TEMPUS
Psicológico

Artículo de Investigación

Desarrollo de instrumentos de evaluación psicológica y educativa desde el modelo B.E.A.R (Berkeley Evaluation and Assessment Research)

Development of psychological and educational assessment instruments based on the B.E.A.R. (Berkeley Evaluation and Assessment Research) model

Claudia Patricia Ovalle Ramírez[1]

Recibido: 26/08/2025 - Aprobado: 20/10/2025 - Publicado: 15/12/2025

Para citar este artículo
Ovalle, C (2025). Desarrollo de Instrumentos de Evaluación Psicológica y Educativa desde el modelo B.E.A.R (Berkeley Evaluation and Assessment Research). Tempus Psicológico, 9(1) - ISSN: 2619-6336 - DOI: https://doi.org/10.30554/tempuspsi.9.1.5419.2026

1 Universidad de Antioquia. Orcid: orcid.org/0000-0002-3664-7290
Correo electrónico: claudia.ovalle@udea.edu.co

DOI: https://doi.org/10.30554/tempuspsi.9.1.5419.2026

Resumen

Se presenta el modelo BEAR (Berkeley Evaluation and Assessment Research) de diseño de instrumentos de medición y sus componentes con base en la lectura de la segunda edición del libro de Wilson y Tan, de la Universidad de Berkley. Se enfatiza el modelo por sus facilidades al unir los conceptos teóricos con el modelo psicométrico de medición, y las posibilidades que representa para el diseño de instrumentos apropiados, generando inferencias correctas cuando se trata de recolectar evidencia sobre los rasgos psicológicos, comportamentales, actitudinales y de habilidad. Se discuten los pasos del proceso y se presentan interrogantes sobre las potencialidades del modelo fuera del supuesto de la Unidimensionalidad.

Palabras clave: *psicometría, modelo BEAR, mapa de Wright, IRT, modelo de Rasch, construcción de instrumentos, independencia local.*

Abstract

The BEAR (Berkeley Evaluation and Assessment Research) model for the design of measuring instruments and their components is presented based on the reading of the second edition of the book by Wilson & Tan, from the University of Berkley. The model is emphasized for its facilities in uniting theoretical concepts with the psychometric measurement model, and the possibilities it represents for the design of appropriate instruments, and therefore correct inferences when it comes to collecting evidence on psychological, behavioral, attitudinal and ability traits. The steps of the process are discussed, and questions are presented about the potentialities of the model outside the assumption of One-dimensionality.

Keywords: *psychometrics, BEAR model, Wright map, IRT, Rasch model, instrument construction, local independence.*

DOI: https://doi.org/10.30554/tempuspsi.9.1.5419.2026

Introducción

Los modelos de medición psicométricos implican el desarrollo de un proceso para lograr producir medidas de diferentes aspectos psicológicos. Las actitudes, las habilidades, la inteligencia son sólo algunos de los rasgos que pueden ser medidos. La elaboración real del instrumento "seguiría un orden, desde una idea inicial sobre la propiedad que se desea medir hasta la recopilación de evidencia que demuestre que el instrumento puede utilizarse con éxito para medir dicha propiedad" (Wilson & Tan, 2023, p.1). Para lograr completar este proceso, Wilson y Tan (2023) proponen que se consideren los siguientes 4 "ladrillos de construcción": el mapa de constructo, el plan de diseño de los ítems, el espacio de resultados y el modelo estadístico de medición. Este proceso se usa para el desarrollo de instrumentos como escalas psicológicas, pruebas de rendimiento, cuestionarios y listas de verificación conductual.

A continuación, se esboza un resumen del proceso de medición propuesto en el libro de Wilson y Tan en su segunda edición (2023), que puede ser útil para el desarrollo de instrumentos de medición, y que es el producto de una experiencia de casi 20 años, ya que la primera versión se editó en 2005.

Inicialmente, Wilson y Tan (2005; 2023) definen la medición como un proceso empírico e informacional, diseñado con un propósito, cuyo insumo es una propiedad empírica de un objeto y que produce información en forma de valores de esa propiedad (Mari et al., 2023, p. 25). Es importante resaltar que la medida implica que más allá de los observables (i.e., ítems), lo que se está evaluando es un constructo latente, sobre el cual el evaluador es responsable de encontrar evidencia por medio de ítems bien construidos, que superen las dificultades del error aleatorio (i.e., debido a las condiciones de aplicación) y el error sistemático (i.e., debido a sesgos en la medición) que afectan la confiabilidad de la medida.

Según Wilson y Tan:

> El enfoque adoptado aquí se basa en la idea de que existe un único atributo subyacente que el instrumento está diseñado para medir. La palabra instrumento se define como una técnica para relacionar algo que observamos en el mundo real (a veces denominado 'manifiesto' u 'observado') con un atributo que estamos midiendo y que existe únicamente como parte de una teoría (a veces denominado 'latente' o 'no observado'). Esta definición es algo más amplia que el uso típico, que se centra en la manifestación más concreta del instrumento—los ítems o preguntas. Se ha elegido esta definición más amplia para revelar los aspectos menos evidentes de la medición. (2023, p. 30)

En general, el enfoque presentado por Wilson y Tan (2023) puede considerarse una manifestación del enfoque sociocognitivo de la medición humana propuesto por Mislevy (2018), así como un ejemplo del Diseño de Evaluaciones Fundamentado (Principled Assessment Design) planteado por diversos autores (Ferrara et al., 2016; Nichols et al., 2016; Wilson & Tan, 2023). Wilson y Tan (2023) aclaran que los procedimientos descritos en el modelo BEAR no son la única forma de realizar mediciones, pues existen otros enfoques; sin embargo, su valor práctico y su sencillez pueden asegurar mejor la evidencia recolectada por el evaluador.

