Innovaciones tecnológicas en la evaluación psicológica: reflexiones sobre su impacto y aplicaciones en la psicometría contemporánea

Guadalupe de la Iglesia*
https://orcid.org/0000-0002-0420-492X
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET),
Universidad de Palermo
Correo electrónico: gdelaiglesia@gmail.com

Recibido: 3 de julio del 2024 / Aceptado: 8 de enero del 2025
doi: https://10.26439/persona2025.n1.7238

RESUMEN. En el ámbito de la evaluación psicológica, se puede observar un viraje en el uso de métodos tradicionales hacia procedimientos que incorporan innovaciones tecnológicas, tales como plataformas digitales, videojuegos, chatbots, realidad virtual e inteligencia artificial. El objetivo de este ensayo es describir las innovaciones tecnológicas actualmente disponibles y las que se encuentran en vías de desarrollo en el ámbito de la evaluación psicológica. Estas herramientas, conocidas por sus siglas en inglés TBA (*technology-based assessment*), se caracterizan por ser evaluaciones basadas en un tipo de tecnología particular: la digital. En este trabajo, se describen los distintos tipos de TBA y se presentan diferentes desarrollos disponibles para evaluar un amplio abanico de variables psicológicas. Se detalla la creciente inclusión de TBA en el ámbito de la evaluación psicológica y se ponderan las ventajas y limitaciones que su uso conlleva. Finalmente, se realiza un recorrido por los lineamientos internacionales que se han consensuado en cuanto al desarrollo de TBA que sean válidas, confiables y justas y su uso. Se concluye que la incorporación de TBA en procesos de evaluación psicológica podría resultar ventajosa en cuanto a la accesibilidad y el *engagement*. Asimismo, este tipo de tecnologías superarían en abundancia y variedad al cúmulo de información obtenido por métodos tradicionales.

Palabras clave: innovaciones tecnológicas / evaluación psicológica / psicometría / inteligencia artificial / *machine learning* / videojuegos / *technology-based assessment*

* El presente trabajo de investigación fue financiado mediante el proyecto PICT 2020 SERIE A, código 0181 de la Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación, y el proyecto PIBAA 2022-2023, Código 28720210100431CO del Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina.

TECHNOLOGICAL INNOVATIONS IN PSYCHOLOGICAL ASSESSMENT: REFLECTIONS ON THEIR IMPACT AND APPLICATIONS IN CONTEMPORARY PSYCHOMETRICS

ABSTRACT. The field of psychological assessment is undergoing a marked transition from traditional methodologies toward approaches that incorporate technological innovations, including digital platforms, serious games, chatbots, virtual reality, and artificial intelligence (AI). This article presents a review of current and emerging technological tools in psychological assessment, with particular emphasis on technology-based assessments (TBAs)—instruments that rely on digital technologies to evaluate a wide range of psychological constructs. The article describes the main types of TBAs, outlines their applications in assessing diverse psychological variables, and analyzes their growing use in psychological assessments. It highlights both the advantages and limitations of these tools and considers international guidelines for developing and implementing TBAs that are valid, reliable, and fair. Overall, the integration of TBAs into psychological assessment represents a promising development, offering greater accessibility, increased engagement, and richer, more diverse data compared to traditional methods.

Keywords: Technology-Based Assessment / psychological assessment / psychometrics / artificial intelligence / machine learning / video games

INTRODUCCIÓN

Desde sus inicios, las herramientas de evaluación psicológica han sido fundamentales tanto en los ámbitos de investigación como en los aplicados. En los primeros, desempeñan un papel central como herramientas de recopilación de datos, mediante las cuales se ponen a prueba hipótesis teóricas y se interpretan los fenómenos bajo estudio. En el ámbito aplicado, los test proporcionan información en diferentes procesos de toma de decisiones, por ejemplo, en el ámbito educativo para evaluar habilidades u orientar vocacionalmente a un individuo, en el clínico para aportar información en procesos diagnósticos o de evolución de tratamientos o en el forense para respaldar la toma de decisiones de índole legal (Cohen & Swerdlik, 2017). Su disponibilidad y buen funcionamiento psicométrico resulta primordial, dado que estas herramientas se utilizan para obtener información, realizar interpretaciones y tomar decisiones que, en la mayoría de los casos, afectan la vida de las personas evaluadas.

Los métodos tradicionales de evaluación psicológica han sido pilares fundamentales tanto en la investigación como en el ámbito aplicado, sin embargo, presentan algunas limitaciones. Su enfoque estático, por ejemplo, dificulta la evaluación de procesos dinámicos complejos. La estandarización, aunque se presenta como una característica necesaria para garantizar la comparabilidad de los resultados, restringe la individualización y la aplicabilidad a poblaciones diversas. Además, la validez ecológica de estos métodos puede ser limitada, ya que los entornos de evaluación artificial dificultan la generalización de los resultados a contextos reales (Fortea et al., 2023). Existen, hoy en día, innovaciones tecnológicas que buscan superar algunas de estas limitaciones.

Este ensayo tiene como objetivo principal describir el panorama actual de la incorporación de innovaciones tecnológicas en la psicometría contemporánea. Para su desarrollo, se llevó a cabo una revisión de fuentes teóricas, investigaciones recientes y ejemplos prácticos extraídos de la literatura académica sobre el uso de la tecnología en la evaluación psicológica, tal como es habitual para este tipo abordajes (Zumbo, 2023). En todos los casos, se buscó articular la información obtenida destacando puntos de convergencia y contrastando las divergencias. Además, se delimitaron las ideas clave en esta intersección del conocimiento para responder a los siguientes objetivos específicos: (1) describir las características de las evaluaciones basadas en tecnología presentando desarrollos en diferentes contextos y poblaciones, (2) detallar el contexto de creciente inclusión de innovaciones tecnológicas en el campo de la evaluación psicológica destacando sus ventajas y desventajas y (3) especificar los principales lineamientos internacionales que se han consensuado en cuanto al desarrollo y uso de TBA que sean válidas, confiables y justas.

EVALUACIONES BASADAS EN TECNOLOGÍA

La integración de la tecnología digital en la evaluación psicológica tiene una larga trayectoria iniciada a mediados del siglo xx (Butcher et al., 2004). Si bien sus primeros usos se limitaban a tareas como la puntuación y el procesamiento de datos, hoy en día se vislumbra una proliferación de herramientas basadas en innovaciones tecnológicas diseñadas para medir una amplia gama de constructos psicológicos (Leong et al., 2016).

La terminología para referirse a este tipo de prácticas es variada. Se habla de *computerized testing* (evaluación computarizada), que enfatiza el rol de las computadoras, de *tele-assessments* o telemetría, que destaca la medición a distancia, de *digital assessments* (evaluaciones digitales) para referirse a la característica digital de los instrumentos usados, de *internet-based tests applications* (aplicaciones de *tests* basadas en internet), que subraya el uso de internet en la evaluación, de *game-based* o *video-game assessments* (evaluaciones basadas en videojuegos), que se refieren en particular al uso de videojuegos como una herramienta de evaluación psicológica, y de: evaluaciones basadas en tecnología o *technology-based assessments* (EBT o TBA), la denominación más consensuada. Este término tiene como ventaja el ser lo bastante amplio como para englobar a los mencionados anteriormente. De hecho, es la terminología elegida por la International Test Commission y la Association of Test Publishers (ITC & ATP, 2022) en su publicación sobre lineamientos para su desarrollo y uso.

La variedad de innovaciones tecnológicas disponibles hoy para usar en procesos de evaluación psicológica es sumamente amplia: servicios de plataformas digitales, evaluaciones adaptativas computarizadas, aplicaciones para dispositivos móviles y PC, chatbots, videojuegos, realidad virtual, entre muchas otras.

