**UNIVERSIDAD DE SALAMANCA**
**DEPARTAMENTO DE ESTADÍSTICA**

**Doctorado en Estadística Multivariante Aplicada**
**Tesis doctoral**

# Mejoras psicométricas en la evaluación de la salud pública: depresión en Costa Rica

Autor: Armando González Sánchez

Directora: M.ª Purificación Vicente Galindo

*Salamanca, noviembre de 2023*

# Mejoras psicométricas en la evaluación de la salud pública: depresión en Costa Rica

Memoria que para optar al **Grado de Doctor, por el Departamento de Estadística de la Universidad de Salamanca**, presenta:

Armando González Sánchez

Salamanca

2023

*Universidad de Salamanca*
*Departamento de Estadística*

# M.ª PURIFICACIÓN VICENTE GALINDO

*Profesora Titular del Departamento de Estadística e Investigación Operativa de la Universidad de Salamanca*

---

CERTIFICA: Que don Armando González Sánchez, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo para optar al Grado de Doctor en Estadística Multivariante Aplicada, que presenta con el título: "Mejoras psicométricas en la evaluación de la salud pública: depresión en Costa Rica", autorizando expresamente su lectura y defensa

Y para que conste, firma el presente certificado en Salamanca, en noviembre de 2023.

M.ª Purificación Vicente Galindo

*No olvides que la ciencia exige del individuo su vida entera. Si tuviéramos dos vidas, tampoco serían suficientes.*

Iván Petróvich Pávlov

# Agradecimientos

Gracias a mi madre, quien en toda su vida ha proveído en todo aquello que ha tenido oportunidad. Gracias por la forma de dar cariño y por cuidar de tus hijos. Gracias por tu disponibilidad y atención. Gracias por las comidas ricas y por los esfuerzos en que todo salga bien.

A mi amigo Grebin por su apoyo moral en las múltiples e incesantes vicisitudes acaecidas en este recodado periplo. Agradezco la labor de amistad y apoyo emocional a mis compañeros Nathalia, Joel y Edith. Con un saludo especial a Amelia.

A Pablo, Paris y Mario. A Camilla, Dante y Vega. A Valentina. Gracias a Ingrid Schultz.

A la memoria de Pedro Montero Gütts, que tu esencia perdure siempre.

Gracias a Enma Patricia Guallo Guevara.

Agradezco a mi familia: Ainara, Aitana, Alex, Alfonso Calvo, Álvaro, Ana, Ángel, Ángel Casero, Arturo, Aylén, Carlos, Carolina, Claudia, Coty, Cristina, David, David (Vero), Edu, Elena, Eric, Guille, Jara, Jorge, José, Juan, Manolo, Mar Calvo, Mar Chapado, María, Marina, Marta, Miguel, Nerea, Nicolás, Nines, Noe, Patricia, Paula, Pavel, Pedro, Pilar, Pili, Rodrigo, Rubén, Samara, Sara, Sergio, Silvia, Sonia, Super, Tere, Tomás, Toño, Valeria y Vero; así como a Carlitos, Carlos, José Santos, Juani y Raúl.

A los integrantes del departamento de Psiquiatría del CAUSA por dejarme ser parte de su equipo que me ha permitido crecer como investigador, así como poder costear el pago a la revista y otros gastos asociados.

Al equipo de investigación con quien tuve el gusto de colaborar en las investigaciones de salud mental en tiempos de COVID de población general y del Ministerio de Educación: Eva, Raúl, Harold y Greibin. Con deseos de pronta recuperación a Eva.

Al Ministerio de Salud de Costa Rica, al Ministerio de Seguro Social de Costa Rica y al Ministerio de Educación de Costa Rica por valorar las aportaciones a la investigación que he realizado en su población. A la Universidad Nacional de Costa Rica, a la Universidad a Distancia de Costa Rica y a sus rectores por permitirme aportar conocimiento para mejorar la salud de las personas. Agradecimientos a Karol Ramirez por su gran labor sobre la difusión del cuestionario a la población a través de las entrevistas que nos hiciste en la radio, la cartelería y la coordinación institucional.

Agradezco a aquellas personas o entidades cuya correlación directa entre el nivel de apoyo recibido y la intención de apoyar haya sido significativa. A las constantes que siempre están ahí, aunque la monotonía de su presencia se las infiera como ausentes o ignoradas.

Que el MEV me haga un sitio en el paraíso.

Pido un minuto de silencio por el artículo: “Propiedades psicométricas del test Fear of COVID-19 en población docente costarricense” que nunca vio la luz por razones políticas.

# Índice

## Contenido

# Índice de tablas

# Índice de figuras

# Presentación

Esta tesis doctoral trata sobre la aplicación de la psicometría enfocada a la aplicación práctica. En concreto se estudiará cómo, a través de la estadística, se podrán mejorar las valoraciones poblacionales sobre salud mental.

Primeramente, en la primera parte, de marco contextual y teórico, se expondrá la situación contextual; ya que este estudio se enmarca en una petición de las autoridades gubernamentales de Costa Rica para evaluar el impacto de la pandemia por COVID-19 en la población. En la primera parte también se hará una introducción al concepto de psicometría, así como al de salud mental.

La segunda parte, de objetivos y metodología, se definirán los objetivos de este proyecto, siempre desde una visión práctica del trabajo; y se definirá la metodología, exponiendo la manera en la que los objetivos serán conseguidos.

En la tercera parte, de resultados y discusión, se hará una exposición de lo encontrado en materia de psicometría en test de depresión, además se comentará lo encontrado ciñéndose a los objetivos planteados y desde una perspectiva práctica.

En la cuarta parte se presentarán las conclusiones del presente estudio contestando a la pregunta sobre la utilidad práctica del mismo. Para las cuestiones que se escapan de los objetivos y que por tanto no entran dentro de la presente investigación se plantean líneas de investigación que retomen lo encontrado.

La quinta parte se reserva para que el lector pueda conocer los test de los que se trata, así como revisar las referencias utilizadas. Esta información queda esquematizada en la siguiente tabla (Tabla 1. Organización de las partes y capítulos de la presente tesis).

Tabla 1. Organización de las partes y capítulos de la presente tesis

| Parte | | Capítulo | |
|---|---|---|---|
| **Nº** | **Nombre** | **Nº** | **Nombre** |
| 1 | Marco contextual y teórico | 1 | Situación contextual |
| | | 2 | Psicometría |
| | | 3 | Salud mental |
| 2 | Objetivos y Metodología | 4 | Objetivos e hipótesis |
| | | 5 | Metodología |
| 3 | Resultados y discusión | 6 | Resultados |
| | | 7 | Discusión |
| 4 | Conclusiones y perspectivas futuras de investigación | 8 | Conclusiones |
| | | 9 | Perspectivas futuras de investigación |
| 5 | Referencias y anexos | 10 | Referencias |
| | | 11 | Anexos |

# Resumen

Durante la emergencia sanitaria por COVID-19 las entidades gubernamentales encargadas de la seguridad sanitaria: el Ministerio de Sanidad, la Caja Costarricense de Seguro Social nos encargaron una serie de estudios periódicos para monitorear el estado de salud mental de la población y actuar en consecuencia. Para ello se conformó un grupo de trabajo compuesto por profesionales expertos en psicología de emergencias, psicometría, psicología social, estadística, matemáticas y biología. Algunos de ellos actuaban en representación de las principales universidades: Autónoma (UNA) Universidad Nacional (UNA) y la Universidad de Educación a Distancia de Costa Rica (UNED) y otros actuaban en nombre propio. Los informes de estas investigaciones fueron expuestos ante las autoridades, quienes usaron esta información para planificar sus políticas de salud pública.

Las conclusiones de los informes debían tratarse con sumo cuidado, ya que la mayoría de los test utilizados no habían sido validados previamente en el país, por lo que podrían tener un funcionamiento diferente en esta población; y es que Costa Rica es el país de Latinoamérica que reporta más felicidad. Debido a su singular felicidad, a las características en salud mental que no habían sido previamente estudiadas a una magnitud poblacional y a sus características lingüísticas y culturales se estimó imprescindible la adaptación tanto cultural como lingüística de las escalas, así como su validación para la población general.

De entre todas las escalas psicométricas utilizadas para evaluar la salud mental, se eligió la depresión por ser una dimensión de gran incidencia mundial. Para estudiar la depresión se utilizó el test llamado PHQ-9 (Patient Health Questionnaire-9), el cual consta de 9 ítems que evalúan diferentes facetas de la depresión recogidas en el DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, cuarta versión). Aunque es un test completo, puede usarse por módulos: pueden retirársele ítems para acortar la escala; de esta forma surgen los test PHQ-8, al retirar el noveno ítem y PHQ-2, al dejar solamente los dos primeros ítems (los

más definitorios de depresión); por lo que ya que se realizaba el trabajo de validar uno, se validarían sus módulos.

Para conocer las propiedades psicométricas de la escala y comprobar si es útil para medir depresión en la población costarricense se utilizaron varias técnicas como la consistencia interna, análisis factorial y análisis de fiabilidad y validez. Además se descubrió que no se estaba detectando adecuadamente la depresión en función del sexo o incluso para población general, por lo que para mejorar esta detección se propusieron nuevos puntos de corte en virtud de una adaptación más adecuada para la población. Así se pudo adaptar la escala para que pueda ser utilizada en la población costarricense de forma confiable y válida.

**Palabras clave:** Depresión, Patient Health Questionnaire, Costa Rica, Estadística Multivariante Aplicada, Big Data, Test, Psicometría, Adaptación Cultural, Adaptación Lingüística, Fiabilidad, Validez, Sensibilidad, Especificidad, Funcionamiento Diferencial, Rasch, Curva ROC, Modelos de Ecuaciones Estructurales, Invarianza de Medida Multigrupo, Análisis Factorial Confirmatorio Multigrupo.

# PRIMERA PARTE: MARCO CONTEXTUAL Y TEÓRICO

# Capítulo 1

## Contexto

En este capítulo se contextualiza la investigación de la cual parte la tesis doctoral. En un primer momento se configuró el proyecto para evaluar la salud mental en la población general costarricense. Tras largas deliberaciones y coordinaciones con los organismos gubernamentales se creó un equipo de investigación de coordinación entre universidades para realizar esta novedosa evaluación. Todo ello bajo el omnipresente miedo a la enfermedad de COVID y un futuro que se vislumbraba difuso.

Se aplicaron pruebas psicológicas conocidas y otras pruebas novedosas. Así surgió la oportunidad tanto de evaluar a la población general como de conocer la eficacia de estas herramientas psicológicas.

## 1.-Situación contextual

A finales del año 2019 se empiezan a reportar casos de COVID-19, así como sus consecuencias a nivel de salud mental (Li et al., 2020). El COVID-19 se trata de una enfermedad provocada por el virus SARS-CoV-2 que se propaga con facilidad. Esta enfermedad ha ido paulatinamente reduciéndose en severidad, disminuyendo las muertes debido a ella, pero para contenerla fue necesario aplicar medidas como el confinamiento de los ciudadanos en sus hogares, restricción de los desplazamientos vehiculares, limitación de relaciones sociales presenciales o uso de mascarillas. La medida del confinamiento fue tomada a nivel global. Además la amenaza constante de la infección y las subsecuentes implicaciones a la vida cotidiana; provocaron repercusión en la salud mental de las personas (Andrés-Olivera et al., 2022; Gamonal-Limcaoco et al., 2021; Roncero et al., 2022). En la población general durante el confinamiento se han reportado desórdenes psicológicos (Liu et al., 2020; Yahya et al., 2020) con prevalencias de ansiedad generalizada en su grado moderado del 29,29% (Qiu et al., 2020)

al 35,1% (Huang & Zhao, 2020). Mientras que los síntomas depresivos se han reportado con una prevalencia del 20,1% (Huang & Zhao, 2020).

El 29 de marzo de 2020 Greibin Villegas, junto con su equipo (Villegas et al., 2020) lanzaron una página web en español para monitorear el avance de la pandemia por COVID-19. En esta página se incluía un mapa con actualizaciones constantes sobre el avance de la pandemia. Yo mismo formé parte de este equipo, y siendo el único psicólogo se me solicitó evaluar la salud mental de los ciudadanos costarricenses, así como sus posibles implicaciones. Las variables seleccionadas fueron depresión, ansiedad, ideación suicida, percepción del miedo a la enfermedad y adherencia a las medidas de protección. Así que se ideó una batería de test que pudieran medir estar variables. Ya que se trataba de nuevos problemas sobre los que no existían cuestionarios se desarrollaron nuevos cuestionarios para medir estos nuevos constructos, como la adherencia a las medidas de protección frente a la COVID-19 o la percepción de miedo social a la COVID-19. Se publicitó en redes sociales en Costa Rica, así como en medios de comunicación del país, y en un solo día se obtuvieron 800 cuestionarios completos. Se hizo un informe de los resultados (Villegas et al., 2020) y los medios como prensa escrita, radio y televisión se hicieron eco de la situación psicológica que se estaba viviendo en esos momentos.

Ante estos resultados el Ministerio de Salud de Costa Rica, a través de la Mesa Técnica de Salud Mental, se puso en contacto con nosotros interesándose por un segundo estudio que incluyese otras variables de interés para esta institución. Además la institución pública encargada de la seguridad social en la República de Costa Rica, la Caja Costarricense de Seguro Social (CCSS), también se interesó. Fruto de este interés conjunto y tras muchas reuniones con instituciones, propuestas de nuevos cuestionarios, manifestación del punto de vista de los representantes gubernamentales y otras figuras de renombre en políticas públicas y en salud mental, se conformó un grupo de investigadores de diferentes especialidades y representando

diferentes instituciones para llevar a cabo un estudio respaldado por el gobierno y que respondiese a las preguntas que se planteaban. Todo ello para poder diseñar políticas públicas de salud mental a través de identificar el estado de la población en temas de salud mental, así como los grupos poblacionales especialmente afectados. Este grupo de investigación estaba compuesto por representantes de las principales universidades públicas del país: la Universidad Nacional (UNA) y la Universidad de Educación a Distancia de Costa Rica (UNED), conformándose como un orgullo nacional el poder coordinarse ante un problema común con una investigación de interés nacional.

Después de muchas reuniones, estudiar las variables a incluir, propuestas de cuestionarios, traducciones y adaptaciones culturales, la redacción del proyecto de investigación, la aprobación de las universidades, el Consejo Nacional de Investigaciones Científicas y el acuerdo de los investigadores se lanzó un cuestionario de salud mental para toda la población costarricense que estuvo disponible desde el 9 al 29 de octubre de 2020. Se promulgó desde la UNED y la solicitud de rellenar el cuestionario se pudo ver en redes sociales y páginas web, para lo cual la CCSS como la UNED, diseñaron cartelería (Figura 1. Cartelería que invitaba a participar en la encuesta sobre salud mental en octubre de 2020).

*Figura 1. Cartelería que invitaba a participar en la encuesta sobre salud mental en octubre de 2020*


*Fuente: Ministerio de Salud (Costa Rica)*

Un año después, en octubre de 2021 el cuestionario se volvió a lanzar al público con nueva cartelería (Figura 2. Cartelería que invitaba a participar en la encuesta sobre salud mental en octubre de 2021), obteniéndose información sobre el estado de salud mental de la población general. Se generaron informes (Carazo-Vargas, Ortega-Moreno, Arias, et al., 2021) que se presentaron a las autoridades, así como ponencias en congresos nacionales e internacionales (Carazo-Vargas, Ortega-Moreno, Villegas Barahona, et al., 2021; González et al., 2021). Por otro lado, con los resultados de las evaluaciones a la población de Costa Rica se

ha podido realizar una evolución y comprender cómo transcurren las psicopatologías en el tiempo y por sectores de población.

*Figura 2. Cartelería que invitaba a participar en la encuesta sobre salud mental en octubre de 2021*


*Fuente: Ministerio de Salud (Costa Rica).*

Los resultados obtenidos se trataron con gran delicadeza, puesto que se trata de una evaluación mediante cribado y no un diagnóstico a los individuos de la población, además algunos de los cuestionarios utilizados originalmente no estaban adaptados a la población costarricense, por lo que hubo que realizar traducciones y adaptaciones en el proceso de generar los cuestionarios. Por otro lado, no existían registros en la bibliografía de adaptaciones de estos test ni tampoco existía bibliografía sobre su adecuación más allá de las culturas a las que fue administrado. Esto implicaba que las conclusiones que se pudieran hacer a partir de estos datos debían tomarse con extrema precaución dado que las escalas no estaban validadas

a la población. De esta forma se evitaban diagnósticos y haciéndose valer de la validez de constructo de los propios test, poder realizar afirmaciones correctas sobre el estado de salud mental de la población.

Después de haberse analizado se comprobó que los test utilizados funcionaban adecuadamente tanto en sus versiones originales como en las adaptaciones a otras culturas e idiomas. Esto se pudo comprobar tras analizar las propiedades psicométricas de los test antes de publicar las conclusiones. Todo ello se hizo para comprobar que se llegaban a conclusiones adecuadas.

En el plan de investigación, entre los objetivos de las investigaciones no figuraba el evaluar las herramientas de medición de psicopatologías, ya que el objetivo principal era la evaluación de la salud mental de la población general costarricense. Esto era porque ya que existía esta necesidad, surgió la oportunidad de evaluar la adecuación de las herramientas de evaluación de la salud mental y ello le dio forma a este trabajo.

# Capítulo 2

# PSICOMETRÍA

*La inteligencia es lo que miden los test de inteligencia.*

Harvard Edwin G. Boring (1886-1968)

En este capítulo se hace una introducción a la psicometría. Medir con precisión constructos psicológicos no es una tarea fácil, ya que carece de forma física y tangencial; por lo tanto se mide de manera indirecta. Los indicadores han de ser lo suficientemente buenos para que midan lo que aseguran que miden, además deben tener precisión para no medir de forma burda. Se tiene que tener en cuenta que ante un mismo estímulo, este puede ser interpretado de forma diferente en función de variables grupales como la cultura o el sexo. Esto es justamente lo que pretende medir la psicometría.

## 2.-Psicometría

Para un adecuado diagnóstico se necesita que un profesional realice una entrevista clínica en la que se recojan los motivos de consulta, los antecedentes personales y familiares, información relativa a otras evaluaciones, el resultado de la interacción entre el profesional y el paciente, así como otros indicios que pudieran aportar información relevante para facilitar un diagnóstico como registro de su conducta o el resultado de pruebas diagnósticas. En este proceso se necesita el valioso tiempo de un profesional, así como el tiempo del paciente. En cambio, si en este proceso se utiliza una prueba que pueda realizar el paciente por sí mismo, el tiempo del profesional sanitario podría enfocarse únicamente en aquellas personas a quienes se les haya detectado el rasgo medido (como por ejemplo depresión, ansiedad o burnout) y no a todas las personas. Esto es el cometido de los test de cribado: servir de tamizaje o de primera valoración como método de selección de aquellos casos en los que sí hay sospechas de que el sujeto presenta el rasgo medido. A partir de la primera detección se puede hacer una valoración más exhaustiva.

Entonces, para poder evaluar a poblaciones o poder hacer evaluaciones individuales, una forma económica, ya sea en tiempo, como en forma de aplicación, son los test. A diferencia de la entrevista, los test dejan preguntas y respuestas cerradas, haciendo que el constructo medido sea completo, exhaustivo y con un resultado fácil de interpretar. Para

homogeneizar y medir constructos mentales, los test psicológicos suelen presentarse en formato escrito; esto es que se trata de presentar a las personas preguntas junto con unas opciones de respuesta para que puedan ser respondidas en función a las instrucciones recibidas.

Básicamente un test es un estímulo estandarizado que puede ser presentado a las personas. En base a sus respuestas puede estudiarse el comportamiento de las personas ante este estímulo. De esta forma puede conocerse la presencia o no, y en su caso el nivel del rasgo que presenta la persona.

Actualmente las herramientas que se utilizan para evaluar aspectos psicopatológicos son la entrevista personal y los inventarios, escalas, test o cuestionarios. Estas herramientas son útiles siempre que midan lo que pretendan medir de forma consistente. En un test de ejecución típica se intenta conocer la diversidad de emociones, valores, referencias y aversiones de las personas; en definitiva los sentimientos, que no pueden ser clasificados en equivocados o acertados. Normalmente, un test de ejecución típica con formato de Likert consiste en una serie de enunciados, los cuales son frases que contienen la actitud o el rasgo a medir, junto con opciones de respuesta que indican la fuerza en que se está de acuerdo o en desacuerdo con la afirmación de poseer ese rasgo.

Los test psicológicos son instrumentos diseñados para medir diferentes aspectos de la vida de las personas. Los resultados de los test ayudan a tomar decisiones como diagnósticos o dictámenes, influyen en los juzgados, describen la evolución de los tratamientos o ayudan a diseñar políticas públicas de salud mediante análisis epidemiológicos.

En términos psicológicos evaluar puede definirse como indagar acerca de la sintomatología. Mientras que diagnosticar ha de hacerse necesariamente de forma personalizada, contemplando varias vías de información como varios test, entrevista personal

u otros indicadores. Ya que no es lo mismo evaluar que diagnosticar, y dadas las características de realizar evaluaciones no individualizadas, cuando se aplica un cuestionario, por definición se está evaluando. En el caso de que se quiera hacer un diagnóstico se deberá contemplar la información que se posee, ya que es posible que aporte más información o incluso que la nueva información obtenida mediante entrevista contradiga la información obtenida mediante cuestionario, haciendo necesaria una mayor indagación para poder realizar un dictamen o diagnóstico psicológico.

El constructo, dimensión, rasgo o cualidad es la entidad que se pretende medir con el test. Por ejemplo, para poder configurar un test que pueda evaluar el constructo de depresión primero se intenta definir qué es depresión y posteriormente intentar contemplar todas las características de la misma; adscribiéndose lo mejor posible al término, idea o constructo. En este sentido la definición de depresión puede ir cambiando con el marco teórico con el que se mida, además de que se pueden llegar a consensos de expertos para concretar el término y sus implicaciones. De esta forma, el concepto de depresión, sus grados y tipologías, así como los criterios para poder diagnosticar a una persona de esta patología, vienen recogidos en el Manual Diagnóstico y Estadístico de los Trastornos Mentales (DMS-5) (Asociación Americana de Psiquiatría, 2014) e incluido en la Clasificación Internacional de Enfermedades (CIE-10) (Ministerio de Sanidad Consumo y Bienestar Social, 2020). Si bien estos manuales se actualizan y con ellos se detallan mejor las características que deben cumplir para concretar lo que se llama depresión, actualmente se ha construido un concepto que le da forma (al pensar en depresión, tenemos cierta idea de lo que es); aunque las actualizaciones vienen encaminadas a una mejor perfilación de lo que es y no es, pero siempre en torno a esta idea o concepto.

Sin embargo para poder aplicar test psicológicos se deben tener en cuenta numerosos detalles que pueden complicar el obtener resultados adecuados. Los factores ambientales pueden influenciar el desempeño de un test: si al momento de aplicarlo existe un gran ruido,

excesiva iluminación y temperatura o el ambiente es incómodo; es posible que se obtenga un resultado diferente si estos factores ambientales incómodos dejasen de estar presentes. De esta forma no se consigue medir lo que se pretende medir. También hay otros factores que pueden influenciar el resultado de un test como una mala explicación de las instrucciones por parte del examinador, el cansancio del examinado o incluso que no se comprenda el propio test por cultura o lenguaje.

Supongamos que queremos conocer el nivel de depresión a una persona. Si aplicamos el test delante de otras personas, pudiendo estas ver las opciones elegidas, es posible que no midamos depresión, sino que los resultados se vean influenciados por la deseabilidad social. Si aplicamos el test de manera verbal y con voz baja, es posible que no midamos depresión, sino que los resultados se vean influenciados por la capacidad auditiva. Si aplicamos el test a una persona nacida en un país diferente es posible que no midamos depresión, sino que los resultados se vean influenciados por la capacidad lingüística o cultural. Si aplicamos el test de depresión a una persona con demencia, quizá no estemos midiendo depresión, sino que los resultados se vean influenciados por la capacidad atencional.

Las formas en que un test pueden fallar pueden provenir de diferentes fuentes y esto incluye el propio test. Es cierto que un test y su aplicación nunca estarán libres de error absoluto, pero se intenta minimizar al máximo las posibles fuentes de error. En definitiva se trata de medir lo que se desea medir con independencia de otros factores que puedan estar influenciando.

Para que un test sea adecuado tiene que medir lo que dice que mide. Además cada uno de los elementos que componen el test deben estar relacionados siempre que pretendan medir lo mismo. Justamente esto es lo que se pretende al analizar las propiedades psicométricas de los test: comprobar que midan lo que aseguran que miden de forma consistente y como dice la teoría que debe ser. Hay muchas formas de comprobar las

propiedades de los test, en la mayoría de ellas interviene la estadística ya que permite analizar cómo los elementos de un test se comportan en la muestra a la que se le ha administrado.

Ya que los test de cribado pueden diseñarse para que se apliquen a grandes cantidades de personas de forma simultánea la estadística ocupa un lugar central tanto en el análisis de los resultados como ante la supervisión de que los test funcionan adecuadamente en la población estudiada. Y en caso de que esto no ocurra, la estadística puede ayudar a mejorar los test.

## 2.1. Fiabilidad y validez

La confiabilidad, (también llamada fiabilidad o grado de acuerdo) puede entenderse como la precisión de una evaluación, o lo que es lo mismo, el grado en el que se encuentra el error aleatorio (resultado de todos los factores aleatorios que pueden darse en una medición). Estos errores pueden producir inconsistencia en la medición y alejar el resultado de la puntuación verdadera (puntuación libre de error). Si una herramienta es precisa se pueden mantener constantes las puntuaciones pese a que sucedan diferentes circunstancias que puedan alterar el error aleatorio.

Tanto la fiabilidad como la validez no explican características del propio test en sí mismo, sino en el comportamiento del test ante una muestra específica y bajo condiciones particulares (American Psychological Association, 1999). La confiabilidad de una prueba aumentará cuanto mayor sea la intercorrelación de sus ítems.

Por ejemplo si una escala está compuesta por los ítems de la fecha de nacimiento, edad, estatura o número de la seguridad social tendrá una muy poca consistencia interna, porque cada uno de los ítems se refiere a diferentes temáticas; pero en cambio al volver a pasar la escala compuesta por estos ítems la fiabilidad debe ser perfecta o casi perfecta. De la

misma forma una escala compuesta por ítems que miden el estado de ánimo tendrá una excelente consistencia interna si tiene varios ítems que miden lo mismo, en cambio tendrá una pobre estabilidad ya que el estado de ánimo es pasajero (McCrae et al., 2011).

Pongamos otro ejemplo: se desea medir la ansiedad de una persona. Para ello se diseña un test de ansiedad con los tres siguientes ítems: “Se me dilatan las pupilas”, “Aumenta mi frecuencia cardiaca” y “mi piel palidece”. Dado que el objetivo es medir ansiedad y uno de los componentes es la reacción fisiológica ante el estímulo ansiógeno, estos ítems podrían medir de alguna forma la respuesta fisiológica. Ocurre que ante la ansiedad se libera una hormona que produce los efectos descritos en los tres ítems, además de otros efectos menos identificables. Entonces la aparición de uno de los efectos fisiológicos implica en mayor o menor medida la aparición de los otros dos, debido a que todos ellos son efectos de la hormona precursora. Si una persona presenta uno de estos efectos debido a la ansiedad, presentará los otros dos en mayor o menor medida. Al analizar las respuestas dadas por las personas es muy probable que quienes hayan respondido a uno los ítems de forma positiva, también lo haya hecho en los demás. Esto implica que el grado de concordancia sea perfecto o casi perfecto. Aunque también implica que la utilidad del test, entendida como validez, se vea reducida porque estaría midiendo lo mismo de diferentes maneras. De esta forma bajaría la validez cuando lo que se desea medir es la ansiedad y se estaría midiendo solo una de las características de la misma o dando más peso a una de las características por sobrerrepresentación de una parte de sus componentes.

En cambio, supongamos otro test de tres ítems, en que el rasgo o sentimiento que se desea medir está lo suficientemente bien recogido en el primer ítem del test, mientras que el segundo y tercer ítem no se recoge información adicional. Dados estos tres ítems, la similitud entre estas tres respuestas dadas, será elevada. Esto es porque los ítems se refieren casi exactamente a lo mismo y por tanto las respuestas dadas serán muy similares. Como uno de

los aspectos que se miden en psicometría es la similitud entre respuestas (el grado de acuerdo entre ellas), al analizar el test, dará elevados índices en fiabilidad. Pero esto no implica que el test sea adecuado, ya que el concepto del rasgo medido no se evalúa de forma completa. Como no se han recogido otros aspectos que definen el rasgo y por lo tanto el concepto de ansiedad no está bien definido a partir de estos tres ítems debido a que le falta indagar acerca de otras características que definen ansiedad. Entonces la fiabilidad aumentará a costa de sacrificar validez, ya que mide muy bien una parte de la ansiedad, pero no consigue evaluar la ansiedad, sino solo una pequeña parte.

Se dice que las personas son demasiado complejas como para obtener una predicción exacta de su rendimiento, así como la forma de haber obtenido su criterio de fiabilidad (Nunnally, 1970). Esto implica que aumentar la confiabilidad tiene un costo, por lo que una mayor confiabilidad no siempre es deseable. Esto se debe a que para aumentar la confiabilidad puede repetirse esencialmente la misma pregunta de diferentes maneras. Esto reduciría la validez, pero aumentaría la fiabilidad.

### 2.1.1. Fiabilidad

#### *2.1.1.1. Dimensiones de fiabilidad*

La fiabilidad de un test es un concepto medible, para ello nos servimos de diferentes técnicas estadísticas que implican el estudio de las respuestas que dio una muestra a un conjunto de ítems. Por esta razón no se estudia el test en sí, sino el funcionamiento del test en la muestra. Esto implica que el test no sea mejor o peor, sino recogiendo mejor o peor el rasgo de la muestra. No todas las técnicas se basan en analizar los mismos datos, sino que para poder aplicar una técnica para conocer la fiabilidad es necesario que los datos hayan sido recogidos de una forma en concreto. Por ello se han desarrollado técnicas estadísticas como de diseño de la investigación para conocer el comportamiento del test en la/s muestra/s.

*Tabla 2. Clasificación de las dimensiones de fiabilidad*

| Dimensión de la confiabilidad | Método | Estadístico |
|---|---|---|
| Estabilidad | Test-retest | r de Pearson (puntuación 1 vs puntuación 2. Mismo sujeto) |
| | Formas paralelas | r de Pearson (forma A vs forma B. Mismo sujeto) |
| Consistencia interna | Formas paralelas | r de Pearson (forma A vs forma B. Sujetos diferentes) |
| | Partición en dos mitades | r y su fórmula de corrección, Spearman-Brown. 1 aplicación. |
| | Coeficiente alfa | Alpha, Kuder-Richardson. 1 aplicación. |
| Confiabilidad entre examinadores | Acuerdo entre examinadores | Kappa, w de Kendall, CCI. 1 aplicación. |

Tabla de elaboración propia.

Estas dimensiones de la fiabilidad se desarrollarán a continuación.

#### *2.1.1.2. Estabilidad*

Es deseable que un test pueda funcionar de la misma forma o de forma muy similar, independientemente de las circunstancias de su aplicación, siempre que el constructo que se desee medir fuese estable. Por ejemplo: si se desea conocer el grado de felicidad que se tiene en estos precisos instantes, es comprensible y esperable que este grado de felicidad fluctúe con el paso del tiempo y en función de los acontecimientos acaecidos, así el grado de felicidad se entiende como un constructo inestable. En cambio, ante evaluaciones de cambios de estado de ánimo no resulta interesante atender a la estabilidad temporal de la prueba porque lo esperable es que exista una inconsistencia: que se den modificaciones en el estado de ánimo. Para poder comprobar la estabilidad se utiliza el método del test-retest y el método de formas equivalentes; ambas formas son aplicadas tras un intervalo de tiempo (American Psychological Association, 1999).

#### 2.1.1.2.1. Test-retest

Mediante esta técnica se estima la capacidad de la herramienta de medición para diferenciar entre sujetos cuando han sido medidos dos veces y bajo las mismas condiciones (Berchtold, 2016).

Parte de la base de que un test es igual a sí mismo, entonces se compara el propio test consigo mismo. Por lo tanto se exige que el test sea administrado en dos ocasiones diferentes y a la misma muestra. Este método de obtención de la fiabilidad se interpreta como la estabilidad temporal de las puntuaciones. Así puede tener los peligros con respecto a la demora en su repetición: que pase poco tiempo y que pase mucho tiempo. Si el test ha sido administrado con poco tiempo entre evaluaciones puede aparecer el fenómeno de que el sujeto pueda recordarlo lo suficiente y además puede que quiera ser consistente con las respuestas en lugar que dio en un primero lugar, en lugar de ser genuinas; por lo que no solo estaría describiéndose el sujeto, sino intentando, a su vez ser consistente con sus respuestas previas. En cambio, si el test ha sido administrado con mucho tiempo entre administraciones, los cambios de las puntuaciones de los sujetos pueden verse influenciados no por el comportamiento del test, sino porque el constructo que se pretende medir realmente haya cambiado. En cualquiera de estas dos opciones no se estaría evaluando lo que el test pretende evaluar, que es el constructo o rasgo.

Cada sujeto de la muestra debe tener dos registros del test (Koo & Li, 2016): uno anterior y otro posterior. Al realizar un análisis de test-retest, el Coeficiente de Correlación Intraclase (ICC, Intraclass Correlation Coefficient) es analizado. Este coeficiente únicamente puede ser utilizado para datos continuos. Indica el grado de concordancia entre observaciones. El CCI puede distinguir el grado de acuerdo absoluto $CCI_A$ y el grado de consistencia $CCI_C$. El primero ($CCI_A$) se refiere a cualquier tipo de diferencia entre las medidas, sin considerar si estas diferencias son constantes, proporcionales u otros; de tal forma que cuanto haya más

diferencia entre medidas, el índice será más bajo. El segundo ($CCI_C$) se refiere a la correlación de una forma aditiva (con una diferencia sistemática).

Existen alternativas para que el método test-retest pueda ser analizado siendo sus datos binarios o categóricos (Crespi et al., 2011; Liljequist et al., 2019).

Para interpretar estos índices puede tomarse como referencia a Fleiss (Fleiss, 1999). Siendo x el coeficiente queda lo siguiente. x=0 ausencia de concordancia. x<0,40 concordancia baja. 0,41<x<0,75 concordancia regular o buena. x>0,75 concordancia muy buena. x=1 concordancia absoluta. Aunque también puede tomarse como referencia a Koo (Koo & Li, 2016), siendo x el coeficiente, quedaría como sigue: x<0,5 pobre. 0,5<x<0,75 moderado. 0,75<x<0,9 bueno. 0,9<x excelente.

#### 2.1.1.2.2. Formas equivalentes

También llamado formas paralelas. Un test sería paralelo o equivalente si mide el mismo constructo que otro test. Este método parte de generar dos test que midan el mismo constructo, con similar dificultad y estructura; de tal forma que la correlación de un test y el segundo test den un indicativo de su confiabilidad. Puede considerarse confiabilidad o estabilidad temporal siempre que las administraciones de los test hayan sido aplicadas tras un trascurso de tiempo. Los resultados obtenidos en un test deben ser comparables en diferentes momentos para poder realizar conclusiones válidas sobre el cambio del individuo. La estabilidad temporal se refiere a la consistencia o grado de acuerdo de los resultados obtenidos al administrar un test. Dicho de otra forma: la capacidad de un test para medir lo mismo en diferentes momentos. Un test tiene estabilidad cuando produce resultados consistentes a lo largo del tiempo.