La aproximación adoptada por Wilson y Tan se conoce como “Modelo de constructo”, y es también aplicada por entidades como la US National Research Council (NRC, 2001). La NRC emplea el modelo del triángulo de medición (Cognición, Observación e Interpretación), el cual consiste en una teoría (la concepción o teoría sobre cómo aprenden las personas, qué saben las personas y cómo el conocimiento y la comprensión progresan a lo largo del tiempo), una tarea o medida (qué tipos de observaciones o tareas son más propensas a provocar demostraciones de conocimientos y habilidades); y por último, los supuestos (suposiciones sobre la mejor manera de interpretar la evidencia de las observaciones para hacer inferencias significativas sobre lo que los evaluados saben y pueden hacer).

Esta aproximación del triángulo de evaluación es adaptado y convertido en un sistema de medición (B.E.A.R Assessment System) (Wilson & Sloane, 2000), el cual cuenta con software propio; sin embargo, puede usarse R para hacer las mismas estimaciones de modelo de medición (Rasch) que este modelo emplea.

1. El Modelo B.E.A.R de construcción de instrumentos

El modelo BEAR (Berkeley Evaluation and Assessment Research) se apoya en cuatro “bloques de construcción” para abordar elementos del Triángulo del NRC (2001): el mapa de constructo, el diseño de ítems, el espacio de resultados y el mapa de Wright. El mapa de constructo es el principio de Cognición (teoría) del triángulo, el diseño de los ítems es el plan para llevar a cabo la Observación (tarea/ medida), y el espacio de resultado y el mapa de Wright facilitan la Interpretación (inferencia).

1.1. El mapa de constructo (¿Cómo se describirá el atributo?)

El mapa de constructo es, en sentido estricto, una representación espacial, ordenada, y de niveles progresivos que muestran las características de un constructo (i.e., “extraversión”, “necesidad de logro”, “razonamiento numérico”, etc.) que debe ser

medido. Para el desarrollo de este mapa, es necesario una definición de constructo bien elaborada por medio del análisis de instrumentos previos y la teoría existente sobre el constructo de interés.

El desarrollo del mapa de constructo según Wilson y Tan, conlleva los siguientes pasos:

> Primero, suponemos que el constructo que deseamos medir tiene una forma particularmente simple: se extiende de un extremo al otro del constructo —por ejemplo, de alto a bajo, de pequeño a grande, de positivo a negativo o de fuerte a débil—. La segunda suposición es que existen puntos cualitativos consecutivos y distinguibles entre esos extremos. Con frecuencia, el constructo se conceptualiza como la descripción de puntos sucesivos en un proceso de cambio, y el mapa de constructo puede considerarse análogo a una 'hoja de ruta' cualitativa de ese cambio a lo largo del constructo. En reconocimiento de esta analogía, estas ubicaciones cualitativamente diferentes a lo largo del constructo se denominarán 'puntos de referencia' (waypoints) —y serán muy importantes y útiles para la interpretación—. Cada punto de referencia tiene una descripción cualitativa por sí mismo, pero, además, adquiere significado en referencia a los puntos anteriores y los que están por encima de él. Tercero, asumimos que los respondientes pueden (en teoría) ubicarse en cualquier punto intermedio entre esos 'puntos de referencia'; es decir, que el constructo subyacente es denso en un sentido conceptual. (2023, p.9)

Ejemplos comunes de lo que representan los waypoints (o puntos de referencia) son la escala de notas escolares (de la A+ a la F, en el sistema americano), y la taxonomía de Bloom, que es una jerarquía de objetivos cognitivos (Bloom, et al., 1956) y afectivos (Krathwohl et al., 1964): por ejemplo, las categorías ordenadas de la taxonomía incluyen de manera ascendente en complejidad cognitiva: recordar, comprender, aplicar, analizar, evaluar y crear.

Aunque un constructo latente no cuente con una medida desarrollada, cuenta con locaciones particulares (los puntos de referencia) que deben ser derivados del contenido o de la teoría del constructo; es decir, el mapa de constructo es una primera aproximación a la consolidación de una escala de medición. El mapa de constructo consiste en una lista ordenada de puntos de referencia que los evaluados pueden alcanzar mientras progresan en una serie de conocimientos, actitudes o comportamientos (Wilson & Tan, 2023).