Plataformas digitales

Las plataformas digitales para la evaluación *online* han proliferado, y hoy en día existen cada vez más opciones para quienes necesitan administrar test psicológicos en este formato. Algunas plataformas no son específicas para llevar a cabo evaluaciones psicológicas, como SurveyMonkey, Qualtrics y Google Forms, pero se pueden adaptar fácilmente a los requisitos de test autoadministrables. Otras plataformas fueron diseñadas específicamente para este propósito, como las opciones digitalizadas ofrecidas por Pearson Assessments, TEACorrige o Paidós DEP. La mayoría de los test con formato de inventario o cuestionario autoadministrable cuentan con versiones en línea y muchas investigaciones se han dedicado a analizar si las versiones digitales son equivalentes a las de lápiz y papel y han encontrado que sí lo son (Butcher et al., 2004; Gnambs, 2022). El panorama es diferente para las técnicas que requieren la manipulación de objetos y una interacción física con el evaluador, tal como ocurre con las escalas Weschler. Sin embargo, se ha estudiado, por ejemplo, la administración digital del WISC-V (Wechsler

Intelligence Scale for Children-Fifth Edition), desarrollada por Pearson Inc. en su plataforma digital *Q-interactive*, con excelentes resultados (Gilbert et al., 2023; Wright, 2020).

Aplicaciones para dispositivos móviles y PC

Las aplicaciones para dispositivos móviles y PC son herramientas valiosas para llevar a cabo evaluaciones invisibles (Rosas et al., 2015), es decir, realizar la medición del constructo de una manera sutil y no intrusiva. Este tipo de abordaje puede resultar útil para el estudio de poblaciones infantiles o en casos de personas para quienes las situaciones de evaluación resultan altamente estresantes.

Entre las innovaciones de este tipo, podemos mencionar el Interactive Child Distress Screener (ICDS; Zieschank et al., 2022), un instrumento de *screening* desarrollado como una aplicación web que consiste en animaciones digitales que miden dificultades emocionales y conductuales en niños y niñas mediante el autorreporte. La evaluación consiste en mostrar dos animaciones contrapuestas que representan a niños y niñas experimentando determinados estados emocionales o mostrando ciertas conductas. Cada ítem tiene una valencia negativa y otra positiva y las negativas son las que suman al puntaje total. Luego de ver ambas animaciones, el evaluado debe cliquear en alguna para responder a la pregunta "¿Cuál es más parecida a vos?". Los estudios psicométricos sobre la herramienta incluyen un análisis factorial exploratorio en el que se aislaron las dimensiones de dificultades internalizantes y externalizantes y, también, análisis de su consistencia interna y su validez convergente con otros constructos relevantes. En ambos casos se obtuvieron excelentes resultados.

Shahamiri y Thabtah (2020) desarrollaron una aplicación para dispositivos móviles que funciona como un *screening* de síntomas del trastorno del espectro autista: el Autism AI System. El sistema ha sido analizado en cuanto a su sensibilidad y especificidad, y se han obtenido excelentes resultados. Una de las fortalezas de estas propuestas es su fácil administración. Al ser una técnica de *screening*, podría ser sumamente útil en la detección temprana de casos y, debido a su carácter portable y automatizado, su administración sería muy sencilla y superaría a las técnicas convencionales disponibles actualmente.

También en esta línea, Drake et al. (2023) diseñaron una simulación para evaluar el autocontrol en niños. La simulación se presenta en una *tablet* y consiste en un supermercado 2D. Los niños deben comprar ocho ítems de una lista con la consigna de buscar la opción más barata de acuerdo con el monto de dinero que se les da. En este caso, para analizar la validez y la confiabilidad se compararon los resultados con las medidas de impulsividad y de desconfianza a la publicidad y se obtuvo evidencia de un factor latente de autocontrol, lo cual se interpretó como un paso inicial en el estudio de este TBA.

Evaluaciones adaptativas computarizadas

En este punto, es importante describir las evaluaciones adaptativas computarizadas (CAT, por las siglas en inglés de *computerized adaptive test*). Las CAT se caracterizan por ser programas de computadora en los cuales los estímulos que se presentan se ajustan en tiempo real a las respuestas y el desempeño del individuo (Granziol et al., 2020). Las CAT son personalizadas, lo cual se logra mediante el uso de algoritmos; este principio fundamental se puede observar en diversas tecnologías. Por ejemplo, en los chatbots, se utilizan para que la interacción con el usuario se ajuste a las respuestas o elecciones de este. Esta adaptabilidad es habitual en el diseño de videojuegos, lo que se conoce como la curva de dificultad (nivel de dificultad, ritmo y sensación de presión) y la búsqueda de garantizar y mantener el *game flow* (Hodent, 2017). Estas evaluaciones dinámicas y adaptativas, en contraste con los clásicos diseños lineales, presentarían múltiples ventajas: mayor precisión, pruebas más cortas, menor tiempo de administración y mayor *engagement* del evaluado (ITC & ATP, 2022).

La CAT-Depression, por ejemplo, es un desarrollo de Tan et al. (2018) que se utiliza para medir síntomas de depresión. El algoritmo detrás de la herramienta selecciona automáticamente ítems que se ajustan a los niveles de depresión de quien está contestando el test. Hay evidencia psicométrica acerca de la buena consistencia interna, la unidimensionalidad del constructo evaluado y la validez de criterio para la CAT-Depression. Un desarrollo similar es el de Flens et al. (2016), quienes administraron un test de ansiedad mediante una CAT y hallaron que era altamente preciso y presentaba poder predictivo.

Las CAT también son útiles para evaluar procesos. Veerbeek y Vogelaar (2023), por ejemplo, desarrollaron un test dinámico computarizado para evaluar los procesos de aprendizaje de razonamiento analógico en niños. Los niños y niñas debían resolver problemas de geometría que requerían la construcción de una respuesta. Mediante la evaluación, se pudo medir el tiempo total empleado, el tiempo dedicado a planificar y revisar la respuesta y la variación en los tiempos de resolución.

Chatbots

También conocidos como agentes conversacionales, los chatbots son programas informáticos que simulan la conversación humana. Sus usos pueden ser variados, por ejemplo, Fan et al. (2023) analizaron el potencial de Juji para evaluar la personalidad. Luego de una conversación *online* con el chatbot, se extrajeron las características textuales del guion de la conversación y, mediante algoritmos de *machine learning*, se realizaron inferencias sobre la personalidad. En cuanto a sus evidencias psicométricas, los puntajes obtenidos mediante el chatbot presentaban alta consistencia interna y estabilidad temporal, una estructura factorial comparable a la de un test clásico de personalidad, evidencias de validez convergente/discriminante y de criterio externo (promedio

académico y adaptación a la universidad) y, además, incrementaba significativamente la varianza explicada por sobre aquella explicada por el test de autorreporte clásico.

Schick et al. (2022) compararon la administración de los test psicológicos de sintomatología general en tres formatos: lápiz y papel, web y chatbots. Hallaron que los tres resultaban medidas psicométricamente apropiadas y que los evaluados indicaron que el uso del chatbot resultaba el método más tedioso. En contraste, en la investigación de Hungerbuehler et al. (2021), la administración de los test psicológicos mediante un chatbot se complementó con elementos de gamificación. El chatbot Viki se utilizó para evaluar la salud mental de empleados de una planta industrial en términos de depresión, ansiedad, estrés, insomnio, agotamiento y estrés laboral. La tasa de respuesta de los empleados fue alta (superior al 80 %), lo que indicó que esta herramienta podría ser útil para evaluar la salud mental en ámbitos laborales.

Realidad virtual

En cuanto al uso de realidad virtual (VR por sus siglas en inglés), cabe destacar que una de sus mayores fortalezas radica en el incremento de la validez ecológica del test, dado que los entornos virtuales pueden replicar entornos físicos reales, lo que permitiría valorar la conducta del evaluado en un contexto similar al real y así incrementar la generalizabilidad del resultado (Fortea et al., 2023). Además, este tipo de evaluaciones permite la recolección multimodal de datos dada por la evaluación del comportamiento (por ejemplo a partir del análisis de los movimientos de los evaluados), la fisiología (la frecuencia cardíaca o la actividad cerebral, por ejemplo) y las respuestas verbales del evaluado (como sus respuestas a ciertas preguntas). Las propuestas para realizar evaluaciones psicológicas con VR son incipientes, pero diversas e incluyen su uso para valorar ansiedad, depresión y habilidades cognitivas (Chitale et al., 2022).

