Máster Universitario en PSICOLOGÍA GENERAL SANITARIA

Trabajo Final de Máster (TFM)

REVISIÓN SISTEMÁTICA DE LAS PROPIEDADES PSICOMÉTRICAS DEL *PATIENT HEALTH QUESTIONNAIRE-4* (PHQ-4)

Presentado por:

Sheila Caro Fuentes

Director:

Juan P. Sanabria-Mazo

Curso: 2022-2023

Convocatoria: Octubre

ÍNDICE DE CONTENIDOS

LISTA DE SIGLAS

Siglas	Significado
AAQ-II	The Acceptance and Action Questionnaire
ACP	Análisis de Componentes Principales
AFC	Análisis Factorial Confirmatorio
AFE	Análisis Factorial Exploratorio
BAI	Beck Anxiety Inventory
BDI	Beck Depression Inventory
BSI	Brief Symptom Inventory
CFI	Comparative Fit Index
DASS	Depression, Anxiety, and Stress Scale
DSM	Diagnostic Statistical Manual of Mental Disorders
GAD	Generalized Anxiety Disorder Scale
GHQ	General Health Questionnaire
GSES	General Self-Efficacy Scale
HADS	Hospital Anxiety and Depression Scale
HAM-A	Hamilton Rating Scale for Anxiety
HAM-D	Hamilton Rating Scale for Depression
NFI	Normed fit index
PANAS	Positive and Negative Affect Schedule
PCQ	Psychological Capital Questionnaire
PHQ	Patient Health Questionnaire
PIB	Producto Interior Bruto
PHQ-ADS	Patient Health Questionnaire Anxiety-Depression Scale
PRISMA	Systematic Reviews and Meta-Analyses
PROSPERO	Prospective Register of Systematic Reviews
PROMIS-ADS	Patient Reported Outcomes Measurement Information System Depression and Anxiety Scale
PSS	Perceived Stress Scale
PSWQ	Penn State Worry Questionnaire
QLS	Questions on Life Satisfaction
RMSEA	Root mean square error of approximation
RS	Resilience Scale
RSES	Rosenberg Self-Esteem Scale
SF-20	Medical Outcomes Study Short-Form General Health Survey
TLI	Tucker-Lewis index
WHO-5	World Health Organisation-Five Well-Being Index

Resumen

Antecedentes. El *Patient Health Questionnaire-4* (PHQ-4) es un instrumento breve que mide síntomas de depresión y ansiedad. Aunque se utiliza frecuentemente, ninguna investigación ha revisado sistemáticamente sus propiedades psicométricas. **Objetivo.** El objetivo principal fue explorar las propiedades psicométricas (fiabilidad y validez) del PHQ-4. **Método.** Este estudio se ajustó según el *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) y se registró en el *Prospective Register of Systematic Reviews* (PROSPERO). Se revisaron cuatro bases de datos (Medline, PsycINFO, Web of Science y SCOPUS), en las que se exploraron artículos publicados entre el 2009 y 2023. Dos revisores realizaron el cribado, la extracción de datos y la evaluación de calidad metodológica de los estudios incluidos. **Resultados.** Se incluyeron 22 estudios con 80129 participantes, con muestras clínicas y no clínicas. La estructura factorial reportada con mayor frecuencia fue la estructura bifactorial, siendo invariante según el género, edad, localización geográfica, ingresos, nivel educativo e idioma. La consistencia interna fue adecuada para el PHQ-4 global (α desde. 72 a .88), así como para el PHQ-2 (α desde .65 a .81) y el GAD-2 (α desde .74 a .84). La estabilidad temporal del instrumento se comprobó mediante la fiabilidad test re-test, reportando una buena correlación entre ambas medidas. Las correlaciones con medidas potencialmente relacionadas fueron significativas y en la dirección esperada. **Conclusión.** El PHQ-4 es un instrumento fiable y válido para cribado de síntomas de depresión y ansiedad en el ámbito de salud, tanto para población clínica como no clínica.

Código de identificación en PROSPERO: CRD42022381809.

Palabras clave. Patient Health Questionnaire-4, PHQ-4, depresión, ansiedad, revisión sistemática, propiedades psicométricas.

Abstract

Background. The Patient Health Questionnaire-4 (PHQ-4) is a brief instrument that measures symptoms of depression and anxiety. Although it is frequently used, no research has systematically reviewed its psychometric properties. **Aim**. The main objective was to explore the psychometric properties (reliability and validity) of the PHQ-4. **Method**. This study was adjusted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and found in the Prospective Register of Systematic Reviews (PROSPERO). Four databases (Medline, PsycINFO, Web of Science, and SCOPUS) were reviewed, in which articles published between 2009 and 2023 were explored. Screening, data extraction, and methodological quality assessment of the studies were performed by two reviewers. Including **Results**. Twenty-two studies with 80,129 participants, with clinical and non-clinical samples, were included. The most frequently reported factorial structure was the bifactorial structure, being invariant according to gender, age, geographic location, income, educational level, and language. The internal consistency was adequate for the global PHQ-4 (α from .72 to .88), as well as for the PHQ-2 (α from .65 to .81) and the GAD-2 (α from .74 to .84). The temporal stability of the instrument was verified through test-retest reliability, reporting a good connection between both measures. Correlations with related measures were significant and in the expected direction. **Conclusion**. The PHQ-4 is a reliable and valid instrument for screening depression and anxiety symptoms in the health field, both for the clinical and non-clinical population.

Identification code in PROSPERO: CRD42022381809.

Keywords. Patient Health Questionnaire-4, PHQ-4, depression, anxiety, systematic review, psychometric properties.

1. Introducción

1.1. Contextualización del estudio

La depresión y la ansiedad se han convertido en un problema de interés creciente para clínicos e investigadores en los últimos años. Informes recientes han destacado que la prevalencia de estos trastornos emocionales en pacientes ambulatorios y en población general son elevados y han señalado, además, la necesidad de disponer de herramientas clínicas que faciliten su detección e intervención temprana (Paykel et al., 2005; Remes et al., 2016; Wittchen et al., 2011).

Este estudio tiene como objetivo revisar sistemáticamente las propiedades psicométricas (fiabilidad y validez) del *Patient Health Questionnaire* (PHQ-4), un instrumento breve de cribado de síntomas de ansiedad y depresión. En las siguientes líneas se definen las principales características de ansiedad y depresión y se reporta la prevalencia, la comorbilidad y las implicaciones de estos trastornos en el sistema sanitario. Adicionalmente, en este trabajo se justifica la importancia de disponer de instrumentos de cribado breves, fiables y válidos, como el PHQ-4. Finalmente, se detallan las propiedades psicométricas exploradas en este estudio de revisión.

1.2. Definición de depresión y ansiedad

La depresión es un trastorno mental afectivo de alta prevalencia en la población (Wittchen et al., 2011). Este trastorno está caracterizado por desórdenes en el estado de ánimo, en las que se presentan síntomas cognitivos y físicos. Concretamente, en la quinta edición del *Diagnostic Statistical Manual of Mental Disorders* (DSM-5) se clasifica el trastorno en función de la temporalidad y del origen de los síntomas, distinguiéndose entre el trastorno depresivo mayor, la distimia y los trastornos bipolares.

Por su parte, la ansiedad se considera una respuesta subjetiva que produce reacciones psicofisiológicas, alerta al organismo y activa mecanismos que preparan a las personas para la acción (Forcadell et al., 2019). Esta reacción típica puede patologizarse y producir trastornos de ansiedad, cuya característica principal es una preocupación desproporcionada y excesiva que genera un malestar intenso en las personas y que influye en su funcionamiento (Tayeh et al., 2016). En la actualidad, el DSM-5 clasifica los trastornos de ansiedad en 11, siendo estos el ”trastorno de ansiedad por separación, mutismo selectivo, fobias específicas, trastorno de ansiedad social, trastorno de pánico, agorafobia, trastorno de ansiedad generalizada, trastorno de ansiedad inducido por sustancias y/o medicamentos, trastorno de ansiedad debido a otra afección médica, trastornos de ansiedad especificados y trastornos de ansiedad no especificados que comparten características clínicas como la angustia y el miedo, aunque dependiendo del tipo se presentan síntomas particulares” (Carballo et al., 2019, p. 2-3).

1.3. Prevalencia de depresión y ansiedad

La ansiedad y la depresión son dos de los trastornos emocionales más comunes (Organización Mundial de la Salud, 2017). Ambas condiciones afectan significativamente a la funcionalidad y al nivel de vida de las personas que sufren estas condiciones de salud (Kelly y Mezuk, 2017).

En un estudio de actualización del estado de salud mental en Europa en 2010 (Wittchen et al., 2011), se estimó que cerca de un 40% de los residentes en Europa sufría alguna enfermedad psicológica, siendo los más frecuentes los trastornos de ansiedad (14%) y el trastorno de depresión mayor (7%).

Sobre los trastornos depresivos, en una revisión sistemática se reportó una prevalencia de depresión cercana al 9% en población general (Cardila et al., 2015) y en un metaanálisis, en el que se revisaron 68 investigaciones de 30 países diferentes, se indicó que aproximadamente un 13% de la población general presenta este trastorno (Lim et al., 2018). Estudios recientes advierten que la carga de enfermedad asociada con estos trastornos aumentará en los próximos años. Además, se espera que la depresión sea la principal causa de incapacidad en todo el mundo para el año 2030 (Vos et al., 2018).

Con respecto a los trastornos de ansiedad, se estima que un 14% de la población general adulta cumple con criterios para ser diagnosticado (Penninx et al., 2021), reportándose cifras en algunos estudios de que aproximadamente 260 millones de personas están afectadas por este trastorno (Organización Panamericana de la Salud, 2017).

Concretamente, la crisis sanitaria de COVID-19 ha generado un incremento significativo en los problemas de salud mental pública, incidiendo en los síntomas de depresión y de ansiedad (Nicolini, 2020; Salari et al., 2020) y visibilizando la necesidad de una detección rápida y eficaz de estos trastornos (Sanabria-Mazo et al., 2023).

1.4. Comorbilidad ansiedad y depresión

Diversas investigaciones han aportado evidencia sobre la elevada comorbilidad entre ansiedad y depresión en diversas poblaciones, tanto con muestras clínicas como con muestras no clínicas (Alonso et al., 2019; Kessler et al., 2007). Recientemente, en un metaanálisis se encontró que la mayoría de las personas que presentan síntomas depresivos a su vez también presentaban síntomas ansiosos, respaldando la alta comorbilidad entre ambos trastornos y destacando la importancia de tratar ambas condiciones simultáneamente (Goodwin, 2015).

1.5. Detección de síntomas de ansiedad y depresión

En la actualidad, la depresión y la ansiedad suponen el problema más común por el que se acude a una consulta en atención primaria según un estudio donde se analizaron los datos

de 41 países distintos (Patten et al., 2020). Sin embargo, varios profesionales de la salud manifiestan que las dificultades de evaluación podrían generar un infradiagnóstico, habiéndose encontrado que la tasa de este en países europeos es del 30% (Fekadu et al., 2022). En un estudio llevado a cabo en Estados Unidos una muestra perteneciente a 15 centros de atención primaria completó un cuestionario de cribado de síntomas de ansiedad. Los resultados de este estudio indicaron que un 47% no habían sido tratados, un 21% recibió un tratamiento farmacológico, un 7% recibió terapia y un 24% un tratamiento mixto (Weisberg et al., 2007).

Otra cuestión que afecta al infradiagnóstico de los casos de ansiedad y de depresión podría ser la gran carga asistencial y la escasez temporal para la asistencia en el ámbito de la atención primaria, siendo la temporalidad media por consulta menor a diez minutos (Latorre Postigo et al., 2012). En este contexto, se considera que España invierte pocos recursos económicos para el gasto sanitario de salud mental (alrededor de un 5%), tratándose de un país en el que se prevé que los trastornos psicológicos aumentarán alrededor de un 20% tras la pandemia del COVID-19 (Cerdá, 2021).

Con respecto a las implicaciones económicas, se ha estimado que los gastos derivados de los trastornos psicológicos en España en 2010 se aproximaron a los 84 mil millones de euros anuales, lo que supone cerca de un 8% del producto interno bruto (PIB), superando así el gasto sanitario total (69 mil millones de euros, un 6% del PIB de 2010). Concretamente, los gastos destinados a los trastornos depresivos y de ansiedad fueron de aproximadamente 23 mil millones de euros, un 2% del PIB (Parés-Badell et al., 2014).

1.6. Instrumentos para medir síntomas de ansiedad y depresión

Por las razones mencionadas, diferentes grupos de investigación han aumentado sus esfuerzos para desarrollar herramientas de evaluación válidas y fiables que permitan detectar síntomas de ansiedad y depresión. Dos de los instrumentos que más se han utilizado para este propósito son el *Patient Health Questionnaire-9* (PHQ-9) y el *Generalized Anxiety Disorder Scale* (GAD-7), ambas medidas disponibles en más de 100 traducciones (Kroenke et al., 2019).

Aunque existen una gran diversidad de instrumentos para medir síntomas de depresión y de ansiedad, cada vez se destaca más la necesidad de disponer de herramientas de cribado ultra breves que permitan disminuir los errores en el diagnóstico (Castro-Rodríguez et al., 2015), agilizar los tiempos de espera de la sanidad pública dando un mejor servicio al paciente (Cano-Vindel et al., 2018) y reducir el coste económico (Latorre et al., 2012).

1.7. Patient Health Questionnaire-4 (PHQ-4)

Uno de los instrumentos ultra breves más utilizados en el ámbito clínico y en investigación para la evaluación de las condiciones de salud que se han descrito en los apartados anteriores, es el *Patient Health Questionnaire-4* (PHQ-4; Kroenke et al., 2009; Maurer et al., 2018;

Stanhope, 2016). El PHQ-4 es un instrumento compuesto por cuatro preguntas: dos de ellas de cribado para medir síntomas de ansiedad (extraídas del GAD-2) y las otras dos para síntomas de depresión (extraídas del PHQ-2) (Wicke et al., 2022). El PHQ-2 es una versión breve del PHQ-9 que evalúa síntomas centrales de la depresión como la disminución del interés en la persona y el estado de ánimo deprimido (Kroenke et al., 2009). La validez de esta escala se ha demostrado en diversos estudios y en diferentes poblaciones (Wicke et al., 2022). Por otro lado, la escala GAD-2 es una versión breve del GAD-7 que evalúa síntomas centrales de la ansiedad como la preocupación constante y los sentimientos nerviosos (Kroenke et al., 2007). Esta versión ha demostrado también su validez en distintas poblaciones y en estudios de carácter psicométrico, con una alta fiabilidad para el cribado de trastornos de ansiedad generalizada, trastornos de estrés postraumático, ansiedad social y trastornos de pánico (Wicke et al., 2022).

El PHQ-4 se responde mediante una escala tipo Likert con 4 puntuaciones (0 = nunca; 1 = varios días; 2 = más de la mitad de los días; 3 = casi todos los días). La puntuación total está comprendida entre 0 y 6 para cada dimensión, siendo 0 ningún nivel de síntomas de ansiedad o depresión y 6 un alto nivel de ansiedad y de depresión (Guerra et al., 2022). Además, es posible el cálculo de una puntuación absoluta para el PHQ-4, que corresponde a la evaluación de distrés o angustia psicológica, obteniéndose una puntuación de 6 o más en toda la escala (Cano-Vindel et al., 2018; Kocalevent et al., 2014; Löwe et al., 2010; Mendoza et al., 2022; Sanabria-Mazo et al., 2023).