Considérese el siguiente ejemplo de un mapa de constructo: la distribución de mediciones repetidas de un atributo de un objeto podría concebirse como la combinación de una cantidad fija del atributo de un respondiente (que podría considerarse la “cantidad verdadera”), y uno o más componentes de error aleatorio en el proceso de medición. Por ejemplo, si se pide a equipos de estudiantes que midan la “envergadura” de su profesor (es decir, la anchura de su alcance cuando extiende ambos brazos), las fuentes de aleatoriedad podrían ser las siguientes:

- Las “brechas” que se producen cuando los estudiantes desplazan sus reglas por la espalda del profesor.
- Los “solapes” que ocurren cuando el extremo de una medición se solapa con el punto de partida de la siguiente.
- La “flacidez” que se produce cuando el profesor se cansa y sus brazos extendidos se hunden.

Los estudiantes pueden ser evaluados en una escala de “capacidad para medir la envergadura de los brazos”, en la cual progresarían desde un nivel bajo del constructo (el punto de referencia 1: descartar la posibilidad del error), a niveles intermedios (puntos de referencia 2 y 3: establecer una o más fuentes de error) y a niveles superiores (punto de referencia 4: no sólo consideran estos efectos aleatorios o de error, sino que también los modelan mediante una representación virtual en computadora).

1.2. El diseño de los ítems

Los formatos más comunes para el diseño de ítems son el de opción múlti- ple, utilizado en pruebas de rendimiento, y el formato tipo Likert de encuestas y escalas de actitud (por ejemplo, con respuestas que van desde “totalmente de acuerdo” hasta “totalmente en desacuerdo”). Ambos son ejemplos del tipo de ítem de “respuesta seleccionada”, en el que al respondiente se le ofrece únicamente un rango limitado de posibles respuestas, y se ve obligado a elegir entre ellas. Existen muchas variantes de este formato, que van desde pregun- tas en cuestionarios hasta la observación de indicadores de conductas (e.g. las clasificaciones de productos por parte de consumidores). Otros tipos de ítems permiten que el respondiente puede generar una “respuesta construida” como un ensayo, una entrevista, una presentación, o una evidencia (por ejemplo, un clavado competitivo, un recital de piano o un experimento científico). Gene- ralmente, estas respuestas se evaluarán mediante una guía de puntuación o rúbrica que funciona de forma similar a como lo hace un mapa de constructo; es decir, a los puntos de referencia del mapa de

constructo se le asigna una puntuación (que podrá considerarse como “pública”) y que da cuenta del nivel del constructo que el evaluado tiene.

1.3. El espacio de resultados

El evaluador necesita construir una estructura para vincular los ítems con el constructo. El problema consiste en que a veces las inferencias son erradas, pues la relación entre ítems y constructo se puede interpretar mal, por ejemplo, haciendo afirmaciones de causalidad erróneas (i.e., afirmar que los ítems causan el constructo o viceversa, pero sin evidencia) o inferencias que no se ajustan (i.e., la variable no observable no se puede inferir en realidad a partir de los ítems diseñados).

El primer paso en el proceso de inferencia, por tanto, debe ser establecer qué aspectos de la respuesta se emplearán como base para la inferencia y cómo esos aspectos serán categorizados y puntuados; es decir, se debe establecer uno de los siguientes Espacios de Resultados (Outcome Space):

(a) La categorización de las respuestas de las preguntas en ‘verdadero’ y ‘falso’ en una prueba (con la puntuación posterior asignada, por ejemplo, como ‘1’ y ‘0’).

(b) El registro de respuestas de tipo Likert (de ‘totalmente de acuerdo’ a ‘totalmente en desacuerdo’) en una encuesta de actitudes, y su puntuación posterior según la valencia de los ítems en relación con el constructo subyacente.

Otros Espacios de Resultados menos comunes serían:

(c) Los protocolos de preguntas y guiones en una entrevista estandarizada de respuesta abierta y la posterior categorización de las respuestas.

(d) La traducción de un desempeño en categorías ordenadas mediante una guía de puntuación (i.e., ‘rúbrica’).

Cualquier conjunto de categorías descritas cualitativamente para registrar y/o evaluar cómo han respondido los participantes a los ítems se denomina el espacio de resultados. Las puntuaciones resultantes de este espacio desempeñan un papel fundamental en el enfoque de mapeo de constructos. Ellas encarnan la “dirección” del mapa de constructo (por ejemplo, las puntuaciones positivas se mueven ‘hacia arriba’ en el mapa de constructo), e indican un nivel mayor del constructo medido.

Una característica fundamental es que el espacio de resultados debe consistir únicamente en un número finito de categorías. Por ejemplo, el espacio de resultados del PF-10 (Prueba de funcionamiento general del adulto mayor) consta de

sólo tres categorías: “Sí, limitado mucho”, “Sí, limitado un poco”, y “No, no limitado en absoluto”, ya que su intención es valorar el nivel de disfunción percibida por el adulto mayor en las actividades cotidianas como vestirse, alimentarse, etc. (White et al., 2011).

El orden de las categorías de respuesta debe estar respaldado tanto por la teoría que sustenta el constructo como por evidencia empírica. La teoría que fundamenta el espacio de resultados debe ser coherente con la teoría del propio constructo. La evidencia empírica puede utilizarse para apoyar el ordenamiento del espacio de resultados. En los ítems de opción múltiple, el procedimiento estándar es asignar una puntuación de 1 a la opción correcta (distractor correcto) y 0 a las incorrectas. Así, cuando el distractor correcto representa efectivamente un ejemplo de un “waypoint” (punto de referencia) particular (y los distractores incorrectos están todos asociados a waypoints situados por debajo de este), entonces la puntuación 1 y

0 tiene sentido. Por supuesto, el desarrollador del instrumento debe asegurarse de que no exista ambigüedad en la asignación de los distractores a los waypoints.

