436 TEXTOS DOCENTS

DISEÑO Y ANÁLISIS DE TEST Y CUESTIONARIOS

Jordi Renom Pinsach

Departamento de Psicología Social
y Psicología Cuantitativa

UNIVERSITAT DE BARCELONA

DISEÑO Y ANÁLISIS DE TEST Y CUESTIONARIOS

Jordi Renom Pinsach

Departamento de Psicología Social
y Psicología Cuantitativa

436 TEXTOS DOCENTS

UNIVERSITAT DE BARCELONA
Edicions

Índice

INTRODUCCIÓN

Esta obra va dirigida a todos aquellos interesados en el desarrollo y control de calidad de test y cuestionarios. El uso de estos instrumentos en psicología, educación y ciencias sociales se ha extendido en los últimos años. A ello han contribuido en gran parte las posibilidades derivadas de las nuevas tecnologías. Sin embargo, esta tendencia no siempre ha ido acompañada de un rigor metodológico que garantice la idoneidad de las herramientas para una correcta evaluación.

El enfoque de este manual es constructivo, pero también crítico con todos aquellos aspectos que pueden provocar un sesgo en los resultados de una evaluación. En el campo de la psicometría, se han publicado muchas obras que exponen los fundamentos de esta disciplina. Sin embargo, a nivel aplicado, no existe una regulación que vele por el seguimiento de esos principios. Por ello, el enfoque de este manual es principalmente técnico y práctico. A efectos formativos, la metodología seguida coincide con otra obra (numero 426) de esta colección, así como con el enfoque de Testing-Quest (www.testing-quest.com) y el concepto de auditoria de test (el test del test).

Para ello, los primeros tres capítulos exponen los principales conceptos empleados en la construcción y análisis de instrumentos psicométricos. Es importante familiarizarse con ellos, dado que luego aparecen en diversas ocasiones en el resto de la obra. A modo de complemento, al final se incluye un glosario con algunos de estos términos, así como una bibliografía de referencia para la consulta en profundidad de los temas tratados.

Capítulo 1

LA MEDICIÓN EN PSICOLOGÍA Y EDUCACIÓN

1.1. Definición, áreas de intervención y ventajas

La primera cuestión al plantear la medición en psicología y educación es su propia necesidad. ¿Por qué medir? ¿Es realmente necesario? La medición facilita el estudio y la evaluación de la conducta humana y mejora con ello la toma de decisiones sobre las personas. Por las características de sus profesiones, tanto psicólogos como educadores a menudo tienen que tomar decisiones acerca de clientes, pacientes, estudiantes, candidatos, etc., lo que hace necesario manejar información objetiva en que basarse. El problema no es exclusivo de estas disciplinas, pero otras ramas científicas ya lo afrontaron directamente desde sus inicios. En física, la necesidad de medir es incuestionable, lo mismo que en medicina, química, biología, etc. Es el fundamento de su propia entidad científica.

La medición en psicología, educación y ciencias sociales también se refiere a números, y en concreto a cuánto de un atributo o rasgo psicológico está presente en un sujeto. Con la medición se combate la ambigüedad y la falta de objetividad en juicios y afirmaciones difíciles de contrastar.

Pero ¿qué es medir? ¿En qué consiste? Medir es tratar de hacer comparables, mediante números, cantidades de una misma propiedad. Para ello, se establecen unidades o patrones (grado centígrado, kilogramo, metro, etc.), de forma que las relaciones entre las cantidades y sus medidas sean equivalentes. La primera acepción del diccionario de la RAE para este término utiliza la idea de unidad como un elemento clave de este concepto. Medir es «comparar una cantidad con su respectiva unidad, con el fin de averiguar cuántas veces la segunda está presente en la primera». En psicología y ciencias sociales esto no es tan fácil. ¿Existe una unidad de inteligencia? ¿Y de introversión, tolerancia a la frustración o tendencias depresivas? Por ello, una definición tradicional en este ámbito ha sido la de Nunnally, que entiende la medición como el establecimiento de un conjunto de reglas para asignar números a (cantidades de) rasgos psicológicos.

Este enfoque difiere del tradicional en otras ciencias, dado que no contempla la existencia de un patrón o unidad de medida, y aquí estriba uno de los principales retos cuando se plantea la medida en contenidos propios de estas disciplinas. Comparar, evaluar, diagnosticar, clasificar, ordenar, etc. son tareas cotidianas en psicología y educación, pero con una dificultad evidente, puesto que ¿cómo se puede comparar numéricamente la inteligencia de dos o más individuos?, ¿basándose en qué nivel de aptitud se decide el candidato más adecuado para un puesto laboral?, ¿qué alumno precisa de mayor ayuda y en qué materia?... Cuantificar todo esto no es sencillo y, de hacerlo, antes hay que aceptar algunas condiciones importantes.