La validez del PHQ-4 se ha demostrado en diversas investigaciones, tanto en muestras clínicas como en muestras no clínicas (Christodoulaki et al., 2022; Kocalevent et al., 2014; Löwe et al., 2010). En una investigación reciente cuyo objetivo fue actualizar la evidencia acerca de la estandarización del instrumento en población general con una muestra representativa en Alemania (incluyendo distintos grupos de edad, género y muestras tanto clínicas y como no clínicas) se concluyó la pertinencia de seguir utilizando el instrumento de evaluación al tratarse de una escala fiable (Wicke, 2022). Además, este cuestionario ha sido adaptado tanto en formato papel y lápiz como en línea, siendo evaluada su validez y fiabilidad en población de habla hispana (Cano-Vindel et al., 2018; Sanabria-Mazo et al., 2023).

1.8. Evaluación de las propiedades psicométricas

Realizar una evaluación de las propiedades psicométricas es indispensable para garantizar la validez y fiabilidad de un instrumento, ayudando a reconocer su utilidad clínica y contribuyendo a que los resultados obtenidos sirvan para decidir acerca de su utilización y en la estimación de inferencias acerca del constructo medido (Muñiz et al., 2013). En las siguientes líneas, se detallan las definiciones de las propiedades psicométricas que se exploraron en esta revisión sistemática:

1. ***Validez de contenido.*** Esta propiedad psicométrica se refiere a la capacidad de una herramienta para medir el constructo que se pretende medir. En otras

palabras, el grado en que los ítems incluidos representan la dimensión del constructo de medición (García de Yébenes Prous et al., 2009; Lamprea et al., 2007; Luján et al., 2015). Para su evaluación se utilizan métodos como el juicio de expertos o el análisis de contenido por el público objetivo. En el juicio de expertos, un grupo de profesionales en el constructo objetivo evalúan que los ítems del cuestionario midan de forma adecuada y completa el constructo de interés. En el análisis de contenido por el público objetivo se realiza la evaluación por una parte representativa de la población objetivo. La importancia de la validez de contenido radica en garantizar que un instrumento de medición mida el constructo que se supone que está midiendo con precisión y representatividad (Urrutia Egaña et al., 2014).

2. ***Estructura factorial.*** Se refiere a la dimensionalidad de la prueba En otras palabras, a la medida en que la estructura del cuestionario representa de una forma adecuada las dimensiones del constructo de medición (Terwee et al., 2007). Para su evaluación se utilizan procedimientos como el análisis factorial exploratorio (AFE) o el análisis factorial confirmatorio (AFC) (Boubeta et al., 2006). El objetivo del AFE es descubrir la estructura factorial de los datos sin supuestos previos sobre el número o la naturaleza de los factores. En el AFE, cuyo objetivo es estudiar la estructura factorial de los datos sin datos previos sobre la estructura del instrumento, se estima la varianza compartida entre los ítems y se agrupan los ítems que tienen una alta correlación en factores comunes. Estos factores se interpretan como dimensiones subyacentes o constructos latentes que explican la varianza en los datos (Martínez et al., 2012). A diferencia del análisis factorial exploratorio, el AFC parte de un modelo teórico definido y busca confirmar la existencia de los factores latentes y la estructura factorial especificada en dicho modelo (Batista-Foguet et al., 2004). Los índices de ajuste a un modelo determinado que más se utilizan son el índice de Tucker-Lewis (TLI), el índice de ajuste normado (NFI) y el índice de ajuste comparativo (CFI), mostrando estos un adecuado ajuste cuando se encuentran en el intervalo de confianza >. 90. Por otro lado, se utiliza la Raíz Cuadrada Media Error de Aproximación (RSMEA) < .08, según Schermelleh-Engel et al. (2003). La importancia de la validez factorial se encuentra en determinar su fiabilidad y validez, además de proporcionar una mejor interpretación de los resultados y dar la posibilidad de compararlos con otras escalas que tienen como objetivo medir el mismo constructo (Boubeta et al., 2006).

3. ***Consistencia interna.*** Se refiere al grado de correlación y coherencia que presentan un conjunto de ítems que miden un mismo constructo. En otras palabras, la forma en que los ítems del instrumento miden de manera coherente la dimensión del constructo de medición (Luján et al., 2015). Para su evaluación, comúnmente, se utilizan los coeficientes alfa de Cronbach, de Omega y de Lambda, indicando, cuando estos son altos, que los ítems que componen el instrumento miden de forma consistente el constructo y que el cuestionario es fiable y tiene alta consistencia. Si por el contrario

estos coeficientes son bajos, el instrumento no es adecuado para el constructo concreto que se pretende medir, encontrándose en este motivo la importancia de una prueba t que tenga una buena consistencia interna (Cascaes da Silva et al., 2015). Para conocer si la herramienta cuenta con una consistencia interna adecuada, se pueden tener en cuenta los siguientes criterios (Nunnally, 1967): (1) cuando alfa de Cronbach es igual o mayor a 0.70, la consistencia interna del instrumento es considerada aceptable; (2) cuando alfa de Cronbach es igual o mayor a 0.90 se considera redundancia o repetición en los ítems que se presentan; y (3) cuando alfa de Cronbach es menor a 0.50, la consistencia interna no se considera aceptable. Así mismo, estos criterios se deben considerar como una guía general y considerar que la validez del instrumento depende a su vez del contexto y/u objetivo de la medición.

4. ***Fiabilidad test-retest***. Se refiere a la propiedad y a la capacidad del instrumento para arrojar puntuaciones similares cuando la aplican distintos evaluadores y en distinto momento a la misma población, siguiendo el mismo método de aplicación. El objetivo es evaluar si los resultados que se obtienen en ambos momentos de medición son similares o no. Para indicar el grado de concordancia entre las puntuaciones obtenidas en ambas ocasiones se utiliza el coeficiente de correlación de Pearson, el coeficiente de correlación intraclase o el coeficiente Kappa. Cuando estos coeficientes de correlación son altos, indica que los resultados obtenidos en los distintos momentos son consistentes y, por lo tanto, el instrumento es fiable (Luján et al., 2015).

5. ***Validez de constructo.*** Esta propiedad estudia en qué medida un instrumento refleja adecuadamente el constructo que se quiere medir y, en consecuencia, es la medida en que las puntuaciones de un cuestionario en concreto se correlacionan positiva o negativamente con las medidas de otras escalas consistentes o inconsistentes con las hipótesis teóricas derivadas de los conceptos que se están midiendo. Existen dos tipos de análisis para evaluar la validez de constructo. Por un lado, la validez de constructo de convergencia, que se trata del grado en que el instrumento correlaciona de forma positiva con otras herramientas que miden el mismo constructo. Por otro lado, la validez de constructo de divergencia, que se refiere al grado en que la herramienta correlaciona de forma negativa con otros instrumentos que miden un constructo diferente o que no tienen relación. Ambas son significativas para evaluar la validez de un instrumento, mostrando la validez de convergencia que la herramienta mide realmente la variable objetivo, y la de divergencia, que no mide constructos diferentes (Luján et al., 2015).

6. ***Efectos de suelo y techo.*** Los efectos de suelo y techo en un instrumento se refieren a una medida para detectar el número de personas que obtienen la puntuación máxima posible de la prueba y el número de personas que obtienen la puntuación mínima posible. Por un lado, los efectos de suelo ocurren cuando un

instrumento no puede detectar cambios significativos en personas que presentan puntuaciones muy bajas en el constructo de medición. Por el contrario, los efectos de techo ocurren cuando es incapaz de detectar cambios significativos en individuos que puntúan muy alto en el constructo medido. La importancia de evaluar esta propiedad se encuentra en garantizar que el instrumento detecta cambios significativos en individuos que se encuentren en el extremo de una distribución (Terwee et al., 2007).

7. ***Interpretabilidad***. Esta propiedad se refiere a la capacidad del instrumento para obtener interpretaciones cualitativas a partir de datos de carácter cuantitativo. Con el objetivo de estudiar la validez del instrumento, se suele utilizar la evaluación de grupos conocidos, donde se observa el rendimiento en un grupo concreto de personas con determinadas características sociodemográficas, pudiendo observar así que la herramienta mide el constructo que se desea medir y que los resultados no varían, siendo estos válidos. En un primer momento, se selecciona un grupo de personas que se conoce que tienen una característica determinada y se les administra la herramienta. Tras ello, se comparan los resultados de la prueba con un grupo que se sabe que no presentan la característica o rasgo determinada. En este sentido, el grupo con el rasgo debería tener un rendimiento significativamente mejor en la prueba que el grupo control. Este procedimiento asegura que la herramienta sea válida y precisa, así como los resultados que se obtienen (Terwee et al., 2007).

1.9. Objetivos

El PHQ-4 es una herramienta potencialmente válida y fiable para la detección de síntomas trastornos de depresión y de ansiedad (Wicke, 2022); sin embargo, hasta la fecha ningún estudio ha revisado sistemáticamente sus propiedades psicométricas. El objetivo general de este estudio fue evaluar, por primera vez, la utilidad clínica del PHQ-4 mediante la exploración sistemática de sus propiedades psicométricas (validez de contenido, consistencia interna, estructura factorial, fiabilidad test-retest, validez de constructo, efectos de suelo y techo e interpretabilidad). Concretamente, en este trabajo se reportan los estudios publicados en la literatura hasta la actualidad. En esta investigación se evaluó, además, la calidad metodológica de los estudios incluidos en la revisión sistemática, basándose en los criterios de Terwee et al. (2007). Los hallazgos de este estudio podrían ayudar a reconocer el potencial de esta herramienta para el cribado de síntomas de depresión y de ansiedad.

2. Método

2.1. Protocolo y registro

Esta revisión sistemática se realizó siguiendo las recomendaciones del *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA). El protocolo de esta revisión se registró en el *Prospective Register of Systematic Reviews* (PROSPERO) el 5 de diciembre de 2022, con el código de identificación CRD42022381809.

2.2. Fuentes de datos y búsquedas

Las búsquedas se realizaron a través de cuatro bases de datos: Medline (PubMed), PsycINFO (ProQuest), Web of Science (Core Collection) y SCOPUS (Elsevier). La estrategia de búsqueda combinó términos relacionados con el nombre del instrumento original (*Patient Health Questionnaire OR PHQ*) con términos asociados a las propiedades psicométricas (*psychometrics OR factor analysis OR factor structure OR reliability OR intra-class OR test-retest OR internal consistency OR validity OR dimensionality OR variance OR known groups OR multi-group OR sensitivity to change OR responsiveness OR normative data OR sensibility OR specificity*). Esta búsqueda se realizó específicamente en los títulos y en los resúmenes. La estrategia de búsqueda utilizada en cada base de datos se presenta en el Anexo 1. Complementariamente, se examinó la lista de referencias de los artículos incluidos con el fin de garantizar que se incluyeran todos los estudios elegibles.

2.3. Criterios de elegibilidad

En la búsqueda en las bases de datos se incorporaron estudios publicados en inglés o en español entre el 2009 (año de publicación del PHQ-4; Kroenke et al., 2009) y el 2023 (actualidad). En esta revisión sistemática se incluyeron todos los estudios que proporcionaron evidencia sobre las propiedades psicométricas del PHQ-4 (validez de contenido, consistencia interna, estructura factorial, validez de constructo, efectos de suelo y techo e interpretabilidad), tanto en población clínica como en población no clínica. Se excluyeron los estudios no originales (editoriales, manuales y revisiones, entre otros), los reportes de caso y los estudios cualitativos. Para garantizar la rigurosidad de la evidencia explorada, los estudios publicados en revistas no arbitradas por pares evaluadores también se excluyeron.

2.4. Gestión de datos y selección de estudios

Se eliminaron automáticamente los artículos que presentaban duplicidad en las bases de datos mediante Mendeley. Después, dos revisores (SCF y JPSM) en paralelo realizaron el cribado de todos los artículos en Rayyan QCRI, basándose en sus títulos y resúmenes. De forma independiente, ambos revisores comprobaron si los textos completos cumplían los criterios de elegibilidad. Por último, la revisora principal (SFC) introdujo la información clave de cada estudio

incluido en una tabla de extracción de datos (autores, año de publicación, país, diseño del estudio, tipo y tamaño de muestra y propiedades psicométricas) y evaluó la calidad de los artículos incluidos. La información que se extrajo durante esta fase y la evaluación de calidad de los estudios fue validada posteriormente por el segundo revisor (JPSM). No se necesitó un revisor adicional para resolver ningún desacuerdo.

2.5. Evaluación de calidad y codificación

La calidad metodológica de los estudios incluidos en esta revisión sistemática se evaluó a partir de los criterios propuestos por Terwee et al. (2007) (ver Tabla 1). Los criterios que se consideraron para la evaluación y la codificación de los estudios incluidos en esta revisión sistemática se especifican a continuación, en función de cada una de las variables de interés evaluadas (propiedades psicométricas). Cada uno de los siete ítems explorados en esta herramienta se puntúa con 2 si se cumplen los criterios, con 1 si se cumplen parcialmente y con 0 si no se cumple ninguno. La puntuación total oscila entre 0 y 14. Los dos revisores consensuaron conjuntamente la calidad metodológica de los estudios incluidos. Los criterios que se consideraron se especifican a continuación, en función de cada una de las propiedades psicométricas evaluadas.

2.5.1. Validez de contenido

- Se asignará una puntuación de *"2"* si el estudio (1) proporciona una descripción clara del objetivo de la medición del cuestionario, (2) describe la población objetivo, (3) define los constructos que pretende medir, (4) detalla el procedimiento de selección de los ítems y (5) indica que participaron expertos en la selección de los elementos.
- Se asignará una puntuación de "1" si (1) faltan algunos de los aspectos mencionados anteriormente, (2) no se describe la población objetivo o (3) el diseño o método de investigación es dudoso.
- Se asignará una puntuación de "0" si no se describe la participación de la población objetivo.

2.5.2. Consistencia interna

- Se asignará una puntuación de *"2"* si en el estudio (1) se han realizado análisis factoriales con un tamaño de muestra adecuado (tamaño igual o mayor a 100), (2) se ha calculado alfa de Cronbach por cada dimensión y (3) si el alfa de Cronbach está entre 0.70 y 0.95.
- Se asignará una puntuación de "1" si no se realiza un análisis factorial y/o el diseño o método es dudoso.
- Se asignará una puntuación de "0" si no se encontró información de consistencia interna.

2.5.3. Estructura factorial

- Se asignará una puntuación de “2” si se ha llevado a cabo un análisis factorial exploratorio y un análisis factorial confirmatorio.
- Se asignará una puntuación de “1” si únicamente se ha llevado a cabo un análisis factorial exploratorio o un análisis factorial confirmatorio.
- Se asignará una puntuación de “0” si no se ha realizado ni un análisis factorial exploratorio ni confirmatorio.

2.5.4. Validez de constructo

- Se asignará una puntuación de “2” si se ha aportado información acerca de la validez convergente y de la validez divergente.
- Se asignará una puntuación de “1” si únicamente se ha aportado información acerca de uno de los conceptos mencionados anteriormente.
- Se asignará una puntuación de “0” si no se aporta ninguna información de la validez de constructo (validez convergente o discriminante).

2.5.5. Fiabilidad test-retest

- Se asignará una puntuación de “2” si el intervalo de tiempo entre la administración de los test está entre una y dos semanas.
- Se asignará una puntuación de “1” si el intervalo de tiempo entre la administración de los test es menor a una semana o mayor a dos semanas.
- Se asignará una puntuación de 0 si no se encuentra información acerca de la fiabilidad test-retest.

