Directrices sobre la evaluación psicológica apoyada en tecnología

Extracto de las directrices sobre la evaluación apoyada en tecnología (EAT) de la Comisión Internacional de Test y de la Asociación de Editores de Test

International Test Commission and Association of Test Publishers (2022). *Guidelines for technology-based assessment.* https://www.intestcom.org/upload/media-library/tba-guidelines-final-2-23-2023-v4-167785144642TgY.pdf*

*Directrices traducidas con el permiso de la ITC (International Test Commission)

Directrices sobre la construcción de test

Directrices sobre la planificación de una evaluación apoyada en la tecnología

1.1 La planificación de la Evaluación Apoyada en Tecnología deben incluir definiciones y descripciones sobre el uso de la tecnología y sobre su impacto en las propiedades métricas y en las características no psicométricas de los test; la planificación ha de considerar los siguientes aspectos:

(a) impacto esperado de la tecnología sobre el propósito del test.

Observaciones: Al adaptar un test del formato papel/lápiz a una Evaluación Apoyada en Tecnología (EAT), es fundamental justificar este cambio detallando motivos como: el aumento de la validez mediante una representación más fidedigna del constructo evaluado, la mejora en eficiencia (reducción del tiempo de aplicación o entrega instantánea de resultados) y la promoción de la equidad (minimizando la varianza no esencial). Estas razones deben contemplarse en todas las fases del proceso evaluativo. Asimismo, es imprescindible examinar los resultados obtenidos para asegurar que la incorporación de la tecnología no genere efectos no anticipados.

(b) cambios previstos que afectarán las propiedades psicométricas del test (por ejemplo, precisión de la medida, comparabilidad entre puntuaciones).

(c) cambios previstos en las características del test que no sean de naturaleza psicométrica (por ejemplo, reducir el coste o aumentar la accesibilidad al test).

(d) cómo afectará la tecnología a la evaluación de los constructos de interés.

Observaciones: Al emplear tecnología para perfeccionar un programa de evaluación existente es importante que los planes de adaptación indiquen claramente si el objetivo es medir los mismos constructos que se medían con el programa de evaluación anterior,

medir nuevos constructos o utilizar la tecnología para una medición más precisa del mismo constructo. Deben explicitarse los beneficios esperados, y debe planificarse un programa de investigación sobre la validez para verificar si se han logrado los resultados esperados. En el caso de que la evaluación se centre en medir nuevos constructos, se debe explicar la razón de esa inclusión (por ejemplo, la necesidad de evaluar una competencia emergente). También se debe declarar explícitamente cualquier factor que pueda impactar en la seguridad de la prueba (positiva o negativamente), ya que estos aspectos pueden afectar los argumentos de validación.

(e) costes asociados a la implementación de la tecnología frente a los beneficios esperados.

Observaciones: Estos costes abarcan la inversión inicial en tecnología, los costes derivados de la formación de los redactores de ítems y constructores de test, los costes de aumentar o cambiar el tamaño del banco de ítems, los costes de informar y formar a las personas evaluadas sobre el uso de la nueva tecnología, los costes de las licencias de uso de la tecnología y los costes de formar a otras partes interesadas, incluyendo los destinatarios finales de los resultados de los test.

(f) cómo la tecnología podría influir en la construcción/redacción de ítems.

Observaciones Los expertos en tecnologías de la información deben analizar exhaustivamente el software empleado en la creación de ítems. Deben dedicarse recursos a formar/reclutar a los redactores de ítems en el uso del nuevo método.

(g) cómo la tecnología podría influir en el diseño y en la realización de las pruebas.

Observaciones: Se debe analizar el impacto potencial que el acceso a la nueva tecnología puede tener sobre las personas evaluadas (por ejemplo, considerar las mejoras necesarias para el acceso a Internet en áreas donde la banda ancha es irregular o tiene limitaciones). Además, se deben considerar soluciones de diseño alternativas. Y, cuando proceda, también se debe describir cómo la tecnología puede facilitar el acceso a las pruebas para personas con discapacidad y para aquellas personas multilingües.

(h) cómo se espera que la tecnología afecte a la puntuación.

Observaciones: Se debe analizar el impacto que la implementación de la nueva tecnología podría tener en la obtención de las puntuaciones (por ejemplo, la corrección

automatizada de preguntas abiertas o, la existencia de múltiples respuestas posibles a un ítem de resolución de problemas). Se deben adoptar medidas para que la corrección no sea percibida como una "caja negra" (por ejemplo, los examinandos pueden realizar muchas acciones, pero no saben cómo se puntúan esas acciones). Se debe analizar y mitigar cualquier posible sesgo asociado al uso de la tecnología.

