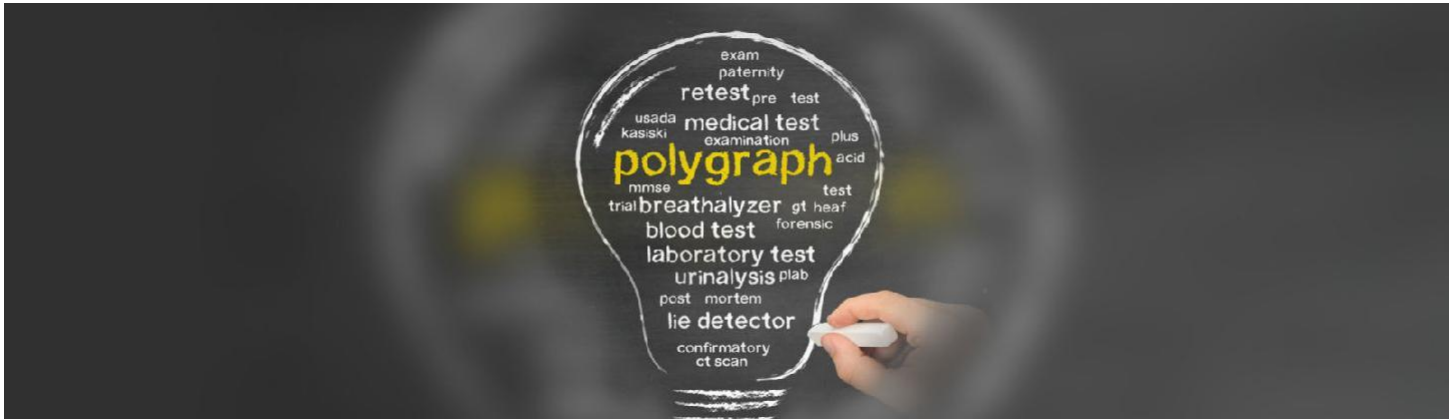




Lección de Cinco Minutos de Ciencia: ¿Qué es un Test Estandarizado?

Raymond Nelson

Una prueba es un procedimiento que se utiliza para caracterizar, cuantificar, identificar o clasificar un fenómeno de interés. Las pruebas también se utilizan para hacer predicciones – acerca de un resultado futuro o una causa pasada que se relaciona con los datos disponibles – que se pueden considerar como una forma especial de clasificación. A menudo, las pruebas se utilizan cuando la medición física directa – que requiere tanto de un fenómeno físico como de una unidad de medición definida – no es posible. Las pruebas científicas suelen hacer uso de información proxy, que tiene una relación teórica o estadística con los fenómenos de interés. Por esta razón, las pruebas científicas a menudo son intrínsecamente probabilísticas – incluso cuando los resultados de las pruebas se reducen a resultados categóricos – y no se espera que sean deterministas o infalibles.

Una característica deseable en una prueba científica es una buena validez. Es decir: los resultados son parte de la realidad de una manera útil. Pero una prueba no puede ser válida si primero no es confiable. Es decir: los resultados de las pruebas son consistentes cuando se repite la prueba o los datos de la

prueba se vuelven a evaluar. Algunas pruebas, como las pruebas de rendimiento académico que se utilizan en las escuelas, están destinadas a ser sensibles al cambio. De igual manera, algunas pruebas psicológicas tienden a ser sensibles a los cambios, como las pruebas de trastornos del estado de ánimo o del afecto. Otras pruebas psicológicas, como las de personalidad y funcionamiento intelectual, pueden estar destinadas a cuantificar características que se consideran estables. Una prueba puede ser confiable (es decir, consistente) pero no válida, pero una prueba no puede ser válida si no es confiable. Por esta razón, se dice que la confiabilidad de una prueba establece el límite superior de su validez. La reproducibilidad de los resultados de las pruebas a menudo está directamente relacionada con el grado de objetividad o subjetividad latente durante su proceso analítico. La estandarización es una estrategia importante para lograr la confiabilidad de las pruebas.

Una *prueba estandarizada* es una prueba científica – ya sea que se use en psicología, educación, industria y gobierno – que se

administra a cada individuo exactamente de la misma manera bajo exactamente las mismas circunstancias. En otras palabras, a cada persona se le hacen las mismas preguntas de la misma manera. Las pruebas estandarizadas se utilizan para medir todo, desde el funcionamiento intelectual, la aptitud, el rendimiento, la personalidad, los intereses, las características sociales, las habilidades físicas y los comportamientos sociales. Otra característica de una prueba estandarizada es que los datos de la prueba se evalúan - puntúan e interpretan - utilizando una metodología consistente para la cual se ha demostrado que los supuestos y procedimientos producen resultados con un grado satisfactorio de validez.

