Validación de un cuestionario para medir la habilidad de la población general para evaluar afirmaciones acerca de tratamientos médicos

Giordano Pérez-Gaxiola[1] y Astrid Austvoll-Dahlgren[2]

[1]Hospital Pediátrico de Sinaloa "Dr. Rigoberto Aguilar Pico", Sinaloa, México; [2]Norwegian Institute of Public Health, Oslo, Noruega

Resumen

Introducción: *Todos los días, las personas se enfrentan a afirmaciones acerca de tratamientos en medios de comunicación, redes sociales o por viva voz.* **Objetivo:** *Validar un cuestionario en español para medir las habilidades de un individuo para evaluar afirmaciones acerca de tratamientos.* **Método:** *Veintidós preguntas de opción múltiple de la base de datos Claim Evaluation Tools fueron traducidas y aplicadas a 172 niños y 268 adultos. Mediante un modelo Rasch se exploró el ajuste promedio e individual por reactivo, el potencial comportamiento diferencial del reactivo (basado en el género, edad y modo de aplicación), la multidimensionalidad y la independencia local.* **Resultados:** *El ajuste promedio por reactivo fue apropiado. Cuatro preguntas de opción múltiple mostraron pobre ajuste. La fiabilidad del cuestionario fue satisfactoria, con un índice de separación de 0.7. Las preguntas de opción múltiple fueron unidimensionales, y no hubo dependencia específica.* **Conclusión:** *Se obtuvo un conjunto de 18 preguntas de opción múltiple con ajuste satisfactorio. El cuestionario es el primero disponible y validado en español para medir las habilidades de los individuos para evaluar afirmaciones acerca de tratamientos.*

PALABRAS CLAVE: *Alfabetización en salud. Educación del paciente. Medicina basada en la evidencia. Toma de decisiones. Estudios de validación.*

Abstract

Introduction: *Every day, people are faced with claims about treatment effects through mass media, social media, or by word of mouth.* **Objective:** *To validate a Spanish-language questionnaire to measure the ability of an individual to assess claims about treatments effects.* **Method:** *A set of 22 multiple choice questions taken from the claim evaluation tools database were translated and applied to 172 children and 268 adults. Using a Rasch model, overall and individual item-person fit was explored, as well as the potential item differential functioning (based on gender, age and mode of administration), multidimensionality and local independence.* **Results:** *Overall item-person fit was appropriate. Four multiple-choice questions showed a poor fit. Reliability of the questionnaire was satisfactory with a person separation index of 0.7. Multiple-choice questions were unidimensional, and there was no specific dependency.* **Conclusion:** *A set of 18 multiple-choice questions with satisfactory fit was obtained. This is the first available questionnaire validated in Spanish to measure individuals' ability to assess claims about treatment effects.*

KEY WORDS: *Health literacy. Patient education. Evidence-based medicine. Decision making. Validation studies.*

Correspondencia:
Giordano Pérez-Gaxiola
E-mail: giordano@sinestetoscopio.com

Fecha de recepción: 28-02-2017
Fecha de aceptación: 20-10-2017
DOI://dx.doi.org/10.24875/GMM.17003340

Gac Med Mex. 2018;154:480-495
Disponible en PubMed
www.gacetamedicademexico.com

Introducción

La población en general se enfrenta todos los días a afirmaciones sobre los efectos de los tratamientos, ya sea a través de los medios de comunicación, redes sociales o familiares y amigos. Estas afirmaciones pueden incluir consejos sobre cómo prevenir enfermedades o sobre los efectos de intervenciones terapéuticas individuales, de salud pública o de sistemas de salud.[1-4] Existen diversas razones por las que estas afirmaciones se realizan, desde buenas intenciones hasta intereses comerciales. Sin embargo, muchas aseveraciones son poco confiables o francamente erróneas y las personas pueden sufrir innecesariamente o desperdiciar recursos.[5-11] Consecuentemente, ayudar a las personas a tomar decisiones compartidas en salud mediante el mejoramiento de sus habilidades para evaluar críticamente dichas afirmaciones es una importante iniciativa de salud pública.[12]

Los esfuerzos de alfabetización en salud en México están dirigidos principalmente a proveer de información fidedigna a través de sitios web específicos.[13] Estos programas y sitios son dirigidos por la Secretaría de Salud. La Comisión Federal para la Protección contra Riesgos Sanitarios, Cofepris, ocasionalmente usa redes sociales, en especial Facebook® y Twitter®, para educar acerca de afirmaciones de dudosa calidad sobre la salud. Por otro lado, la enseñanza de la atención sanitaria basada en la evidencia y las habilidades para leer críticamente literatura científica o afirmaciones sobre tratamientos se realiza solo en algunas facultades de medicina y no se dirige a la población general. Por lo tanto, existe la necesidad de brindar este tipo de capacitación también a los pacientes. Añadido a esto, hasta donde tenemos conocimiento no existe un instrumento validado disponible en español para medir la habilidad de las personas para analizar críticamente afirmaciones sobre tratamientos.

El desarrollo y la aplicación de la base de datos Claim Evaluation Tools

Aunque existe un creciente número de recursos educativos para mejorar el pensamiento crítico de las personas hacia las afirmaciones sobre tratamientos, son pocos los que han sido propiamente validados.[14] Una revisión sistemática reciente concluyó que esas intervenciones pedagógicas no han sido medidas consistentemente y no existen instrumentos completos para evaluar las habilidades de pensamiento crítico.[15] Debido a esto, la base de datos Claim Evaluation Tools (CET), un conjunto de preguntas de opción múltiple (POM), fue desarrollada por el proyecto Informed Health Choices (IHC) (www.informedhealthchoices.org). El proyecto IHC es una colaboración internacional de investigadores de Uganda, Ruanda, Kenia, el Reino Unido, Australia y Noruega que ha desarrollado intervenciones educativas para que las personas puedan evaluar afirmaciones sobre tratamientos.[16,17] La base de datos CET se elaboró inicialmente para medir los resultados de los ensayos aleatorios que formaron parte del proyecto IHC, pero puede ser usada en ámbitos escolares para crear cuestionarios o para realizar estudios transversales para medir las habilidades de pensamiento crítico en diferentes poblaciones.[16,17]

Las POM fueron elaboradas con base en retroalimentación cualitativa y cuantitativa tanto de expertos en metodología como de la población general de varios países.[18] Se han realizado análisis psicométricos de las POM en Noruega y Uganda y se encontró que tienen un constructo válido y confiable en esos entornos, donde se incluyeron niños, adultos, profesionales de la salud y personas de bajos recursos y bajo nivel educativo.[19]

Como punto inicial para el desarrollo de recursos educativos y de la base de datos CET, el grupo IHC desarrolló una lista de conceptos clave que las personas deben entender para evaluar afirmaciones sobre tratamientos.[20] Esta lista sirve como currículo a investigadores y maestros para desarrollar intervenciones. La lista se revisa y se actualiza anualmente. La base de datos CET incluye entre cuatro y seis POM por cada concepto clave.[18] Estas POM fueron escritas en inglés y posteriormente se han traducido a noruego, luganda, alemán y chino. La base de datos CET está hospedada en el sitio web Testing Treatments interactive (www.testingtreatments.org), plataforma interactiva traducida a múltiples idiomas y cuya versión en español se dio a conocer en 2012.

El objetivo de este estudio es describir la validación psicométrica de un conjunto de POM que miden las habilidades de la población general para analizar afirmaciones sobre tratamientos, traducido al español y aplicado en Culiacán, Sinaloa, ciudad de aproximadamente 850 000 habitantes en el noroeste de México.

Método

Para este estudio se seleccionó una muestra pragmática de POM que abarcan 11 de los 32 conceptos

clave originales de la base de datos CET para su validación en español.[20] Estos 32 conceptos clave estuvieron divididos en su primera versión en seis grupos:

1. Reconocer la necesidad de comparaciones imparciales de tratamientos.
2. Juzgar si una comparación entre tratamientos es imparcial.
3. Entender el papel que desempeña el azar.
4. Considerar todas las comparaciones imparciales relevantes.
5. Entender los resultados de las comparaciones imparciales de tratamientos.
6. Juzgar si las comparaciones imparciales de los tratamientos son relevantes.