1.2 La planificación de la EAT debe incluir estudios para analizar la experiencia de las personas evaluadas con la nueva tecnología, que incluyan su facilidad de uso, eficacia y cualquier fallo o contratiempo tecnológico.

Observaciones: Se deben recoger datos para asegurar que los inconvenientes relacionados con la usabilidad o el acceso no afectaron a la realización de la prueba (por ejemplo, análisis de las tasas de omisión o fallos en la visualización de ítems de un nuevo formato). Se deben implementar estrategias para reducir la potencial desorientación que las personas evaluadas pueden experimentar al utilizar la nueva tecnología con el objetivo de minimizar la introducción de varianza irrelevante en el proceso evaluativo.

1.3 La planificación de la EAT debe identificar a los colectivos cuyo desempeño puede estar influenciado de manera desigual por el uso de la tecnología con el fin de determinar y minimizar la introducción de varianza irrelevante.

Observaciones: Es probable que la experiencia con la tecnología interactúe con la cultura, la discapacidad, el estatus socioeconómico y otras características de las personas evaluadas. Las entidades evaluadoras, ya sean organizaciones, instituciones o empresas, deben analizar la diversidad de la muestra para detectar y gestionar cualquier interacción potencial, con el objetivo de proponer soluciones adecuadas. Del mismo modo, debe incluirse el diseño universal de pruebas (UTD, Universal Test Design; un enfoque para el desarrollo de evaluaciones que intenta maximizar la accesibilidad de un test para todos sus posibles destinatarios), con el fin de eliminar barreras para personas con discapacidad o problemas de acceso.

1.4 La planificación de la EAT debe incluir la creación de tutoriales diseñados para facilitar la familiarización con los elementos del test.

Comentarios: Es importante ofrecer tutoriales para que las personas evaluadas se familiaricen con la interfaz de usuario del sistema de evaluación, con el fin de evitar

efectos no deseados sobre los resultados (Preparación, práctica y orientación de las personas evaluadas sobre el sistema de evaluación).

Directrices sobre la calidad psicométrica y técnica

Directrices sobre precisión, comparabilidad y equiparación de las puntuaciones

7.1 Se debe definir y documentar con suficiente detalle las fases relacionadas con la aplicación y estandarización de las EAT, de modo que sirvan de apoyo a las acciones destinadas a mitigar las amenazas contra la calidad de las medidas.

Observaciones: Las amenazas contra la calidad de la medida incluyen la varianza irrelevante en las puntuaciones, la infrarrepresentación del constructo o el aumento del error de medida debido al uso de la tecnología. La infraestructura de hardware y software forman parte de la estandarización de las condiciones de aplicación de las pruebas.

7.2 Se deben aportar evidencias sobre la precisión de la medida (fiabilidad) a lo largo de todo el rango de la escala, para que estas respalden los usos e interpretaciones previstos.

Observaciones: Numerosas EAT implementan tecnología adaptativa, en la cual no se estima la fiabilidad para grupos específicos de ítems o formas del test. En este contexto, las funciones de información del test y las funciones condicionales del error típico de medida resultan idóneas para informar sobre la precisión de la medida. Cuando sea posible, se debe identificar y dar cuenta de las principales fuentes del error de medida y aportar evidencias sobre la fiabilidad y precisión de la medida para subgrupos relevantes de personas evaluadas. Las estimaciones de fiabilidad tradicionales pueden utilizarse para formas de test lineales (pruebas fijas, no adaptativas), independientemente del modo de aplicación del test.

7.3 En los casos en que la EAT contemple el uso de formas múltiples de un test, se deberán utilizar métodos de equiparación adecuados para asegurar que las formas equiparadas miden el mismo constructo con un nivel comparable de dificultad y precisión.

Observaciones: Los métodos utilizados en el proceso de equiparación de puntuaciones de las EAT deben ser apropiados para la aplicación en cuestión, con relación a las declaraciones sobre la validez y con relación a al uso; por ejemplo, cuando una prueba se aplica utilizando diferentes modalidades o dispositivos o bajo distintas condiciones de aplicación; cuando se aplican dos o más formas de la misma prueba; cuando se implementa una forma adaptativa de una prueba; cuando existen diferencias de diseño que podrían afectar a los constructos o al rendimiento de las personas (diferentes tiempos, diferentes opciones de respuesta, adaptaciones, etc.); y cuando se rediseña o actualiza una prueba (cambios en el proyecto, tipos de ítems, constructo, tiempo).

7.4 Las evidencias sobre la comparabilidad de las puntuaciones deben respaldar la interpretación de las puntuaciones obtenidas por medio de diferentes tecnologías, dispositivos y condiciones de aplicación, así como, cuando sea procedente, por medio de diferentes formas e ítems.