#### *2.1.1.2.3. Consistencia interna*

Refleja la coherencia o redundancia de los componentes de la herramienta de medición. Esta medida se utiliza para conocer la fuente de error de la medición debida a la elección de la muestra de ítems utilizados en el test. Esta forma de confiabilidad hace referencia al grado en que distintas partes del test miden el mismo constructo o dominio. Para evaluar la consistencia interna se usan los métodos de formas equivalentes (cuando se compara el test en el mismo espacio temporal), el método de las dos mitades y el método del coeficiente de Alpha de Cronbach.

El peligro de esta técnica radica en que el solo fijarse en una consistencia interna elevada como indicador de buena consistencia también denota alta redundancia entre ítems o lo que es lo mismo: poca variabilidad de contenido. A esto se le conoce poco la paradoja de la atenuación de la teoría clásica de los test y está señalada como no deseable los ítems equivalentes en los test (Loevinger, 1954).

Considere que todos los ítems de un test son equivalentes, miden el mismo rasgo, tienen la misma fiabilidad y validez, así como el mismo grado de dificultad. En este caso los ítems tendrán la misma intercorrelación con iguales coeficientes de fiabilidad. Si la fiabilidad aumenta a 1, todas las correlaciones entre ítems también serán 1 y la persona que conteste a un ítem es como si contestara a todos los ítems por igual.

De aquí se infiere que, si un ítem está lo suficientemente bien construido, en aras de reducir la longitud de los test al mínimo, la consistencia interna no aportaría mayor información, porque si lo que se pretende conocer es si un ítem mide determinado constructo por medio de la consistencia interna, si ese ítem está constatado que mide este constructo, no sería necesario añadir ítems paralelos o aumentar la extensión de ítems que midan el mismo constructo, porque ya está constatado que se mide adecuadamente.

Entonces, además de fijarse en la consistencia interna, también es deseable explorar la construcción de los ítems buscando.

Supongamos que queremos conocer cómo se ha recaudado financiación para costear un proyecto de investigación. Este proyecto admite subvenciones de diferentes fuentes. Supongamos que una institución gubernamental aporta un porcentaje de un 2,0% al proyecto y hay que poner su sello y agradecer su aporte. Este es un porcentaje bajo que explica poco la fuente de financiación, pero sabemos exactamente el monto, la institución y cualesquiera otros detalles. Con la salvedad de que al hablar de probabilidad se habla con un grado de certidumbre, de forma similar, cuando afirmamos que existe una correlación altamente significativa es que sabemos con cierta seguridad la fuente del aporte, aunque este aporte sea escaso. Así se comprenderán mejor las correlaciones: puede que los ítems 4 y 9 tengan correlaciones bajas, en cambio estas son altamente significativas.

#### 2.1.1.2.3.1. Partición en dos mitades

Puede administrarse un test en una ocasión a una muestra de sujetos. Este test puede partirse a posteriori en dos mitades, de tal forma que un mismo sujeto puede tener dos puntuaciones en cada mitad del test. La correlación de ambas mitades se tomaría como un indicador de estabilidad.

#### 2.1.1.2.3.2. Correlación ítem-total

También llamado índice de homogeneidad mide si un ítem evalúa lo mismo que el resto de ítems, entendido como el conjunto de ellos. O lo que es lo mismo: si un ítem puntúa alto, cuánto puntuará el resto de ítems del test. Si un ítem puntúa alto en este índice, se infiere que mide lo mismo que en la escala; de la misma manera que si un ítem puntúa bajo, se infiere

que no mide lo mismo que el resto de la escala. Se establece como valores menores de 0,2 que los ítems no estarían midiendo lo mismo que el resto de la escala y son indicaciones de buen funcionamiento valores mayores de 0,3 en esta puntuación.

#### 2.1.1.2.3.3. Método de las dos mitades

Debido a que no es necesario generar más ítems de un mismo test ni tampoco es necesario pasar el test en varias ocasiones, satisface el principio de parsimonia mejor que otros métodos. Esto hace que sea un procedimiento predilecto por su simplicidad. Se asume que en un test hay suficientes ítems como para que, al seleccionar la mitad de sus ítems, ambas mitades sigan midiendo el mismo constructo. El coeficiente de fiabilidad se estima por la correlación de ambas mitades. Entonces indica la consistencia interna. Este método puede estar afectado por la selección de los ítems que componen cada mitad. Comúnmente se seleccionan los ítems pares para conformar la primera mitad y establecer la correlación con la segunda mitad compuesta por los ítems impares.

#### 2.1.1.2.3.4. Coeficiente de Alpha de Cronbach

Mediante este coeficiente, Lee (Cronbach, 1951) expresa la fiabilidad del test en función del número de ítems y de la proporción de la varianza total que viene explicada por la covarianza de los ítems. De esta forma, cuanto mayor covarianza inter-ítems, mayor será este indicador de consistencia interna.

$$\alpha = \frac{n}{n-1}(1 - \frac{\sum S_j^2}{S_x^2})$$

Donde:

$n$ es el número de ítems.

$\sum S_j^2$ es el sumatorio de las varianzas de los ítems.

$S_x^2$es la varianza total de las puntuaciones del test.

#### 2.1.1.2.3.5. KR20

Cuando los ítems son dicotómicos se presenta este estimador (Kuder & Richardson, 1937) en el que se sustituye el sumatorio de varianzas de los ítems $\sum S_j^2$ por la suma de la varianza de cada ítem $\sum p_h q_h$, entendida como proporción de aciertos y fallos. Al igual que con el Alpha de Cronbach, cuanto mayor sea el número de ítems, mayor será su fiabilidad. En esta expresión la detección del rasgo se representa con un 1 y la no detección del rasgo como un 0.

$$KR_{20} = \frac{n}{n-1}(1 - \frac{\sum p_h q_h}{S_x^2})$$

Donde:

$n$ es el número de elementos del test.

$p_h$es la proporción de aciertos en el elemento $h$.

$q_h$es la proporción de errores en el elemento $h$. Así $p_h q_h$es la varianza del ítem $h$.

$S_x^2$es la varianza total del test.

#### *2.1.1.2.4. Confiabilidad inter-examinadores*

Se utiliza para comprobar en qué grado la medición de un rasgo medido en un individuo puede variar en función de la subjetividad del examinador. Lo deseable es que el sujeto obtenga puntuaciones idénticas con independencia de qué examinador le esté realizando la prueba. Para comprobar esta forma de confiabilidad se utiliza el método de acuerdo interjueces.

##### 2.1.1.2.4.1. Acuerdo interjueces

Consiste en administrar una herramienta de evaluación una sola vez y posteriormente un comité de expertos, llamados jueces, serán quienes puntúen cada uno de forma independiente al resto. Posteriormente se comparan los resultados de los diferentes jueces para establecer el grado de acuerdo entre ellos. Este procedimiento es útil para, por ejemplo, entrevistas semiestructuradas, ya que existe un fuerte criterio subjetivo del evaluador, quien debe estar formado en la administración de la prueba. En cambio, carece de sentido ante pruebas que puntúan de manera objetiva.

Puede medirse con el Coeficiente de Correlación Intraclase (CCI) el cual se ha aceptado como índice de concordancia para datos continuos y ordinales. También pueden usarse el índice de Kappa para datos nominales.

### 2.1.1.3. Validez

#### *2.1.1.3.1. Dimensiones de validez*

Persigue verificar que lo que pretende medir el test es lo realmente medido por éste. De esta forma, las inferencias que puedan darse a partir de las puntuaciones de un test realmente reflejan la puntuación en ese constructo.

Pueden diferenciarse tres tipos de validez: de constructo, de contenido y predictiva. Aunque realmente puede entenderse como la evidencia que concretiza un constructo.

#### 2.1.1.3.1.1. Fuentes internas de validez

##### 2.1.1.3.1.1.1. Contenido del test

Este tipo de evidencia consiste en demostrar que los ítems representan un constructo, para lo cual es posible reunir evidencia de que, efectivamente los ítems miden este constructo mediante un comité de expertos que evalúen el grado en que el contenido del test es relevante y representativo del constructo.

##### 2.1.1.3.1.1.2. Estructura interna del test

El Análisis Factorial fue desarrollado para identificar constructos psicológicos y es un buen indicador para conocer la estructura interna de un test (López-Aguado & Gutiérrez-Provecho, 2019). Mide las intercorrelaciones entre datos observables, de tal forma que, si ciertos grupos de ítems se correlacionan más entre sí mismos, pueden agruparse definiendo así constructos. Es destacable que el supuesto de normalidad del Análisis Factorial Confirmatorio se puede incumplir en las escalas tipo Likert (Flora & Curran, 2004).

Se ha elaborado un decálogo (Ferrando et al., 2022) para ayudar a realizar un Análisis Factorial de forma más sistemática y rigurosa, reduciendo así las fuentes de error. Puede resumirse en lo siguiente:

- Adecuación de los datos y de la muestra. Se refiere a la obtención de datos con una muestra adecuada y suficiente.
- Cálculo de los Estadísticos descriptivos univariados. Es importante este cálculo, ya que informará de cómo se comportan los ítems en la muestra. Puede

informar de ítems que no funcionan adecuadamente al ser constantes (sin variabilidad en sus opciones de respuesta) o también si no se han comprendido bien las instrucciones o cualesquiera otra circunstancia extraña.

- Justificación del análisis. Se refiere al análisis factorial, ya que para que pueda darse necesita que existan intercorrelaciones entre el conjunto de los ítems que se quieren estudiar. Por ello se debe estudiar la matriz de intercorrelaciones, así como índices que explican la comunalidad de las relaciones entre los ítems. Si se utiliza el índice KMO (Kaiser & Rice, 1974), se sugiere que dé un mínimo de 0,75 para poder interpretar el Análisis Factorial tenga sentido y cuanto mayor sea este índice implica que los ítems se refieran al mismo constructo.

- Selección de los ítems analizables. No todos los ítems, del que es *a priori* el mismo grupo, deben incluirse siempre en un Análisis Factorial. Y esto es que, aunque inicialmente el conjunto global de ítems pueda explicar un mismo constructo, al analizarse uno a uno con algún otro índice, como el MSA (Measure of Sampling Adequacy) (Lorenzo-Seva & Ferrando, 2021; Meyer et al., 1977; Shirkey & Dziuban, 1976) y es posible que de forma aislada no aporte lo suficiente al constructo, como para valorar el permanecer entre los ítems seleccionados. Este índice sugiere la eliminación de cada uno de los ítems siempre que no supere el índice sugerido de 0,5.

- Decidir el tipo de modelo factorial. Se trata de seleccionar entre un modelo lineal y uno no lineal.

    - Para el modelo lineal se toma el supuesto de que las puntuaciones de los ítems son continuas e ilimitadas. También se asume lo mismo para los niveles de los factores. Esto hace que las regresiones ítem-factor sean lineales y la matriz de correlaciones inter-ítem sea una matriz de correlaciones de Pearson.

- Para el modelo no lineal las puntuaciones de los ítems ya no son continuas, sino que son discretas y limitadas. Las regresiones entre los ítems y los factores no describen una línea, sino que describen una ojiva (figura formada por dos arcos de círculo iguales, que se cortan en un extremo y presentan una concavidad enfrentada) o describen una S. Esto hace que la matriz de correlaciones inter-ítem tenga una solución de correlaciones policóricas en el caso de opción múltiple y tetracóricas en el caso de opciones de respuesta binaria (Hoffmann et al., 2013).

- Elegir la solución factorial más adecuada. Dentro del Análisis Factorial conviene distinguir entre el Análisis Factorial Confirmatorio y el Exploratorio. Aunque es considerado como dos extremos de un mismo conjunto continuo (Ferrando et al., 2022).

  - El Análisis Factorial Exploratorio no tiene suficientes restricciones como para ser única y de una solución inicial, generalmente se transforma mediante rotación para ser más fácilmente interpretada. Es importante aclarar que ante un conjunto de un solo factor, la rotación no es útil al saturar todos los ítems en un solo factor. Esto implica que al no existir posibles cambios debidos a la rotación, la solución del Exploratorio y del Confirmatorio sería exactamente la misma (Ferrando et al., 2022).

  - El Análisis Factorial Confirmatorio, por su parte, no necesita de rotación debido a que es lo suficientemente restrictiva y por ello su solución inicial también es su solución final.

- Adecuación de la solución factorial. Pese a que la teoría implique la dependencia o independencia, conviene poner a prueba esta hipótesis para no restringir las posibilidades e imponer una solución cuando no es correcta (Lorenzo-Seva & Ferrando, 2020). Al realizar la rotación existen dos principales soluciones:

- Solución rotada ortogonal. Se utiliza en factores independientes.
- Solución rotada oblicua. Se utiliza para factores dependientes. Sean los factores dependientes o independientes entre sí se recomienda la solución oblicua (Browne, 1972).

- Estimación de los parámetros. Aunque en principio tanto el Exploratorio como el Confirmatorio pueden estimarse gracias a los mismos parámetros, debe tenerse en cuenta de que el Confirmatorio se estima en una etapa, mientras que el explotatorio en dos etapas. Las formas para hacer estas estimaciones:
  - Mínimos Cuadrados no Ponderados (Unweigthed Least Squares, ULS). Consiste en un método simple, fiable y computacionalmente eficiente que funciona bien tanto para el modelo lineal como el no lineal (Krijnen, 1996).
  - Máxima verosimilitud (Maximum Likelihood, ML). Se trata de un modelo sólido, pero restringido para el modelo lineal (Krijnen, 1996).
  - Mínimos cuadrados ponderados diagonalmente (Diagonally Weighted Least Squares, DWLS). Constituye el procedimiento más frecuentemente utilizado para soluciones de Análisis Factorial Confirmatorio (Muthén, 1993).

- Evaluar la adecuación de la solución factorial. Se trata de comprobar si los datos obtenidos tienen sentido. Para ello se debe evaluar el grado de ajuste de los datos, la claridad, fuerza y grado de determinación del tipo de solución obtenida, así como la calidad y precisión de las puntuaciones resultantes de la solución. Un índice a tener en cuenta es la raíz Media Cuadrática Residual (Root Mean Square of Residuals, RMSR) (Harman, 1976), el Goodness of Fit Index (McDonald & Mok, 1995) o el Root Mean Square Error of Approximation (RMSEA) (Tennant & Pallant, 2012). Por otro

lado, la hipótesis que evalúa el Análisis Factorial es la dimensionalidad, para lo que se utilizan los autovalores o el análisis paralelo (Lorenzo-Seva et al., 2011). En soluciones multidimensionales simples, claras e interpretables, lo ideal sería que cada ítem saturase únicamente en un único factor, aunque justamente esto es poco común. El peso saliente indica que el ítem evalúa principalmente este factor. Los pesos otorgados a otros factores indican que efectivamente ese ítem explica ese factor, aunque en menor medida o de manera residual.

- Versión final del test. Se realiza a través de la selección de los ítems que conforman la versión final del test. Esto es, considerando los ítems que se aconsejan retirar, aquellos que saturan adecuadamente en un factor, aquellos con valores contantes, así como las agrupaciones suficientes para explicar un factor, puede refinarse el conjunto de ítems. Por otro lado, configurar un test de forma ideal que permita una mayor discriminación de los sujetos, tendría ítems cuyos índices de dificultad fuesen distribuidos en un 75% por valores comprendidos entre 0,4 y 0,6. Y en un 12,5% superior y otro inferior a índices de dificultad menores y mayores a los índices propuestos de 0,4 y 0,6. Aquellos ítems indeseables en un test son los que ofrecen información redundante: que pese a que tengan unos buenos indicadores estadísticos, la forma o el contenido del ítem se asemejan tanto que conceptualmente son indiferenciables o con escasas diferencias. Da lugar a un mal ajuste y distorsiona la solución factorial. Los ítems que aportan ruido tienen pesos bajos en las saturaciones de los factores, ya sea porque se responden al azar, porque son ambiguos o incomprensibles, estos ítems son detectados mediante el indicador MSA, por ejemplo. Por otro lado, los ítems complejos son aquellos que saturan de forma similar en más de un factor a la vez, complicando la interpretabilidad y definición del ítem a un factor. En cambio, los ítems buscados son aquellos que saturan fuertemente en un solo factor

y son únicos (no tienen duplicidades o redundancia), son los llamados marcadores (Cattell, 1988; Eysenck, 1952).

Como aclaraciones al Análisis Factorial y al decálogo, obtiene importancia el hecho de que el ideal inalcanzable devolviese una solución única y unifactorial. En cambio la realidad de los datos hace que un índice de asociación entre los ítems sea inevitable, incluso entre factores diferentes, ya que la naturaleza de los ítems no es pura, sino que entraña complejidad (Cattell, 1988). Cuando se realiza el análisis de un test con una estructura ya conocida, sus factores deben ser muy específicos y su longitud debe ser corta para limitarse a realizar un Análisis Factorial Confirmatorio. Además, aunque pueda parecer un proceso lineal, este es más bien cíclico, al existir la necesidad de rehacerse cuando uno de las modificaciones es realizada, para asegurar que esta modificación no influya en el conjunto. Además, las modificaciones realizadas tienen diferentes niveles de exigencia. Por otro lado, el repetir los pasos no se limitaría a analizar los datos, sino a seleccionar y adecuar la muestra, así como verificar la estructura del test.

#### 2.1.1.3.1.1.3. Proceso de respuesta

Es un método poco investigado que consiste en entrevistar a los individuos para conocer los procesos internos que acontecen para la emisión de una respuesta: examinar las diferentes estrategias que pueden ser utilizadas para así enriquecer la comprensión del constructo medido. Esto puede dar valor a identificar diferencias entre grupos, más allá de las propias respuestas, el proceso para llegar a ellas.

#### 2.1.1.3.1.2. Fuentes externas de evidencia

Para conocer las fuentes externas de evidencia o de validez, tiene su base en estudiar la relación del test con otras variables externas al mismo. Por ejemplo las medidas de algún criterio que pretende medir el propio test; las puntuaciones de test que midan el mismo constructo; las puntuaciones de test que midan constructos diferentes, pero de alguna forma relacionados.

##### 2.1.1.3.1.2.1. Evidencia convergente-divergente

Un test debe correlacionar de forma más elevada con un test que mida el mismo constructo y de igual forma debe correlacionar de forma más baja con otro test que mida un constructo diferente.

##### 2.1.1.3.1.2.2. Relaciones entre las puntuaciones de un test y criterios externos

Se basa en que, si el test que pretende medir un constructo y se tiene el conocimiento de que el sujeto está previamente diagnosticado o presenta un rasgo inequívoco de la presencia de ese constructo, el test medirá adecuadamente si logra detectar la presencia del constructo y esto puede comprobarse con este criterio externo.

En el caso en que, el criterio se presente como una variable continua puede usarse el coeficiente de correlación de Pearson. Si se trata de una variable dicotómica con una continua puede utilizarse el coeficiente de correlación biserial-puntual. De igual forma puede usarse Spearman en el caso de que exista alguna variable ordinal.

#### 2.1.1.4. Modelo de Rasch

La Teoría de Respuesta al Ítem (TRI) es una familia de modelos psicométricos utilizados con el fin de evaluar la habilidad de los individuos en un constructo. En la TRI se pueden distinguir cuatro modelos distintos en función del número de parámetros que se estiman:

- Modelo de un parámetro, también llamado modelo de Rasch (Aryadoust et al., 2021). Indica la dificultad del ítem. Se marca con la letra "b". Expresa cuán difícil sería para un sujeto "acertar" el ítem. Aplicado a sintomatología indicaría lo difícil que le resulta al sujeto ser detectado en el constructo medido. En una gráfica de la Curva Característica del Ítem (CCI) puede representarse el parámetro b como la posición de la curva en el plano horizontal; de tal modo que si el eje y representa la probabilidad de acertar el ítem y el x la dificultad del ítem, el parámetro b describiría la posición en el eje x de la curva.
- Modelo de dos parámetros (Muraki, 1992). Se estima la dificultad del ítem "b" y el índice de discriminación del ítem "a". El parámetro "a" indica la capacidad de un ítem de poder clasificar a los sujetos en los diferentes niveles del ítem. En la CCI describiría la inclinación de la curva, cuyo aplanamiento implicaría menor discriminación del ítem y una mayor pendiente implica una mayor discriminación.
- Modelo de tres parámetros, o modelo de Birnbaum (van den Brink, 1982). A los dos parámetros anteriores se le añade el parámetro "c" de pseudoazar o adivinación. Indica la probabilidad de "acertar" el ítem, solo interpretable para ítems de pruebas en que, una opción de respuesta es cierta y el resto falsas. En los exámenes, para eliminar el parámetro azar se ideó la siguiente fórmula:

$$Nota=A - E/(k-1)$$

Donde:

*A* representa el número de aciertos.
*E*: representa el número de errores.
*k*: número de alternativas de cada pregunta.

- El modelo de 4 parámetros (Thurstone, 1927) indica el grado de diferencia entre las respuestas, una medida de variabilidad. A mayor valor, más distancia entre mediciones. Está expresado en desviaciones típicas.

## 2.2. Funcionamiento Diferencial

### 2.2.1. Influencia de la cultura en los test

En 1951 se llevó a cabo la primera investigación sobre el sesgo de los ítems (Eells et al., 1951), que por su familiaridad para ciertos grupos socioeconómicos estarían reflejando no la variable que se pretende medir, sino las oportunidades para aprender de estos grupos. Esto se llama sesgo cultural.

En 1969 se publica el artículo de Jensen defendiendo que el CI está determinado genéticamente, que las diferencias raciales importan tanto como las diferencias genéticas entre individuos y que la educación compensatoria no es una forma adecuada de abordar el problema de diferencias en la educación otorgada (Jensen, 1969).

Dar aspecto de cientificidad a elementos racistas sin haber realizado pruebas pertinentes implica caminar por terrenos farragosos. Uno de los peligros de asumir que las pruebas funcionan igual de bien para cualesquiera grupos, puede implicar sesgos importantes. Cuando una explicación alternativa pudiera ser que la herramienta o prueba que se realiza para evaluar la competencia entre los grupos no sea lo suficientemente eficiente como para evaluar el desempeño de cada grupo de forma adecuada, sino adaptada a cada necesidad que tuviese cada grupo. En las sociedades cazadoras, la inteligencia bien pudiera haberse medido como la capacidad de recoger alimento con escasos recursos; mientras que hoy en día la inteligencia se estima como capacidades de resolución de problemas lingüísticos, matemáticos y espaciales.

En sociedades meritocráticas prima la competencia del individuo frente a si pertenece o no a un grupo aventajado. Los test psicométricos persiguen conocer a los sujetos sin que se vean influenciados por el grupo al que pertenezcan. Con independencia de la naturaleza de este grupo, como el nivel socioeconómico u otros como la edad, el sexo, el nivel educativo o la procedencia. Esto obliga a que los test pretendan ser insesgados, objetivos, pretendiendo la

igualdad de oportunidades que evalúen adecuadamente a los sujetos con independencia de los grupos a los que se adscriban. De tal forma que el test solo mida constructos de las personas y no esté influenciado por el grupo al que pertenezca el sujeto. O dicho de otro modo: que no discrimine a los grupos poblacionales, sino a las personas en sus constructos medidos. Para evitar que un test favorezca o sea más fácil detectar rasgos en miembros de un grupo sobre otras personas, y que las puntuaciones obtenidas sean fruto del nivel de los sujetos y no de otras características, se proponen correcciones a los test para facilitar o dificultar la detección del rasgo medido.

Un test está sesgado cuando individuos con un mismo nivel de un rasgo o constructo presentan diferentes probabilidades de que sea detectado este rasgo debido al grupo al que pertenecen (grupos sociales, culturales, étnicos…).

Los test solo muestran diferencias reales en la habilidad o rasgo medido, pero en el caso en que, esta habilidad o rasgo medido esté influenciada por las características inherentes al grupo de pertenencia, se deberán identificar estas características y homogeneizar el test para que no esté influenciado por el grupo de pertenencia.

Pongamos el ejemplo de la modestia. Con el objetivo de medir el rendimiento académico, se preguntó a una serie de estudiantes sobre su desempeño en las áreas de lectura y matemáticas (Tanzer, 1995). Los estudiantes procedían tanto de Singapur como de Australia. La prueba administrada presentaba la misma estructura factorial en ambos grupos, pero algunos componentes de la prueba mostraban un funcionamiento diferente. Se conjeturó que era debido a una variable extraña que originalmente no estaba contemplada en el estudio: el factor modestia. Y es que, comparando con sus notas tanto en lectura como en matemáticas, ante igual nivel de competencia, los estudiantes manifestaban que su rendimiento académico era diferente. Esto es porque la autopercepción recoge connotaciones culturales, ya que en la cultura asiática la modestia es una virtud deseable; entonces los estudiantes asiáticos

reflejaban una mayor resistencia a mostrarse con una mayor seguridad en sí mismos. Este fenómeno reducía la puntuación de un grupo frente a otro, pese a tener exactamente el mismo nivel de competencia.

Un ejemplo de diferencias culturales podría ser una hipotética investigación que midiese capacidad semántica. Imaginemos que se presenta una imagen de un copo de nieve. Sería más probable que ante un copo de nieve, en culturas acostumbradas a la nieve, se tenga más sustantivos dedicados a esta entidad. En contraposición a culturas cuyo contacto con la nieve sea nulo debido a su ubicación por latitud o condiciones climatológicas. Esto es debido a la familiaridad del estímulo presentado. En el caso de querer medir capacidad semántica, se estaría cometiendo un sesgo cultural al intentar medir esta capacidad mediante la designación de entidades cuyo peso cultural sea diferencial. Esto haría que la herramienta utilizada no sea efectiva para medir el rasgo que se intenta medir. Por lo tanto, a pesar de que se estuviera utilizando la misma prueba, integrantes de la cultura Inuit tendrían más facilidad para emitir semántica adecuada que integrantes de la cultura saharaui.

En el libro Introducción a la Psicometría (Tornimnbeni et al., 2008) ofrecen otro ejemplo sobre diferencias entre los grupos y la capacidad de detección del rasgo en un test. En un test de depresión, el ítem “Lloro fácilmente” aunque está enfocado a detectar depresión y es un elemento propio de la sintomatología depresiva, la forma de expresar depresión no es igual en todas las personas. Si bien en la mayoría de las culturas los hombres están condicionados culturalmente a no expresar tanto sus emociones como las mujeres, la característica de llorar podría medirse igual con independencia del sexo: detectaría bien si varones y mujeres lloran. Pero si llorar es un medio para detectar depresión no sería tan útil este ítem para detectar depresión, ya que la depresión puede estar presente mientras que el llanto ofrece menos probabilidad de presentarse en los varones. Esto dejaría que solo contestaran positivamente a este ítem aquellos varones cuyo rasgo depresivo fuese mucho

más elevado que el resto. Además, en el caso de no compensar este ítem, y el test estuviera compuesto por ítems similares de detección culturalmente sesgada del rasgo depresivo, se estaría supradiagnosticando de depresión a la mujer e infradiagnosticando de depresión al varón; pese a poseer exactamente el mismo nivel de depresión en ambos grupos.

En definitiva el test intenta medir el rasgo de las personas con independencia de influencias culturales. No obstante si estos desacuerdos no son culturales, sino que responde a explicaciones biologicistas no estaría favoreciendo sistemáticamente a un grupo frente a otro, sino que el rasgo medido sería el que verdaderamente se encuentra presente. Es importante señalar que el "impacto del test" mide sin sesgos culturales o de otro tipo, sino que realmente mide de forma adecuada el rasgo. Imaginemos las pruebas que se realizan a los candidatos de los Cuerpos de Seguridad del Estado, en concreto las pruebas físicas de la Policía Nacional en España. Una prueba concreta es la de la carrera de 1.000 metros. En el caso en que las desigualdades en cuanto al desempeño de la carrera entre varones y mujeres respondieran a disparidades culturales, podría presentarse unos baremos diferenciados para poder ajustar estas disimilitudes culturales. No obstante, si las disparidades corresponden a diferencias reales en el desempeño de la carrera y no a un error a la hora de registrar los tiempos o evaluar el desempeño de la carrera, se estaría hablando de "impacto del test". El impacto del test hace alusión a que el constructo medido se hace de forma adecuada: se mide sin sesgo cultural el desempeño real de las personas. Esta ausencia de sesgo de medición implica que las diferencias que pudieran establecerse entre sexos no estarían atendiendo a reducir las influencias culturales propias de los grupos, y por tanto al carecer de razones culturales que impliquen una dispar medición entre grupos, no quedaría justificada una baremación diferencial por razón de sexo para reducir las diferencias culturales de la herramienta utilizada para medir.

De lo que se trata es de que la herramienta utilizada para medir pueda medir de forma adecuada, sin influencias culturales. Ya que se dice que el sesgo de medición existe cuando las diferencias individuales en las puntuaciones de un test no reflejan las diferencias reales en un rasgo o habilidad.

En el caso de que se detecten estas diferencias culturales que impliquen que los grupos tengan un desempeño diferente o que el rasgo medido sea más o menos difícil de detectar en función del grupo influenciado por la cultura, se deberán aplicar correcciones a la herramienta con la que se mide, para que pueda medir realmente el rasgo sin influencias culturales.

En el caso de poder corregir el sesgo diferencial, al presentar el test se estaría garantizando que la detección del rasgo se realice sin discriminación alguna por motivos de pertenencia a grupos como el sexo, género, origen racial o étnico, nacionalidad, religión o creencias, salud, edad, clase social, orientación sexual, identidad sexual, discapacidad, estado civil, migración o situación administrativa, o cualquier otra condición o circunstancia personal o social. En caso de no poder corregir este sesgo se estaría violando el supuesto de equidad cultural que deben cumplir los test

En psicología la existencia de un sesgo de medición en un test puede implicar resultados gravemente erróneos infiriendo la presencia de un trastorno cuando realmente no se encuentra presente.

Estas modificaciones a los test conllevan implicaciones éticas y a nivel social y jurídico, ya que pueden infravalorar sistemáticamente la detección del rasgo y dificultar por sistema que un grupo poblacional, por el mero hecho de pertenecer a ese grupo poblacional, no quede adecuadamente reflejado su nivel de rasgo por el test.

Con la finalidad de conocer las actitudes de las personas con independencia de que su grupo de pertenencia esté más o menos familiarizado con el constructo medido o que su

grupo tenga características que lo confieran como aventajado y por tanto sea más fácil para ese grupo contestar a este test y no por las características de la persona, sino por el mero hecho de pertenecer a ese grupo; se realizan las correcciones a los test. Por ejemplo una población que por sus circunstancias no esté acostumbrada a hacer una valoración de sus propios sentimientos, le resultará más difícil conocer la forma en que se expresa, los indicativos o expresión de sus emociones; y tendrá unos resultados diferentes pese a que teóricamente el constructo de depresión sea intrínsecamente igual que en una población de comparación.

En el caso en que un cuestionario se haya diseñado bajo unas circunstancias y adaptado a determinada población, es posible que no funcione de forma similar en otras poblaciones o que los constructos no sean los mismos o existan alteraciones tales que al funcionar de forma diferente sea necesaria alguna modificación o incluso descartar o desaconsejar el uso de ese cuestionario para esa población. Normalmente si un cuestionario tiene un constructo robusto solo son necesarias algunas modificaciones para poder aplicar el cuestionario a diferentes poblaciones a las que originalmente se diseñó. Las modificaciones pueden ser de contenido con traducciones o adaptación de términos o expresiones, o psicométricas como conocer cómo funcionan los ítems en función del subgrupo poblacional (a esto se le llama funcionamiento diferencial de los ítems) o también pueden eliminarse ítems por no coincidir en un constructo medido (no funciona de la forma en que originalmente se espera) o incluso alterar los puntos de corte de tal forma que se detecte adecuadamente el rasgo medido.

Entonces las escalas o test intentan detectar un rasgo concreto. Este rasgo para poder adscribirse a la definición, tal y como se plantea en el mismo periodo en que fueron ideadas, y también pueden venir de la definición del rasgo que se encuentra en los manuales de los que se ha obtenido el concepto. También puede hacerse por juicio de expertos, que no es más que

una forma paralela a los manuales de definición del rasgo. No obstante, a continuación se plantea un problema y es que es posible que el cuestionario se haya ideado para medir el tipo de depresión que presenta una población. En concreto los datos utilizados para poder hacer estudios y definir las características de las patologías vienen de sociedades occidentales. E incluso para hacer estudios de campo, es común que los grupos poblacionales estudiados sean estudiantes universitarios. Entonces la definición de un constructo es posible que esté asociada al entendimiento de este constructo en una sociedad en concreto. Es por eso que al presentar este mismo test que pretende medir el constructo ideado en una sociedad, sea necesaria una evaluación del test para comprobar si el constructo se mide de forma similar en sociedades diferentes.

Por estos motivos se hace necesario el adecuar tanto culturalmente como psicométricamente los cuestionarios a la población a la que se desea aplicar. Comúnmente se han encontrado diferencias entre países, aunque en culturas parecidas se tiende a tener resultados parecidos. Cuanto más robustos sean los constructos y mejor sean los cuestionarios que los pretenden medir, menos variabilidad entre culturas se espera.

### 2.2.2. Influencia del sexo en la depresión

La depresión es una dimensión de salud mental que presenta diferencias en función del sexo. Estas diferencias presentan variabilidad en función del país y muestra que se estudie, pero un metaanálisis demuestra que la probabilidad de poseer un diagnóstico de depresión mayor siendo mujer es 1,95 veces más elevada que con respecto a los varones (n=3.638.259; OR=1,95, 95% CI [1,88-2,03], d=0,27 [0,26-0,29]; además estas diferencias por sexo se acrecientan en edades juveniles de en la adolescencia (Salk et al., 2017). No obstante, el diagnóstico de la depresión en varones permanece infradiagnosticada, ya que, aunque se realicen la mitad de diagnósticos de depresión en los varones (National Institute of Mental

Health, 2022) se producen 3,4 veces más de suicidios entre los varones con respecto a las mujeres, quizá por la letalidad de los métodos utilizados (Mergl et al., 2015). A pesar de todo, la depresión es el mejor indicador conocido para predecir suicidio (Gonzalez, 2008). Entre las explicaciones se encuentran la expresión diferencial de emociones en cuanto al sexo (R. F. Levant et al., 2014), las diferencias producidas por la socialización (Call & Shafer, 2018) y la expresión diferencial de las conductas depresivas como trabajar demasiado, abuso de sustancias o agresión (Martin et al., 2013). Además, los mecanismos de afrontamiento a la depresión utilizados también poseen sus diferencias sexuales como la autosuficiencia y el control emocional (Hoy, 2012). Aunque sigue siendo común estudiar la depresión sin diferenciar por sexos, existen iniciativas que implican estudiar por sintomatología predominante en función del sexo, dando equidad a la dimensión de depresión (Shi et al., 2021).