Si otras ciencias tienen en la medición la base de su desarrollo, en psicología y educación también existe la necesidad de disponer de procedimientos rigurosos que aporten medidas (números) con la máxima objetividad posible en este campo. La observación, la entrevista y la evaluación mediante test y cuestionarios sirven a tal fin, y cada estrategia toma mayor relevancia según las características y el nivel de control de las condiciones que intervienen en la situación estudiada.

De acuerdo con el propósito de este manual, los siguientes apartados irán tratando los diferentes aspectos del proceso de desarrollo de un nuevo test o instrumento de medida, así como del control de calidad de uno ya existente. El objetivo es que el lector conozca los elementos que ga-

rantizan las cualidades de estos instrumentos y de las medidas que proporcionan. Para ello, y pese a sus diferencias, que se tratarán más adelante, el concepto de test, cuestionario y examen se empleará indistintamente a lo largo de esta obra, aglutinándose bajo la designación general de instrumentos o test psicométricos.

Antes de entrar en detalle sobre estos instrumentos, es conveniente conocer las áreas de intervención en que la medición aporta un mayor servicio:

- Enseñanza: para comprobar el nivel de aprendizaje de los alumnos y verificar los objetivos alcanzados. También en la acreditación/certificación de nivel de conocimientos y competencia o en pruebas de admisión. En su concepción más clásica, la evaluación mediante test o exámenes en educación ha ido muy asociada a un enfoque sumativo, si bien, como instrumentos de medida, estas herramientas también sirven en la evaluación formativa.
- Diagnóstico: identificar problemas, trastornos, detectar deficiencias o dificultades de aprendizaje. Esta es una de las áreas en donde el uso de pruebas como instrumentos de exploración precede a la intervención (tratamiento, formación, etc.) del profesional. También sirven una vez acabada esta, a fin de comprobar el cambio experimentado en los individuos en los rasgos, aptitudes y habilidades evaluados.
- Recursos humanos y organizaciones: la medición en el ámbito de las organizaciones y recursos humanos sirve para perfilar competencias, aptitudes y rasgos de personalidad, así como para evaluar la adecuación a un puesto laboral. Los cuestionarios y test también se emplean para conocer las actitudes, expectativas, creencias, etc. de los miembros de la organización.
- Orientación: las pruebas proporcionan información objetiva sobre el potencial y las motivaciones de personas en momentos clave de su vida. La orientación académica y profesional ayuda a los estudiantes y profesionales a enfocar el futuro hacia las áreas en que pueden alcanzar un mayor potencial.
- Otros ámbitos especializados: la utilización de test es frecuente en otras áreas o subáreas de las anteriores. A modo de ejemplo, en el ámbito laboral existen disciplinas como la ergonomía (estudio de la adaptación del trabajo a las capacidades y posibilidades del individuo) que a menudo utilizan cuestionarios que proporcionen datos cuantitativos objetivos sobre los trabajadores. Igual sucede en otros ámbitos recientes como la psicología del deporte.

En otro orden de cosas, las evaluaciones y acreditaciones a gran escala de tipo profesional (MIR, PIR, etc.), de idiomas (First Certificate, TOEFL, etc.) o para el manejo de vehículos se han basado tradicionalmente en test y exámenes.

Una vez perfiladas las áreas de intervención, podemos plantear cuáles son los aspectos que justifican el interés por la medición. Las ventajas pueden clasificarse en cuatro apartados:

- Objetividad y estandarización: la medición aporta objetividad. Los resultados de uno o más test administrados a un individuo informan de su perfil de personalidad o de sus aptitudes intelectuales en unas escalas normalizadas que facilitan la interpretación de su situación particular. Los test proporcionan medidas recogidas en condiciones estandarizadas por igual para todos los sujetos evaluados. Estos resultados reflejan sus diferencias individuales, y lo hacen en métricas también estandarizadas que permiten situar cada caso respecto al grupo de referencia del individuo.
- Facilidad de análisis: las herramientas de análisis de datos son cada vez más accesibles y la disponibilidad de medidas facilita su tratamiento estadístico. Por ejemplo, el impacto de una intervención psicológica o formativa puede evaluarse fácilmente comparando las puntuaciones de un cuestionario aplicado a un grupo de escolares antes y después de esta. También se pueden identificar diferencias entre grupos (por sexos), perfilar colectivos, etc. Estas

y otras muchas posibilidades aumentan en el caso de pruebas administradas en línea, donde la elaboración de la base de datos (respuestas y puntuaciones) es directa y automática. Así se agiliza el proceso de análisis, puesto que no se requieren máquinas de corrección (lectoras ópticas o escáneres) ni tabulación manual, etc.