Observaciones: Cuando el propósito de una EAT requiere la comparabilidad de las puntuaciones entre distintas versiones del test, puede ser necesaria la equiparación de las mismas. Deben aportarse evidencias de que el constructo se mide de forma comparable entre los grupos de personas evaluadas (p. ej., estudios de invarianza de medida, análisis de funcionamiento diferencial de los ítems, análisis de alineación). Factores como la tecnología, los dispositivos, la modalidad de aplicación, la plataforma y otras condiciones de la aplicación y ambientales pueden influir en el contenido, el tiempo, la representación, la respuesta y los procesos cognitivos subyacentes. Se debe analizar en qué medida estos factores afectan la comparabilidad entre puntuaciones. Las pruebas que sustentan la comparabilidad entre puntuaciones deben incluir una variedad de fuentes de evidencias de validez. El constructor del test tiene la responsabilidad de aportar evidencias que respalden cualquier argumento de comparabilidad, mientras que quienes aplican la prueba son responsables de garantizar que no se introduzcan variaciones que puedan afectarla durante o después de su aplicación.

7.5 Se deben definir las modificaciones esperadas y se debe aportar documentación que detalle el mantenimiento o mejora la calidad de las medidas.

Observaciones: En algunas circunstancias, el objetivo no es la comparabilidad entre puntuaciones y debe especificarse claramente el grado en que las variaciones afectan a la interpretación de la puntuación de la prueba.

7.6 Las evidencias psicométricas que respaldan la comparabilidad entre puntuaciones en la EAT deben incluir un análisis de las características y variaciones entre las distribuciones de las puntuaciones, así como de la fiabilidad y el error típico de medida.

Observaciones: El efecto del modo de aplicación puede estudiarse tanto a nivel del test como del ítem. Los argumentos sobre la equivalencia entre las puntuaciones podrían apoyarse en la equivalencia entre las distribuciones de las puntuaciones, la equivalencia de constructo, la equivalencia predictiva y la invarianza poblacional entre modalidades y dispositivos. En la equivalencia de constructo, el constructo entre modalidades/dispositivos sigue siendo el mismo; en la equivalencia predictiva (correlacional) las relaciones con las variables externas son similares; puede examinarse la invarianza poblacional de las funciones de equiparación entre los principales subgrupos si se dispone de muestras suficientes. Cuando se utilizan bancos de ítems, para que las puntuaciones sean intercambiables de un banco de ítems alternativo a otro, estos deben construirse para garantizar la generación de formas del test que cumplan las mismas especificaciones relacionadas con el contenido y con las características estadísticas.

7.7 La documentación que respalda la comparabilidad o equivalencia entre puntuaciones debe incluir información sobre la recogida de datos, las características de las muestras, así como sobre los métodos y análisis realizados.

Observaciones: Las recomendaciones sobre la documentación hacen referencia a la inclusión de información sobre procedimientos de recogida de datos, descripciones de las muestras, métodos y análisis realizados, así como cualquier limitación o precaución en la interpretación de los resultados.

Directrices sobre la medida del cambio y el crecimiento

7.8 Las métricas y los índices empleados para medir cambio o crecimiento deben ser fiables y válidas con relación a los fines previstos, además de contar con el respaldo de documentación pertinente.

Observaciones: Cuando se hacen inferencias sobre cambios en el rendimiento de los evaluados en el constructo o su crecimiento, es importante que dichas inferencias estén respaldadas por evidencias de fiabilidad y validez, con referencia a un modelo subyacente de progreso del aprendizaje. En ausencia de dichas evidencias o cuando las puntuaciones de crecimiento o cambio se consideren poco fiables, se deben abstener de presentar conclusiones o indicadores de crecimiento o cambio. Las interpretaciones de los índices derivados de la EAT deben sustentarse en evidencias de validez. Además, los datos procedentes de sistemas de evaluación o de cualquier otro sistema complementario empleado para respaldar inferencias sobre el crecimiento o el cambio deben ser representativos o transformarse a una escala y nivel de agregación que respalden tales inferencias.

Directrices sobre la validación de la evaluación basada en la tecnología

7.9 Se deben definir con claridad los usos y fines previstos de la EAT.

Observaciones: Los usos y finalidades previstos de las puntuaciones de un test dictan los tipos de evidencias de validez que deben recopilarse, analizarse y comunicarse para justificar el uso de un test. Por lo tanto, estos usos y finalidades deben exponerse con claridad a los aplicadores del test, a las personas que lo realizan y a todas las partes interesadas.