Desde su primera publicación, y simultáneamente con este estudio, la lista de conceptos clave ha sido actualizada. Para este estudio se usó la numeración de la primera publicación.[20] La actualización de la lista original de los conceptos clave (Anexo 1) no tiene impacto en los métodos usados o en las implicaciones de los resultados de este estudio. Los conceptos seleccionados incluyen los seis grupos originales (Tabla 1). La versión más reciente contiene 34 conceptos clave divididos en tres grupos y está disponible en línea (http://www.testingtreatments.org/wp-content/uploads/2016/10/Key-Concepts-2nd-edition-with-TTI-short-titles_14122016.pdf).

Para permitir la eliminación de POM potencialmente problemáticas se seleccionaron dos reactivos por cada concepto clave de interés. Todos los conceptos clave y las POM fueron traducidos por uno de los autores. La traducción fue revisada y actualizada después de retroalimentación de un médico y un paciente. Un ejemplo de una POM traducida se muestra en la Tabla 2. Aparte de las POM, el cuestionario incluyó preguntas sobre características demográficas como edad, sexo, experiencia científica (dirigida a adultos: haber participado antes en un ensayo clínico o haber tomado un curso de medicina basada en evidencias) y nivel socioeconómico/educativo (usando la pregunta ¿cuántos libros tienen en casa? como medición indirecta en niños).

Tabla 1. Conceptos clave seleccionados de la base de datos Claim Evaluation Tools y traducidos para su validación en español

1.2. Las anécdotas no son evidencia confiable.
1.3. Asociación no necesariamente significa causalidad.
1.7. Ten cuidado con los conflictos de intereses.
1.10. Evita expectativas poco realistas.
2.1. Las comparaciones de tratamientos son necesarias
2.2. Las comparaciones deben hacerse entre grupos similares.
2.5. Cuando sea posible, las personas no deben saber qué tratamiento están recibiendo.
3.1. Los estudios pequeños pueden tener resultados engañosos.
4.1. Los resultados de un solo estudio pueden ser engañosos.
5.1. Los tratamientos pueden tener tanto efectos beneficiosos como dañinos.
6.1. Los estudios deben medir desenlaces que sean importantes.

Tabla 2. Ejemplo del formato de las preguntas de opción múltiple

Jorge tiene dolor de estómago. La última vez que Jorge tuvo dolor de estómago fue hace 2 meses. Aquella vez, él tomó un poco de leche tibia y después de una hora se le quitó el dolor. Por eso, Jorge dice que la leche tibia cura el dolor de estómago.
Pregunta: ¿Tiene razón Jorge?
Opciones:
A. No se puede saber. Es posible que el dolor se le hubiera quitado sin tomar la leche
B. No se puede saber, pero es probable que sea verdad basado en que Jorge tuvo esa experiencia
C. Sí, la experiencia de Jorge es suficiente para demostrar que la leche tibia alivia el dolor de estómago
Respuesta:

Descripción de los participantes y aplicación de los cuestionarios

El cuestionario final fue aplicado a adultos en línea y a niños de 10 a 15 años de edad en papel. El cuestionario en línea se realizó usando QuestBack® y los participantes fueron reclutados usando redes sociales (Facebook®, Twitter® y Whatsapp®). El cuestionario en papel se aplicó en dos escuelas secundarias en Culiacán, Sinaloa: Escuela Activa Integral, escuela privada con 350 estudiantes, y Escuela Secundaria Federal #2 “General Antonio Rosales Flores”, escuela pública con 600 estudiantes. Las escuelas fueron seleccionadas por conveniencia de la zona central de la ciudad. Un salón de cada uno de los tres grados de secundaria fue seleccionado al azar de cada escuela para la aplicación del cuestionario. En la literatura no existe un consenso sobre el tamaño de muestra necesario conforme a la metodología del modelo Rasch.[21] El tamaño se determinó a partir de un juicio pragmático que toma en cuenta el número de reactivos evaluados, y el poder estadístico necesario para evaluar el ajuste en el modelo Rasch.

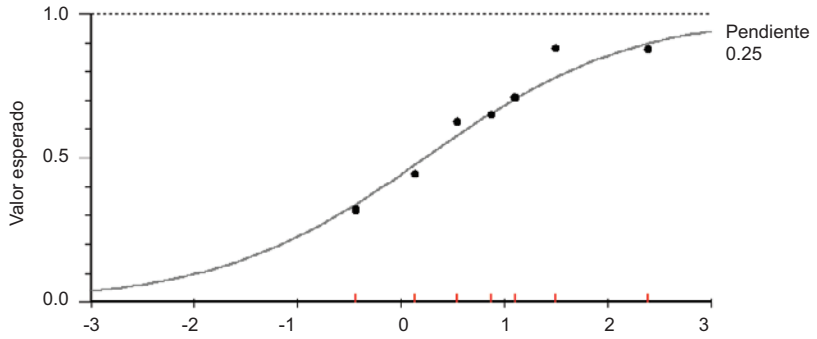

Figura 1. *Curva de características de los reactivos.*

A partir de estudios previos se encontró que una muestra > 300 participantes es suficiente cuando se prueba un solo cuestionario que incluya aproximadamente 22 POM y para probar el ajuste y sesgo por sexo o edad del participante.[18,19]

Modelo Rasch

El modelo Rasch es una forma dinámica y unificada para evaluar diversos aspectos de medición para validar un instrumento, incluyendo validez interna del constructo (probando multidimensionalidad), invarianza de la medida (interacción reactivo-persona) y sesgo por reactivo (diferencial por reactivo).[22,23] Puede ser usado para datos dicotómicos o politómicos.[23-25] Como los reactivos en la base de datos CET tienen un formato de opción múltiple se califican dicotómicamente. Los datos fueron recabados en Excel y el análisis Rasch[23] se realizó con el programa RUMM2030® (RUMM Laboratory).

Se exploró la estructura del intervalo de clase (número y tamaño de grupos por habilidad) de la muestra y se realizó el análisis inicial del reactivo seleccionado para explorar la interacción reactivo-persona.

En un modelo Rasch, la relación o *ratio* entre dos reactivos debe ser constante entre los diferentes grupos de habilidades (habilidades para analizar críticamente afirmaciones acerca de tratamientos). Los patrones de respuesta hacia un reactivo se comparan con lo que se esperaría según el modelo, forma probabilística del escalograma de Guttman.[23] En otras palabras, entre más fácil es el reactivo es más probable que sea respondido correctamente y entre más hábil es la persona es más probable que conteste adecuadamente.[26]

En RUMM2030®, la interacción reactivo-persona se presenta en unidades de medida denominadas logit, que expresan la distancia entre los parámetros del modelo. La ubicación promedio del reactivo siempre es 0. Si el instrumento tiene un nivel de dificultad adecuado (ni muy fácil ni muy difícil), la ubicación de las personas también estará alrededor de cero.[23] Si la ubicación de la persona queda por arriba de 0, el instrumento es fácil, si queda por debajo de 0 indica que es difícil.

El ajuste residual por reactivo y persona analiza el grado de divergencia entre los datos observados y los esperados para cada persona-reactivo cuando se suma para todos los reactivos y personas, respectivamente. En RUMM2030® esto se reporta como una puntuación z (*z-score*) aproximada, que representa una distribución normal estandarizada.[27] Idealmente, el ajuste por reactivo o persona debería tener un promedio de 0 y una desviación estándar (DE) de 1.[23]

También se investigó el ajuste por persona para identificar aquellas con posible ajuste pobre. Si todas las personas responden como se espera, caerán dentro de un ajuste residual de ± 2.5. Se valoró excluir a las personas con ajuste pobre del análisis porque podrían ocasionar sesgo. Algunas razones para un ajuste pobre podrían ser que la persona adivine, copie o provea "respuestas demasiado perfectas", como seleccionar siempre la opción A en todas las preguntas del cuestionario.

Se exploró el ajuste individual por reactivo en el modelo Rasch con chi cuadrado y el ajuste residual. Las POM con probabilidades de chi cuadrado significativo no se ajustaron al nivel de significación de 0.01 en el modelo. Las POM con ajuste residual en un rango ± 2.5 se consideraron potencialmente problemáticas y requirieron mayor análisis.