- Comunicación: la estandarización implícita en la medida reduce uno de los problemas más frecuentes en el trabajo de psicólogos y educadores: la dificultad de comunicación. Expresiones como «poco ansioso», «bastante motivado», «moderada comprensión lectora» y «ligero retraso» pueden llegar a transmitir impresiones e interpretaciones muy diferentes entre distintos profesionales con distintos niveles de experiencia en el tema evaluado.

  Con el uso de medidas, los informes, memorias, diagnósticos y exploraciones psicológicas se expresan en términos comprensibles y asimismo interpretables para el colectivo de profesionales. En cierto modo, un perfil de personalidad puede compararse con el resultado de un hemograma o análisis de sangre. Un hemograma informa numéricamente del estado sanguíneo del individuo, es decir, proporciona medidas (indicadores) estandarizadas de diferentes atributos como son los glóbulos rojos, glóbulos blancos, plaquetas, etc. Un ejemplo equivalente sería el de un perfil de resultados de un test de personalidad, que también aporta medidas estandarizadas sobre rasgos como introversión, impulsividad, tolerancia, responsabilidad, etc.
- Economía: desde comienzos del siglo xx se constató que las medidas estandarizadas en psicología y educación ahorraban tiempo y esfuerzo, y facilitaban la comunicación entre profesionales. El uso de pruebas, en especial colectivas, agiliza cualquier proceso de evaluación. Un ejemplo simple, aunque obvia muchos aspectos, podría ser este: (A) en un proceso de evaluación de 30 candidatos que optan a un puesto de trabajo, se efectúan 30 entrevistas individuales de 30 minutos (un mínimo de 15 horas de presencia de un profesional). Una alternativa (B) a este procedimiento sería entrevistar a solo 4 candidatos (2 horas de presencia de un profesional) que antes hayan superado con el mejor resultado una batería de pruebas administrada colectivamente en solo 2 horas (2 horas de presencia de un profesional) al grupo total completo de 30 candidatos y corregida en línea (1 hora de dedicación del profesional).

  En el caso (A), tendríamos 15 horas presenciales más un tiempo, como mínimo, equivalente (15 horas) para que el profesional revise sus anotaciones y valore las entrevistas; en total, alrededor de 30 horas de dedicación. La alternativa (B) comporta 5 horas de dedicación.

  Este es un ejemplo simple y quizás reduccionista, pero que ya pone en evidencia el atractivo del uso de test en evaluaciones colectivas. Además, la ventaja aumenta muchísimo si se trata de pruebas en línea que se responden y puntúan dentro de un mismo proceso sin costes de impresión, hojas, cuadernos, etc.

En términos generales, el uso de test ofrece al profesional un ahorro de tiempo, mayor agilidad en la evaluación y garantías de objetividad (a veces aparentes). Todo ello ha generado una industria alrededor de la producción de test que respondan a las necesidades de evaluación de los profesionales. De hecho, los únicos catálogos de productos comerciales para psicólogos son de test. Este fenómeno ha llevado a considerar cada vez más la importancia de las evidencias sobre las garantías de calidad de los instrumentos.

La calidad de un test es un requisito fundamental para su uso. Todas las ventajas anteriores lo son bajo la condición de que los instrumentos posean unas garantías de calidad, es decir, que no incorporen sesgos, sean precisos y representen aquello que realmente dicen medir. Dada la importancia de este aspecto, el enfoque de esta obra se centrará tanto en la creación de test con garantías como en la verificación de la calidad de instrumentos ya en uso.

En este terreno, es importante señalar la inexistencia de una normativa de obligado cumplimiento a la hora de garantizar las cualidades de un instrumento psicométrico. Sí hay directrices,

recomendaciones y propuestas de organizaciones profesionales y académicas que velan por las buenas prácticas en la construcción y uso de test. Sin embargo, la aplicación de estas recomendaciones queda a criterio del autor del test, del canal de distribución (servicios web, editoriales, etc.) y del nivel de conocimientos del usuario final.