El Virtual Seminar Room (VSR; Wiebe et al., 2023), por ejemplo, tiene como el objetivo evaluar síntomas asociados al trastorno por déficit de atención con hiperactividad. En específico, busca evaluar atención, impulsividad e hiperactividad de manera multimodal. Hasta el momento, se ha analizado el uso del VSR en sujetos control. Los evaluados se ven inmersos en un aula virtual y se les solicita que realicen una tarea de ejecución continua (*continuous performance task* o CPT). Esta consiste en presionar rápidamente la barra espaciadora de un teclado (físico) cada vez que una combinación de dos letras aparece en una pizarra (virtual). Los evaluados deben abstenerse de presionar la barra ante combinaciones de otras letras y también deben mantenerse concentrados ante la presencia de estímulos distractores visuales, auditivos y audiovisuales (por ejemplo, un compañero de clase que saluda, un perro que ladra, una ambulancia que pasa). El aspecto multimodal de la herramienta está dado por sus variadas mediciones: desempeño en la tarea, medidas de electroencefalografía (EEG) y movimientos de la cabeza (actigrafía).

Otro desarrollo similar es el de Kourtesis et al. (2020), quienes diseñaron el Virtual Reality Everyday Assessment Lab (VR-EAL) para evaluar funciones ejecutivas. Se encontró que el resultado de la evaluación con VR era similar al de las pruebas tradicionales con lápiz y papel. Además, se propuso que el VR-EAL posee mayor validez ecológica, es de más rápida administración y resulta una experiencia más agradable para los evaluados.

Videojuegos

Las propuestas de evaluación psicológica con el uso de videojuegos son múltiples y, en muchos casos, están diseñadas como evaluaciones adaptativas estandarizadas. Los videojuegos también pueden funcionar como herramientas de evaluación invisible, además de contar con elementos que buscan que el usuario tenga una experiencia positiva, se divierta y se comprometa con la tarea.

Kiki en equilibrio (Gnosis Kids, 2022), por ejemplo, es un videojuego cuyo objetivo es evaluar funciones ejecutivas en población infantil, en específico la inhibición cognitiva y la memoria de trabajo. El desafío del juego consiste en ayudar a Kiki a ordenar sus juguetes. Esta herramienta no solo está pensada para la evaluación, sino que también tiene un potencial uso como tratamiento para entrenar el desarrollo de esas funciones. Se ha realizado un estudio de casos clínicos con un diseño pre-post en el que se observó un incremento en las habilidades de los sujetos (evaluadas con la batería Yellow-Red) luego de la indicación de utilizar el videojuego cuatro veces por semana durante un mes (M. Caputo & R. Magariños, comunicación personal, 31 de enero del 2024).

Nawaiam, por otra parte, es un videojuego cuyo objetivo es brindar información sobre candidatos a un puesto laboral (Nawaiam Ltd., 2019). El juego consiste en una aventura, ambientada en un mundo apocalíptico posterior al cambio climático, donde el jugador debe resolver, en quince minutos, diferentes tareas para salvar la mayor cantidad de vidas. Mediante las conductas de juego del evaluado (como la toma de decisiones, el tiempo y el manejo de recursos), se elabora un perfil profesional que luego se utiliza para tomar decisiones en la búsqueda de personal e, incluso, para evaluar al personal ya activo en una organización. El videojuego evalúa cuatro dimensiones principales: asertividad, sociabilidad, tolerancia y reglas. Los investigadores del proyecto SEGASPI (Serious Games para una Selección de Personal Inclusiva) analizaron las propiedades psicométricas de la evaluación de *Nawaiam* y encontraron que predecía la *performance* adaptativa con un porcentaje de varianza significativamente mayor que la obtenida por el modelo de los cinco grandes rasgos de personalidad (Ramos-Villagrasa & Fernández del Río, 2023). El juego ha tenido muy buena recepción en el ámbito aplicado, pues ya se ha utilizado en diecisiete países (H. Llovet, comunicación personal, 31 de enero del 2024), lo que va en línea con uno de los puntos señalados

por la ITC y la ATP (2022) en cuanto que este tipo de tecnologías podría facilitar las comparaciones transculturales e incrementar el alcance global de los instrumentos.

Cabe destacar también el desarrollo del videojuego *NoABS* (Bravo García & Losada, 2023), diseñado para diagnosticar vulnerabilidad y prevenir el abuso sexual infantil. El videojuego sitúa al jugador en distintos escenarios cotidianos (casa, escuela, club), en donde se plantean situaciones en las que se le solicita que elija una respuesta ("no", "no sé", "sí"). Con esas respuestas se calcula un puntaje total de grado de vulnerabilidad frente al abuso sexual de manera global y según distintas dimensiones: familiar, social, escolar y tecnológica. Al momento se han obtenido evidencias de su consistencia interna y validez convergente interna, además de que se ha analizado la distribución de los puntajes según edad y género (Losada & Gaggino, 2023).

Otra propuesta novedosa es la de Quwaider et al. (2023), quienes desarrollaron un *first person shooter* (*The Protector*) para la evaluación de los cinco grandes rasgos de la personalidad mediante las conductas dentro del juego. Los indicadores utilizados fueron variados: cantidad de disparos, elección de escenarios, customización del personaje, elección de compañeros de juego, misiones completadas, entre varios otros. Una vez obtenidos los datos, recurrieron a distintos modelos de *machine learning* para modelar las conductas observadas y clasificar a los sujetos según sus rasgos de personalidad y predecir sus conductas en futuras rondas de juego.

Por otra parte, Wang et al. (2023) desarrollaron un videojuego, al que llamaron *Fisherman*, con el objetivo de medir (y posiblemente también entrenar) funciones ejecutivas en adultos mayores. *Fisherman* consiste en pescar con una red la vida marina que aparece en burbujas. Cuenta con tres subjuegos que apuntan a una función ejecutiva particular: *Pescador cauteloso* busca medir inhibición, *Pescador ágil* pretende medir la capacidad de cambio entre tareas y *Pescador sabio* busca estimar la memoria visoespacial. Los peces aparecen aleatoriamente en las burbujas y los pescadores deben recordar dónde surgieron y el orden en el que lo hicieron. El análisis psicométrico del test incluyó el estudio de su consistencia interna, su estructura factorial y su asociación con medidas tradicionales del mismo constructo. Todos los resultados fueron satisfactorios, de modo que, como paso futuro, queda el desarrollo de los baremos.

Otros desarrollos

Javed et al. (2023) revisaron las herramientas más prometedoras para automatizar las evaluaciones de salud cognitiva y concluyeron que son las siguientes: aprendizaje supervisado y no supervisado de *machine learning*, *deep learning*, algoritmos de procesamiento natural del lenguaje y técnicas de procesamiento de imágenes. En esta área existen proyectos que, con esas herramientas, proponen identificar marcadores para realizar diagnósticos tempranos y predecir la enfermedad de Alzheimer. Otras TBA se centran

en el uso de imágenes o indicadores de actividad cerebral para realizar diagnósticos psicológicos. Se ha analizado, por ejemplo, la posibilidad de estimar un coeficiente intelectual mediante el análisis con IA de ondas cerebrales (Jahidin et al., 2015). En esta línea, Zhu et al. (2024) investigaron el uso de *machine learning* para analizar neuroimágenes estructurales y predecir síntomas de psicosis. También existen propuestas para realizar diagnósticos tempranos de depresión mediante el análisis del comportamiento en redes sociales (Cacheda et al., 2019) y de trastorno por déficit de atención a través de datos portátiles como el ritmo cardíaco y el registro de patrones de sueño (Kim et al., 2023).

CRECIENTE INCLUSIÓN DE INNOVACIONES TECNOLÓGICAS EN LA EVALUACIÓN PSICOLÓGICA: VENTAJAS Y DESVENTAJAS

La revolución digital, impulsada por la expansión de Internet y el desarrollo exponencial del *software*, ha transformado radicalmente el campo de la evaluación psicológica. La disponibilidad de herramientas digitales para la administración de pruebas *online* y retroalimentación personalizada ha ampliado significativamente las posibilidades de evaluación. La pandemia de COVID-19 y la consecuente imposibilidad de llevar a cabo evaluaciones psicológicas de manera presencial aceleraron drásticamente la adopción de herramientas digitales en la evaluación psicológica (O'Brien & McNicholas, 2020) e impulsaron la rápida adaptación de técnicas tradicionales a formatos *online* y al desarrollo de nuevas soluciones tecnológicas. Este movimiento apresurado pero necesario trajo consigo tanto beneficios como dificultades.

