PSICODIAGNOSTICAR – VOLUMEN 33: 7/22. Rosario, 2025

EVALUACIÓN COMPARATIVA DE MODELOS DE INTELIGENCIA ARTIFICIAL (IA) GENERATIVA EN EL DISEÑO DE INSTRUMENTOS DE EVALUACIÓN DE LA PERSONALIDAD

Horacio D. García[1], Araceli G. Aguilera[2] y Deborah E. Lucero[3]

RESUMEN

Este estudio evaluó la calidad psicométrica de instrumentos de evaluación de personalidad basados en el modelo de los Cinco Grandes Factores (Big Five), generados por tres modelos de inteligencia artificial: OpenAI O3, Mistral y Deepseek R1. Mediante un análisis con 384 adultos argentinos, se encontró que los tres instrumentos mostraron consistencia interna aceptable o buena, aunque con diferencias notables: mientras OpenAI O3 presentó los mejores indicadores, Deepseek R1 mostró limitaciones en la fiabilidad del factor Amabilidad (α = .624). En cuanto a la estructura factorial, todos los modelos explicaron una varianza razonable (45.2%–48.9%), pero el de OpenAI O3 demostró superioridad con un ajuste excelente (RMSEA = .046, NNFI = .952, CFI = .961) y mejor parsimonia (BIC más bajo), frente a un ajuste apenas aceptable en Mistral y Deepseek R1 (RMSEA = .059/.051; NNFI/CFI < .95). Los resultados evidencian el potencial de la Inteligencia Artificial (IA) para generar instrumentos psicométricos, pero resaltan la necesidad de validación rigurosa y supervisión especializada para asegurar su calidad aplicada.

Palabras clave: *Inteligencia Artificial generativa, Modelo de los Cinco Grandes Factores, evaluación de la personalidad, psicometría*

[1] Facultad de Psicología. Universidad Nacional de San Luis. San Luis, Argentina. Proyecto de investigación PROICO 12–0420. UNSL. E–mail: hdgarcia69@gmail.com.
[2] Facultad de Psicología. Universidad Nacional de San Luis. San Luis, Argentina. Proyecto de investigación PROICO 12–0420. UNSL
[3] Facultad de Psicología. Universidad Nacional de San Luis, San Luis, Argentina. Proyecto de investigación PROICO 12–0420. UNSL

ABSTRACT

This study assessed the psychometric quality of personality assessment instruments based on the Big Five model, generated by three artificial intelligence systems: OpenAI O3, Mistral, and Deepseek R1. Using a sample of 384 argentine adults, analyses revealed that all three instruments demonstrated acceptable to good internal consistency, though with notable differences: while OpenAI O3 outperformed the others, Deepseek R1 showed limitations in reliability for the Agreeableness factor (α = 0.624). Regarding factorial structure, all models accounted for reasonable variance (45.2%–48.9%), but OpenAI O3 exhibited superior fit (RMSEA = 0.046, NNFI = 0.952, CFI = 0.961) and better parsimony (lowest BIC), whereas Mistral and Deepseek R1 achieved only adequate fit (RMSEA = 0.059/0.051; NNFI/CFI < 0.95). These findings highlight Artificial Intelligence's (AI) potential for automating psychometric tool development while underscoring the need for rigorous validation and expert oversight to ensure applied quality.

Keywords: *Big Five, Generative Artificial Intelligence, personality assessment, psychometrics*

INTRODUCCIÓN

La evaluación de la personalidad mediante modelos psicométricos robustos constituye un pilar fundamental en la investigación psicológica contemporánea. Comprender las diferencias individuales estables en patrones de pensamiento, sentimiento y comportamiento permite predecir una amplia gama de resultados vitales, como el bienestar, el rendimiento académico y laboral, y la salud física y mental (Soto, 2019) o sobre la mortalidad, el divorcio y el rendimiento ocupacional (Roberts et al., 2007). Dentro de los múltiples marcos teóricos propuestos para estructurar el estudio de la personalidad, el Modelo de los Cinco Grandes Factores (Big Five) ha emergido como el paradigma dominante en las últimas décadas (McCrae & John, 1992). Este modelo postula que la mayor parte de la varianza en los rasgos de personalidad humana puede subsumirse en cinco dimensiones amplias y relativamente universales: Apertura a la Experiencia, Responsabilidad, Extraversión, Amabilidad y Neuroticismo. La robustez empírica del modelo se ha demostrado a través de análisis factoriales de datos de lenguaje natural (léxicos) y cuestionarios en diversas culturas e idiomas, aunque su universalidad exacta sigue siendo objeto de debate, particularmente en poblaciones no industrializadas (Gurven et al., 2013), lo que refuerza la necesidad de validar instrumentos adaptados culturalmente. Los cinco factores han mostrado una notable estabilidad a lo largo de la vida adulta (Atherton et al., 2022) y se asocian consistentemente con variables relevantes en distintos dominios de la vida de la persona (Soto, 2019; Wilmot & Ones, 2019).

Tradicionalmente, el desarrollo de instrumentos de evaluación psicológica, como los cuestionarios de personalidad basados en el Big Five, ha sido un proceso laborioso que requiere una considerable inversión de tiempo y esfuerzo experto en la redacción, revisión y validación de ítems (Downing, 2006). La psicometría enfrenta persistentemente desafíos como la optimización de ítems, la reducción de sesgos y la adaptación a contextos multiculturales (Van der Linden, 2018).