Con frecuencia, los cuestionarios y exámenes son un medio para otro fin en proyectos y entornos interdisciplinarios. Un profesional experto en una materia, pero ajeno a la psicología o la educación, puede crear con facilidad un cuestionario o un examen sobre dicha materia sin tener en cuenta que existe una metodología específica para desarrollarlos. Más tarde, a partir de los resultados y datos recogidos, podrá efectuar análisis a mayor nivel, establecer conclusiones y tomar decisiones sobre los individuos evaluados. El punto clave de todo este proceso consiste en la asunción de que el instrumento de medida funciona de manera correcta. En muchas ocasiones, se acepta que, por defecto y de modo natural, el simple hecho de crearlo conlleva ya unas garantías para su aplicación sin sesgos ni disfunciones. Algo parecido ocurre cuando un profesional utiliza un test ya existente asumiendo que alguien ha verificado que tiene unas mínimas garantías. En otras ocasiones, simplemente desconoce que existen una metodología y unos indicadores de calidad que informan sobre la idoneidad del test para la evaluación a que se destina.

1.2. Problemas metodológicos

Tras ver cuáles son las áreas de aplicación y los principales valores asociados a la medida, ahora el enfoque cambiará, ya que, en los test, exámenes y cuestionarios, no todo son ventajas. La medida aporta un valor siempre y cuando cumpla con una serie de condiciones que garanticen unas cualidades básicas. Igual sucede con otros procedimientos de recogida de información como la entrevista y la observación. Existen diferentes opciones a la hora de estructurar una entrevista o de sistematizar una observación. En ambos casos, en función del nivel de objetividad que se requiera, el profesional debe seguir unas pautas de actuación y de control de la situación que garanticen la representatividad y la utilidad de los datos recogidos para la evaluación.

En el caso de los test psicométricos, se combinan diversos problemas de orden metodológico que afectan en profundidad a la calidad de sus medidas. Todo usuario debería tenerlos presentes al emplearlos, puesto que ayudan a entender el motivo de los controles de calidad en un proceso de construcción y la actitud preventiva a la hora de utilizarlos. En los siguientes apartados veremos cómo la medición se enfrenta al hecho de ser relativa, probabilística e indirecta. De estos tres problemas estructurales se derivan otros dos: el sesgo y el mal uso de las medidas.

Estos cinco bloques de problemas justifican el contenido y el enfoque del resto de este manual. No se trata de problemas que se puedan resolver del todo. Tampoco son nuevos. Históricamente, empiezan con el propio origen de los test psicométricos hace más de un siglo y sus efectos siguen presentes.

El grado de afectación varía según el tipo de prueba y circunstancias de uso. Según se trate de exámenes o test psicológicos, algunos problemas ganan protagonismo. No obstante, la estrategia para afrontarlos pasa siempre por tomar conciencia de su existencia y conocer los procedimientos y directrices de construcción y de manejo que reducen en lo posible los efectos negativos.

1.2.1. Medición relativa

En la mayoría de los test psicométricos, la interpretación de la medida está referida al resultado del grupo de sujetos al que pertenece el individuo evaluado. Este grupo se toma como referencia y viene definido por características comunes como el sexo, la edad, el nivel escolar, el perfil profesional, etc. de los individuos que lo forman. Dicho de otro modo, el significado asociado (alto,

medio, bajo, etc.) a la puntuación obtenida por un individuo depende de cómo se ha distribuido el conjunto de puntuaciones del grupo al que pertenece. De este modo, la puntuación obtenida en el test se transforma en otra relativa que sitúa al individuo respecto al grupo y que será la empleada en su perfil profesional.

Este procedimiento es el más habitual en pruebas psicológicas y da nombre a una forma de entenderlas: pruebas o test de norma de grupo (TNG). En este enfoque, el grupo se utiliza como norma (grupo normativo) para calcular la puntuación final del individuo a partir de la puntuación directa original (aciertos, número de afirmaciones, etc.). Para ello, se emplean baremos o tablas de equivalencia que asocian una puntuación con la otra. Por tanto, aunque la puntuación obtenida por un sujeto en un cuestionario aporta información, esta no es suficiente para situarlo. El modelo de TNG es el más común en psicología, pero no en educación. En el ámbito educativo, y en todos aquellos en los que se realizan exámenes, no suelen emplearse grupos normativos. La puntuación de un examinado se interpreta en relación con un criterio o baremo externo establecido por un docente, comisión o programa formativo (apto, no apto, puntuación mínima que se ha de superar, equivalencia entre número de aciertos y calificación asociada, etc.). Aquí, el baremo es independiente de las puntuaciones obtenidas por el grupo de referencia.