Entre las ventajas vinculadas a la incorporación de estas innovaciones tecnológicas podemos mencionar la eficiencia y la precisión. Dado que la administración y la puntuación se automatizan, se reduce el tiempo de evaluación y se minimiza la tasa de error humano, lo que incrementa la precisión y la confiabilidad de los resultados obtenidos (Butcher et al., 2004). Los inventarios digitalizados, por ejemplo, ofrecen la ventaja de una administración estandarizada en términos de formato y contenido de los ítems. Permiten, en general, un control automático de las respuestas, ya que se puede programar que todas estas sean de respuesta obligatoria. También previenen los inconvenientes relacionados con respuestas múltiples, ya que la evaluación puede configurarse para que solo se pueda elegir una opción entre las disponibles. Además, se pueden obtener puntuaciones automatizadas inmediatas e, incluso, en muchos casos, puntuaciones transformadas a puntajes estandarizados utilizando baremos internos ya cargados en el sistema. Esto no solo reduce los tiempos de trabajo del evaluador, sino también los inevitables errores humanos de cálculo. Además, la precisión se dará en la posibilidad de adaptar la evaluación durante el proceso de acuerdo con las respuestas que se van obteniendo del examinado. Estas evaluaciones adaptativas no solo son más precisas, sino que también suelen ser más cortas (Granziol et al., 2020).

Además, las TBA pueden ampliar el alcance y accesibilidad, dado que son flexibles en cuanto al momento y lugar de evaluación. Esto evidencia un impacto significativo en la ampliación del alcance de la evaluación psicológica a poblaciones que, debido a su ubicación geográfica o a dificultades de traslado, previamente no tenían acceso a ella (Wosik et al., 2020).

Asimismo, emergen nuevas posibilidades de evaluación: los entornos virtuales simulados, como aquellos generados en realidad virtual o mediante *serious games*, pueden resultar más realistas y atractivos y favorecer la inmersión y el compromiso con la tarea. Estas evaluaciones habilitan la valoración de indicadores que de otro modo no podrían ser evaluados: datos de proceso, tiempos de respuesta y patrones de respuesta (Zumbo et al., 2023). La complejidad en el procesamiento de los datos también puede proveer información integrada que de otra manera no hubiera sido posible analizar (López Steinmetz & Godoy, 2023). Tan es así, que las innovaciones tecnológicas posibilitan la evaluación ecológica momentánea (EMA, por las siglas en inglés de *ecological momentary assessment*), que es un método de evaluación psicológica que se caracteriza por la medición repetida de variables (como comportamientos, cogniciones y emociones) en el momento y en el lugar donde ocurren. Estas evaluaciones contribuyen a la reducción del sesgo de memoria, a la recolección de información del evento en tiempo real y al análisis de conductas o eventos poco frecuentes o imposibles de ser evaluados mediante métodos tradicionales (Wrzus & Neubauer, 2023).

Otro beneficio significativo asociado con este tipo de evaluaciones es el uso de baremos para comparar las puntuaciones brutas. El almacenamiento inmediato en bases de datos facilita la actualización periódica de estos baremos, lo que se puede realizar en plazos mucho más cortos en comparación con las evaluaciones *offline*. Esto permite la comparación con normas más precisas y actualizadas en tiempos récord, algo que no había sido posible hasta ahora. Y, finalmente, podemos destacar la mejora en la experiencia del evaluado: este tipo de tecnologías suelen contar con un desarrollo específico destinado a valorar y garantizar que la interfaz sea atractiva, amigable y fácil de usar. Sumado a ello, estas tecnologías brindan la posibilidad de retroalimentación inmediata y de ajuste personalizado al examinado (Wools et al., 2019).

Sin embargo, las TBA también pueden presentar desventajas. La amenaza a la validez resulta la principal y está dada por varios factores. Uno de ellos es el incumplimiento de los lineamientos de administración estandarizada de pruebas que no fueron diseñadas ni validadas en su formato digital (Krach et al., 2020). Además, los datos que operacionalizan los constructos pueden no ser completamente claros o fáciles de validar, lo cual impactaría directamente en la operacionalización del constructo (Wools et al., 2019). Asimismo, las TBA pueden presentar problemas de equidad, pues dependen del acceso a dispositivos digitales e internet. Algunas personas pueden no contar con

el *software* necesario, no disponer de una conexión a internet que permita una administración fluida o utilizar pantallas demasiado pequeñas para una lectura adecuada de los reactivos. Adicionalmente, una falla, ya sea del sistema o por problemas de conexión a internet, que genere una interrupción en la administración, podría derivar en datos incompletos y puntuaciones inválidas (ITC & ATP, 2022). Esto podría ocasionar problemas en el proceso de evaluación psicológica, ya que muchos test no deben administrarse nuevamente en un corto periodo de tiempo debido a los efectos de aprendizaje o la familiarización con los contenidos. Además, la administración a distancia no supervisada puede ser un desafío a las garantías de seguridad y privacidad de los datos y puede presentar una dificultad para la valoración de indicadores no verbales y el control del espacio físico en el cual el individuo está respondiendo al test (Sawyer, 2021). Todos estos aspectos han sido tomados en cuenta por la comunidad científica dedicada al avance de los desarrollos psicométricos, y en la mayoría de los casos se han establecido lineamientos y recomendaciones para abordarlos.

LINEAMIENTOS PARA EN DESARROLLO Y USO DE INNOVACIONES TECNOLÓGICAS EN LA EVALUACIÓN PSICOLÓGICA

Dados los desarrollos mencionados, sus ventajas y limitaciones, y el contexto de creciente inclusión de los mismos, diferentes publicaciones buscaron dar respuestas a la necesidad de lineamientos para el desarrollo y el uso de TBA, como Wright et al. (2020), ITC & ATP (2022), Farmer et al. (2021), Wigdorowitz et al. (2021) y Wools et al. (2019). El objetivo de estas guías radica en intentar garantizar que las evaluaciones sean válidas, confiables y justas. A continuación, se describen algunos de sus puntos centrales.

Evidencias psicométricas

Tal como ocurre con toda técnica que busque operacionalizar un constructo latente mediante un indicador empírico, las TBA deben contar con adecuadas evidencias de validez y confiabilidad. Al igual que con las técnicas tradicionales, la búsqueda de evidencias de validez puede guiarse por las cinco fuentes sugeridas por los Standards for Educational and Psychological Testing publicados por la AERA (American Educational Research Association): 1) el contenido de la prueba, 2) el proceso de respuesta, 3) la estructura interna, 4) la relación de las puntuaciones del test con otras variables y 5) las consecuencias de la evaluación.

En relación con las evidencias basadas en el contenido del test, la selección de indicadores puede seguir un enfoque orientado por la teoría (*theory-driven*, el más habitual) o un enfoque orientado por los datos (*data-driven*), también conocidos como enfoques de arriba hacia abajo (*top-down*) o de abajo hacia arriba (*bottom-up*). Los enfoques *bottom-up* pueden verse favorecidos por las TBA, ya que suelen facilitar la recolección y

análisis masivo de datos. Sin embargo, en ambos casos es fundamental contar con definiciones teóricas claras, ya sean preexistentes o emergentes del diseño, que incluyan las teorizaciones y los debates psicológicos actuales. Además, se debe garantizar que los indicadores abarquen todos los aspectos relevantes del constructo que se pretende medir y asegurar la comparabilidad de las puntuaciones entre los evaluados considerando casos como las CAT, en donde las combinaciones de estímulos que se presenten pueden diferir (Cooper, 2023). El diseño y la selección de ítems basados en tecnología (IBT) se describen con mayor detalle en el subapartado siguiente.

Asimismo, es fundamental monitorear el riesgo de que las TBA introduzcan una varianza irrelevante del constructo (Arslan et al., 2020), es decir, que las mediciones obtenidas contengan variabilidad relacionada con variables ajenas a la que se pretende medir. Un ejemplo sería la presencia de puntuaciones bajas que reflejen dificultades en el uso de la TBA, en lugar de representar niveles bajos del constructo que se está midiendo. En relación con esto, la ITC y la ATP (2022) subrayan la importancia de trabajar en colaboración con expertos en UX (experiencia de usuario) y UI (interfaz de usuario), así como el desarrollo de tutoriales tanto para evaluadores como para usuarios. En este contexto, también es crucial evaluar los procesos de respuesta y las evidencias de validez aparente. Por ello, las pruebas piloto (o *alpha* y *beta testing*, en términos técnicos) representan una valiosa fuente de información para el diseño y ajuste de las TBA.