Las preguntas con formato de respuesta tipo Likert en encuestas y cuestionarios suelen puntuarse según el número de categorías de respuesta disponibles—si hay cuatro categorías como “Totalmente de acuerdo”, “De acuerdo”, “En desacuerdo” y “Totalmente en desacuerdo”, entonces suelen puntuarse como 0, 1, 2 y 3, respectivamente (o, a veces, como 1, 2, 3 y 4). Cuando un conjunto de respuestas tipo Likert se usa con un mapa de constructo de actitudes o comportamientos, pueden surgir dificultades para interpretar cómo se asignan “De acuerdo” y “En desacuerdo” a los waypoints. En los conjuntos de opciones con una valencia negativa respecto del constructo, la puntuación generalmente se invierte, asignándose 3, 2, 1 y 0, respectivamente (lo que se conoce como “reverse scoring”).

En el caso de ítems de respuesta abierta, las categorías de resultado deben ordenarse en categorías ordinales cualitativamente distintas. Al igual que con los ítems tipo Likert, tiene sentido considerar cada uno de estos niveles ordinales como puntuados usando enteros sucesivos, por ejemplo:

- Crítica completa o comparación de dos argumentos = 3 puntos
- Una justificación completa o un contraargumento = 2
- Una afirmación o evidencia = 1
- Sin evidencia = 0
- Sin oportunidad de responder = sin dato.

DOI: https://doi.org/10.30554/tempuspsi.9.1.5419.2026

1.4. El mapa de Wright

El espacio de resultados produce un conjunto de datos compuesto por los códigos o puntuaciones de cada persona en la muestra. El segundo paso en la inferencia consiste en relacionar estas puntuaciones con el constructo. Esto se hace mediante el cuarto bloque de construcción, denominado el mapa de Wright. El mapa de Wright pone en una misma escala las características de los individuos (el constructo latente) y las características de los ítems (en particular, su dificultad), gracias a las propiedades del modelo psicométrico de Rasch. Este modelo estadístico se utiliza para transformar los códigos basados en los ítems de acuerdo con los "waypoints" (puntuados con enteros 0, 1, 2, etc.) y estimar así la ubicación de los respondientes en una métrica que permite comparar los resultados entre diferentes respondientes.

Más exactamente, el mapa de Wright se apoya en una característica central del modelo Rasch: las estimaciones de ubicación de los respondientes a lo largo del constructo que subyace al mapa de constructo se pueden emparejar con las ubicaciones estimadas de las categorías de respuesta de los ítems (establecidas a partir de su dificultad). Esto permite relacionar las hipótesis acerca de los ítems que han sido diseñados para vincularse con puntos de referencia específicos del mapa de constructo mediante las categorías de respuesta determinadas. En palabras de Wilson:

> Esta característica (tener la misma escala para los ítems y para el constructo) del modelo de Rasch es fundamental tanto para la teoría como para la práctica de la medición en un contexto determinado: (a) desde el punto de vista teórico, proporciona una forma de examinar empíricamente la estructura inherente al mapa de constructo, y añade este análisis como un elemento poderoso en el estudio de la validez del uso de un instrumento; y(b) desde el punto de vista práctico, permite a quienes realizan la medición 'ir más allá de los números' al informar los resultados de la medición a profesionales y usuarios, y los capacita para utilizar el mapa de constructo como un recurso interpretativo clave. (Wilson & Tan, 2023, p. 45)

En el modelo de Rasch, la probabilidad de la respuesta al ítem i, se modela como una función de la ubicación del respondiente (su habilidad ó θ \theta) y de la ubicación del ítem (su dificultad ó δi/delta), donde ambas ubicaciones se conciben a lo largo de una escala común. Las puntuaciones de los ítems obtenidas a partir de una muestra de respondientes se utilizan para estimar los parámetros de los respondientes (i.e., habilidad) y de los ítems (i.e., dificultad) en una escala mediante un modelo estadístico, y luego, la correspondencia entre las ubicaciones de los ítems

en esa escala y los waypoints del mapa de constructo se utiliza para establecer referencias (por ejemplo, puntuaciones) para la escala (Mari et al., 2023).

La lógica del modelo de Rasch es que el respondiente posee una cierta "cantidad" del constructo, indicada por θ\theta, y que un ítem también posee una cierta "cantidad" del mismo constructo, indicada por δi\delta. Sin embargo, estas cantidades interactúan en direcciones opuestas—por eso lo que realmente importa es la diferencia entre el respondiente y el ítem: θ−δi \ theta - delta. La cantidad θ\ theta del respondiente debe compararse con la cantidad δi\delta del ítem para determinar la probabilidad de una respuesta '1' o respuesta correcta (en lugar de una respuesta '0'):