2.5.6. Efectos de suelo y techo

- Se asignará una puntuación de “2” si menos del 15% de los participantes lograron las puntuaciones más altas o más bajas posibles.
- Se asignará una puntuación de “1” si más del 15% de los participantes lograron las puntuaciones más altas o más bajas posibles o el método o diseño es dudoso.
- Se asignará una puntuación de “0” si no se encuentra información acerca de los efectos de suelo y techo.

2.5.7. Interpretabilidad

- Se asignará una puntuación de “2” si se reportan puntuaciones medias y desviación estándar de cuatro o más grupos normativos.
- Se asignará una puntuación de “1” si se reportan puntuaciones medias y desviación estándar de menos de cuatro grupos normativos o no se presenta desviación estándar y/o método dudoso.
- Se asignará una puntuación de “0” si no se encuentra información acerca de la interpretabilidad.

2.6. Síntesis de datos

Los resultados obtenidos en esta revisión sistemática se describieron de forma general y diferenciadas según las características de las submuestras, teniendo en cuenta si se tratan de muestras clínicas o no clínicas. Se realizó, inicialmente, una síntesis narrativa para describir las principales características de los artículos incluidos y, después, se reportaron los hallazgos obtenidos en función de cada propiedad psicométrica objetivo de evaluación.

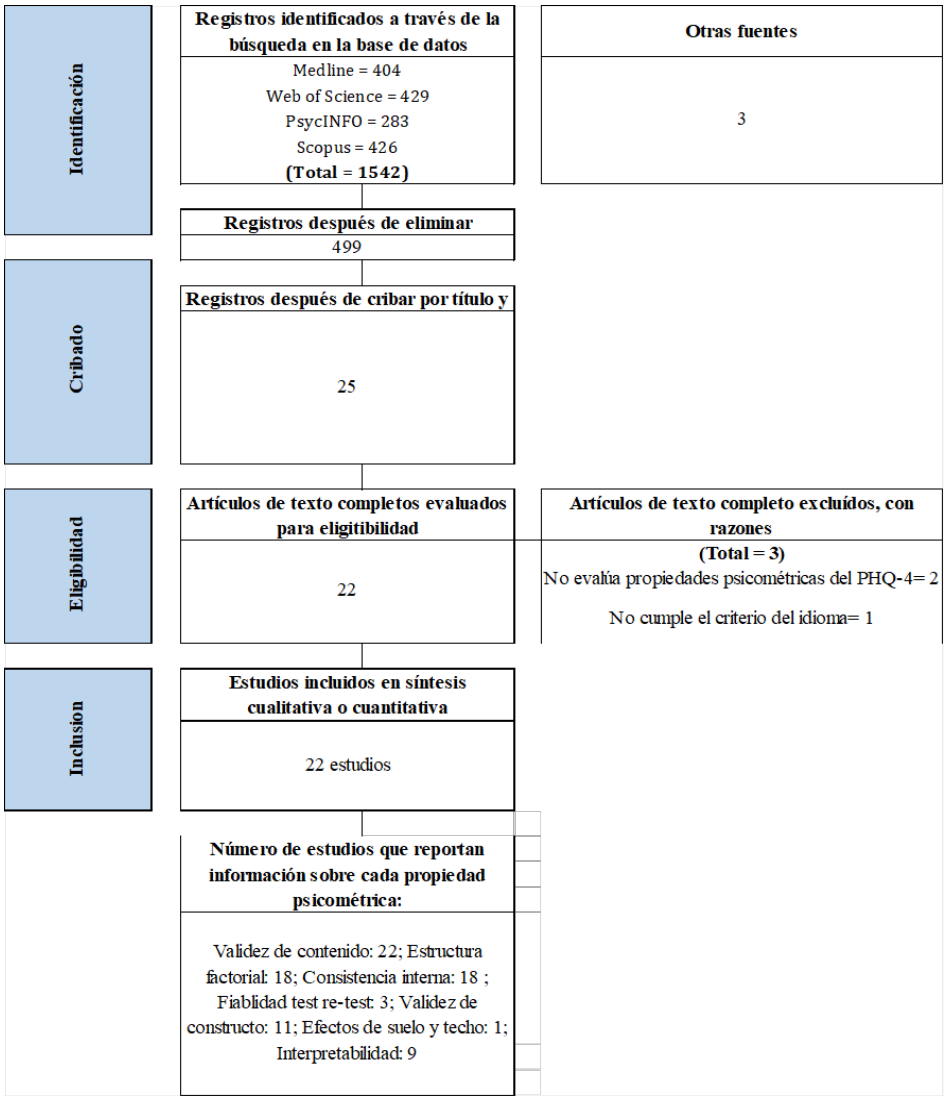
3. Resultados

3.1. Selección e inclusión de estudios

Como se muestra en la Figura 1, la primera búsqueda en las bases de datos lanzó un total de 1542 artículos publicados. Además, se incluyeron 3 artículos a partir de la revisión de la lista de referencias de los artículos incluidos. Después de eliminar los duplicados, se revisaron 496 títulos y resúmenes, de los que 25 se escogieron para la revisión del texto completo. Tras ello, se excluyeron 3 artículos, 2 de ellos por no evaluar las propiedades psicométricas del PHQ-4 y un tercero por no cumplir los criterios que se establecieron acerca del idioma. Finalmente, un total de 22 artículos formaron parte de esta revisión sistemática.

Figura 1

Diagrama de flujo de los registros identificados y seleccionados en la revisión sistemática.



Nota: Elaboración propia.

3.2. Características de los estudios incluidos

En relación con la calidad metodológica de los estudios, la calidad metodológica fue moderada-baja (ver Tabla 1). En una escala de 0 a 14 puntos, 7 estudios obtuvieron una puntuación total de 5 o menos (n = 7), 14 de los artículos obtuvieron una puntuación de entre 6 y 9 (n = 14) y tan solo uno de ellos obtuvo una puntuación de 10 en la evaluación de la calidad metodológica (n = 1).

Respecto a las características de los 22 estudios incluidos en esta revisión sistemática (ver Tabla 2), cabe mencionar que se desarrollaron en 11 países diferentes: España (n = 3), Estados Unidos (n = 7), Alemania (n = 3), Colombia (n = 2), Dinamarca (n = 1), Irán (n = 2), Tanzania (n = 1), Corea (n = 1), Grecia (n = 1), Ecuador (n = 1) y Filipinas (n = 1). El rango del tamaño de las muestras osciló entre 116 y 28774 participantes. El tipo de muestra más reportada fue la no probabilística (n = 16). El porcentaje de mujeres que componían la muestra se reportó en la mayoría de los estudios incluidos (n = 21), excepto en el de Tibubos y Kröger (2020); de estos, en algunos la muestra estaba compuesta exclusivamente por mujeres (n = 3). El porcentaje de mujeres que componían la muestra en el resto de los estudios (n = 19) osciló entre el 23% y el 75%. La edad media de las muestras se reportó en varios de los estudios (n = 16), con un rango de edad entre los 20 (DE = 19.7) y los 60 años (DE = 59.5)

En la mayoría de los estudios las muestras fueron no clínicas (n = 12) frente a las muestras clínicas (n = 10). El rango del tamaño de las muestras de los artículos que utilizaron muestras clínicas (n = 10) osciló entre 116 y 2852 participantes y el porcentaje de mujeres varió entre el 23% y el 75%. La edad media se reportó en más de la mitad de los estudios (n = 6), con un rango de 32 a 60 años.

En relación con los estudios con muestras no clínicas (n = 12), el rango del tamaño de la muestra varió entre 584 y 28774 participantes. El porcentaje de mujeres participantes se reportó en la mayoría de los artículos (n = 9) y osciló entre el 43% y el 75%, excepto en 3 de ellos (Barrera et al., 2021; Materu et al., 2020; Rodríguez-Muñoz et al.,2022) en los que la muestra pertenecía en su totalidad a mujeres. La edad media de los participantes se reportó en todos los estudios (n = 10), con un rango de 20 y 53 años.

Tabla 1

Evaluación de la calidad de los estudios incluidos utilizando los criterios propuestos por Terwee et al. (2007).

Artículos	Validez de contenido	Estructura Factorial	Consistencia interna	Fiabilidad test re-test	Validez de constructo	Efectos de suelo y techo	Interpretabilidad	Puntuación final (0-14)
Población clínica								
Kroenke et al. (2009)	2	1	2	0	2	0	0	7
Kerper et al. (2014)	2	1	1	0	2	0	0	6
Cano-Vindel et al. (2018)	2	1	2	0	0	0	0	5
Ahmadi et al. (2019)	2	1	2	0	0	0	0	5
Kroenke et al. (2019)	2	0	0	0	2	0	0	5
Ghaheri et al. (2020)	2	1	2	0	2	0	2	9
Kim et al. (2021)	2	2	1	2	1	0	2	10
Basaraba et al. (2022)	0	0	0	0	0	0	2	2
Población no clínica								
Löwe et al. (2010)	2	1	2	0	1	0	2	8
Kocalevent et al. (2014)	2	1	1	0	2	0	2	8
Mills et al. (2015)	2	1	2	0	1	0	0	6
Khubchandani et al. (2016)	2	0	2	2	0	0	0	6
Materu et al. (2020)	2	1	1	0	0	2	0	7
Tibubos y Kröger (2020).	2	1	0	0	0	0	0	3
Barrera et al. (2021)	2	1	0	0	0	0	2	5
Lenz y Li (2021)	2	1	2	0	0	0	2	7
Christodoulaki et al. (2022).	2	1	1	2	0	0	0	6
Guerra et al. (2022)	2	2	2	0	2	0	0	8
Mendoza et al. (2022)	2	1	2	0	2	0	0	6

La calidad de los estudios incluidos se evaluó utilizando los criterios propuestos por Terwee et al. (2007) para medidas sanitarias. Cada uno de los criterios recibe una puntuación de 2 si se cumplen los criterios, 1 si se cumplen parcialmente y 0 si no se cumple ningún criterio. La puntuación total puede oscilar entre 0 y 14.

Nota: Elaboración propia.

Tabla 1.1

Evaluación de la calidad de los estudios incluidos utilizando los criterios propuestos por Terwee et al. (2007) (continuación).

Artículos	Validez de contenido	Estructura Factorial	Consistencia interna	Fiabilidad test re-test	Validez de constructo	Efectos de suelo y techo	Interpretabilidad	Puntuación final (0-14)
Población no clínica								
Rodríguez-Muñoz et al. (2022).	2	2	2	0	1	0	2	9
Sanabria-Mazo et al. (2023)	2	1	2	0	0	0	2	7
Wicke et al. (2022)	2	0	2	0	0	0	0	4

La calidad de los estudios incluidos se evaluó utilizando los criterios propuestos por Terwee et al. (2007) para medidas sanitarias. Cada uno de los criterios recibe una puntuación de 2 si se cumplen los criterios, 1 si se cumplen parcialmente y 0 si no se cumple ningún criterio. La puntuación total puede oscilar entre 0 y 14.
Nota: Elaboración propia.

Tabla 2

Características de los estudios incluidos en esta revisión (n = 22).

Autor (año); país	Población objetivo	Tamaño de la muestra	Tipo de muestra	Edad, años M (DE)	Género, %mujeres
Población clínica					
Kroenke et al. (2009); Estados Unidos	Pacientes de atención primaria de 15 clínicas	2149	No probabilística	47.2 (15.4)	66%

Nota: NR: No reportado. Elaboración propia.

Tabla 2.1

Características de los estudios incluidos en esta revisión (n = 22) (continuación).

Autor (año); país	Población objetivo	Tamaño de la muestra	Tipo de muestra	Edad, años M (DE)	Género, %mujeres
Población clínica					
Kerper et al. (2014); Dinamarca.	Pacientes quirúrgicos en fase preoperatoria de clínicas de valoración anestesiológica	2852	No probabilística	47 (NR)	52%
Cano-Vindel et al. (2018); España	Pacientes con sospecha de ansiedad, depresión o síntomas somáticos de atención primaria	1052	No probabilística	NR (NR)	77%
Ahmadi et al. (2019); Irán	Pacientes con enfermedad coronaria	279	No probabilística	59.5 (9.9)	45%
Kroenke et al. (2019); Estados Unidos	Pacientes de atención primaria inscritos en un ensayo de teleasistencia de trastornos musculoesqueléticos crónicos, dolor, depresión y/o ansiedad comórbida	294	Probabilística	57.4 (NR)	23%
Ghaheri et al. (2020); Irán	Pacientes infértiles de un centro de fertilidad	539	No probabilística	32.9 (5.3)	54%
Kim et al. (2021); Corea	Pacientes adultos ambulatorios	116	No probabilística	NR (NR)	56%
Basaraba et al. (2022); Estados Unidos	Personas con y sin VIH en dos clínicas de atención primaria y un hospital general	911	Probabilística	32.0 (11.3)	63%

Nota: NR: No reportado. Elaboración propia.

Tabla 2.2

Características de los estudios incluidos en esta revisión (n = 22) (continuación).

Autor (año); país	Población objetivo	Tamaño de la muestra	Tipo de muestra	Edad, años M (DE)	Género, %mujeres
Población no clínica					
Löwe et al. (2010); Alemania	Población general de Alemania	5030	No probabilística	48.4 (18)	54%
Kocalevent et al. (2014); Colombia	Población general de Colombia	1500	No probabilística	41.8 (16.2)	52%
Mills et al. (2015); Estados Unidos	Población hispanoamericana de habla inglesa y española	436 (Grupo de habla inglesa: 210; Grupo de habla española:226)	No probabilística	Grupo de habla inglesa: 38.5 (13.7) Grupo de habla española: 46.2 (13.4)	Grupo de habla inglesa: 49% Grupo de habla española: 50%
Khubchandani et al. (2016); Estados Unidos	Estudiantes universitarios	934	Probabilística	20.3 (2.7)	63%
Materu et al. (2020); Tanzania	Mujeres adolescentes fuera de la escuela y mujeres jóvenes	2426	No probabilística	19.7 (2.5)	100%
Tibubos y Kröger (2020); Alemania	Población general, emigrantes y refugiados	28774	No probabilística	46.9 (17.8)	NA

Nota: NR: No reportado. Elaboración propia.

Tabla 2.3

Características de los estudios incluidos en esta revisión (n = 22) (continuación).

Autor (año); país	Población objetivo	Tamaño de la muestra	Tipo de muestra	Edad, años M (DE)	Género, %mujeres
Población no clínica					
Barrera et al. (2021); Estados Unidos	Mujeres embarazadas	1148	No probabilística	27.7 (6.5)	100%
Lenz & Li (2021); Estados Unidos	Residentes de un condado rural	711	No probabilística	NR (NR)	63%
Christodoulaki et al. (2022); Grecia	Población general	584	Probabilística	52.9 (18)	43%
Guerra et al. (2022); Ecuador	Estudiantes universitarios	1732	No probabilística	20.44 (2.3)	55%
Mendoza et al. (2022); Filipinas	Adultos de Filipinas en los primeros meses del brote de la pandemia de COVID-19	4524	No probabilística	27.6 (7.6)	75%
Rodríguez-Muñoz et al. (2022); España	Mujeres embarazadas	845	No probabilística	Nr (NR)	100%
Sanabria-Mazo et al. (2023); España	Población general de Colombia en los primeros meses del brote de la pandemia de COVID-19	18061	No probabilística	NR (NR)	75%
Wicke et al. (2022); Alemania	Población general de Alemania	5022	Probabilística	NR (NR)	53%

Nota: NR: No reportado. Elaboración propia.