La obtención de evidencias sobre la estructura interna de las TBA puede lograrse mediante las técnicas factoriales convencionales y el análisis individual del comportamiento de los indicadores (Cohen & Swerdlik, 2017). Una de las fuentes de evidencia de validez más accesibles al momento del diseño es la dada por la relación de los resultados obtenidos mediante las TBA y otras variables. Esos criterios externos podrán ser *gold standards* ya establecidos, a los cuales la TBA busca superar en cuanto a diseño u otra característica; podrían ser criterios que se desean predecir, otras medidas psicométricas o criterios relacionados (convergentes o discriminantes). Sobre las evidencias relacionadas a las consecuencias de la evaluación, es esencial definir claramente los alcances y limitaciones de la TBA, así como el uso adecuado de los resultados obtenidos.

Además, las TBA deberán contar con evidencias de confiabilidad para garantizar la precisión y estabilidad de las puntuaciones. Para ello, se pueden emplear las estimaciones clásicas de consistencia interna (alfas, división por mitades, formas paralelas) y test-*retests* (Cohen & Swerdlik, 2017). Sin embargo, es posible que en algunas TBA su valoración no sea factible. Por ejemplo, la ITC y la ATP (2022) sugieren que, cuando se empleen técnicas de teoría de respuesta al ítem (TRI), se reporten alternativamente las curvas de los errores estándar condicionales y la función de información del test. También puede suceder que, como con muchas técnicas clásicas, sea inviable utilizar

test-*retests* dada la familiaridad con el contenido en la segunda toma. En estos casos, se sugiere buscar evidencias psicométricas por otras vías.

Diseño universal de los test y globalización

Un aspecto clave en el diseño de estas nuevas tecnologías es el enfoque en el diseño universal de los test, que se refiere a la construcción de instrumentos accesibles para el mayor número posible de evaluados. Este enfoque considera no solo cuestiones relacionadas con discapacidades físicas o intelectuales, sino también factores como la edad, el género y la cultura. En este aspecto vuelve a cobrar relevancia el trabajo en conjunto con equipos de UX/UI.

Asimismo, las TBA tienen como fortaleza el poder ser fácilmente distribuidas por todo el mundo. Este beneficio acarrea, sin embargo, la responsabilidad de garantizar que las medidas obtenidas en otros contextos para los cuales la TBA no fue analizada psicométricamente sean válidas, confiables y justas. Tal como plantean la ITC y la ATP (2022), es fundamental asegurar que 1) se esté midiendo efectivamente el mismo constructo, 2) la TBA sea justa y libre de sesgo, 3) sea confiable y 4) sea válida para el propósito para el cual fue diseñado.

Aquí cobra relevancia la equidad intercultural. Es posible que métodos automatizados no contemplen particularidades culturales o sociodemográficas y valoren diferencialmente grupos que son distintos en cuanto a esos aspectos (Ashford et al., 2023). En otras palabras, las TBA podrían producir resultados que no reflejen adecuadamente el constructo que se pretende medir, sino diferencias provenientes de otras características del evaluado. Por ello, es fundamental considerar desde el inicio la portabilidad del constructo, es decir, la posibilidad de que la TBA sea utilizado en múltiples poblaciones. Esto implica no solo considerar diferencias lingüísticas, sino también culturales y, especialmente, diferencias en cuanto al acceso y al uso de la tecnología. En este aspecto, resultará necesario, además de recurrir a profesionales de UX y UI, contar con localizadores expertos en este tipo de procesos y obtener evidencias psicométricas específicas en las poblaciones en las que se pretende utilizar la TBA.

Ítems basados en tecnología

Los indicadores, elementos o ítems de las TBA son denominados *ítems basados en tecnología* (IBT; TEI, por las siglas en inglés de *technology-enhanced items*) y constituyen el dato que será recolectado e interpretado. Su formato puede ser muy diverso e incluir lo siguiente: respuestas a opciones múltiples o tareas de rendimiento digitalizadas; el registro del uso del cursor, incluyendo clics y tiempos de interacción (por ejemplo, mapas de calor o *heat maps*); tiempos de reacción y pausas en el uso del teclado; acciones ejecutadas en un videojuego; texto generado por el evaluado en una conversación con un

chatbot; análisis de expresiones faciales, movimientos oculares, tono de voz y contenido del habla; y datos fisiológicos como ritmo cardiaco, pulsaciones y temperatura corporal. Además, se pueden considerar registros de conductas en línea o en redes sociales, información proveniente de dispositivos digitales interconectados (internet de las cosas), así como datos sociodemográficos, médicos, psicológicos u otro tipo de información proveniente de bases de datos de centros de salud u organizaciones pertinentes, entre muchas otras opciones.

Huff y Sireci (2001) indicaron que la fortaleza principal de muchos IBT parece es el incremento del *engagement* y de la validez ecológica, ya que, en muchos casos, elementos y tareas que replican lo que ocurre en el mundo físico y esto incluso puede incrementar la validez de contenido y superar la baja representación que algunos ítems tradicionales a veces pueden tener. En cuanto a los paradatos (información adicional a las respuestas directas del evaluado), pueden ser muy útiles para evaluar la validez de las respuestas, las cuales, como en toda evaluación, pueden estar afectadas por factores como la inatención o el descuido (Cannata et al., 2022). Un ejemplo relevante es el trabajo de Pokropek et al. (2023), quienes analizaron distintas medidas generadas a partir de los movimientos del cursor durante la respuesta a un cuestionario *online*. En su estudio, encontraron una relación entre estos movimientos y los indicadores tradicionales de respuestas inválidas, y observaron que algunas medidas específicas del movimiento del cursor tenían potencial para utilizarse como indicadores de la validez de las respuestas.

Al diseñar o seleccionar IBT, el enfoque debe centrarse en el constructo que se pretende medir. Aunque los indicadores elegidos no representen de una manera tradicional el constructo a medirse, su elección debe basarse en un análisis que garantice que realmente operacionalicen el constructo (Lindner & Grieff, 2023). Este análisis puede realizarse de diversas maneras; la más viable es el contraste del funcionamiento del indicador nuevo con los indicadores clásicos que pretende reemplazar. Además, se recomienda recurrir a un juicio de expertos que permita evaluar la validez de contenido del IBT y obtener medidas del grado de acuerdo entre jueces. La elección de los indicadores debe ser precisa, dado que su formato podría introducir varianza irrelevante en la puntuación o ser fuente de sesgo en algún subgrupo de individuos específico. Este comportamiento diferencial podrá evaluarse mediante el análisis estadístico individual de cada elemento (por ejemplo, mediante el análisis del funcionamiento diferencial de cada ítem). Esto será también importante cuando los TBA sigan la lógica de las CAT y los IBT que se presenten varíen en tiempo real de acuerdo con las respuestas obtenidas del evaluado. En este punto, la variabilidad y la versatilidad de los IBT podrían tener un impacto positivo en todo el proceso de evaluación psicológica dado que se presentarían estímulos personalizados a las características del evaluado.

Como se mencionó previamente, las TBA favorecen la elección de indicadores mediante metodologías *bottom-up* o *data-driven*. Ello está alineado con el uso de análisis que recurren al *machine learning*, en los que se seleccionan los indicadores que mejor predicen determinado criterio. En estos estudios, la selección es meramente estadística y exclusivamente basada en los datos. El riesgo de recurrir únicamente a este tipo de análisis está en 1) seleccionar indicadores desconectados de las teorías psicológicas vigentes y que no sean interpretables (carencia de validez teórica), o 2) seleccionar indicadores redundantes o extremadamente simplificados que no abarcan por completo la descripción del constructo (carencia de validez de contenido). Sin embargo, este tipo de análisis puede ser una fuente valiosa de información para reconsiderar definiciones conceptuales y teóricas. De cualquier modo, un proceso iterativo entre lo empírico y lo teórico debe ser la práctica habitual. En este punto, Zumbo et al. (2023) destacaron que resulta esencial contar con un marco teórico coherente que conecte los datos con los indicadores elegidos y el constructo que se pretende medir.