En años recientes, los modelos de inteligencia artificial generativa, como GPT–3, Mistral y Deepseek, han emergido como herramientas prometedoras para automatizar tareas complejas, particularmente en el análisis y generación de textos. Estos sistemas, entrenados en corpus lingüísticos masivos, pueden producir textos coherentes y adaptados a especificaciones teóricas, lo que sugiere su potencial para optimizar procesos de diseño psicométrico y particularmente en la generación de ítems psicológicos. En este sentido, Floridi & Chiriatti (2020) señalan que estos modelos de lenguaje tienen la capacidad de emular estructuras lingüísticas humanas con precisión estadística, si bien destacan que carecen de comprensión semántica profunda.

El potencial de los grandes modelos de lenguaje (GML) en psicometría es multifacético. Se exploran aplicaciones en la generación automática de ítems, la adaptación de pruebas, la calificación automatizada de respuestas abiertas, e incluso la simulación de respuestas de participantes para pre–testear instrumentos (Russell–Lasalandra et al., 2024). Específicamente, en el ámbito de la generación automática de ítems, los GML pueden recibir instrucciones detalladas sobre el constructo a medir (por ejemplo, una faceta específica de la Amabilidad), el formato del ítem (por ejemplo, escala Likert), la población objetivo y las propiedades psicométricas deseadas, para luego generar un conjunto inicial de reactivos. Este enfoque promete acelerar significativamente el proceso de desarrollo, permitir la creación de grandes bancos de ítems y facilitar la generación de formas paralelas de tests (Hernandez & Nie, 2022; Li et al., 2024).

Potencialidades y Desafíos de la Inteligencia Artificial (IA) en la Psicometría de la Personalidad

La aplicación de grandes modelos de lenguaje para generar ítems de personalidad, como se investiga en el presente estudio, ofrece ventajas significativas. Puede reducir la carga de trabajo de los expertos, generar ítems con formulaciones novedosas y asegurar que cubran adecuadamente el dominio del constructo definido teóricamente. Investigaciones preliminares sugieren que estos modelos pueden desarrollar reactivos con calidad superficial comparable a los redactados por humanos e incluso producir escalas con propiedades psicométricas prometedoras (Laverghetta et al., 2024; Russell–Lasalandra et al., 2024). No obstante, la integración de la IA en la psicometría no está exenta de desafíos y requiere una validación rigurosa. Una preocupación central es la calidad psicométrica de los ítems generados automáticamente, dado que la calidad de los instrumentos generados depende críticamente de la capacidad de los modelos para emular el conocimiento experto y evitar sesgos semánticos o redundancias (Falcão et al., 2023). Este estudio pretende contribuir a la validación de IA generativa en psicometría, partiendo del interrogante de si los instrumentos desarrollados por los modelos poseen propiedades psicométricas aceptables.

OBJETIVOS

El presente estudio tiene como objetivo principal evaluar y comparar la calidad psicométrica de instrumentos de evaluación de la personalidad generados por diferentes modelos de inteligencia artificial generativa. A continuación, se detallan los objetivos específicos:

- Evaluar si los cuestionarios de personalidad creados por modelos de IA (Open AI o3, Deepseek R1 y Mistral–medium) poseen propiedades psicométricas aceptables en términos de validez de constructo, fiabilidad y ausencia de sesgo.
- Determinar si los modelos de IA pueden generar ítems que cubran adecuadamente los dominios teóricos de los cinco grandes factores de la personalidad (Apertura a la Experiencia, Responsabilidad, Extraversión, Amabilidad y Neuroticismo).
- Investigar si los modelos de IA pueden producir ítems con la suficiente calidad para formar escalas con propiedades psicométricas prometedoras.
- Proveer evidencia empírica sobre la viabilidad de utilizar modelos de IA generativa en el diseño de instrumentos psicométricos.

MATERIAL y MÉTODO

Diseño

Se empleó un diseño instrumental, comparativo y transversal cuyo objetivo fue contrastar la calidad psicométrica de tres versiones de un instrumento de evaluación de la personalidad (Open AI o3, Deepseek R1 y Mistral–medium), todas basadas en la teoría de los cinco grandes factores.

Participantes

La muestra inicial la conformaron 425 personas adultas (310 mujeres y 115 hombres) con edades entre 18 y 64 años (M = 32.02 años DE = 9.04 años), todas con, al menos, nivel secundario completo y residentes en la provincia de San Luis, Argentina. Tras la detección y eliminación de 41 casos por presentar valores atípicos (vía boxplots), inconsistencia semántica o sospecha de

deseabilidad social, la muestra final quedó en N = 384 participantes, manteniendo una proporción aproximada de 71.60% mujeres y 28.40% hombres. La selección fue de tipo accidental y, en su mayoría (68.45%), incluyó estudiantes universitarios de la región.

Instrumentos

Cada uno de los tres modelos generó un cuestionario de 50 reactivos: 10 por factor (Apertura a la Experiencia, Responsabilidad, Extraversión, Amabilidad y Neuroticismo) con escala Likert de 5 puntos (1 = "Totalmente en desacuerdo" a 5 = "Totalmente de acuerdo"). Por factor se incluyeron ocho ítems con puntuación directa y dos invertidos. Además, se agregaron dos reactivos de control para detectar inconsistencia semántica y dos para deseabilidad social.