Las pruebas estandarizadas se utilizan hoy en día en la psicología, educación, industria y el gobierno. El primer uso de pruebas estandarizadas fue en China durante la dinastía Han (206 a.C. a 220 d.C.) con el objetivo de seleccionar personas para puestos de servicio civil basándose en el mérito en lugar de la conexión familiar. Las pruebas estandarizadas se han utilizado en las escuelas de EE. UU. desde finales del siglo XIX, a menudo para determinar la preparación o la admisión a un colegio o universidad. Las pruebas estandarizadas de inteligencia se desarrollaron por primera vez a principios del siglo XX - reemplazando intentos erróneos previos de medir la capacidad intelectual en función del tamaño físico del cráneo humano - con el fin de identificar a los niños en edad escolar que están rezagados con respecto a otros y que beneficiarse con una asistencia adicional. Las pruebas de inteligencia en adultos se hicieron más comunes con la entrada de los EE. UU. En 1917 durante la Primera Guerra Mundial, cuando un grupo de psicólogos ayudó a desarrollar las pruebas Army Alpha y Army Beta, para reclutas alfabetizados y analfabetos, que se

administrarían a más de 1,7 millones de personas en el año siguiente.

Los datos de los resultados de las pruebas individuales también se pueden comparar contra un criterio estándar. Por ejemplo: ¿un aspirante a bombero puede arrastrar un maniquí de rescate de 165 libras por 100 pies en menos de 60 segundos? O: ¿puede un aspirante a bombero completar una carrera de obstáculos cronometrada, que requiere de una combinación de fuerza, agilidad y resistencia? Los datos de los criterios se pueden acumular en grupos, lo que permite una comparación estadística. ¿Qué proporción de aspirantes a bombero puede completar con éxito las tareas de fuerza, agilidad y resistencia? Los datos de cualquier individuo también se pueden comparar contra la norma estadística. Por ejemplo: ¿qué proporción de solicitantes de bomberos requieren más o menos tiempo para completar una carrera de obstáculos?

Las normas estadísticas se pueden desarrollar en función de la teoría subyacente de una prueba. Por ejemplo: una universidad puede evaluar el sesgo del género institucional comparando el número de admisiones masculinas y femeninas contra el número de solicitudes masculinas y femeninas mediante una prueba binomial de proporciones y comparando el resultado con una distribución de referencia estadística para obtener resultados no sesgados. El resultado probabilístico se puede considerar como una densidad – la probabilidad de obtener los datos de prueba observados bajo un modelo no sesgado (nulo) – o como la probabilidad acumulada de una mayor o menor verosimilitud, que la de los datos observados. Los datos de las pruebas estandarizadas también se pueden utilizar para investigar hipótesis y desarrollar teorías.

Los datos de referencia estadística también se pueden desarrollar empíricamente agregando las puntuaciones de las pruebas de numerosos grupos de muestras utilizando el teorema del límite central (CLT). El CLT establece que las distribuciones de la muestra tendrán una repartición aproximadamente normal, independientemente de si la población está sesgada. Otro aspecto del CLT es que si los grupos de muestra se seleccionan al azar, la distribución resultante de las medias de la muestra convergerá hacia la media de la población desconocida. De esta manera podemos calcular estimaciones de parámetros estadísticos reproducibles para grandes poblaciones que no pueden ser medidas fácilmente. Por ejemplo, ¿cuál es la altura promedio de los adultos estadounidenses? Los datos de cualquier individuo se pueden comparar con la distribución de referencia empírica para responder preguntas tales como la proporción de personas que se espera que alcancen una determinada puntuación, o la proporción de personas que se espera que logren una puntuación mayor o menor que una única puntuación individual.

Las pruebas estandarizadas no son la única forma de prueba. Las pruebas *No Estandarizadas* son aquellas en las que la administración de la prueba puede variar de un individuo a otro, y aquellas para las que el análisis y la interpretación de los datos de la prueba pueden variar de una persona a otra. Las pruebas no estandarizadas se caracterizan por la ausencia de criterios objetivos y datos de referencia estadística como base para un análisis e interpretación reproducible. Los resultados de las pruebas no estandarizadas pueden estar sujetos a críticas por su ambigüedad o subjetividad, y por su relación sin sentido o desconocida con los resultados prácticos de interés

En el campo de la psicología profesional, algunas pruebas se describen como pruebas *proyectivas*, porque están diseñadas para crear un contexto, mediante el uso de estímulos y procedimientos analíticos ambiguos, para que un administrador de pruebas proyecte su propia intuición experta - que puede estar sujeta a formas múltiples de sesgo - en el análisis y el resultado. Ejemplos de pruebas proyectivas son la prueba de manchas de tinta de Rorschach, la prueba de percepción temática, las pruebas del dibujo humano y la de oraciones incompletas. Las pruebas proyectivas se consideran poco fiables y, por esta razón, tienen una validez cuestionable. Debido a que pueden obtener información matizada que puede no ser capturada por medidas psicológicas estandarizadas - que a menudo se basan en ítems FALSO/VERDADERO y escalas Likert - las pruebas proyectivas todavía se consideran un aspecto útil en una evaluación psicológica completa. Las pruebas proyectivas no deben usarse aisladamente y, cuando se utilizan, se integran a una evaluación psicológica completa que incorpora los resultados de las mediciones de pruebas objetivas, datos de la entrevista y la historia de desarrollo social de un individuo para dar un diagnóstico, narrativa y plan de tratamiento que intenta abordar una cuestión de referencia.