> (a) Cuando las cantidades θ\theta y δi\delta son iguales (es decir, están en el mismo punto del mapa de Wright), las respuestas '0' y '1' tienen la misma probabilidad—por lo tanto, la probabilidad de una respuesta '1' es 0,50. Por ejemplo, el respondiente tiene la misma probabilidad de estar de acuerdo o en desacuerdo con el ítem en una pregunta de actitud; o, en una pregunta de logro, tiene igual probabilidad de responder correcta o incorrectamente.
>
> (b) Cuando el respondiente posee más del constructo que el ítem (es decir, θ>δi ó theta > delta), la probabilidad de una respuesta '1' es mayor a 0,50. En este caso, es más probable que el respondiente esté de acuerdo (en una pregunta de actitud) o que responda correctamente (en una pregunta de logro).
>
> (c) Cuando el ítem posee más del constructo que el respondiente (es decir, θ<δi ó theta< delta), la probabilidad de una respuesta '1' es menor a 0,50. Aquí, el respondiente tiene más probabilidad de estar en desacuerdo (en una pregunta de actitud) o de responder incorrectamente (en una pregunta de logro).

En el contexto de pruebas de logro, diríamos que la 'habilidad' del respondiente es:

(a) igual a,

(b) mayor que, o

(c) menor que la 'dificultad' del ítem.

En el contexto de medición de actitudes, diríamos que:

(a) el respondiente y la afirmación son igualmente positivos,

(b) el respondiente es más positivo que el ítem, y

(c) el respondiente es más negativo que el ítem.

DOI: https://doi.org/10.30554/tempuspsi.9.1.5419.2026

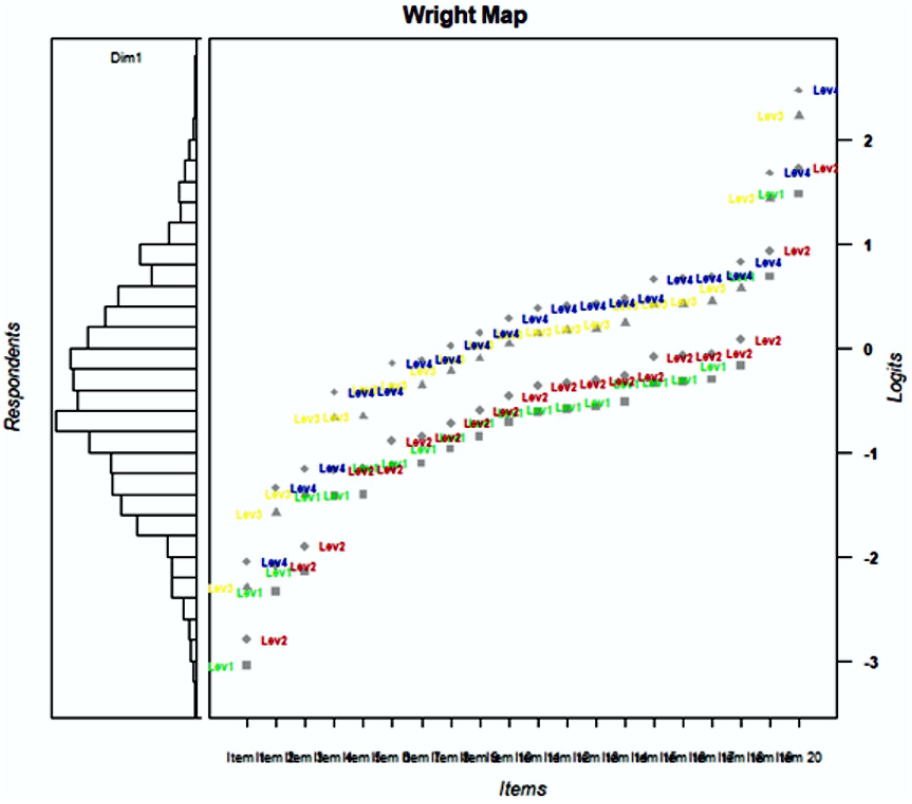

Figura 1. Mapa de Wright

Nota. El mapa relaciona los ítems (y su dificultad) y a los individuos (y su habilidad). Los ítems aparecen graficados como puntos a la derecha (de acuerdo con su nivel de dificultad) y las habilidades de los individuos se representan a la izquierda como una distribución de frecuencia del rasgo e (habilidad). El mapa de Wright permite establecer categorías de sofisticación conceptual dependiendo de la dificultad y la habilidad de los individuos: a mayor dificultad y habilidad, mayor sofisticación conceptual.

La figura 1 presenta un ejemplo de mapa de Wright. Para ítems dicótomos, un modelo de Rasch (modelo logístico 1PL) permite identificar la relación entre la capacidad (latente) del individuo y la dificultad de los ítems. Por tanto, la probabilidad de que un individuo responda a un ítem de forma correcta se modela como función de la diferencia entre la capacidad del individuo y a la vez de la dificultad del ítem. Este modelo permite la comparación visual entre ítems, y el análisis de su relación con la capacidad de los respondientes por medio del mapa de Wright. En dicho mapa se representa gráficamente la habilidad del individuo en la izquierda (histograma de la distribución de frecuencias de habilidad) y la dificultad del ítem a la derecha (medida con el modelo de Rasch) (Wilson, 2005; Irribarra, 2021). Por inspección visual del mapa, se puede concluir cuáles ítems deberían integrar la prueba ya que se facilita determinar la relación entre un ítem y la capacidad del individuo en términos de una misma escala, permitiendo también la construcción de categorías de desempeño (Hontagas, et al., 1998).