Cuando los datos se ajustan al modelo de Rasch, la capacidad se mide consistentemente mediante la escala de rasgos, con un error de medición bajo. Consecuentemente, el ajuste en el modelo Rasch implica confiabilidad. En un análisis Rasch esto se

muestra con un índice de separación e indica la capacidad del constructo para diferenciar entre personas con alta o baja habilidad. Esto también indica qué puede confiarse en las estadísticas de ajuste.[27] Un valor de 0.7 en el índice de separación se consideró aceptable.

Para la identificación de POM con ajuste pobre se empleó la curva de características de los reactivos (CCI), que indica la probabilidad esperada de contestar correctamente un reactivo en función de la habilidad en el análisis de rasgos latentes (Figura 1). La línea en la CCI representa las puntuaciones esperadas y los puntos, las puntuaciones observadas en cada intervalo de clase. Existe un buen ajuste cuando las puntuaciones observadas siguen las esperadas; los reactivos con ajuste pobre indican error de medición.[23] Se revisó individualmente la CCI para todas las POM en busca de ajustes pobre.

El análisis de las características de los reactivos puede usarse también para identificar el diferencial por reactivo (DIF). Idealmente, con excepción de la habilidad para evaluar afirmaciones sobre tratamientos, se espera que las POM de la base de datos CET funcionen igual para ambos sexos y los distintos grupos etarios. Como las versiones de los cuestionarios de este estudio fueron administrados a distintas poblaciones (electrónico a adultos y en papel a niños), también se exploró el potencial DIF para esta variable. Existen dos tipos de DIF: el uniforme se presenta cuando la diferencia entre los grupos respecto a un reactivo es uniforme, por ejemplo, que los adultos tuvieran mayor habilidad en comparación a los niños; en el no uniforme, la diferencia entre los grupos de habilidad puede ser inconsistente.[23] Se excluyeron las POM que mostraran un DIF no uniforme.

En el modelo Rasch se asume que cualquier subconjunto de reactivos administrados a una persona debe proveer la misma estimación sobre la habilidad, por lo tanto, la evaluación de la multidimensionalidad es esencial, la cual puede realizarse comparando la ubicación de la persona con base en dos subconjuntos de reactivos del cuestionario: se identificaron los dos subconjuntos de POM más divergentes dentro del cuestionario y se compararon con la ubicación de la persona con una prueba independiente de t de Student.

Además, se exploró la dependencia local, es decir, hasta qué punto un reactivo influye en la respuesta de otros reactivos. Esto se realizó con la función de correlaciones residuales en RUMM2030®. Una correlación residual entre 0.2 y 0.3 por arriba del promedio de las correlaciones residuales de todos los reactivos fue considerada potencialmente problemática. Se consideró modificar o eliminar las POM con ajuste pobre.

Tabla 3. Características demográficas de los participantes en la validación de un instrumento que explora la habilidad para evaluar afirmación sobre tratamientos médicos (formatos web e impreso)

	Cuestionario web (n = 268)	Cuestionario impreso (n = 172)
Grupo de edad (años)		
10-15	4	172
16-25	26	0
26-35	88	0
36-45	91	0
46-55	26	0
56-65	27	0
66-75	6	0
Sexo femenino	107 (40 %)	86 (50 %)
Participaron antes en un ensayo clínico	70 (26 %)	NA
Adultos: habían tomado un curso de lectura crítica o medicina basada en evidencias	105 (39 %)	NA
Niños: ¿cuántos libros tienes en casa?		
Muchos	NA	84 (49 %)
Algunos	NA	71 (41 %)
Pocos	NA	17 (10 %)

Resultados

Se incluyeron 268 participantes adultos en el cuestionario en línea y 172 niños, entre 10 y 15 años, en la versión en papel. Todos los estudiantes de los salones seleccionados participaron: 99 niños de la escuela pública y 73 de la privada; 67 % de los adultos se encontró entre los 25 y 45 años, 26 % había participado en algún ensayo clínico y 39 % había tomado cursos de lectura crítica. Las características demográficas se muestran en la Tabla 3.

Se identificó un intervalo de clase de siete grupos de habilidad. El ajuste residual por reactivo y persona fue satisfactorio y cercano a 0: −0.247 ± 2.083 y −0.0643 ± 0.60, respectivamente, sin embargo, idealmente el error estándar para el ajuste residual por reactivo debía estar por debajo de 1.4.

La ubicación promedio de la habilidad de la persona fue de 1.348 logits, lo que indicó que en general los participantes encontraron la prueba algo fácil (tuvieron mayor habilidad que el nivel de dificultad de las POM).

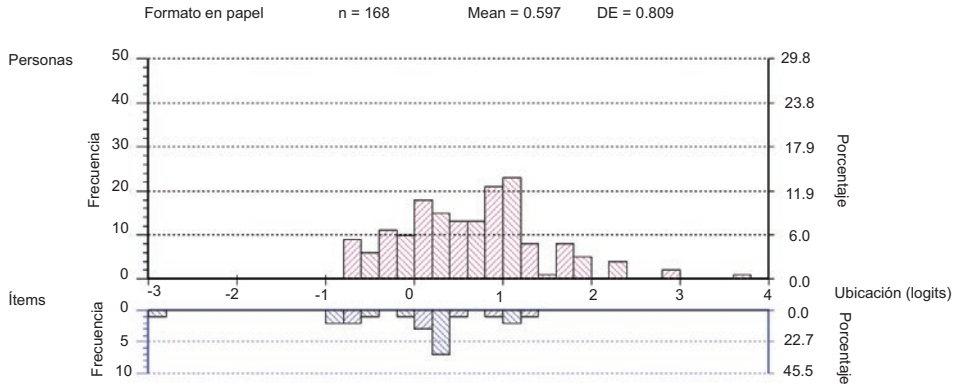

Figura 2. *Mapa de la interacción promedio de reactivo-persona en los niños. Agrupación establecida a una longitud de intervalo de 0.20 que deriva en 35 grupos.*

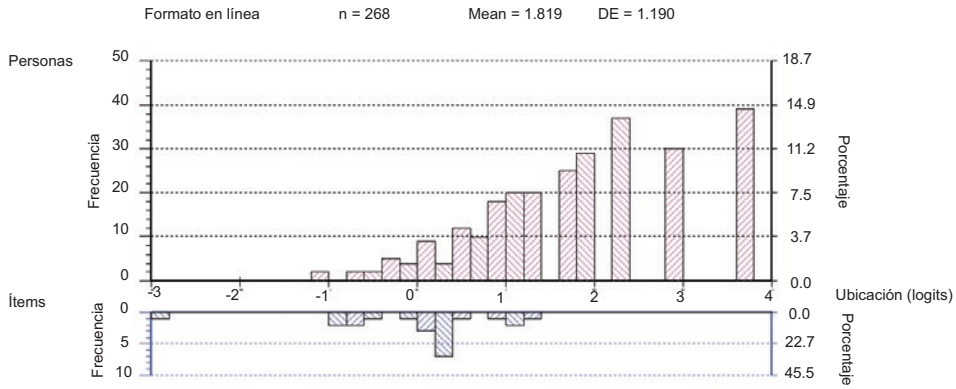

Figura 3. *Mapa de la interacción promedio de reactivo-persona en los adultos. Agrupación establecida a una longitud de intervalo de 0.20 que deriva en 35 grupos.*

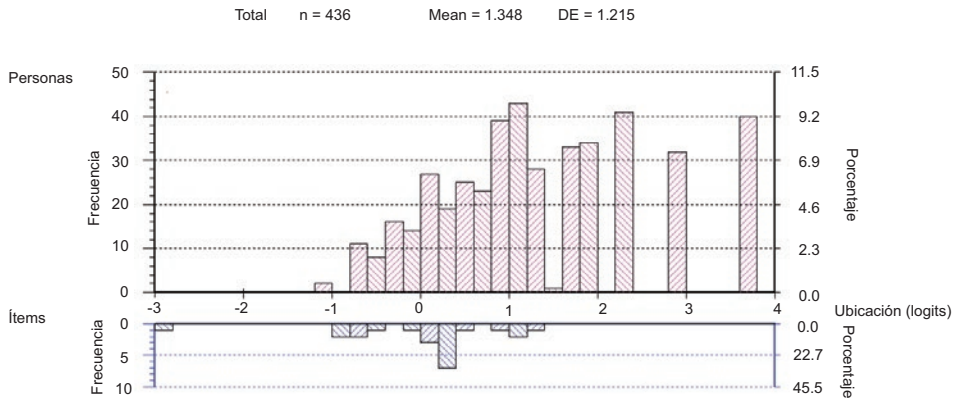

Figura 4. *Mapa de distribución persona-reactivo. Agrupación establecida a una longitud de intervalo de 0.20 que deriva en 35 grupos.*

Cuando se evaluaron por separado las muestras, la focalización fue mejor en los niños y la habilidad, en los adultos (Figuras 2 y 3), probablemente porque una proporción de adultos había recibido antes entrenamiento sobre los conceptos clave y lectura crítica.