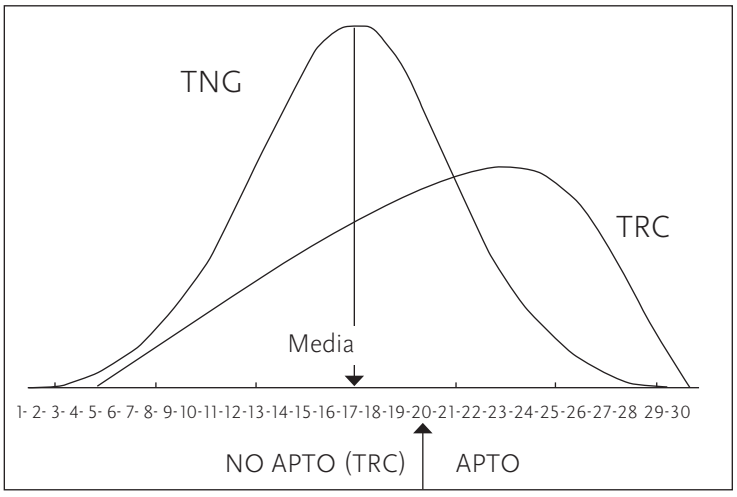
Este otro enfoque caracteriza los llamados test teferidos al criterio (TRC), en que la situación de un sujeto ya no depende de los otros, sino de un criterio externo al grupo.

Las diferencias entre los TNG y los TRC van más allá de cómo se interpretan los resultados de un test. Ambas orientaciones se distinguen por otros aspectos importantes. Los TNG no disponen de un criterio externo que determine si una puntuación es baja o alta. Su objetivo es clasificar y conocer las diferencias (distancias) entre sujetos en función de un criterio estadístico. Por ello requieren devariabilidad (varianza) en las respuestas de los sujetos y que la distribución de las puntuaciones del test se aproxime a la curva normal (se justifica en el capítulo 2).

En cuanto a los TRC, se orientan más hacia la evaluación de aprendizajes y competencias que hacia la detección de diferencias individuales. No precisan tanto la variabilidad de las respuestas ni tampoco una determinada distribución como es la curva normal.

La figura 1.1 muestra una gráfica de distribución esperada de puntuaciones totales de un grupo de sujetos obtenidas en un cuestionario tipo TNG que mide el nivel de ansiedad (a partir de la suma de respuestas sí/no a cada enunciado) y otra de un examen final, creado como TRC, de un curso de inglés (suma de aciertos). En el eje de abscisas, aparece el rango de puntuaciones directas, que en este ejemplo es común a ambos test y oscila de 1 a 30 puntos o aciertos según la prueba.

Figura 1.1. Ejemplo de distribuciones de un TNG y un TRC.



La distribución del TNG es simétrica, centrada y abarca todo el rango ajustándose al modelo de la curva normal. La del TRC tiende a concentrarse hacia la banda de puntuaciones altas y confirma que la tendencia del grupo evaluado ha sido obtener más puntuaciones altas que bajas. El objetivo del TRC es evaluar el nivel adquirido al acabar el curso y no espera ninguna distribución normal, sino una acorde con los objetivos del curso.

Para el examen de inglés, en la figura se destaca la puntuación 20 (aciertos) como punto de corte entre apto y no apto (establecido por un comité, profesor de curso, escuela de idiomas, etc.). Aquí no importa la forma de la distribución ni si el grupo tiene más o menos nivel: cada examinado queda clasificado sin necesidad de compararlo con otros. En esta situación sería aceptable, y aún más deseable, que la gráfica se desplazase hacia la derecha, pues indicaría un mayor nivel de aprendizaje tras el curso.

En el caso del TNG, no existe un comité o un experto que determine una puntuación de corte sobre la que interpretar los resultados de ansiedad. La media y la amplitud de la distribución (variabilidad) son las referencias para situar el nivel de ansiedad de los individuos evaluados.

Como conclusión principal de todo lo dicho, se deriva que en los TNG es fundamental disponer de un grupo normativo representativo (grande) del sujeto evaluado. Esto comporta una serie de requerimientos importantes a la hora de garantizar la calidad de las medidas.