La formulación del modelo de Rasch difiere de la de la teoría clásica de los tests en varios aspectos fundamentales. En primer lugar, el modelo de Rasch se expresa tanto a nivel de ítem como de instrumento, y no sólo a nivel de instrumento como en el caso de la teoría clásica de los tests; es decir, en la Teoría Clásica de los Test (CTT), la puntuación total en el instrumento X se expresaba en términos de T(puntuación) y E (error). Por el contrario, en el modelo de Rasch, es la respuesta del ítem para el ítem i, Xi (pronunciado «X-sub-i») la que se modelará centrándose en el ítem. En segundo lugar, el modelo de Rasch tiene tanto un parámetro de persona, que está a nivel de instrumento, como parámetros de ítem, que están a nivel de ítem; por tanto, puede considerarse un modelo multinivel. En tercer lugar, el modelo de Rasch centra la atención en modelar la probabilidad de las respuestas observadas en lugar de modelar la suma de las respuestas, como es el caso del CTT.

2. Vinculando el mapa de constructo y el mapa de Wright: ¿cómo ayuda el mapa de Wright a dar significado a la escala?

El mapa de Wright aporta significado a la escala al representar gráficamente, en una misma métrica, tanto las ubicaciones de los respondientes como las ubicaciones de los ítems en relación con el constructo que se desea medir. Este vínculo con el mapa de constructo es crucial porque permite interpretar los resultados de la medición más allá de simples puntuaciones numéricas. Wilson (2023) presenta el ejemplo de Galileo, quien desarrolló termoscopios que permitían la transducción del calor de los objetos a diversos dispositivos similares a los termómetros actuales, pero de una manera idiosincrática que no permitía una comparación general entre diferentes termoscopios. El dilema de cómo vincular las indicaciones de estos diferentes dispositivos tomó muchos años (incluido el desarrollo de diversas técnicas de estandarización). Sin embargo, el desarrollo crucial fue la fijación de las diferentes indicaciones a puntos críticos generalmente disponibles (públicos) e interpretables, específicamente los puntos de congelación y ebullición del agua.

El evaluador desea tener una base referenciada por criterios para establecer tales valores de referencia públicos (como los puntos de ebullición del agua u otro criterio de referencia). Sin embargo, existen muchas formas en que la correspondencia entre los waypoints' y las estimaciones de sus parámetros correspondientes puede fallar. A veces, las estimaciones no se agrupan de la manera en que el mapa de constructo lo predeciría; otras veces, se agrupan, pero no en el orden previsto. Por este motivo, se requiere hacer nuevas recolecciones de datos, que permitan con nuevas muestras establecer las correspondencias más afinadas entre el mapa de constructo y el mapa de Wright. El enfoque general adoptado es utilizar el mapa de Wright como un vínculo conceptual entre la intención teórica del mapa de constructo

y la evidencia empírica de las estimaciones de los ítems y los respondientes. Este debe ser constantemente revisado y mejorado desde la teoría y desde la evidencia empírica (varias aplicaciones de los ítems a diferentes submuestras).

3. Más de dos categorías de puntuación: datos politómicos. ¿Cómo extendemos el modelo estadístico de Rasch a más de dos categorías?

Para extender el modelo estadístico de Rasch a más de dos categorías, se utiliza el modelo Rasch de respuesta múltiple o modelo politómico. En lugar de tener sólo dos posibles respuestas (por ejemplo, correcto o incorrecto), este modelo permite que las respuestas se clasifiquen en varias categorías ordenadas. Esto es útil en situaciones donde los ítems tienen respuestas graduales o escaladas, como en escalas de actitud o en preguntas de logro con más de dos niveles de dificultad.

El modelo politómico de Rasch se basa en la misma idea fundamental del modelo de Rasch para dos categorías, pero con un ajuste para manejar múltiples categorías de respuesta. Este modelo utiliza umbrales o puntos de corte entre las categorías para estimar las probabilidades de que un respondiente se ubique en una categoría específica. Los parámetros δik (delta_ik) se conocen como “parámetros de paso”— que describen la probabilidad de dar el paso de una categoría de puntuación a la siguiente, por ejemplo, de la puntuación k− 1, a la puntuación k (Masters, 1982; Wright & Masters, 1982). Al considerar las probabilidades para cada una de las categorías de respuesta, las curvas de probabilidad resultantes se pueden denominar funciones de respuesta de categoría (CRF), el equivalente de las curvas de información por ítem del modelo de Rasch, pero ahora aplicado por categoría.

La información presentada en el mapa de Wright ofrece una visión del éxito en el desarrollo de la medición, ya que permite observar rápidamente cuán bien se ajusta la distribución de los ítems a la distribución de los participantes. En el caso politómico esto es relevante por dos razones: (a) la proximidad de los participantes a los umbrales entre categorías de respuesta influye en los errores estándar (por ejemplo, “falsos positivos”) y (b) las limitaciones en el rango de los umbrales pueden indicar limitaciones en la definición del constructo (y, por tanto, errores en el mapa de constructo) implícito en el conjunto de ítems.