3.3. Validez de contenido

El constructo de interés se incluyó en la mayor parte de los estudios que forman parte de esta revisión (n = 21), excepto en uno de ellos (Basaraba et al., 2022). En el resto, todos proporcionaron una descripción concisa constructo que se pretende medir (n = 21), describieron la población objetivo del estudio (n = 21), definieron los constructos que pretendían medir con el instrumento (n = 21), detallaron el procedimiento de selección de los ítems (n = 21) e indicaron que participaron expertos en la selección de los elementos (n = 21).

3.4. Estructura factorial

Un total de 18 estudios reportaron información acerca de la estructura factorial (n = 18) (ver Anexo 2). De estos, 2 realizaron un análisis de componentes principales, 12 un análisis factorial confirmatorio, 3 un análisis factorial exploratorio y confirmatorio y 1 un análisis de componentes principales y un análisis factorial confirmatorio. Todos los artículos mostraron consistencia en un mejor ajuste al modelo bifactorial (n = 16), excepto dos que mostraron índices de ajuste favorables al modelo unifactorial (n = 2; Materu et al., 2020; Lenz y Li, 2021).

Las cargas factoriales estándar se reportaron en 10 estudios y oscilaron entre 0.66 y 0.92. Las cargas factoriales de los ítems se indicaron en todos los artículos que estudiaron la validez factorial (n = 18) y oscilaron entre 0.62 y 1.35. Los índices de ajuste que se estimaron fueron CFI en 16 estudios, TLI en 8, RMSEA en 15 y NFI en 4. Los resultados del CFI oscilaron entre 0.97 y 1.00, los del TLI entre 0.90 y 1.00, los del RMSEA entre 0.00 y 0.14 y los de NFI entre 0.90 y 0.99.

De los estudios que utilizaron muestras clínicas (n = 10), 6 exploraron la dimensionalidad del PHQ-4. De estos, 2 realizaron un análisis de componentes principales, 3 un análisis factorial confirmatorio y 1 un análisis factorial exploratorio y un análisis factorial de componentes principales. Todos mostraron consistencia en el ajuste al modelo bifactorial (n = 6). Las cargas factoriales estándar de los ítems se indicaron en 4 estudios y oscilaron entre 0.66 y 0.76. Las cargas factoriales de los ítems se reportaron en todos (n = 6) y variaron entre 0.67 y 0.96. Los índices de ajuste al modelo bifactorial se estimaron en 4 estudios, encontrándose el CFI y el RMSEA en todos, el TLI en 3 y el NFI en 1. Las puntuaciones del CFI oscilaron entre 0.98 y 1.00, el TLI entre 0.90 y 0.99, el RMSEA entre 0.001 y 0.14 y el NFI obtuvo una puntuación de 0.99.

En relación con los artículos que utilizaron muestras no clínicas (n = 12), todos reportaron información acerca de la dimensionalidad del PHQ-4. De estos, 9 realizaron un análisis factorial confirmatorio, 2 un análisis factorial exploratorio y 1 un análisis de componentes principales y un análisis factorial confirmatorio. De los estudios que exploraron la estructura factorial unidimensional, así como la estructura factorial bidimensional, todos mostraron consistencia en cuanto a la dimensionalidad del PHQ-4 y reportaron índices de ajuste favorables al modelo de

dos factores (n = 10), excepto 2 que solo aportaron evidencia para el modelo unidimensional. La carga factorial de los ítems se indicó en 6 estudios y osciló entre 0.78 y 0.92. Los índices de ajuste al modelo bifactorial se reportaron en 11 estudios, encontrándose el CFI en todos, el TLI en 5, el NFI en 4 y el RMSEA en 10. El rango de puntuaciones entre las que osciló el CFI fue de entre 0.973 y 1.000, el TLI entre 0.923 y 1.00, RMSEA entre 0.000 y 0.145; y NFI entre 0.90 y 0.996.

Por otro lado, 11 estudios aportaron evidencia acerca de la invarianza del PHQ-4 (ver Anexo 3); de estos, solo 1 correspondía a muestra de población no clínica. Concretamente, 7 artículos arrojaron datos acerca de la invarianza según el género, 7 según los grupos de edad, 2 según la localización geográfica, 1 según el nivel económico, 1 según el nivel académico y 3 según el idioma. Todos mostraron consistencia en la invarianza del PHQ-4 en función de estas variables (n = 11).

3.5. Consistencia interna

La consistencia interna se evaluó en 17 de los estudios incluidos (ver Tabla 3). Se evaluó tanto para el PHQ-4 global (n = 17) como para sus subescalas (n = 14). En 16 de los estudios se reportó mediante el coeficiente de consistencia interna α de Cronbach, en 3 mediante el coeficiente McDonald's ω, en 1 mediante el coeficiente Kappa y en 1 mediante el coeficiente Guttman's λ2. Un total de 3 de los estudios incluidos en esta revisión reportó más de un coeficiente de consistencia interna para el PHQ-4 global o para sus subescalas.

El α de Cronbach en los estudios que exploraron este coeficiente osciló entre 0.74 y 0.879, incluyendo el PHQ-4 global y sus subescalas. Solo dos de los estudios reportaron este coeficiente más bajo en la subescala del PHQ-2, arrojando un α de Cronbach de 0.66 y 0.65 (Kerper et al., 2014; Materu et al., 2020). El McDonald 's ω en los tres estudios que exploraron este coeficiente varió entre 0.78 y 0.88.

En relación con los estudios que utilizaron muestras clínicas, 6 exploraron la consistencia interna del PHQ-4; de estos, 5 reportaron la consistencia interna del PHQ-4 global y de las subescalas. En todos, se utilizó el coeficiente α de Cronbach, excepto en 1 que se indicó el coeficiente de correlación de Spearman (Cano-Vindel et al., 2018). El α de Cronbach varió entre 0.66 y 0.85.

Con respecto a los estudios que utilizaron muestras no clínicas, 11 exploraron este coeficiente; de estos, 9 aportaron evidencia sobre la consistencia interna de las subescalas del PHQ-4. La mayoría utilizó el coeficiente α de Cronbach (n = 10), reportando resultados entre 0.65 y 0.88. El coeficiente McDonald's ω se estimó en 4 estudios, con resultados entre 0.78 y 0.88.

Tabla 3

Consistencia interna del PHQ-4 y sus subescalas.

Autor (año)	PHQ-4	PHQ-2	GAD-2
Población clínica			
Kroenke et al. (2009)	α = 0.85	α = 0.81	α = 0.82
Kerper et al. (2014)	α = 0.83	α = 0.66**	α = 0.78
Cano-Vindel et al. (2018)	*α = 0.72	*α = 0.86	*α = 0.76
Ahmadi et al. (2019)	α = 0.78	α = 0.79	α = 0.75
Ghaheri et al. (2020)	α = 0.81	α = 0.77	α = 0.78
Kim et al. (2021)	α = 0.79	NR	NR
Población no clínica			
Löwe et al. (2010)	α = 0.78	α = 0.75	α = 0.82

*Cano-Vindel et al. (2018): Cálculo del coeficiente alfa de Cronbach a partir del método de correlación de Spearman. NR: No reportado.
Nota: Elaboración propia.

Tabla 3.1

Consistencia interna del PHQ-4 y sus subescalas (continuación).

Autor (año)	PHQ-4	PHQ-2	GAD-2
Población no clínica			
Kocalevent et al. (2014)	α = 0.84	NR	NR
Mills et al. (2015)	α = 0.86	α = 0.81	α = 0.80
Khubchandani et al. (2016)	α = 0.81	α = 0.76 Kappa = 0.69	α = 0.82 Kappa = 0.77
Materu et al. (2020)	α = 0.81	α = 0.65**	α = 0.74
Christodoulaki et al. (2022)	0.80	NR	NR
Guerra et al. (2022)	α = 0.88 ω = 0.88	α = 0.78 ω = 0.78	α = 0.84 ω = 0.84
Mendoza et al. (2022)	α = 0.82	α = 0.71	α = 0.83

Cano-Vindel et al. (2018): Cálculo del coeficiente alfa de Cronbach a partir del método de correlación de Spearman. NR: No reportado.

Nota: Elaboración propia.

Tabla 3.1

Consistencia interna del PHQ-4 y sus subescalas (continuación).

Autor (año)	PHQ-4	PHQ-2	GAD-2
Población no clínica			
Rodríguez-Muñoz et al. (2022)	α = 0.77	α = 0.70	α = 0.75
Sanabria-Mazo et al. (2023)	α = 0.86 ω = 0.86 λ2 = 0.86	α = 0.79 ω = 0.81 λ2 = 0.80	α = 0.83 ω = 0.83 λ2 = 0.82
Wicke et al. (2022)	ω = 0.85	ω = 0.77	ω = 0.78

*Cano-Vindel et al. (2018): Cálculo del coeficiente alfa de Cronbach a partir del método de correlación de Spearman. NR: No reportado.
Nota: Elaboración propia.

3.6. Fiabilidad test-retest

La estabilidad temporal en la administración del PHQ-4 se evaluó en 3 estudios, calculandose en 2 el coeficiente de correlación intraclase (ICC) y en 1 el coeficiente Kappa. La fiabilidad test-retest se estudió con muestras clínicas en uno de los artículos incluídos en esta revisión, en el cual se reportaron los resultados utilizando el ICC (Kim et al., 2021) y con población no clínica en dos de ellos, indicándose en uno Kappa (Khubchandani et al., 2016) y el ICC en otro (Christodoulaki et al., 2022). En relación con los resultados reportados, los rangos entre los que se encuentra Kappa están entre 0.69 y 0.81 y el ICC entre 0.83 y 0.96 (ver Anexo 4).

3.7. Validez de constructo

Se reportó información acerca de la validez de constructo en 11 de los artículos incluidos en esta revisión (ver Tabla 4); de estos, 9 reportaron información acerca de la validez convergente del test, 6 acerca de la validez divergente y 5 ambas.

La validez convergente se exploró con cuestionarios que medían variables como la salud y el malestar psicológico, como el *Medical Outcomes Study Short-Form General Health Survey* (SF-20) (n = 1), el *Brief Symptom Inventory* (BSI) (n = 1) y el *Patient Reported Outcomes Measurement Information System Depression and Anxiety Scale* (PROMIS-ADS) (n = 1); y se encontró que sus medidas correlacionaron de forma significativa y positiva con el PHQ-4 y con sus subescalas. La correlación entre estos cuestionarios y el PHQ-4 osciló entre 0.36 y 0.80. Los rangos de la correlación con el PHQ-2 fueron de entre 0.33 y 0.72 y con el GAD-2 de entre 0.28 y 0.72. El PHQ-4 correlacionó también de forma positiva con cuestionarios que medían constructos de ansiedad y de depresión, como el *Patient Health Questionnaire Anxiety-Depression* Scale (PHQ-ADS) (n = 1), el *Hospital Anxiety and Depression Scale* (HADS) (n = 2) y el PHQ-9 (n = 1); y se encontró una correlación positiva entre los instrumentos en un rango de entre 0.46 y 0.83. El PHQ-2 correlacionó también de forma positiva con estos constructos, obteniendo un rango de correlación de entre 0.47 y 0.67 con las escalas HADS, *Depression, Anxiety, and Stress Scale* (DASS) (n = 1) y PHQ-9.

La validez convergente también se investigó con otros instrumentos que medían el constructo de ansiedad únicamente o, por el contrario, depresión. Se obtuvieron correlaciones significativamente altas y positivas con instrumentos que medían síntomas ansiosos como *Hamilton Anxiety Scale* (HAM-A) (n = 1), *Beck Anxiety Inventory* (BAI) (n = 1) y *Penn State Worry Questionnaire* (PSWQ) (n = 1), tanto con el PHQ-4 (0.68, 0.72 y 0.56) como con sus subescalas GAD-2 (0.64, 0.71 y 0.54) y PHQ-2 (0.54, 0.56 y 0.451). También se obtuvieron correlaciones altas con instrumentos que medían síntomas depresivos, como entre el *Beck Depression Inventory* (BDI) (n = 1) y el *Hamilton Rating Scale for Depression* (HAM-D) (n = 1). Estos instrumentos también correlacionaron significativamente con el PHQ-4 (0.76 y 0.72), con el PHQ-2 (0.65 y 0.65) y con GAD-2 (0.67 y 0.61).

En relación a la validez divergente, en la correlación entre el PHQ-4 y sus subescalas e instrumentos que miden constructos opuestos a la ansiedad y la depresión, como el bienestar general (*World Health Organisation-Five Well-Being Index*; WHO-5) (n = 1), autoestima (*Rosenberg Self-Esteem Scale;* RSES) (n = 1), calidad de vida (*Questions on Life Satisfaction*; QLS) (n = 2), capital psicológico (*Psychological Capital Questionnaire*; PCQ-12) (n = 1) y resiliencia (*Resilience Scale-11*; RS-11) (n = 1) se obtuvieron correlaciones negativas y significativas con un rango de entre -0.56 y -0.26 en el PHQ-4, de entre -0.26 y 0.51 en PHQ-2 y de entre -0.21 y -0.54.

Tabla 4

Validez de constructo del PHQ-4.

Autor (año)	Validez convergente	Validez discriminante
Población clínica		
Kroenke et al. (2009)	NR	PHQ-4 y SF-20- Salud Mental (r=0.80) PHQ-2 y SF-20- Salud Mental (r=0.72) GAD-2 y SF-20- Salud Mental (r=0.72) PHQ-4 y SF-20- Funcionamiento Social (r=0.52) PHQ-2 y SF-20- Funcionamiento Social (r=0.50) GAD-2 y SF-20- Funcionamiento Social (r=0.44) PHQ-4 y SF-20- Percepción General de Salud (r=0.48) PHQ-2 y SF-20- Percepción General de Salud (r=0.47) GAD-2 y SF-20- Percepción General de Salud (r=0.40) PHQ-4 y SF-20- Funcionamiento (r=0.37) PHQ-2 y SF-20- Funcionamiento (r=0.37) GAD-2 y SF-20- Funcionamiento (r=0.29) PHQ-4 y SF-20- Dolor corporal (r=0.36) PHQ-2 y SF-20- Dolor corporal (r=0-33) GAD-2 y SF-20- Dolor corporal (r=0.32) PHQ-4 y SF-20- Funcionamiento Físico (r=0.36) PHQ-2 y SF-20- Funcionamiento Físico (r=0.36) GAD-2 y SF-20- Funcionamiento Físico (r=0.28)
Kerper et al. (2014)	PHQ-4 y BSI (r = 0.69) PHQ-2 y BSI (r = 0.60) GAD-2 y BSI (r = 0.67)	NR

*Aunque en el estudio de Mills et al. (2015) se exploraron correlaciones, no se exploraron entre las medidas de otros instrumentos. NR: no reportado.

AAQ-II: Cuestionario de aceptación y compromiso; BAI: Beck Anxiety Inventory; BDI:Beck Depression Inventory; BSI: Brief Symptom Inventory; DASS-S: Depression, Anxiety, and Stress Scale; GHQ-12: 12-item General Health Questionnaire; GSES: General Self-Efficacy Scale; HADS: Hospital Anxiety and Depression Scale; HAM-A: Hamilton Anxiety Scale; HAM-D_17: Hamilton Rating Scale for Depression; PANAS: Positive and Negative Affect Schedule; PCQ-12:Psychological Capital Questionnaire-12; PHQ-ADS: Patient Health Questionnaire Anxiety-Depression Scale; PROMIS-ADS: Patient Reported Outcomes Measurement Information System Depression and Anxiety Scale; PSS-10: Escala de estrés percibido; PSWQ: Penn State Worry Questionnaire; QLS: Questions on Life Satisfaction; RS-11: Resilience Scale; RSES: Rosenberg Self-Esteem Scale; SF-20: Medical Outcomes Study Short-Form General Health Survey; WHO-5: World Health Organisation-Five Well-Being Index.