PROCEDIMIENTO

Se describe a continuación la secuencia de tareas implementadas desde la generación de los reactivos, hasta la recolección de datos.

a) Diseño de los instrumentos: Inicialmente se solicitó a cada IA recomendaciones acerca de la estructura óptima (número de ítems, escala de respuesta y composición reactivos de control) que debería tener un instrumento que evaluara personalidad, desde el enfoque de los cinco grandes factores, para que presentara excelentes indicadores psicométricos en una muestra de adultos en Argentina. Las respuestas de los modelos fueron concordantes en indicar que con una estructura de 50 ítems (10 por cada factor) y cuatro ítems de control (dos de Deseabilidad Social y dos de Inconsistencia semántica), con opciones de respuesta en una escala Likert de cinco puntos, se podría obtener un instrumento breve y con capacidad para evaluar las dimensiones teóricas propuestas. A continuación, se le solicitó a cada modelo de IA la redacción de los reactivos con instrucciones claras respecto de las condiciones que deberían cumplir: cantidad de ítems, escala de respuesta, que la redacción fuera original, ajustada a las expresiones de la población diana, y que a la vez, fueran presentados para lograr propiedades psicométricas óptimas en cuanto a validez (contenido, constructo, criterio, discriminante, facial, convergente e incremental) y confiabilidad (consistencia interna, partición por mitades, acuerdo interjueces), teniendo especial cuidado para que cada reactivo reflejara únicamente el constructo asignado y minimizar cargas factoriales cruzadas. Finalmente se les solicitó que distribuyeran los ítems aleatoriamente para asegurar el tema de aquiescencia.

b) Confección: La propuesta de cada modelo fue revisada por dos expertos en el campo del estudio de la personalidad y de las diferencias individuales, con el objeto de asegurar la pertinencia de los reactivos al constructo propuesto y su aplicabilidad a los participantes del estudio. No fue necesaria ninguna modificación sustancial de los reactivos. A continuación, se construyó un formulario Google donde además de atender las cuestiones relacionadas con los aspectos éticos (consentimiento informado) se incluyeron reactivos que indagaron aspectos sociodemográficos. Finalizado este proceso se realizó una prueba piloto con 20 participantes del mismo contexto, no detectándose necesidad de ajustes.

c) Administración: Durante abril de 2025, el cuestionario se dispuso en línea vía Google Forms, bajo acceso por invita-

ción individual. Los ítems se presentaron en orden fijo, según aleatorización previa por modelo; se incluyó un bloque inicial de instrucciones; no hubo límite de tiempo.

Procedimiento estadístico

a) **Depuración de datos**: Inicialmente se excluyeron casos con valores atípicos (gráficos de caja), respuestas claramente inconsistentes o indicios de sesgo por deseabilidad social.
b) **Análisis de datos:** Se calcularon medidas de tendencia central, dispersión y se examinó la distribución de cada reactivo identificándose que la mayoría presentó asimetría significativa. Este motivo, sumado al hecho de que el nivel de medición de cada reactivo es ordinal, justificó la decisión de implementar el análisis factorial mediante matrices policóricas con el software Factor (12.06.08) (Lorenzo–Seva & Ferrando, 2025). Se indagaron indicadores de adecuación de la matriz factorial, así como de cada reactivo (Determinante de la matriz de correlación policórica, Test de Bartlett, Kaiser–Meyer–Olkin (KMO) y Medida de adecuación muestral (MSA)), en tanto que el análisis factorial se realizó solicitando el ajuste a cinco factores con rotación Promin, implementando complementariamente el Análisis paralelo donde se compararon los valores propios reales frente al percentil 95 de la distribución de autovalores obtenida mediante simulación aleatoria. Finalmente, para evaluar los índices de ajuste de los modelos propuestos se analizaron: RMSEA (Root Mean Square Error of Approximation), NCP (Non–Centrality Parameter), Test Approximate Fit, χ^2 mínimo y χ^2 de LOSEFER, NNFI (Tucker–Lewis Index) / TLI (Tucker–Lewis Index), CFI (Comparative Fit Index), BIC (Bayesian Information Criterion) y Pruebas de close–fit, poder estadístico y distribución de residuos. Para obtener indicadores de confiabilidad, se calculó el alfa de Cronbach para cada factor mediante el software Jamovi (2.6.26) (The Jamovi project, 2024).

Aspectos éticos

El protocolo se diseñó estrictamente conforme a los lineamientos de la Guía Ética para las Investigaciones de la Facultad de Psicología de la Universidad Nacional de San Luis, aprobada mediante Ord. 002/20. De este modo, se garantizó el anonimato de los participantes y la confidencialidad en el tratamiento de los datos, mediante la firma del correspondiente consentimiento informado.