Por otro lado, el asumir los roles de género tradicionales implica claras barreras para diagnosticar depresión en varones, ya que la idea de que las mujeres padecen más depresión que los varones invisibiliza u oculta la posibilidad de que los varones sufran de depresión. De hecho el ser mujer se presenta como un factor de riesgo para sufrir depresión (McCarron et al., 2016). Aunque sufrir depresión puede verse como un estigma, lo hace con más hincapié en los varones, ya que las actitudes de masculinidad tipificadas con rudeza y estoicismo pueden tener mayores dificultades de conocer y expresar los sentimientos (Addis & Hoffman, 2017). Y justamente los varones que asumen masculinidades tradicionales son particularmente más susceptibles de sufrir depresión (Good & Wood, 1995). Además estos perfiles implican una menor búsqueda de ayuda para afrontar la depresión (R. Levant et al., 2011; Wilson & Durbin, 2010). Todo ello, implica un infradiagnóstico de la depresión en varones con respecto a mujeres. Es por ello que las diferencias de depresión en cuanto al sexo deben ser consideradas. Y por ello se proponen puntos de corte diferenciales para una mejor detección

de la depresión en los varones. Por todo ello son necesarios moderadores para incorporar equidad de género (Salk et al., 2017). (Swetlitz, 2021).

## 2.3. Puntos de corte

Para categorizar una variable cuantitativa pasándola a cualitativa estableciendo los puntos de corte de un test se parte de su puntuación realizando una categorización de las puntuaciones, de tal forma que, de un rango de puntuaciones al siguiente, el paciente que haya obtenido cualquier puntuación que se encuentre dentro del dicho rango de puntuaciones, podrá asociársele la categoría asignada. De esta forma si un test mide un constructo y a mayor puntuación del test se le atribuye mayor intensidad de ese constructo, el establecer categorías en función de sus puntuaciones ayuda a etiquetar la intensidad de la presencia del constructo en el paciente. Esto ayuda a la toma de decisiones como considerar la urgencia en administrar terapia a los pacientes.

Los puntos de corte ayudan a establecer tanto la presencia o ausencia del trastorno, como su nivel de intensidad. Por criterio clínico puede afirmarse que el paciente que sufre un trastorno, que vendría entendido como un menoscabo significativo en la calidad de vida implicando la incapacidad del desempeño de tareas cotidianas, así como la capacidad de desarrollar un trabajo.

El categorizar una variable cuantitativa supone una pérdida importante de información. Además, en el caso en que los puntos de corte se eligiesen de forma sistemática, puede dar lugar a que las decisiones tomadas no sean extrapolables a otros contextos. Debido a ello, se recomienda la utilización de puntos de corte en el caso de existir un modelo de decisión con reglas sencillas de actuación y siempre teniendo en cuenta que los puntos de corte pueden verse influidos por criterios diferentes a la sensibilidad y especificidad (Molinero, 2003).

En algunos test, como por ejemplo aquellos que se utilizan en procesos selectivos de las administraciones públicas, se tiene la obligatoriedad de dar a conocer a los participantes de las pruebas selectivas, las baremaciones que se utilizarán en las pruebas. Entonces los puntos de corte o criterios selectivos se deberán ser fijados y documentados con anterioridad a la realización de las pruebas. No obstante, en las pruebas de ejecución típica (pruebas de personalidad, actitudes o valores) podrían desvirtuar la finalidad de estas pruebas, especialmente cuando no requieran de ningún tipo de preparación, sino que estas pruebas pretenden conocer cómo la persona piensa, siente y actúa. Entonces a través de la ocultación de estos puntos de corte o atributos buscados (o repudiados) se estaría garantizando el principio de igualdad, en el que todos los individuos tienen las mismas probabilidades de ser seleccionados con independencia del grupo al que pertenezcan. No obstante y para que no se tomen valores arbitrarios, los puntos de corte, aunque no sean públicos, deben marcarse de antemano, siempre con criterios razonados y documentados y que estos criterios establecidos sean aplicados a los candidatos siempre de la misma forma (Hernández et al., 2015).

Para establecer los puntos de corte se pueden utilizar diferentes métodos que se detallan a continuación.

### 2.3.1. Naturales

Cuando se conocen las variables estudiadas y puedan establecerse agrupaciones consecutivas que se formarían por sí solas puede hablarse de cortes naturales. Tómese como ejemplo las edades como variable continua y los estudios obtenidos. De esta forma educación infantil puede establecerse de 1 a 6 años, primaria de 6 a 12 y así sucesivamente.

### 2.3.2. Equiprobables

Se definen los grupos para que cada uno de los grupos tenga un número similar de sujetos, haciendo que la probabilidad de seleccionar un sujeto al azar no venga determinado por su grupo de pertenencia. Por ejemplo al hacer grupos de trabajo en clase, se asignará un número similar de sujetos a cada uno de los trabajos a realizar de tal forma que las diferencias entre el número de sujetos por cada grupo sean mínimas.

### 2.3.3. Regla de Freedman-Diaconis (intervalo fijo)

Dado que una variable cuantitativa puede representarse con un histograma, esta regla está diseñada para encontrar un número similar de sujetos que obtuvieron determinada puntuación en un determinado rango (Freedman & Diaconis, 1981). De esta forma puede obtenerse una clasificación de categorías optimizadas utilizando el número de sujetos contenido en cada categoría teniendo en cuenta una amplitud de categorías uniformes o fijas (la misma amplitud para todas las categorías).

La ecuación general de esta regla es la siguiente:

$$Amplitud\ de\ las\ categorías = 2\frac{AI(x)}{\sqrt[3]{n}}$$

Donde $AI(x)$ representa la Amplitud Intercuartílica de las puntuaciones y $n$ representa el número de puntuaciones de la muestra $(x)$.

### 2.3.4. Regla de Denby-Mallows (intervalo variable)

Establece una amplitud variable de las categorías en función de las puntuaciones (Denby & Mallows, 2009) obteniéndose categorías con amplitudes variables. De esta forma se

adecúan las puntuaciones a las categorías. Utilizando esta técnica existe el problema de las puntuaciones extremas que pueden mover las amplitudes de las categorías.

### 2.3.5. Métodos de selección de extremos

El 95%. Este método utiliza puntos de corte concentrándose en los valores superiores e inferiores de una distribución, fijándose en los percentiles. Así el 5% de la distribución al considerar los valores superiores e inferiores se parte en 2 obteniéndose los valores 2,5 y 97,5. Estas puntuaciones se consideran puntos de corte lo suficientemente extremos como para considerar a los sujetos que toman estas puntuaciones, como puntuaciones extremas tanto para valores inferiores como superiores.

Cuartiles. Al dividir la muestra en cuartiles un criterio puede ser considerar como puntuaciones elevadas aquellas que sobrepasan el tercer cuartil, mientras que serían puntuaciones reducidas aquellas que fuesen inferiores a lo marcado por el primer cuartil.

Deciles. Al igual que los dos métodos previos, este consiste en considerar como extremos los valores por encima y por debajo a 9 y 1, respectivamente.

Los métodos del 95%, de los cuartiles como de los deciles necesitan tamaños de muestra elevados para evitar sesgos.

### 2.3.6. Punto de corte óptimo o valor mínimo de p

Se basa en realizar pruebas Chi cuadrado para cada punto de corte propuesto, obteniéndose una tabla cuadrada 2x2 de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. El punto de corte se va alterando, obteniéndose una tabla cada vez. Esta técnica establece como el punto de corte óptimo aquel valor de Chi cuadrado más

elevado de entre todos los realizados; esto implica que el valor de p sea el menor, por lo que ambos resultados son coincidentes.

## 2.4. Métodos con curva ROC

La Curva ROC (Característica Operativa del Receptor o Receiver Operating Characteristic Curve, por sus siglas en inglés) es una herramienta originalmente utilizada en los radares de la Segunda Guerra Mundial para ayudar a clasificar qué eran misiles, aviones u otros tipos de ataques y qué era ruido (Armesto, 2011; del Valle Benavides, 2017; Lusted, 1971). El 7 de diciembre de 1941 la Armada Imperial Japonesa ordenó un ataque a la base norteamericana de Pearl Harbor, en Hawai. Este ataque tenía carácter preventivo para impedir el avance de la Armada Norteamericana en el Sudeste Asiático. El ejército nipón empleó 353 aviones causando daños en 188 aviones, 2043 muertos y 1178 heridos, además de enormes daños en destructores, cruceros, buques y demás flota armada. Esto conmocionó al pueblo estadounidense queriendo evitar futuros ataques por sorpresa. Así se desarrolló un programa que tenía como fin mejorar los procesos por los cuales las señales detectadas por los radares se clasificaban como potencialmente destructivas o simplemente ruido.

*Figura 3. Ejemplo de receptor de señal de radar*

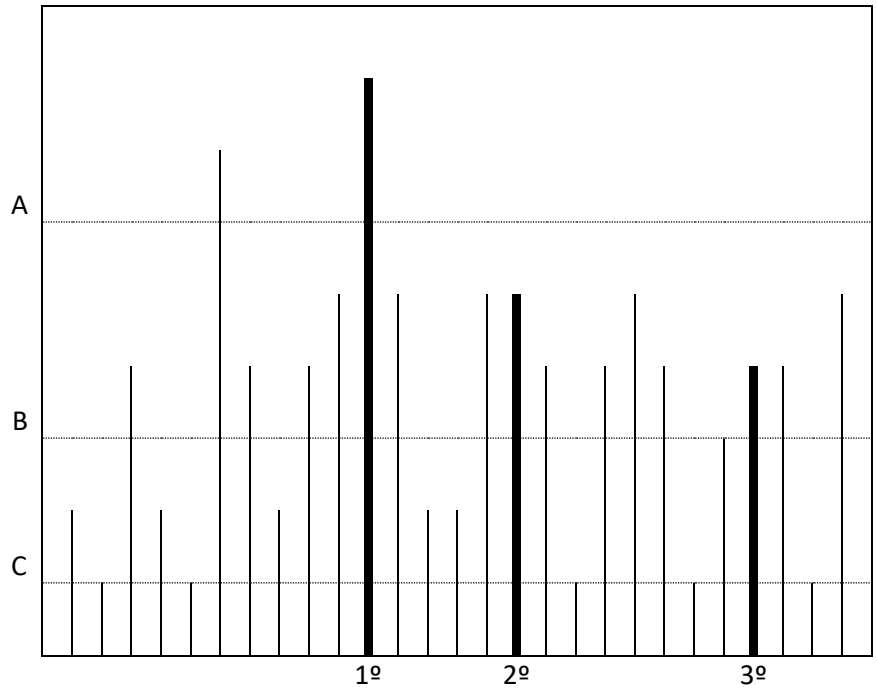



Figura de elaboración propia.

La figura (Figura 3. Ejemplo de receptor de señal de radar) representa las señales recogidas por un radar junto con unos umbrales propuestos. Las columnas representan todas las señales captadas por el radar y aquellas que están destacadas por los ordinales 1º, 2º y 3º representan señales que emiten misiles reales; mientras que el resto de columnas representan otros acontecimientos como pájaros u otras señales de radio detectables. Las líneas horizontales representan los posibles umbrales propuestos para detectar adecuadamente los misiles, que es el evento de interés. En el caso de seleccionar el criterio A detectaría el primer misil (verdadero positivo) más una señal de ruido (falsa alarma), pero no detectaría los misiles 2º y 3º (falsos negativos), dejando la mayoría de las señales que no son misiles, sin detectar (verdaderos negativos). En el caso de seleccionar el criterio B se clasificarían como misiles todos ellos, pero también habría otras 13 señales clasificadas como misiles cuando en realidad no lo son. En el caso de seleccionar el criterio C se clasificarían como misiles todas y cada una de las señales captadas no habiendo diferenciación entre lo que se desea clasificar. Lo

deseable es clasificar como misiles aquellos que lo son, ya que si se detectan, se podrían lanzar señuelos que impedirían que los misiles lograran su objetivo. En caso de no detectar alguno de los misiles no se lanzarían los señuelos y podría tener consecuencias desastrosas.

Este es el ejemplo del uso primigenio de la teoría de detección de señales con la Curva ROC. En este ejemplo es deseable detectar la mayor cantidad de misiles posible, aunque ante las señales clasificadas como misiles es preferible lanzar un señuelo por precaución para evitar males mayores.

El Incidente del Equinoccio de Otoño (o la Falsa Alarma que salvó al mundo de la III Guerra mundial) (Thomson, 2013) también es un ejemplo de las consecuencias de falsas alarmas. Los satélites de alerta temprana (Oko) están diseñados para detectar el lanzamiento de misiles balísticos permitiendo así poder efectuar una acción militar defensiva. Debido a las tensiones militares entre EEUU y la Unión Soviética, en septiembre de 1983 se había dado el aviso de prepararse ante una guerra nuclear: esto es aperturarse a la posibilidad de recibir y enviar misiles nucleares. Si la Unión Soviética recibiese estos misiles, el contraataque consistiría en lanzar misiles nucleares a los EEUU. En este contexto de tensión, un satélite dio la alarma detectando un misil intercontinental estadounidense lanzado desde la Base de la Fuerza Aérea de Malmstrom, tardando en llegar 20 minutos a la Unión Soviética. El teniente coronel Stanislav Yevgráfovich Petrov, encargado del centro de mando de la inteligencia militar soviética tenía por cometido comprobar posibles ataques y dar el aviso de estos a sus superiores. Al recibir esta alarma, Stanislav sabía que si diera el aviso a sus superiores, se iniciaría una guerra nuclear, ya que la Unión Soviética lanzaría misiles nucleares contra EEUU como contraataque. No obstante, decidió esperar a tener más datos que pudieran corroborar esta alarma, y recibió el aviso de que se estarían recibiendo cuatro misiles más. Sin embargo razonó que una guerra nuclear no se empieza con solamente 5 misiles, cuando EEUU tiene miles de ellos. Entonces en lugar de transmitir la advertencia a la cadena de mando, consideró

estas alarmas como falsas alarmas. Esta decisión evitó un ataque nuclear de represalia contra Estados Unidos y sus aliados de la OTAN, lo cual, probablemente habría resultado en una escalada de guerra nuclear a escala mundial. Posteriores investigaciones determinaron que las alertas del satélite habían funcionado mal.

No en todos los contextos es preferible clasificar los eventos de interés por precaución. Supongamos un hospital que desea detectar problemas con el sueño de sus empleados, pero tiene recursos limitados para hacer frente a los resultados que previsiblemente arrojará una población cuyo trabajo consiste en trabajar en un horario destinado a dormir. Entonces como se prevé que un porcentaje de sus trabajadores tenga problemas con el sueño, se considerará se tiene un problema con el sueño solo a aquellos trabajadores cuyo problema con el sueño sea moderado y severo, dejando sin tratar a aquellos que tienen un problema leve. Entonces el valor umbral seleccionado depende de lo preferible que ocurra en cada caso si es preferible detectar los casos en los que verdaderamente ocurra el evento en detrimento de tomar como verdaderos casos falsos; o en cambio detectar solo aquellos casos en los que es más seguro detectar el evento en detrimento de no clasificar como eventos algunos casos que verdaderamente lo son. El problema de dónde situar el punto de corte implica conocer el contexto donde se aplica y las connotaciones de obtener más falsos positivos o falsos negativos. Entonces se deben valorar los contextos de aplicación de los puntos de corte, así como las consecuencias de obtener resultados con mayor o menor porcentaje verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Otro ejemplo son las pruebas diagnósticas para la detección del virus SARS-CoV-2 ¿qué es preferible: poner en cuarentena a todos los casos sospechosos o en cambio poner en cuarentena solo a aquellas personas con síntomas graves de la enfermedad provocada por el virus? Para resolver esta pregunta se deben valorar las implicaciones de cada opción. Una vez

resuelta la pregunta se podrá ajustar el punto de corte que permita clasificar los considerados casos de los que no lo son.

Cuando se habla de verdaderos/falsos positivos/negativos se está considerando la sensibilidad y la especificidad y en una curva ROC se representan ambos. Además, según Zhou y su equipo (Zou et al., 1997) las curvas ROC permiten comparar dos o más test simultáneamente y establecer cuál de ellos ofrece mejor capacidad discriminante.

Entonces pueden clasificarse sujetos como sigue:

*Tabla 3. Tabla de contingencia entre el patrón oro (Gold Standard) y el resultado del test*

| | | Resultado del test | |
|---|---|---|---|
| | | Positivo | Negativo |
| Gold Standard | Depresivo | Verdadero positivo ($V^+$) | Falso negativo ($F^-$) |
| | No depresivo | Falso positivo ($F^+$) | Verdadero negativo ($V^-$) |

Tabla de elaboración propia.

De esta forma se consideran 4 grupos y para explicarlo se toma como el trastorno depresivo y la detección del mismo mediante un test psicológico:

- Los verdaderos positivos ($V^+$). Fracción de Verdaderos Positivos (FVP). Realmente tienen depresión y además esta depresión se detectó en el test.
- Los verdaderos negativos ($V^-$). Fracción de Verdaderos Negativos (FVN). Los integrantes de este grupo no sufren de depresión además el test los clasificó correctamente no detectando depresión.
- Falsos positivos ($F^+$). Fracción de Falsos Positivos (FFP) o también designado como error tipo I. Este grupo se compone de aquellos sujetos que no sufren de depresión, no presentan los síntomas, pero en cambio el test ha detectado que sí sufren de depresión. Cuanto menor sea este grupo, mejores clasificaciones hará el test.
- Falsos negativos ($F^-$). Fracción de Falsos Negativos (FFN) o también designado como error tipo II. Al igual que los falsos positivos, este grupo se refiere a una incorrecta

clasificación de sus integrantes. Está compuesto por quienes sí sufren de depresión, pero el test no ha logrado detectarlos.

### 2.4.1. Sensibilidad y especificidad

De una forma con enfoque cualitativo, en el ámbito de la entrevista psicológica forense, la década de 1980 fue llamada como "era de la sensibilidad", ya que existía un interés excesivo en tener certeza de que no se dejara sin detectar a ninguna víctima de abusos sexuales. Esto incrementaba la tasa de falsos positivos.

A mediados de 1980, en EEUU se inició un cambio legislativo en el que el abuso sexual se incluía como maltrato. En este entonces se tomaba siempre como cierto el discurso del menor y suficiente para enjuiciar a los acusados. Además, la forma de obtener la información, de hacer entrevistas forenses y recabar otros datos, no estaban lo suficientemente refinados. En estas entrevistas existían sugestiones y se daba el heurístico de disponibilidad, que viene siendo que cuanto más accesible y disponible sea un suceso, más frecuente y probable parecerá. También existía perseverancia en la creencia, esto es que al tomar una decisión, normalmente preservamos esta decisión inicial pese a los acontecimientos que puedan darse después, como nuevas evidencias. En aquel entonces, se daba la tendencia a verificar la hipótesis única o profecía autocumplida que implicaba aceptar y no falsar la hipótesis. No se cuidaba el contexto pudiendo darse la entrevista en una atmósfera tensa. Todo ello podía facilitar que el relato correspondiese a una situación no vivenciada, dándose falsa memoria, fabulación, mentira u olvido u otros sesgos importantes que condicionaban la forma tanto de entrevistar como la forma de obtener información (Fisher & Geiselman, 2010; Saywitz et al., 1992).

En esta época aumentaron las detecciones de casos de abusos cuando en realidad las personas a quienes se les detectó, no lo sufrieron. El lema de entonces era *Cuéntame tu secreto para que pueda ayudarte*. El caso Mc. Martin (Garven et al., 1998) fue un gran ejemplo

en evidenciar la importancia de una minimización de errores en cuanto a la detectar incorrectamente abusos sexuales en niños.

Tanto la normativa como la forma de realizar las pruebas a los menores hicieron que aparecieran casos de abusos sexuales en las guarderías que implicaban a profesores y directores; aunque no siempre estas acusaciones eran ciertas. Hubo un gran revuelo mediático y se empezó a cuestionar el método de entrevistar. Se crearon modelos de investigación, se inició el primer centro de entrevistas a menores (Children's Advocacy Center) y los métodos de entrevista cambiaron. Entonces a raíz del caso Mc. Martin, hubo un movimiento en el polo opuesto, y en la década de 1990 se inició la *era de la especificidad* en la que la prevención de errores de falsos positivos se convirtió en una prioridad, aumentando consecuentemente la tasa de falsos negativos. Esto intentaba minimizar la victimización de los niños. El lema por entonces era *Si afirmas haber sido abusado sexualmente, tendrás que convencerme* (Garven et al., 1998; Saywitz et al., 1992). Posteriormente se evidenció el daño de una incorrecta clasificación impulsando asimismo la corrección de los índices de sensibilidad y especificidad de las pruebas ya que tanto los falsos positivos como los falsos negativos son motivo de preocupación (Faller, 2014). Por otro lado, el paradigma actual liderado por el actual protocolo de entrevista investigativa sobre abusos sexuales a niños, el llamado protocolo NICHD (Karni-Visel et al., 2019) pretende darles importancia a los errores producidos tanto por falsos negativos como por falsos positivos, cuidando ambos errores. Entonces se refuerza el apoyo emocional que pudiera ofertarse en vez de poner el énfasis en descubrir o averiguar la presencia o ausencia de abusos, sino que el paradigma actual es contemplar y atender las referencias sobre abuso sexual que el infante pudiera hacer. Y tanto si se encuentra algo como si no, se le dará apoyo emocional. A nivel terapéutico se estaría reconociendo y validando las emociones del sujeto. El lema en la actualidad sería: "Te voy a acompañar, te voy a ayudar; y cuando se encuentren detalles, si es que aparecen, también se recogerán estos detalles para poder obtener más información sobre la situación en la que te encuentras" (Grupo de trabajo

e investigación de la sección de psicología jurídica y forense del COPC, 2014; Ramón et al., 2018).

De forma más formal, la sensibilidad y especificidad son indicadores estadísticos que evalúan el grado de eficacia de una prueba diagnóstica y miden la discriminación diagnóstica de una prueba en relación a un criterio de referencia que es considerado genuinamente verdadero (*gold standard*) (Yerushalmy, 1896). Por ello, son valores que informan de cómo discrimina el test.

La Sensibilidad, también llamada Fracción de Verdaderos Positivos (FVP) se refiere a la probabilidad de una correcta clasificación positiva. Esto es, que si el sujeto presenta depresión y además se le detecta depresión, sería la probabilidad de ser detectada esta circunstancia. Expresa cómo de sensible es la prueba para detectar la enfermedad. Puede entenderse como la probabilidad de que el test clasifique como depresivo a la persona que realmente lo está. Y puede formalmente definirse como la probabilidad resultante de dividir el número de casos de verdaderos positivos entre el número de positivos, según el *gold standard*.

$$Sensibilidad = FVP = \frac{V^+}{V^+ + F^-} = \frac{Verdaderos\ positivos}{Total\ de\ depresivos}$$

La especificidad, la Fracción de Verdaderos Negativos (FVN) se refiere a la probabilidad de una correcta clasificación negativa. Dicho de otra forma: un sujeto que no presenta depresión de forma correcta no se le detecta depresión en la prueba, indicando la capacidad de identificar a los sujetos sanos. Puede entenderse como la probabilidad de detectar de forma correcta una persona sin depresión o exactitud negativa y formalmente es la probabilidad de casos verdaderos negativos entre el número total de no depresivos.

$$Especificidad = FVN = \frac{V^-}{V^- + F^+} = \frac{Verdaderos\ negativos}{Total\ de\ no\ depresivos}$$

Un adecuado nivel de Sensibilidad y Especificidad hacen de una prueba adecuada siempre considerando su contexto. Esto es porque el contexto de la prueba marca los objetivos a perseguir.

Debido a que se desea conocer el índice de severidad y no una clasificación en función del número de sujetos que se encuentre en cada categoría, una opción adecuada es utilizar el criterio de *closest top-left* o el índice de Youden (Youden, 1950) que utilizan criterios externos en pruebas de diagnóstico. En cuanto a establecer la presencia o ausencia de un trastorno se utiliza un criterio externo. De tal forma que si está constatado que un paciente padece determinado trastorno, se espera que la puntuación obtenida en el test lo refleje. Ambos tienen el inconveniente que refleja los resultados de manera dicotómica y que no siempre son coincidentes (Perkins & Schisterman, 2006).

### 2.4.2. Valor predictivo

Cuando aún no se ha confirmado si el resultado del test es correcto o incorrecto, el valor predictivo indica la probabilidad de que el resultado obtenido sea correcto o incorrecto. Estos valores dependen de la prevalencia de su condición poblacional. Entonces el valor predictivo de un sujeto dependerá de lo frecuente o infrecuente que sea la enfermedad en su población.

El valor predictivo se utiliza para conocer la probabilidad de que, el resultado del test sea positivo, esto es que se haya detectado un paciente con depresión, aunque no siempre sea correcta esta detección. En sentido contrario el valor predictivo indica la probabilidad de que una persona que no haya sido detectada con depresión por el test, esté realmente sana (Molina Arias, 2013).

El Valor Predictivo Positivo (VPP) es la probabilidad de que un sujeto clasificado como positivo (deprimido) sea en realidad positivo (con diagnóstico de depresión). Es calculado a

partir de la razón de sujetos positivos detectados con el test sobre el número de total de sujetos con la enfermedad (verdaderos positivos entre todos los diagnosticados con depresión).

$$VPP = \frac{V^+}{V^+ + F^-}$$

Valor Predictivo Negativo (VPN) es la probabilidad de que un sujeto con resultado negativo en el test (sano) sea en realidad negativo (sin diagnóstico de depresión). Es la proporción de verdaderos negativos entre quienes no tienen diagnóstico de depresión. Se calcula como sigue:

$$VPN = \frac{V^-}{V^- + F^+}$$

### 2.4.3. Índices de ratio y riesgo

La razón de momios, también llamado *Odds ratio* en inglés, de una prueba indica la probabilidad real de ser correctamente clasificado. Para conocer este índice se utiliza la probabilidad previa, así como el resultado del test.

Razón de verosimilitud positiva o Cociente de Probabilidad Positivo ($CP^+$) indica la probabilidad de detectar como positivo a un sujeto verdaderamente positivo (Molina Arias, 2013). Dicho de otro modo: es la probabilidad de que un sujeto con diagnóstico de depresión se le detecte depresión con el test. Esta razón viene determinada por la expresión:

$$CP^+ = \frac{Sensibilidad}{1 - Especificidad}$$

La Razón de verosimilitud negativa o Cociente de Probabilidad Negativo ($CP^-$) indica cuántas veces es más probable de detectar como negativo a un sujeto verdaderamente

negativo. O también, la probabilidad de que un sujeto sin depresión se le detecte efectivamente sin depresión. Su expresión es la siguiente:

$$CP^{-} = \frac{(1 - Sensibilidad)}{Especificidad}$$

La Razón De Momios Diagnóstica (DOR) es una medida de efectividad del test. Es definida como la razón entre las probabilidades de que la prueba tenga resultado positivo si el sujeto es realmente positivo y las probabilidades de que la prueba sea positiva si el sujeto no es realmente positivo. Los *Odds* ratio, así como los Valores Predictivos dependen de la prevalencia de la enfermedad en la población de los sujetos. Su expresión es la siguiente:

$$DOR = \frac{CP^{+}}{CP^{-}} = \frac{V^{+}}{F^{+}} / \frac{F^{-}}{V^{-}}$$

Este DOR no puede calcularse cuando $F^{+} = 0$ o cuando $F^{-} = 0$. En cambio cuando $F^{+} = F^{-} = 0$ la prueba se considera perfecta.

Los *Odds* ratio pueden tomar valores de 0 a $+\infty$. El valor nulo es el 1 indicando que el test no tiene capacidad discriminatoria. En cambio, valores superiores a la unidad indicarían capacidad discriminatoria positiva. Los valores menores a la unidad indica que el test puede mejorarse invirtiendo el resultado de la prueba (el test mide de forma inversa: no detecta depresión, sino que detecta la ausencia de esta). Cuando mayor sea este índice, indicará un mayor rendimiento de la prueba.

El DOR es utilizado en metaanálisis por su simplicidad y facilidad para poder establecer un valor entre Especificidad y Sensibilidad, además que puede combinarse con otros estudios y poder establecer la precisión de las pruebas gracias a ello (Moses et al., 1993).

### 2.4.4. Gráficos de la curva ROC

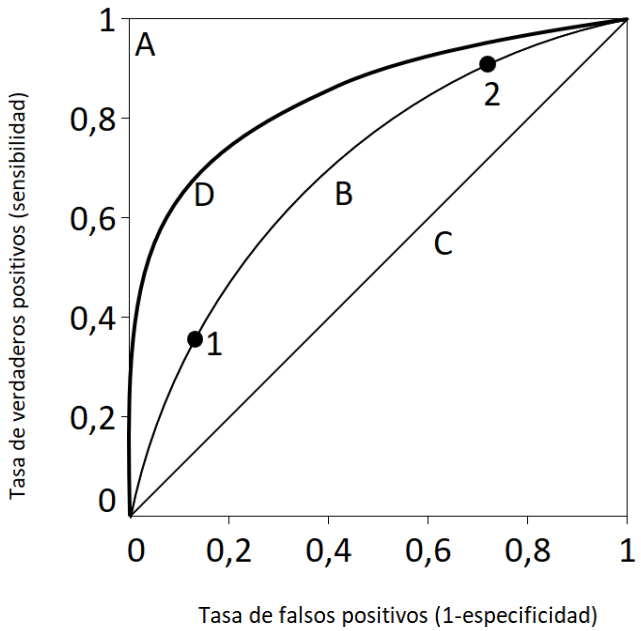
Esta técnica de visualización permite organizar y clasificar los resultados de los test en función de su desempeño en la muestra aplicada. Es una técnica ampliamente utilizada en la teoría de detección de señales, así como para establecer un punto por el cual los falsos positivos y los falsos negativos se equilibren de manera conjunta (Swets, 1988; Swets et al., 2000). Por estos motivos se emplea en ciencias sanitarias para poder tomar decisiones sobre herramientas diagnósticas (Zou et al., 2007).

Esta curva representa conjuntamente la $Sensibilidad$ y $1 - Especificidad$ en todos y cada uno de los puntos posibles de corte posibles.

$$ROC(c) = \begin{cases} y = S(c) \\ x = 1 - E(c) \end{cases}$$

Este gráfico bidimensional (Figura 4. Curvas ROC hipotéticas) se construye a partir de la representación conjunta de la sensibilidad y la especificidad. En concreto en el eje X se representa el complementario de la tasa de verdaderos positivos $1 - Especificidad$. Mientras que en el eje Y se representa la tasa de verdaderos positivos o $Sensibilidad$.

*Figura 4. Curvas ROC hipotéticas*




La curva A corresponde a una perfecta clasificación (AUC=1). La curva C corresponde a una clasificación al azar (AUC=0,5). La curva B corresponde a una clasificación típica con baja capacidad de clasificación (AUC=0,65). La curva D corresponde a una clasificación típica con adecuada capacidad de clasificación (AUC=0,80). Los puntos 1 y 2 corresponden a propuestas de puntos de corte de la línea B, siendo el 1 más conservador y el 2 más liberal. Figura de elaboración propia.

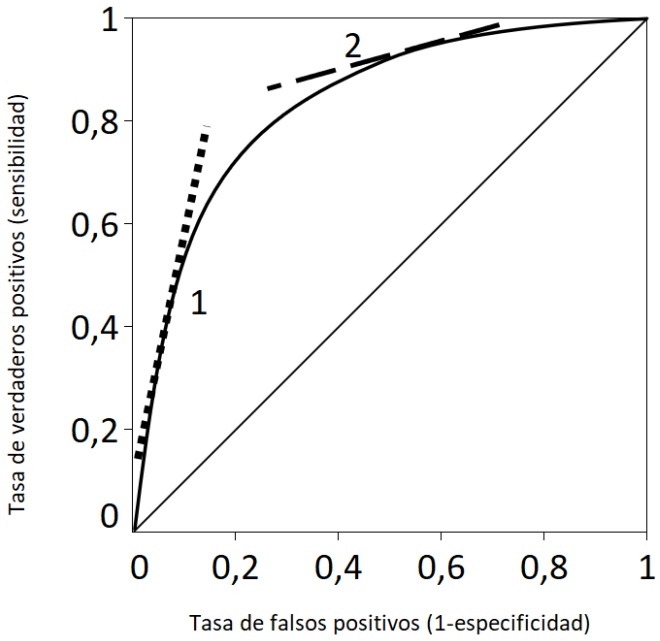
El espacio ocupado entre la línea diagonal, de 45º, y los catetos inferior y derecho no es utilizado en ningún momento, dejando únicamente el espacio utilizable, el complementario del cuadrado: las líneas derecha (0,0-0,1) y superior (0,1-1,1); generalmente las curvas ROC producidas por los test o cualquiera otra herramienta a evaluar se encuentre entre estos dos extremos; ya que la representación en el cuadrado de la sensibilidad y (1-especificidad) del patrón oro coincide con estas líneas. Mientras que la línea diagonal corresponde a una probabilidad de 0,5 de clasificación correcta.

En este espacio, del triángulo superior, se representa la especificidad y sensibilidad de forma conjunta de uno o varios test, recordando que ante la especificidad se representa su complementario con la unidad (1-especificidad). Esta representación viene descrita como una sucesión de puntos, que unidos, forman una línea la cual ocupa el espacio de la diagonal

superior. Cada uno de estos puntos es el valor que toma la sensibilidad y (1-especificidad) para cada valor del punto de corte. Esta línea, cuanto más se aleje de la diagonal y más se acerque a las líneas vertical y horizontal utilizables del cuadrado, dará indicaciones de una mejor clasificación. Entonces para obtener un punto de corte se ha de elegir una puntuación cuya tasa de falsos positivos y verdaderos positivos sea aceptable. Comúnmente se considera aceptable una tasa por la cual supere el 0,8 del área del espacio. A este área se le designa como Área Bajo la Curva (AUC) siendo la probabilidad igual al azar el 0,5 y el 1 a la perfecta clasificación. A medida que aumenta la precisión del test este índice se aproxima más a la unidad.

De forma informal, un punto de la línea descrita en el espacio ROC es mejor que otro si se posiciona en el nordeste de la figura (más próximo al punto (0,1); ya que este punto representa la perfecta clasificación (T. Fawcett, 2006). De forma similar, los puntos de corte que se sitúen en la parte más a la izquierda del gráfico, cercanos al eje X (como el punto 1 de la figura (Figura 5. Ejemplos de pendientes de la recta tangente a la curva ROC)), serán considerados más conservadores ya que se hará una clasificación positiva solo cuando se marca una elevada evidencia. De esta forma se cometen pocos errores debidos a falsos positivos. En cambio, aquellos puntos de corte más cercanos al eje Y en la parte superior de la derecha (como el punto 2 de la figura (Figura 4. Curvas ROC hipotéticas)), serán considerados más liberales haciendo que se la mayoría de positivos se clasifiquen correctamente; esto implica que la tasa de falsos positivos aumente.

*Figura 5. Ejemplos de pendientes de la recta tangente a la curva ROC*




El puntos de corte de la pendiente 1 es considerado como más conservador siendo la prueba más específica, mientras que el punto de corte de la pendiente 2 es considerado como más liberal en los diagnósticos aumentando la sensibilidad de la prueba. Figura de elaboración propia.

