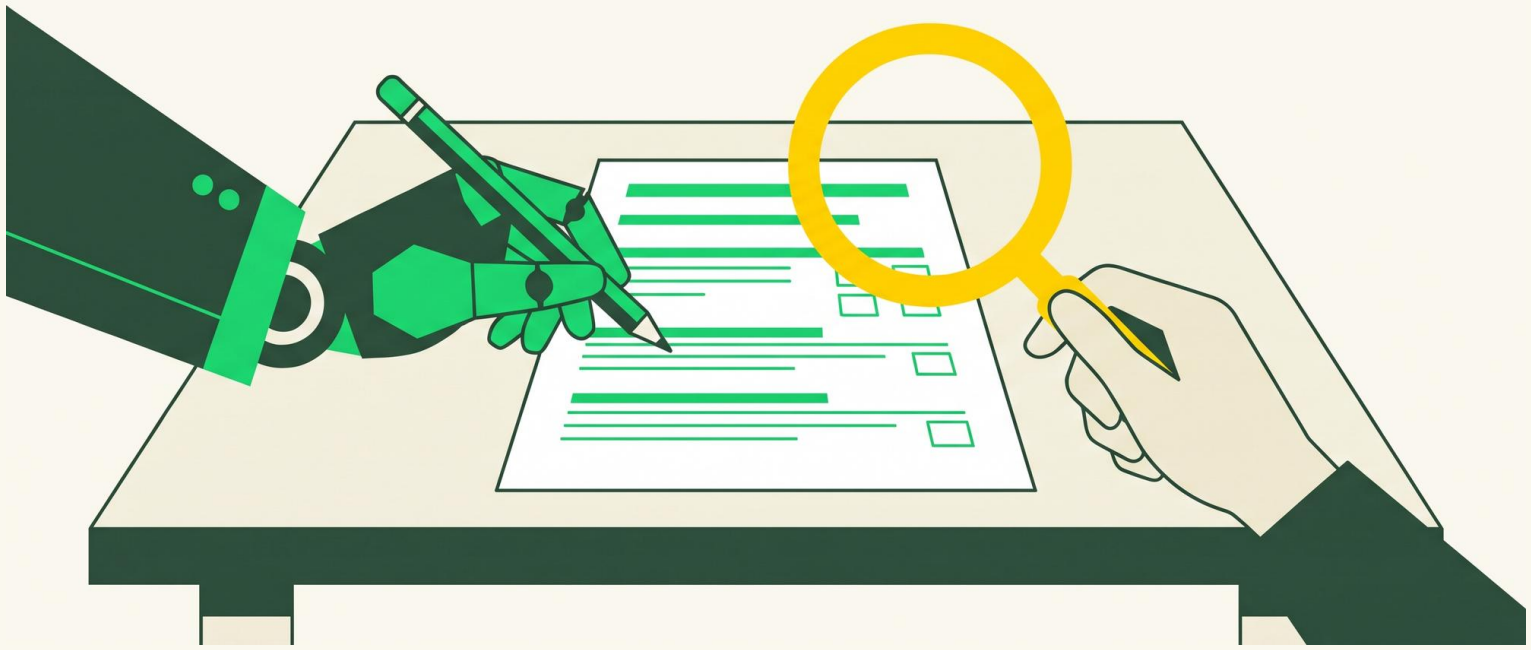
Dr. José Ventura-León

Universidad Privada del Norte, Perú

Ponencia magistral · miércoles 25 de noviembre de 2026

USS Vicerrectorado de Investigación

¿Puede una máquina escribir un test... y además juzgarlo?



Hoy ya hace ambas cosas: redacta ítems, los califica como un panel de jueces y anticipa su estructura.

La pregunta ética es otra: ¿qué evidencia hace falta antes de creerle?

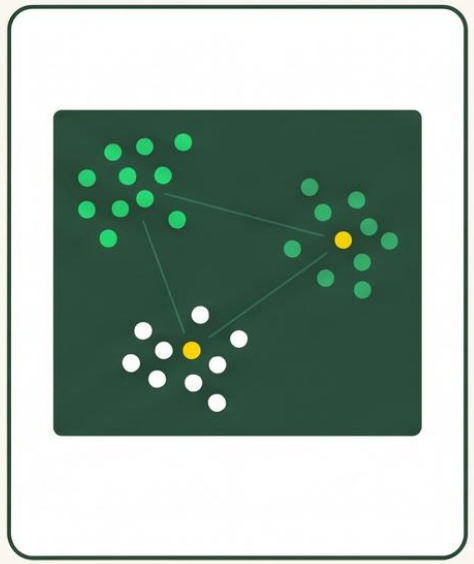
La inteligencia artificial ya entró a la psicometría