Nota: Elaboración propia.

Tabla 4.1

Validez de constructo del PHQ-4 (continuación).

Autor (año)	Validez convergente	Validez discriminante
Población clínica		
Kroenke et al. (2019)	PHQ-4 y PHQ-ADS (r=0.83) PHQ-4 y PROMIS-ADS (r= 0.71) PHQ-4 y PROMIS-ADS-8 (r=0.67)	PHQ-4 y SF-12 MCS (r=-0.57) PHQ-4 y SF-36 Mental (r=-0.66)
Ghaheri et al. (2020)	PHQ-4 y HADS-A (r = 0.717) PHQ-2 y HADS-A (r = 0.573) GAD-2 y HADS-A (r = 0.700) PHQ-4 y HADS-D (r = 0.535) PHQ-2 y HADS-D (r = 0.491) GAD-2 y HADS-D (r = 0.458) PHQ-4 y PSWQ (r = 0.560) PHQ-2 y PSWQ (r = 0.451) GAD-2 y PSWQ (r = 0.545)	PHQ-4 y WHO-5 (r =-0.559) PHQ-2 y WHO-5 (r = -0.518) GAD-2 y WHO-5 (r = -0.545)
Kim et al. (2021)	PHQ-4 y BDI (r = 0.762) PHQ-2 y BDI (r = 0.656) GAD-2 y BDI (r = 0.678) PHQ-4 y HAM-D_17 (r = 0.720) PHQ-2 y HAM-D_17(r = 0.654) GAD-2 y HAM-D_17(r = 0.615) PHQ-4 y BAI (r = 0.728) PHQ-2 y BAI (r = 0.567) GAD-2 y BAI (r = 0.716) PHQ-4 y HAM-A (r = 0.680) PHQ-2 y HAM-A (r = 0.548) GAD-2 y HAM-A (r = 0.647)	NR

*Aunque en el estudio de Mills et al. (2015) se exploraron correlaciones, no se exploraron entre las medidas de otros instrumentos. NR: no reportado.

AAQ-II: Cuestionario de aceptación y compromiso; BAI: Beck Anxiety Inventory; BDI:Beck Depression Inventory; BSI: Brief Symptom Inventory; DASS-S: Depression, Anxiety, and Stress Scale; GHQ-12: 12-item General Health Questionnaire; GSES: General Self-Efficacy Scale; HADS: Hospital Anxiety and Depression Scale; HAM-A: Hamilton Anxiety Scale; HAM-D_17: Hamilton Rating Scale for Depression; PANAS: Positive and Negative Affect Schedule; PCQ-12:Psychological Capital Questionnaire-12; PHQ-ADS: Patient Health Questionnaire Anxiety-Depression Scale; PROMIS-ADS: Patient Reported Outcomes Measurement Information System Depression and Anxiety Scale; PSS-10: Escala de estrés percibido; PSWQ: Penn State Worry Questionnaire; QLS: Questions on Life Satisfaction; RS-11: Resilience Scale; RSES: Rosenberg Self-Esteem Scale; SF-20: Medical Outcomes Study Short-Form General Health Survey; WHO-5: World Health Organisation-Five Well-Being Index.

Nota: Elaboración propia.

Tabla 4.2

Validez de constructo del PHQ-4 (continuación).

Autor (año)	Validez convergente	Validez discriminante
Población no clínica		
Löwe et al. (2010)	NR	PHQ-4 y RSES (r=−0.49) PHQ-2 y RSES (r=−0.48,) GAD-2 y RSES (r=−0.40) PHQ-4 y QLS (r=−0.39) PHQ-2 y QLS (r=−0.37) GAD-2 y QLS (r=−0.32) PHQ-4 y RS-11 (r=−0.35) PHQ-2 y RS-11 (r=−0.34) GAD-2 y RS-11 (=−0.28)
Kocalevent et al. (2014)	PHQ-4 y HADS (r = 0.46) PHQ-2 y HADS (r = 0.47) GAD-2 y HADS (r = 0.37) PHQ-4 y GHQ-12 (r = 0.44) PHQ-2 y GHQ-12 (r = 0.42) GAD-2 y GHQ-12 (r = 0.37)	PHQ-4 y GSES (r = -0.26) PHQ-2 y GSES (r = -0.26) GAD-2 y GSES (r = -0.21) PHQ-4 y QLS (r = -0.29) PHQ-2 y QLS (r = -0.27) GAD-2 y QLS (r = -0.25)
Mills et al. (2015)	NR	NR

*Aunque en el estudio de Mills et al. (2015) se exploraron correlaciones, no se exploraron entre las medidas de otros instrumentos. NR: no reportado.
AAQ-II: Cuestionario de aceptación y compromiso; BAI: Beck Anxiety Inventory; BDI:Beck Depression Inventory; BSI: Brief Symptom Inventory; DASS-S: Depression, Anxiety, and Stress Scale; GHQ-12: 12-item General Health Questionnaire; GSES: General Self-Efficacy Scale; HADS: Hospital Anxiety and Depression Scale; HAM-A: Hamilton Anxiety Scale; HAM-D_17: Hamilton Rating Scale for Depression; PANAS: Positive and Negative Affect Schedule; PCQ-12:Psychological Capital Questionnaire-12; PHQ-ADS: Patient Health Questionnaire Anxiety-Depression Scale; PROMIS-ADS: Patient Reported Outcomes Measurement Information System Depression and Anxiety Scale; PSS-10: Escala de estrés percibido; PSWQ: Penn State Worry Questionnaire; QLS: Questions on Life Satisfaction; RS-11: Resilience Scale; RSES: Rosenberg Self-Esteem Scale; SF-20: Medical Outcomes Study Short-Form General Health Survey; WHO-5: World Health Organisation-Five Well-Being Index.
Nota: Elaboración propia

Tabla 4.3

Validez de constructo del PHQ-4 (continuación).

Autor (año)	Validez convergente	Validez discriminante
Población no clínica		
Guerra et al. (2022)	PHQ-4 y AAQ-II (r=0.673) PHQ-4 y PSS-10 (r=0.673)	PHQ-4 y PCQ-12 (r = -0,296)
Mendoza et al. (2022)	PHQ-2 y DASS-S (r = 0.60) GAD-2 y DASS-S (r=0.64) PHQ-2 y PANAS p (r = 0.58) GAD-2 y PANAS n (r=0.64)	PHQ-2 y PANAS p (r = -0.48) GAD-2 y PANAS n (r=-0.29)
Rodríguez-Muñoz et al. (2022)	PHQ-4 y PHQ-9 (r = 0.76) PHQ-2 y PHQ-9 (r=0.67) GAD-2 y PHQ-9 (r=0.63)	NR

*Aunque en el estudio de Mills et al. (2015) se exploraron correlaciones, no se exploraron entre las medidas de otros instrumentos. NR: no reportado.
AAQ-II: Cuestionario de aceptación y compromiso; BAI: Beck Anxiety Inventory; BDI:Beck Depression Inventory; BSI: Brief Symptom Inventory; DASS-S: Depression, Anxiety, and Stress Scale; GHQ-12: 12-item General Health Questionnaire; GSES: General Self-Efficacy Scale; HADS: Hospital Anxiety and Depression Scale; HAM-A: Hamilton Anxiety Scale; HAM-D_17: Hamilton Rating Scale for Depression; PANAS: Positive and Negative Affect Schedule; PCQ-12:Psychological Capital Questionnaire-12; PHQ-ADS: Patient Health Questionnaire Anxiety-Depression Scale; PROMIS-ADS: Patient Reported Outcomes Measurement Information System Depression and Anxiety Scale; PSS-10: Escala de estrés percibido; PSWQ: Penn State Worry Questionnaire; QLS: Questions on Life Satisfaction; RS-11: Resilience Scale; RSES: Rosenberg Self-Esteem Scale; SF-20: Medical Outcomes Study Short-Form General Health Survey; WHO-5: World Health Organisation-Five Well-Being Index.
Nota: Elaboración propia

3.8. Efectos de suelo y techo

Los efectos de suelo y techo se analizaron en un solo estudio (Materu et al., 2020). No se observaron efectos de techo y todos los ítems presentaron efectos de suelo.

3.9. Interpretabilidad

La interpretabilidad se analizó desde un enfoque de validez de grupos conocidos en 8 estudios (ver Anexo 5). Las principales variables sociodemográficas que se estudiaron fueron el género (n=6), el nivel educativo (n = 4), la edad (n = 6), el estado civil (n = 5), la situación laboral (n = 5) y el nivel económico (n = 2). De los estudios que reportaron diferencias en función de la variable género, 4 encontraron diferencias significativas en las puntuaciones en PHQ-4, PHQ-2 y GAD-2, obteniendo las mujeres puntuaciones más altas. En relación con el nivel educativo, 3 estudios identificaron diferencias significativas, concluyendo que las personas con menor nivel educativo obtuvieron puntuaciones más altas en PHQ-4 en uno de los estudios y en PHQ-4, PHQ-2 y GAD-2 en dos. Con respecto a la edad, 4 estudios reportaron puntuaciones más altas en PHQ-4 en personas jóvenes frente a personas de mayor edad, dos de estos reportando además esos datos en PHQ-2 y GAD-2. Por otro lado, uno de los artículos concluyó puntuaciones más altas en PHQ-4, GAD-2 y PHQ-2 en personas de mayor edad. De los estudios que encontraron diferencias significativas en relación con el estado civil y las puntuaciones en PHQ-4, PHQ-2 y GAD-2, todos fueron consistentes en que las personas separadas o solteras obtienen mayores puntuaciones en estas escalas. Se encontraron diferencias significativas en relación con la actividad laboral en 4 estudios, de los cuales 3 obtuvieron unas puntuaciones significativamente más altas en personas desempleadas tanto en PHQ-4, GAD-2 y PHQ-2 y uno puntuaciones significativamente más altas en personas con incapacidad laboral en el PHQ-4. Por último, en relación con el nivel económico, ambos estudios reportaron puntuaciones significativamente más altas en personas con un nivel económico bajo en PHQ-4 en un estudio y en PHQ- PHQ-2 y GAD-2 en el otro.

4. Discusión

La depresión y la ansiedad son dos condiciones de salud altamente prevalentes que requieren la atención de investigadores, clínicos y políticos. Disponer de herramientas de cribado ultra breves que detecten los síntomas de estas condiciones de salud podría contribuir a disminuir los errores en el diagnóstico, agilizar los tiempos de espera de la sanidad pública y reducir los costes económicos. Por lo anterior, el presente trabajo tuvo como objetivo llevar a cabo, por primera vez, una revisión sistemática de las propiedades psicométricas del PHQ-4 evaluando la utilidad del instrumento en la práctica clínica. Tras completar la búsqueda en bases de datos y aplicar los criterios de selección de esta revisión, se identificaron 22 artículos (publicados entre 2009 y 2023) con 80129 participantes, con muestras tanto clínicas como no clínicas. Todos estos estudios tuvieron como propósito evaluar las propiedades psicométricas del PHQ-4 como un cuestionario de cribado breve de síntomas de ansiedad y depresión. Las conclusiones extraídas de la evidencia reportada en estos estudios se describen a continuación.

El PHQ-4 se trata de un instrumento validado actualmente en muestras de una gran variedad de países (Estados Unidos, Dinamarca, España, Irán, Corea, Alemania, Colombia, Tanzania, Grecia, Ecuador y Filipinas). La población objetivo de investigación en estos estudios ha variado entre muestras clínicas y muestras no clínicas. Entre las muestras no clínicas se incluyeron la población general de distintos países (Alemania, Colombia y Grecia), así como subgrupos poblacionales como los estudiantes universitarios, los emigrantes, los refugiados, las mujeres embarazadas, las personas residentes en condados rurales o las mujeres jóvenes. Las muestras clínicas incorporaron pacientes de atención primaria, pacientes quirúrgicos, pacientes con sospecha de ansiedad o depresión y/o pacientes con alguna patología.

La validez de contenido del PHQ-4 fue evaluada casi en la totalidad de los estudios, encontrándose tan solo uno donde no se reportaba esta información. En general, los artículos han reportado una descripción clara de los constructos de ansiedad y depresión como objetivos de medición y han descrito la población objetivo de estudio. Además, han detallado el procedimiento mediante el que se han seleccionado los ítems del cuestionario, así como el tipo participación de expertos o académicos durante la validación. Estos hallazgos suscitan a concluir que el cuestionario es una herramienta válida en la práctica clínica.

Sobre la estructura factorial del PHQ-4, reportada en 18 estudios, se encontró que la mayor parte de la evidencia apoya el modelo de dos factores, siendo consistentes con la medición de los constructos de depresión y ansiedad propuestos originalmente por Kroenke et al. (2009). Únicamente dos estudios con muestras no clínicas reportaron evidencia para el modelo unifactorial. Sin embargo, en estas investigaciones solo se exploró el modelo unifactorial y, por tanto, se desconoce si los índices de ajustes son más favorables en modelo o en otro (Lenz y Li, 2021; Materu et al., 2020). Estas conclusiones y la evidencia aportada en torno al modelo de dos factores (ansiedad y depresión), en las que se observa consistencia en las muestras clínicas, sugieren que el PHQ-4 es un instrumento válido para la detección de síntomas de

ansiedad y depresión en la práctica clínica. Concretamente, los profesionales del ámbito sanitario pueden obtener una evaluación específica de los síntomas de depresión (PHQ-2) y de ansiedad (GAD-2), así como una global de distrés o angustia psicológica (PHQ-4). Además, al tratarse de una herramienta de cribado ultrabreve, el tiempo de respuesta es muy bajo, permitiendo la detección temprana de estos síntomas, el seguimiento del proceso y la reducción de costos sanitarios. La estructura factorial de este instrumente permite la personalización de los tratamientos en función de las puntuaciones que obtiene el paciente en cada uno de los factores del instrumento.

La consistencia interna del instrumento total (PHQ-4) y de sus subescalas (PHQ-2 y GAD-2) se reportó en 14 de los estudios incluidos y mostró ser alta en todos los estudios revisados según los criterios de Terwee et al. (2007) y de Nunnally (1967), exceptuando la puntuación de alfa de Cronbach obtenida para el PHQ-2 en uno de los artículos (Materu et al., 2020). El comportamiento de esta propiedad psicométrica fue similar tanto en población de carácter clínico como en población de carácter no clínico, sugiriendo la validez y fiabilidad del instrumento para cualquier tipo de muestras, así como su pertinencia de uso en la práctica clínica.

En cuanto a la estabilidad temporal del cuestionario, solamente 3 estudios reportaron información acerca de esta propiedad realizando un procedimiento test re-test, en los que los índices Kappa e ICC fueron aceptables según los criterios de Terwee et al., (2007). Sin embargo, hay escasez de estudios longitudinales que aborden las características de este instrumento, por lo que esta propiedad no se ha evaluado de forma exhaustiva. Se requiere, por tanto, más evidencia para obtener conclusiones sólidas sobre esta propiedad psicométrica.

La validez de constructo del PHQ-4, reportada en 11 de los 22 artículos incluidos en esta revisión, está respaldada por la correlación significativa y positiva existente con otras medidas de los constructos de ansiedad y depresión (BSI, PROMIS-ADS, PHQ-ADS, HADS, DASS, PHQ-9, HAM-A, HAM-D, BAI, PSWQ y BDI), así como por la correlación significativa y negativa con medidas de constructos teóricamente divergentes (WHO-5, SF-20, RSES, QLS, PCQ-12 y RS-11). En conclusión, esta herramienta tiene la capacidad para medir de forma precisa los constructos de ansiedad y depresión, aumentando la fiabilidad y la validez de su administración en el ámbito sanitario.