### 2.4.5. Métodos ROC de estimación paramétricos y no paramétricos

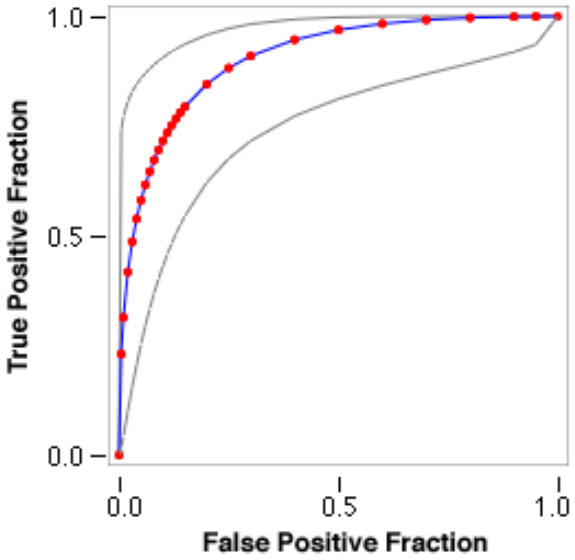
Para los métodos no paramétricos de la creación de una curva ROC se emplean los puntos generado por cada punto de corte posible, en el que se pueden definir su especificidad junto con su (1-sensibilidad) (Hsieh & Turnbull, 1996). Cualquier curva ROC que se genere a partir de un conjunto finito de pares de puntuaciones generará una función escalonada, la cual se irá suavizando conforme más elementos contenga (T. Fawcett, 2006).

Como alternativa a los métodos no paramétricos, los modelos paramétricos asumen que ambas medidas tienen distribuciones normales e independientes con diferentes medias y desviaciones típicas. Uno de estos modelos es la distribución exponencial binegativa implementada en el programa SPSS y más usada en estudios de análisis de supervivencia.

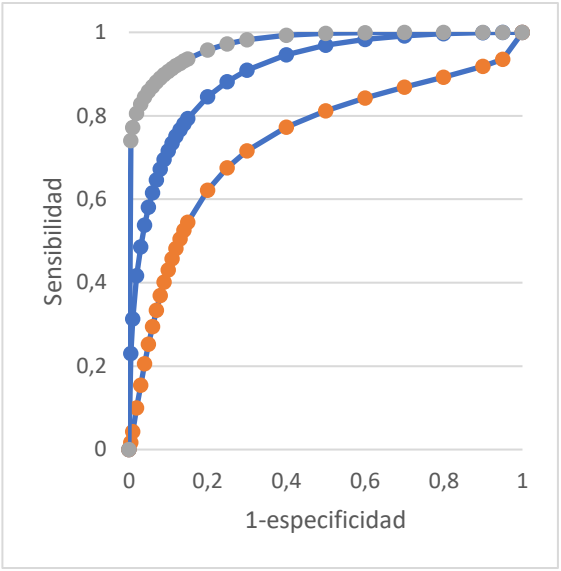
A continuación, y con la misma base de datos, se ofrece una figura (Figura 6. Representación de la curva ROC con los mismos datos y diferentes programas) compuesta por 3 gráficos generados bajo 3 programas diferentes. Los 3 gráficos representan la misma curva ROC en la que, de 50 sujetos, 42 de ellos fueron clasificados correctamente. De entre los casos clasificados incorrectamente, 3 corresponden a positivos no detectados o falsos negativos ($F^-$), mientras que 5 corresponden a negativos no detectados o falsos positivos ($F^+$). Los datos de este ejemplo ofrecen una precisión del 84% y un AUC=0,892.

*Figura 6. Representación de la curva ROC con los mismos datos y diferentes programas*

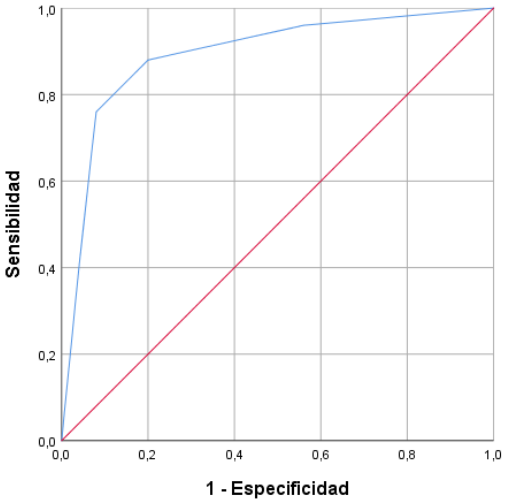
1:




2:




*3:*




Elaboración de 1 con: ROC Analysis. Online ROC Curve Calculator. Elaboración de 2 con: Excel. Elaboración de 3 con: SPSS. Nota: La línea central y los puntos corresponden a la curva ROC ajustada. Las líneas que marcan el contorno de la línea central marcan los intervalos de confianza al 95% de la curva ROC ajustada. Los datos de ejemplo fueron obtenidos de la propia página web. Fuente: (Eng, 2014).

Pueden verse los datos que permitieron la generación de los gráficos de la figura (Figura 6. Representación de la curva ROC con los mismos datos y diferentes programas) en la siguiente tabla (Tabla 4. Ejemplo de puntuaciones de un test de un test de depresión junto con el diagnóstico de depresión. Base para la generación de curva una ROC). Esta tabla se compone

de dos columnas. La columna de Gold Standard está destinada para acoger las puntuaciones del patrón oro en formato dicotómico: la presencia (1) o ausencia (0) de la depresión. En la columna de Puntuaciones del test se contemplan las puntuaciones obtenidas de cada sujeto en un test de depresión ficticio, con un rango de respuestas que va desde el 1 hasta el 6, en números enteros; e indica que cuanta más elevada sea la puntuación, la severidad de la depresión es mayor, según este test. Cada fila corresponde a la puntuación conjunta de un sujeto tanto en su presencia o no de depresión, en este caso Gold Standard, como su puntuación del test; no obstante, se ha partido la tabla en dos para economizar espacio. Para identificar a qué registro corresponde cada par de puntuaciones se agregó una columna de número (Nº), de tal forma que la tabla tiene 6 columnas, repitiendo una vez 3 de ellas.

*Tabla 4. Ejemplo de puntuaciones de un test de un test de depresión junto con el diagnóstico de depresión. Base para la generación de curva una ROC*

| Nº | Gold Standard | Puntuaciones del test | Nº | Gold Standard | Puntuaciones del test |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 26 | 1 | 1 |
| 2 | 0 | 1 | 27 | 1 | 2 |
| 3 | 0 | 1 | 28 | 1 | 2 |
| 4 | 0 | 1 | 29 | 1 | 3 |
| 5 | 0 | 1 | 30 | 1 | 3 |
| 6 | 0 | 1 | 31 | 1 | 3 |
| 7 | 0 | 1 | 32 | 1 | 4 |
| 8 | 0 | 1 | 33 | 1 | 4 |
| 9 | 0 | 1 | 34 | 1 | 4 |
| 10 | 0 | 1 | 35 | 1 | 4 |
| 11 | 0 | 1 | 36 | 1 | 4 |
| 12 | 0 | 2 | 37 | 1 | 4 |
| 13 | 0 | 2 | 38 | 1 | 4 |
| 14 | 0 | 2 | 39 | 1 | 4 |
| 15 | 0 | 2 | 40 | 1 | 4 |
| 16 | 0 | 2 | 41 | 1 | 5 |
| 17 | 0 | 2 | 42 | 1 | 5 |
| 18 | 0 | 2 | 43 | 1 | 5 |
| 19 | 0 | 2 | 44 | 1 | 5 |
| 20 | 0 | 2 | 45 | 1 | 5 |
| 21 | 0 | 3 | 46 | 1 | 5 |
| 22 | 0 | 3 | 47 | 1 | 5 |
| 23 | 0 | 3 | 48 | 1 | 5 |
| 24 | 0 | 4 | 49 | 1 | 5 |
| 25 | 0 | 5 | 50 | 1 | 5 |

Tabla de elaboración propia.

Para generar estos gráficos (Tabla 5. Coordenadas de los puntos de la curva ROC ajustada con IC al 95%) se han tenido en cuenta tanto la FFP o Fracción de Falsos Positivos como la FVP o Fracción de Verdaderos Positivos. Además se añadió el Intervalo de Confianza al 95%. Por lo que el resultado se compone de 3 sucesiones de puntos que conforman 3 líneas: la primera línea conformada tras especificar en el plano cada uno de los puntos de FFP junto con los de FVP; la segunda y tercera línea (límite inferior y superior) resultado de representar cada punto de estos límites con FVP. Para poder realizar tanto los cálculos como el ajuste se utilizó el programa en línea *ROC Analysis Online ROC Curve Calculator* (Eng, 2014).

*Tabla 5. Coordenadas de los puntos de la curva ROC ajustada con IC al 95%*

| FFP | FVP | 95% Inferior | 95% Superior |
|---|---|---|---|
| 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 0.0050 | 0.2301 | 0.0169 | 0.7407 |
| 0.0100 | 0.3135 | 0.0430 | 0.7718 |
| 0.0200 | 0.4168 | 0.0996 | 0.8061 |
| 0.0300 | 0.4860 | 0.1545 | 0.8282 |
| 0.0400 | 0.5384 | 0.2056 | 0.8449 |
| 0.0500 | 0.5807 | 0.2523 | 0.8587 |
| 0.0600 | 0.6159 | 0.2949 | 0.8705 |
| 0.0700 | 0.6461 | 0.3337 | 0.8808 |
| 0.0800 | 0.6723 | 0.3690 | 0.8901 |
| 0.0900 | 0.6955 | 0.4012 | 0.8985 |
| 0.1000 | 0.7161 | 0.4306 | 0.9062 |
| 0.1100 | 0.7347 | 0.4575 | 0.9132 |
| 0.1200 | 0.7515 | 0.4821 | 0.9198 |
| 0.1300 | 0.7668 | 0.5047 | 0.9258 |
| 0.1400 | 0.7809 | 0.5255 | 0.9314 |
| 0.1500 | 0.7938 | 0.5447 | 0.9366 |
| 0.2000 | 0.8454 | 0.6214 | 0.9577 |
| 0.2500 | 0.8822 | 0.6757 | 0.9723 |
| 0.3000 | 0.9096 | 0.7160 | 0.9824 |
| 0.4000 | 0.9466 | 0.7727 | 0.9934 |
| 0.5000 | 0.9691 | 0.8119 | 0.9978 |
| 0.6000 | 0.9832 | 0.8424 | 0.9994 |
| 0.7000 | 0.9918 | 0.8684 | 0.9999 |
| 0.8000 | 0.9967 | 0.8927 | 1.0000 |
| 0.9000 | 0.9992 | 0.9189 | 1.0000 |
| 0.9500 | 0.9998 | 0.9357 | 1.0000 |
| 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Tabla de elaboración propia.

Una vez obtenidas las coordenadas y representada en el plano la curva ROC pueden estimarse los puntos de corte para optimizar la proporción de FVP y FFP asumible por el test conforme a las características deseadas. La estimación de puntos de corte es motivo del siguiente apartado.

### 2.4.6. Métodos de optimización del punto de corte mediante curva ROC

Dada una distribución aleatoria en la que se encuentren sujetos con depresión y sujetos sin depresión, un estado ideal sería que ambos grupos no se solapasen, existiendo una adecuada separación entre dichos grupos y conformándose como dos distribuciones claramente diferenciadas. De esta forma solo se obtendrían verdaderos positivos y verdaderos negativos o una sensibilidad y especificidad máximas, haciendo evidente el punto de corte. Pero en la realidad los sujetos con depresión y sin depresión pueden tener puntuaciones del test entremezcladas, y dependerá del punto de corte del test, el hecho de que a los sujetos se les atribuya o no depresión.

Por lo tanto tenemos dos distribuciones, una de enfermos y otra de sanos, y la intención es marcar un punto de corte tal, que separe a ambos grupos de la forma más eficiente posible. La eficacia de este punto de corte puede medirse en la minimización de la detección de falsos positivos o falsos negativos, aunque, como se ha comentado anteriormente, dependiendo de los intereses a satisfacer, puede llegar a ser más interesante aumentar la sensibilidad en detrimento de la detección de falsos positivos o aumentar la (1-especificidad), aumentando consecuentemente la detección de negativos cuando realmente no lo son.

El punto de corte es una puntuación en el test a partir de la cual, el sujeto que la supere, se considere como enfermo, mientras que si no la supera, se le considera sano. En test

que tuvieran las puntuaciones no dicotomizadas, la definición de punto de corte sería únicamente dicotómica al separar dos grupos y no más de dos. En el caso de contar con un test cuyos resultados fuesen categóricos de más de dos opciones, podrían estimarse tantos puntos de corte como diferencias entre pares de grupos consecutivos. Aunque para poder lograr este escenario, el *gold standard* debería tener marcadas también las categorías deseables y consecutivas aumentando en severidad de la enfermedad.

Para la explicación se considera que existen dos distribuciones normales en sendos grupos identificados: uno de sanos y otro de enfermos (no depresivos y depresivos, respectivamente). En la muestra, ambos grupos se representan con distribuciones normales y varianzas iguales, además estas distribuciones se encuentran separadas. Esta separación dependerá de cuán similares sean las medias de estos grupos y consecuentemente de cuánto se entremezclen entre sí (Figura 7. Curvas ROC: Elección de puntos de corte y área bajo la curva (AUC)).

*Figura 7. Curvas ROC: Elección de puntos de corte y área bajo la curva (AUC)*

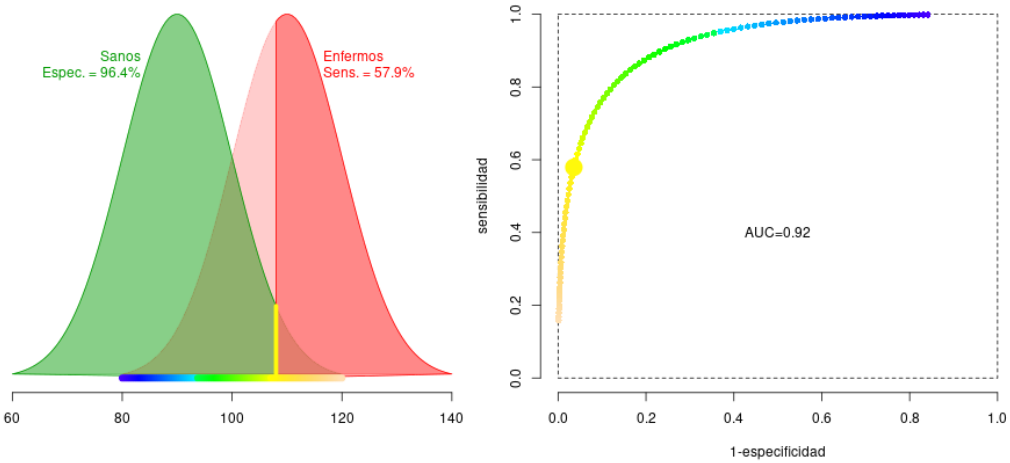



Figura de elaboración propia con programa informático. Fuente: (Barón López, 2022)

Una vez calculados los índices de sensibilidad y especificidad, puede generarse la curva ROC de forma complementaria a las distribuciones de frecuencias.

En el ejemplo de la figura (Figura 7. Curvas ROC: Elección de puntos de corte y área bajo la curva (AUC)) entre ambas puntuaciones los grupos se entremezclan, por lo que considerando una selección taxativa, quedará necesariamente errado el diagnóstico. No obstante, puede estimarse este error o tasa de sujetos incorrectamente clasificados teniendo en cuenta el contexto del test y la naturaleza de la complementariedad de $V^+$, $V^-$, $F^+$y $F^-$. Esto es debido a que al alterar el punto de corte, la tasa de sujetos clasificados mediante el test se modifica, ya que $Sensibilidad$ y $(1 - Especificidad)$ toman valores diferentes en función del punto de corte seleccionado.

Pongamos un ejemplo. Sea un test de depresión y un conjunto de sujetos, de entre los cuales 100 corresponden a sujetos sanos y otros 100 corresponden a sujetos enfermos. Las distribuciones de frecuencias de las puntuaciones de estos sujetos en el test para detectar depresión tienen varianzas iguales, pero medias diferentes. De esta forma y dado que las puntuaciones del test para detectar depresión se interpretan en que cuanto mayor sea la puntuación, más depresión es detectada, las puntuaciones de los sujetos enfermos se espera que se sitúen con una media superior a la distribución conformada por sujetos enfermos. En el caso de que la separación entre grupos sea, por ejemplo de 20 puntos, indica que las medias de ambos grupos se separan por 20 puntos en las puntuaciones del test. Ahora supongamos que marcamos el punto de corte en 90 puntos. La clasificación de los sujetos queda como sigue (Tabla 6. Ejemplo de clasificación de los sujetos con punto de corte de alta sensibilidad).

*Tabla 6. Ejemplo de clasificación de los sujetos con punto de corte de alta sensibilidad*

| Resultado del test | Gold Standard | Frecuencia |
|---|---|---|
| Positivo | Enfermo | 98 |
| Negativo | Enfermo | 2 |
| Positivo | Sano | 50 |
| Negativo | Sano | 50 |

*Nota*. La curva ROC de este ejemplo tiene un punto de corte 90, n=200 y su diferencia entre medias es de 20 unidades. Tabla de elaboración propia.

Una vez generada la tabla a partir de los datos del ejemplo se puede calcular su sensibilidad.

$$Sensibilidad = FVP = \frac{V^+}{V^+ + F^-} = \frac{Verdaderos\ positivos}{Total\ de\ depresivos} = \frac{98}{100} = 0{,}980$$

Y su especificidad.

$$Especificidad = FVN = \frac{V^-}{V^- + F^+} = \frac{Verdaderos\ negativos}{Total\ de\ no\ depresivos} = \frac{50}{100} = 0{,}500$$

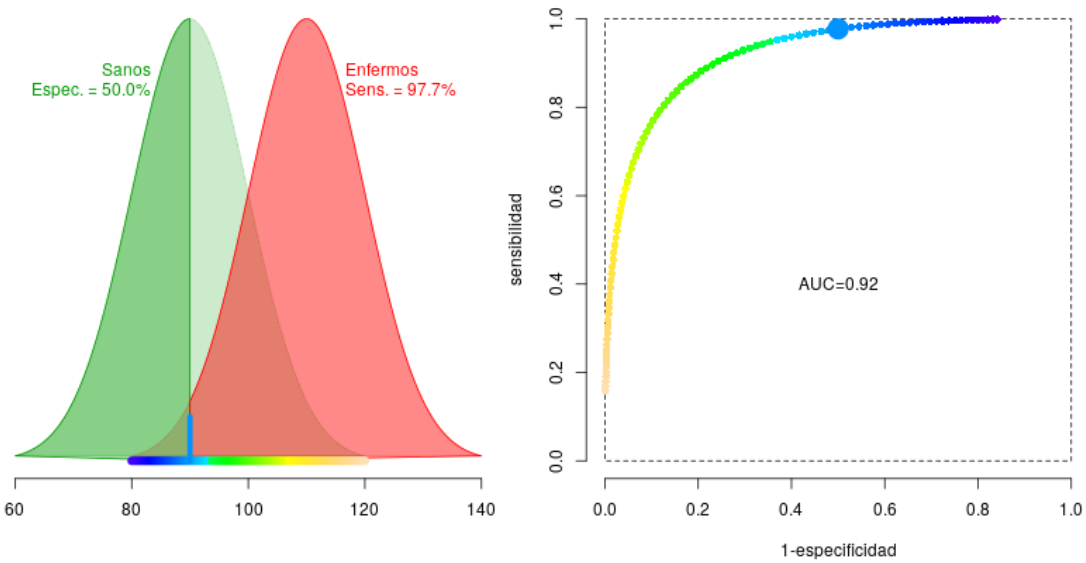
Conviene recordar que para representar la especificidad se utiliza su complementario a la unidad, pero en este caso resulta en el mismo número de 0,500.

Como podemos observar en la siguiente figura (Figura 8. Distribución de frecuencias y AUC con punto de corte de alta sensibilidad) el área bajo la curva ROC es adecuado al superar el punto de 0,8 (AUC=0,92). Dado que esta área indica la probabilidad del test de clasificar correctamente, este índice puede considerarse adecuado. No obstante, el punto de corte seleccionado en la figura (Figura 8. Distribución de frecuencias y AUC con punto de corte de alta sensibilidad) y representado por un círculo situado encima de la curva ROC; implica que la discriminación de sanos y enfermos realizada a partir de este punto de corte, presenta una gran sensibilidad, cercana a la unidad. Haciendo que los positivos se detecten fácilmente. Implica que, consecuentemente, deje un reducido índice de verdaderos positivos no detectados. Por otro lado, el punto de corte seleccionado implica que la especificidad sea del 50%: el punto de azar; teniendo la misma probabilidad de que debido a las puntuaciones del test se clasifique a un sujeto verdaderamente sano tanto sano o como enfermo. Sin distinción entre ellos. De esta forma puede decirse que este test con este punto de corte tiene una gran sensibilidad, pero nula especificidad. En el caso en que la especificidad baje de ese 50% quiere decir que la capacidad del test para detectar sanos sería más baja que el azar. Dicho de otra forma: si el test ha detectado a un sujeto como sano y tomando como información únicamente

la pertenencia al grupo y no sus puntuaciones, habría más probabilidad de que este sujeto clasificado como sano, en realidad estuviese enfermo.

Puede observarse en la figura que se dibuja una línea vertical que representa el punto de corte, separando las distribuciones de frecuencias considerando sanos o enfermos aquellos sujetos cuyas puntuaciones se sitúen a la derecha o izquierda de esta línea, respectivamente. De forma complementaria, al modificar este punto de corte representado como la línea vertical en las distribuciones de frecuencias, también se modifica el círculo situado encima de la curva ROC. Dado que la separación entre distribuciones se mantiene constante, el AUC también se mantiene constante.

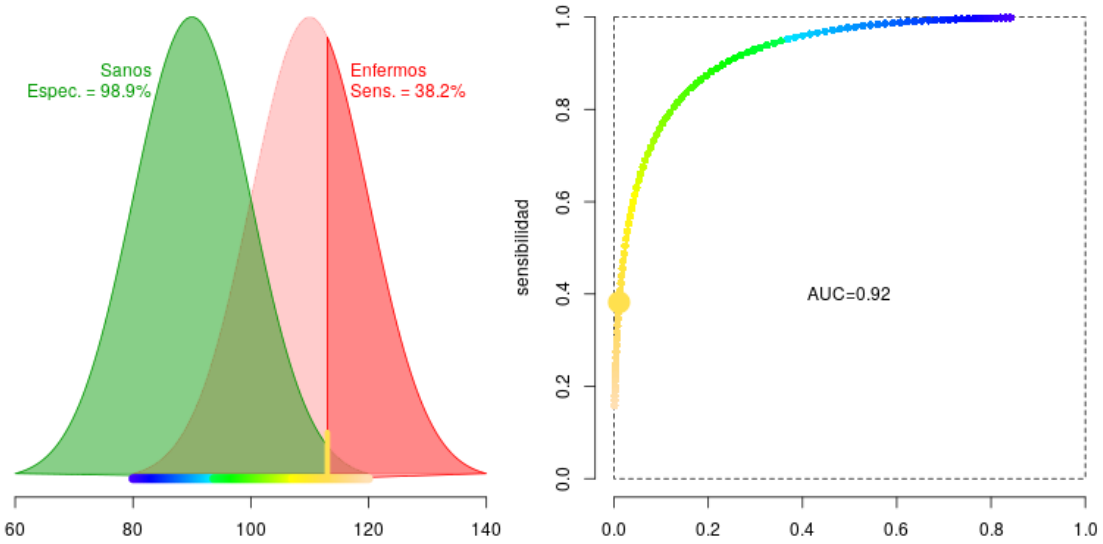
*Figura 8. Distribución de frecuencias y AUC con punto de corte de alta sensibilidad*




*Nota*. El punto de corte fijado implica que el test tenga elevada capacidad para detectar positivos, pero nula capacidad para detectar negativos. Figura de elaboración propia con programa informático. Fuente: (Barón López, 2022).

En el caso en que el punto de corte se modifique, manteniendo constantes los demás datos, tanto la sensibilidad como la especificidad cambiarán (Figura 9. Distribución de frecuencias y AUC con punto de corte de alta especificidad).

*Figura 9. Distribución de frecuencias y AUC con punto de corte de alta especificidad*

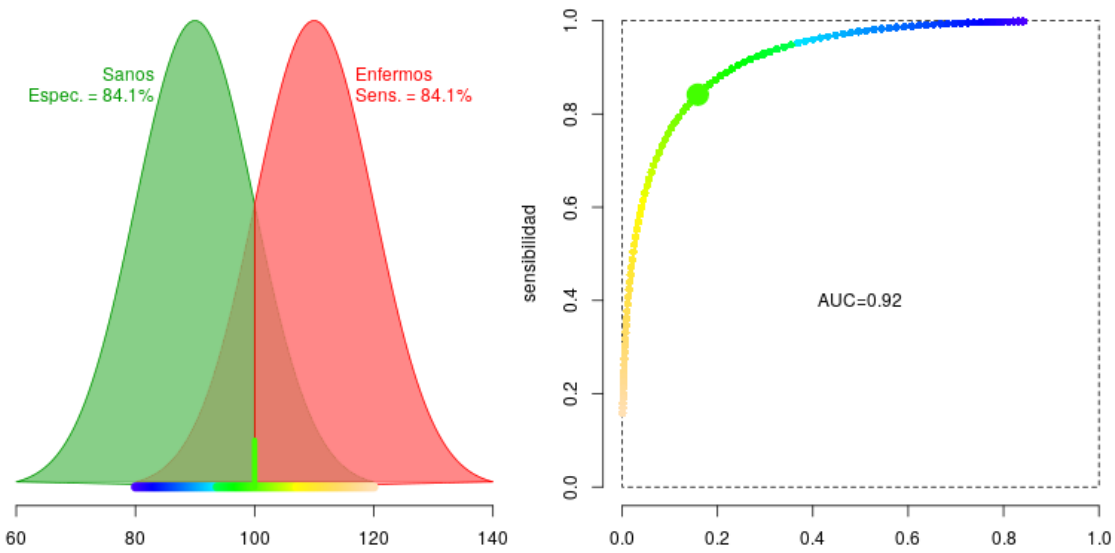



*Nota.* El punto de corte se ha situado en 113. Figura de elaboración propia con programa informático. Fuente: (Barón López, 2022).

Dado que el punto de corte se ha modificado, se ha reducido el espacio destinado a clasificar correctamente a los positivos, aumentando consecuentemente el error tipo II.

De manera visual, un punto de corte adecuado, dándole la misma importancia a los dos errores, un punto equilibrado es aquel en el que la línea tangente que corte la curva ROC tenga 45º o se sitúe en la parte más izquierda posible a la vez que se sitúa en el punto más elevado posible. Con estas instrucciones se genera la figura (Figura 10. Distribución de frecuencias y AUC con punto de corte maximizado).

*Figura 10. Distribución de frecuencias y AUC con punto de corte maximizado*




Nota. El punto de corte se ha situado en 100. Figura de elaboración propia con programa informático. Fuente: (Barón López, 2022).

En este caso el punto de corte se encuentra situado en el punto closest-top-left optimizado.

No obstante se siguen cometiendo errores ya que las campanas se superponen, lo que implica la existencia de errores tipo I y tipo II. Para estimar la probabilidad de estos errores

*Tabla 7. Ejemplo de clasificación de los sujetos curva ROC con punto de corte closest-top-left*

| Resultado del test | Gold Standard | Frecuencia |
|---|---|---|
| Positivo | Enfermo | 84 |
| Negativo | Enfermo | 16 |
| Positivo | Sano | 16 |
| Negativo | Sano | 84 |

Tabla de elaboración propia.

Esto implica que los correctamente clasificados tanto de enfermos como de sanos coincida en 84% ($Sensibilidad = Especificidad = 0{,}84$), mientras que los errores también son coincidentes en su complementario ($Error\ tipo\ I = Error\ tipo\ II = 0{,}16$). Dicho de otra forma: el test tiene una capacidad de clasificar correctamente tanto los positivos como los negativos del 84%. De entre las personas enfermas, el test ha clasificado correctamente al 84%

de ellas. Como el porcentaje es el mismo, puede afirmarse que de entre las personas sanas, el test ha clasificado correctamente al 84% de ellas. Por otro lado se tiene una probabilidad de no clasificar correctamente al 16% de los sujetos.

En determinados contextos, como una prueba de detección de Virus de la Inmunodeficiencia Humana, se pretende extremar la sensibilidad, detectando anticuerpos cuando están presentes. Esta sensibilidad se fija en torno al 99%. Esto implica que la especificidad se reduzca, aumentando los falsos positivos (García et al., 2011). Por lo que ante un resultado positivo, este debe ser corroborado por otro test de mayor especificidad. Teniendo así un test de cribado con alta sensibilidad y un test confirmatorio con alta especificidad (Buttò et al., 2010). Entonces la calidad diagnóstica de los test viene determinada por la selección cuidadosa de los puntos de corte y sobre la calidad del propio test.

También hay que tener en cuenta de que, aunque el test haya funcionado correctamente y haya conseguido detectar lo que pretende detectar, las implicaciones que se le dan, no siempre son las correctas. Supongamos que una muestra biológica de orina interacciona con un test de embarazo. En el caso de que el resultado del test sea positivo se puede llegar a la conclusión de que el sujeto del que se extrajo la muestra de orina tiene la condición de embarazo. En circunstancias normales estas asociaciones con adecuadas, aunque no siempre se da el caso y no es debido a un falso positivo, sino que el test efectivamente detectó la sustancia para la cual fue diseñado, solo que la condición de embarazo no es aplicable a un humano de género masculino. Esto es lo que ocurre con la detección de cáncer de testículos en varones, ya que el test detecta una hormona producida por el embrión y la placenta, en el embarazo (Carstairs, 2013). Esta hormona es la Gonadotropina Coriónica Humana. No obstante, esta hormona es también emitida por las células tumorales de los testículos. Entonces se da el caso, que, aunque sea un verdadero positivo del test, en función del contexto, pueden llegarse a unas conclusiones o a otras muy diferentes.

#### *2.4.6.1. Índice de Youden*

También llamado Versión 2 de probabilidad corregida de detectar enfermedad. Con este índice se compara la diferencia entre la tasa de verdaderos positivos con la de falsos positivos, estimándose que un buen test debe tener una elevada diferencia. La unidad de este índice representa que la prueba discierne de forma perfecta y valores cercanos a la unidad se consideran mejores índices, mientras cuanto más cercano a cero sea el índice, indica una peor capacidad discriminante (Youden, 1950). Su fórmula es la siguiente:

$$Índice\ de\ Youden = Sensibilidad + Especificidad - 1$$

El procedimiento consiste en que se aplica esta fórmula a cada punto de corte, seleccionando aquel punto de corte cuyo Índice de Youden sea más cercano a 1.

Como la propiedad buscada es que se tenga una alta sensibilidad y especificidad puede realizarse este método, teniendo en cuenta su limitación de que se estaría escogiendo un punto de corte que, teniendo en cuenta ambos indicadores, seleccionase un punto dando el mismo peso a ambos indicadores.

Este estadístico se centra en la selección de un punto de corte basándose en los indicadores ya dados de sensibilidad y especificidad, por lo tanto, previamente a la aplicación del índice de Youden debe explorarse que la prueba sea adecuada en términos de Sensibilidad y especificidad de forma conjunta.

En cambio, en el caso de desear que exista una discrepancia de pesos y por tanto una prueba cuyo punto de corte tenga mayor precisión en uno de los dos indicadores (especificidad o sensibilidad), el índice de Youden no estaría contemplando esta circunstancia y por tanto otro índice deberá ser elegido.

#### 2.4.6.2. Closest top left

Aplicándolo de forma similar al Índice de Youden estima como punto de corte óptimo aquel que se sitúa en la curva ROC en la parte izquierda más elevada. Este índice toma como punto de corte más adecuado aquel que tenga un valor menor (Hanley & McNeil, 1982). Para calcularlo se aplica la siguiente fórmula:

$$Closest\ top\ left = \sqrt{(1-S)^2 + (1-E)^2}$$

Los conceptos de Sensibilidad, Especificidad, así como los estadísticos Índice de Youden y Closest top left mejoran su comprensión si se expresan gráficamente.

## 2.5. Modelos de Ecuaciones Estructurales

Los Modelos de Ecuaciones Estructurales (SEM) son aproximaciones estadísticas a la comprobación de las hipótesis conceptuales entre relaciones de variables observadas y latentes con respecto a la realidad. Esto quiere decir que primeramente se parte de una teoría y posteriormente se evalúa si la teoría se adapta a la realidad de los datos. Las relaciones teóricas se refieren a cómo debería comportarse el constructo, en nuestro caso: cómo debería comportarse el instrumento de medida. Si conceptualmente se ha diseñado el instrumento para desvelar molestias físicas y de ahí se han generado ítems que se refieren a las molestias físicas que puedan ser producidas por la depresión; si seguidamente ha ocurrido lo mismo con dimensiones cognitivas, con las implicaciones de la depresión, con las emociones producidas, u otras. O en cambio no pueden establecerse distinciones lo suficientemente fuertes como para establecer subgrupos de dimensiones entre los ítems del test. Ésta es la teoría que pone a prueba el SEM y la contrasta con la realidad de los datos: ¿existen grupúsculos de ítems que se relacionen más entre sí que entre otros? De ser así, podríamos estar ante una dimensión o también llamado constructo latente. Porque la depresión entendida desde el SEM es un constructo, así como las dimensiones somáticas y emocionales. Porque ¿qué comparten en común los ítems de: me duele la cabeza y siento excesivo cansancio? La tarea de abstracción y conocer qué tienen en común o a qué categoría pueden pertenecer estos ítems, corresponde al investigador; el cual, después de una adecuada comprensión del problema y tras un adecuado análisis de la literatura al respecto, podrá nombrar la agrupación más adecuada que pueda definir la categoría en la que se engloban estos dos ítems. Y la tarea se vuelve más difícil cuando se torna evidente la comunalidad de los ítems, pero uno de ellos trastoca los esquemas, obligando a reformular la dimensión para que de forma efectiva se pueda incluir el nuevo ítem a la dimensión y no quede desdefinido tras su inclusión. Para ejemplificar este caso, supongamos que tenemos 3 ítems: me sudan las manos, me duele la cabeza y me suicidaría si tuviera la oportunidad. Los dos primeros ítems corresponden a consecuencias

físicas de malestar, mientras que el tercer ítem corresponde a un razonamiento con intencionalidad. Aunque la categoría de correspondencia de los 2 ítems parezca clara, si los datos sugieren que existe una sola categoría que englobe los 3 ítems, debería generarse una dimensión tal que pudiera satisfacer a los 3 ítems de forma que no cupiesen dudas en la inclusión de cualesquiera ítems en esta dimensión.

El SEM se compone de la especificación del modelo, que es la fase por la cual se detallan las relaciones entre variables observadas y latentes. En esta fase se sirve de la teoría para conocer de cuántas variables latentes debería estar compuesto el modelo, así como sus relaciones. Éste es un proceso delicado, ya que desde la compresión de las variables desde la teoría puede tener sentido que las relaciones entre variables sean en un sentido, cuando tras ser revisado por otro prisma, las relaciones cambien. Este proceso implica tanto conocer las relaciones del modelo propuesto como sus alternativas y retractores, ya que, tras comprobar un ajuste pobre, sería lógico reinterpretar el modelo desde una segunda perspectiva y así sucesivamente. De esta forma se trataría de encontrar un modelo particular cuyas variables observadas encajen en los campos dando una solución interpretable y ofreciendo una interpretación plausible de estas variables observadas.