GENERAR ÍTEMS



JUZGAR EL CONTENIDO



MAPEAR SIGNIFICADOS



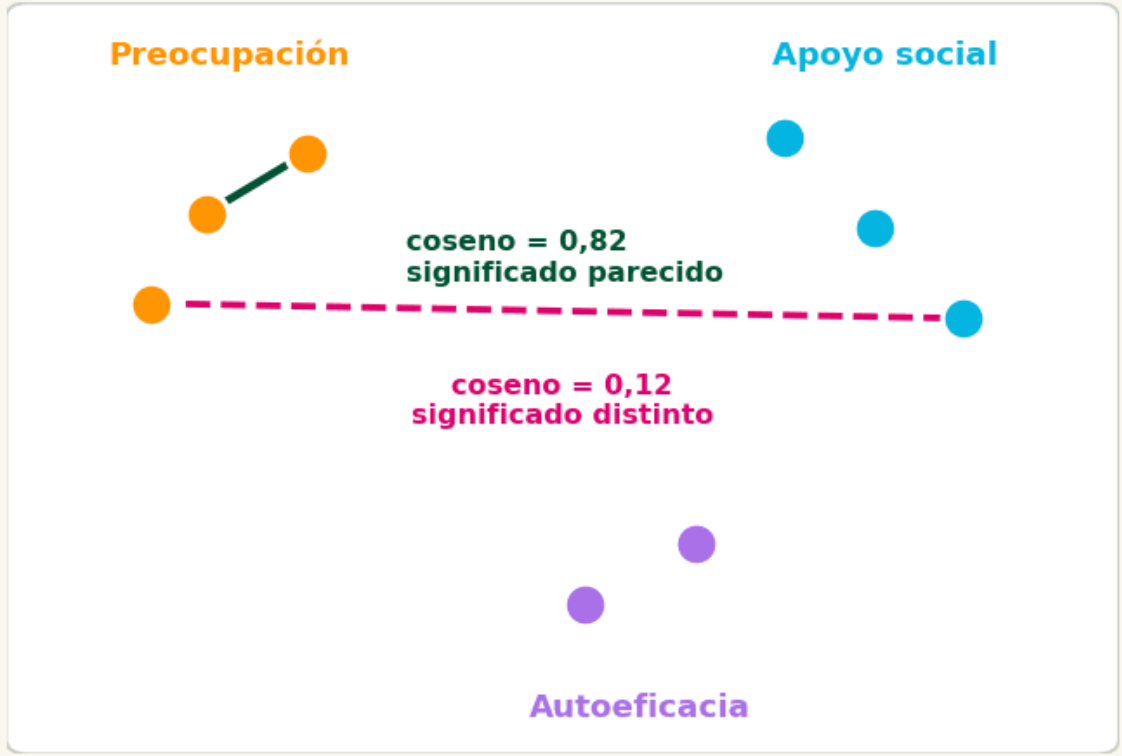
ANTICIPAR LA ESTRUCTURA

Generar (Russell-Lasalandra et al., 2026) · Juzgar (Milano et al., 2026; Poźniak et al., 2026) · Mapear (Hommel & Arslan, 2025; Wulff & Mata, 2025) · Anticipar (Voss et al., 2026)

Un embedding convierte cada ítem en un punto de un mapa de significados

- Me preocupo antes de un examen
- Me cuesta dormir si tengo una evaluación
- Pienso que voy a fallar en la prueba
- Mis amigos me ayudan cuando lo necesito
- Tengo a quién contarle mis problemas
- Mi familia me escucha
- Puedo resolver problemas difíciles
- Confío en mi capacidad para aprender

Cerca en el mapa = significado parecido.
La IA mide la redacción, todavía no a las personas.



Valores ilustrativos

Un modelo de lenguaje predijo, solo desde el texto, las correlaciones entre ítems (precisión fuera de muestra de 0,59) y la confiabilidad de las escalas (0,84) en un estudio prerregistrado (Hommel & Arslan, 2025).