Imaginemos que el cuestionario de ansiedad del ejemplo anterior se administra en la actualidad a un grupo de 1000 personas, del mismo sexo y franja de edad, y que obtenemos la gráfica A de su distribución de puntuaciones (figura 1.2). Como sucede en muchas pruebas comercializadas, imaginemos que este cuestionario lleva tiempo en uso y que se dispone del baremo de una muestra, similar a la actual, que lo respondió hace años. Las respuestas de esos sujetos produjeron la distribución B.

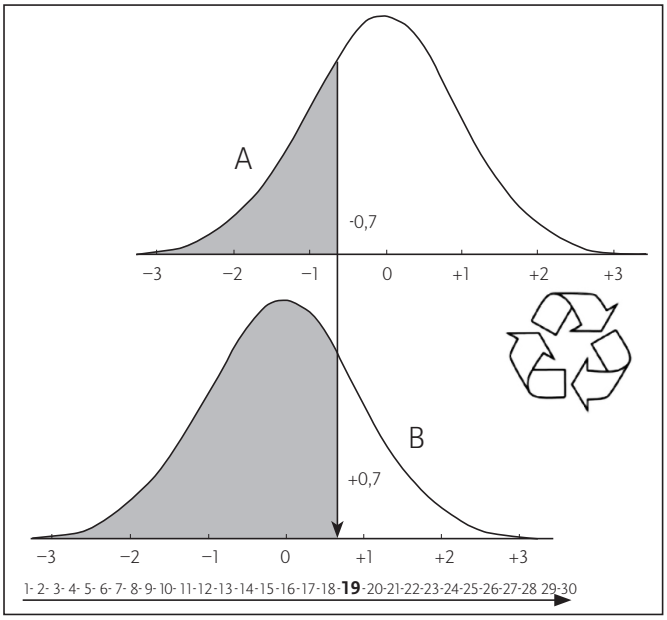
Si tomamos como referencia una puntuación directa (19 puntos), en la distribución B (antigua) corresponde a una situación por encima de la media, concretamente a un valor de +0,7 en la escala *z*, mientras que en la distribución A actual queda por debajo y equivale a $z = -0{,}7$. Esta diferencia indica que, hace años, obtener una puntuación 19 en el cuestionario se interpretaba como un nivel de ansiedad mayor que hoy en día. Quizás el cambio en el ritmo de vida actual, la sociedad, etc. ha convertido el valor 19 en menos de lo que representaba hace años. En cualquier caso, el ejemplo evidencia que la puntuación 19 por sí misma no informa con garantías si antes no se relativiza respecto al grupo de referencia.

Las gráficas A y B corresponden a dos grupos normativos y a dos baremos distintos. Una persona que puntúe actualmente 19 en el cuestionario queda situada a la baja si se la interpreta compárandola con el baremo B. De esto se deriva que los baremos de un test pueden caducar y es necesario actualizarlos para evitar sesgos. Muchas pruebas son sensibles a cambios sociales, de creencias, de costumbres, de formas de vida, etc., y requieren que los baremos se reciclen. Otras son más estables, aunque conviene verificar periódicamente que sus baremos se mantienen vigentes.

Por tanto, al decir que la medida en psicología es relativa, nos referimos a que su valor e interpretación puede variar con el tiempo, pero también en función de otros factores. La diferencia de las distribuciones A y B de la figura 1.2 podría servir también para un ejemplo entre países, culturas o características personales relevantes. Diversos test y cuestionarios comercializados en diferentes países proporcionan baremos calculados con grupos normativos de su zona geográfica y ámbito cultural. De este modo, una puntuación en un país puede no representar lo mismo en otro.

En otros casos, la distribución de puntuaciones varía en función del sexo y, en ese caso, las pruebas ofrecen baremos distintos para hombres y mujeres. Igual sucede con los baremos por franjas de edad, nivel de estudios, perfil profesional, etc.

Figura 1.2. Interpretación relativa de una misma puntuación directa.



En este punto es importante considerar que, cuando se combinan varias características, se multiplica la cantidad de baremos necesarios. Por ejemplo, si se tienen en cuenta la edad y el sexo, habrá que elaborar baremos de hombres de 20 a 30 años y de mujeres de 20 a 30 años, de hombres de 30 a 40 años y de mujeres de 30 a 40 años, etc. Esta combinación mejora la interpretación de las puntuaciones, pero comercialmente también encarece el test, ya que el número de grupos normativos necesarios se multiplica.

En psicometría, no se suele trabajar con muestras pequeñas. Los análisis requieren de muchos sujetos, y más cuando se trata de baremos con centenares de individuos en cada uno. En la práctica, esto se traduce en costes, ya que es preciso muestrear y gestionar la participación de grandes colectivos que han de aportar su colaboración en condiciones controladas.