Los promedios de los ajustes residuales por reactivo y persona fueron satisfactorios y cercanos a 0, aunque con errores estándar más altos que el ideal (< 1.4); el mapa de persona-reactivo se muestra en la Figura 4, en el cual la parte de superior representa los grupos participantes y su nivel de habilidad y la parte de inferior, la ubicación de los reactivos y su distribución.[27] La confiabilidad del cuestionario fue satisfactoria, con un índice de separación de 0.7.

El análisis del ajuste por persona mostró que casi todos los reactivos quedaron dentro del rango diferencial ± 2.5. Al explorar el ajuste por reactivo, solo una POM cayó fuera del rango residual ± 2.5 y el valor de chi cuadrado alcanzó significación estadística.

Al inspeccionar las CCI de las POM para determinar la habilidad de cada pregunta para discriminar, encontramos que casi todas las POM tuvieron ajuste satisfactorio en la curva, con excepción de dos.

Se exploró el DIF por edad, sexo y por mecanismo de aplicación del cuestionario. Ninguna POM mostró DIF para sexo. En cuanto a la edad, se encontró un DIF no uniforme en dos POM y DIF uniforme en cuatro. Seis POM tuvieron un DIF para el mecanismo de aplicación, en el cual los participantes de la versión en línea tuvieron más habilidad, es decir, los adultos, quienes habían tenido entrenamiento previo en lectura crítica.

Mediante t de Student se encontró que las POM fueron satisfactorias y unidimensionales. Además, no existió dependencia local importante, lo cual sugiere que no hubo redundancia en los reactivos.

Por lo anterior, se decidió eliminar las POM con ajuste pobre en la curva CCI, con DIF no uniforme, con valor de chi cuadrado significativo y ajuste residual ± 2.5 (cuatro POM), con lo cual se obtuvo un conjunto de 18 preguntas de opción múltiple con ajuste satisfactorio en el modelo Rasch y que abarcaban todos los conceptos clave seleccionados.

Discusión

Este estudio explora las propiedades psicométricas de un cuestionario desarrollado usando una muestra selecta de preguntas de opción múltiple de la base de datos CET y traducidas al español. En general, el cuestionario tuvo un buen ajuste al modelo y una confiabilidad satisfactoria. Con base en los hallazgos del análisis Rasch, se eliminaron las POM individuales con pobre ajuste.

Todas las POM de dicha pueden obtenerse para uso no comercial, previa expresa solicitud, en la página web de Testing Treatments interactive (www.testingtreatments.org). La información para la evaluación psicométrica en diferentes entornos está disponible también bajo solicitud.

Ha habido interés en validar conjuntos de POM de la base de datos en diferentes países. Actualmente se están realizando validaciones en China, Alemania, Reino Unido y Noruega y existen planes para validar la traducción al español del resto de los conceptos clave, así como estudios transversales en Chile y España. Esto permitirá explorar la validez y aplicabilidad en otras regiones de habla hispana. Nuestra intención también es reformular las cuatro preguntas con pobre ajuste en futuros estudios.

Para este estudio evaluamos las POM que consideramos importantes en el contexto mexicano, con lo que abarcamos 11 de los 32 conceptos clave originales. Con excepción de los conceptos clave 1.10 y 6.1, las POM restantes también fueron priorizadas en las intervenciones educativas del proyecto IHC en Uganda.[16,17] Aunque estos juicios de priorización se realizaron separadamente, se llegó a conclusiones casi idénticas.

Los juicios sobre la relevancia de cada concepto clave deben realizarse conforme el contexto y la intención del uso de las POM, por ejemplo, al evaluar herramientas educativas en una población particular. Algunos reactivos se consideran más básicos que otros y podrían ser punto de partida cuando se crea el currículo para una intervención educativa.

Esta investigación es la primera iniciativa para traducir y validar las POM de esta base de datos en español, sin embargo, se necesitan más estudios para evaluar las POM que abordan el resto de los conceptos clave, para proveer a investigadores, clínicos y maestros de un conjunto completo que sirva para hacer evaluaciones en múltiples entornos en los países de habla hispana. Actualmente falta evidencia acerca de las habilidades de la población general para evaluar afirmaciones sobre tratamientos, lo cual puede evaluarse con un estudio transversal usando las POM que validamos.

Conclusión

Basados en el modelo Rasch, se encontró que el cuestionario fue confiable; la mayoría de las preguntas de opción múltiple mostró ajuste satisfactorio y validez interna en el constructo. El producto final es un conjunto de 18 POM que pueden usarse en México y países de habla hispana para estudios de validez y aplicabilidad en otras regiones, así como para estudios transversales que evalúen la habilidad de la población en general, niños y adultos, para analizar afirmaciones sobre tratamientos.

Agradecimientos

Estamos profundamente agradecidos con los niños y maestros de la Escuela Activa Integral y la Escuela Secundaria Federal #2 “General Antonio Rosales

Flores", así como con todas las personas que contribuyeron a este proyecto. Agradecemos también a sir Iain Chalmers y Andrew D. Oxman, por su apoyo para que este estudio fuera posible.