Una vez detallado el modelo se calculan las estimaciones de los parámetros libres del conjunto de datos observados. Para ello se pueden usar técnicas de ANOVA o diseños de regresión múltiple. Se utilizan métodos iterativos que involucran una serie de pruebas para obtener los estimadores de parámetros libres que implican una matriz de covarianzas como la observada. Después de cada iteración se comparan la matriz de covarianzas con la matriz de observaciones para obtener una matriz residual. Las iteraciones se suceden hasta que la matriz residual no pueda minimizarse más, esto es lo que se conoce como convergencia. Cuando se ha llegado a una solución, cuando se ha convergido se emite un valor que indica el valor de

ajuste de la función. Cuando este valor se acerca a 0 quiere decir que el error es mínimo y cuanto más alejado de 0, más grande será este error. A esto se le llama el ajuste del modelo.

Cuando un modelo está ajustado quiere decir que encaja con los datos. Lo que implica que los datos encajan con el modelo. Para evaluar la calidad de ajuste, lo bien que encaja uno con el otro, se suele utilizar el índice $\chi^2$ de prueba de bondad de ajuste. Esta se calcula mediante el producto del valor de la función de ajuste y el tamaño de la muestra menos 1.

$$F(N-1)$$

No obstante, debería llamarse maldad de ajuste, ya que, cuanto mayor sea $\chi^2$, peor ajuste tendrá y mientras menor sea su valor, mejor ajuste tendrá. De todas formas se ha de calcular conforme a sus grados de libertad y obtener un p-valor que indique el error dispuesto a asumir para considerar aceptable la discrepancia entre el modelo teórico y el de los datos. Por otro lado, el SEM ofrece otros parámetros de ajuste, además del $\chi^2$ pero estos índices de ajuste no poseen puntos de corte estadísticos, ya que no definen valores críticos (Rick H. Hoyle, 2023). Esto implica que los valores resultantes sean orientativos al carecer de decisiones sobre valores que indiquen lo aceptable o no del ajuste del modelo.

Posteriormente y tras analizados los índices de ajuste del modelo se evalúa si se debería o no realizar modificaciones al modelo teórico, por ejemplo si se ha comprobado que 2 factores en vez de uno, explican mejor la relación entre los ítems, sería recomendable considerar modificar o reconsiderar el modelo a 2 variables latentes confiando que el ajuste del modelo será mayor; aunque otros cambios pueden ser realizados, no sin controversia al respecto de estas modificaciones (MacCallum et al., 1992).

Para interpretar los resultados obtenidos primeramente se puede utilizar el indicador de $\chi^2$, que es útil para evaluar el ajuste general al modelo, pero también se ofrecen otros índices que otorgan información específica del ajuste al modelo. Frecuentemente se detallan los parámetros estandarizados o no estandarizados. Los primeros solo pueden interpretarse

con un escalamiento de referencia sus las variables. De esta forma se indican las unidades de cambio que se dan en la variable dependiente por cada unidad de cambio en la variable independiente cuando el resto de variables se mantienen inalterables. Mientras que los parámetros estandarizados comparten la misma unidad de medida, mejorando la interpretabilidad y comparación entre ellos; así, las estimaciones de parámetros estandarizados corresponden a estimaciones del tamaño del efecto.

La información que ofrece un diagrama SEM es una serie de relaciones entre variables observadas y latentes que configuran la estructura de modelo de ecuaciones. Se compone de rectángulos, que representan a las variables observadas; elipses, que representan a variables latentes como los errores o variables dependientes o independientes; y flechas, que indican la asociación entre combinaciones de los elementos anteriores.

La ventaja principal del SEM frente a otros métodos como el ANOVA o loa regresión múltiple para probar hipótesis causales radica en que se consideran las condiciones necesarias para demostrar una relación causal: la asociación, la direccionalidad y el aislamiento (Bollen, 1989). Para demostrar la asociación la causa y el efecto deben estar relacionados. De esta forma el SEM no implica una mejora frente a otras técnicas. Para demostrar el aislamiento de otras causas que produzcan el efecto, como las variables extrañas o variables confusas, se realizan una asignación aleatoria de los niveles de la variable causal. Otras formas de realizar este aislamiento de variables son la correlación parcial, el ANOVA y la regresión múltiple; estas técnicas pueden usarse para aislar relaciones causales de otras variables; pero el SEM es más flexible y comprensivo que otras aproximaciones, proporcionando aproximaciones mediante el control de variables extrañas, confusas y midiendo el error (Rick H. Hoyle, 2023). Para demostrar la direccionalidad se torna más difícil, porque lo que se puede comprobar es la relación entre variables, y no qué variable preludió a cuál, por lo que es necesario el uso de la teoría para interpretar los resultados.

Para realizar los SEM pueden usarse los programas como AMOS (Arbuckle, 2016) o LISREL (Scientific Software International, 2023).

### 2.5.1. Invarianza de medida Multigrupo (MGCFA)

El Análisis Factorial Confirmatorio Multi Grupo (o MGCFA por sus siglas en inglés), es una técnica de modelización basado en la covarianza que prueba la heterogeneidad en un modelo de medida mediante un Análisis Factorial Confirmatorio. La principal diferencia de AFC y MGCFA radica en que en el segundo, se parte la muestra en dos o más submuestras para comprobar el modelo. También es llamado prueba de invarianza de medida o prueba multigrupo de invarianza de medida. Su hipótesis nula indica que no existen diferencias entre dos o más grupos (Angell, 2019).

Para llevar a cabo un MGCFA primero se debe establecer el modelo base y comprobar su invarianza parcial. Esto implica que el modelo que ha sido probado previamente para la muestra completa también es evaluado su ajuste para cada submuestra, comprobando si el modelo es válido para cada subgrupo de la muestra. En este paso es posible que el modelo se ajuste para una submuestra, pero no para otra. Posteriormente se establece la invarianza configuracional, esta se realiza solo cuando existe un ajuste del modelo con las mismas variables latentes para cada subgrupo. Por último se comprueba si el patrón de variables observadas explicadas por un factor es idéntico entre los grupos.

La invarianza se basa en la prueba de diferencias de $\chi^2$ ($\Delta\chi^2$), que representa el cambio de valores de $\chi^2$ que se dan al evaluar la métrica del modelo. El resultado da una $\chi^2$ con sus respectivos grados de libertad. Su H0 implica que no existen diferencias en el ajuste del modelo cuando las cargas factoriales están restringidas para que sean consideradas iguales entre los grupos.

# Capítulo 3

## Salud mental

*Qué notable desconcierto, tener celos de un muerto.*

*Lope de Vega, 1625*

*Los que no han sufrido no saben nada; no conocen ni el bien ni el mal; ni conocen a los hombres ni se conocen a sí mismos.*

*François de Salignac de la Mothe*

# 3.-Salud mental

Es importante señalar que la salud mental puede verse disminuida por muchas alteraciones diferentes a la de la depresión, por lo que hablar de depresión no es sinónimo que hablar de algún grado de menoscabo en salud mental.

## 3.1. Depresión

Uno de los rasgos medidos en esta investigación es la depresión. La depresión se clasifica como una alteración del estado de ánimo. Esta clasificación puede hacerse desde varias vertientes, pero los manuales más aceptados han sido el de Criterios Diagnósticos de Investigación (RDS), la Clasificación Internacional de Enfermedades (CIE) y el *Diagnostic and Statistical Manual of Mental* Disorders (DSM).

Dentro de los estados de ánimo, el DSM en su versión quinta clasifica la depresión con 7 tipologías: trastorno depresivo mayor, distimia, trastorno de desregulación disruptiva del estado de ánimo, trastorno disfórico premenstrual, trastorno depresivo debido a otra afección médica, trastorno depresivo inducido por sustancias o medicamentos y otros trastornos depresivos especificados o no especificados (Asociación Americana de Psiquiatría, 2014).

Aunque la depresión esté bien definida como una entidad, es común que se encuentre presente junto a otras patologías como esquizofrenia, trastornos de alimentación, trastornos de síntomas somáticos, disfunciones sexuales y disforia de género. Además se encuentra más entre las personas que también sufren de ansiedad, trastorno obsesivo-compulsivo y trastorno por estrés postraumático (Morrison & James, 2014).

El DSM-5 (Asociación Americana de Psiquiatría, 2014) mide la depresión como un trastorno del estado de ánimo que tiene como rasgo más común la presencia de un ánimo triste, vacío o irritable. Si bien es cierto que existe variabilidad en cuanto a la expresión de la

depresión, en lo que acontece a este estudio, se hará hincapié en el Trastorno Depresivo Mayor.

La depresión es un trastorno mental. Esto quiere decir que es un síndrome con relevancia clínica. O lo que es lo mismo: una colección de síntomas que causan a la persona que lo padece discapacidad o malestar en el desempeño social, personal o laboral (Morrison & James, 2014). Entonces evaluar a una persona con depresión (a expensas de un diagnóstico diferencial que confirme la patología) implica que la persona ve alterada de manera significativa su vida debido a esta condición. Si bien es cierto que el grado de severidad, así como el tiempo que lleve con la enfermedad, así como las causas pueden hacer que el concepto de depresión sea un término global cuya vivencia sea muy diferente entre quienes lo padecen.

Con el objetivo de operativizar y poder trabajar con el concepto de depresión, primero en este trabajo se adscribirá al concepto de depresión como trastorno definido en el DSM-5. Por otro lado, el trastorno de depresión mayor será el concepto a utilizar, dada su gravedad. Con independencia de que la gravedad pueda ser puntual o persistente en el tiempo. Esto es porque en el caso de que sea persistente en el tiempo se trataría de una depresión crónica o también llamado trastorno distímico, en que por lo general su duración se prolonga durante años, aunque es común que esta afectación sea relativamente leve. Se ha codificado al trastorno depresivo mayor de episodio único con F32, mientras que el trastorno depresivo mayor de episodio recurrente se codifica con F33, mientras que la distimia se clasifica como un trastorno del estado de ánimo persistente con el código F34.1. Cada uno de los trastornos tiene subtipos como el F32.0, F32.1 y F32.2 que se refieren a la gravedad, siendo Trastorno depresivo mayor leve, moderado y grave, respectivamente. También tanto para episodio único como para recurrentes, el trastorno depresivo se puede catalogar como con características psicóticas, en remisión parcial o total y por último el trastorno depresivo no especificado. En

este último se engloban características no recogidas en las anteriores, pero referidas a depresión.

Para poder clasificar un trastorno depresivo mayor se deben descartar síntomas maníacos (episodios de alegría anómala) ya que en ese caso se estaría hablando de un trastorno bipolar (antes llamado enfermedad maníacodepresiva).

La forma que tiene el DSM-5 de identificar un trastorno es por la presencia de un listado de síntomas. Si bien el listado puede referirse a la definición amplia de los constructos que se deseen evaluar, es posible que la persona evaluada no deba cumplir todos para poder cumplir con el diagnóstico, sino que al satisfacerse un número determinado de síntomas, podría realizarse el diagnóstico como tal.

De esta forma, el DSM-5 (Asociación Americana de Psiquiatría, 2014) indica que para poder diagnosticar a una persona de trastorno depresivo mayor se deben cumplir 5 o más criterios, demarcándose temporalmente en las 2 semanas previas. Además de forma necesaria uno de los síntomas ha de ser un estado de ánimo deprimido y el otro una pérdida de interés o de placer. Los síntomas han de causar malestar clínicamente significativo; esto es que causen malestar en los ámbitos social, laboral u otras áreas importantes. Además los síntomas no deben ser causados por los efectos fisiológicos de medicamentos o de otras enfermedades médicas.

Los síntomas a los que se refiere anteriormente para diagnosticar depresión son los siguientes:

1. Encontrarse deprimido la mayor parte del día y además durante casi todos los días de las dos últimas semanas. Además, se puntualiza que en niños y adolescentes la irritabilidad y la depresión pueden ser intercambiables.
2. Disminución del interés o placer por hacer actividades. Esta falta de interés debe ser importante, englobar la mayoría de las actividades, estar presente en la mayor parte

del día y estar presente la mayor parte de los días de las dos últimas semanas. En la depresión casi siempre ocurre una pérdida de interés o de placer. Esto es que no se siente placer por actividades que fueron placenteras en algún momento, que no se siente ilusión por festividades, pasatiempos. También es posible disminuir el interés por el sexo o deseo sexual. Se registra mayor aislamiento social y abandono de actividades que se disfrutaban.

Estos dos primeros síntomas son centrales en el trastorno depresivo mayor. Pero además deben satisfacerse síntomas adicionales que causen significativo malestar y sean nuevos o que hayan empeorado de forma clara. Además debe implicar malestar clínicamente significativo en las áreas social, laboral o de estudios u otras áreas importantes de la vida. En trastorno de depresión mayor, en un estadio leve, es posible que las actividades puedan realizarse, aunque implique un alto esfuerzo.

3. Cambio en brusco de peso. Esto es aumentar o disminuir el peso un 5% del peso normal en un mes (unos 3,5kg para una persona de 70kg). Además, la presencia de pérdida o aumento de apetito durante dos semanas también puede ser intercambiable con el cambio brusco de peso. En algunas personas simplemente no les apetece comer careciendo de apetito, mientras que en otras sienten ansia por determinadas comidas. Esto conlleva aumento o pérdida significativa de peso; sin embargo, ya que los niños necesitan crecer, el hecho de no aumentar el peso que se supone que debería de ganar, ya se considera significativo y tomado en cuenta como síntoma depresivo.
4. Tanto sueño excesivo como pérdida de la calidad del sueño. El sueño excesivo es poco frecuente y se manifiesta en un cansancio con somnolencia excesiva durante el día. Mientras que las dificultades para dormir implican conciliar el sueño, mantenerlo y despertarse antes de tiempo. Es común que el motivo de solicitar tratamiento sean las

alteraciones del sueño, ignorando la depresión, aunque la depresión sea el diagnóstico primario.

5. Agitación o retraso psicomotor. Debe ser observable por otras personas y ocurrir casi todos los días. La agitación es más propia de una depresión atípica, englobándose en la categoría de otros trastornos depresivos no especificados. Pero en realidad se trata de un espectro que puede estar alterado, estando en este caso tanto agitado como retrasado en el ámbito psicomotor. Esto es enlentecimiento, falto de reflejos, conversación pausada o en su polo opuesto: una agitación de tensión, nerviosismo, inquietud o dificultad para concentrarse. Como en otros síntomas, este debe ser clínicamente significativo y ya que se puede comprobar fácilmente: estas alteraciones deben ser lo suficientemente graves como para ser observables por otras personas.
6. Fatiga o pérdida de energía casi todos los días. Esto quiere decir que la persona siente un cansancio excesivo conllevando un esfuerzo considerable el realizar tareas cotidianas. Este esfuerzo puede ser extenuante y la tarea a realizar, requerir más tiempo del habitual.
7. Sentirse inútil, excesivamente culpable o culpable sin realmente serlo. Esto debe ocurrir casi todos los días y no ser simplemente autorreproche o autoculpabilidad por estar enfermo. Puede implicar no sentirse merecedor de lo que se tiene, la reevaluación de pensamientos o ideas que causan malestar, llevado de forma pesimista, con culpabilidad o minusvaloración (rumiación). De un acontecimiento originalmente neutro se le puede dar una interpretación negativa minusvalorándose, magnificando los errores o defectos, atribuyéndose autoría de estos hechos y consecuentemente atribuyéndose los efectos negativos.
8. Disminución de la capacidad de concentración, de pensar adecuadamente o de tomar decisiones, aunque sean pequeñas decisiones. Pueden incluir problemas de atención y de memoria. Incapacidad para desempeñar tareas de gran demanda cognitiva. Esto

debe ocurrir todos o casi todos los días. En niños o adolescentes puede reflejarse en un descenso abrupto de las notas. En ancianos puede confundirse con demencia o conllevar el inicio de una demencia.

9. Pensamientos recurrentes de: muerte, ideas suicidas o intentos de suicidio con o sin plan para llevarlo a cabo. Estas ideas de muerte han de ser diferentes al miedo a morir, sino tomando la muerte como algo atractivo o como escape al sufrimiento. La idea de morir es resultado de una profunda evaluación de su estado actual y la proyección de futuro en la que la desesperación limita las opciones sobre la desesperanza ante el futuro, haciendo el suicidio como una salida atractiva. Es un método de escape ante la situación que envuelve al sujeto y de la que sus recursos no son suficientes para afrontar la situación. Este pensamiento puede venir del deseo de terminar un dolor emocional terriblemente doloroso, incapacidad para imaginar un futuro diferente al que provoca dolor. También puede venir acentuado por sentirse una carga para los seres queridos, entendiéndose como, tras una profunda valoración, estimar que la situación de los seres queridos mejorará tras la muerte del paciente.

Sin embargo, el hecho de que una persona cumpla con cinco o más de alguno de los criterios arriba expuestos no tiene por qué considerarse un diagnóstico de depresión mayor; sino que, se tiene que haber descartado que estos síntomas puedan explicarse por el uso de sustancias o enfermedades o afecciones médicas. Algunas enfermedades o afecciones médicas, debido a su naturaleza, comparten sintomatología como operaciones bariátricas de pérdida brusca de peso.

Se debe descartar también un episodio maníaco (o hipomaníaco). Además se deben descartar el diagnóstico de un trastorno esquizoafectivo, esquizofrenia, un trastorno esquizofreniforme, un trastorno delirante u otro trastorno especificado o no especificado del espectro de la esquizofrenia y otros trastornos psicóticos.

Por otro lado estos síntomas no deben ser una respuesta adecuada a una pérdida significativa de alguna cualidad de la vida de las personas, ya que estos síntomas pueden llegar a ser comprensibles ante una pérdida (como ruina económica, catástrofes naturales u otros). No obstante el considerarse una respuesta normal o anormal ante pérdidas significativas se debe valorar en entrevista, junto con historia clínica, así como el contexto y el sentido de estas pérdidas.

Para realizar un diagnóstico de trastorno depresivo mayor pueden tomarse datos no solamente de autoinforme, sino de las expresiones y conductas que presente el paciente. Incluso reconocer estos sentimientos puede resultar difícil para el propio paciente o incluso manifestar dolores corporales en lugar de sentimientos de tristeza. Manifestaciones exageradas de irritabilidad, ira persistente, culpabilizar a otros, un exagerado sentimiento de frustración; todo ello desencadenado por acontecimientos menores, también son manifestaciones de sentimientos que se deben tener en cuenta en la entrevista diagnóstica. Cuando se cronifican los síntomas es común que se desarrollen trastornos de personalidad, ansiedad y consumo de sustancias.

El trastorno depresivo mayor tiene una prevalencia alrededor del 7% en los Estados Unidos. Siendo las mujeres quienes presentan entre 1,5 a 3 veces más este trastorno que los varones. Además, en el grupo de edad de los 18 a los 29 años se presenta 3 veces más que en edades de 60 años o más.

La prevalencia del trastorno depresivo mayor tiene diferencias de hasta 7 veces más en unas culturas que en otras. Pero se mantiene en proporciones similares en función del sexo y la edad. Aunque no implica que se hayan podido establecer patrones entre culturas. En atención primaria no suele ser motivo inicial de asistencia, sino que se refiere más bien a síntomas somáticos como el insomnio o la pérdida de energía. En mujeres y hombres no hay diferencias claras en cuanto a la sintomatología presentada, en cambio el intento de suicidio

de las mujeres es mayor, mientras en los varones los suicidios consumados son mayores. Sin embargo, entre quienes padecen trastorno depresivo mayor, estas diferencias por sexo aminoran.

### 3.1.1. Patient Health Questionnaire

#### *3.1.1.1. PHQ-9*

El test PRIME-MD (Spitzer et al., 1994) se diseñó para detectar cinco trastornos más comunes: depresión, ansiedad, trastorno somatomorfo, trastornos por uso de alcohol y trastornos alimenticios. Se trata de un test de cribado junto con una entrevista: se les pedía a los sujetos que rellenasen un test autoadministrado y en los trastornos que dieran positivo, se les aplicaba una entrevista estructurada sobre esos trastornos. Esto implicaba unos 5 minutos para los sujetos que no se les detectaban patologías clínicas, mientras que implicaba unos 12 minutos para sujetos con que padecían de alguno de los trastornos estudiados. De este cuestionario, se desarrolló el PHQ y su familia de test (Spitzer, 1999; Spitzer et al., 2000). Siendo el trastorno de ansiedad, los trastornos somatomorfos, y la depresión como más atractivos por los investigadores sobre todo por tratarse de test autoadministrados. Esta familia de test, al resultar del estudio de estas principales 5 patologías más comunes, se disgregaron en test específicos para cada una de las patologías, creando módulos. Entonces se crearon subtest que evaluarían constructos diferentes y que a la vez son complementarios; por lo que pueden ser usados aislados, por módulos y como parte del PHQ completo(Kroenke et al., 2001; Kroenke & Spitzer, 2002; Löwe et al., 2004). Entonces se disgregó en el test de depresión PHQ-9 y el test de ansiedad GAD-7 (Spitzer et al., 2006). Por otro lado la relación entre depresión y ansiedad y las quejas somáticas, al ser especialmente comunes estos tres constructos, se pudo recoger mediante la herramienta PHQ-SADS (la triada ADS, Ansiedad, Depresión y Somático) (Löwe et al., 2008).

En este estudio, para evaluar la depresión se utilizó la escala PHQ-9 (Patient Health Questionnaire) (Kroenke et al., 2001). Esta escala es una versión del test Primary Care Evaluation of Mental Heath Disorders (PRIME-MD) (Spitzer et al., 1994) basada en el DSM-IV (Bogduk, 2013) y sus ítems siguen la ordenación original que la que aparecen los síntomas especificados en dicha versión del manual. Entre las versiones del DSM 4 al 5 no aparecen apenas cambios en lo que concierne al trastorno depresivo mayor, salvo la ordenación de los síntomas, que coincide con la ordenación de los ítems del PHQ-9.

El test PHQ-9 es un test de depresión en el que, al no haber diferencias importantes en cuanto a la depresión, se considera que está bien definido por la parte del DSM-5 que trata sobre el trastorno depresivo mayor. En el DSM-5 se especifica que se deben cumplir 5 de los 9 síntomas expuestos, junto con características como que deben estar presentes alguno de los dos primeros ítems, que se deben descartar otras patologías, que se refieren a las dos últimas semanas, así como que el malestar causado debe ser significativo y no presentes como una respuesta natural a acontecimientos adversos. Estas características se recogen con laxitud en la redacción de los ítems del PHQ-9, ya que se trata de un test de cribado y la comprensión prima sobre la complejidad del trastorno. Además la característica de tamiz hace necesaria la intervención de un sanitario para poder realizar de forma efectiva un diagnóstico.

Los ítems del PHQ-9 se han ordenado de tal forma que los dos primeros, sobre anhedonia y desesperanza, constituyan el estado nuclear de la depresión. De forma extremadamente reducida se ha conseguido condensar estos dos ítems para poder detectar la depresión, conformándose el test PHQ-2 (Inagaki et al., 2013; Mitchell et al., 2016; Richardson et al., 2010). De igual forma el PHQ-8 se trata de los primeros 8 ítems del PHQ, no incluyendo el ítem de suicidio. En el caso en que ante el ítem 9 se haya dado positivo, esto es que exista cualquier puntuación mayor de 0 en este ítem, se recomienda ahondar más en este constructo con un cuestionario aledaño de 4 ítems para evaluar el riesgo autolítico: las 4 P (Pasados

intentos de suicidio, Plan, Probabilidad de realizar el suicidio y factores Protectores) (Dube et al., 2010). Es importante comprender que se hace referencia a la forma en que se recoge la expresión de suicidio tratándose de la frecuencia de ocurrencia de estos pensamientos. Su importancia radica en que tratándose de un pensamiento repetido de autolisis, se tiene como un pensamiento recurrente y no pasajero. Aunque la predicción del suicidio es muy baja, solo el 5% de quienes lo manifiestan lo cometen. Para que se dé el suicidio tienen que darse dos características. La primera es la existencia de capacidad cognitiva de darse cuenta de que el suicidio es una salida; mientras que la segunda es tener problemas.

En el PHQ-9 se ha estudiado que posee una estructura bifactorial siendo los ítems 1, 2, 6, 7, 8 y 9 los que forman el factor cognitivo-afectivo, mientras que los ítems 3, 4 y 5 los que forman el factor somático (Chilcot et al., 2013; Teymoori et al., 2020), aunque estos estudios no son consistentes (Merz et al., 2011). Se trata de una escala de la Teoría Clásica de los Test que ha sido ideada para medir depresión y cuyas opciones de respuesta son las frecuencias sobre la molestia que le producen determinados eventos en un periodo de los últimos 14 días; siendo estas opciones de respuesta: para nada, varios días, más de la mitad de los días y casi todos los días, cada una tomando valores de 0 a 3. En la versión original se propusieron los puntos de corte de 5, 10, 15 y 20. Quedando una clasificación como sigue: depresión mínima 0-4; depresión media 5-9; depresión moderada 10-14; depresión de moderada a severa 15-19 y depresión severa 20-27.

En cuanto a la cesión de su uso, en el propio test se especifica que “No se necesita autorización para su reproducción, traducción, muestra o distribución”.

Además de los 9 ítems, el PHQ añade un ítem extra el cual se responde únicamente en el caso en que se hubieran señalado la presencia de cualquier problema. En concreto se pregunta sobre el grado de dificultad que causaron en las áreas: laboral, cuidado del hogar o convivencia con otras personas.

#### *3.1.1.2. Test nucleares del PRIME-MD*

Como se ha explicado anteriormente, el test Prime-MD se constituye como el predecesor del PHQ. Este test combina un test de cribado autoadministrado con preguntas de entrevista clínica (Spitzer et al., 1994). De este se deriva el PHQ, que tiene cinco módulos que cubren 5 tipos comunes de trastornos mentales. Estos son: depresión, ansiedad, somatomorfo, alcohol y trastornos de la alimentación. Provisionalmente se han seleccionado diagnósticos para todos los tipos de trastornos, con excepción a los somatomorfos, según el DSM-IV (Spitzer, 1999; Spitzer et al., 2000). Posteriormente, el test PHQ-9 es en realidad la subescala de depresión del test PHQ original. Cada uno de sus 9 ítems se puntúa del 0 al 3, teniendo una puntuación máxima y mínima de 0 a 27 (Kroenke et al., 2001, 2010; Kroenke & Spitzer, 2002; Löwe et al., 2004; Spitzer et al., 1994).

Dado que entre los módulos del PHQ original se encuentra uno dedicado a la ansiedad, se considera el test GAD-7 como uno de estos módulos y será contemplado en un epígrafe diferente.

También se creó el PHQ-15 que derivó de los estudios PHQ originales y es cada vez más utilizado para evaluar los síntomas somáticos (Kroenke et al., 2002, 2010). El test PHQ-15 se trata de una versión más extensa, ideada para evaluar la severidad y la presencia potencial de la somatización y los desórdenes somatomorfos de forma más específica. A los ítems del test le acompañan otros ítems adicionales que no se utilizan para calcular la puntuación del test, sino que representa la impresión del paciente sobre cómo la depresión afecta a las áreas de su vida. Su utilidad radica en adecuar el tratamiento dado el deterioro que puede causar la depresión a su calidad de vida.

Por último, entre los test primarios del PHQ se encuentran el PHQ-SADS, que incluye el PHQ-9, el GAD-7 y el PHQ-15 junto con la parte de pánico del PHQ original (Kroenke et al., 2010). Al tratarse de un test combinado, se corrige como sus partes de forma aislada.

#### *3.1.1.3. Variantes del PHQ*

El test Brief PHQ (Spitzer et al., 2000) se trata de un test enfocado a las mujeres. Se compone de las escalas PHQ-9 y la escala de pánico del PHQ original. Además se le agregan ítems sobre factores estresantes y de la salud femeninos entre los que se encuentra el abuso sexual, problemas sobre menstruación, embarazo o parto. Los elementos estresantes y de salud de la mujer no son diagnósticos ni se puntúan. Las otras escalas se puntúan por separado.

El test resultante enfocado a adolescentes es el PHQ-A (J. G. Johnson et al., 2002). Se trata de una versión sustancialmente modificada del PHQ original, pero adaptada para su uso en adolescentes. El manual y las puntuaciones que permiten hacer diagnóstico están disponibles bajo petición expresa a los autores.

El test PHQ-2 es un test de cribado de solo dos ítems (Inagaki et al., 2013; Kroenke et al., 2003; Mitchell et al., 2016; Richardson et al., 2010; Scoppetta et al., 2021). Con una estructura monofactorial, presenta puntuaciones totales de 0 a 6 y de sus ítems aislados de 0 a 3. Se trata de los dos ítems más salientes del PHQ-9, que corresponden con los dos primeros.

El test PHQ-4 (Kroenke, Spitzer, et al., 2009) se trata de una construcción en la que se incluyen los test PHQ-2 (Richardson et al., 2010) y GAD-2 (García-Campayo et al., 2012; Kroenke et al., 2010; Plummer et al., 2016).

El test PHQ-8 (Kroenke, Strine, et al., 2009) se trata del test Patient Health Questionnaire habiendo eliminado el último ítem relativo al suicidio. Entre las razones de

eliminar este ítem para configurar una versión más reducida, ahorrándose un ítem, están las de reducción de longitud y tratar el suicidio como una entidad aparte. Esto es porque el ítem de suicidio es el que tiene menor correlación entre el resto de ítems, aunque esta correlación es significativa y entra dentro del mismo constructo unidimensional que es la depresión (Kroenke, Strine, et al., 2009). Para tratar el suicido de forma más precisa, se aconsejan otros test complementarios, aunque el ítem 9 de forma aislada, está considerado un buen indicador, habiéndose fundamentado bien como tamizaje de pensamientos suicidas (Razykov et al., 2012).

Así, el PHQ se conforma como un test modular que puede adaptarse recortando o aumentando el número de ítems sopesando la potencia de detección con el tiempo invertido para realizarlo, incluir constructos que desean medirse o utilizarse conjuntamente con otros test para abarcar otros trastornos, mejorando la comprensión de la situación del paciente.

## 3.2. Ansiedad

La ansiedad puede considerarse como una respuesta de activación frente a un estímulo amenazante, el cual puede ser real o imaginario. Se caracteriza por una preocupación excesiva y difícil de controlar. La respuesta del individuo acompaña a preparaciones fisiológicas y psicológicas. Es por ello, por lo que pueden encontrarse: preocupación, inquietud, taquicardia, aceleración de la respiración, irritabilidad, tensión muscular o quedarse con la mente en blanco, entre otras. Además puede darse en diferentes niveles de intensidad, encontrándose una ansiedad adaptativa o patológica e incompatible con un adecuado desempeño y por lo tanto incapacitante. Cuando la ansiedad está injustificada, es desproporcionada y escapa al control voluntario de la persona, tiene un carácter intenso y recurrente, genera malestar significativo o interfiere en la vida normal del individuo, puede

categorizarse como ansiedad desadaptativa; ya que impide un adecuado desempeño de las capacidades de la persona.

La ansiedad puede presentarse bajo los siguientes síntomas:

Entre los síntomas psicofisiológicos de la ansiedad se destacan en las cefaleas y mareos, los sofocos y escalofríos, la hipertensión arterial y opresión torácica, dificultades para respirar, náuseas, vómitos, diarrea o estreñimiento, dolores musculares, dificultad de coordinación, fatiga, alteraciones del ciclo menstrual, disminución del deseo sexual, problemas con la eyaculación.

Entre los síntomas cognitivos se destacan la disminución del rendimiento en procesos mentales superiores como la atención, concentración, aprendizaje y memoria; dificultades en la toma de decisiones, pérdida de confianza en sí mismo, indefensión, despersonalización, sentimientos de inferioridad, sensación de desorganización, entre otros.

Con respecto a los síntomas motores se destacan el enlentecimiento motor, reacciones de sobresalto, irritabilidad y escala tolerancia a la frustración y agresividad, aislamiento social, dificultades con el sueño, conductas de riesgo como tabaquismo, abuso de alcohol y sustancias, así como problemas con la alimentación.

Además la ansiedad puede patologizarse en diferentes trastornos, entre los que se encuentran los trastornos de:

- Ansiedad, con crisis de angustia, agorafobia, trastorno de angustia, trastorno de ansiedad generalizada, fobia específica, fobia social, trastornos obsesivo-compulsivos, trastorno por estrés agudo, trastorno por estrés postraumático, trastorno adaptativo con ansiedad y trastorno de ansiedad debido a enfermedad médica.
- Somatomorfos, con el trastorno por somatización, de conversión, por dolor y la hipocondría.

- Psicofisiológicos, con trastornos cardiovasculares, gastrointestinales, respiratorios, dermatológicos, del aparato locomotor, genitourinarios y de la disminución de la respuesta inmunitaria.

### 3.2.1. GAD-7

Para medir ansiedad se utilizó la escala GAD-7 (Spitzer et al., 2006), que es un cuestionario en parte ideado a partir del DSM-IV (Bogduk, 2013)(Bogduk, 2013). Se estima que en población general, la prevalencia de la depresión varía entre 2,8% y 8,5% (Spitzer et al., 2006). Esta escala mide el desorden de ansiedad generalizada mediante 7 ítems más un ítem que hace referencia al grado de dificultad que estos problemas le han resultado en las áreas: laboral, cuidado del hogar y relaciones con otras personas. Este cuestionario hace referencia a las dos últimas semanas, por lo que puede considerarse un estado actual de ansiedad. Se han propuesto como punto de corte el 10, detectando una sensibilidad de 89% y especificidad del 82%, afirmando que la sensibilidad está casi maximizada.

Las opciones de respuesta son para nada, varios días, más de la mitad de los días y casi todos los días, cada una tomando valores de 0, 1, 2 y 3 respectivamente. Los puntos de corte propuestos son de 0-4 ansiedad mínima, de 5-9 ansiedad media, de 10-14 ansiedad moderada y de 15 a 21 ansiedad severa. Para su corrección se realiza una suma de puntos.

Aunque existen discrepancias en cuanto a la estructura del GAD-7, se ha estudiado una estructura unifactorial cuando la muestra es heterogénea (población clínica y no clínica) (S. U. Johnson et al., 2019)(S. U. Johnson et al., 2019).

En cuanto a la cesión de su uso, en el propio test se especifica que “No se necesita autorización para su reproducción, traducción, muestra o distribución”.

### 3.2.2. *El GAD-2:* Variante del General Anxiety Disorder 7

Como se ha mencionado anteriormente, el GAD-7 forma parte del PHQ-SADS. Además el GAD-2, junto con el PHQ-2 configuran el PHQ-4 estudiando el constructo ansiedad-depresión.

Esta escala permite realizar una valoración rápida de síntomas de Ansiedad Generalizada. Ideado para evaluar el trastorno de Depresión Generalizada (y por sus siglas en inglés *General Anxiety Disorder)* se generó como una entidad aparte, conformándose como un test ampliamente utilizado. Tras varios análisis se logró una escala de solo dos ítems, dando lugar al GAD-2 (García-Campayo et al., 2012; Plummer et al., 2016). En su versión española presenta una consistencia interna α Cronbach=0,875. Con ítems cuyas puntuaciones parciales van del 0 al 3 y cuyas puntuaciones totales van del 0 al 6, un punto de corte de 3 mostró sensibilidad adecuada (91,5%) y especificidad (85,8%), con un área bajo la curva normal estadísticamente significativa (ROC = 0,937, p <0,001), para distinguir a los pacientes con Trastorno de Ansiedad Generalizada de los controles. La validez concurrente también fue alta y significativa con las escalas HAM-A (Escala de Ansiedad de Hamilton) (Hamilton, 1959) (0,806, p <0,001), HADS (Escalas de Depresión y Ansiedad en Hospital) (Cabrera et al., 2015) (dominio de ansiedad, 0,825, p <0,001) y WHO-DAS-II (Escala de evaluación de discapacidad de la Organización Mundial de la Salud II) (McArdle et al., 2005) (0,642, p<0,001).