Análisis masivo de datos e interpretación de resultados

En relación con el procesamiento de datos, cada vez existen más propuestas que involucran el uso de la inteligencia artificial (IA), el *deep learning* (DL) y *el machine learning* (ML) que permiten interpretar grandes volúmenes de datos útiles en los procesos de evaluación psicológica (ITC & ATP, 2022; López Steinmetz & Godoy, 2023; Pellert et al., 2024; Zhou et al., 2022). Por ejemplo, Graham et al. (2019) realizaron una revisión sobre los usos de la IA para la evaluación de salud mental y concluyeron no solo que estas técnicas podrían ser útiles para el diagnóstico, sino que también podrían hasta redefinir los diagnósticos tal como los entendemos hoy. Una de las razones principales por la que estas herramientas resultan atractivas para el área de la evaluación psicológica es su foco en un aspecto clave que comparten con ellas: la predicción de algún criterio (Iliescu et al., 2022). En la mayoría, se recurre a técnicas de aprendizaje supervisado y no supervisado, a algoritmos de procesamiento natural del lenguaje o a técnicas de procesamiento de imágenes con el fin de predecir o estimar algún tipo de criterio.

Algunas TBA presentan el desafío de ofrecer resultados interpretables a partir del análisis de una gran cantidad de datos generados por la *performance* del evaluado. Lindner y Grieff (2023) señalan que el procesamiento del *log data* (registro de todos los eventos dentro del sistema) puede llegar a ser muy complejo y demandar mucho tiempo, ya que implica preprocesar los datos y realizar múltiples análisis secuenciales para otorgarles sentido. Estos procesos requieren la combinación de técnicas de minería de datos y análisis masivo de información. Es posible que con los avances de la IA estos procesos se automaticen y se logre hacerlos más ágiles y eficientes.

Además de los tradicionales cálculos de puntuaciones compuestas, las TBA permiten otro tipo de *scoring*, denominado respuesta construida (*construct-reponse*), que requiere del uso de algoritmos más complejos. El *input* en este tipo de cálculos pueden ser datos más complejos como texto o habla. El diseño de estos sistemas de puntuación se realiza en tres pasos (ITC & ATP, 2022): primero, se normalizan las respuestas (se transforman las respuestas a un formato que pueda utilizarse en el cálculo); segundo, se extraen los indicadores que operacionalizan lo que se desea medir; y, tercero, se entrena al modelo estadístico y se elige aquel con mejor *performance*. Este proceso debe realizarse en una muestra que represente a la población objetivo y también en subgrupos con el fin de verificar la posible existencia de sesgos.

En cuanto a los resultados obtenidos mediante las TBA, la ITC y la ATP (2022) han subrayado la importancia de minimizar el riesgo de obtener puntuaciones basadas en cajas negras. Se sugiere que los resultados deriven de procesos transparentes, éticos y libres de sesgos. Además, es fundamental que estos procesos sean explícitos y evaluables, que permitan el análisis entre evaluadores y la comparación de los resultados obtenidos por la TBA y los calculados por evaluadores humanos. Esto implica "abrir la caja negra" y comprender el proceso que conduce al resultado producido por la TBA.

En relación con ello, Lindner y Grieff (2023) propusieron identificar las áreas que requieren investigación en el procesamiento de datos en evaluaciones computarizadas y el riesgo asociado con la caja negra. Su análisis concluyó que los principales desafíos son 1) la validez teórica y empírica de los indicadores, 2) el diseño de procedimientos estandarizados en la recolección y análisis de los datos, y 3) el delineamiento de estándares éticos relacionados con la recolección, el uso y la interpretación.

El énfasis recae especialmente en la necesidad de conferir sentido a los datos y asegurar que la interpretación se corresponda con algún constructo relevante, lo cual se logrará mediante un análisis teórico-empírico de los indicadores. Además, a medida que aumenta la complejidad y el caudal de parámetros recolectados, se vuelve más necesaria la base teórica que guíe su interpretación y la validación empírica. Como en todo proceso de evaluación psicológica, en general, los resultados se utilizan para tomar decisiones relacionadas con la vida de las personas (validez consecuencial), la responsabilidad asociada a ello debe estar siempre en el centro de la acción.

Administración, confidencialidad y seguridad

El último punto está relacionado con pautas sobre la administración de las TBA y las garantías de seguridad y confidencialidad de los contenidos de la TBA y los resultados obtenidos.

En cuanto a la administración, como con toda evaluación psicológica, se debería solicitar el debido consentimiento informado del evaluado y el asentimiento en el caso de menores de edad (Lindner & Grieff, 2023). Esto será especialmente importante cuando se utilicen herramientas que favorezcan la realización de evaluaciones invisibles (Rosas et al., 2015). A pesar de que el foco de ese tipo de intervenciones está en que el evaluado no experimente el estrés relacionado con estar siendo evaluado, será fundamental obtener su consentimiento o asentimiento previo. Este consentimiento debe incluir los elementos habituales de garantías de privacidad, transparencia en cuanto a los objetivos de la evaluación y el almacenamiento de los datos y el uso que se les dará. El procedimiento para obtener el consentimiento puede llevarse a cabo antes del uso de la TBA o integrarse como parte inicial de la misma. Además, es importante considerar las leyes locales en las que la evaluación se realizará, ya que pueden variar las regulaciones referidas, por ejemplo, a la transferencia de datos personales y al uso de datos biométricos. En este punto, el desarrollo de la TBA requerirá de la consulta a expertos en derecho.

Además, al igual que con las técnicas tradicionales, es importante que quienes administren TBA tengan las credenciales y el entrenamiento necesarios para la tarea, que involucrará, en este caso, el manejo de tecnología (O'Brien & McNicholas, 2020) y la búsqueda de una administración estandarizada. Se recomienda invertir en tutoriales e instrucciones claras para el uso y la administración, tanto para evaluadores como para evaluados. Las consignas deben ser claras y adaptables a los distintos tipos de dispositivos que se puedan utilizar, así como a los diferentes entornos en donde se pueda administrar la TBA (ITC & ATP, 2022). En este sentido, las TBA deberían establecer pautas de encuadre para la administración. Algunas de estas pautas serán de índole técnica (requerimientos tecnológicos necesarios en cuanto a *hardware*, *software* y conectividad, por ejemplo); otras estarán relacionadas con los ambientes en que se administre presencial o remotamente la TBA (por ejemplo, posibles distractores, compañía dentro del lugar de la administración, interacciones con otras personas, uso de otros dispositivos); y otras estarán ligadas a las conductas del evaluador (consignas, disponibilidad durante la evaluación, respuestas ante consultas del evaluado, entre otras). También se deberían establecer opciones claras en caso de que la evaluación se vea interrumpida. Las pautas de encuadre también impactarán en la seguridad de la TBA: el cuidado del contenido y de los datos obtenidos. Los sistemas de reconocimiento facial, además de poder ser una fuente útil de información para la evaluación, podrían utilizarse para la autenticación de la identidad del evaluado y para monitorear su conducta durante la administración. Se podría considerar el uso de sistemas de monitoreo remoto (como el *proctoring*) para la supervisión en vivo de la evaluación.

En cuanto al resguardo de los contenidos y los estímulos que se presenten en la TBA, debe recordarse que la American Psychiatric Association (APA, 2017) ha establecido

que los profesionales de la psicología están éticamente obligados a asegurar que los ítems de los test estén asegurados y no se divulguen. Este lineamiento involucra a todo el contenido de la TBA y a todos los actores involucrados en el desarrollo o uso de la técnica. En cuanto a la protección de los resultados y su acceso, es fundamental consultar con expertos en seguridad informática para evaluar los riesgos y proponer medidas preventivas. Será necesario implementar los protocolos contra el *hacking*, el robo de datos y otras amenazas que puedan comprometer la seguridad de las TBA y la información generada a través de su uso (ITC & ATP, 2022). Se recomienda que tanto la recolección de información como la puntuación se realice en un servidor y no en el dispositivo que esté utilizando el evaluado. Además, se sugiere la encriptación de los datos, dado que en muchos casos se trata de información altamente sensible como datos biométricos, de salud y provenientes de menores de edad.

También se deben establecer las vías de acceso a los resultados: ¿se proporcionará el resultado al evaluado, al profesional o a ambos? Asimismo, es importante considerar si es apropiado que el evaluado reciba los resultados de su evaluación tal como se obtienen mediante la TBA, o si se debiera implementar algún otro tipo de devolución. Algunas TBA ofrecen resultados inmediatamente después de haber completado la administración, pero esta no debería ser una práctica común. La entrega directa de los resultados a una persona sin la debida formación en su interpretación puede provocar impactos indeseados.