Bibliografía

1. Lewis M, Orrock P, Myers S. Uncritical reverence in CM reporting: assessing the scientific quality of Australian news media reports. Health Sociol Rev. 2010;19:57-72.
2. Glenton C, Paulsen E, Oxman A. Portals to Wonderland? Health portals lead confusing information about the effects of health care. BMC Med Inform Decis Mak. 2005;5:7:8.
3. Moynihan R, Bero L, Ross-Degnan D, Henry D, Lee K, Watkins J, et al. Coverage by the news media of the benefits and risks of medications. N Engl J Med. 2000;342:1645-1650.
4. Wolfe R, Sharp LK, Lipsky MS. Content and design attributes of antivaccination web sites. JAMA. 2002;287:3245-3248.
5. Woloshin S, Schwartz L, Byram S, Sox H, Fischhoff B, Welch H. Women's understanding of the mammography screening debate. Arch Int Med. 2000;160:1434-1440.
6. Fox S, Duggan M. Health Online 2013. [Consultado 2013 Apr 09]. Disponible en: http://www.pewinternet.org/Reports/2013/Health-online.aspx
7. Robinson EJ, Kerr CE, Stevens AJ, Lilford RJ, Braunholtz DA, Edwards SJ, et al. Lay public's understanding of equipoise and randomisation in randomised controlled trials. Health Technol Assess. 2005;9:1-192.
8. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med. 2007; 64:1853-1862.
9. Horsley T, Hyde C, Santesso N, Parkes J, Milne R, Stewart R. Teaching critical appraisal skills in healthcare settings. Cochrane Database Syst Rev. 2011;9:CD001270.
10. Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev. 2011;1:CD001431.
11. Evans I, Thornton H, Chalmers I, Glasziou P. Testing treatments: better research for better healthcare. Second edition. London: Pinter & Martin; 2011.
12. The BMJ Opinon. [Blog]. Chalmers I, Glasziou P, Badenoch D, Atkinson P, Austvoll-Dahlgren A, Oxman A. Evidence Live 2016: promoting informed healthcare choices by helping people assess treatment claims. Disponible en: https://blogs.bmj.com/bmj/2016/05/26/evidence-live-2016-promoting-informed-healthcare-choices-by-helping-people-assess-treatment-claims
13. Roundtable on Health Literacy, Board on Population Health and Public Health Practice, Institute of Medicine. Health literacy: improving health, health systems, and health policy around the world: workshop summary. Washington (DC): National Academies Press; 2013. Disponible en: https://www.ncbi.nlm.nih.gov/books/NBK202445
14. Menzin J, Lang KM, Levy P, Levy E. A general model of the effects of sleep medications on the risk and cost of motor vehicle accidents and its application to France. Pharmacoeconomics. 2001;19:69-78.
15. Austvoll-Dahlgren A, Nsangi A, Semakula D. Interventions and assessment tools addressing key concepts people need to know to appraise claims about treatment effects: a systematic mapping review. Syst Rev. 2016;5:215.
16. Semakula D, Nsangi A, Oxman M, Austvoll-Dahlgren A, Rosenbaum S, Kaseje M, et al. Can an educational podcast improve the ability of parents of primary school children to assess claims about the benefits and harms of treatments? Protocol for a randomized trial. BMC. 2017;18(31). Disponible en: https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1745-y
17. Informed Health Choices. [Sitio web]. Nsangi A, Semakula D, Oxman M., Austvoll-Dahlgren A, Rosenbaum S, Kaseje M, et al. Resources to teach children in low income countries to assess claims about treatment effects. Protocol for a randomized trial. Informed Health Choices; 2016. Disponible en: http://www.informedhealthchoices.org/wp-content/uploads/2016/08/IHC-Process-Evaluation-School-resources_final-1.pdf
18. Austvoll-Dahlgren A, Semakula D, Nsangi A, Oxman A, Chalmers I, Rosenbaum S, et al. Measuring ability to assess claims about treatment effects: the development of the 'Claim Evaluation Tools'. BMJ Open. 2017;7(5). Disponible en: https://bmjopen.bmj.com/content/7/5/e013184
19. Austvoll-Dahlgren A, Guttersrud G, Nsangi A, Semakula D, Oxman A, group. TI. Measuring ability to assess claims about treatment effects: a latent trait analysis of the "Claim Evaluation Tools" using Rasch modelling. BMJ Open. 2017;7:e013185.
20. Austvoll-Dahlgren A, Oxman AD, Chalmers I, Nsangi A, Glenton C, Lewin S, et al. Key concepts that people need to understand to assess claims about treatment effects. J Evid Based Med. 2015;8:112-125.
21. Linacre J. Sample size and item calibration (stability. Rasch Measurement Transactions. 1994;7(4):328.
22. Rasch-analysis.com. http://www.rasch-analysis.com
23. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007;57:1358-1362.
24. Guttersrud O, Dalane JO, Pettersen S. Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-specific 'critical nutrition literacy' scales. Public Health Nutr. 2014;17:877-883.
25. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. Arthritis Rheum. 2007;57:1363-1367.
26. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health. 2004;7:S22-S26.
27. Psylab Group. Introductory Rasch analysis using RUMM2030. The Section of Rehabilitation Medicine. University of Leeds; 2016

Apéndice 1

Conceptos clave que las personas necesitan entender para analizar las afirmaciones acerca de los efectos de los tratamientos

1. Reconocer la necesidad de comparaciones imparciales de tratamientos

Las decisiones bien informadas sobre tratamientos necesitan información fiable. No todas las afirmaciones sobre los efectos de los tratamientos son confiables.

Conceptos	Explicaciones	Implicaciones
1.1 Los tratamientos podrían llegar a hacer daño	Las personas frecuentemente exageran los beneficios de tratamientos e ignoran o menosprecian los potenciales daños. Sin embargo, pocos tratamientos son 100% seguros.	Siempre debe considerarse la posibilidad de que un tratamiento tenga efectos dañinos.
1.2 Las experiencias personales o anécdotas (historias) son una base poco confiable para analizar los efectos de la mayoría de los tratamientos	Las personas frecuentemente creen que las mejorías en un problema de salud (por ejemplo, recuperación de una enfermedad) se deben a un tratamiento. De forma similar, pueden creer que un desenlace de salud indeseado se debe al tratamiento. Sin embargo, el hecho de que un individuo mejoró después de recibir un tratamiento no significa que el tratamiento causó la mejoría, o que otro que recibe el mismo tratamiento también mejorará. La mejoría (o al revés, el desenlace de salud no deseado) pudieron ocurrir aún sin tratamiento.	Las afirmaciones acerca de los efectos de un tratamiento pueden ser engañosas si se basan en historias sobre cómo ayudó a personas individuales o si esas historias atribuyen mejoras a tratamientos que no han sido evaluados en revisiones sistemáticas de comparaciones imparciales.
1.3 Un desenlace después de un tratamiento podría estar asociado con el tratamiento, pero no causado por el mismo	El hecho de que un desenlace (beneficioso o dañino) esté asociado con un tratamiento no significa que el tratamiento causó el desenlace. Por ejemplo, las personas que buscan y reciben un tratamiento podrían estar más sanas y vivir en mejores condiciones que aquellas personas que no buscan o reciben el tratamiento. Por eso, podría parecer que las personas que reciben el tratamiento se mejoran gracias a él, pero su mejoría se podría deber a estar previamente sanas o a vivir en mejores condiciones, en lugar de deberse únicamente al tratamiento.	Hasta que se hayan descartado otras razones que expliquen la asociación entre un desenlace y un tratamiento por medio de una comparación imparcial, no debe asumirse que el desenlace fue causado por el tratamiento.
1.4 Tratamientos muy usados o que han sido usados desde hace mucho tiempo no necesariamente son benéficos o seguros	Hay tratamientos que no han sido evaluados apropiadamente pero como han sido usados ampliamente y por mucho tiempo se asume que funcionan. A veces, estos tratamientos podrían ser dañinos o de cuestionable beneficio.	No debe asumirse que los tratamientos son beneficiosos o seguros solo por el hecho de que son muy usados o se han empleado desde hace mucho tiempo, a menos que esto se haya encontrado en una revisión sistemática de comparaciones imparciales del tratamiento.
1.5 Tratamientos nuevos, de marca, o caros, pudieran no ser mejores que las alternativas existentes	Muchas veces se asume que los tratamientos nuevos son mejores solo por ser nuevos o por su alto costo, sin embargo, solo existe una pequeña posibilidad de que sean mejores que los existentes. Algunos efectos secundarios tardan en aparecer y no se puede saber si aparecerán hasta que se haga un seguimiento a largo plazo.	No se debe asumir que un tratamiento es beneficioso o seguro solo por el hecho de que es nuevo, de marca o caro.
1.6 Las opiniones de expertos o autoridades por sí solas no son una base fiable para decidir sobre los beneficios o daños de los tratamientos	Doctores, investigadores, organizaciones de pacientes y otras autoridades muchas veces están en desacuerdo acerca de los efectos de los tratamientos. Esto puede ser a que sus opiniones no siempre se basan en revisiones sistemáticas de comparaciones imparciales de tratamientos.	No debe confiarse en las opiniones de expertos o autoridades sobre los efectos de los tratamientos, a menos que su opinión claramente se base en los hallazgos de revisiones sistemáticas de comparaciones imparciales de tratamientos.