A partir de estos ejemplos, es evidente la importancia de emplear baremos ajustados al momento temporal, país y perfil personal del sujeto evaluado. De lo contrario, el usuario asume un alto riesgo de sesgo en las evaluaciones y esto no es exclusivo de la psicometría. En medicina, las evaluaciones que hace un pediatra en el seguimiento del peso y talla de los recién nacidos se basan en baremos (curvas evolutivas) calculados para su zona geográfica de referencia. Aplicar a niños escandinavos curvas evolutivas obtenidas en poblaciones del Sudeste asiático sería otra forma de sesgo. Antropométricamente, las tallas y pesos medios de los bebés evolucionan mes a mes de manera diferente. El mismo ejemplo valdría si el pediatra emplea en 2021 curvas de desarrollo de 1990.

En conclusión, al decir que la medida es relativa se acepta que no es invariante. Si trasladamos el ejemplo a la física, rayaría en lo absurdo, pues sería como decir que un kilogramo de peso no representa el mismo peso en Finlandia que en Uruguay.

1.2.2. Medición probabilística

En toda medida existe siempre un componente de error cuya distribución (oscilación) y características hay que conocer. Esto ocurre con cualquier instrumento de medida.

Una cinta métrica no metálica suele tener cierta elasticidad y, al medir diversas veces, una misma distancia puede dar valores que oscilen en un pequeño intervalo. Debido al uso y desgas-

te, la cinta proporciona lecturas que no acaban de coincidir exactamente en un mismo valor. Algo similar ocurriría con una báscula mecánica. En ambos casos se trata del problema de la precisión y de cómo una medida obtenida no debe entenderse de entrada como un valor exacto.

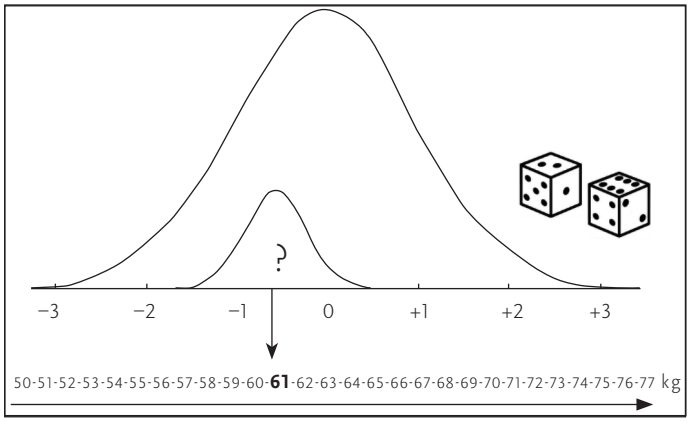
Supongamos que se toman las medidas de peso en kilogramos de un grupo de 500 adultos con un rango de valores entre 50 y 77 kg. Al trazar la distribución de los valores obtenidos (figura 1.3), esta se asemeja a la curva normal con un peso medio global cercano a 63 kg. Conscientes de que la báscula tiene problemas en su mecanismo, podríamos comprobar su nivel de imprecisión repitiendo las medidas para un sujeto concreto que tiene un peso por debajo de la media. En caso de pesar 100 veces a esta persona con la misma báscula, se generaría una distribución de nuevos resultados con un valor medio individual de 63 kg. Esta nueva distribución se denomina distribución del error de medida, y en el gráfico de la figura 1.3 está inscrita dentro de la distribución general obtenida de toda la muestra.

Para esta persona, la báscula ha dado en ocasiones valores hacia la banda alta y baja de peso, pero la mayoría de las 100 medidas repetidas se han concentrado cerca de 61 kg (su media).

Aunque forzado y simple, este ejemplo ayuda a tomar conciencia del intervalo de incerteza de este instrumento, de su imprecisión y de algunas repercusiones para los instrumentos psicométricos. De entrada, podemos suponer que el peso de esa persona es 61 kg, pero no asegurarlo al cien por cien. Lo único seguro es que el peso verdadero se encuentra dentro de un intervalo cuyo centro es 61 kg, y que, cuanto más estrecho sea el intervalo, mejor. Si esta báscula funcionara peor (con menos precisión), el intervalo de incerteza sería más amplio.

En el ejemplo, también se asume que el error de medida de la balanza es aleatorio. Dicho de otro modo, que la tendencia a apartarse del valor central es simétrica hacia ambos lados y constante para cualquier peso medido. De esta forma, la misma distribución del error valdría para personas más o menos pesadas.