7.10 Se deben definir con claridad el constructo o constructos medidos por la EAT.

Observaciones: Las personas que realizan las pruebas y los destinatarios de sus resultados (por ejemplo, profesores, empleadores, investigadores, organismos de certificación, etc.) deben comprender que mide una EAT. Una definición clara del constructo medido debería incluir descripciones de los dominios cognitivos y de contenido medidos en las pruebas educativas; los dominios de conocimientos y destrezas medidos por los exámenes de acreditación, las dimensiones de la personalidad que se

miden en las evaluaciones de personalidad, actitudes que se miden en las encuestas, etc. Las especificaciones que describen estas áreas y dominios, y que sirven como definiciones operativas de los constructos medidos, deben ponerse a disposición de los evaluados y de quienes interpretan las puntuaciones de los test. Verificar que una EAT mide el constructo o constructos que pretende medir es un paso fundamental en la validación de la evaluación.

7.11 Se deben aportar evidencias de validez que respalden los usos previstos de las puntuaciones en la EAT.

Observaciones: La validación de la EAT debe comenzar con la consideración de los tipos de evidencias que confirmarían que el test: (a) mide con precisión los constructos previstos; y (b) no mide constructos no previstos. Es poco probable que un único estudio proporcione pruebas suficientes para respaldar el uso de una test para los fines previstos. Más bien, deben sintetizarse múltiples fuentes de evidencias de validez en un argumento de validez coherente que apoye el uso del test. Por ejemplo, las evaluaciones educativas, laborales, clínicas y de acreditación pueden centrarse en diferentes constructos, resultados previstos o alineación con dominios de contenido.

7.12 La validación de la EAT debe confirmar que la infraestructura necesaria para realizar el test no afecta el rendimiento de las personas evaluadas.

Observaciones: Un argumento exhaustivo de validez para la EAT debe verificar que las personas que realizan el test entienden cómo interactuar con el sistema para acceder sin problemas al test y proporcionar sus respuestas. Se debe eliminar la influencia de los conocimientos informáticos como fuente de varianza irrelevante. Además,se debe evaluar la interfaz de usuario para garantizar que no causa un estrés o una carga cognitiva indebida a las personas que realizan el test al enfrentarse a los ítems y responderlos adecuadamente.

7.13 La validación de la EAT debe tener en cuenta la diversidad de la población evaluada, así como la equidad en las interpretaciones de las puntuaciones del test entre los diferentes grupos evaluados.

Observaciones: Es probable que las personas que realizan el test difieran entre sí en muchos aspectos, como el género, la raza/cultura, el estatus socioeconómico, la discapacidad, la edad y otras características personales. Los estudios de invarianza a

nivel de ítems (por ejemplo, funcionamiento diferencial de los ítems) y a nivel de test (por ejemplo, funcionamiento diferencial de los test), así como la validez de criterio de las puntuaciones de los test (por ejemplo, validez predictiva diferencial), pueden ayudar a evaluar aspectos potenciales de sesgo e injusticia entre grupos de evaluados. Los análisis cualitativos, como los protocolos de reflexión en voz alta o las entrevistas, también pueden arrojar luz sobre la equidad y sobre cómo mejorar los programas de evaluación para que sean lo más inclusivos posible. La consideración de la diversidad y la equidad comienza en las primeras fases del desarrollo de las pruebas. La construcción de pruebas culturalmente sostenibles puede mejorar la validez de las puntuaciones garantizando que el contenido y los contextos abarquen la totalidad de la variación cultural dentro de la población evaluada.

7.14 La validación de la EAT debe garantizar que los tiempos fijados para la realización de test son claros y razonables.

Observaciones: Las personas evaluadas deben disponer de tiempo suficiente para completar todos los ítems del test y mostrar todo su potencial en relación con los constructos medidos. Si la velocidad de respuesta forma parte explícitamente del constructo medido, el test debe definir claramente como se mide este aspecto y comunicarlo claramente a las personas que realizan el test. Además, estos deben recibir instrucciones sobre cómo administrar su tiempo durante la prueba y cómo esto puede influir en sus puntuaciones. La comprensión de las reglas relacionadas con el tiempo y puntuación por parte de los evaluados puede constituir una prueba de validez importante.

7.15 Se deben realizar estudios de validación de la EAT de forma periódica para (a) confirmar que el uso del test permanece justificado y (b) perfeccionar el propio el test.

Observaciones: La validación de la EAT debe ser continua, abarcando tanto la evaluación formativa como sumativa, para adaptarse a la naturaleza dinámica de la evaluación y la población evaluada. Los estudios de validez suelen identificar fortalezas y áreas de mejora del test; estas últimas ofrecen información valiosa para su mejora. No obstante, la determinación final sobre la idoneidad del test para su propósito previsto debe basarse en evidencias actualizadas de validez. Esa determinación debe actualizarse en función de las nuevas pruebas de validez aportada.