Se solicitó y fue concedida la autorización de investigaciones con humanos a través del Consejo Nacional de Investigaciones en Salud, el organismo CONIS. La concesión fue bajo el Acuerdo Nº 22 de la sesión ordinaria Nº 38 del 26 de agosto de 2020. Con estos requisitos previos, se podía iniciar el estudio.

## 3.3. Escalas de comparación

Para comprobar la validez desde fuentes de validez externa, se pueden emplear otros test o escalas utilizadas en el estudio de "Salud mental y relaciones con el entorno en tiempos de COVID 19 en la población costarricense". Por lo que se explican a continuación:

### 3.3.1. PROQOL

El cuestionario PROQOL (Hemsworth et al., 2018) permite identificar los efectos positivos y negativos del trabajo o la relación con personas que han experimentado situaciones adversas o estresantes. Consta de tres constructos: CS (Satisfacción compasiva), BO (Burnout) y STS (Estrés Traumático Secundario). Se compone de 30 ítems, distribuidos en 10 ítems por cada dimensión. Las correlaciones entre estos constructos son: CS-BO van de r=-0,81 a r=-0,88, para STS-BO sus correlaciones van de r=0,79 a r=0,88 y para el par CS-STS sus correlaciones van desde r=-0,13 hasta r=0,26. Con respecto a las correlaciones inter-ítems en a la primera muestra, CS tenía un $\alpha$ = 0,90, fiabilidad compuesta de 0,92; STS en $\alpha$ = 0,82, fiabilidad compuesta de 0,88; y Burnout en $\alpha$ = 0,80, fiabilidad compuesta de 0,83. Con respecto a la segunda muestra, CS tenía un $\alpha$ = 0,91, fiabilidad compuesta de 0,93; STS en $\alpha$ = 0,85, fiabilidad compuesta de 0,89; y Burnout en $\alpha$ = 0,75, fiabilidad compuesta de 0,79. Con respecto a la tercera muestra estudiada, CS tenía un $\alpha$ = 0,89, fiabilidad compuesta de 0,93; STS en $\alpha$ = 0,78, fiabilidad compuesta de 0,86; y Burnout en $\alpha$ = 0,74, fiabilidad compuesta de 0,78. Por lo tanto, cada escala mostró una fiabilidad con valores superiores a 0,70.

### 3.3.2. Fear of COVID-19 Scale (FCV-19S)

Esta escala construida en el marco de la situación mundial por COVID, se enfoca en el impacto que esta situación ha tenido en las personas, específicamente miedo, preocupación y ansiedad. Desarrollada por Ahorsu y su equipo (Ahorsu et al., 2020), posee buenas propiedades psicométricas en cuanto al constructo medido. La retención de los factores de carga fue significativa en los 7 ítems de los que se compone el test. Con una correlación que va desde 0,66 a 0,74 de cada ítem con el resto. No se encontró funcionamiento diferencial de los ítems en cuanto a género ni tampoco en cuanto a edad. Se dan valores de consistencia interna de Alpha de Cronbach de 0,82 y una validez de constructo de 0,88. En cuanto a la validez simultánea con las pruebas HADS (The Hospital Anxiety and Depression Scale) y PVDS (Perceived Vulnerability to Disease Scale) ambas resultan adecuadas con alta significación ($p<0,001$).

### 3.3.3. Escala de Miedo Social a la COVID-19 (SFCV-19S)

Se trata de una adaptación de un cuestionario que originalmente se diseñó para detectar miedo social ante acontecimientos sobre inseguridad ciudadana (Vozmediano et al., 2008), se adaptó este cuestionario para detectar el miedo social que pudieran experimentar las personas ante el contagio por COVID-19. Para detectar frecuencias de episodios de miedo lo bastante importantes para tener una relevancia en la vida diaria de los sujetos, se incluyó el cuestionario de miedo de Vozmediano adaptado para que se registre la cabida del miedo social a la COVID-19. En su publicación original evalúa el miedo social al delito, y en la adaptación se evalúa el miedo a la enfermedad por COVID-19, en las áreas social, laboral y afectiva. Se trata de un test de solamente 3 ítems que en su publicación original presenta una adecuación muestral KMO de 0,75, encontrando dos factores que explican el 62% de la varianza. En el caso

de emplear “miedo hacia la integridad física” (Fear of COVID-19 Scale) como escala de referencia para la comparación, la fiabilidad es de 0,75. En el caso de utilizar la escala “miedo hacia fuera de la integridad física”, de 0,74. Empleando la escala completa, el índice se eleva a 0,81.

### 3.3.4. How Stressed Are You?

“How Stressed Are You?” (HSAY). Se trata de la escala accesible desde: https://www.headington-institute.org/ la cual no aparece en publicaciones científicas a la fecha de redacción de este manuscrito. Tiene versiones traducidas al inglés, camboyano, francés y árabe.

La escala “How stressed are you” de The Headington Institute no ha sido validada psicométricamente, aunque las dimensiones que incorpora (cognitiva, emocional, corporal y conductual) se relacionan con las dimensiones presentes en cuestionarios validados para la medición de consecuencias en personas que han experimentado situaciones de estrés tales como el Daily Stress Inventory (Brantley et al., 1988), el Derogatis Stress Profile (Derogatis, 1987), o el Perceived Stress Scale (Cohen et al., 1994)(Cohen et al., 1994). Asimismo, en recientes estudios se confirma la utilidad de cuestionarios autoaplicables (como la escala “How stressed are you”), indicando que los ítems presentes correlacionan con conceptualizaciones compartidas entre teorías sobre estrés (Amirkhan, 2012).

Para poder administrar este instrumento primero sufrió un proceso de traducción y contratraducción por parte de personal bilingüe. Además se adaptó la escala traducida para adaptarse a la población costarricense. Todo este proceso fue supervisado por dos psicólogos expertos en construcción de cuestionarios. Como resultado se obtuvieron 25 ítems que conforman la escala “How Stressed Are You?”.

### 3.3.5. The 14-ítem Resilient Scale (RS14)

Se utilizó la versión española de la escala de resiliencia RS-14 (Robles-bello, 2014), la cual fue una traducción de la escala original de EE.UU. (Wagnild & Collins, 2009). Se anuncia como la escala más popular de medición de la resiliencia (G. Wagnild, 2022). La escala consta de catorce (14) ítems. Se estudió una estructura unifactorial explicando el 75,97% de la variabilidad con una consistencia interna de 0,74, aunque en el artículo original se evalúa la presencia de dos factores: competencia personal y aceptación de uno mismo y de la vida. En diferentes traducciones, presenta inestabilidad factorial y unidimensionalidad en la versión española (Robles-bello, 2014).

El test de resiliencia hace referencia a afirmaciones que describen a la persona, quien debe seleccionar su grado de acuerdo o desacuerdo que mejor indique sus sentimientos con respecto a estas afirmaciones. Para ello, se ofrece una escala de siete (7) opciones de respuesta que van desde el desacuerdo al de acuerdo.

Se ha estudiado una relación inversa entre depresión ($r=-0,79$, $p<0,001$) y ansiedad rasgo ($r=-0,64$, $p<0,001$). En población española, se obtuvieron unas puntuaciones de media 71 (DE=32,81), teniendo las puntuaciones máximas y mínimas teóricas de 98 y 24. También se señala que la edad correlaciona con la resiliencia apoyando la teoría de que la resiliencia se toma como un proceso dinámico modificable.

# SEGUNDA PARTE: OBJETIVOS Y METODOLOGÍA

# Capítulo 4

# Objetivos e hipótesis

*Si no sabes a dónde vas, probablemente terminarás en otro lugar.*

Lawrence J. Peter

# 4.-Objetivos e hipótesis

## 4.1. Objetivos

El objetivo principal de este trabajo es otorgar a la población costarricense una herramienta de calidad, válida, fiable, objetiva y no discriminatoria por sexo para la evaluación de la depresión.

Este objetivo general se concreta en los siguientes objetivos específicos:

1. Adaptar cultural y lingüísticamente los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 en la población costarricense.
2. Comprobar las diferencias que los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 puedan tener en población costarricense con respecto al sexo.
3. Conocer el comportamiento de los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 con respecto a otros test psicológicos: la Escala de Miedo a la COVID-19 (FCV-19S), Escala de Miedo Social a la COVID-19 (SFCV-19S) test de Estrés (HSAY), efectos emocionales del trabajo (PROQOL) y Resiliencia (RS14).
4. Determinar las propiedades psicométricas y la estructura factorial y de los instrumentos PHQ-9, PHQ-8, PHQ-4 y PHQ-2 en la población costarricense.
5. Proponer nuevos puntos de corte mediante el estudio de la sensibilidad y especificidad de los instrumentos PHQ-9, PHQ-8, PHQ-4 y PHQ-2 tanto para la población costarricense general como en función del sexo.

## 4.2. Hipótesis

Para conseguir los objetivos propuestos se plantean las siguientes hipótesis que pretende responder este estudio. Se espera encontrar que:

1. Ligeras adaptaciones de los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 son suficientes para que los test sean adecuados cultural y lingüísticamente para la población costarricense.
2. La invarianza de medida multigrupo de los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 presentan diferencias en función del sexo.
3. Los test de depresión PHQ-9, PHQ-8, PHQ-4 y PHQ-2 poseen una correlación significativa y directa con la Escala de Miedo a la COVID-19 (FCV-19S), Escala de Miedo Social a la COVID-19 (SFCV-19S) test de Estrés (HSAY), efectos emocionales del trabajo (PROQOL) e inversa con Resiliencia (RS14).
4. Los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 presentan adecuadas propiedades psicométricas para poder evaluar depresión en población costarricense. Los test PHQ-9, PHQ-8 y PHQ-2 presentan estructura unifactorial, mientras que el PHQ-4 presenta estructura bifactorial.
5. Los puntos de corte propuestos no implican diferencias estadísticamente significativas con respecto a los resultados obtenidos con los puntos de corte originales.

# Capítulo 5

# Metodología

*La mejor estructura no garantizará los resultados ni el rendimiento. Pero la estructura equivocada es una garantía de fracaso.*

Peter Ferdinand Drucker

# 5.-Metodología

Para este trabajo de tesis doctoral se presentó un Protocolo de investigación biomédica observacional, según lo establecido en el artículo 47 del reglamento No. 39533-S a la Ley Reguladora de Investigación Biomédica No. 9234, según legislación costarricense. Este protocolo se realizó para la investigación denominada: "Salud mental y relaciones con el entorno en tiempos de COVID 19 en la población costarricense".

El mentado protocolo fue redactado y consensuado por el equipo de investigación, estableciéndose como último responsable la Universidad Estatal a Distancia (UNED) y teniendo como investigadora principal y coordinadora del equipo a Eva Carazo Vargas. Otras instituciones participantes también participan. Entre ellas, figuran la Universidad Nacional de Costa Rica (UNA), la Caja Costarricense de Seguro Social (CCSS), el Ministerio de Salud y el Ministerio de Educación Pública. Todos estos organismos estuvieron presentes en las diversas reuniones mantenidas para recoger las necesidades que mantenían cada una de estas instituciones, para así poder adaptar el protocolo, de tal forma que se satisfacieran las demandas realizadas en cuanto a conocer el estado de salud mental y poder realizar inferencias y mecanismos de acción que permitieran minimizar los daños que pudiera sufrir la población.

Además cuenta con el patrocinio tanto de la Universidad Estatal a Distancia de Costa Rica, con la participación de la Escuela de Psicología de la Universidad Nacional de Costa Rica, el Ministerio de Salud y la Caja Costarricense de Seguro Social, pese a que no se destinó presupuesto alguno para el desarrollo de esta investigación en ninguna de sus fases.

Según la Norma de Atención Integral de la Salud Mental y de Abordaje Psicosocial en Situaciones de Emergencias y Desastres en los Escenarios de Servicios de Salud y en la Comunidad N° 41599 – S (Oficialización de La Norma de Atención Integral de La Salud Mental y de Abordaje Psicosocial En Situaciones de Emergencias y Desastres En Los Escenarios de

Servicios de Salud y En La Comunidad N° 41599 - S, 2019), se definió el concepto de salud mental y abordaje psicosocial como sigue:

- Abordaje Psicosocial. Proceso de apoyo y seguimiento articulado, interdisciplinario, interinstitucional, intersectorial y comunitario basado en la gestión del riesgo, orientado a restablecer la cotidianeidad de las personas, la integridad emocional y reactivar sus redes sociales, con un enfoque participativo, de derechos y de género, brindado por personal especializado y no especializado".
- Salud Mental. Proceso de bienestar y desempeño personal y colectivo caracterizado por la autorrealización, la autoestima, la autonomía, la capacidad para responder a las demandas de la vida en diversos contextos: familiares, comunitarios, académicos, laborales y disfrutar la vida en armonía con el medio ambiente.

Además, los enfoques utilizados se definieron como sigue:

- Enfoque de Salud Mental. Este enfoque permite visualizar la salud mental como una construcción colectiva, histórica y social que surge de las condiciones de vida e interacción entre las personas, grupos sociales, comunidades y el ambiente en que se encuentran, con lo cual las acciones se orientan hacia la promoción y fortalecimiento de factores protectores desde las comunidades y grupos, evitando la excesiva medicalización.
- Enfoque Psicosocial. Lo psicosocial no es una dimensión desarticulada del proceso de atención de población afectada ante situaciones de emergencias o desastres, es un abordaje transversal en todas las decisiones que se toman en este tipo de eventos, ya que se producen diferentes problemáticas en los niveles: individual, familiar, comunitario y social, deteriorando los mecanismos de protección y aumentan los riesgos de que se presenten numerosos tipos de problemas sociales o de conducta.

Por otro lado, y como complemento de la definición previa, según la Organización Mundial de la Salud a través de su Comité interinstitucional Permanente, la Salud mental y apoyo psicosocial (SMAPS) (IASC Inter-Agency Standing Committee, 2006) son utilizados para describir una amplia gama de acciones que permiten abordar problemas sociales, psicológicos y psiquiátricos preexistentes o inducidos por una emergencia. Estas acciones son implementadas en contextos muy distintos por organizaciones y personas con distintos bagajes profesionales, pertenecientes a distintos sectores y con distintos tipos de recursos. Todos estos múltiples actores - y sus donantes - necesitan evaluaciones prácticas que conduzcan a recomendaciones que se puedan utilizar inmediatamente para mejorar la salud mental y el bienestar de las personas.

El estudio presentado y del que se realiza esta tesis doctoral consiste en una exploración de las situaciones que se están dando en la salud mental de la población costarricense en el contexto de la emergencia sanitaria por COVID-19. Teniendo en cuenta que las condiciones para disminuir el riesgo de contagio, a través de medidas para la prevención y retención de la propagación de la enfermedad. Estas medidas son el confinamiento, restricción de la movilización, distanciamiento físico o el cumplimiento de otras medidas planteadas por el gobierno como la restricción de movilidad vehicular. Todo ello junto con el cambio en la realidad social cotidiana, han podido tener efecto en el equilibrio emocional de las personas habitantes en Costa Rica. La investigación se realizó por medio de una encuesta administrada por medios telemáticos (ordenador, Tablet, móvil) en la que se incluye un cuestionario autoadministrado. En este cuestionario se pregunta sobre una serie de variables de interés, como: el impacto cognitivo, conductual, emocional y físico, el riesgo percibido y miedo al contagio, adherencia a las medidas y capacidad y estrategias de afrontamiento. Todo ello teniendo en cuenta las características psicosociales de la población.

De esta forma se realizó un análisis multivariante para caracterizar el estado de salud mental de la población costarricense con el fin de estimar la prevalencia de las dimensiones más relevantes para explicar cómo la población está gestionando sus emociones. Todo ello con el fin de favorecer la construcción de recomendaciones y sugerencias para la atención de la población y, en especial de los grupos más vulnerables, quienes resultan más afectados por la enfermedad de COVID-19 y por las medidas sanitarias que la misma requiere, como el confinamiento. Las dimensiones incluidas la evaluación han sido categorizadas en cinco áreas relacionadas con lo afectivo, lo cognitivo, lo conductual, lo interpersonal y lo somático. Los ítems que se incluyen para la medición de las diferentes dimensiones provienen de pruebas validadas.

De la batería de cuestionarios y otras preguntas como las sociodemográficas, se seleccionó el cuestionario PHQ para su validación y motivo de esta tesis doctoral. Además, los cuestionarios GAD y la pregunta de padecimiento de depresión, fueron utilizados para el análisis del artículo, aunque en esta tesis doctoral se amplía la información incluyendo las correlaciones con estos otros test.

Aunque en el test original de PHQ-9 consta de 9 ítems, el propio test se añade una pregunta extra que hace referencia a las dificultades experimentadas en alguna de las áreas social, familiar o laboral, a efectos de análisis se ha disgregado esta pregunta en tres diferentes para explorar cada área por separado. Los 9 primero ítems son los ítems a través de cuyas puntuaciones se obtiene un punto de corte, pero a efectos de sumatorio de puntuaciones, la última puntuación se deshecha, residiendo su valor en la práctica clínica, ya que a efectos psicométricos y en las instrucciones originales de la herramienta, no se analiza.

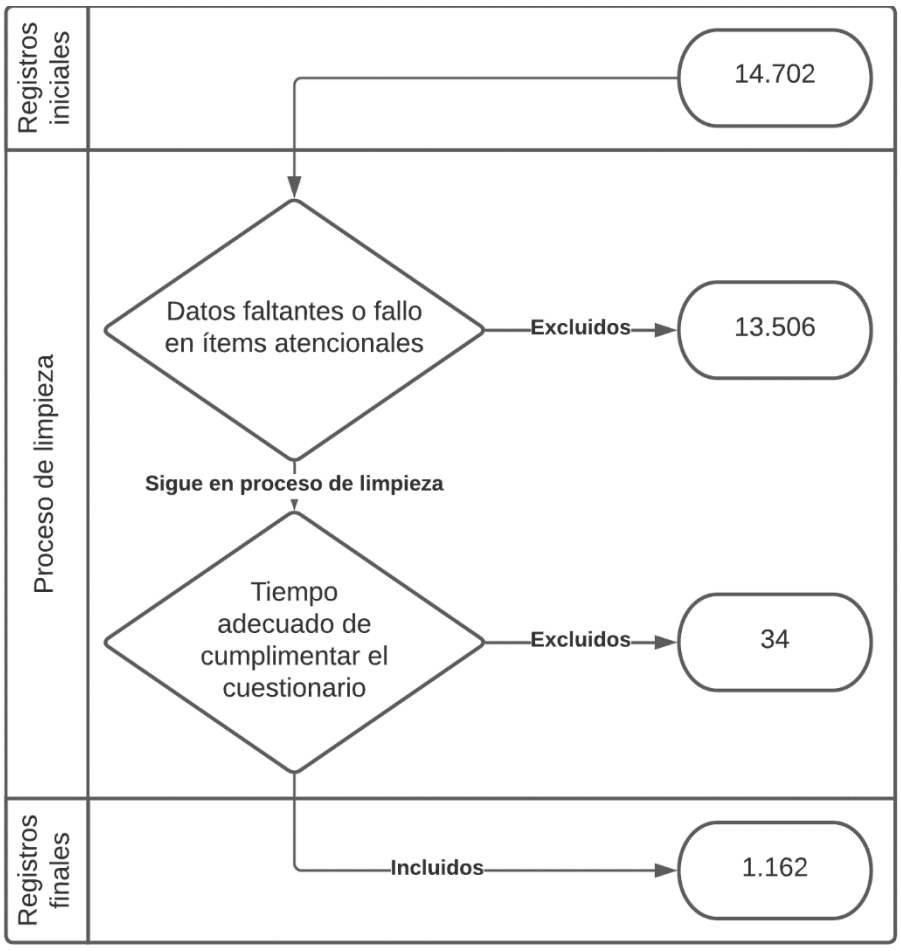
## 5.1. Participantes

Los criterios de inclusión fueron ser costarricense residente en Costa Rica, tener un nivel de idioma suficiente como para leer y comprender el español y aceptar el consentimiento

informado. Los criterios de exclusión fueron no haber superado los 4 ítems atencionales y no terminar el cuestionario con un tiempo de respuesta válido.

Originalmente se obtuvieron 14.702 registros en la plataforma de encuestas, no obstante 13.506 registros fueron descartados por tener algún dato faltante o haber fallado los ítems atencionales. Una vez realizado este cribado y tras un estudio del tiempo empleado en cumplimentar las preguntas, se eliminaron las respuestas de 34 sujetos que utilizaron demasiado poco tiempo en completar las respuestas. De esta forma la muestra final se compone de 1.162 cuestionarios completos (Figura 11. Flujograma del proceso de limpieza de los datos).

*Figura 11. Flujograma del proceso de limpieza de los datos*



Fuente: elaboración propia.

Para conocer el tamaño de la muestra requerido para realizar un adecuado Modelo de Ecuaciones Estructurales se utilizó una calculadora específica (Soper, 2022) fijando los siguientes parámetros:

- Tamaño del efecto anticipado: 0,1
- Nivel de poder estadístico deseado: 0,8
- Número de variables latentes: 1
- Número de variables observadas: 9
- Nivel de probabilidad: 0,05

Dando como resultado un tamaño mínimo recomendado de 1.100 participantes, un tamaño de muestra mínimo para el modelo estructural de 1.100 y una muestra mínima para detectar diferencias de 87.

Los participantes fueron en su mayoría mujeres, de situación civil soltero, sin estudios universitarios, de la provincia de San José y desempleados.

*Tabla 8. Características sociodemográficas de la muestra*

| | |
|---|---|
| **Muestra completa** | (n=1.162) |
| **Edad (media, DT)** | 35,52 (12,18) Rango 15-78 |
| **Género (n. %)** | |
| Varón | 247 (21,3) |
| Femenino | 915 (78,7) |
| **Estado civil (n. %)** | |
| Soltero | 561 (48,3) |
| Casado | 361 (31,1) |
| En pareja | 116 (10,0) |
| Divorciado | 94 (8,1) |
| Separado | 19 (1,6) |
| Viudo | 9 (0,8) |
| Sin datos | 2 (0,2) |
| **Nivel educativo (n. %)** | |
| Primaria I | 6 (0,5) |
| Primaria C | 17 (1,5) |
| Secundaria I | 46 (4,0) |
| Secundaria C | 132 (11,4) |
| Universidad I | 338 (29,1 |

| | |
|---|---|
| Universidad C | 326 (28,1) |
| Máster I | 52 (4,5) |
| Máster C | 153 (13,2) |
| Doctorado I | 14 (1,2) |
| Doctorado C | 16 (1,4) |
| Posdoctorado | 3 (0,3) |
| Estudios técnicos I | 8 (0,7) |
| Estudios técnicos C | 51 (4,4) |
| **Provincia (n. %)** | |
| San José | 459 (39,5) |
| Alajuela | 176 (15,1) |
| Cartago | 205 (17,6) |
| Heredia | 149 (12,8) |
| Guanacaste | 94 (8,1) |
| Puntarenas | 42 (3,6) |
| Limón | 37 (3,2) |
| **Situación laboral (n. %)** | |
| Empleado | 651 (56,0) |
| Desempleado | 105 (9,0) |
| Otros | 406 (34,9) |

Nota. Los estudios de secundaria incluyen estudios técnicos a nivel de secundaria. Los estudios técnicos se refieren a estudios una vez se haya superado los estudios de secundaria. I: incompleto. C: completo.

De la muestra 30 sujetos (2,6% del total) tenían menos de 18 años en el momento de realizar la encuesta (1 sujeto de 15 años, 3 sujetos de 16 años y 26 sujetos de 17 años). Además solo 10 sujetos (un 0,9% del total) presentaba más de 65 años.

## 5.2. Instrumentos

La escala de depresión PHQ-9 (Diez-Quevedo et al., 2001; Kroenke et al., 2001) es utilizada para detectar sintomatología concurrente con depresión mediante cribado. Tiene excelentes propiedades psicométricas tanto por su consistencia interna (α= .89) como por su confiabilidad test-retest (ICC=.84). Está basado en la Teoría Clásica de los Test y sus ítems tienen puntuaciones que van de 0 a 3. La escala PHQ-9, en su corrección, tiene puntuaciones totales que van de 0 a 27. Estas se obtienen sumando la puntuación aislada de cada uno de sus ítems. Para su corrección se toman los puntos de corte de 5, 10, 15 y 20, quedando una clasificación de la siguiente manera: depresión mínima 0-4; depresión media 5-9; depresión moderada 10-14; depresión moderada a severa 15-19 y depresión severa 20-27.

El test PHQ-9 consta de 9 ítems principales además de otros 3 ítems en los que se pregunta si ha tenido molestias por alguno de los problemas mencionados en los ítems principales, referido a la dificultad que ocasionan estos problemas en tres contextos: trabajo, hogar o relaciones sociales. Estos 3 últimos ítems tienen opciones de respuesta que van desde 0, sin dificultad, hasta 3, extrema dificultad. En la escala original se toman estos tres ítems a la vez, pero a efectos de análisis se han separado. De esta manera se puede conocer el área específica que se ve afectada, en lugar de conocer la alteración del conjunto de las tres áreas.

Para este estudio se han modificado algunos ítems del PHQ-9 versión de la española:

1: “Poco interés o alegría por hacer cosas” por la siguiente redacción: “Poco interés o placer en hacer las cosas”.

2: “Sensación de estar decaído/a, deprimido/a o desesperanzado/a” por la siguiente redacción: “Sentirse desanimado/a, deprimido/a o sin esperanzas”.

Esta nueva redacción no altera el contenido del ítem, sino que busca una comprensión más cercana a la comprensión costarricense. De igual forma, en la sección de grado de

dificultad en realizar determinadas tareas, se ha seleccionado la versión argentina del PHQ-9, ya que es estimó como la más próxima al estilo comunicativo costarricense. Además esta parte se disgregó en tres ítems diferentes para separar y concretar el área en la cual se efectúa esta dificultad; ya que en la redacción original no realiza tal distinción.

La escala PHQ-2 se encuentra dentro de la escala PHQ-9 ya que corresponde a sus dos primeros ítems, conformando una versión reducida de la versión principal (Camargo et al., 2022; Inagaki et al., 2013; Kroenke et al., 2003; Mitchell et al., 2016; Richardson et al., 2010). Además, los dos ítems incluidos se consideran los criterios básicos del trastorno depresivo. Su rango de puntajes va de 0 a 6. Los puntajes mayores o iguales a 3 son un indicador de depresión.

La versión PHQ-8 incluye todos los ítems de la versión PHQ-9 excepto el noveno ítem, relacionado con el suicidio que indica depresión actual en un punto de corte ≥10 (Kroenke, Strine, et al., 2009; Razykov et al., 2012).

La versión PHQ-4 consiste en la unión de los dos primeros ítems del PHQ-9 junto con los dos primeros ítems del GAD-7 (General Anxiety Disorder-7), examinando así la depresión y la ansiedad (Kroenke, Spitzer, et al., 2009; Löwe et al., 2010; Wicke et al., 2022).

El GAD-7 es una escala de detección para medir el trastorno de ansiedad generalizada (S. U. Johnson et al., 2019; Ong et al., 2022; Plummer et al., 2016; Spitzer et al., 2006). Con fuerte consistencia interna (α=.94) y buena confiabilidad test-retest (ICC=.83), utiliza 7 ítems cuyas opciones de respuesta toman valores de 0 a 3. La puntuación total de la prueba resulta de la suma de sus ítems, teniendo así puntuaciones mínimas y máximas de 0 a 21. Los puntos de corte propuestos en su escala original son de 0-4 ansiedad mínima, de 5-9 ansiedad media, de 10-14 ansiedad moderada y de 15 a 21 ansiedad severa.

Se han modificado los ítems del GAD-7 de la siguiente forma:

1. Se mantiene el ítem de la versión española.
2. Se mantiene el ítem de la versión colombiana.
3. Se mantiene el ítem de la versión colombiana.
4. Se mantiene el ítem de la versión colombiana.
5. Se mantiene el ítem de la versión mexicana, peruana o puertorriqueña, ya que son el mismo.
6. Se modifica de “Enfadarse o irritarse con facilidad” a “Molestarse o enojarse fácilmente”.
7. Se mantiene el ítem de la versión colombiana.

El cuestionario GAD-2 está incluido en el GAD-7, correspondiente a sus dos primeros ítems. Una puntuación ≥3 indica relevancia clínica del trastorno de ansiedad (Johnson et al., 2019; Kroenke et al., 2007; Plummer et al., 2016; Spitzer et al., 2006). Posee un buen ajuste para una estructura bifactorial (RMSEA .02; 90% CI .023-.032) (Löwe et al., 2010).

El test de Resiliencia RS-14 mide el grado de adaptación individual a situaciones adversas. Esta escala está compuesta por 14 ítems con puntuaciones totales teóricas que van de 98 a 14 y que indican diferentes niveles de resiliencia según este rango de puntuación: 98-82=Resiliencia muy alta; 81-64=Alta resiliencia; 63-49=Regular; 48-31=Baja resiliencia; y 30-14=Resiliencia muy baja (Cénat et al., 2018; W. Chen et al., 2020; G. M. Wagnild & Collins, 2009; Zelviene et al., 2021).

La Escala de Miedo al Covid-19 [Fear of COVID-19 Scale] ha sido validada en numerosos países (Ahorsu et al., 2020, 2022; I. H. Chen et al., 2022; Espejo & Checa, 2021; Griffiths, 2001; Huarcaya-Victoria et al., 2020; Iversen et al., 2021; Kempen et al., 2008; Lang & Harrington, 2020; Magano et al., 2021; Mahmood et al., 2020; Malik et al., 2021; Martínez-Lorca et al., 2020; Midorikawa et al., 2021; Mohsen et al., 2022; Muller et al., 2021; Nikopoulou et al.,

2022; Pakpour et al., 2020; Pang et al., 2022; Perz et al., 2022; J. Piqueras et al., 2020; J. A. Piqueras et al., 2021; Reznik et al., 2020, 2021; Sakib, Bhuiyan, Hossain, & Mamun, 2020; Sakib, Bhuiyan, Hossain, al Mamun, et al., 2020; Soares et al., 2021; Soraci et al., 2020; Soto-Briseño et al., 2021; Stănculescu, 2021; Tzur Bitan et al., 2020; Tzur et al., 2020; Wakashima et al., 2020; Winter et al., 2020). Esta escala mide la intensidad de temor personal que pueden tener los individuos frente a la enfermedad por SARS-CoV-2. Tiene unas adecuadas propiedades psicométricas: aceptable correlación ítem-total corregida (de 0.47 a 0.56), fuertes factores de carga (de 0.66 a 0.74) consistencia interna (α=.82), fiabilidad test-retest (ICC=.72), además de validez concurrente con test de ansiedad (r=.42) y depresión (r=.51) a través del Hospital Anxiety and Depression Scale (HADS) (Annunziata et al., 2020; Bjelland et al., 2002; Garaiman et al., 2021; Herrero et al., 2003; Herrmann, 1997; Matza et al., 2010; Yamamoto-Furusho et al., 2018; Zemła et al., 2019). Consta de 7 ítems con afirmaciones ante las cuales se ha de seleccionar el grado de acuerdo con las mismas, con 5 opciones de respuesta que van desde el muy en desacuerdo con una puntuación de 1 al muy de acuerdo cuya puntuación es 5. La puntuación total del cuestionario es una suma de sus ítems particulares yendo estas puntuaciones desde el 7 hasta el 35 e indicando que a mayor puntuación, mayor severidad de temor indicado. Esta escala carece de puntos de corte recomendados o establecidos.

Como criterio de validez externa se incluyó un ítem en el cuestionario: “De esta lista de padecimientos, por favor seleccione aquellos que usted padece en estos momentos” entre la lista de padecimientos se incluyó la depresión; pero también se incluyeron los criterios sobre exploración médico-psicológica para cribado en conductores (Jubal et al., 2007; Mirabet et al., 2022). Estos criterios son de padecer ansiedad, de tomar medicamentos en los últimos 30 días (“¿Ha tomado en los últimos 30 días pastillas para los nervios, la depresión o para dormir?”), de encontrarse en tratamiento psicológico o psiquiátrico en los últimos 30 días (“¿Ha estado en tratamiento psiquiátrico o psicológico en los últimos 30 días?”) o de padecer de alteraciones del sueño.

## 5.3. Procedimiento

Durante 8 meses se recabaron datos para el estudio de validación del PHQ-9, desde el 22 de marzo al 22 de septiembre de 2021; esto comprende el segundo año de pandemia por COVID-19 en el país. Los datos fueron recabados mediante un cuestionario en línea que los propios participantes se autoaplicaban. Se introdujeron 4 ítems atencionales en todo el cuestionario, de los cuales uno de ellos se encontraba en el propio test PHQ-9. Su redacción era como sigue: "Si está leyendo esto, por favor, seleccione la siguiente opción: Para nada/Varios días/Más de la mitad de los días/Casi todos los días". De entre las cuatro opciones se seleccionaba una de ellas. Los sujetos que no hubieran marcado la casilla requerida fueron excluidos del análisis.

Para desarrollar el cuestionario adaptado a la población costarricense se dieron muchos pasos para obtener una versión final. Dado que en Costa Rica se habla español, primero se revisaron todas las versiones de los cuestionarios disponibles en la página web de Pfizer (https://www.phqscreeners.com/). Tanto para consultar como para descargar o reproducir estas versiones de los test no se necesita autorización expresa y así se detalla en su web. Los cuestionarios en español disponibles en la fecha de acceso fueron para las poblaciones de Argentina, Chile, Colombia, México, Perú, Puerto Rico, España y USA. La redacción de los ítems sufrió ligeras modificaciones. Una de ellas fue la redacción en cuanto al sexo. Aunque en la redacción de textos el masculino se trata de una forma neutral para referirse a las personas, en el test PHQ-9 existe desdoblamiento de sexo (deprimido/a, dormido/a…). Esta doble mención separada por una barra, aunque no es incorrecta sí resulta artificiosa o alejada del lenguaje natural. Para seguir el movimiento social de modificación del lenguaje, así como para evitar duplicar el término con su desdoble para masculino y femenino se optó porque las afirmaciones del cuestionario no tuvieran una terminación con terminación

femenina, modificando así los términos (de deprimido/a a con depresión, o problemas para quedarse dormido/a a problemas para iniciar el sueño, etc.), obteniendo una forma indesdoblable. Por otro lado, se seleccionaron ítems completos de las versiones argentina y mexicana, pero para otras redacciones se tuvieron que reescribir los ítems manteniendo su esencia, pero adaptándose a un lenguaje más entendible por la población costarricense. El resultado de este proceso fue un cuestionario supervisado por 3 psicólogos clínicos: una psicóloga clínica nativa costarricense, un psicólogo especializado en metodología y un psicólogo de emergencias. Se constató la versión final mediante acuerdo de expertos, posteriormente fue consolidado por nativos costarricenses. Se expuso este cuestionario a una muestra local preguntando sobre la comprensión, redacción y claridad del texto y no se encontraron propuestas hacia la modificación de la versión presentada. Tampoco se encontraron propuestas de modificación de este test en la versión de prueba a voluntarios para conocer cualquier problema relacionado con el desempeño del cuestionario en línea.