VII Congreso Internacional de Ciencia, Tecnología e Innovación USS 2026

"Inteligencia artificial para la investigación y la innovación responsable"

SeMiLLa: la herramienta detrás de esta charla



SeMiLLa

SEmantic Measurement of Items via LLM Assistance

(Ventura-León, 2026a, 2026b) · github.com/jventural/SeMiLLa

REDACTA
ítems a partir de un constructo

AUDITA
redundancia, contenido y estructura prevista

ADVIERTE
antes de gastar una muestra

Por qué la construí

Generar, analizar y adaptar ítems con IA se hacía en flujos dispersos. Faltaba uno abierto, reproducible y en español.

Qué es

Un paquete de R de código abierto (licencia MIT), con una aplicación web.

Qué no es

Un reemplazo de las personas que responden.

Léelo completo
Preprint en PsyArXiv
osf.io/s3nkm

USS Vicerrectorado de Investigación

SeMiLLa por dentro: seis pasos y una frontera

Versión 2.12.6: los seis primeros círculos trabajan con el texto; el séptimo, con personas



Su propia documentación lo advierte: **estas pruebas describen la redacción, no la conducta de personas reales, y no sustituyen un estudio de campo** (Ventura-León, 2026b).

En su preprint (versión 1.1), asignó el 74 % de los ítems al mismo factor que los datos reales (ARI medio = 0,37); en la Escala de Celos, en cambio, la coherencia semántica y la carga factorial se relacionaron al revés, r = −0,73 (Ventura-León, 2026a, p. 1).

DESAFÍO 1

Validez

¿Mide lo que dice medir… o solo lo que se le pidió optimizar?

Universidad Señor de Sipán

La validez se argumenta con evidencia

Lo que la IA puede aportar sin datos

- Un borrador de evidencia de **contenido** (jueces simulados).
- Una **estructura prevista** a partir del significado de los ítems.
- Alertas de redundancia, ambigüedad y deseabilidad.

Son hipótesis sobre la redacción.

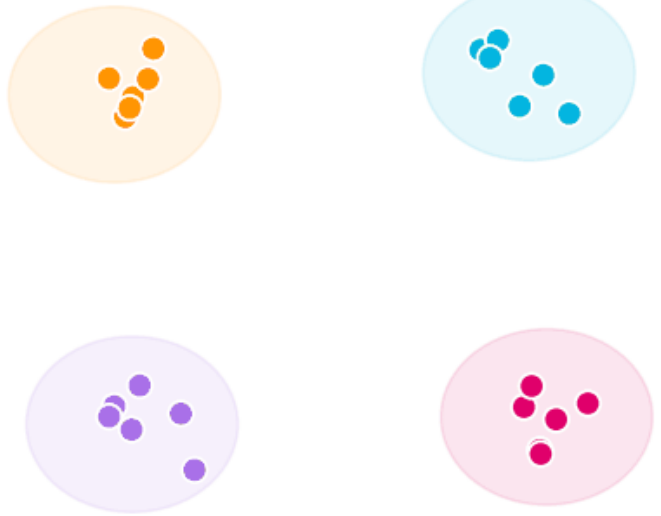
Lo que solo aportan las personas

- Los **procesos de respuesta** (cómo entienden el ítem).
- La **estructura interna** observada.
- Las **relaciones con otras variables**.
- Las **consecuencias** del uso de las puntuaciones.

Fuentes de evidencia de validez (American Educational Research Association et al., 2014; Messick, 1995).

Refinar con IA: la estructura mejora… y el constructo se vacía

Tres ciclos de refinamiento (trayectoria ilustrativa)



16 de 24 ítems reescritos por el modelo

Precisión de la estructura	**66,7 → 95,8 %** ▲
Pares de ítems redundantes	**7 → 14** ▲
Ítems con el molde "Cuando…"	**4 → 15 de 24** ▲

Facetas de la definición del constructo

✗	No juzgar / aceptación	**4 → 0 ítems**
✗	Cuidar el cuerpo: dormir y comer	**1 → 0 ítems**

**La precisión subió porque los ítems se parecen más entre sí.
El algoritmo cumplió la meta; la meta no era todo lo que importa.**

Caso real: escala de regulación emocional para adolescentes (24 ítems, cuatro dimensiones), tres ciclos de refinamiento automático.
Regla que adoptamos: tras optimizar, volver a auditar la redundancia y **leer los ítems**.

Un panel de jueces simulados no es un panel de expertos

Un solo modelo, muchas opiniones

En SeMiLLa, un solo modelo responde en una consulta "como si" fueran diez jueces. Las valoraciones no son independientes: es pseudorreplicación.