Figura 1.3. Distribución del error de las medidas repetidas de peso.



Si esto sucede en un ejemplo en el plano físico, la situación se agrava con los instrumentos psicométricos. Las puntuaciones de un test o cuestionario nunca son valores exactos, sino estimaciones puntuales del verdadero valor de la medida.

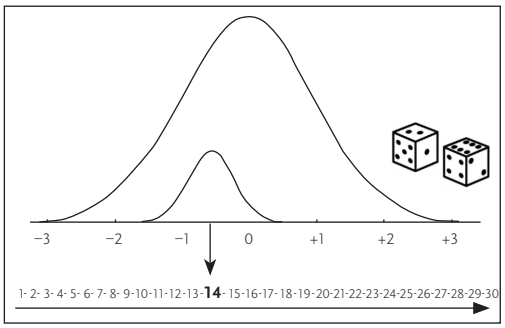
Si hacemos una traslación de lo que sucede en la báscula, el ejemplo es plenamente válido para un cuestionario que mide la ansiedad (figura 1.4). En la figura 1.3, el eje de abscisas muestra una doble escala, la del peso y su tipificación (escala *z*). En la figura 1.4, ya no aparece la escala física (kg), sino la de puntos de 1 a 30 obtenidos en el cuestionario. Esta escala no dispone de una unidad de medida física, sino de simples puntuaciones que supuestamente representan diferentes niveles de ansiedad de las personas, pero de las que se desconoce si existe un patrón o unidad (fal-

ta el kg). En esta situación, que es la habitual en los test psicométricos, la unidad o patrón es estadística y viene dada por la escala *z*. Este cambio se verá con más detalle en el siguiente capítulo.

Haciendo una comparación con el ejemplo anterior, la figura 1.4 refleja cómo se distribuyen las puntuaciones de un grupo de personas que han respondido al test (grupo normativo) y la distribución del error de medida para el caso concreto de una que hubiera respondido la prueba diversas veces. Tras responder repetidas veces, parece que esta persona tiene un promedio de 14 puntos, pero este dato no es más que el valor central de un intervalo en que ha obtenido puntuaciones superiores e inferiores. Es probable que la puntuación verdadera esté en ese intervalo, pero no podemos asegurar cuál es en concreto.

Todos los resultados de pruebas psicométricas son valores aproximados y no exactos. Además, en el ejemplo de peso se repiten 100 veces las mediciones, pero esto es inviable en un test. Por tanto, la estimación del intervalo de error es menos consistente y suele proporcionar amplitudes más grandes.

Figura 1.4. Distribución del error de las puntuaciones repetidas de un test.



1.2.3. Medición indirecta

Un tercer problema metodológico en los test psicométricos es que la medición refleja indirectamente aquello que pretende expresar. Las pruebas no se enfrentan a magnitudes físicas, sino a construcciones teóricas (constructos), a conceptos que intentan representar aspectos del comportamiento humano, de las actitudes, de la personalidad de los individuos, etc. Las pruebas pretenden medir características psicológicas, conocimientos, habilidades, etc. muchas veces difíciles de concretar y con indicadores poco definidos. Los test no se centran en características antropométricas que disponen de unidades como el kilogramo, el centímetro o el grado centígrado. Al decir que la medición es indirecta nos referimos a la representatividad de los indicadores que reflejan las características de la personalidad, de las competencias y capacidades, de las actitudes, etc. que se quiere medir.

¿Qué indicadores reflejan lo ansiosa que es una persona? ¿Y su capacidad de tolerancia a la frustración o de razonamiento espacial? ¿Un cuestionario con frases que plantean lo introvertida y reservada que es una persona acaba proporcionando una medida que expresa en qué grado lo es realmente? ¿Un examen teórico para obtener el carné de conducir informa del nivel de preparación del futuro conductor? ¿Qué preguntas se pueden hacer a alguien para determinar su nivel de resiliencia o su actitud emprendedora?

Todas estas preguntas reflejan dudas sistemáticas sobre lo que se denomina la validez de las puntuaciones del test. A menudo existe incerteza sobre lo que realmente reflejan las puntuaciones de un cuestionario o de un examen. Desde sus inicios, el hecho de garantizar su validez ha sido el principal reto de los test.

Si tomamos como ejemplo un cuestionario sobre introversión, puede haber confusión con otros conceptos cercanos como son las habilidades sociales del individuo. Las respuestas a los