Existe poca evidencia en relación con los efectos de suelo y techo de la herramienta. Al tratarse de un cuestionario de cribado ultrabreve, esta información podría ser de gran relevancia para evitar sesgos de interpretación. La evidencia disponible actualmente invita a cuestionarse si esta herramienta de cribado es lo suficientemente sensible como para detectar cambios significativos en los extremos inferiores o superiores en población clínica o no clínica. Se necesita un mayor número de evidencias en esta propiedad psicométrica para disponer de conclusiones sólidas.

La interpretabilidad del PHQ-4 y de sus subescalas se evaluó mediante 8 estudios. La variabilidad de las muestras aporta evidencia de que en general las mujeres, así como las

personas con un nivel económico y/o académico bajo, presentan más síntomas de depresión y ansiedad. Reconocer que existen grupos más afectados según determinadas variables sociodemográficas permite conocer si existen grupos de riesgo, la adaptación de las intervenciones y el diseño de estrategias de prevenciones ajustadas a necesidades específicas. Además, estos hallazgos permiten aumentar la calidad de la asistencia y atención de los pacientes en el contexto sanitario.

Aunque existen muchos instrumentos de cribado para medir síntomas de ansiedad (Escala de ansiedad de Hamilton- HAS; Inventario de ansiedad de Beck- BAI; Inventario de ansiedad de Zung- ZAI), síntomas de depresión (Escala de Depresión de Hamilton- HDRS; Inventario de Depresión de Beck- BDI; Inventario de Depresión de Zung- SDS) y síntomas de ansiedad y depresión (Escalas de Depresión Ansiedad Estrés- DASS-21), el PHQ-4 es un instrumento especialmente útil debido a la brevedad de su administración y a la rapidez en el cribado de los síntomas. Estas herramientas de cribado se utilizan para identificar, primero, de forma rápida y eficiente las personas con riesgo de padecer síntomas relacionados con estos trastornos y, luego, para derivarlas a una evaluación que permita establecer un diagnóstico más preciso, con instrumentos de referencia como la Entrevista Diagnóstica Internacional Compuesta (Entrevista CIDI) y la Entrevista clínica estructurada para los trastornos del eje I del DSM-IV (SCID-I).

4.1. Limitaciones y fortalezas

Este estudio es pionero en revisar sistemáticamente las propiedades psicométricas del PHQ-4 desde su publicación (2009). Se trata de una herramienta de cribado que se ha aplicado en una gran variedad de poblaciones, recalcando así la importancia de conocer su validez y fiabilidad para su administración en el ámbito sanitario. A lo largo de esta revisión sistemática, se ha proporcionado una síntesis narrativa de la evidencia hasta la actualidad (2023), con un registro prospectivo del propósito del estudio en PROSPERO para una mayor transparencia científica sobre el método planteado inicialmente. La búsqueda de evidencia basada en la estrategia PICOS, así como el procedimiento de extracción de datos y de evaluación de calidad metodológica se desarrolló de manera sistemática y exhaustiva por un revisor principal y un segundo revisor, minimizando de esta forma posibles sesgos de selección e interpretación. Para aumentar la rigurosidad metodológica de este estudio, en este trabajo se siguieron las recomendaciones de PRISMA para revisiones sistemáticas. Los criterios de evaluación de calidad metodológica se apoyaron, además, en una propuesta ampliamente aceptada por la comunidad científica para la valoración de las propiedades psicométricas de instrumentos en salud (Terwee et al., 2007).

Sin embargo, la extensión de las conclusiones derivadas de este estudio puede estar parcialmente limitada al haberse incluido solo artículos escritos en inglés o español, siendo una posible futura línea de investigación replicar este procedimiento aumentando los criterios de inclusión con respecto al idioma. Por otro lado, los resultados de la evaluación de la calidad

metodológica de los artículos incluidos en esta revisión indican que podrían existir sesgos de evaluación de las propiedades psicométricas del PHQ-4, lo que limita una exploración más precisa de algunas de las características de esta herramienta de cribado como, por ejemplo, la estabilidad temporal y los efectos de suelo y techo.

4.2. Recomendaciones

Aunque el PHQ-4 ha mostrado tener una alta fiabilidad y validez, es importante considerar que se trata de un instrumento de cribado y no de diagnóstico de depresión y ansiedad. Por tanto, sus resultados no sustituyen a un diagnóstico completo y deben ser interpretados por expertos en el constructo en un contexto clínico concreto.

Por el número de evidencias aportadas, este estudio podría haber incluido un metaanálisis, que por límites temporales del TFM no fue posible, siendo una posible futura línea de investigación sintetizar cuantitativamente las evidencias disponibles sobre las propiedades psicométricas del PHQ-4. Concretamente, este metaanálisis podría ser importante porque permitirían realizar una evaluación sistemática y objetiva de los hallazgos de los 22 artículos incluidos. Al combinar los datos de varios estudios mediante esta técnica estadística, los investigadores y los clínicos podrían obtener una medida más precisa del efecto global de interés y determinar si las diferencias observadas en los estudios individuales son significativas. La síntesis cuantitativa también contribuiría a identificar patrones o tendencias en los datos que son más difíciles de detectar en una revisión sistemática tradicional.

4.3. Conclusiones

El PHQ-4 es una herramienta de cribado breve de síntomas de ansiedad y de depresión válida y fiable para su utilización en el ámbito de salud, tanto para población clínica como no clínica. Este cuestionario, sin sustituir una posterior evaluación diagnóstica, puede ser especialmente útil para, además, continuar investigando sobre ambas condiciones de salud.

En este estudio se revisaron cuatro bases de datos (Medline, PsycINFO, Web of Science y SCOPUS) explorando artículos publicados entre el año de creación de la herramienta (2009) y la actualidad (2023). Siguiendo las recomendaciones de PRISMA, el cribado de artículos, la extracción de datos y la evaluación de la calidad metodológica se llevó a cabo por dos revisores independientes.

La mayor parte de los datos obtenidos de esta revisión reportaron valores adecuados en relación con la fiabilidad y validez del PHQ-4, siendo las propiedades psicométricas más estudiadas la validez de contenido, la consistencia interna, la estructura factorial, la validez de constructo y la interpretabilidad, respectivamente. Los hallazgos con respecto a la estructura factorial reportaron la existencia de dos factores, uno de ansiedad y otro de depresión, además de la invarianza de la herramienta para las variables de género, grupos de edad, localización geográfica, nivel de ingresos, nivel educativo e idioma. La consistencia interna reportó valores

adecuados tanto para el PHQ-4 como para sus subescalas (PHQ-2 y GAD-2). En relación con la fiabilidad test re-test, se encontró una buena correlación entre los dos momentos de administración de la prueba, aunque fueron escasamente exploradas en los estudios incluidos en esta revisión. Con respecto a la validez de constructo, el PHQ-4 correlacionó de forma significativa y positiva con medidas teóricamente convergentes y de forma significativa y negativa con medidas teóricamente divergentes. Al aplicar los criterios de calidad metodológica de Terwee et al. (2007), se observó que algunas de las propiedades psicométricas no han sido ampliamente estudiadas, como es el caso de los efectos de suelo y techo y la fiabilidad test re-test.

En resumen, este estudio aporta evidencia acerca de la dimensionalidad y la fiabilidad del PHQ-4, habiéndose reportado en un estudio reciente que, además, su presentación en formato en línea no altera los resultados (Sanabria-Mazo et al., 2023). Los hallazgos de esta revisión sistemática señalan la utilidad clínica de esta herramienta para el cribado de síntomas de depresión y de ansiedad, tanto en muestras clínicas como no clínicas. Debido a la brevedad de su administración, el PHQ-4 podría contribuir a la optimización de los recursos de salud.

5. Referencias

Ahmadi, S. M., Masjedi Arani, A., Bakhtiari, M., y Davazdah Emamy, M. H. (2019). Psychometric properties of Persian version of Patient Health Questionnaires-4 (PHQ-4) in coronary heart disease patients. *Iranian Journal of Psychiatry and Behavioral Sciences, 13*(4), e85820. https://doi.org/10.5812/ijpbs.85820

Alonso, M. Y. G., Llamazares, M. D. C. E., Martín, M. Á. M., Gómez, M. B. M., y Val, E. M. (2019). Comorbilidad de los trastornos alimentarios con ansiedad y depresión en estudiantes universitarios: revisión sistemática. *Revista Argentina de Clínica Psicológica, 28*(4), 375-384. https://dialnet.unirioja.es/servlet/articulo?codigo=7786589

Barrera, A. Z., Moh, Y. S., Nichols, A., y Le, H. N. (2021). The Factor Reliability and Convergent Validity of the Patient Health Questionnaire-4 Among an International Sample of Pregnant Women. *Journal of women's health (2002), 30*(4), 525–532. https://doi.org/10.1089/jwh.2020.8320

Basaraba, C. N., Stockton, M. A., Sweetland, A., Medina-Marino, A., Lovero, K. L., Oquendo, M. A., Greene, M. C., Mocumbi, A. O., Gouveia, L., Mello, M., Dos Santos, P., Suleman, A., Mabunda, D., Mandlate, F., Xavier, A., Fumo, W., Massinga, L., Khan, S., Feliciano, P., Kann, B., Salem, A.F., Bezuidenhout, C., Mootz, J.J., Duarte, C.S., Cournos, F., Wall, M.M., y Wainberg, M. L. (2023). Does It Matter What Screener We Use? A Comparison of Ultra-brief PHQ-4 and E-mwTool-3 Screeners for Anxiety and Depression Among People With and Without HIV. *AIDS and behavior, 27*(4), 1154–1161. https://doi.org/10.1007/s10461-022-03852-w

Batista-Foguet, J. M., Coenders, G., y Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. *Medicina clínica*, *122*(1), 21-27. https://www.elsevier.es/es-revista-medicina-clinica-2-articulo-analisis-factorial-confirmatorio-su-utilidad-validacion-cuestionarios-13057542

Boubeta, A. R., Mallou, J. V., Piñeiro, J. A., y Lévy, J. (2006). El análisis factorial. *Modelización con estructuras de covarianzas en ciencias sociales: temas esenciales, avanzados y aportaciones especiales*, *119*. https://books.google.es/books?hl=es&lr=&id=WEfC1TGVJBgC&oi=fnd&pg=PA119&dq=validez+factorial+afc+afe&ots=frI1VS9GXw&sig=pibsN0YwH53BLThOqQA-cyxyA6M#v=onepage&q=validez%20factorial%20afc%20afe&f=false

Cano-Vindel, A., Muñoz-Navarro, R., Medrano, L. A., Ruiz-Rodríguez, P., González-Blanch, C., Gómez-Castillo, M. D., Capafons-Bonet, A., Chacón, F., Santolaya, F., y PsicAP Research Group. (2018). A computerized version of the Patient Health Questionnaire-4 as an ultra-brief screening tool to detect emotional disorders in primary care. *Journal of Affective Disorders, 234*, 247-255. https://doi.org/10.1016/j.jad.2018.01.030

Carballo, M. M., Estudillo, C. P., Meraz, L. L., Parrazal, L. B., y Valle, C. M. (2019). Trastornos de ansiedad: revisión bibliográfica de la perspectiva actual. *Eneurobiología, 10*(24). https://dialnet.unirioja.es/servlet/articulo?codigo=7735542

Cardila, F., Martos, A., Barragán, A. B., Pérez-Fuentes, M. D. C., Molero, M. D. M., y Gázquez, J. J. (2015). Prevalencia de la depresión en España: Análisis de los últimos 15 años. *European Journal of Investigation in Health, Psychology and Education, 5*(2), 267-279. https://dialnet.unirioja.es/servlet/articulo?codigo=5388786

Cascaes da Silva, F., Gonçalves, E., Valdivia Arancibia, B. A., Bento, G. G., Silva Castro, T. L. D., Soleman Hernandez, S. S., y Silva, R. D. (2015). Estimadores de consistencia interna en las investigaciones en salud: el uso del coeficiente alfa. *Revista Peruana de medicina experimental y salud pública*, *32*(1), 129-138. http://pepsic.bvsalud.org/pdf/psicousf/v7n2/v7n2a03.pdf

Castro-Rodríguez, J. I., Olariu, E., Garnier-Lacueva, C., Martín-López, L. M., Perez-Sola, V., Alonso, J., Forero, C. G., y INSAyD Investigators. (2015). Diagnostic accuracy and adequacy of treatment of depressive and anxiety disorders: A comparison of primary care and specialized care patients. *Journal of Affective Disorders, 172*, 462-471. https://doi.org/10.1016/j.jad.2014.10.020

Cerdá, J. C. M. (2021). Salud mental en tiempos de pandemia para tiempos sin pandemia. Revista Española de Enfermería de Salud Mental, (14). https://doi.org/10.6018/reesm.447311

Christodoulaki, A., Baralou, V., Konstantakopoulos, G., y Touloumi, G. (2022). Validation of the Patient Health Questionnaire-4 (PHQ-4) to screen for depression and anxiety in the Greek general population. Journal of psychosomatic research, 160, 110970. https://doi.org/10.1016/j.jpsychores.2022.110970.

Fekadu, A., Demissie, M., Birhane, R., Medhin, G., y Thornicroft, G. (2022). Under detection of depression in primary care settings in low and middle-income countries: A systematic review and meta-analysis. Systematic Reviews, 11(1), 21. https://doi.org/10.1186/s13643-022-01893-9

Forcadell López, E., Lázaro García, L., Fullana Rivas, M. A., y Lera Miguel, S. (2019, January 29). *Trastornos de ansiedad: definición y síntomas.* Portal CLÍNIC. https://www.clinicbarcelona.org/asistencia/enfermedades/trastornos-de-ansiedad/definicion

García de Yébenes Prous, M. A., Rodríguez Salvanés, F., y Carmona Ortells, L. (2009). Validación de cuestionarios [Validation of questionnaires]. *Reumatologia clinica, 5*(4), 171-177. https://doi.org/10.1016/j.reuma.2008.09.007.

Ghaheri, A., Omani-Samani, R., Sepidarkish, M., Hosseini, M., y Maroufizadeh, S. (2020). The Four-item Patient Health Questionnaire for Anxiety and Depression: A Validation Study in Infertile Patients. *International journal of fertility & sterility, 14*(3), 234–239. https://doi.org/10.22074/ijfs.2020.44412

Goodwin G. M. (2015). The overlap between anxiety, depression, and obsessive-compulsive disorder. *Dialogues in clinical neuroscience, 17*(3), 249–260. https://doi.org/10.31887/DCNS.2015.17.3/ggoodwin

Guerra, V. M. L., Mejía, A. J. A., y Alcedo, J. M. G. (2022). Propiedades psicométricas y estructura factorial del cuestionario de salud del paciente PHQ-4 en estudiantes universitarios ecuatorianos. *Revista Cubana de Enfermería, 38*(3). Recuperado de https://revenfermeria.sld.cu/index.php/enf/article/view/4885/895.