RESULTADOS

Análisis de Confiabilidad

En el análisis de la confiabilidad interna de los tres instrumentos se calcularon los coeficientes alfa de Cronbach para cada dimensión y para el instrumento general (Tabla 1). Para el factor Apertura a la experiencia, los coeficientes oscilaron entre .745 (Mistral) y .749 (Deepseek R1), con Open AI o3 mostrando un valor de .746, situándose todos en el rango aceptable. En Responsabilidad, los valores alfa fueron superiores a .81 en los tres modelos, destacando Mistral con .849. Extroversión también mostró valores ≥ .82, siendo Mistral el más alto (.855). Para Amabilidad, los resultados variaron significativamente: Deepseek R1 obtuvo el valor más bajo (.624), clasificado como cuestionable, mientras que

Tabla 1
Índices de confiabilidad mediante alfa de Cronbach

	Open AI o3	Mistral	Deepseek R1
Apertura a la experiencia	.746	.745	.749
Responsabilidad	.815	.849	.815
Extroversión	.821	.855	.826
Amabilidad	.713	.748	.624
Neuroticismo	.855	.818	.865
General	.824	.768	.764

Open AI o3 (.713) y Mistral (.748) alcanzaron niveles aceptables. Neuroticismo fue el factor con mayor confiabilidad en todos los modelos, con coeficientes de .855 (Open AI o3), .818 (Mistral) y .865 (Deepseek R1), todos por encima de .81.

A nivel general, el alfa de Cronbach para Open AI o3 fue de .824, clasificado como bueno, mientras que Mistral y Deepseek R1 registraron .768 y .764, respectivamente, dentro del rango aceptable. Estos hallazgos indican que, aunque todos los modelos presentan niveles variables de consistencia interna, Open AI o3 exhibió la mayor confiabilidad general, seguido por Mistral y Deepseek R1. No obstante, el desempeño de Deepseek R1 en el factor Amabilidad sugiere una menor homogeneidad entre sus ítems en esta dimensión.

Tabla 2
Indicadores de la adecuación de la matriz de correlaciones

Indicador	Open AI o3	Mistral	Deepseek R1
Determinante de la matriz de correlación policórica	< .000001	< .000001	< .000001
Test de Bartlett χ^2 (df) P valor	4208.0 (1225), p = .000010	4184.9 (1225), p = .000010)	4449.7 (1225), p = .000010
Kaiser–Meyer–Olkin (KMO) valor (BCa 95% CI)	.4185 (.116 – .571)	.76971 (.351– .816)	.7958 (.569 – .821)
Cantidad de Ítems propuestos para remover (MSA)	19	0	0

Tal como lo muestra la Tabla 2 el Determinante de la matriz policórica fue inferior a .000001 en los tres casos, lo que sugiere una alta multicolinealidad entre las variables y, por tanto, una estructura interna susceptible de ser factorizada, o bien, si no se acompaña de un buen ajuste global y una solución factorial clara puede indicar redundancia de los reactivos.

La prueba de esfericidad de Bartlett resultó estadísticamente significativa en todos

los modelos: Open AI o3: χ^2 (1225) = 4208.0, p < .001; Mistral: χ^2 (1225) = 4184.9, p < .001; y Deepseek R1: χ^2 (1225) = 4449.7, p < .001, confirmando la viabilidad del análisis factorial en los tres modelos.

Sin embargo, el índice KMO, que evalúa la proporción de varianza común entre los ítems respecto a su varianza total, mostró resultados divergentes. El instrumento generado por Open AI o3 obtuvo un valor de KMO = .4185 (IC 95% BCa = [.116, .571]), que se encuentra por debajo del umbral mínimo aceptable de .50, lo que indica una insuficiente adecuación muestral global y sugiere que los ítems comparten poca varianza común. Además, se identificaron 19 ítems con valores MSA individuales por debajo del criterio recomendado, algunos extremadamente bajos (por ejemplo, ítem 31: MSA = .12565), lo que compromete seriamente la validez del modelo factorial propuesto por este instrumento.

En contraste, los instrumentos generados por Mistral y Deepseek R1 mostraron una adecuación muestral aceptable a buena. El índice KMO de Mistral fue de .7697 (IC 95% BCa = [.351, .816]), y el de Deepseek R1 alcanzó un valor ligeramente superior, KMO = .7958 (IC 95% BCa = [.569, .821]), ambos dentro del rango considerado meritorio según los estándares de Kaiser (1974) y sin ítem con valores MSA por debajo del umbral de .50, lo que respalda una estructura interna más coherente en términos de covariación.

En conjunto, estos resultados sugieren que, mientras los instrumentos de Mistral y Deepseek R1 cumplen con los requisitos psicométricos básicos para ser sometidos a análisis factorial exploratorio, el instrumento generado por Open AI o3 presenta serias limitaciones en cuanto a la adecuación de su matriz de correlaciones, tanto a nivel global como por ítem, lo que pone en duda la utilidad de su estructura factorial sin una revisión sustancial de los reactivos involucrados.

Determinación del número de factores

Para asegurar la concordancia teórica con la estructura de los cinco grandes factores de personalidad, se solicitó inicialmente la retención de cinco factores en cada instrumento. A continuación, se resumen los resultados de la varianza explicada y del análisis paralelo.

Tabla 3

Autovalores y varianza explicada de los factores identificados en cada modelo

F	Open AI o3			Mistral			Deepseek R1		
	Autov.	% Var.	% Acum.	Autov.	% Var.	% Acum.	Autov.	% Var.	% Acum.
1	8.2296	16.46	16.46	8.5953	17.19	17.19	8.3802	16.76	16.76
2	6.4040	12.81	29.27	5.7204	11.44	28.63	4.8992	9.80	26.56
3	4.4134	8.83	38.09	4.4610	8.92	37.55	4.1877	8.38	34.93
4	3.0684	6.14	44.23	3.1121	6.22	43.78	2.8265	5.65	40.59
5	2.1451	4.29	48.52	2.5498	5.10	48.88	2.3078	4.62	45.21

Tal como se puede observar en la Tabla 3, los porcentajes de Varianza acumulada para los cinco factores solicitados superan el 45% en los tres modelos. En la propuesta de Open AI o3, los factores presentaron valores propios (autovalores) de 8.23, 6.40, 4.41, 3.07 y 2.15, que explican el 48.5% de la varianza total acumulada. Para el modelo de Mistral, los valores propios fueron 8.60, 5.72, 4.46, 3.11 y 2.55, alcanzando una varianza acumulada del 48.9%, mientras que para el de Deepseek R1, se observaron autovalores de 8.38, 4.90, 4.19, 2.83 y 2.31 para los cinco primeros factores, con una varianza acumulada del 45.2%.