Conceptos	Explicaciones	Implicaciones
1.7 Los conflictos de intereses pueden resultar en afirmaciones tendenciosas acerca de los efectos de los tratamientos	Las personas con interés en promover un tratamiento (además de querer ayudar a las personas), como ganar dinero, promover tratamientos exagerando los beneficios e ignorando efectos dañinos potenciales. De manera inversa, personas podrían oponerse a un tratamiento por una variedad de razones, como prácticas culturales.	Debe investigarse si las personas que hacen afirmaciones sobre tratamientos tienen conflictos de intereses. Si los tienen, no hay que dejarse llevar por sus afirmaciones acerca de los efectos de esos tratamientos.
1.8 Aumentar la cantidad de un tratamiento no necesariamente significa que su beneficio se va a incrementar y podría causar daño	Aumentar la dosis de un tratamiento (por ejemplo, la cantidad de pastillas de vitaminas) frecuentemente causa daño sin aumentar los efectos benéficos.	Si se cree que un tratamiento es beneficioso, no debe asumirse que cuanto más, mejor.
1.9 Una detección temprana no necesariamente es mejor	Las personas frecuentemente asumen que la detección temprana de una enfermedad conlleva mejores desenlaces. Sin embargo, cribar personas para detectar enfermedades solo funciona si se cumplen dos condiciones. Primero, debe existir un tratamiento efectivo. Segundo, le debe ir mejor a las personas tratadas antes de que la enfermedad se haga aparente. Las pruebas de tamizaje pueden ser poco exactas (por ejemplo, pueden clasificar a personas sanas como si estuvieran enfermas), causar daño por etiquetar a personas como si estuvieran enfermas cuando no lo están y por los efectos adversos de las pruebas diagnósticas o de los tratamientos.	No debe asumirse que una detección temprana vale la pena si no se ha evaluado en revisiones sistemáticas de comparaciones imparciales entre personas que han sido cribadas y personas que no lo han sido.
1.10 La esperanza o el miedo pueden generar expectativas irreales acerca de los efectos de los tratamientos	La esperanza es buena, pero a veces las personas necesitadas o desesperadas desean que los tratamientos funcionen y asumen que no harán ningún daño. Del mismo modo, el miedo puede llevar a la gente a usar tratamientos que pueden no funcionar o causar daño. Como resultado, es posible que se pierda tiempo y dinero en tratamientos cuya utilidad no se ha demostrado o que pueden causar daño.	Un tratamiento no es beneficioso o seguro, o vale la pena cueste lo que cueste, simplemente por la creencia o la esperanza de que puede ser útil.
1.11 Las creencias de cómo funcionan los tratamientos no son predictores confiables de los verdaderos efectos de los tratamientos	Los tratamientos que deberían funcionar en teoría frecuentemente no funcionan en la práctica, o incluso puede resultar dañinos. Una explicación de cómo o porqué un tratamiento podría funcionar no prueba que verdaderamente funciona o sea seguro.	No debe asumirse que las afirmaciones acerca de un tratamiento basadas en una explicación de cómo podría funcionar son correctas si no se han evaluado en una revisión sistemática de pruebas imparciales sobre tratamientos.
1.12 Son raros los efectos grandes y dramáticos de tratamientos	Efectos grandes (en los que todos o casi todos los que reciben un tratamiento experimentan un beneficio o un daño) son fáciles de detectar sin comparaciones imparciales, pero pocos tratamientos tienen efectos tan grandes que no requieran comparaciones imparciales.	Las afirmaciones de grandes efectos probablemente sean erróneas. Hay que esperar que tengan efectos moderados, pequeños o triviales, no tanto efectos dramáticos. No debe confiarse en las afirmaciones sobre efectos pequeños o moderados de un tratamiento si no provienen de una revisión sistemática de comparaciones imparciales de tratamientos.

2. Juzgar si una comparación entre tratamientos es imparcial

Las decisiones bien informadas sobre tratamientos requieren comparaciones imparciales de tratamientos, es decir, comparaciones diseñadas para minimizar el riesgo de errores. No todas las comparaciones de tratamientos son imparciales.

Conceptos	Explicaciones	Implicaciones
2.1 Evaluar los efectos de tratamientos requiere comparaciones apropiadas	Si un tratamiento no se compara con otra cosa, no es posible saber qué hubiera pasado sin el tratamiento, por lo tanto, es difícil atribuirle un desenlace a este.	Siempre debe preguntare cuáles son las comparaciones cuando se consideren afirmaciones sobre los efectos de tratamientos. No son confiables las afirmaciones que no se basan en comparaciones apropiadas.
2.2 Aparte de que se comparen los tratamientos, los grupos en los que se comparan deben ser similares	Si las personas en los grupos de comparación de los tratamientos difieren en otras cosas además del tratamiento que está siendo comparado, los aparentes efectos de podrían deberse a esas diferencias y no a efectos reales de los tratamientos. Las diferencias en las características de las personas en los grupos comparados podrían resultar en estimaciones de los efectos de tratamientos que parecen mayores o menores de las que realmente son. Un método como el asignar las personas a los diferentes tratamientos usando números aleatorios (el equivalente a lanzar una moneda al aire) es la mejor manera de asegurar que los grupos que se están comparando sean similares en términos de características estudiadas e incluso no estudiadas.	No debe confiarse en los resultados de comparaciones de tratamientos que no sean al azar (por ejemplo, en las que las personas que se van a comparar escojan qué tratamiento recibir). Es necesario tener cuidado adicional si no se está seguro de que las características de los grupos comparados son similares. Si las personas no fueron asignadas a su tratamiento al azar, hay que preguntar si hubo diferencias importantes en los grupos que pudieran haber resultado en que las estimaciones de los efectos de tratamientos parecieran más grandes o más pequeñas de lo que lo realmente son.
2.3 Lo que pase a las personas de una comparación de tratamientos debe contarse en el grupo al que fueron asignados	La asignación al azar asegura que los grupos de tratamiento y comparación tengan características similares. Sin embargo, a veces hay personas que no reciben o toman los tratamientos que les asignan. Las características de dichas personas muchas veces difieren de las que sí toman sus tratamientos asignados. Por eso, si se excluyen del análisis a las personas que no se toman sus tratamientos pudiera pasar que ya no se esté comparando igual con igual.	Es necesario tener cuidado al confiar en los resultados de comparaciones de tratamientos si los desenlaces de los pacientes no se cuentan en el grupo al que fueron asignados. Por ejemplo, en una comparación entre una cirugía y una medicina, las personas que mueren esperando la cirugía deben contarse en el grupo de cirugía aun cuando no la hayan recibido.
2.4 Las personas en los grupos que se están comparando deben de atenderse de manera similar (aparte de los tratamientos que están siendo comparados)	Aparte de los tratamientos que están siendo comparados, las personas en los grupos de tratamiento y comparación deben recibir cuidados similares. Si, por ejemplo, las personas en uno de los grupos reciben mayor atención y cuidados que las personas en el grupo de comparación, las diferencias de los desenlaces podrían deberse a la cantidad de atención que recibieron los grupos y no a los tratamientos comparados. Una manera de prevenir esto es que los que proveen el cuidado no sepan (estén "cegados") a qué tratamiento fueron asignadas las personas.	Es conveniente tener cuidado con los resultados de comparaciones de tratamientos si las personas de los grupos que están comparando no se cuidaron de manera similar (aparte de los tratamientos que estaban siendo comparados). Los resultados de comparaciones así pudieran ser engañosos.
2.5 Cuando sea posible, las personas no deben de saber cuál de los tratamientos están recibiendo	Las personas en un grupo de tratamiento pueden tener mejorías (por ejemplo, menos dolor) solo porque creen que están recibiendo un mejor tratamiento, aun cuando el tratamiento no es mejor (esto se llama efecto placebo) o porque se comportan de manera distinta (debido a que saben qué tratamiento reciben, comparado con cómo se hubieran comportado si no lo supieran). Si los individuos saben que están recibiendo (es decir, que no están "cegados") un tratamiento que ellos creen que es mejor, algunos o todos los aparentes efectos del tratamiento se podrían deber al efecto placebo o porque los participantes se han comportado diferente.	Es necesario tener cuidado con los resultados de comparaciones de tratamientos si los participantes supieron qué tratamiento estaban recibiendo. Esto puede haber afectado sus expectativas o su comportamiento. Los resultados de comparaciones así pudieran ser engañosos.

Conceptos	Explicaciones	Implicaciones
2.6 Los desenlaces deben ser medidos de la misma forma (de manera imparcial) en los grupos de tratamiento que están siendo comparados	Si un desenlace es medido de manera diferente en dos grupos de comparación, las diferencias en ese desenlace podrían deberse a cómo fue medido el tratamiento en cada grupo. Por ejemplo, si los que evalúan el desenlace creen que un tratamiento en particular funciona y saben qué pacientes recibieron ese tratamiento, pueden ser más propensos a observar mejores resultados en los que han recibido dicho tratamiento. Una manera de prevenir esto es que los evaluadores del desenlace no sepan (estén "cegados") qué personas reciben cada tratamiento.	Es necesario tener cuidado con los resultados de comparaciones de tratamientos si los desenlaces no fueron medidos de la misma manera de los diferentes grupos de tratamiento. Los resultados de estas comparaciones pudieran ser engañosos.
2.7 Es importante que se midan los desenlaces en todos los que se incluyeron en los grupos de tratamiento que están siendo comparados	Las personas a quienes no se les da seguimiento hasta el final del estudio de diferentes tratamientos podrían tener peores desenlaces que las personas a las que sí se les da seguimiento. Por ejemplo, podrían haber abandonado el estudio porque el tratamiento no estaba funcionando o porque tuvieron efectos adversos. Si esas personas se excluyen, los resultados del estudio pueden ser engañosos.	Es necesario tener cuidado con los resultados de comparaciones de tratamientos si muchas personas se perdieron en el seguimiento, o si hubo una diferencia grande en los porcentajes de pérdidas o abandonos de personas entre los grupos que están siendo comparados. Los resultados de estas comparaciones pudieran ser engañosos.