Kelly, K. M., y Mezuk, B. (2017). Predictors of remission from generalized anxiety disorder and major depressive disorder. *Journal of Affective Disorders, 208*, 467-474. https://doi.org/10.1016/j.jad.2016.10.042

Kerper, L., Spies, C., Tillinger, J., Wegscheider, K., Salz, A., Weiss-Gerlach, E., Neumann, T., y Krampe, H. (2014). Screening for Depression, Anxiety, and General Psychological Distress in Pre-operative Surgical Patients: A Psychometric Analysis of the Patient Health Questionnaire 4 (Phq-4). *Clinical Health Promotion, 4*, 5-14. https://doi.org/10.29102/clinhp.14002

Kessler, R. C., Merikangas, K. R., y Wang, P. S. (2007). Prevalence, comorbidity, and service utilization for mood disorders in the United States at the beginning of the twenty-first century. *Annual Review of Clinical Psychology, 3,* 137-158. https://doi.org/10.1146/annurev.clinpsy.3.022806.091444

Khubchandani, J., Brey, R., Kotecki, J., Kleinfelder, J., y Anderson, J. (2016). The Psychometric Properties of PHQ-4 Depression and Anxiety Screening Scale Among College Students. *Archives of psychiatric nursing, 30*(4), 457–462. https://doi.org/10.1016/j.apnu.2016.01.014

Kim, H. W., Shin, C., Lee, S. H., y Han, C. (2021). Standardization of the Korean Version of the Patient Health Questionnaire-4 (PHQ-4). *Clinical psychopharmacology and neuroscience : the official scientific journal of the Korean College of Neuropsychopharmacology, 19*(1), 104–111. https://doi.org/10.9758/cpn.2021.19.1.104

Kocalevent, R. D., Finck, C., Jimenez-Leal, W., Sautier, L., y Hinz, A. (2014). Standardization of the Colombian version of the PHQ-4 in the general population. *BMC psychiatry, 14*, 205. https://doi.org/10.1186/1471-244X-14-205.

Kroenke, K., Baye, F., y Lourens, S. G. (2019). Comparative validity and responsiveness of PHQ-ADS and other composite anxiety-depression measures. *Journal of affective disorders, 246*, 437–443. https://doi.org/10.1016/j.jad.2018.12.098

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., y Löwe, B. (2007). Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. *Annals of internal medicine, 146*(5), 317-325. https://doi.org/10.7326/0003-4819-146-5-200703060-00004

Kroenke, K., Spitzer, R. L., Williams, J. B., y Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The PHQ–4. *Psychosomatics, 50*(6), 613-621. https://doi.org/10.1176/appi.psy.50.6.613.

Lamprea M, Julio Alejandro, y Gómez-Restrepo, Carlos. (2007). Validez en la evaluación de escalas [Validity in the assessment of scales]. *Revista Colombiana de Psiquiatría, 36*(2),

340-348. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0034-74502007000200013&lng=en&tlng=es.

Latorre Postigo, J. M., Navarro Bravo, B., Parra Delgado, M., Salguero, J. M., Mae Wood, C., y Cano Vindel, A. (2012). Evaluación e intervención de los problemas de Ansiedad y Depresión en Atención Primaria: Un Problema sin resolver. *Revista Clínica de Medicina de Familia, 5*(1), 37-45. https://dx.doi.org/10.4321/S1699-695X2012000100007

Lenz, A. S., y Li, C. (2022). Evidence for Measurement Invariance and Psychometric Reliability for Scores on the PHQ-4 From a Rural and Predominately Hispanic Community. *Measurement and Evaluation in Counseling and Development, 55*(1), 17-29. https://doi.org/10.1080/07481756.2021.1906157

Lim, G.Y., Tam, W.W., Lu, Y., Ho, C.S., Zhang, M.W., y Ho, R.C. (2018). Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific Reports, 8*, 2861. https://doi.org/10.1038/s41598-018-21243-x

Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., Schneider, A., y Brähler, E. (2010). A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of affective disorders, 122*(1-2), 86-95. https://doi.org/10.1016/j.jad.2009.06.019.

Luján-Tangarife, J. y Cardona-Arias, J. (2015). Construcción y validación de escalas de medición en salud: revisión de propiedades psicométricas. *Archivos de Medicina, 11* (3), 1-10. https://dialnet.unirioja.es/servlet/articulo?codigo=5178935

Martínez, C. M., y Sepúlveda, M. A. R. (2012). Introducción al análisis factorial exploratorio. *Revista colombiana de psiquiatría*, *41*(1), 197-207. https://doi.org/10.1016/S0034-7450(14)60077-9

Materu, J., Kuringe, E., Nyato, D., Galishi, A., Mwanamsangu, A., Katebalila, M., Shao, A., Changalucha, J., Nnko, S., y Wambura, M. (2020). The psychometric properties of PHQ-4 anxiety and depression screening scale among out of school adolescent girls and young women in Tanzania: a cross-sectional study. *BMC Psychiatry, 20*. https://doi.org/10.1186/s12888-020-02735-5

Maurer, D. M., Raymond, T. J., y Davis, B. N. (2018). Depression: screening and diagnosis. *American family physician, 98*(8), 508-515. https://pubmed.ncbi.nlm.nih.gov/30277728/

Mendoza, N. B., Frondozo, C. E., Dizon, J. I. W. T., y Buenconsejo, J. U. (2022). The factor structure and measurement invariance of the PHQ-4 and the prevalence of depression and anxiety in a Southeast Asian context amid the COVID-19 pandemic. *Current Psychology*. https://doi.org/10.1007/s12144-022-02833-5

Mills, S. D., Fox, R. S., Pan, T. M., Malcarne, V. L., Roesch, S. C., y Sadler, G. R. (2015). Psychometric Evaluation of the Patient Health Questionnaire-4 in Hispanic Americans. *Hispanic journal of behavioral sciences, 37*(4), 560–571. https://doi.org/10.1177/0739986315608126

Muñiz, J., Elosua, P., y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: Segunda edición. [Guidelines for the translation and adaptation of tests: Second edition]. *Psicothema, 25*(2), 151-157. http://dx.doi.org/10.7334/psicothema2012.218

Nicolini, H. (2020). Depresión y ansiedad en los tiempos de la pandemia de COVID-19. *Cirugía y Cirujanos, 88*(5), 542-547. https://doi.org/10.24875/CIRU.M20000067

Nunnally, J. C. (1967). Psychometric Theory. McGraw-Hill. https://www.scirp.org/(S(oyulxb452alnt1aej1nfow45))/reference/ReferencesPapers.aspx?ReferenceID=1298318

Organización Mundial de la Salud (2017). "Depression: let's talk" says who, as depression tops list of causes of ill health. *Saudi Medical Journal, 38*(5), 565.

Organización Panamericana de la Salud. (2017). *Depresión y otros trastornos mentales comunes*. Estimaciones sanitarias mundiales. Washington: OPS. https://iris.paho.org/bitstream/handle/10665.2/34006/PAHONMH17005-spa.pdf

Parés-Badell, O., Barbaglia, G., Jerinic, P., Gustavsson, A., Salvador-Carulla, L., y Alonso, J. (2014). Cost of disorders of the brain in Spain. *PloS one, 9*(8), e105471. https://doi.org/10.1371/journal.pone.0105471

Patten, S. B., Williams, J. V., Lavorato, D. H., Wang, J. L., McDonald, K., y Bulloch, A. G. (2015). Descriptive epidemiology of major depressive disorder in Canada in 2012. Canadian journal of psychiatry. *Revue canadienne de psychiatrie, 60*(1), 23–30. https://doi.org/10.1177/070674371506000106

Paykel, E. S., Brugha, T., y Fryers, T. (2005). Size and burden of depressive disorders in Europe. *European neuropsychopharmacology, 15*(4), 411-423. https://doi.org/10.1016/j.euroneuro.2005.04.008

Penninx, B. W., Pine, D. S., Holmes, E. A., y Reif, A. (2021). Anxiety disorders. *The Lancet, 397*(10280), 914-927. https://doi.org/10.1016/S0140-6736(21)00359-7

Remes, O., Brayne, C., Van Der Linde, R., y Lafortune, L. (2016). A systematic review of reviews on the prevalence of anxiety disorders in adult populations. *Brain and behavior, 6*(7), e00497. https://doi.org/10.1002/brb3.497

Rodríguez-Muñoz, M. F., Ruiz-Segovia, N., Soto-Balbuena, C., Le, H. N., Olivares-Crespo, M. E., y Izquierdo-Méndez, N. (2020). The Psychometric Properties of the Patient Health Questionnaire-4 for Pregnant Women. *International journal of environmental research and public health, 17*(20), 7583. https://doi.org/10.3390/ijerph17207583

Salari, N., Hosseinian-Far, A., Jalali, R., Vaisi-Raygani, A., Rasoulpoor, S., Mohammadi, M., Rasoulpoor, S., y Khaledi-Paveh, B. (2020). Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: A systematic review and meta-analysis. *Globalization and Health, 16*(1), 57. https://doi.org/10.1186/s12992-020-00589-w

Sanabria-Mazo, J. P., Gómez-Acosta, A., Castro-Muñoz, J. A., Rojas, Y. F., Soler, A. F., Luciano, J. V., y Sanz, A. (2023). Dimensionality and reliability of the online version of the Patient

Health Questionnaire-4 in a large Colombian sample: Results from the PSY-COVID study. *Current Psychology*.

Schermelleh-Engel, K., Moosbrugger, H., y Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74. https://psycnet.apa.org/record/2003-08119-003

Stanhope, J. (2016). Patient Health Questionnaire-4. *Occupational Medicine, 66*(9), 760-761. https://doi.org/10.1093/occmed/kqw165

Tayeh, P., Agámez, P., y Chaskel, R. (2016). Trastornos de ansiedad en la infancia y la adolescencia. *Precop SCP, 15*(1), 6-18. https://www.aepap.org/sites/default/files/ansiedad_0.pdf

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M., y de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Tibubos, A. N., y Kröger, H. (2020). A cross-cultural comparison of the ultrabrief mental health screeners PHQ-4 and SF-12 in Germany. *Psychological assessment, 32*(7), 690–697. https://doi.org/10.1037/pas0000814

Urrutia Egaña, M., Barrios Araya, S., Gutiérrez Núñez, M., y Mayorga Camus, M. (2014). Métodos óptimos para determinar validez de contenido. *Educación médica superior, 28*(3), 547-558. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=s0864-21412014000300014

Vos, T., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., ... y Abdulle, A. M. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *PLoS medicine, 15*(11), e1002624. https://doi.org/10.1016/S0140-6736(18)32279-7

Weisberg, R. B., Dyck, I., Culpepper, L., y Keller, M. B. (2007). Psychiatric treatment in primary care patients with anxiety disorders: A comparison of care received from primary care providers and psychiatrists. *American Journal of Psychiatry, 164*(2), 276-282. https://doi.org/10.1176/ajp.2007.164.2.276

Wicke, F. S., Krakau, L., Löwe, B., Beutel, M. E., y Brähler, E. (2022). Update of the standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders, 312*, 310-314. https://doi.org/10.1016/j.jad.2022.06.054.

Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, A., van Os, J., Preisig, M., Salvador-Carulla, L., Simon, R., y Steinhausen, H. C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology, 21*(9), 655-679. https://doi.org/10.1016/j.euroneuro.2011.07.018

6. Anexos

Anexo 1

Estrategia de búsqueda en bases de datos.

Medline (*n* = 404)
((TI=("Patient Health Questionnaire") OR TI=("PHQ")) AND (TI=("psychometric") OR AB=("psychometric") OR TI=("factor analysis") OR AB=("factor analysis") OR TI=("factor structure") OR AB=("factor structure") OR TI=("reliability") OR AB=("reliability") OR TI=("intraclass") OR AB=("intraclass") OR TI=("test-retest") OR AB=("test-retest") OR TI=("internal consistency") OR AB=("internal consistency") OR TI=("validity") OR AB=("validity") OR TI=("dimensionality") OR AB=("dimensionality") OR TI=("variance") OR AB=("variance") OR TI=("known group") OR AB=("known group") OR TI=("multigroup") OR AB=("multigroup") OR TI=("sensitivity to change") OR AB=("sensitivity to change") OR TI=("responsiveness") OR AB=("responsiveness") OR TI=("normative data") OR AB=("normative data") OR TI=("sensibility") OR AB=("sensibility") OR TI=("specificity") OR AB=("specificity")))
Web of Science (*n* = 429)
((TI=("Patient Health Questionnaire") OR TI=("PHQ")) AND (TI=("psychometric") OR AB=("psychometric") OR TI=("factor analysis") OR AB=("factor analysis") OR TI=("factor structure") OR AB=("factor structure") OR TI=("reliability") OR AB=("reliability") OR TI=("intraclass") OR AB=("intraclass") OR TI=("test-retest") OR AB=("test-retest") OR TI=("internal consistency") OR AB=("internal consistency") OR TI=("validity") OR AB=("validity") OR TI=("dimensionality") OR AB=("dimensionality") OR TI=("variance") OR AB=("variance") OR TI=("known group") OR AB=("known group") OR TI=("multigroup") OR AB=("multigroup") OR TI=("sensitivity to change") OR AB=("sensitivity to change") OR TI=("responsiveness") OR AB=("responsiveness") OR TI=("normative data") OR AB=("normative data") OR TI=("sensibility") OR AB=("sensibility") OR TI=("specificity") OR AB=("specificity")))
PsycINFO (*n* = 283)

((TITLE("Patient Health Questionnaire") OR TITLE("PHQ")) AND (TITLE("psychometric") OR ABSTRACT("psychometric") OR TITLE("factor analysis") OR ABSTRACT("factor analysis") OR TITLE("factor structure") OR ABSTRACT("factor structure") OR TITLE("reliability") OR ABSTRACT("reliability") OR TITLE("intraclass") OR ABSTRACT("intraclass") OR TITLE("test-retest") OR ABSTRACT("test-retest") OR TITLE("internal consistency") OR ABSTRACT("internal consistency") OR TITLE("validity") OR ABSTRACT("validity") OR TITLE("dimensionality") OR ABSTRACT("dimensionality") OR TITLE("variance") OR ABSTRACT("variance") OR TITLE("known group") OR ABSTRACT("known group") OR TITLE("multigroup") OR ABSTRACT("multigroup") OR TITLE("sensitivity to change") OR ABSTRACT("sensitivity to change") OR TITLE("responsiveness") OR ABSTRACT("responsiveness") OR TITLE("normative data") OR ABSTRACT("normative data") OR TITLE("sensibility") OR ABSTRACT("sensibility") OR TITLE("specificity") OR ABSTRACT("specificity")))

Scopus (*n* = 426)

((TITLE("Patient Health Questionnaire") OR TITLE("PHQ")) AND (TITLE("psychometric") OR ABS("psychometric") OR TITLE("factor analysis") OR ABS("factor analysis") OR TITLE("factor structure") OR ABS("factor structure") OR TITLE("reliability") OR ABS("reliability") OR TITLE("intraclass") OR ABS("intraclass") OR TITLE("test-retest") OR ABS("test-retest") OR TITLE("internal consistency") OR ABS("internal consistency") OR TITLE("validity") OR ABS("validity") OR TITLE("dimensionality") OR ABS("dimensionality") OR TITLE("variance") OR ABS("variance") OR TITLE("known group") OR ABS("known group") OR TITLE("multigroup") OR ABS("multigroup") OR TITLE("sensitivity to change") OR ABS("sensitivity to change") OR TITLE("responsiveness") OR ABS("responsiveness") OR TITLE("normative data") OR ABS("normative data") OR TITLE("sensibility") OR ABS("sensibility") OR TITLE("specificity") OR ABS("specificity")))

Los filtros siguientes se aplicaron en todas las bases de datos que fue posible: articles, 2009 to 2022, English and Spanish.
Nota: Elaboración propia.

Anexo 2

Estructura factorial del PHQ-4.