El cuestionario se alojó en la plataforma LimeSurvey y estuvo accesible mediante internet. Los participantes eran animados a participar mediante anuncios gubernamentales en televisión, radio y redes sociales. Se realizaron varias campañas para que pudiera participar la población desde el ministerio de sanidad y la Caja Costarricense de Seguro Social. Esta última institución es una institución semejante al Ministerio de Derechos Sociales y Agenda 2030 (también conocido como Ministerio de Asuntos Sociales) de España.

## 5.4. Análisis de datos

El test PHQ-9 consta de 10 ítems: los 9 primeros referidos a las dimensiones de la depresión y el décimo referido a las implicaciones en la vida diaria de la gravedad de la depresión. El ítem 10 se refiere a cómo afecta la depresión en las áreas social, familiar y laboral. Para una mejor comprensión de la afección de la depresión en estas áreas, estas se contemplaron por separado, conformando un ítem por cada área. De esta forma el ítem 10

corresponde al área laboral (o cómo afecta para hacer su trabajo), el 11 al área familiar (o cómo afecta al desempeño de las tareas del hogar) y el 12 al área social (o cómo afecta en sus relaciones con los demás).

Se obtuvieron estadísticos descriptivos correspondientes a frecuencias, medias, desviaciones estándar, asimetría y curtosis mediante el programa estadístico SPSS v.25 (SPSS Corp., 2017). Para estudiar la coherencia interna de los ítems en el test se utilizó el estadístico Omega de McDonald's (Oω) utilizando la macro Omega para el paquete SPSS beta 0.2 (Hayes & Coutts, 2020). Esto es debido a que se ha estudiado que el α no es una medida óptima de confiabilidad porque su cálculo no está disponible para los programas estadísticos populares y porque está optimizado para variables cualitativas como los diferentes niveles de los ítems de test psicológicos (Hayes & Coutts, 2020). Para estimar la validez concurrente se utilizaron correlaciones a través de Rho de Spearman.

En la distribución de frecuencias la curtosis con un valor de 0 expresa que la distribución se ajusta a una curva normal, tomando los valores negativos forma platicúrtica y los positivos leptocúrtica. La asimetría toma como valor de referencia el 0 para una distribución simétrica.

Para el cálculo de correlaciones entre los test se utilizan las puntuaciones de los test. Esto quiere decir que se trata de los test corregidos. Ello implica que el test PHQ-9 sea un sumatorio de sus 9 primeros ítems, sin considerar las áreas de afección. Por otro lado, el test PROQOL al no tener una medida general de análisis, se consideraron sus subescalas por separado.

En cuanto a las correlaciones del test PHQ-9 con respecto a las respuestas sobre los criterios externos se compararon las puntuaciones del test PHQ-9 con el porcentaje de muestra que dio positivo en ese criterio junto con la puntuación del test, para todas y cada una de las puntuaciones del test.

Para el cálculo del criterio Depresión-Ansiedad se tuvieron en cuenta las puntuaciones del criterio depresión y la del criterio ansiedad; posteriormente se creó una variable resultante de la suma de sus puntuaciones. Así se obtuvo esta variable con 3 opciones de respuesta: 0 para aquellos sujetos que contestaron ausencia de padecimiento de depresión y de ansiedad conjuntamente; 1 para aquellos sujetos que contestaron padecimiento de depresión o ansiedad disjuntamente; y 2 para aquellos sujetos que contestaron padecimiento de depresión o ansiedad conjuntamente.

Tanto para realizar los descriptivos del PHQ-9, como para las comparaciones de Chi cuadrado de los criterios con respecto al test PHQ-9 se realizaron los análisis utilizando las directrices originales sobre la corrección de este test.

Para el cálculo de los estadísticos Chi cuadrado de Pearson se tuvieron en cuenta los estadísticos de asociación, de tal forma que fue utilizado el estadístico $\chi^2$ de tendencia lineal para aquellos pares de variables con ordenaciones binarias para la variable resultado y de 3 o más categorías para la variable de exposición ordenadas; así como los estadísticos de asociación Gamma y R de Spearman para pares de variables ordinales (Rivas-Ruiz et al., 2013). Para los pares de variables ordinal-nominal, se utilizó el índice eta para conocer el grado de relación entre variables (Kennedy, 1970; Levine & Hullett, 2002; Olejnik & Algina, 2003).

Para los análisis de validez interna y del modelo de Rasch se utilizó el programa FACTOR v.12.01.02 (Lorenzo-Seva & Ferrando, 2006). El análisis factorial se elaboró mediante correlación policórica con el modelo factorial de Análisis Factorial de Rango Mínimo (MRFA) (Timmerman & Lorenzo-Seva, 2011). Para determinar el índice de Kaiser-Meyer-Olkin (KMO) se utilizó la muestra completa como dos submuestras aleatorias con el método de Solomon (Lorenzo-Seva, 2021). La varianza explicada se obtuvo a través de sus valores propios. Además, se ofrece el estadístico [Measure of Sampling Adequacy (MSA)] (Lorenzo-Seva & Ferrando, 2021), que aconseja la eliminación de ítems con valores inferiores a 0,50 (Lorenzo-Seva &

Ferrando, 2006). El análisis factorial se realizó siguiendo los estándares recomendados (Ferrando et al., 2022). Este programa también ofrece el Índice de Dificultad Relativa (IDR) de los ítems; se aconseja que exista un 75% de los ítems con dificultades medias (entre 0,4 y 0,6), mientras que los porcentajes restantes se distribuyan uniformemente en ambas colas de la distribución; no obstante en pruebas de cribado es interesante un porcentaje mayor de ítems en los extremos (Lorenzo-Seva & Ferrando, 2006). El índice de CMI se refiere al Cuartil de Medias Ipsativas, el cual se refiere a las medias de las variables que se colocan en la distribución de los valores registrados para cada sujeto, de tal forma que ofrece la misma información que el IDR, pero en cuartiles. Los estadísticos de la Teoría de Respuesta al Ítem se ofrecieron según ojiva normal de 4 parámetros (Samejima, 1969).

Para el cálculo de los puntos de corte óptimos de depresión se obtuvieron resultados similares para los criterios de tomar medicación, estar en tratamiento psiquiátrico o psicológico o padecer depresión. Asimismo, para establecer los puntos de corte del test PHQ-4 se obtuvo el criterio de padecer depresión al mismo tiempo que padecer ansiedad. Se consideró que no se detectaba el rasgo de depresión (o depresión-ansiedad) según los puntos de corte de los test en sus publicaciones originales. Para la elaboración de los puntos de corte nuevos se realizó la misma técnica que para la creación del test PHQ-8: se la distancia de los rangos intermedios, reduciendo o incrementando los rangos extremos.

Se utilizaron los criterios del índice J de Youden y el criterio *Closest top left* (el punto más cercano a la izquierda superior de la curva ROC). Ambos indicadores no siempre coinciden, pero se opta por el estadístico J de Youden cuando existen discrepancias en estos índices (Perkins & Schisterman, 2006). Además, se presentan la sensibilidad, la especificidad y el área bajo la curva ROC. De manera similar, los puntos de corte para el PHQ-8 se establecieron en su versión original, respetando la longitud de severidad de cada tramo a partir del punto óptimo de corte (Kroenke, Strine, et al., 2009); para calcular los puntos de

corte que indican la severidad de la depresión en cada test, se han sumado las bases correspondientes a los puntos de corte obtenidos utilizando los índices Youden J y Closest top left de la curva ROC para respetar los rangos de longitud de cada punto de corte para establecer la gravedad de los síntomas depresivos.

Para realizar los modelos de ecuaciones estructurales se utilizó el programa AMOS (Arbuckle, 2016), probando los diferentes modelos de ecuaciones estructurales encontrados en la literatura acerca de la estructura factorial del PHQ-9 (Lamela et al., 2020). Se comprobó la invarianza de medida de la estructura del test mediante Análisis Confirmatorio Multi Grupo (MGCFA)

Para el análisis multivariante se utilizó la técnica de HJ-Biplot (Galindo, 1986) con el programa MultBiplot versión (11/01/2018) 18.1101 (Vicente-Villardón, 2016).

Se marcó un nivel de confianza de 0,05 como significativo para todos los análisis, incluidos los intervalos de confianza (IC).

# TERCERA PARTE: RESULTADOS Y DISCUCIÓN

# Capítulo 6

# Resultados

*El pensamiento estadístico será un día tan necesario para el ciudadano eficiente como la capacidad de leer y escribir.*

Herbert George Wells

# 6.-Resultados

Parte de estos resultados se encuentran publicados como artículo científico (González-Sánchez et al., 2023).

## 6.1. Adaptación cultural y lingüística

En la adaptación lingüística los jueces que evaluaron los ítems no encontraron dificultad en su comprensión, conceptualización u otras complicaciones.

El cuestionario no había sido validado previamente en Costa Rica. Aunque el español es el idioma principal que se habla en Costa Rica, sus características culturales únicas requirieron algunos cambios menores en el cuestionario para una mejor comprensión y fácil comprensión por parte de la población objetivo. Esto no implicó traducir el cuestionario de un idioma a otro, sino realizar pequeños ajustes, como se había hecho anteriormente con el test PHQ-9.

Se estudiaron todas las versiones disponibles y considerando que las pruebas son esencialmente las mismas, pero con cambios menores, se seleccionaron aquellos ítems que pudieran ser mejor comprendidos por los costarricenses. De esta manera, se seleccionaron los ítems de las diferentes versiones del español y en aquellos en los que se consideró necesario se introdujeron cambios menores, como eliminar artículos o añadir comas.

El equipo reconoció la necesidad de estas modificaciones para que el cuestionario fuera más comprensible para la población costarricense, como se ha hecho en otros países de habla hispana. En el sitio web oficial de Patient Health Questionnaires Screeners (https://www.phqscreeners.com/) se encuentran disponibles ocho versiones del test PHQ-9 en español de entre las cuales se seleccionaron los ítems más comprensibles para configurar una nueva selección de esos mismos ítems.

## 6.2. Descriptivos depresión

De las respuestas obtenidas (n=1.162) se obtuvieron los estadísticos descriptivos que se observan en la tabla (Tabla 9. Estadísticos descriptivos de los ítems del PHQ-9).

*Tabla 9. Estadísticos descriptivos de los ítems del PHQ-9*

| Estadístico | | Ítem 1 | Ítem 2 | Ítem 3 | Ítem 4 | Ítem 5 | Ítem 6 | Ítem 7 | Ítem 8 | Ítem 9 | Laboral | Familiar | Social |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 | 1162 |
| Media | | 1,2 | 1,18 | 1,36 | 1,62 | 1,27 | 1,12 | 0,94 | 0,63 | 0,46 | 1,11 | 1,01 | 1,02 |
| Error estándar de la media | | 0,03 | 0,031 | 0,033 | 0,031 | 0,032 | 0,033 | 0,029 | 0,027 | 0,025 | 0,024 | 0,026 | 0,026 |
| Mediana | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| Moda | | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Desviación típica | | 1,013 | 1,063 | 1,133 | 1,061 | 1,103 | 1,135 | 1,001 | 0,922 | 0,859 | 0,829 | 0,875 | 0,901 |
| Varianza | | 1,025 | 1,13 | 1,283 | 1,126 | 1,217 | 1,288 | 1,003 | 0,851 | 0,738 | 0,687 | 0,765 | 0,811 |
| Asimetría | | 0,478 | 0,509 | 0,288 | 0,055 | 0,367 | 0,532 | 0,78 | 1,368 | 1,879 | 0,385 | 0,544 | 0,544 |
| Error estándar de asimetría | | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 | 0,072 |
| Curtosis | | -0,848 | -0,964 | -1,313 | -1,298 | -1,192 | -1,152 | -0,506 | 0,827 | 2,486 | -0,395 | -0,432 | -0,528 |
| Error estándar de curtosis | | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 | 0,143 |
| Rango | | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Mínimo | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Máximo | | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Suma | | 1395 | 1375 | 1580 | 1884 | 1481 | 1303 | 1098 | 737 | 535 | 1289 | 1170 | 1180 |
| Percentil 10 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Percentil 20 | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Percentil 25 | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Percentil 30 | | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Percentil 40 | | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Percentil 50 | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| Percentil 60 | | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| Percentil 70 | | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 0 | 1 | 1 | 1 |
| Percentil 75 | | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Percentil 80 | | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 1 | 1 | 2 | 2 | 2 |
| Percentil 90 | | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 |
| Frec. | 0 | 324 | 365 | 323 | 172 | 351 | 468 | 486 | 696 | 838 | 277 | 368 | 381 |
| | 1 | 459 | 423 | 386 | 441 | 381 | 303 | 384 | 283 | 186 | 546 | 490 | 463 |
| | 2 | 201 | 170 | 165 | 204 | 190 | 173 | 162 | 95 | 65 | 274 | 232 | 237 |
| | 3 | 178 | 204 | 288 | 345 | 240 | 218 | 130 | 88 | 73 | 65 | 72 | 81 |

Frec.: Frecuencias de las puntuaciones.

En esta tabla (Tabla 9. Estadísticos descriptivos de los ítems del PHQ-9) aparecen los estadísticos descriptivos de los ítems del PHQ-9 conformado por los 9 ítems del cuestionario, más los tres ítems que de igual forma son parte del cuestionario referentes a la dificultad del desempeño de las áreas laboral, familiar y social.

Haciendo referencia a los 9 ítems, el ítem con menor media es el ítem 9, como era de esperar, ya que se trata del suicidio, que en determinados casos es una cualidad extrema de depresión, pero que no siempre se encuentra presente en el diagnóstico de depresión. Los siguientes ítems con puntuaciones menores son el ítem 8, (que trata sobre el enlentecimiento

o inquietud) y seguidamente el ítem 7 (sobre dificultades en la concentración). En la parte opuesta de los ítems con mayor media se sitúa el ítem 4 (referente al cansancio) con una media de 1,62. Aunque el cansancio es un buen indicador de depresión, no tiene relación biunívoca, por lo que puede ser explicado por otras muchas razones y no exclusivamente depresión. Por ello y pese a tener la media más elevada, no es el ítem más predictor de depresión.

La anhedonia, contemplada por el ítem 1 se sitúa como el tercer ítem con mayor media. Esto indica que, pese a ser una de las características más salientes de la depresión no tiene por qué expresarse de forma más saliente. Las implicaciones de esto conllevan a entender la anhedonia como una parte de la depresión definitoria de esta, pero no tiene por qué estar presente en su estadio más liviano. En cambio, el suicidio, representado por el ítem 9, recoge la expresión de una forma de depresión. Sin embargo, puede existir depresión sin pensamientos de suicidio y pensamientos de suicidio sin depresión.

La homogeneidad del ítem indicada mediante el error estándar, la varianza y desviación típica, indican la variabilidad observada siendo la más elevada en los ítems 3 y 5 y menor en el ítem 9. Para interpretar este índice hay que tener en cuenta que el rango de respuestas va de 0 a 3 y que las frecuencias de respuesta difieren entre los ítems. El ítem 9 presenta menor variación, ya que sus respuestas se concentran en frecuencias bajas ya que su distribución ofrece puntuaciones del 72,11% para la primera opción de ausencia de pensamientos suicidas, dejando un 16,00% para la opción de algunos días; un 5,59% para la opción de más de la mitad de los días y por último un 6,28% para la opción de casi todos los días. Esto implica un 27,88% de puntuación mayor que 1.

En cuanto a la distribución de frecuencias de los ítems puede observarse que en los cinco primeros ítems las frecuencias más elevadas corresponden a la puntuación 1 de varios días; mientras que el resto de los ítems presentan mayor frecuencia de puntuaciones en el

extremo inferior de 0 (para nada). Resalta el ítem 4 (sobre pérdida de energía) ya que es el ítem que presenta más puntuaciones elevadas. Así se refleja en su media, la cual es superior al resto de ítems. Por lo tanto puede afirmarse que en este estudio el rasgo más compartido por la muestra es la pérdida de energía.

El ítem 3 presenta una curtosis más baja que el resto de ítems indicando que la discrepancia entre las puntuaciones de sus extremos con las del centro no son tan diferentes. De forma contraria los ítems 8 y 9 presentan una curtosis positiva, a diferencia del resto de ítems. No obstante, el ítem 9 tiene un apuntamiento extremo denotando datos atípicos, lo cual es explicado por la concentración de sus frecuencias en el rango inferior de opciones de respuesta. Ya que el 72,11% de la muestra seleccionó 0 en el ítem 9, se refleja que hasta el percentil 70 las puntuaciones sean de 0.

En cuanto a la asimetría, todos los ítems resultaron con asimetría positiva. Esto indica una concentración en puntuaciones bajas. La asimetría del ítem 9 marca una gran cola a la izquierda, seguido por el ítem 8 con también una asimetría positiva; siendo los únicos dos ítems con asimetrías con puntuaciones mayores de 1. En el polo opuesto y con menor asimetría destaca el ítem 4 que presenta una asimetría positiva pero cercana a 0.

Pasando a analizar los descriptivos de las áreas afectadas, el área con una media más elevada es la laboral ya que es la que menos se ha seleccionado la opción 0 (de ninguna dificultad), en cambio es la que presenta más simetría y la más mesocúrtica de las tres opciones. Aunque sus puntuaciones no difieren mucho de las áreas familiar y social, es la que más puntuación acumulada tiene y menor variabilidad. Por tanto podría decirse que es el área más afectada de las tres. No obstante, la opción 3 de extremada dificultad en el desempeño del trabajo fue seleccionada en menor medida que con respecto a las áreas familiar y social. Las áreas social y familiar obtuvieron puntuaciones muy similares tanto en su variabilidad como distribución de frecuencias.

En la siguiente tabla aparece el porcentaje de la muestra que presentaba cada uno de los estadios de depresión según criterio original de puntos de corte.

*Tabla 10. Estadios de depresión en población general, según puntos de corte originales*

| Estadio depresivo | n | Porcentaje |
|---|---|---|
| Mínima | 333 | 28,7 |
| Media | 308 | 26,5 |
| Moderada | 214 | 18,4 |
| De moderada a severa | 148 | 12,7 |
| Severa | 159 | 13,7 |

Puede observarse en la tabla (Tabla 11. Estadísticos de dificultad según el modelo de Rasch) cómo el ítem 2 tiene un valor de discriminación mucho mayor que el resto, de forma que se posiciona como el mejor indicador de depresión. Sobre el parámetro b cabe destacar que los ítems 8 y 9 poseen índices positivos, indicando lo difícil que es puntuar en estos ítems. Además puede ser un indicador de una sintomatología más positiva y extrema. El parámetro c no implica interpretación adecuada tratarse de un test de ejecución típica. El parámetro d al indicar variabilidad indica que el ítem 9 posee un grado de variabilidad mucho mayor que el resto, mientras que el ítem 4 posee la menor variabilidad.

*Tabla 11. Estadísticos de dificultad según el modelo de Rasch*

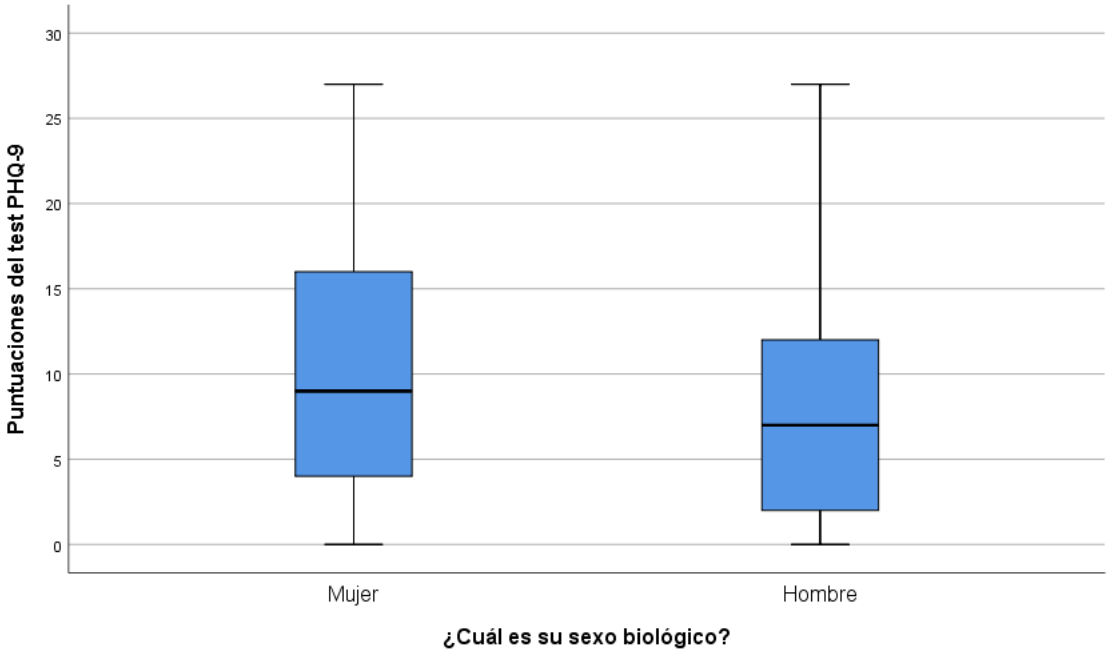
| Nº Ítem | a | b | c | d |
|---|---|---|---|---|
| Ítem 1 | 1,736 | -0,677 | 0,520 | 1,180 |
| Ítem 2 | 2,247 | -0,530 | 0,506 | 1,021 |
| Ítem 3 | 1,437 | -0,717 | 0,341 | 0,830 |
| Ítem 4 | 1,676 | -1,217 | 0,080 | 0,621 |
| Ítem 5 | 1,528 | -0,620 | 0,396 | 0,978 |
| Ítem 6 | 1,843 | -0,280 | 0,480 | 1,009 |
| Ítem 7 | 1,440 | -0,251 | 0,816 | 1,481 |
| Ítem 8 | 1,263 | 0,320 | 1,282 | 1,830 |
| Ítem 9 | 1,120 | 0,786 | 1,583 | 2,053 |

### *6.2.1.* Diferencias entre grupos *a priori*

#### *6.2.1.1. Depresión en función del sexo*

Con una media de 10,22 (7,38) para las mujeres y de 8,21 (7,42) para los varones, estas diferencias fueron estadísticamente significativas (p<,001): las puntuaciones de las mujeres en depresión son superiores a la de los varones.

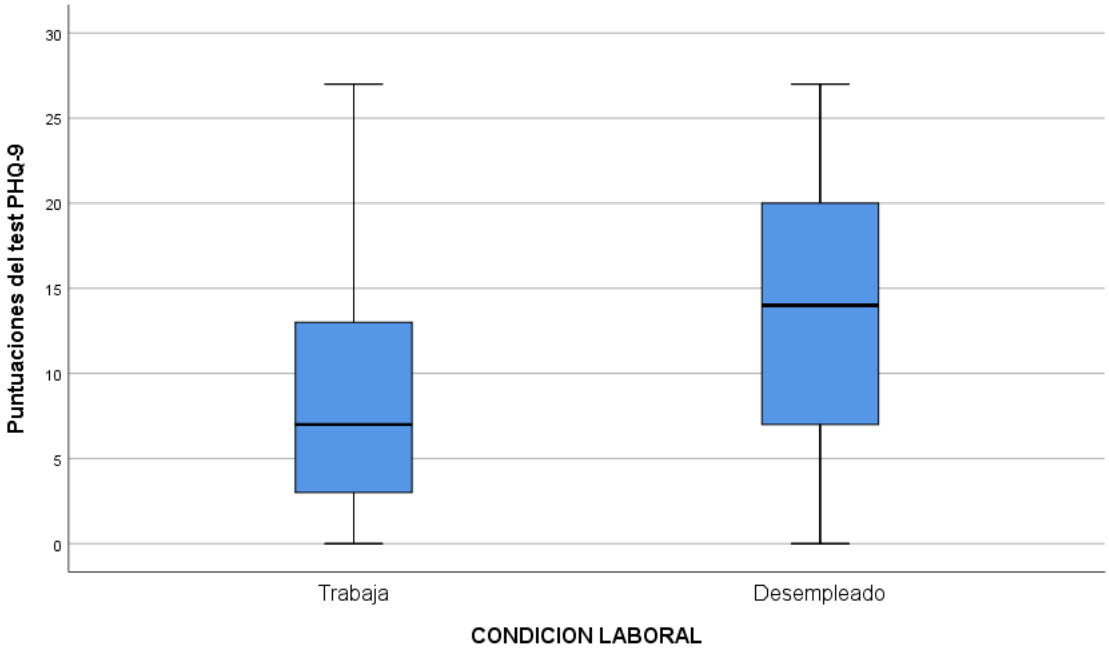
*Figura 12. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto al sexo*




#### *6.2.1.2. Depresión en función de la situación laboral*

Presentaban mayor depresión aquellas personas que se encontraban en situación de desempleo en comparación con las que tenían trabajo (p<,001) con medias (y desviaciones típicas) de 13,80 (8,06) y 8,73 (7,07), respectivamente. Lo que quiere decir que las puntuaciones en depresión de quienes están desempleados son superiores a quienes se encuentran trabajando.

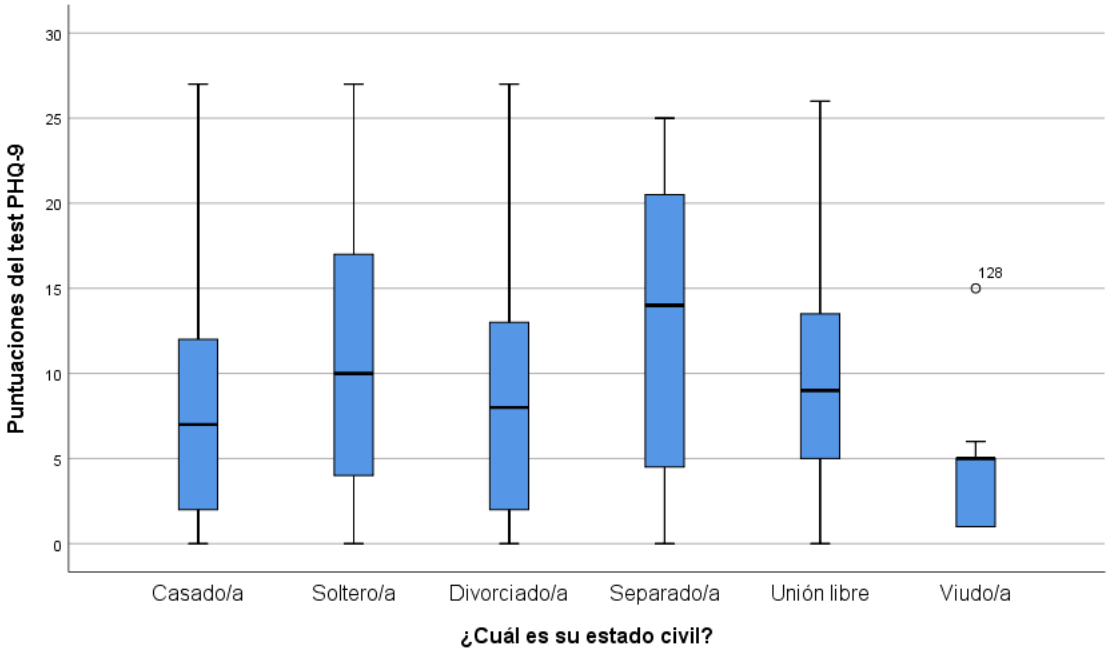
*Figura 13. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto a la situación laboral*




#### *6.2.1.3. Depresión en función del estado civil*

Las personas casadas presentaron una media y desviaciones típicas en la escala de depresión de 8,20 y 6,93; las personas solteras 11,02 y 7,71; las personas divorciadas 8,74 y 7,26; las personas separadas 12,68 y 8,88; las personas de unión libre 9,62 y 6,42; y las personas divorciadas de 4,88 y 4,31 (Figura 14. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto al estado civil).

*Figura 14. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto al estado civil*




Aunque las puntuaciones de cada grupo difieren y la mayor discrepancia de puntuaciones brutas se observa entre los grupos de personas solteras y enviudadas, las diferencias estadísticamente significativas solo se presentaron entre los grupos de casados y solteros (p<,001); siendo estos últimos quienes presentaban un mayor grado de depresión con respecto al grupo de casados.

#### *6.2.1.4. Depresión en función del nivel de estudios*

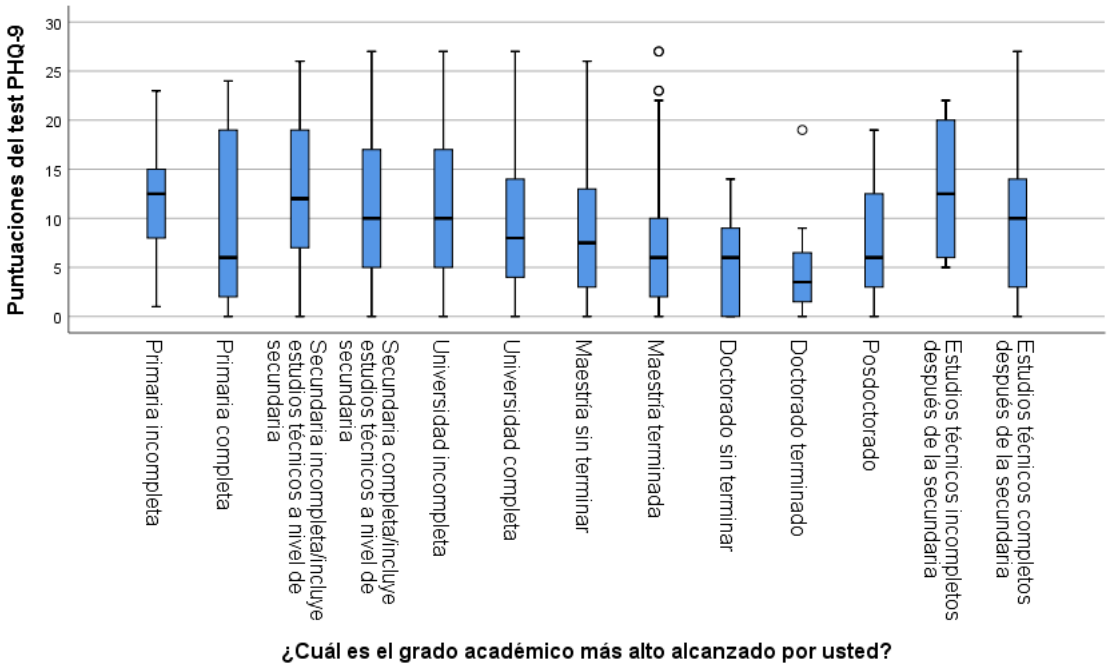
Las puntuaciones del test PHQ-9 en función de los grupos conformados por el nivel de estudios presentan medias dispares, siendo el grupo de doctorado terminado el que menos depresión presentó, y el grupo de estudios técnicos y secundaria sin completar el que más puntuación presentó (Tabla 12. Puntuaciones de depresión en función del nivel de estudios). No se encontraron diferencias entre los grupos de nivel de estudios en función de las puntuaciones en el test de depresión a excepción de los siguientes pares de grupos: doctorado terminado y secundaria completa (p=,042), doctorado terminado y secundaria incompleta (p=,009), Máster terminado y universidad completa (p=,027), Máster terminado y Universidad

completa (p<,001), Máster terminado y Secundaria completa (p<,001) y Máster terminado y Secundaria incompleta (p<,001) (Tabla 12. Puntuaciones de depresión en función del nivel de estudios).

*Tabla 12. Puntuaciones de depresión en función del nivel de estudios*

| Nivel de estudios | Media | Desviación Típica |
|---|---|---|
| Primaria incompleta | 12,00 | 7,32 |
| Primaria completa | 9,11 | 8,76 |
| Secundaria incompleta/incluye estudios técnicos a nivel de secundaria | 12,67 | 7,46 |
| Secundaria completa/incluye estudios técnicos a nivel de secundaria | 11,14 | 7,40 |
| Universidad incompleta | 10,86 | 7,63 |
| Universidad completa | 9,52 | 7,31 |
| Maestría sin terminar | 9,00 | 7,31 |
| Maestría terminada | 6,96 | 6,27 |
| Doctorado sin terminar | 5,57 | 4,46 |
| Doctorado terminado | 4,75 | 4,72 |
| Posdoctorado | 8,33 | 9,71 |
| Estudios técnicos incompletos después de la secundaria | 13,00 | 7,25 |
| Estudios técnicos completos después de la secundaria | 10,11 | 7,71 |

*Figura 15. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto al nivel de estudios*



#### 6.2.1.5. Depresión en función de la provincia de residencia

En cuanto a la provincia de procedencia de los participantes y pese a que el rango de puntuaciones por provincia toma valores medios de 8,03 a 11,70 no se observan diferencias asociadas a la ubicación ya que no se encontraron diferencias de las puntuaciones de depresión en función de la provincia de residencia (p=0,075) (Figura 16. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto a la provincia de residencia).

*Figura 16. Diagrama de caja y bigotes de las puntuaciones PHQ-9 con respecto a la provincia de residencia*

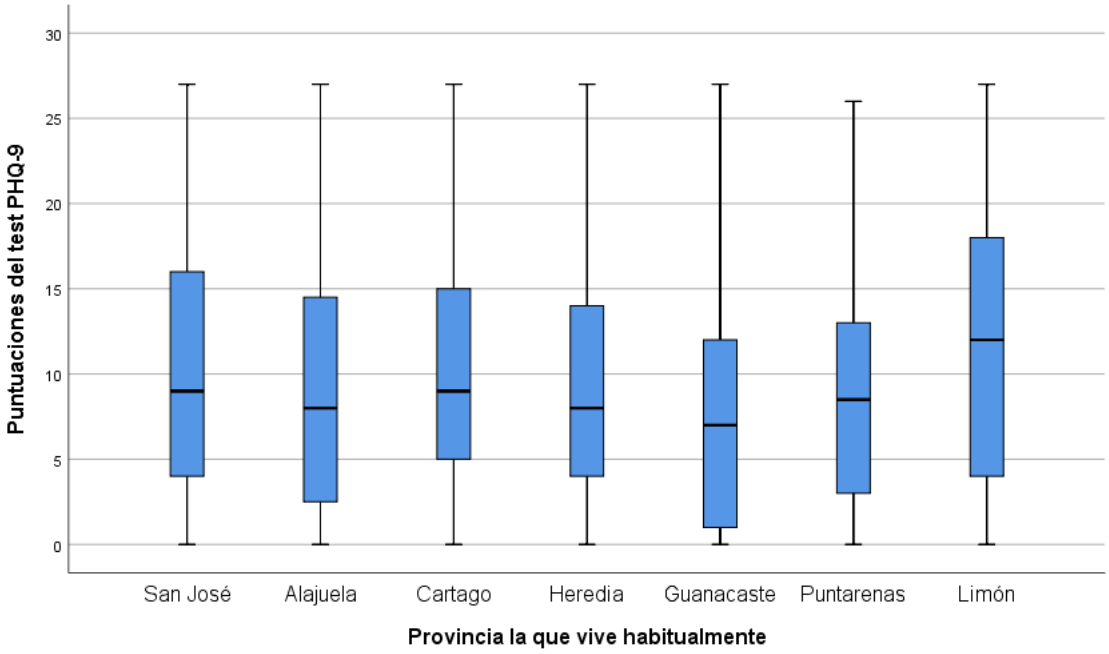



*Tabla 13. Puntuaciones de depresión en función del nivel de la provincia de residencia*

| Provincia | Media | Desviación Típica |
|---|---|---|
| San José | 10,23 | 7,55 |
| Alajuela | 9,26 | 7,62 |
| Cartago | 9,97 | 7,07 |
| Heredia | 9,54 | 7,15 |
| Guanacaste | 8,03 | 7,17 |
| Puntarenas | 9,76 | 7,99 |
| Limón | 11,70 | 7,80 |

## 6.3. Fiabilidad

Tras estudiar las propiedades psicométricas de los ítems, se observa cómo el ítem 4 se configura como el ítem más fácil, incluso cuando los ítems 1 y 2 configuran el núcleo de la evaluación de la depresión.