La V de Aiken describe al modelo

El índice y su intervalo se calculan igual, pero informan sobre el modelo, no sobre un consenso de expertos.

Útiles como complemento

Algunos LLM detectan ítems claramente irrelevantes con alto acuerdo interno. La literatura los propone como complemento del juicio humano.

Ética mínima: declarar el modelo, el prompt y la no independencia como limitación.

(Aiken, 1985; Mara & Rybak, 2026; Milano et al., 2026; Penfield & Giacobbi, 2004; Poźniak et al., 2026; Ventura-León, 2026a, p. 4; Ventura-León et al., 2025, p. 9)

¿Acierta la compuerta previa al campo? Nueve escalas publicadas

Escala	DASS-21	RSES	BigFive	ECR	NPI	SD3	DD	CFCS	RIASEC
Compuerta (solo texto)	Aplicar p = 1,00	Revisar p = 0,00	Revisar p = 0,00	Revisar p = 0,00	Aplicar p = 0,93	Revisar p = 0,03	Revisar p = 0,77	Revisar p = 0,00	Aplicar p = 1,00
Campo (n = 1 500)	Funciona	No ajusta	No ajusta	No ajusta	Funciona	Funciona	No ajusta	Funciona	No ajusta
¿Coinciden?	✓	✓	✓	✓	✓	✗	✓	✗	✗

Acierta 6 de 9 (67 %) frente al 56 % de responder siempre "revisar". Con n = 9, la compuerta orienta; no decide. Dos errores son pesimistas (SD3 y CFCS) y uno optimista (RIASEC, RMSEA 0,081).

Criterio: CFI ≥ 0,95 y RMSEA ≤ 0,08 en un AFC con 1 500 respuestas reales por escala (datos: Open-Source Psychometrics Project). SeMiLLa 2.9.13, calibrada con estas mismas escalas: describe, no valida. AUC = 0,65 (0,50 es el azar), con un intervalo muy ancho. El preprint (Ventura-León, 2026c) usa otra versión y otro criterio: allí CFCS cuenta como fusión.

Una premisa no puede volver como resultado

0,296

error medio al estimar la correlación entre factores (ocho escalas)

0,285

error de devolver, sin simular, el supuesto de partida (0,47)

La correlación "esperada" entre dimensiones era, en realidad, el supuesto con que arranca la simulación.

Además, en la escala CFCS los factores correlacionan 0,86 y el modelo de dos factores ajusta bien (CFI = 0,994): un corte fijo no decide por sí solo.

Ética de la comunicación: llamar **supuesto** a lo que es supuesto. En la herramienta se cambió "correlación esperada" por "correlación en el escenario simulado".

El preprint del método lo delimita: la probabilidad es simulada y condicional a los supuestos declarados, no una probabilidad calibrada para una población real (Ventura-León, 2026c, p. 8).

DESAFÍO 2

Sesgos

¿A quién se parece el mapa de significados con que la IA lee nuestros ítems?

Universidad Señor de Sipán

El mapa se aprende de lo que ya escribimos, sesgos incluidos



Los vectores aprendidos de textos de la web reproducen sesgos conocidos, como los de género y etnia, y reflejan el reparto de género en las ocupaciones (Caliskan et al., 2017).

En psicometría: el ítem se redacta y se juzga con el lenguaje mayoritario del corpus, sobre todo en inglés.

Riesgo: ítems que suenan naturales para unos grupos y extraños para otros.

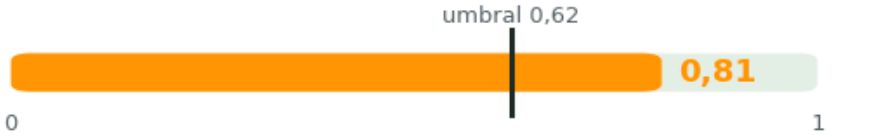
Los datos de entrenamiento pueden reproducir sesgos y excluir a grupos poco representados; por eso la salida de la IA se revisa antes de usarla (Ventura-León et al., 2025, p. 5).