Otra alternativa de una prueba estandarizada se conoce como *evaluación clínica*, que se refiere al uso de un juicio experto no estructurado basado en la observación. La evaluación clínica, como la *evaluación holística*, es un proceso de síntesis e integración de varias fuentes de información administra a cada individuo exactamente de la misma manera bajo para las que no se define procedimiento estructurado alguno. La evaluación clínica tendrá una vulnerabilidad inherente hacia los sesgos, junto con las limitaciones inherentes en torno a la confiabilidad

(y, por extensión, las limitaciones inherentes a la validez). En ocasiones, se ha descubierto que las comparaciones científicas de los tamaños del efecto para los juicios expertos no estructurados no son diferentes de los juicios de los no expertos, y son superadas sistemáticamente por las soluciones estructuradas y estandarizadas.

Los desarrolladores de pruebas estandarizadas intentan eliminar las fuentes del sesgo que pueden resultar de diferencias o inconsistencias en la administración de pruebas y en el análisis de los datos pruebas. En la medida en que se basen en información proxi y no midan realmente lo que pretenden medir, siempre existe la posibilidad de que factores no intencionales influyan en el resultado de una prueba. Cuando estos otros factores se pueden explotar, existe oportunidad para falsificar. No existe ninguna prueba que sea completamente inmune a posibles falsificaciones y, por esta razón, muchas pruebas incorporan características, tanto de adquisición como de análisis de datos, para evaluar la probabilidad de falsificación. Algunas pruebas están cargadas con información verbal, de modo que la evaluación de otras habilidades puede resultar difícil si la comunicación verbal es difícil. Del mismo modo, las pruebas que se basan en información cultural pueden ser injustas para las personas inmigrantes. En los casos más extremos, las pruebas estandarizadas en el ámbito educativo a veces han sido criticadas por evaluar la capacidad de una persona para tomar pruebas estandarizadas.

Las pruebas estandarizadas se caracterizan en parte por la publicación de estándares para su administración y el análisis de datos de prueba. Otra característica de las pruebas estandarizadas es la estandarización del conocimiento, la capacitación, las habilidades y la experiencia requeridas para el administrador de la prueba. La estandarización de la

competencia profesional — mediante una educación, capacitación y certificación rigurosa — ha sido, durante la mayor parte del siglo pasado, una de las estrategias más importantes para aumentar la confiabilidad y validez de una prueba. Algunas pruebas estandarizadas pueden ser administradas por un controlador o supervisor, mientras que el análisis de datos lo realiza un experto o psicometrista.

Hoy en día, con la disponibilidad de microcomputadoras potentes y económicas prácticamente en todas partes, las inconsistencias en la administración y el análisis de datos de pruebas pueden reducirse en gran medida en el factor de la confiabilidad de la aplicación de la prueba mediante el uso de diseños de *pruebas asistidas por computadora* (CAT) en las que los estímulos de prueba e incluso las instrucciones pueden a menudo ser automatizadas. CAT también se usa ampliamente para reducir el error humano y la subjetividad en el análisis de datos de prueba. En el futuro — debido a una mayor disponibilidad de algoritmos de aprendizaje automático / algoritmos de inteligencia artificial que pueden desarrollar y probar una hipótesis rápidamente sin poca o ninguna intervención del operador — es probable que observemos un mayor uso de las *pruebas adaptada por computadora*. Por ejemplo: un algoritmo de prueba puede seleccionar preguntas de estímulo de prueba subsecuentes, basadas en respuestas pasadas que tuvieron éxito o no, ajustando el contenido para que la probabilidad de éxito sea igual o superior a un nivel del 50% (o algún otro nivel que se adapte al contexto y al objetivo de la prueba).

La idea de las pruebas estandarizadas es que les da a todos la misma oportunidad de producir un resultado satisfactorio — con el resultado práctico de que se asume que el individuo está representado razonablemente en el desarrollo

de la prueba, los datos de referencia estadística y los estándares de criterio, que nos permite asumir que las diferencias observadas en los resultados individuales son un reflejo de las diferencias individuales. Independientemente de la objetividad pretendida de las pruebas estandarizadas, el potencial de sesgo desconocido es tal que siempre habrá una discusión en curso sobre el grado en que las pruebas pueden lograr el ideal igualitario de precisión imparcial.

A medida que nos adentramos en una era en la que existe una mayor interacción entre humanos y máquinas inteligentes, también habrá una discusión importante sobre el grado en que los profesionales humanos, ya sean administradores de pruebas o desarrolladores de máquinas / algoritmos, sean considerados responsables de las conclusiones profesionales en las que pueden basarse parte de los resultados de pruebas estandarizadas. Independientemente de si las administran expertos humanos, técnicos-operadores o sistemas automatizados, las pruebas estandarizadas – incluyendo la administración y los métodos de análisis de datos de pruebas estandarizados – aseguran que los resultados sean reproducibles y confiables, lo que lleva a la capacidad de lograr una comprensión más satisfactoria de la validez de criterio y el valor práctico de los resultados de prueba observados