Otro error típico con los ítems politómicos es que pueden presentarse efectos de “piso y techo” en cualquiera de los lados (es decir, del lado de los ítems o del lado de los participantes) del mapa de Wright. Por ejemplo, los participantes pueden

ubicarse muy por encima de los umbrales más altos de los ítems, o muy por debajo de los umbrales más bajos. Esto puede llevar al desarrollador de la medición a cuestionar si ha creado una gama suficientemente amplia de ítems (o incluso de puntos de referencia en el mapa del constructo) para representar adecuadamente todo el rango del constructo. Alternativamente, el rango de los participantes puede ser bastante estrecho en comparación con el rango de los umbrales de los ítems, lo cual podría llevar a los desarrolladores de la medición a preguntarse si realmente se necesita una gama tan amplia de umbrales y si deberían concentrar los ítems en aquellos que se ajusten mejor al rango de los participantes.

Este tipo de consideraciones debe ser evaluado cuidadosamente en cada nuevo contexto. Puede suceder, por ejemplo, que la muestra actual esté artificialmente limitada, y que el uso futuro del instrumento incluya efectivamente participantes en los extremos del rango, los cuales no pueden ser medidos por los efectos de techo y piso.

4. El problema de la unidimensionalidad

Aunque Wilson (2023) no comenta sobre la dimensionalidad, es necesario dejar planteado el tema para los profesionales que planean desarrollar o adaptar un instrumento. “La unidimensionalidad se define como la existencia de un solo rasgo latente subyacente a los datos” (Hattie, 1985, p. 139). La unidimensionalidad implica que un conjunto de respuestas a un set de ítems es unidimensional si, y sólo si, la matriz de respuestas a los ítems es localmente independiente después de eliminar un único factor latente común. Sin embargo, lograr esto no es tan sencillo en la realidad, y aún así, se insiste en formar a los nuevos psicómetras como si siempre se cumpliera la unidimensionalidad, “McDonald caracteriza la visión predominante sobre la posibilidad de que los datos se ajusten estrictamente a un modelo unidimensional: ‘tal caso no ocurrirá en la aplicación de la teoría’” (1981, p. 102).

Ante este hecho, los investigadores han dedicado mucho esfuerzo a: a) estudiar el grado en que las estimaciones de parámetros de la TRI (Teoría respuesta al ítem) son robustas (es decir, aproximadamente correctas) frente a distintos niveles de violación de la Unidimensionalidad, y b) desarrollar criterios estadísticos para juzgar si los datos se aproximan razonablemente al Modelo A de la figura 2 (por ejemplo, un rasgo general “fuerte” o unidimensional). Los modelos que pueden relacionar los ítems y la variable latente pueden ser variados y esto implicaría diferencias en las relaciones de causalidad y en la posibilidad de tener ítems en realidad ortogonales y que no comparten varianza con diferentes variables latentes.

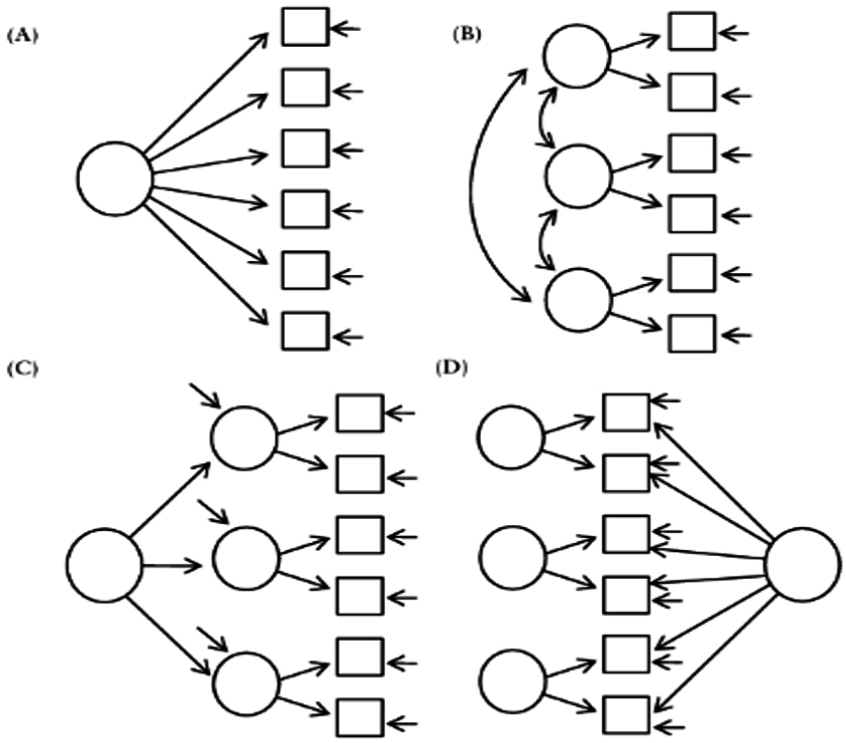

Figure 2.1 Alternative models: A—unidimensional model, B—correlated traits model, C—second-order factor model, D—bifactor model

Figura 2. Modelos que relacionan ítems y factores.
Nota. Tomado de: Reise & Revicki (2015).