CONCLUSIONES

En este contexto, se prevé que la incorporación de innovaciones tecnológicas para la evaluación psicológica seguirá creciendo, impulsada por la disponibilidad de nuevos desarrollos específicos para este propósito. Estos avances requerirán de la formación de equipos interdisciplinarios, dado que los psicometristas no podrán dar respuesta a todos los desafíos que involucran. Los equipos de trabajo requerirán de profesionales que combinen conocimientos de psicología, psicometría, análisis de datos, diseño y arte digital, programación, experiencia del usuario, diseño de interfaz, seguridad informática, derecho y localización, entre muchas otras áreas.

Las TBA, al igual que cualquier técnica utilizada en un proceso de evaluación psicológica, deben ser consideradas como una herramienta más entre los distintos procedimientos de evaluación psicológica y nunca utilizarse como herramientas diagnósticas aisladas. La información obtenida por las TBA debe ser siempre valorada y contextualizada por un humano. En este sentido, es crucial recordar la responsabilidad de los evaluadores en cuanto al uso de TBA que cuenten con evidencias psicométricas apropiadas, al uso apropiado de los resultados obtenidos y a la ponderación con otras fuentes de información relevantes que sean parte de una batería diagnóstica (Lindner & Grieff, 2023).

La tecnología ha demostrado ser un medio útil para medir constructos que, por métodos tradicionales, resultaba demasiado complejo analizar (ITC & ATP, 2022). Por esta razón, es probable que estos nuevos desarrollos constituyan una de las vías centrales para expandir los límites actuales de la evaluación psicológica. Estas innovaciones podrían ampliar el bagaje de constructos que se pueden evaluar y, adicionalmente, facilitar y democratizar el acceso a servicios de evaluación psicológica. No obstante, dado que aún gran parte de estos desarrollos son incipientes y la experiencia con TBA, al momento, es escasa, será fundamental continuar investigando, revisando y reformulando los lineamientos aquí mencionados para garantizar el uso apropiado de las TBA.

CONFLICTO DE INTERESES

Declaración de conflicto de intereses: Guadalupe de la Iglesia es socia de la Asociación de Desarrolladores de Videojuegos Argentina (ADVA).

REFERENCIAS

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. http://www.apa.org/ethics/code/index.html

Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I., & Yan, F. (2020). The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational measurement: Issues and practice*, *39*(2), 96-106. https://doi.org/10.1111/emip.12326

Ashford, L. J., Spivak, B. L., Ogloff, J. R. P., & Shepherd, S. M. (2023). Statistical learning methods and cross-cultural fairness: Trade-offs and implications for risk assessment instruments. *Psychological Assessment*, *35*(6), 484-496. https://dx.doi.org/10.1037/pas0001228

Bravo García, L. Y., & Losada, A. V. (2023). Vulnerability to child sexual abuse: NOABS tool. *Quaderns de Psicologia*, *25*(1), Artículo e1742. https://doi.org/10.5565/rev/qpsicologia.1742

Butcher, J. N., Perry, J., & Hahn, J. (2004). Computers in clinical assessment: Historical developments, present status, and future challenges. *Journal of Clinical Psychology*, *60*(3), 331-345. https://doi.org/10.1002/jclp.10267

Cacheda, F., Fernandez, D., Novoa, F. J., & Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, *21*(6), Artículo e12554. https://doi.org/10.2196/12554

Cannata, D., Breil, S. M., Lepri, B., Back, M. D., & O'Hora, D. (2022). Toward an integrative approach to nonverbal personality detection: Connecting psychological and

artificial intelligence research. *Technology, Mind, and Behavior*, *3*(2), 1-16. https://doi.org/10.1037/tmb0000054

Chitale, V., Baghaei, N., Playne, D., Liang, H. N., Zhao, Y., Erensoy, A., & Ahmad, Y. (2022). The use of videogames and virtual reality for the assessment of anxiety and depression: A scoping review. *Games for Health Journal*, *11*(6), 341-354. https://doi.org/10.1089/g4h.2021.0227

Cohen, R. J., & Swerdlik, M. E. (2017). Psychological testing and assessment: An introduction to tests and measurement (9.ª ed.). McGraw-Hill.

Cooper, C. (2023). *An introduction to psychometrics and psychological assessment: Using, interpreting and developing tests* (2.a ed.). Routledge.

Drake, P., Hartig, J., Froitzheim, M., Mau, G., Schramm-Klein, H., & Schuhen, M. (2023). Theory-based behavioral indicators for children's purchasing self-control in a computer-based simulated supermarket. *European Journal of Psychological Assessment*, *39*(4), 289-298. https://doi.org/10.1027/1015-5759/a000757

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, *108*(8), 1277-1299. https://doi.org/10.1037/apl0001082

Farmer, R. L., McGill, R. J., Dombrowski, S. C., Benson, N. F., Smith-Kellen, S., Lockwood, A. B., Powell, S., Pynn, C., & Stinnett, T. A. (2021). Conducting psychoeducational assessments during the COVID-19 crisis: The danger of good intentions. *Contemporary School Psychology*, *25*(1), 27-32. https://doi.org/10.1007/s40688-020-00293-x

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., Spinhoven, P., & de Beurs, E. (2019). Development of a Computerized Adaptive Test for Anxiety based on the Dutch-Flemish version of the PROMIS item bank. *Assessment*, *26*(7), 1362-1374. https://doi.org/10.1177/1073191117746742

Fortea, L., Tortella-Feliu, M., Juaneda-Seguí, A., De la Peña-Arteaga, V., Chavarría-Elizondo, P., Prat-Torres, L., Soriano-Mas, C., Lane, S. P., Radua, J., & Fullana, M. A. (2023). Development and validation of a smartphone-based app for the longitudinal assessment of anxiety in daily life. *Assessment*, *30*(4), 959-968. https://doi.org/10.1177/10731911211065166

Gilbert, K., Benson, N. F., & Kranzler, J. H. (2023). What does the digital administration format of the Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V) measure? *Contemporary School Psychology*, *27*, 623-633. https://doi.org/10.1007/s40688-022-00447-z

Gnambs, T. (2022). The web-based assessment of mental speed: An experimental study of testing mode effects for the trail-making test. *European Journal of Psychological Assessment*, *39*(5), 349-353. https://doi.org/10.1027/1015-5759/a000711

Gnosis Kids. (2022). *Kiki en equilibrio* [Software].

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports*, *21*(11), Artículo 116. https://doi.org/10.1007/s11920-019-1094-0

Granziol, U., Brancaccio, A., Pizziconi, G., Spangaro, M., Gentili, F., Bosia, M., Gregori, E., Luperini, C., Pavan, C., Santarelli, V., Cavallaro, R., Cremonese, C., Favaro, A., Rossi, A., Vidotto, G., & Spoto, A. (2020). On the implementation of computerized adaptive observations for psychological assessment. *Assessment*, *29*(2), 225-241. https://doi.org/10.1177/1073191120960215

Hodent, C. (2017). The gamer's brain: How neuroscience and UX can impact video game design. Crc Press. https://doi.org/10.1201/9781315154725

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational measurement: Issues and Practice*, *20*(3), 16-25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

Hungerbuehler, I., Daley, K., Cavanagh, K., Garcia Claro, H., & Kapps, M. (2021). Chatbot-based assessment of employees' mental health: Design process and pilot implementation. *JMIR Formative Research*, *5*(4), Artículo e21678. https://doi.org/10.2196/21678

Iliescu, D., Greiff, S., Ziegler, M., & Fokkema, M. (2022). Artificial intelligence, machine learning, and other demons. *European Journal of Psychological Assessment*, *38*(3), 163-164. https://doi.org/10.1027/1015-5759/a000713

International Test Commission and Association of Test Publishers. (2022). *Guidelines for technology-based assessment*. ATP Global.