3. Entender el papel que desempeña el azar

Las decisiones bien informadas sobre tratamientos requieren información acerca del riesgo de ser engañados por el papel que desempeña el azar.

Conceptos	Explicaciones	Implicaciones
3.1 Los estudios pequeños y con pocos desenlaces usualmente no son informativos y sus resultados pueden ser engañosos	Cuando hay pocos desenlaces, las diferencias en las frecuencias de los desenlaces en los grupos de comparaciones pueden haber ocurrido por azar o atribuirse erróneamente a diferencias entre los tratamientos.	Deben tomarse con cautela los resultados de comparaciones de tratamientos que tengan pocos desenlaces. Los resultados de estas comparaciones pudieran ser engañosos.
3.2 El uso de los valores-p para indicar la probabilidad de que algo haya ocurrido por azar puede ser engañoso. Los intervalos de confianza son más informativos	La diferencia en un desenlace es la mejor manera de estimar qué tan bueno y seguro es un tratamiento (o pudiera ser si la comparación hubiera sido hecha en muchas más personas). Sin embargo, por el papel que desempeña el azar, la diferencia podría ser más grande o más pequeña. El intervalo de confianza es el rango en el cual podría caer la verdadera diferencia, después de haber tomado en cuenta el papel del azar. Aunque un intervalo de confianza (margen de error) es más informativo que un valor-p, este último se reporta frecuentemente. Los valores-p frecuentemente se interpretan mal para decir que los tratamientos tienen o no efectos importantes.	Entender un intervalo de confianza puede ser necesario para entender la confiabilidad de la estimación del efecto de un tratamiento. Cuando sea posible, hay que considerar los intervalos de confianza cuando se analicen las estimaciones de los efectos de tratamientos. No hay que dejarse llevar por los valores-p.
3.3 Decir que una diferencia es estadísticamente significativa o estadísticamente no significativa puede ser engañoso	La significación estadística comúnmente se confunde con la importancia. El punto de corte para considerar que un resultado es estadísticamente significativo es arbitrario, y resultados no significativos podrían ser informativos (al mostrar que es poco probable que un tratamiento tenga un efecto importante) o inconcluyentes (al mostrar que los efectos de un tratamiento son inciertos).	Las afirmaciones acerca de que los resultados fueron significativos o no significativos frecuentemente se refiere a si fueron estadísticamente significativos o no. Esto no es lo mismo a ser o no importantes. No hay que dejarse llevar por afirmaciones así.

4. Considerar todas las comparaciones imparciales relevantes

Las decisiones sobre tratamientos bien informadas requieren revisiones sistemáticas de la evidencia. Las revisiones que no son sistemáticas pueden ser engañosas.

Conceptos	Explicaciones	Implicaciones
4.1 Los resultados de una sola comparación de tratamientos pueden ser engañosos	Una sola comparación de tratamientos rara vez provee suficiente evidencia y frecuentemente hay resultados de otras comparaciones de los mismos tratamientos. Estas otras comparaciones pueden tener resultados diferentes o ayudar a las personas a tener estimaciones de los efectos de tratamientos más confiables y precisas.	Los resultados de una sola comparación de tratamientos pueden ser engañosos.
4.2 Las revisiones de tratamientos que no usan métodos sistemáticos pueden ser engañosas	Revisiones que no usan métodos sistemáticos pueden generar resultados tendenciosos o imprecisos de los efectos de tratamientos, porque la selección de estudios para su inclusión en la revisión puede tener sesgos, o los métodos pueden hacer que no se encuentren algunos estudios. Además, el análisis de algunos estudios puede tener sesgos, o el resumen de los resultados de los estudios que se seleccionaron pueden ser inadecuados o inapropiados.	Cuando sea posible, hay que usar revisiones sistemáticas sobre comparaciones imparciales de tratamientos en vez de revisiones no-sistemáticas para sustentar las decisiones.
4.3 Revisiones sistemáticas bien hechas frecuentemente revelan una falta de evidencia o pruebas, pero proveen la mejor base para hacer juicios acerca de la certeza de la evidencia	La certeza de las pruebas (la medida en que la investigación proporciona una buena indicación de los posibles efectos de los tratamientos) puede afectar las decisiones sobre tratamientos que las personas toman. Por ejemplo, alguien podría decidir no usar o pagar por un tratamiento si la certeza que tiene sobre la evidencia es baja o muy baja. Qué tanta certeza existe sobre las pruebas depende de la imparcialidad de las comparaciones, el riesgo de haber sido engañado por el papel del azar, y por qué tan directamente relevante es la evidencia. Las revisiones sistemáticas proveen la mejor base para estos juicios y deben reportar un análisis de la certeza que se tiene sobre las pruebas con base en estos juicios.	Cuando se usen los hallazgos de una revisión sistemática para sustentar las decisiones, siempre hay que considerar el grado de certeza que se tiene sobre las pruebas.

5. Entender los resultados de las comparaciones imparciales de los tratamientos

Las decisiones bien informadas sobre los tratamientos requieren información acerca del tamaño de los efectos. Los resultados de las investigaciones pueden ser presentados de formas engañosas.

Conceptos	Explicaciones	Implicaciones
5.1 Los tratamientos frecuentemente tienen efectos beneficiosos y dañinos	Como los tratamientos pueden tener efectos beneficiosos y dañinos, las decisiones deben ser informadas por el balance entre beneficios y riesgos. Los costes también deben considerarse.	Hay que tener siempre en cuenta las ventajas y desventajas entre los beneficios potenciales, los daños potenciales y los costos de los tratamientos.
5.2 Los efectos relativos de los tratamientos por sí solos pueden ser engañosos	Los efectos relativos (ej. la relación de la probabilidad de un desenlace en un grupo de tratamiento comparado con la del grupo de comparación) son insuficientes para juzgar la importancia de la diferencia (entre las probabilidades del desenlace). Un efecto relativo puede dar la impresión de que la diferencia es mayor de lo que es en realidad cuando la probabilidad del desenlace es pequeña desde un inicio. Por ejemplo, si un tratamiento reduce la probabilidad de un infarto cerebral por 50% pero también puede ocasionar daños, y tu riesgo de desarrollar un infarto cerebral es de 2 en 100, entonces el tratamiento sí pudiera valer la pena. Pero si tu riesgo de padecer un infarto cerebral es de 2 en 10,000, entonces posiblemente no valga la pena el tratamiento aun cuando el efecto relativo sea el mismo.	Hay que tener siempre en cuenta los efectos absolutos de los tratamientos, es decir, la diferencia en los desenlaces entre los grupos que están siendo comparados. No hay que tomar decisiones sobre un tratamiento con base solo en los efectos relativos.

Conceptos	Explicaciones	Implicaciones
5.3 Las diferencias de los promedios entre tratamientos pueden ser engañosas	Para desenlaces que se miden con una escala (ej. ganancia de peso, dolor) la diferencia entre el promedio de un grupo de tratamiento y el promedio del grupo de comparación puede no dejar claro cuántas personas experimentaron un cambio que verdaderamente valga la pena (en peso, o en dolor) como para que lo noten o para que lo consideren importante.	Cuando los desenlaces se miden en una escala, no se puede suponer que todas las personas experimentaron el efecto promedio del tratamiento.

6. Juzgar si las comparaciones imparciales de los tratamientos son relevantes

Las decisiones bien informadas sobre tratamientos requieren información relevante. Los resultados de comparaciones imparciales pudieran no ser relevantes para un individuo determinado.