Estos porcentajes se sitúan dentro de rangos aceptables para el Análisis Factorial Exploratorio cuando se trabaja con constructos amplios (Field, 2013), aunque el instrumento de Deepseek R1 muestra una explicación algo más moderada en comparación con Open AI o3 y Mistral.

Análisis paralelo

Se compararon los valores propios reales frente al percentil 95 de la distribución de autovalores obtenida mediante simulación aleatoria (Horn, 1965) (Tabla 4).

Tabla 4

Autovalores y percentil 95 del Análisis paralelo en cada modelo

Factor	Open AI o3	Mistral	Deepseek R1
1	8.23 * / 1.98**	8.60 * / 1.96**	8.38 * / 1.92**
2	6.40 * / 1.86**	5.72 * / 1.85**	4.90 * / 1.81**
3	4.41 * / 1.78**	4.46 * / 1.78**	4.19 * / 1.74**
4	3.07 * / 1.71**	3.11 * / 1.71**	2.83 * / 1.68**
5	2.15 * / 1.66**	2.55 * / 1.66**	2.31 * / 1.63**
6	1.59 * / 1.61**	2.04 * / 1.61**	1.92 * / 1.58**

Nota. *Autovalores real / **percentil 95

En todos los casos, el análisis paralelo indicó que hasta el sexto factor los autovalores reales superan el umbral del percentil 95 de la aleatoriedad, lo que sugiere la retención de seis factores en cada instrumento, aunque sólo cinco corresponden a la estructura teórica deseada.

A pesar de que el análisis paralelo respalda la retención de seis factores, la decisión final debe equilibrar la evidencia empírica con la coherencia teórica. Dado que la teoría de los cinco grandes factores postula cinco dimensiones claramente definidas, se optó por la extracción de cinco factores en cada instrumento. Esta elección está fundamentada en:

1. Consistencia teórica: La estructura de cinco dimensiones es el referente en investigación de personalidad (McCrae & John, 1992).
2. Simplicidad interpretativa: Un modelo de cinco factores facilita la asignación de ítems a cada rasgo y la comparación entre instrumentos.
3. Varianza explicada razonable: Los porcentajes acumulados de varianza para cinco factores, aunque moderados

(45.2 – 48.9%), son adecuados para mediciones de constructos amplios y permiten mantener la parsimonia del modelo.

Por lo tanto, aunque la evidencia de análisis paralelo indique la posibilidad de un sexto factor en todos los casos, la retención de cinco factores se considera la opción más apropiada para alinear los resultados con el marco teórico con el cual fueron creados.

A continuación, se presenta el análisis del ajuste y la validación de la solución factorial de cinco factores para cada instrumento, incluyendo índices absolutos, de aproximación y de ajuste incremental, así como el criterio de información bayesiano (BIC).

Tabla 5

Indicadores de ajuste de cada modelo

Índice	Open AI o3	Mistral	Deepseek R1
RMSEA (LOSEFER) BCa 95% CI; criterio	.046 (.044 – .046; close)	.059 (.057 – .059; fair)	.051 (.050 – .051; fair)
NCP / df	943.138 / 985	938.213 / 985	994.850 / 985
Prueba de ajuste aproximado p (H_0: RMSEA < 0.05)	p = 1.000	p = 1.000	p = 1.000
χ^2 mínimodf = 985; p	1318.123 p = .00001	1485.908 p = .00001	1495.355 p = .00001
LOSEFER χ^2 df = 985; p	1777.487 p = .00001	2294.141 p = .00001	2032.171 p = .00001
NNFI (TLI) BCa 95% CI	.952 (.951 – .961)	.921 (.920 – .931)	.930 (.929 – .941)
CFI BCa 95% CI	.961 (.960 – .968)	.937 (.936 – .944)	.944 (.943 – .953)
BIC	–5213.857	4077.767	3833.337

En la Tabla 5 es posible apreciar que, en términos absolutos, el modelo de Open AI o3 exhibió un ajuste excelente, con un RMSEA de .046 (IC 95% = .044–.046), dentro del rango de "close fit", y un NCP/df cercano a .96, lo que sugiere un modelo parsimonioso que reproduce adecuadamente las covarianzas observadas. La prueba de ajuste aproximado (H_0: RMSEA < .05) fue no significativa (p = 1.000), confirmando que el error de aproximación es inferior a .05.

Por su parte, los instrumentos de Mistral y Deepseek R1 presentaron RMSEA en rangos de "fair fit" (.059 y .051, respectivamente), con NCP/df también próximos a 1.00, pero distanciados del criterio de ajuste "close". Aun así, ambos modelos superan el umbral de aceptabilidad en la prueba de ajuste

RMSEA (p = 1.000).