Jahidin, A. H., Taib, M. N., Tahir, N. M., & Megat Ali, M. S. A. (2015). IQ classification via brainwave features: Review on artificial intelligence techniques. *International Journal of Electrical & Computer Engineering*, *5*(1), 84-91. https://ijece.iaescore.com/index.php/IJECE/article/view/5620/4784

Javed, A. R., Saadia, A., Mughal, H., Gadekallu, T. R., Rizwan, M., Maddikunta, P. K. R., Mahmud, M., & Hussain, A. (2023). Artificial intelligence for cognitive health assessment: State-of-the-art, open challenges and future directions. *Cognitive Computation*, *15*(6), 1767-1812. https://doi.org/10.1007/s12559-023-10153-4

Kim, W. P., Kim, H. J., Pack, S. P., Lim, J. H., Cho, C. H., & Lee, H. J. (2023). Machine learning–based prediction of attention-deficit/hyperactivity disorder and sleep problems with wearable data in children. *JAMA Network Open*, *6*(3), Artículo e233502. https://doi.org/10.1001/jamanetworkopen.2023.3502

Krach, S. K., Paskiewicz, T. L., & Monk, M. M. (2020). Testing our children when the world shuts down: Analyzing recommendations for adapted tele-assessment during COVID-19. *Journal of Psychoeducational Assessment*, *38*(8), 923-941. https://doi.org/10.1177/0734282920962839

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2020). Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity. *Journal of the International Neuropsychological Society*, *27*(2), 181-196. https://doi.org/10.1017/S1355617720000764

Leong, F. T. L., Bartram, D. Cheung, F. M., Geisinger, K. F., & Iliescu, D. (2016). *The ITC international handbook of testing and assessment*. Oxford University Press.

Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment. *European Journal of Psychological Assessment*, *39*(4), 241-251. https://doi.org/10.1027/1015-5759/a000790

López Steinmetz, L. C. & Godoy, J. C. (2023). Posibles aplicaciones prácticas del uso de Machine Learning (ML) en la investigación y práctica de la clínica psicológica. *Acta Psiquiátrica y Psicológica de América Latina*, *69*(4), 266-272. https://ri.conicet.gov.ar/handle/11336/234969

Losada, A. & Gaggino, M. (2023). *Avances acerca del juego interactivo NOABS de diagnóstico de vulnerabilidad y prevención del ASI* [Presentación de escrito]. XVII Jornadas de Informática en Salud, Buenos Aires, Argentina.

Nawaiam Ltd. (2019). *Nawaiam* [Software].

O'Brien, M., & McNicholas, F. (2020). The use of telepsychiatry during COVID-19 and beyond. *Irish Journal of Psychological Medicine*, *37*(4), 250-255. https://doi.org/10.1017/ipm.2020.54

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, *19*(5), 808-826. https://doi.org/10.1177/17456916231214460

Pokropek, A., Żółtak, T., & Muszyński, M. (2023). Mouse chase. Detecting careless and unmotivated responders using cursor movements in web-based surveys. *European Journal of Psychological Assessment*, *39*(4), 299-306. https://doi.org/10.1027/1015-5759/a000758

Quwaider, M., Alabed, A., & Duwairi, R. (2023). Shooter video games for personality prediction using five factor model traits and machine learning. *Simulation Modelling Practice and Theory*, *122*, Artículo 102665. https://doi.org/10.1016/j.simpat.2022.102665

Ramos-Villagrasa, P. J., & Fernández del Río, E. (2023). Predictive validity, applicant reactions, and influence of personal characteristics of a gamefully designed assessment. *Journal of Work and Organizational Psychology*, *39*(3), 169-178. https://doi.org/10.5093/jwop2023a18

Rosas, R., Ceric, F., Aparicio, A., Arango, P., Arroyo, R., Benavente, C., Escobar, P., Olguín, P., Pizarro, M., Ramírez, M. P., Tenorio, M., & Véliz, S. (2015). ¿Pruebas Tradicionales o Evaluación Invisible a Través del Juego? Nuevas Fronteras de la Evaluación Cognitiva. *Psykhe*, *24*(1), 1-11. https://doi.org/10.7764/psykhe.24.1.724

Sawyer, S. M. (2021). Psychosocial assessments after COVID-19. *Journal of Adolescent Health*, *68*(3), 429-430. https://doi.org/10.1016/j.jadohealth.2020.12.126

Schick, A., Feine, J., Morana, S., Maedche, A., & Reininghaus, U. (2022). Validity of chatbot use for mental health assessment: Experimental study. *JMIR MHealth and UHealth, 10*(10), Artículo e28082. https://doi.org/10.2196/28082

Shahamiri, S. R., & Thabtah, F. (2020). Autism AI: A new autism screening system based on artificial intelligence. *Cognitive Computation*, *12*, 766-777. https://doi.org/10.1007/s12559-020-09743-3

Tan, Q., Cai, Y., Li, Q., Zhang, Y., & Tu, D. (2018). Development and validation of an item bank for depression screening in the Chinese population using computer adaptive testing: A Simulation study. *Frontiers in Psychology*, *9*, Artículo 1225. https://doi.org/10.3389/fpsyg.2018.01225

Veerbeek, J., & Vogelaar, B. (2023). Computerized process-oriented dynamic testing of children's ability to reason by analogy using log data. *European Journal of Psychological Assessment*, *39*(4), 263-273. https://doi.org/10.1027/1015-5759/a000749

Wang, P., Fang, Y., Qi, J.-Y., & Li, H.-J. (2023). FISHERMAN: A serious game for executive function assessment of older adults. *Assessment*, *30*(5), 1499-1513. https://doi.org/10.1177/10731911221105648

Wiebe, A., Kannen, K., Li, M., Aslan, B., Anders, D., Selaskowski, B., Ettinger, U., Philipsen, A., & Braun, N. (2023). Multimodal virtual reality-based assessment of adult ADHD: A feasibility study in healthy subjects. *Assessment*, *30*(5), 1435-1453. https://doi.org/10.1177/10731911221089193

Wigdorowitz, M., Rajab, P., Hassem, T., & Titi, N. (2021). The impact of COVID-19 on psychometric assessment across industry and academia in South Africa. *African*

Journal of Psychological Assessment, *3*, Artículo a38. https://doi.org/10.4102/ajopa.v3i0.38

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments – threats and opportunities. En B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Springer International Publishing. https://doi.org/10.1007/978-3-030-18480-3_1

Wosik, J., Fudim, M., Cameron, B., Gellad, Z. F., Cho, A., Phinney, D., Curtis, S., Roman, M., Poon, E. G., Ferranti, J., Katz, J. N., & Tcheng, J. (2020). Telehealth transformation: COVID-19 and the rise of virtual care. *Journal of the American Medical Informatics Association*, *27*(6), 957-962. https://doi.org/10.1093/jamia/ocaa067

Wright, A. J. (2020). Equivalence of remote, digital administration and traditional, in-person administration of the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V). Psychological Assessment, 32(9), 809-817. https://doi.org/10.1037/pas0000939

Wright, J., Mihura, J., Pade, H., & McCord, D. (2020). *Guidance on psychological tele-assessment during the COVID-19 crisis*. American Psychological Association. https://www.apaservices.org/practice/reimbursement/health-codes/testing/tele-assessment-covid-19

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment, 30*(3), 825-846. https://doi.org/10.1177/10731911211067538

Zhou, S., Zhao, J., & Zhang, L. (2022). Application of artificial intelligence on psychological interventions and diagnosis: An overview. *Frontiers in Psychiatry*, *13*, Artículo 811665. https://doi.org/10.3389/fpsyt.2022.811665

Zhu, Y., Maikusa, N., Radua, J., Sämann, P. G., Fusar-Poli, P., Agartz, I., Andreassen, O. A., Bachman, P., Baeza, I., Chen, X., Choi, S., Corcoran, C. M., Ebdrup, B. H., Fortea, A., Garani, R. R. G., Glenthoj, B. Y., Glenthoj, L. B., Hass, S. S., Hamilton, H. K., … & Koike, S. (2024). Using brain structural neuroimaging measures to predict psychosis onset for individuals at clinical high-risk. *Molecular Psychiatry*, *29*, 1465-1477. https://doi.org/10.1038/s41380-024-02426-7

Zieschank, K., Ireland, M. J., Day, J., & March, S. (2022). Psychometric evaluation of a new digitally animated child self-report assessment instrument: The Interactive Child Distress Screener. *Assessment*, *30*(3), 907-922. https://doi.org/10.1177/10731911211072907

Zumbo, B. D. (2023). A dialectic on validity: Explanation-focused and the many ways of being human. *International Journal of Assessment Tools in Education*, *10*, 1-96. https://doi.org/10.21449/ijate.1406304

Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and product in computer-based assessments: Clearing the ground for a holistic validity framework. European Journal of Psychological Assessment, 39(4), 252-262. https://doi.org/10.1027/1015-5759/a000748