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población clínica					
Kroenke et al. (2009)	ACP	Modelo de dos factores	Índices de ajuste adecuados en el modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.82-0.90	NR
Kerper et al. (2014)	ACP	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.81-0.96	NR

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.1

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población clínica					
Cano-Vindel et al. (2018)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.71 Rango de cargas factoriales de los ítems: 0.77-0.87	CFI = 0.99 TLI = 0.98 RMSEA = 0.08 AIC = 218.81 BIC = 258.49
Ahmadi et al. (2019)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.66 Rango de cargas factoriales de los ítems: 0.69-0.92	RMSEA = 0.03 NFI = 0.99 IFI = 0.99 TLI = 0.99 CFI = 0.99

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.2

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población clínica					
Ghaheri et al. (2020)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.74 Rango de cargas factoriales de los ítems: 0.74-0.90	CFI = 1.00 RMSEA < 0.001 SRMR = 0.001
Kim et al. (2021)	AFE/AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.76 Rango de cargas factoriales de los ítems: 0.67-0.83	RMSEA = 0.141 SRMR = 0.023 TLI = 0.901 CFI = 0.983

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.3

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población no clínica					
Löwe et al. (2010)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.79 Rango de cargas factoriales de los ítems: 0.76-0.87	CFI=0.984 TLI=0.988 RMSEA=0.027
Kocalevent et al. (2014)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.70-0.85	GFI = 0.989 NFI = 0.987 TLI = 0.923 CFI = 0.987 RMSEA = 0.145

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.4

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población no clínica					
Mills et al. (2015)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.74-1.35	CFI = 0.973 RMSEA = 0.082 SRMR = 0.071
Materu et al. (2020)	AFC/ACP	Modelo de un factor	Índices de ajuste adecuados en el modelo unidimensional	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.67-0.77	CFI = 0.995 SRMR = 0.013 RMSEA = 0.054

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.5

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población no clínica					
Tibubos y Kröger (2020)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.67-0.77	CFI = 0.994; RMSEA = 0.062 SRMR = .0.011
Barrera et al. (2021)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.88 Rango de cargas factoriales de los ítems: 0.62-0.79	SRMR = 0.006 RMSEA = 0.030 CFI = 0.999

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.7

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población no clínica					
Lenz & Li (2021)	AFC	Modelo de un factor	Índices de ajuste adecuados en el modelo de un factor	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.66-0.85	CFI = 0.981 SRMR=0.021
Christodoulaki et al. (2022)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.83 Rango de cargas factoriales de los ítems: 0.80-0.91	CFI = 0.998 TLI = 0.986 RMSEA = 0.100

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.8

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población no clínica					
Guerra et al. (2022)	AFE/AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.78 Rango de cargas factoriales de los ítems: 0.64-0.87	CFI = 0.996 GFI = 0.996 NFI = 0.996 PNFI = 0.166 AIC = 14621 RMSEA = 0.087
Mendoza et al. (2022)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.92 Rango de cargas factoriales de los ítems: 0.74-0.86	CFI = 1.000 TLI = 1.000 RMSEA = 0.000 SRMR = 0.001

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 2.9

Estructura factorial del PHQ-4 (continuación).

Autor (año)	Tipo de análisis (AFE/AFC/ACP)	Modelos explorados	Modelo con mejores índices de ajuste	Cargas factoriales de los ítems	Índices de ajuste
Población no clínica					
Rodríguez-Muñoz et al. (2022)	AFE/AFC	Modelo de un factor. Modelo de dos factores	Índices de ajuste favorables del modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: NR Rango de cargas factoriales de los ítems: 0.77-0.91	RMSEA = 0.069 CFI = 0.99 NFI = 0.90 AIC = 31.006 ECVI = 0.037 PNFI = 0.09
Sanabria-Mazo et al. (2023)	AFC	Modelo de un factor y de dos factores	Índices de ajuste favorables al modelo de 2 factores	Carga factorial estándar en el modelo de dos factores: 0.83 Rango de cargas factoriales de los ítems: 0.71-0.92	CFI = 0.99 TLI = 0.99 NFI = .99 RMSEA = 0.04

AFE: Análisis factorial exploratorio; AFC: Análisis factorial confirmatorio; ACP: Análisis de componentes principales. NR: No reportado.

Nota: Elaboración propia.

Anexo 3

Estudio de la invarianza del PHQ-4 en función de determinadas variables sociodemográficas.

Autor (año)	Género	Grupos de edad	Localización geográfica	Nivel de ingresos	Nivel educativo	Idioma
Población clínica						
Cano-Vindel et al. (2018)	No existe invarianza para esta variable	No existe invarianza para esta variable	NR	NR	NR	NR
Población no clínica						
Löwe et al. (2010)	No existe invarianza para esta variable	No existe invarianza para esta variable	NR	NR	NR	NR
Kocalevent et al. (2014)	No existe invarianza para esta variable	No existe invarianza para esta variable	NR	NR	NR	NR
Mills et al. (2015)	NR	NR	NR	NR	NR	No existe invarianza para esta variable

*Los resultados del estudio de Tibubos y Kröger (2020) aportan evidencia es invariante a nivel transcultural. NR: No reportado.

Nota: Elaboración propia.

Anexo 3.1

Estudio de la invarianza del PHQ-4 en función de determinadas variables sociodemográficas (continuación).

Autor (año)	Género	Grupos de edad	Localización geográfica	Nivel de ingresos	Nivel educativo	Idioma
Población no clínica						
Tibubos y Kröger (2020)*	NR	NR	NR	NR	NR	NR
Barrera et al. (2021)	NR	NR	NR	NR	NR	No existe invarianza para esta variable
Lenz & Li (2021)	No existe invarianza para esta variable	No existe invarianza para esta variable	NR	NR	NR	No existe invarianza para esta variable
Christodoulaki et al. (2022)	No existe invarianza para esta variable	No existe invarianza para esta variable	NR	NR	NR	NR

*Los resultados del estudio de Tibubos y Kröger (2020) aportan evidencia es invariante a nivel transcultural. NR: No reportado.

Nota: Elaboración propia.

Anexo 3.2

Estudio de la invarianza del PHQ-4 en función de determinadas variables sociodemográficas (continuación).

Autor (año)	Género	Grupos de edad	Localización geográfica	Nivel de ingresos	Nivel educativo	Idioma
Población no clínica						
Guerra et al. (2022)	No existe invarianza para esta variable	NR	NR	NR	NR	NR
Mendoza et al. (2022)	No existe invarianza para esta variable	No existe invarianza para esta variable	No existe invarianza para esta variable.	NR	NR	NR
Sanabria-Mazo et al. (2023)	No existe invarianza para esta variable	No existe invarianza para esta variable	No existe invarianza para esta variable.	No existe invarianza para esta variable.	No existe invarianza para esta variable.	NR

*Los resultados del estudio de Tibubos y Kröger (2020) aportan evidencia es invariante a nivel transcultural. NR: No reportado.

Nota: Elaboración propia.

Anexo 4

Fiabilidad test retest del PHQ-4.

Autor (año)	Tiempo inter-test	Coeficiente de fiabilidad
Población clínica		
Kim et al. (2021).	2 semanas	Coeficiente de correlación intraclase (ICC)= 0.827
Población no clínica		
Khubchandani et al. (2016)	10 días	Coeficiente Kappa= Entre 0.69 y 0.81.
Christodoulaki et al (2022).	1 semana	Coeficiente de correlación intraclase (ICC)= 0.96

Nota: Elaboración propia.

Anexo 5

Interpretabilidad del PHQ-4 y sus subescalas dependiendo de variables sociodemográficas.

Autor (año)	Resultados según variables sociodemográficas
Población clínica	
Ghaheri et al. (2020)	**Género:** Las mujeres obtuvieron puntuaciones significativamente más altas que los hombres en el PHQ-4 (M= 5.45, DE= 3.18 vs. M= 3.67, DE=3.11), en el PHQ-2 (M=2.86, DE=1.77 vs. M=1.90, DE=1.83) y en el GAD-2 (M=2.59, DE=1.86 vs. M=1.78, DE=1.70). **Nivel educativo:** Las personas con menor nivel educativo obtuvieron puntuaciones significativamente más altas que las de un nivel educativo más alto en el PHQ-4 (M=5.16, DE=3.36 vs. M=4.56,DE=3.24), en el PHQ-2 (M=2.54,DE=1.91 vs. M=2.41, DE= 1.86) y en el GAD-2 (M=2.62, DE=1.99 vs. M=2.15, DE=1.76).
Kim et al. (2021)	**Género**: No hay diferencias significativas. **Edad**: Las personas de menor edad obtuvieron puntuaciones significativamente más altas que las de una edad mayor en el PHQ-4 (M=7.33, DE=3.371 vs. M=4.87, DE=3.357). **Estado civil:** Las personas separadas y solteras obtuvieron puntuaciones significativamente más altas que las personas casadas y viudas en el PHQ-4 (M=7.86,DE=1.864 y M=7.26; DE=3.257 vs. M=5.00, DE=3.618 y M=3.67, DE=2.082). **Estado laboral:** No hay diferencias significativas. **Nivel económico**: Las personas con un nivel económico bajo obtuvieron puntuaciones significativamente más altas que las personas con un nivel económico alto en el PHQ-4 (M=8.83, DE=2.980 vs. M=5.28, DE=3.308).

Nota: Elaboración propia.

Anexo 5.1

Interpretabilidad del PHQ-4 y sus subescalas dependiendo de variables sociodemográficas (continuación).

Autor (año)	Resultados según variables sociodemográficas
Población no clínica	
Löwe et al. (2010)	**Género**: Las mujeres obtuvieron puntuaciones significativamente más altas que los hombres en el PHQ-4 (M=1.96, DE=2.12 vs. M=1.56, DE=1.97), en PHQ-2 (M=1.00, DE=1.22 vs. M=0.87, DE=1.17) y GAD-2 (M=0.93, DE=1.14 vs. M=0.70, DE=1.03). **Edad**: Las personas de mayor edad obtuvieron puntuaciones más altas que las personas de menor edad en el PHQ-4 (M=2.22, DE=2.25 vs. M=1.55, DE=1.95), en el PHQ-2 (M=1.27, DE=1.41 vs. M=0.83, DE=1.11) y en el GAD-2 (M=0.95, DE=1.10 vs. M=0.72, DE=1.08). **Estado civil:** Las personas no viven en parejas puntuaron significativamente más alto que aquellas que viven en pareja en el PHQ-4 (M=2.00, DE=2.23 vs. M=1.59, DE=1.92), en el PHQ-2 (M=1.08, DE=1.30 vs. M=0.84, DE=1.12) y en el GAD-2 (M=0.92,DE=1.17 vs. M=0.76, DE=1.04). **Estado laboral:** Las personas desempleadas obtuvieron puntuaciones significativamente más altas que las personas empleadas en el PHQ-4 (M=3.55, DE=2.81 vs. M=1.65, DE=1.95), en el PHQ-2 (M=1.93, DE=1.57vs. M=0.88, DE=1.15) y en el GAD-2 (M=1.60, DE=1.52 vs. M=0.77, DE=1.05). **Nivel económico**: Las personas con menor nivel económico obtuvieron puntuaciones significativamente más altas que las de un nivel económico más alto en el PHQ-4 (M=3.00, DE=2.73 vs. M=1.40, DE=1.76), en el PHQ-2 (M=1.61, DE=1.57 vs. M=0.73, DE=1.04) y en el GAD-2 (M=1.38, DE=1.47 vs. M=0.67, DE=0.96).
Kocalevent et al. (2014)	**Género:** Las mujeres puntuaron significativamente más alto que los hombres en el PHQ-4 (M=1.41, DE=2.09 vs. M=1.13, DE=1.90). **Edad:** No existen diferencias significativas. **Estado civil:** No existen diferencias significativas**.** **Nivel educativo:** Las personas con un nivel educativo más bajo puntuaron más alto que las de un nivel educativo superior en el PHQ-4 (M=1.03, DE=1.76 vs. M=1.77, DE=2.26)**.** **Estado laboral:** Las personas minusválidas puntuaron significativamente más alto que las personas empleadas en el PHQ-4 (M=2.71, DE=3.29 vs. M=1.12; DE=2.93).

Nota: Elaboración propia.

Anexo 5.2

Interpretabilidad del PHQ-4 y sus subescalas dependiendo de variables sociodemográficas (continuación).

Autor (año)	Resultados según variables sociodemográficas
Población no clínica	
Barrera et al. (2021)	**Estado civil**: Las mujeres solteras obtuvieron puntuaciones significativamente más altas que las mujeres casadas en el PHQ-4 (M=5.44, DE=3.47vs. M= 4.54, DE= 3.33), en el PHQ-2 (M=2.68, DE=1.92 vs. M= 2.32, DE= 1.91) y en el GAD-2 (M=2.76, DE=1.93 vs. M= 2.22, DE= 1.80).
Lenz y Li (2021)	**Género:** No existen diferencias significativas. **Edad:** Las personas más jóvenes obtuvieron puntuaciones significativamente más altas que las personas de más edad en el PHQ-4 (M=3.01, DE=3.19 vs. M=2.69, DE=3.07).
Rodríguez-Muñoz et al. (2022)	**Edad**: Las mujeres jóvenes obtuvieron puntuaciones significativamente más altas que las mujeres mayores en el PHQ-4 (M=3.15, DE= 2.78 vs. M= 2.37, DE= 2.45), en el PHQ-2 (M=1.82, DE=1.67 vs. M=1.33, DE=1.37) y en el GAD-2 (M=1.29, DE=1.53 vs. M=1.09, DE=1.32). **Nivel educativo:** Las mujeres jóvenes con un nivel educativo bajo obtuvieron puntuaciones significativamente más altas que las mujeres con un nivel educativo alto en el PHQ-4 (M= 2.83, DE= 2.66 vs. M=2.12, DE= 2.07), en el PHQ-2 (M=1.59, DE=1.61 vs. M=1.14, DE=1.22) y en el GAD-2 (M=1.26, DE=1.48 vs. M=1.00, DE=1.26). **Estado Laboral:** Las mujeres desempleadas obtuvieron puntuaciones significativamente más altas que las mujeres empleadas en el PHQ-4 (M=2.86, DE=2.72 vs. M=2.42, DE=2.36), en el PHQ-2 (M=1.63, DE= 1.67 vs. M=1.32, DE=1.38) y en el GAD-2 (M=1.24, DE=1.51 vs. M=1.12, DE=1.34). **Estado civil:** No se encontraron diferencias significativas.
Sanabria-Mazo et al. (2023)	**Género:** Las mujeres obtuvieron puntuaciones significativamente más altas que los hombres en el PHQ-4 (M=4.42, DE=3.00 vs. M=3.86, DE=2.99), en el PHQ-2 (M=2.34, DE=1.60 vs. M=2.09, DE=1.63) y en el GAD-2 (M=2.08, DE=1.68 vs. M=1.77, DE=1.63). **Edad**: Las personas de menor edad obtuvieron puntuaciones más altas que las personas de mayor edad en el PHQ-4 (M=5.41, DE=3.14 vs. M=3.08, DE=2.51), en el PHQ-2 (M=2.98, DE=1.65 vs. M=1.53, DE=1.37) y en el GAD-2 (M=2.43, DE=1.81 vs. M=1.55, DE=1.40). **Nivel educativo:** No hay diferencias significativas. **Estado laboral:** Las personas desempleadas obtuvieron puntuaciones significativamente más altas que las personas empleadas en el PHQ-4 (M=4.52, DE=3.09 vs. M=4.13, DE=2.94), en el PHQ-2 (M=2.43, DE=1.65 vs. M=2.18, DE=1.57) y en el GAD-2 (M=2.09, DE=1.72 vs. M=1.95, DE=1.64).

Nota: Elaboración propia.