En cuanto al contraste χ^2, en los tres casos el LOSEFER χ^2 corregido fue sustancialmente mayor que el χ^2 no corregido, reflejando la penalización por no centralidad e indicando la necesidad de dicha corrección en muestras de tamaño moderado.

Los índices incrementales muestran que solo el modelo de Open AI o3 alcanza valores de NNFI y CFI superiores a .95, mientras que Mistral y Deepseek R1 se sitúan por debajo de este umbral (.921–.944), si bien siempre por encima de .90, lo cual es considerado aceptable en estudios exploratorios de constructos amplios.

Finalmente, el BIC más bajo (negativo) para Open AI o3 indica una mejor parsimonia comparada con los otros dos modelos, cuyos BIC positivos sugieren una peor compensación entre ajuste y complejidad. En conjunto, estos hallazgos apuntan a que el instrumento de Open AI o3 no solo cumple con los requisitos teóricos –cinco factores conforme al modelo de los Big Five– sino que además presenta un ajuste estadístico superior para reproducir la estructura factorial observada.

Tabla 6

Análisis de la robustez del ajuste factorial y la idoneidad del modelo para reproducir las covarianzas observadas

Indicador	Open AI o3	Mistral	DeepSeek
Test Close–fit RMSEA; p (df = 985)	.046; p = .98	.059; p = .24	.051; p = .24
Análisis de potencia H_0: RMSEA=0.05 vs H_1: RMSEA=0.08; β	β > .99	β > .99	β > .99
Distribution of Residuals Nº residuos; mín., med., máx., media, var.	1225; –0.1976, –0.0097, 0.1367; –0.0073, 0.0028	1225; –0.1910, –0.0037, 0.5366; –0.0042, 0.0032	1225; –0.2028, –0.0070, 0.1908; –0.0085, 0.0029
RMSR (CI) WRMR (CI)	0.0530 (0.053) 0.0446 (0.045)	0.0564 (0.056) 0.0484 (0.048)	0.0550 (0.055) 0.0467 (0.047)

Pruebas de close–fit, poder estadístico y distribución de residuos

La Tabla 6 exhibe que la prueba de ajuste cercano (H_0: RMSEA < .05) indicó que únicamente el modelo de Open AI o3 cumple holgadamente este criterio (RMSEA = 0.046; p = .98), seguido por Mistral (RMSEA = 0.059; p = .24), mientras que el modelo de Deepseek R1 (RMSEA = 0.051) no alcanza el umbral de ajuste cercano (p < .05), lo cual sugiere que, aunque los tres modelos presentan RMSEA aceptables, solo el modelo de Open AI o3 exhibe un error de aproximación tan bajo que no puede diferenciarse de un valor de .05 en este test. Contar con un test de close–fit tan favorable como el de Open AI o3 fortalece la argumentación de que ese instrumento emula muy bien la estructura factorial teórica. Para Mistral y

Deepseek R1, podría discutirse la necesidad de revisar algunos ítems o la dimensionalidad para acercarse aún más al estándar de "cerca perfecto" que ofrece Open AI o3.

Para el contraste de hipótesis H_0: RMSEA = 0.05 versus H_1: RMSEA = 0.08, todos los modelos mostraron un poder estadístico muy elevado ($\beta > .99$), lo que indica una probabilidad prácticamente nula de cometer un error tipo II al detectar un ajuste deficiente si existiera una desviación real del valor de RMSEA bajo la hipótesis nula.

Respecto de la distribución de residuos, los valores de RMSR y WRMR, junto con la simetría en la mediana cercana a cero y la limitada varianza de los residuos, sugieren que no existen sesgos sistemáticos en la reproducción de las covarianzas observadas y que el modelo de Open AI o3 presenta el ajuste más homogéneo, seguido muy de cerca por Deepseek R1 y Mistral.

Con estos resultados, se corrobora la robustez del modelo factorial de Open AI o3 en todas las dimensiones evaluadas, mientras que Mistral y Deepseek R1, aunque aceptables, muestran un ajuste ligeramente inferior, especialmente en la prueba de close–fit y en el nivel de discrepancia promedio de sus residuos.

DISCUSIONES y CONCLUSIONES

El presente estudio tuvo como objetivo principal evaluar comparativamente la calidad psicométrica de instrumentos de evaluación de la personalidad basados en el modelo de los Cinco Grandes Factores (Big Five), generados por tres modelos distintos de inteligencia artificial (IA) generativa: Open AI o3, Mistral y Deepseek R1. Los hallazgos ofrecen una visión matizada sobre el potencial y los desafíos actuales del uso de grandes modelos de lenguaje en el desarrollo de instrumentos psicométricos.

En términos de fiabilidad, los tres modelos demostraron ser capaces de generar escalas con consistencia interna generalmente aceptable o buena para la mayoría de los factores. Destaca la alta fiabilidad obtenida para Neuroticismo en los tres modelos ($\alpha > .81$) y para Responsabilidad ($\alpha > .81$), así como para Extroversión ($\alpha > .82$). Sin embargo, se observaron inconsistencias, particularmente en el factor Amabilidad generado por Deepseek R1, cuya fiabilidad fue cuestionable ($\alpha = .624$), sugiriendo una menor homogeneidad en los ítems de esta dimensión para ese modelo específico (Nunnally & Bernstein, 1994). El instrumento de Open AI o3 mostró la mayor confiabilidad general ($\alpha = .824$). Estos resultados se alinean con investigaciones previas que indican que los GML pueden generar escalas con propiedades psicométricas prometedoras, aunque la calidad puede ser variable (Laverghetta et al., 2024; Russell–Lasalandra et al., 2024).