Un coseno alto no garantiza el mismo significado, ni uno bajo lo descarta

1. Marca de más

"Me permito sentir emociones aunque no me gusten"

"Me doy cuenta de las emociones de los demás sin juzgarlas"



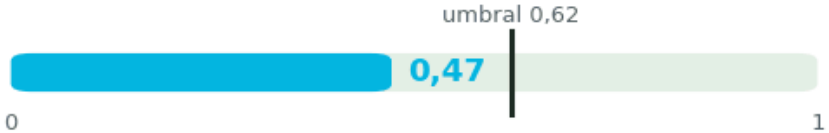
Detector automático: supera el umbral, "redundantes"

✗ **Falso positivo: uno mira hacia adentro y el otro hacia los demás; los une el vocabulario ("emociones", "juzgar"), no el significado.**

2. Deja pasar de menos

"Les pido a mis padres que respeten mi espacio"

"Les pido a mis amigos que no invadan mi privacidad"



Detector automático: bajo el umbral, "distintos"

✗ **Falso negativo: es el mismo ítem con otro destinatario y sinónimos. Con este modelo, solo lo detectó la lectura humana.**

Ítems de una escala real (el segundo par, levemente adaptado). Con un modelo de embeddings más grande el par 2 sube a 0,725 y sí se detecta: el modelo elegido cambia el diagnóstico.

La compuerta detectaba deseabilidad, no constructo

Mismo defecto, distinta deseabilidad

Deseabilidad	Probabilidad de "escala limpia"
La real, según el juez	**0,000**
Constante para todos los ítems	**0,960**

Lo encontramos introduciendo el defecto a propósito, con la verdad conocida. La versión 2.9.36 añadió un control de la asignación de cada ítem.

Al quitar la diferencia de deseabilidad, el veredicto se invirtió.

Funcionaba solo cuando una dimensión era muy deseable y la otra muy poco (brecha de 0,72). En los otros dos estudios la brecha era de 0,07 y 0,08: **ahí no veía nada.**

Un sesgo de respuesta se coló como señal de calidad.

Coincide con el preprint: de seis sesgos de respuesta, la deseabilidad fue el único cuyo punto de quiebre distinguió entre escalas (Ventura-León, 2026c, p. 29). Datos de la tabla: Item Response Warehouse.

Adaptar entre culturas: el modelo no viaja solo

Idioma del entrenamiento

Los modelos aprendieron sobre todo en inglés. El español de Perú no es un caso promedio.

El indicador que no avisa

En una adaptación polaca del PSS-10 al trabajo, dos de diez ítems quedaron más cerca de otro ítem que de su original; el indicador de deriva los marcó "bajo" y "moderado".

Corregir a ojo puede empeorar

Reescribir por intuición movió un ítem de riesgo moderado a alto. Hay que medir cada variante.

La invarianza se prueba con datos

El tamizaje semántico de DIF orienta; la equivalencia se prueba con un AFC multigrupo.

La adaptación asistida por IA entrega un borrador rápido; no reemplaza al panel bilingüe de expertos (Ventura-León, 2026a, p. 9).

DESAFÍO 3

Integridad

¿Podemos confiar en el proceso, y no solo en el resultado?

Universidad Señor de Sipán

Referencias inventadas: un riesgo medido

55 %
de las citas de GPT-3.5 no existían

18 %
de las citas de GPT-4 tampoco existían

Y entre las citas reales, el 43 % y el 24 % tenían errores sustantivos.

636 referencias de 84 revisiones breves escritas por los modelos.

Práctica: cada referencia se verifica contra Crossref u OpenAlex (DOI, año, volumen, páginas) y se lee la fuente antes de citarla.

(Ventura-León et al., 2025, p. 9; Walters & Wilder, 2023)

Si cambia el modelo, cambia el diagnóstico

Los mismos 32 ítems, dos modelos de embeddings

	384 dim.	1 536 dim.
Pares redundantes	**18**	**13**
Similitud máxima	**0,81**	**0,78**
Depurados sobre 0,62 (*)	**5 de 8**	**4 de 8**

Fallas que se leen como resultados

Un juez que no llegó a correr informaba "0 gemelos confirmados".

Un componente que no se cargaba dejaba el análisis con dos de tres algoritmos, **sin avisar**.

Declarar siempre: modelo y versión, fecha, semilla, caché de llamadas, versión del paquete y las fallas encontradas.

(*) Pares que depuramos y superan el umbral; solo 2 coinciden. Tabla: caso propio. Qué declarar: Ventura-León (2026a, pp. 3–4; 2026c, pp. 31–32) y Ventura-León et al. (2025, p. 14).