En la figura 2 se muestran modelos alternativos que pueden resultar de tener un constructo latente y varias estructuras factoriales. Por ejemplo, el modelo A sería el unidimensional con una sola variable latente, y este es el modelo en el que se aplica análisis de TRI. El modelo B implica que los rasgos o variables latentes están correlacionados y no son independientes, de modo que no es posible hacer rotaciones ortogonales, sino que los ítems pueden caracterizar más de un factor. El modelo C implica que una variable latente es causa de variables de segundo orden (factores) los cuales a su vez son independientes y están configurados por distintos grupos de ítems que no están relacionados entre sí. El último modelo (D) tiene tanto la causalidad del rasgo latente como de los factores de segundo nivel. Si la multidimensionalidad se debe a múltiples dimensiones latentes moderadamente correlacionadas, o si existe un factor general fuerte, los modelos de TRI, según algunos autores son relativamente robustos y se pueden usar sin distinción de que haya Unidimensionalidad o multidimensionalidad; sin embargo, en la literatura se recomienda hacer análisis de bondad de ajuste comparando modelos, y también se sugiere hacer un análisis factorial para establecer las variables subyacentes (unidimensionalidad vs. multidimensionalidad) (Reise & Revicki, 2015).

Conclusiones

El desarrollo de mediciones de rasgos psicológicos es un proceso que requiere múltiples pasos, claridad en los conceptos y ajuste a un modelo de medición sin abusar de los supuestos que se requiere cumplir. Los estudiantes de psicometría requieren identificar que el modelo unidimensional no es el más común, de modo que tendrán que recurrir a modelos de medición alternativos, que permitan establecer las características que subyacen a los constructos involucrados.

El proceso de creación de instrumentos requiere el uso de procedimientos como el planteado por Wilson (2023), quien por más de 20 años ha perfeccionado y enseñado su modelo BEAR. Los cuatro ladrillos de construcción del modelo son claves para el desarrollo de instrumentos con validez interna y que sean confiables; sin embargo, queda la inquietud sobre la restricción al modelo unidimensional y sobre las dificultades que se hallan cuando el evaluador es quien de manera autónoma determina los puntos de referencia de su constructo. No obstante, es una buena aproximación para avanzar en el desarrollo de la teoría y de las formas de medición de los constructos psicológicos.

Referencias

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals* (Vol. Handbook I: Cognitive domain). Davidson.

Ferrara, S., Lai, E., Reilly, A. & Nichols, P. D. (2016). Principled approaches to assessment design, development, and implementation. In A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, pp. 41–74. Wiley-Blackwell. https://doi.org/10.1002/9781118956588.ch3

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hontagas, P., Ponsoda, V., Olea, J. &Revuelta, J. (1998). Representación de funciones características de ítems dicotómicos y politómicos. *Psicothema, 10*(2), 475-479. https://www.redalyc.org/pdf/727/72710219.pdf

Irribarra, D. (2021). *A Pragmatic Perspective of Measurement.* Springer. https://link.springer.com/book/10.1007/978-3-030-74025-2

Krathwohl, D. R., Bloom, B. S. & Masia, B. B. (1964). *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook II: Affective Domain.* David Mckay Company Incorporated.

Masters, G. N. (1982) A Rasch model for partial credit scoring. *Psychometrika 47*(2), 149-174. https://doi.org/10.1007/BF02296272

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*(1), 100–117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

Mari, L., Wilson, M. & Maul, A. (2023). *Measurement Across the Sciences: Developing a Shared Concept System for Measurement (second edition).* Springer.

Mislevy, R. (2018). *Sociocultural Foundations of Educational Measurement*. Routledge

National Research Council (NRC). (2001). Knowing what students know: The science and design of educational assessment (Committee on the foundations of assessment). In J. Pellegrino, N. Chudowsky & R. Glaser (Eds.*), Division on Behavioral and Social Sciences and Education*. National Academy Press.

Nicholls, P. D, Ferrara, S., Lai, E. & Reilly, A (2016) Principled Approaches to Assessment Design, Development, and Implementation. In Leighton, J & Rupp, A (Eds) *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. Wiley Handbooks in Education. https://doi.org/10.1002/9781118956588.ch3

Reise, S. & Revicki, A. (2015). *Handbook of IRT Modeling: Applications to typical performance assessment.* Routledge.

White, D. K., Wilson, J. C. & Keysor, J. J. (2011). Measures of adult general functional status: SF-36 Physical Functioning Subscale (PF-10), Health Assessment Questionnaire (HAQ), Modified Health Assessment Questionnaire (MHAQ), Katz Index of Independence in Activities of Daily Living, Functional Independence Measure (FIM), and Osteoarthritis-Function-Computer Adaptive Test (OA-Function-CAT). *Arthritis Care & Research*, *63*(11), S297-S307. https://doi.org/10.1002/acr.20638

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Routledge, Taylor & Francis Group.

Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181–208. https://doi.org/10.1207/S15324818AME1302_4

Wilson, M. & Tan, S. (2023). Test development: Principled assessment design. In D. Mc Caffrey & A. Rupp (Eds.), *International Encyclopedia of Education* (fourth edition, Volume 10: Quantitative Research/ Educational Measurement, pp. 146–162). Oxford: Elsevier Ltd.

Wright, B. D. & Masters, G. N. (1982). Rating Scale Analysis. *Psychology, 7*(2). https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1779536