El análisis de la adecuación para la factorización reveló resultados divergentes. Mientras que los instrumentos de Mistral (KMO = .770) y Deepseek R1 (KMO = .796) mostraron una adecuación muestral aceptable a buena, sin ítems problemáticos (MSA < .50), el instrumento de Open AI o3 presentó un KMO inaceptablemente bajo (KMO = .419) y múltiples ítems con pobre adecuación muestral individual (19 ítems con MSA < .50). Este hallazgo es crítico, ya que un KMO bajo cuestiona la idoneidad de la matriz de correlaciones para el análisis factorial, sugiriendo que los ítems comparten poca varianza común o que existen problemas de multicolinealidad o singularidad no deseada. Este hallazgo subraya la importancia de combinar la generación automatizada con la supervisión experta y la depuración de ítems (Falcão et al., 2023). Paradójicamente, y a pesar de esta pobre adecuación inicial, el modelo de Open AI o3 fue el que demostró un mejor ajuste estructural a la solución de cinco factores.

Respecto a la estructura factorial, aunque el análisis paralelo sugirió consistentemente la retención de seis factores para los tres modelos, se optó por forzar una solución de cinco factores para mantener la coherencia con el marco teórico del Big Five, una decisión justificada por la interpretabilidad y la parsimonia, dado que la varianza explicada acumulada por los cinco factores fue razonable (45.2% – 48.9%). Futuras investigaciones podrían explorar la naturaleza de ese sexto factor emergente.

Los índices de ajuste del modelo factorial (realizado mediante rotación Promin sobre matrices policóricas) indicaron una clara superioridad del instrumento generado por Open AI o3. Este modelo alcanzó un ajuste excelente ("close fit") según el RMSEA (.046) y superó los umbrales recomendados para NNFI (.952) y CFI (.961), además de presentar el BIC más favorable (indicativo de mejor parsimonia). Las pruebas de ajuste cercano (close–fit) y la distribución de residuos corroboraron la robustez de este modelo. En contraste, los modelos de Mistral y Deepseek R1 mostraron un ajuste aceptable ('fair fit' según RMSEA: .059 y .051 respectivamente) pero inferior, con índices NNFI y CFI por debajo de .95 (Field, 2013).

La aparente contradicción entre la pobre adecuación muestral inicial (KMO/MSA) y el excelente ajuste final del modelo de Open AI o3 merece una reflexión. Podría indicar que, si bien algunos ítems individuales generados por este modelo de lenguaje pueden ser subóptimos o compartir poca varianza común de forma aislada, el conjunto de ítems logra capturar colectivamente la estructura latente de los cinco factores de manera muy precisa al ser modelados conjuntamente. Esto subraya la importancia de evaluar no sólo los ítems individuales, sino la calidad psicométrica del instrumento completo generado por IA. También sugiere que los GML, como Open AI o3, podrían estar operando bajo principios de generación de contenido que no se alinean perfectamente con los supuestos clásicos de la psicometría para ítems individuales, pero que resultan efectivos a nivel estructural global. Investigaciones como las de Laverghetta et al. (2024) y Russell–Lasalandra et al. (2024) están desarrollando marcos para optimizar y validar estos procesos, integrando GML con técnicas psicométricas avanzadas y enfoques iterativos de mejora.

Estos hallazgos contribuyen a la creciente literatura sobre la aplicación de IA en psicometría (Hernandez & Nie, 2022; Li et al., 2024; Russell–Lasalandra et al., 2024). Demuestran el potencial de los GML para acelerar el desarrollo de instrumentos, generar contenido novedoso y cubrir dominios teóricos, tal como se proponía en los objetivos. Sin embargo, también evidencian la necesidad imperativa de una validación psicométrica rigurosa y la supervisión de expertos (Falcão et al., 2023), ya que la calidad y las propiedades de los instrumentos generados pueden variar significativamente entre modelos y factores. La preocupación por la calidad psicométrica, los sesgos potenciales y la falta de comprensión profunda de los GML (Floridi & Chiriatti, 2020) sigue siendo relevante.

Es importante reconocer las limitaciones de este trabajo. Primero, la muestra se compuso principalmente de estudiantes universitarios de una región específica de Argentina, lo que limita la generalización de los resultados a otras poblaciones. Segundo, se evaluaron únicamente tres GML específicos; modelos más recientes o diferentes podrían arrojar resultados distintos. Tercero, el estudio se centró exclusivamente en el modelo Big Five; la capacidad de la IA para generar ítems para otros constructos psicológicos requiere investigación adicional. Cuarto, el diseño transversal no permite evaluar la estabilidad temporal de los instrumentos generados. Finalmente, aunque se instruyó a las IA para optimizar propiedades psicomé-

tricas, el proceso interno mediante el cual los GML generan los ítems permanece en gran medida como una "caja negra", dificultando la comprensión detallada de por qué un modelo supera a otro en ciertas métricas.

