RECURSO GRATUITO · CHECKLIST Y GUÍA

Cómo saber si un test de RRHH es fiable y válido

Por **Equipo Beetrics**, con dirección científica de Víctor Ciudad-Fernández, doctor por la Universitat de València.

Cada pocas semanas, alguien intenta venderte un test. Un cuestionario de personalidad para selección, una herramienta de colores para tus equipos, una encuesta que promete medir el compromiso. Y casi todos los folletos dicen lo mismo: "validado científicamente y altamente fiable". El problema es que esas dos palabras, fiable y válido, tienen un significado técnico muy concreto, y se usan a la ligera continuamente.

Este recurso es un checklist para distinguir la ciencia del marketing en diez minutos, sin ser psicómetra, y para llevarlo a tu próxima reunión con un proveedor. No necesitas creerte lo que te digan. Necesitas saber qué preguntar.

Fiabilidad y validez: las dos palabras que lo deciden casi todo

Son los dos pilares sobre los que se juzga cualquier instrumento que mide a personas, y están definidos en los **estándares profesionales de evaluación psicológica y educativa**, la referencia internacional en la materia. La **fiabilidad** es la consistencia: que el test mida de forma estable y no cambie de forma caprichosa. La **validez** es lo que de verdad importa para tomar una decisión: que el test mida lo que dice medir y que sirva para el uso que le vas a dar.

Un detalle que se olvida a menudo: un test puede ser muy fiable y, aun así, completamente inútil, si mide con precisión algo que no te sirve. Por eso las dos preguntas van juntas, pero es la validez la que justifica gastar dinero o tomar una decisión con el resultado.

El checklist: seis preguntas antes de comprar

Estas son las seis preguntas que recomendamos hacer ante cualquier test, evaluación o encuesta antes de pagar por él o de tomar una decisión con sus resultados. Cada una incluye qué pedir al proveedor y la señal de alarma que conviene vigilar.

1 Fiabilidad: ¿mide de forma consistente?

Un test fiable da resultados estables: si la misma persona lo repite a las pocas semanas, sale algo parecido (fiabilidad test-retest) y sus preguntas miden lo mismo de forma coherente (consistencia interna).

Qué pedir: Pide el coeficiente de fiabilidad. Como convención, en consistencia interna se suele esperar al menos en torno a 0,70, y más alto cuanto más peso tenga la decisión.

Señal de alarma: Que te vendan como virtud que "el test cambia contigo". Si el resultado baila cada vez que se hace, no estás midiendo un rasgo, estás midiendo el día que tuvo la persona.

2 Validez: ¿mide lo que dice y predice lo que importa?

La validez es que el test mida de verdad el rasgo que dice medir (validez de constructo) y que sirva para el uso que le vas a dar, por ejemplo predecir el desempeño (validez de criterio o predictiva).

Qué pedir: Pregunta contra qué criterio se ha validado y con qué resultados. La pregunta clave no es "¿describe bien a la gente?", sino "¿predice algo que te importe?".

Señal de alarma: La palabra "validado" suelta, sin decir contra qué ni con qué muestra. Y la validez aparente: que el informe suene a verdad no prueba nada, es el efecto que hace que los horóscopos parezcan acertar.

3 Baremos: ¿comparado con quién?

Una puntuación solo significa algo comparada con un grupo de referencia, el baremo. "Puntúas alto en X" depende de alto respecto a quién.

Qué pedir: Pregunta con qué población está baremado el test y de cuándo es esa referencia. Lo ideal es que sea representativa de tu contexto, idealmente España y, si se puede, tu sector.

Señal de alarma: Que no haya baremos, o que sean de otra cultura, otro idioma o de hace décadas. Estarías comparando a tu gente con una regla que no es la suya.

4 Transparencia: ¿hay manual técnico?

Las herramientas serias publican un manual técnico con su evidencia de fiabilidad y validez, y se someten a revisión independiente. La ciencia se enseña, no se esconde.

Qué pedir: Pide el manual técnico o la documentación de propiedades psicométricas. Que exista, y que puedas leerlo, ya dice mucho.

Señal de alarma: "Es nuestro algoritmo propietario, no lo compartimos". El secretismo es una bandera roja, no un signo de sofisticación.

5 Equidad: ¿funciona igual para todos?

Un test que se usa para decidir (seleccionar, promocionar) tiene que medir de forma justa entre grupos. Si penaliza de forma sistemática a unos frente a otros, además de injusto es un riesgo legal.

Qué pedir: Pregunta si se ha estudiado el sesgo y el posible impacto adverso entre grupos. Los estándares profesionales de evaluación lo exigen para los usos de decisión.

Señal de alarma: Que nadie haya mirado nunca si el test trata distinto por sexo, edad u origen. "Nunca nos lo han preguntado" no es una respuesta tranquilizadora.

6 Uso adecuado: ¿para qué sí y para qué no?

Ningún test sirve para todo. Uno puede ser una buena base para una conversación de desarrollo y, a la vez, una pésima base para decidir a quién contratar o despedir.

Qué pedir: Pregunta para qué usos está validado, en sus propias palabras, y respeta ese límite. Un buen proveedor te dirá también para qué NO conviene usarlo.

Señal de alarma: "Sirve para selección, desarrollo, equipos, liderazgo y clima". Cuando algo vale para todo, suele no valer del todo para nada.