Seis principios para una psicometría con IA responsable

1	**Humano en el circuito**	Mantener la supervisión humana y el juicio crítico: todo resultado se revisa antes de incorporarlo.	Ventura-León et al. (2025, p. 8)
2	**Declarar modelo y prompt**	Declarar el uso y entregar el prompt con sus resultados como material suplementario.	Ventura-León et al. (2025, pp. 9, 14)
3	**Validar con datos reales**	Un pronóstico optimista indica ausencia de alarma en la redacción, no una garantía de validez.	Ventura-León (2026c, p. 28)
4	**Auditar sesgos**	Revisar la salida de la IA con sentido crítico y comprobar que no contenga sesgos inapropiados.	Ventura-León et al. (2025, p. 5)
5	**Reproducir con semilla y caché**	Semilla fija, temperatura 0 y caché en disco: el mismo resultado sin volver a llamar al modelo.	Ventura-León (2026a, pp. 3–4)
6	**Verificar cada referencia**	Verificar con rigor: la IA puede crear información o referencias falsas que parecen reales.	Ventura-León et al. (2025, p. 9)

Propuesta del ponente, con respaldo en Ventura-León (2026a, 2026c) y Ventura-León et al. (2025).

VII Congreso Internacional de Ciencia, Tecnología e Innovación USS 2026

"Inteligencia artificial para la investigación y la innovación responsable"



La IA acelera el borrador.

La validez se gana en el campo.

La responsabilidad ética no se delega: se documenta.

La decisión final y la responsabilidad por el contenido son intransferibles (Ventura-León et al., 2025, p. 10).

USS Vicerrectorado de Investigación

Referencias (1 de 2)

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Hommel, B. E., & Arslan, R. C. (2025). Language models accurately infer correlations between psychological items and scales from text alone. *Advances in Methods and Practices in Psychological Science, 8*(4), Artículo 25152459251377093. https://doi.org/10.1177/25152459251377093

Mara, C., & Rybak, T. (2026). Leveraging large language models in patient-reported outcome measure development: Practical opportunities, cautions, and a human-in-the-loop roadmap. *Journal of Patient-Reported Outcomes, 10*(1), Artículo 93. https://doi.org/10.1186/s41687-026-01112-2

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milano, N., Ponticorvo, M., & Marocco, D. (2026). Human expertise and large language model embeddings in the content validity assessment of personality tests. *Educational and Psychological Measurement, 86*(1), 30–53. https://doi.org/10.1177/00131644251355485

Penfield, R. D., & Giacobbi, P. R., Jr. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3

Poźniak, S., Kleka, P., & Szwal, A. (2026). *Are LLMs better expert judges? Rethinking content validity assessment in the age of AI* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/4hsyx_v2

Referencias (2 de 2)

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2026). Generative psychometrics via AI-GENIE: Automatic item generation and validation with network-integrated evaluation. *Behavior Research Methods, 58*(8), Artículo 217. https://doi.org/10.3758/s13428-026-03082-1

Ventura-León, J. (2026a). *SeMiLLa: An R package and Shiny app for LLM-assisted construction and pre-empirical semantic screening of psychometric scales* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/s3nkm_v1

Ventura-León, J. (2026b). *SeMiLLa: Drafting and auditing psychometric items with large language models* (Versión 2.12.6) [Paquete de R]. GitHub. https://github.com/jventural/SeMiLLa/tree/compuerta-rapida

Ventura-León, J. (2026c). *The stress test: A mechanistic simulation of response bias and item fragility in psychological scales* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/cn3ue_v1

Ventura-León, J., Gamboa-Melgar, G., Ruiz-Castro, J., Tocto-Muñoz, S., Sánchez-Villena, A. R., & Lino-Cruz, C. (2025). *Use of artificial intelligence in scientific publishing: Good practice guide for authors and institutions* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/bmhu2_v1

Voss, N. M., Wu, F. Y., Javalagi, A. A., & Kell, H. J. (2026). Integrating ensemble clustering and text embeddings for estimating the factor loadings of self-report scales. *Educational and Psychological Measurement, 86*(6), 1223–1252. https://doi.org/10.1177/00131644261430762

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*(1), Artículo 14045. https://doi.org/10.1038/s41598-023-41032-5

Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour, 9*(5), 944–954. https://doi.org/10.1038/s41562-024-02089-y

Declaración de uso de IA: las ilustraciones se generaron con Gemini (Google) y fueron revisadas por el ponente, responsable del contenido.

Gracias

Dr. José Ventura-León
jose.ventura@upn.pe
ORCID 0000-0003-2996-4244 · github.com/jventural/SeMiLLa

Universidad Señor de Sipán