La investigación futura debería abordar estas limitaciones. Sería valioso replicar estos hallazgos con muestras más diversas y representativas, incluyendo diferentes grupos culturales y etarios, para evaluar la invarianza de medida de los instrumentos generados por IA. Comparar un espectro más amplio de GML, incluyendo modelos de código abierto y aquellos con diferentes arquitecturas o datos de entrenamiento, proporcionaría una visión más completa del panorama actual. Investigar la aplicación de estos métodos a otros constructos de personalidad y dominios psicológicos es fundamental. Estudios longitudinales son necesarios para evaluar la fiabilidad test–retest y la validez predictiva a largo plazo de las escalas generadas por IA. Además, es crucial desarrollar metodologías para "abrir la caja negra" de los GML en el contexto de la generación de ítems, quizás mediante el análisis de las representaciones internas (embeddings) o el desarrollo de técnicas de prompting más sofisticadas y transparentes, como las exploradas por Laverghetta et al. (2024) y Russell–Lasalandra et al. (2024). Finalmente, la investigación debe continuar abordando los desafíos éticos, incluyendo la mitigación de sesgos algorítmicos, la garantía de la privacidad y equidad en el uso de estas herramientas en la evaluación psicológica.

En conclusión, este estudio demuestra que los modelos de IA generativa, como Open AI o3, Mistral y Deepseek R1, poseen una capacidad considerable para generar ítems destinados a evaluar los cinco grandes factores de la personalidad, logrando en general instrumentos con niveles aceptables de fiabilidad y validez estructural. El modelo Open AI o3 destacó por producir un instrumento con un ajuste factorial superior a la estructura teórica del Big Five, a pesar de las preocupaciones iniciales sobre la adecuación de sus ítems individuales. Los modelos Mistral y Deepseek R1 también generaron instrumentos mayormente funcionales, aunque con un ajuste estructural ligeramente inferior y, en el caso de Deepseek R1, con problemas de fiabilidad en una de las dimensiones.

Estos resultados subrayan el potencial transformador de la IA para la psicometría, particularmente en la automatización y optimización de las fases iniciales del desarrollo de tests. Sin embargo, la variabilidad en el desempeño entre modelos y la aparición de resultados psicométricos anómalos (como la dicotomía KMO/ajuste en Open AI o3 o la baja fiabilidad de Amabilidad en Deepseek R1) resaltan que la IA generativa no es una panacea y su aplicación requiere una validación empírica exhaustiva y una supervisión experta continua. La integración exitosa de la IA en la evaluación psicológica dependerá del desarrollo de marcos de validación robustos, la atención a consideraciones éticas y un diálogo constante entre los avances tecnológicos y los principios psicométricos fundamentales.

REFERENCIAS BIBLIOGRÁFICAS

Atherton, O.E., Sutin, A.R., Terracciano, A., & Robins, R.W. (2022). Stability and change in the Big Five personality traits: Findings from a longitudinal study of Mexican–origin adults. *Journal of Personality and Social Psychology, 122*(2), 337–350. https://doi.org/10.1037/pspp0000385

Downing, S.M. (2006). Twelve steps for effective test development. En S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Lawrence Erlbaum Associates Publishers.

Falcão, F., Pereira, D.M., Gonçalves, N., Champlain, A., Costa, P. & Pêgo, J. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education, 28*, 1441–1465. https://doi.org/10.1007/s10459–023–10225–y

Field, A.P. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4.ª ed.). SAGE Publications.

Floridi, L., & Chiriatti, M. (2020). GPT–3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines, 30*(4), 681–694. https://doi.org/10.1007/s11023–020–09548–1

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five–factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology*, *104*(2), 354–370. https://doi.org/10.1037/a0030841

Hernandez, I. & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology 76*, 1011–1035. https://doi.org/10.1111/peps.12543

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36. https://doi.org/10.1007/BF02291575

Laverghetta, A., Jr., Luchini, S., Linell, A., Reiter–Palmon, R., & Beaty, R. (2024). The creative psychometric item generator: a framework for item generation and validation using large language models. *ArXiv*. https://doi.org/10.48550/arXiv.2409.00202

Li, C.–J., Zhang, J., Tang, Y., & Li, J. (2024). Automatic item generation for personality situational judgment tests with large language models. *ArXiv*. https://doi.org/10.48550/arXiv.2412.12144

Lorenzo–Seva, U., & Ferrando, P.J. (2025). Factor (Version 12.06.08) [Computer software]. Universitat Rovira i Virgili. https://psico.fcep.urv.cat/utilitats/factor/

McCrae, R.R., & John, O.P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467–6494.1992.tb00970.x

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). McGraw–Hill.

Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life

outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345. https://doi.org/10.1111/j.1745–6916.2007.00047.x

Russell–Lasalandra, L.L., Christensen, A.P., & Golino, H. (2024). Generative Psychometrics via AI–GENIE: Automatic Item Generation and Validation via Network–Integrated Evaluation. *ArXiv*. https://doi.org/10.31234/osf.io/fgbj4

Soto, C.J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science, 30*(5), 711–727. https://doi.org/10.1177/0956797619831612

The Jamovi project (2024). Jamovi (Version 2.6.26) [Computer Software]. https://www.jamovi.org

Van der Linden, W.J. (2018). *Handbook of item response theory.* Vol II: Statistical tools. CRC Press.

Wilmot, M.P., & Ones, D.S. (2019). A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences*, *116*(46), 23004–23010. https://doi.org/10.1073/pnas.1908430116

Fecha recepción: 19/05/2025
Fecha aceptación: 29/08/2025