Qué predice de verdad el desempeño (y qué no)

Si vas a usar un test para seleccionar o promocionar, la pregunta no es si el informe "describe bien" a la persona, sino si **predice** algo que te importe. Y aquí la investigación es bastante clara. La síntesis de décadas de estudios de Schmidt y Hunter (1998), matizada por la revisión más reciente de Sackett y colaboradores (2022), sitúa entre los predictores más sólidos del desempeño las **entrevistas estructuradas**, las **pruebas de trabajo** (hacer una muestra real de la tarea) y la **capacidad cognitiva**, con la **responsabilidad o concienciación** de los Cinco Grandes aportando valor por encima de lo anterior (Barrick y Mount, 1991).

En el otro extremo, los cuestionarios de "tipos" o colores apenas predicen el rendimiento real. Traducido a tu lunes: una entrevista bien estructurada y una prueba parecida al trabajo valen más que la tipología más vistosa del mercado.

> **Que a tu equipo le encante un test no lo hace válido. Lo que cuenta es si predice lo que te importa, no si gusta.**

El error que vemos una y otra vez

La trampa más común no es comprar un test malo a sabiendas, es comprarlo por la razón equivocada: porque el informe "suena a verdad" y a la gente le resulta entretenido. Eso tiene nombre, validez aparente, y es justo el efecto que hace que los horóscopos parezcan acertar: frases lo bastante generales como para que cualquiera se reconozca en ellas. Sentirse identificado no es evidencia. La evidencia es el manual técnico, los baremos y los estudios de validez. Lo vimos con detalle en el caso del test DISC, una herramienta muy popular y psicométricamente débil.

Cómo medimos en Beetrics

En Beetrics aplicamos este mismo listón a nuestro propio trabajo. Medimos con instrumentos validados y, además, añadimos una lente que la mayoría de los tests ignora: no solo cómo es cada persona de forma aislada, sino cómo está conectada en la red real de trabajo, es decir, de quién depende de verdad la organización para funcionar. Puedes ver la base científica en la que nos apoyamos en la ciencia detrás de Beetrics.

Qué hacer el lunes

- Antes de tu próxima reunión con un proveedor, lleva estas seis preguntas por escrito.
- Pide siempre el manual técnico y los datos de fiabilidad y validez. Que existan, y que te los enseñen, ya filtra a la mitad.
- Pregunta contra qué criterio se ha validado el test y con qué población está baremado.
- Separa el fin: lo que vale para una conversación de desarrollo no tiene por qué valer para decidir una contratación.
- Si vas a tomar una decisión con consecuencias, apóyate en métodos estructurados y parecidos al trabajo antes que en una tipología.

Medir bien no es un lujo académico, es lo que separa una decisión defendible de una corazonada con apariencia de dato. Si quieres profundizar, puedes seguir por la guía de rotación de personal o ver por qué el DISC se queda corto y qué usar en su lugar.

Referencias

- American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Schmidt, F. L. y Hunter, J. E. (1998). The Validity and Utility of Selection Methods in Personnel Psychology. *Psychological Bulletin*, 124(2), 262-274.
- Sackett, P. R., Zhang, C., Berry, C. M. y Lievens, F. (2022). Revisiting Meta-Analytic Estimates of Validity in Personnel Selection. *Journal of Applied Psychology*, 107(11), 2040-2068.
- Barrick, M. R. y Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology*, 44(1), 1-26.

Preguntas frecuentes

¿Qué diferencia hay entre fiabilidad y validez?

La fiabilidad es consistencia: que el test dé resultados estables y no cambie de forma caprichosa. La validez es que mida de verdad lo que dice medir y sirva para el uso que le vas a dar. Un test puede ser muy fiable y aun así inútil: una báscula descalibrada que siempre marca dos kilos de más es muy fiable, porque repite el mismo número, pero no es válida para saber tu peso. Necesitas las dos cosas, y la validez es la que de verdad justifica una decisión.

¿Qué nivel de fiabilidad debería exigir a un test?

No hay un número mágico, pero como convención habitual en consistencia interna se suele pedir al menos en torno a 0,70, y bastante más alto cuanto más importante sea la decisión que vas a tomar con el resultado. Para una conversación de desarrollo el listón es más bajo que para decidir una contratación o un despido. Lo importante es que el proveedor te dé el dato; si no lo tiene o no lo enseña, esa es ya tu respuesta.

¿Qué método predice mejor el desempeño en una selección?

La investigación acumulada (Schmidt y Hunter, 1998, matizada por la revisión de Sackett y colaboradores, 2022) apunta a las entrevistas estructuradas, las pruebas de trabajo y la capacidad cognitiva como los predictores más sólidos, con la responsabilidad o concienciación de los Cinco Grandes aportando valor adicional. Los cuestionarios de "tipos" o colores, en cambio, apenas predicen el desempeño real. La conclusión práctica: estructura tus entrevistas y apóyate en pruebas parecidas al trabajo antes que en una tipología vistosa.

¿Es legal usar tests de personalidad en selección en España?

Se pueden usar, pero con condiciones: el tratamiento de datos tiene que ser proporcionado y con una finalidad clara, informando a la persona y respetando el RGPD y los criterios de la AEPD. No somos asesores legales, así que para un caso concreto conviene validarlo con tu asesoría, pero la regla general es medir lo justo para la finalidad declarada y no acumular datos sensibles sin base. Lo desarrollamos en nuestra guía sobre cómo medir sin vigilar.

¿El DISC es fiable y válido?

Como punto de partida para una conversación de equipo puede resultar útil y a la gente suele gustarle, pero su fiabilidad y su validez predictiva son débiles, así que no es una buena base para decidir a quién contratar, promocionar o despedir. Lo analizamos en detalle, y qué usar en su lugar (Big Five y HEXACO), en un artículo aparte.