Conceptos	Explicaciones	Implicaciones
6.1 Las comparaciones imparciales de tratamientos deben medir desenlaces importantes	Pacientes, profesionales de la salud e investigadores pueden tener diferentes puntos de vista en cuanto a qué desenlaces son importantes. Los estudios frecuentemente miden desenlaces como irregularidades del latido cardiaco como sustitutos de desenlaces importantes como infarto cardiaco o muerte. Sin embargo, los efectos de tratamientos en desenlaces sustitutos o indirectos no necesariamente proveen un indicador confiable de los efectos en los desenlaces importantes.	Hay que considerar con cautela los desenlaces indirectos o sustitutos.
6.2 Las comparaciones imparciales de tratamientos realizadas en animales o grupos de personas altamente selectivos podrían no ser relevantes	Las revisiones sistemáticas de estudios que solo incluyen animales o a una minoría muy selectiva de personas son poco probables de proveer resultados que sean relevantes para la mayoría de las personas.	Los resultados de revisiones sistemáticas de estudios en animales o grupos de personas muy selectivos pueden ser engañosos.
6.3 Los tratamientos evaluados en comparaciones imparciales podrían no ser relevantes o aplicables	Una comparación imparcial sobre los efectos de un procedimiento quirúrgico realizado en un hospital especializado podría no proveer de una estimación confiable sobre los efectos de dicho procedimiento en otros entornos. Igualmente, comparar un nuevo medicamento con un medicamento o una dosis que no se usa comúnmente (y que pudiera ser menos efectiva o segura que las que comúnmente se usan) no daría una buena estimación de cómo se compararía el nuevo medicamento con lo que comúnmente se hace.	Si las circunstancias de quienes consultan una información son suficientemente diferentes, los resultados de revisiones sistemáticas de comparaciones imparciales podrían no ser aplicables a esos individuos.
6.4 Los resultados de un grupo selectivo de personas dentro de comparaciones imparciales pueden ser engañosos	Las comparaciones de tratamientos a menudo reportan resultados para un grupo seleccionado de participantes en un esfuerzo para evaluar si el efecto de un tratamiento es diferente para diferentes tipos de personas (por ejemplo, hombres y mujeres o diferentes grupos etarios). Estos análisis frecuentemente están mal planeados y reportados. La mayoría de los efectos sugeridos por estos "resultados de subgrupos" posiblemente se deben al papel del azar y no reflejan efectos verdaderos.	Los hallazgos basados en resultados de subgrupos de personas dentro de una comparación de tratamientos pueden ser engañosos.

Glosario

Asignación	Es la asignación de los participantes en comparaciones de tratamientos a los diferentes tratamientos (grupos) que están siendo comparados.
Asociación	Asociación es la relación entre dos atributos, por ejemplo usar un tratamiento y tener un desenlace.
Azar	En el contexto de comparaciones de tratamientos, el azar es la ocurrencia de diferencias entre los grupos comparados que no se deben a los efectos de los tratamientos o a los sesgos. El papel del azar (error aleatorio) puede llevar a conclusiones incorrectas sobre los efectos de los tratamientos si muy pocos desenlaces ocurren en los estudios.
Certeza de las pruebas	La certeza de las pruebas, o de la evidencia, es un análisis de qué tan buena indicación provee una revisión sistemática sobre el efecto de un tratamiento. Por ejemplo, la posibilidad de que el efecto sea sustancialmente diferente de lo que se encontró en los estudios (suficientemente diferente como para afectar una decisión). Los juicios acerca de la certeza de las pruebas se basan en factores que pueden reducir esta certeza (el riesgo de sesgos, inconsistencias, evidencia indirecta, imprecisión y sesgo de publicación) y factores que pueden incrementar la certeza.
Comparación de tratamientos	Las comparaciones de tratamientos son estudios sobre los efectos de los tratamientos.
Confiable	La fiabilidad de una afirmación o una prueba sobre un tratamiento es la medida en la que es fiable o puede ser de confianza. Cabe señalar que la fiabilidad a menudo tiene un significado diferente en el contexto de la investigación, en el cual es el grado en que los resultados obtenidos por un procedimiento de medición se pueden replicar.
Desenlace	Un desenlace es un beneficio o un daño potencial de un tratamiento medido en una comparación de tratamientos. Una medida de desenlace es cómo el desenlace fue medido en un estudio.
Desenlaces sustitutos o indirectos	Un desenlace indirecto o sustituto es una medida del desenlace que no es de importancia práctica directa pero que se cree que refleja desenlaces importantes. Por ejemplo, la presión arterial no es directamente importante para los pacientes pero a menudo se usa como un desenlace en los estudios porque es un factor de riesgo para infartos cardiacos o cerebrales.
Diferencia promedio	La diferencia promedio se usa para expresar diferencias de tratamientos en desenlaces continuos como peso, presión arterial o dolor medido en una escala. Es la diferencia entre el valor promedio de una medida de desenlace (por ejemplo, kilogramos) en un grupo de tratamiento y el valor promedio en el grupo de comparación.
Efecto placebo	Es un beneficio sentido, medido u observado en la salud o el comportamiento, no atribuible al tratamiento administrado.
Efectos absolutos	Los efectos absolutos son las diferencias entre desenlaces en los grupos comparados, por ejemplo, si 10 % (10 por 100) experimenta un desenlace en uno de los grupos de comparación y 5 % (5 por 100) experimenta el desenlace en el otro grupo, el efecto absoluto sería 10 %-5% = una diferencia de 5 %.
Efectos relativos	Los efectos relativos son razones o relaciones. Por ejemplo, si 10 % (10 por 100) experimenta un desenlace en uno de los grupos de comparación y 5 % (5 por 100) experimenta el desenlace en el otro grupo, el efecto relativo sería 5/10 = 0.50.
Escala	Una escala es un instrumento para medir o calificar un desenlace que pudiera tener un número infinito de posibles valores en un rango determinado, como el peso, la presión arterial, el dolor o la depresión.
Estudio	Un estudio es una investigación que utiliza métodos específicos para evaluar algo. Diferentes tipos de estudios pueden ser utilizados para evaluar los efectos de los tratamientos. Algunos son más fiables que otros.
Intervalo de confianza	Un intervalo de confianza es una medida estadística del rango en el cual existe una alta probabilidad (usualmente 95 %) de que caiga el valor verdadero. Intervalos de confianza amplios indican menor confianza, e intervalos de confianza estrechos indican mayor confianza.
Placebo	Un placebo es un tratamiento que no contiene ingredientes activos que ha sido diseñado para ser indistinguible del tratamiento activo que está siendo evaluado.
Probabilidad	Una probabilidad es la posibilidad de que algo pase, como el riesgo de que un desenlace ocurra. Ver riesgo.
Prueba imparcial	Las pruebas imparciales de tratamientos son comparaciones diseñadas para minimizar el riesgo de errores sistemáticos (sesgos) y errores aleatorios (que resultan por el papel del azar).
Revisión sistemática	Una revisión sistemática es un sumario de las pruebas generadas por investigaciones (estudios) que usan métodos sistemáticos y explícitos para resumir los hallazgos. Estas revisiones abordan una pregunta claramente formulada usando un abordaje estructurado para identificar, seleccionar y evaluar críticamente estudios relevantes, y para recoger y analizar datos de los estudios incluidos en la revisión.

Riesgo	Riesgo es la probabilidad de que un desenlace ocurra. Ver probabilidad.
Significación estadística	La significación estadística es una diferencia poco probable (por debajo de un determinado nivel de confianza-típicamente 5%) para ser explicada por el papel de azar.
Subgrupo	Un subgrupo es una subdivisión de un grupo de personas; un grupo selecto dentro de un grupo, por ejemplo, en estudios o revisiones sistemáticas de los efectos de tratamiento a menudo se hacen preguntas sobre si existen diferentes efectos para diferentes subgrupos de personas, como las mujeres y los hombres, o las personas de diferentes edades.
Teoría	Una teoría es una suposición o un conjunto de ideas que intentan explicar algo.
Tratamiento	Un tratamiento es cualquier intervención (acción) que intenta mejorar la salud, incluyendo intervenciones preventivas, terapéuticas o de rehabilitación e intervenciones de salud pública o sistemas de salud.
Valor-p	El valor-p es la probabilidad (que va del cero al uno) de que los resultados observados en un estudio (o resultados más extremos) pudieran haber pasado solo por azar si realmente no existiera diferencia entre los tratamientos.