*Tabla 14. Propiedades de los ítems del test PHQ-9*

| Ítem | Factores de carga de los ítems | CMI | IDR | MSA normalizado | Media | Desviación típica | CI 95% | Mínimos Cuadrados no Ponderados | Comunalidad |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,84 | 3 | 0,4 | 0,94 | 1,20 | 1,01 | (1,12-1,28) | 0,87 | 0,70 |
| 2 | 0,87 | 3 | 0,39 | 0,90 | 1,18 | 1,06 | (1,10-1,26) | 0,93 | 0,76 |
| 3 | 0,80 | 3 | 0,45 | 0,95 | 1,36 | 1,13 | (1,27-1,44) | 0,83 | 0,64 |
| 4 | 0,82 | 3 | 0,54 | 0,94 | 1,62 | 1,06 | (1,54-1,70) | 0,87 | 0,66 |
| 5 | 0,82 | 3 | 0,42 | 0,94 | 1,27 | 1,10 | (1,19-1,36) | 0,84 | 0,67 |
| 6 | 0,84 | 3 | 0,37 | 0,95 | 1,12 | 1,14 | (1,04-1,21) | 0,88 | 0,71 |
| 7 | 0,80 | 3 | 0,31 | 0,95 | 0,94 | 1,00 | (0,87-1,02) | 0,83 | 0,63 |
| 8 | 0,73 | 2 | 0,21 | 0,95 | 0,63 | 0,92 | (0,56-0,70) | 0,79 | 0,53 |
| 9 | 0,66 | 2 | 0,15 | 0,94 | 0,46 | 0,86 | (0,40-0,52) | 0,76 | 0,68 |

CMI: Cuartil de Medias Ipsativas. IDR: Índice de Dificultad Relativa. MSA: Measure of Sampling Adequacy.

El Alpha de Cronbach de 0,928 obtenido mediante SPSS v25 revela una elevada consistencia interna. Además se realizó el análisis ítem-total, exponiendo que los 8 primeros ítems correlacionan más entre ellos, en cambio el noveno ítem, al ser eliminado, elevaba la consistencia interna del resto de ítems. Para hacer esta afirmación se debe considerar un cuarto decimal, ya que el valor de Alpha de Cronbach aparece truncado (0,9283) y el valor del Alpha al ser eliminado el ítem 9, aparece redondeado, siendo su valor de 0,9286. Pese a ser una diferencia mínima (0,0003) según el criterio de la consistencia interna medida a través de la correlación ítem-total, el ítem 9 sería un candidato a eliminar. Cabe destacar que el valor de Omega basado en los factores de carga, al forzar una solución unifactorial de máxima verosimilitud ofrece un valor de 0,931. Y cuando este índice es resultado de valores estandarizados, su valor es de 0,929. Por lo tanto la candidatura a la eliminación del ítem 9 queda en entredicho al no aumentar la fiabilidad del test tras la eliminación del ítem.

*Tabla 15. Descriptivos de fiabilidad del test PHQ-9*

| Nº ítem | Media de escala si el elemento se ha suprimido | Varianza de escala si el elemento se ha suprimido | Correlación total de elementos corregida | Correlación múltiple al cuadrado | Alfa de Cronbach si el elemento se ha suprimido |
|---|---|---|---|---|---|
| 1 | 8,60 | 43,784 | 0,782 | 0,642 | 0,917 |
| 2 | 8,62 | 42,665 | 0,827 | 0,726 | 0,914 |
| 3 | 8,44 | 43,030 | 0,738 | 0,588 | 0,920 |
| 4 | 8,18 | 43,527 | 0,759 | 0,630 | 0,919 |
| 5 | 8,53 | 43,052 | 0,761 | 0,602 | 0,919 |
| 6 | 8,68 | 42,353 | 0,788 | 0,653 | 0,917 |
| 7 | 8,86 | 44,465 | 0,735 | 0,565 | 0,920 |
| 8 | 9,17 | 46,151 | 0,661 | 0,467 | 0,925 |
| 9 | 9,34 | 47,604 | 0,585 | 0,390 | 0,929 |

Por otro lado, el Omega de McDonald es coincidente con el Alpha de Cronbach en 0,929.

Se calcularon los índices para averiguar la consistencia interna y la estructura factorial, aunque para el test PHQ-2 el cálculo del índice Omega de McDonald no pudiera ser posible dado que al constar solo de dos ítems la reducción de dimensiones no es posible y por ello la prueba de Bartlett resultó no significativa (Tabla 16. Propiedades psicométricas de los distintas versiones del test PHQ-9). De todas formas, todas las pruebas de Bartlett restantes resultaron altamente significativas (p<,001) y en todos los casos, incluido el PHQ-4, un solo factor fue extraído. Las diferencias en cuanto a los índices de fiabilidad se deben a los cálculos que utilizan los diferentes programas estadísticos, para generar la tabla (Tabla 16. Propiedades psicométricas de los distintas versiones del test PHQ-9) se utilizó el programa FACTOR.

*Tabla 16. Propiedades psicométricas de los distintas versiones del test PHQ-9*

| Estadísticos | PHQ-2 | PHQ-4 | PHQ-8 | PHQ-9 |
|---|---|---|---|---|
| Varianza explicada | 91.47 | 76.84 | 75.43 | 73.39 |
| KMO | 0.5 | 0.780 | 0.930 | 0.930 |
| Bartlett | 1,350.2 NS | 3,110.5** | 8,881.6** | 9,957.6** |
| Alpha (α) | 0.853 | 0.899 | 0.929 | 0.928 |
| Omega de McDonald (Oω) | No calculado | 0.898 | 0.930 | 0.931 |
| Ítems a eliminar según MSA | Ninguno | Ninguno | Ninguno | Ninguno |

MSA: Measure of Sampling Adequacy.

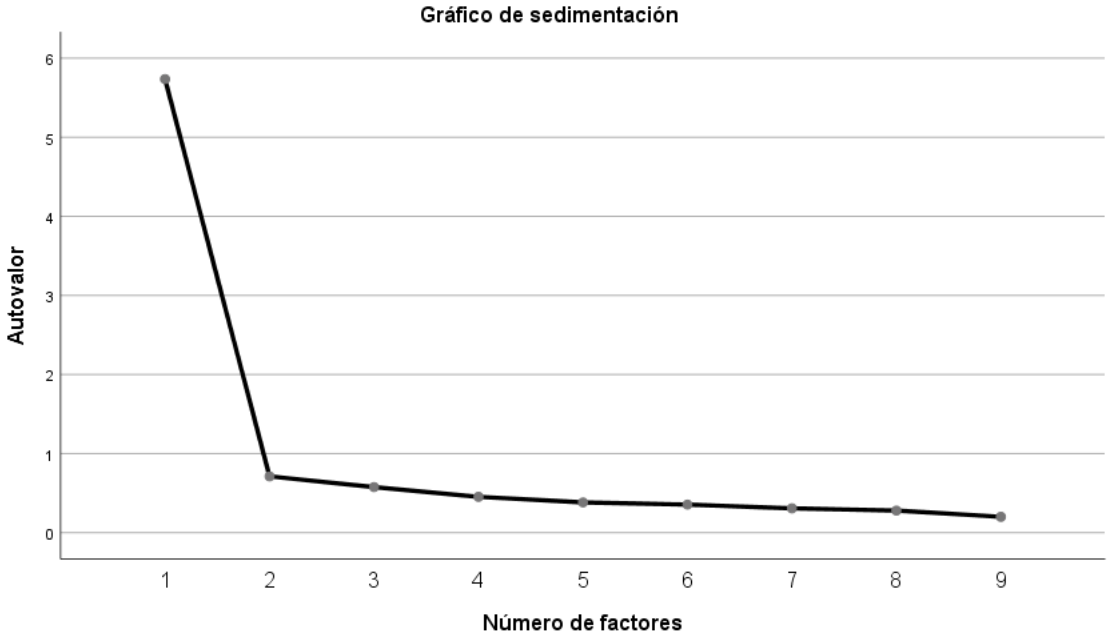
Siendo la prueba de esfericidad altamente significativa (p<0,001), un índice Kaiser–Meyer–Olkin (KMO) elevado (0,93), un determinante de la matriz de correlaciones muy bajo (0,001) y una bondad de ajuste altamente significativa ($\chi^2$=385,217; gl=27; p<0,001) se realizó un Análisis Factorial Exploratorio, el cual lograba explicar un 63,72% de la varianza total con un solo factor (Tabla 17. Varianza total explicada mediante el método de máxima verosimilitud).

*Tabla 17. Varianza total explicada mediante el método de máxima verosimilitud*

| Factor | Autovalores iniciales | | | Sumas de extracción de cargas al cuadrado | | |
|---|---|---|---|---|---|---|
| | Total | % de varianza explicada | % de varianza acumulada | Total | % de Varianza | % de Varianza acumulada |
| 1 | 5,735 | 63,727 | 63,727 | 5,344 | 59,373 | 59,373 |
| 2 | 0,712 | 7,908 | 71,635 | | | |
| 3 | 0,576 | 6,400 | 78,035 | | | |
| 4 | 0,453 | 5,035 | 83,070 | | | |
| 5 | 0,382 | 4,247 | 87,317 | | | |
| 6 | 0,354 | 3,938 | 91,255 | | | |
| 7 | 0,307 | 3,415 | 94,670 | | | |
| 8 | 0,280 | 3,106 | 97,775 | | | |
| 9 | 0,200 | 2,225 | 100 | | | |

En la figura (Figura 17. Gráfico de sedimentación de los ítems del PHQ-9) se muestra cómo se respeta la regla del codo en el número de factores seleccionado, aportando escasa información conforme se van añadiendo nuevos factores.

*Figura 17. Gráfico de sedimentación de los ítems del PHQ-9*




Ya que todos los ítems comparten factor, este puede llamarse factor depresión. La carga factorial que es explicada por cada uno de los ítems puede observarse en la siguiente tabla (Tabla 18. Factores de carga y errores de la varianza estandarizados y estimados de los ítems al factor depresión). Las cargas factoriales estandarizadas de la mayoría de los ítems superan el valor de 0,7. Esto implica que estos ítems quedan representados por un factor en particular. Sin embargo los ítems 8 (sobre enlentecimiento o inquietud) y 9 (sobre suicidio) no superan este estándar pese a tener valores cercanos a 0,7. Por ello y pese a que todos los ítems quedan explicados por un solo factor, los ítems 8 y 9 son los que menos aportan al constructo.

*Tabla 18. Factores de carga y errores de la varianza estandarizados y estimados de los ítems al factor depresión*

| Ítems | Estandarizados | | Estimados | |
|---|---|---|---|---|
| | Carga | Error de la varianza | Carga | Error de la varianza |
| 1 | 0,823 | 0,323 | 0,833 | 0,331 |
| 2 | 0,866 | 0,250 | 0,921 | 0,282 |
| 3 | 0,767 | 0,412 | 0,869 | 0,528 |
| 4 | 0,792 | 0,373 | 0,840 | 0,420 |
| 5 | 0,784 | 0,385 | 0,865 | 0,469 |
| 6 | 0,827 | 0,316 | 0,939 | 0,407 |
| 7 | 0,754 | 0,431 | 0,755 | 0,433 |
| 8 | 0,678 | 0,540 | 0,626 | 0,459 |

| | | | | |
|---|---|---|---|---|
| 9 | 0,612 | 0,626 | 0,525 | 0,462 |

## 6.4. Validez

### 6.4.1. Validez externa

Se realizó una tabla cruzada (Tabla 19. Porcentajes de respuesta de los criterios en función de las puntuaciones del test PHQ-9) con las puntuaciones del PHQ-9 con los diferentes criterios externos recogidos. Puede verse cómo el porcentaje de síes aumenta conforme aumenta la puntuación del test.

*Tabla 19. Porcentajes de respuesta de los criterios en función de las puntuaciones del test PHQ-9*

| Puntuación en el test PHQ-9 | Depresión | | Ansiedad | | Criterio Depresión-Ansiedad | | | Psicofármacos | | Alteraciones del sueño | | Tratamiento salud mental | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No | Sí | No | Sí | 0 | 1 | 2 | No | Sí | No | Sí | No | Sí |
| 0 | 100 | 0 | 96,4 | 3,6 | 96,4 | 3,6 | 0 | 90,2 | 9,8 | 99,1 | 0,9 | 94,6 | 5,4 |
| 1 | 100 | 0 | 72,5 | 27,5 | 72,5 | 27,5 | 0 | 90,2 | 9,8 | 90,2 | 9,8 | 96,1 | 3,9 |
| 2 | 100 | 0 | 93,1 | 6,9 | 93,1 | 6,9 | 0 | 93,1 | 6,9 | 87,9 | 12,1 | 98,3 | 1,7 |
| 3 | 96,2 | 3,8 | 57,7 | 42,3 | 57,7 | 38,5 | 3,8 | 80,8 | 19,2 | 80,8 | 19,2 | 96,2 | 3,8 |
| 4 | 95,0 | 5,0 | 75,0 | 25,0 | 73,3 | 23,3 | 3,3 | 81,7 | 18,3 | 83,3 | 16,7 | 81,7 | 18,3 |
| 5 | 85,5 | 14,5 | 56,5 | 43,5 | 51,6 | 38,7 | 9,7 | 82,3 | 17,7 | 77,4 | 22,6 | 85,5 | 14,5 |
| 6 | 90,8 | 9,2 | 64,6 | 35,4 | 61,5 | 32,3 | 6,2 | 83,1 | 16,9 | 61,5 | 38,5 | 93,8 | 6,2 |
| 7 | 82,9 | 17,1 | 58,6 | 41,4 | 52,9 | 35,7 | 11,4 | 77,1 | 22,9 | 65,7 | 34,3 | 84,3 | 15,7 |
| 8 | 84,7 | 15,3 | 49,2 | 50,8 | 45,8 | 42,4 | 11,9 | 76,3 | 23,7 | 62,7 | 37,3 | 93,2 | 6,8 |
| 9 | 82,7 | 17,3 | 42,3 | 57,7 | 36,5 | 51,9 | 11,5 | 67,3 | 32,7 | 50,0 | 50,0 | 82,7 | 17,3 |
| 10 | 86,4 | 13,6 | 39,0 | 61,0 | 33,9 | 57,6 | 8,5 | 71,2 | 28,8 | 50,8 | 49,2 | 79,7 | 20,3 |
| 11 | 76,6 | 23,4 | 25,5 | 74,5 | 21,3 | 59,6 | 19,1 | 61,7 | 38,3 | 38,3 | 61,7 | 87,2 | 12,8 |
| 12 | 61,4 | 38,6 | 29,5 | 70,5 | 20,5 | 50,0 | 29,5 | 70,5 | 29,5 | 45,5 | 54,5 | 81,8 | 18,2 |
| 13 | 80,0 | 20,0 | 37,1 | 62,9 | 34,3 | 48,6 | 17,1 | 68,6 | 31,4 | 51,4 | 48,6 | 82,9 | 17,1 |
| 14 | 72,4 | 27,6 | 31,0 | 69,0 | 27,6 | 48,3 | 24,1 | 62,1 | 37,9 | 34,5 | 65,5 | 79,3 | 20,7 |
| 15 | 64,5 | 35,5 | 22,6 | 77,4 | 19,4 | 48,4 | 32,3 | 58,1 | 41,9 | 25,8 | 74,2 | 80,6 | 19,4 |
| 16 | 64,5 | 35,5 | 38,7 | 61,3 | 25,8 | 51,6 | 22,6 | 61,3 | 38,7 | 35,5 | 64,5 | 83,9 | 16,1 |
| 17 | 40,7 | 59,3 | 18,5 | 81,5 | 7,4 | 44,4 | 48,1 | 44,4 | 55,6 | 22,2 | 77,8 | 74,1 | 25,9 |
| 18 | 46,7 | 53,3 | 26,7 | 73,3 | 10,0 | 53,3 | 36,7 | 50,0 | 50,0 | 13,3 | 86,7 | 80,0 | 20,0 |
| 19 | 44,8 | 55,2 | 13,8 | 86,2 | 6,9 | 44,8 | 48,3 | 51,7 | 48,3 | 31,0 | 69,0 | 75,9 | 24,1 |
| 20 | 51,9 | 48,1 | 14,8 | 85,2 | 11,1 | 44,4 | 44,4 | 59,3 | 40,7 | 22,2 | 77,8 | 85,2 | 14,8 |
| 21 | 26,7 | 73,3 | 20,0 | 80,0 | 6,7 | 33,3 | 60,0 | 33,3 | 66,7 | 33,3 | 66,7 | 86,7 | 13,3 |
| 22 | 40,0 | 60,0 | 20,0 | 80,0 | 10,0 | 40,0 | 50,0 | 60,0 | 40,0 | 20,0 | 80,0 | 75,0 | 25,0 |
| 23 | 34,5 | 65,5 | 13,8 | 86,2 | 13,8 | 20,7 | 65,5 | 48,3 | 51,7 | 6,9 | 93,1 | 79,3 | 20,7 |
| 24 | 50,0 | 50,0 | 36,4 | 63,6 | 22,7 | 40,9 | 36,4 | 63,6 | 36,4 | 31,8 | 68,2 | 68,2 | 31,8 |
| 25 | 41,2 | 58,8 | 11,8 | 88,2 | 5,9 | 41,2 | 52,9 | 70,6 | 29,4 | 17,6 | 82,4 | 70,6 | 29,4 |
| 26 | 20,0 | 80,0 | 0 | 100 | 0 | 20,0 | 80,0 | 40,0 | 60,0 | 30,0 | 70,0 | 70,0 | 30,0 |
| 27 | 15,8 | 84,2 | 5,3 | 94,7 | 5,3 | 10,5 | 84,2 | 36,8 | 63,2 | 26,3 | 73,7 | 52,6 | 47,4 |
| Total | 76,7 | 23,3 | 49,5 | 50,5 | 45,2 | 35,8 | 19,0 | 72,1 | 27,9 | 57,3 | 42,7 | 85,5 | 14,5 |

Por otro lado, las tablas de contingencia de los puntos de corte originales del PHQ-9 con respecto a los criterios pueden observarse a continuación (Tabla 20. Tabla de contingencia de la corrección del test PHQ-9 con los criterios externos).

*Tabla 20. Tabla de contingencia de la corrección del test PHQ-9 con los criterios externos*

| Corrección del PHQ-9 | Depresión | | Ansiedad | | Criterio Depresión-Ansiedad | | | Psicofármacos | | Alteraciones del sueño | | Tratamiento salud mental | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No | Sí | No | Sí | 1 | 2 | 3 | No | Sí | No | Sí | No | Sí |
| Mínima | 328 | 5 | 274 | 59 | 273 | 56 | 4 | 41 | 292 | 300 | 33 | 22 | 311 |
| Media | 263 | 45 | 169 | 139 | 155 | 122 | 31 | 69 | 239 | 197 | 111 | 37 | 271 |
| Moderada | 163 | 51 | 70 | 144 | 59 | 115 | 40 | 70 | 144 | 96 | 118 | 38 | 176 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| De moderada a severa | 78 | 70 | 36 | 112 | 21 | 72 | 55 | 69 | 79 | 38 | 110 | 31 | 117 |
| Severa | 59 | 100 | 26 | 133 | 17 | 51 | 91 | 75 | 84 | 35 | 124 | 41 | 118 |
| Total | 891 | 271 | 575 | 587 | 525 | 416 | 221 | 324 | 838 | 666 | 496 | 169 | 993 |

Considerando tanto los porcentajes de presentación de estos criterios como las puntuaciones brutas del test corregido, en todas ellas se observa una tendencia a incrementar las puntuaciones en función de las puntuaciones (o corrección del test). Todas las correlaciones fueron altamente significativas (p<,001). Además, la correlación entre el sumatorio de los ítems de depresión y el criterio de padecimientos de depresión resultó altamente significativa con un r de Pearson de 0,953.

*Tabla 21. Correlaciones de Pearson (inferior) y Spearman (superior) del test con los criterios externos y entre ellos*

| | Puntuaciones PHQ-9 | Depresión | Ansiedad | Psicofármacos | Sueño | Tratamiento Salud Mental |
|---|---|---|---|---|---|---|
| Puntuaciones PHQ-9 | 1 | ,957 | ,922 | ,850 | ,900 | ,805 |
| Depresión | ,953 | 1 | ,935 | ,913 | ,890 | ,758 |
| Ansiedad | ,903 | ,880 | 1 | ,876 | ,895 | ,723 |
| Psicofármacos | ,854 | ,912 | ,879 | 1 | ,837 | ,639 |
| Sueño | ,909 | ,853 | ,921 | ,839 | 1 | ,733 |
| Tratamiento Salud Mental | ,816 | ,778 | ,731 | ,687 | ,682 | 1 |

La significancia de las relaciones se constata tanto para las correlaciones de las puntuaciones con el test como para las pruebas Chi cuadrado. Ya que todas las pruebas Chi cuadrado realizadas entre la corrección del PHQ-9 con sus criterios resultaron altamente significativas (p<,001). Se obtuvieron valores de eta indicando relación media entre variables. Los valores de eta resultantes del análisis de las puntuaciones del PHQ-9 con respecto a los criterios fueron de 0,493 para la depresión, de 0,472 para la ansiedad, de 0,292 para el uso de psicofármacos, de 0,503 para alteraciones con el sueño, de 0,188 para tratamiento en salud mental. Por otro lado, el estadístico gamma fue de 0,693 para el criterio depresión-ansiedad.

Se realizaron pruebas con la Curva ROC para comprobar el funcionamiento de los criterios. Así, las áreas descritas utilizando diferentes criterios se muestran en la siguiente tabla (Tabla 22. Área bajo la curva normal de los diferentes test en función de sus criterios).

*Tabla 22. Área bajo la curva normal de los diferentes test en función de sus criterios*

| Puntuaciones de los test | Depresión | Ansiedad | Psicofármacos | Sueño | Tratamiento Salud Mental | Depresión-ansiedad |
|---|---|---|---|---|---|---|
| PHQ-2 | 0,798 | 0,745 | 0,658 | 0,741 | 0,633 | 0,800 |
| PHQ-4 | 0,801 | 0,794 | 0,696 | 0,772 | 0,661 | 0,813 |
| PHQ-8 | 0,820 | 0,788 | 0,696 | 0,806 | 0,658 | 0,824 |
| PHQ-8 | 0,829 | 0,789 | 0,694 | 0,804 | 0,661 | 0,834 |

Debido a que la correlación más alta fue la existente entre padecer depresión y el test de depresión, la capacidad explicativa del test considerando una sola dimensión, se utilizó el criterio de depresión como criterio externo para los siguientes análisis.

### 6.4.2. Validez interna

El rango de correlaciones entre los 9 primeros ítems tiene valores que van de 0,404 para la pareja de ítems 4-9 (energía y suicidio) a 0,760 para los ítems 2-6 (anhedonia y minusvaloración). Al considerar también los ítems 10, 11 y 12 la correlación entre el ítem 9 y el 11 cae a 0,394 y el resto de correlaciones añadidas no superan la puntuación de 0,715 para el par 10-11 (repercusiones en el trabajo y el hogar). Todas y cada una de estas correlaciones son altamente significativas ($p<0,001$). Esto indica una adecuada consistencia interna.

Además, el grado de variación conjunta se puede observar en la mitad superior de la tabla (Tabla 23. Matriz compuesta con las varianzas de los ítems (diagonal), las correlaciones inter-ítem (mitad inferior) y las covarianzas (mitad superior)), mostrando siempre valores positivos expresando relación directa. Los pares de ítems 4-9 y 8-9 presentan las covarianzas más pequeñas entre los 9 primeros ítems (0,368 y 0,362, respectivamente).

*Tabla 23. Matriz compuesta con las varianzas de los ítems (diagonal), las correlaciones inter-ítem (mitad inferior) y las covarianzas (mitad superior)*

| Ítems | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,025 | ,801 | ,693 | ,706 | ,711 | ,776 | ,642 | ,497 | ,410 | ,515 | ,494 | ,533 |
| 2 | ,744 | 1,130 | ,788 | ,758 | ,728 | ,917 | ,660 | ,561 | ,531 | ,540 | ,512 | ,608 |
| 3 | ,604 | ,654 | 1,283 | ,844 | ,809 | ,763 | ,633 | ,532 | ,423 | ,500 | ,504 | ,542 |
| 4 | ,657 | ,672 | ,702 | 1,126 | ,791 | ,754 | ,603 | ,491 | ,368 | ,511 | ,528 | ,531 |
| 5 | ,637 | ,621 | ,647 | ,675 | 1,217 | ,812 | ,676 | ,555 | ,423 | ,509 | ,536 | ,536 |
| 6 | ,675 | ,760 | ,593 | ,626 | ,649 | 1,288 | ,698 | ,571 | ,530 | ,531 | ,537 | ,605 |
| 7 | ,634 | ,620 | ,558 | ,567 | ,612 | ,614 | 1,003 | ,571 | ,423 | ,481 | ,478 | ,515 |
| 8 | ,532 | ,572 | ,509 | ,502 | ,546 | ,545 | ,618 | ,851 | ,362 | ,400 | ,409 | ,428 |
| 9 | ,471 | ,581 | ,434 | ,404 | ,446 | ,543 | ,492 | ,457 | ,738 | ,318 | ,296 | ,365 |
| 10 | ,613 | ,613 | ,533 | ,581 | ,557 | ,565 | ,580 | ,523 | ,447 | ,687 | ,518 | ,487 |
| 11 | ,558 | ,551 | ,509 | ,569 | ,555 | ,541 | ,545 | ,507 | ,394 | ,715 | ,765 | ,508 |
| 12 | ,584 | ,635 | ,532 | ,556 | ,539 | ,592 | ,571 | ,515 | ,472 | ,652 | ,645 | ,811 |

### 6.4.3. Validez concurrente

A continuación (Tabla 24. Correlaciones de Pearson (inferior) y Spearman (superior) entre los test) pueden verse las correlaciones de Pearson y Spearman existentes entre los test. Todas las correlaciones resultaron altamente significativas (p<,001) a excepción 2. Las correlación de Pearson entre la Escala de Miedo Social a la COVID-19 y la Escala de afrontamiento de la Resiliencia 14, ya que su correlación de Pearson se estableció en -0,093 con un p valor=0,001. Y la correlación entre la subescala de PROQL de Satisfacción por compasión y la Escala de Miedo Social a la COVID-19, resultó no significativa (p=0,373) con un valor de -0,026.

*Tabla 24. Correlaciones de Pearson (inferior) y Spearman (superior) entre los test*

| | Test | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | PHQ2 | 1 | 0,927 | 0,910 | 0,911 | 0,733 | 0,767 | 0,246 | 0,213 | 0,809 | 0,601 | 0,744 | -0,558 | -0,625 |
| B | PHQ4 | 0,926 | 1 | 0,916 | 0,916 | 0,931 | 0,917 | 0,338 | 0,289 | 0,866 | 0,657 | 0,748 | -0,532 | -0,620 |
| C | PHQ8 | 0,912 | 0,913 | 1 | 0,998 | 0,796 | 0,843 | 0,310 | 0,255 | 0,884 | 0,663 | 0,771 | -0,551 | -0,636 |
| D | PHQ9 | 0,911 | 0,911 | 0,996 | 1 | 0,796 | 0,843 | 0,301 | 0,250 | 0,887 | 0,664 | 0,779 | -0,563 | -0,647 |
| E | GAD2 | 0,725 | 0,931 | 0,786 | 0,783 | 1 | 0,937 | 0,387 | 0,324 | 0,804 | 0,626 | 0,650 | -0,435 | -0,534 |
| F | GAD-7 | 0,761 | 0,916 | 0,836 | 0,834 | 0,938 | 1 | 0,418 | 0,349 | 0,858 | 0,681 | 0,696 | -0,461 | -0,560 |
| G | FCV-19S | 0,251 | 0,349 | 0,316 | 0,303 | 0,396 | 0,432 | 1 | 0,558 | 0,384 | 0,391 | 0,204 | -0,123 | -0,199 |
| H | SFCV-19S | 0,211 | 0,287 | 0,250 | 0,242 | 0,320 | 0,348 | 0,569 | 1 | 0,307 | 0,336 | 0,147 | -0,048 | -0,103 |
| I | HSAY | 0,797 | 0,856 | 0,868 | 0,868 | 0,793 | 0,846 | 0,397 | 0,312 | 1 | 0,723 | 0,766 | -0,547 | -0,638 |
| J | PROQOL-STS | 0,593 | 0,651 | 0,658 | 0,655 | 0,616 | 0,675 | 0,401 | 0,334 | 0,719 | 1 | 0,660 | -0,316 | -0,450 |
| K | PROQOL-B | 0,735 | 0,741 | 0,758 | 0,764 | 0,642 | 0,686 | 0,208 | 0,149 | 0,768 | 0,664 | 1 | -0,688 | -0,672 |
| L | PROQOL-CS | -0,546 | -0,515 | -0,532 | -0,547 | -0,412 | -0,436 | -0,102 | -0,026 | -0,533 | -0,282 | -0,686 | 1 | 0,704 |
| M | RS14 | -0,587 | -0,584 | -0,604 | -0,618 | -0,500 | -0,524 | -0,184 | -0,093 | -0,606 | -0,422 | -0,643 | 0,674 | 1 |

Nota. A cada test se ha asociado con una letra para reducir el ancho de las columnas.

El test RS-14, así como la subescala de PROQOL sobre Satisfacción por Compasión obtuvieron correlaciones negativas al compararlos con los otros test, pero no entre ellos.

Es importante recordar que el test PHQ-2 está compuesto por los 2 primeros ítems del PHQ-9. De igual forma el PHQ-2 y el PHQ-4 solo difieren en que a este último se le añaden los 2 primeros ítems de ansiedad del test GAD-7. De esta forma, los test de la *A* a la *F* comparten cierta información entre sí; a excepción del PHQ-9 y el GAD-7, ya que se trata de 2 test independientes y de constructos diferentes (depresión y ansiedad). Esto explicaría la alta correlación entre los test. Por otro lado, la relación ente depresión y ansiedad también se considera elevada (0,843).

En un segundo orden de análisis, la relación entre los test de miedo con la depresión no es elevada (0,303 para el par PHQ-9 y FCV-19S; y 0,242 para el par PHQ-9 y SFCV-19S). La relación de correlación entre estos test y la ansiedad, aunque es superior, tampoco es elevada (0,432 para el par PHQ-9 y FCV-19S; y 0,348 para el par PHQ-9 y SFCV-19S).

En cuanto al test de Estrés, las correlaciones más elevadas se encuentran con la depresión (0,868) y la ansiedad (0,846), de carácter moderado con los test de miedo (0,397 para el par HSAY y FCV-19S; y 0,312 para el par HSAY y SFCV-19S).

Las subescalas de PROQOL presentan correlaciones elevadas para depresión y ansiedad, y moderadas para miedo. Cabe destacar que su correlación más elevada corresponde a la subescala de Burnout con la escala de estrés.

Por último, la escala de Resiliencia presenta valores por encima de 0,5 en todas las comparaciones con excepción de 3: la relación con miedo social señalada anteriormente como

no significativa; la relación con miedo físico (FCV-19S); y la de la subescala de ProQOL de Estrés Traumático secundario.

## 6.5. Sensibilidad y especificidad

### 6.5.1. Población general

Utilizando el criterio del padecimiento de la depresión, al realizar la curva ROC para los test se obtuvieron los siguientes resultados para cada uno de los test.

Para el PHQ-2, los criterios de Youden (selección del valor más grande) y del punto más cercano arriba a la izquierda (selección del valor más pequeño) son coincidentes en el punto 2,5 (Tabla 25. Puntos de corte para PHQ-2. Población general). Esto indica que una persona que presentase valores más pequeños que 3 se podría clasificar como no depresivo, dado que está más asociado a que no presenta depresión con respecto a que sí la presenta. De esta forma se puede detectar depresión con una sensibilidad de 73,1% y una especificidad del 70,4%.

*Tabla 25. Puntos de corte para PHQ-2. Población general*

| Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
|---|---|---|---|---|
| -1,0 | 1,000 | 1,000 | 0,000 | 1,000 |
| 0,5 | 0,993 | 0,699 | 0,293 | 0,699 |
| 1,5 | 0,952 | 0,567 | 0,385 | 0,569 |
| 2,5 | 0,731 | 0,296 | *0,434* | *0,400* |
| 3,5 | 0,609 | 0,181 | 0,428 | 0,431 |
| 4,5 | 0,406 | 0,103 | 0,303 | 0,603 |
| 5,5 | 0,288 | 0,053 | 0,235 | 0,714 |

Para el PHQ-8, su valor de referencia que discierne el positivo en depresión del negativo se situaría en la puntuación 10,5 (Tabla 26. Puntos de corte para PHQ-8. Población general). Dejando así unas sensibilidad y especificidad del 75,6% y 73,0% respectivamente.

*Tabla 26. Puntos de corte para PHQ-8. Población general*

| Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
|---|---|---|---|---|
| -1 | 1,000 | 1,000 | 0,000 | 1,000 |
| 0,5 | 1,000 | 0,873 | 0,127 | 0,873 |
| 1,5 | 1,000 | 0,815 | 0,185 | 0,815 |
| 2,5 | 0,996 | 0,751 | 0,245 | 0,751 |
| 3,5 | 0,993 | 0,691 | 0,301 | 0,691 |
| 4,5 | 0,974 | 0,631 | 0,343 | 0,631 |
| 5,5 | 0,945 | 0,569 | 0,376 | 0,572 |
| 6,5 | 0,915 | 0,496 | 0,419 | 0,503 |

| | | | | |
|---|---|---|---|---|
| 7,5 | 0,871 | 0,435 | 0,435 | 0,454 |
| 8,5 | 0,834 | 0,370 | 0,464 | 0,406 |
| 9,5 | 0,797 | 0,324 | 0,473 | 0,383 |
| 10,5 | 0,756 | 0,270 | *0,486* | *0,364* |
| 11,5 | 0,708 | 0,223 | 0,485 | 0,367 |
| 12,5 | 0,664 | 0,198 | 0,467 | 0,390 |
| 13,5 | 0,627 | 0,168 | 0,459 | 0,409 |
| 14,5 | 0,579 | 0,146 | 0,433 | 0,445 |
| 15,5 | 0,535 | 0,119 | 0,416 | 0,480 |
| 16,5 | 0,476 | 0,100 | 0,376 | 0,533 |
| 17,5 | 0,421 | 0,089 | 0,332 | 0,586 |
| 18,5 | 0,365 | 0,071 | 0,295 | 0,639 |
| 19,5 | 0,303 | 0,049 | 0,253 | 0,699 |
| 20,5 | 0,229 | 0,043 | 0,186 | 0,772 |
| 21,5 | 0,173 | 0,029 | 0,144 | 0,827 |
| 22,5 | 0,137 | 0,020 | 0,116 | 0,864 |
| 23,5 | 0,103 | 0,015 | 0,089 | 0,897 |

En cambio, para el PHQ-9 se obtiene una discrepancia entre lo sugerido por el índice de Youden y el índice ofrecido por el criterio del punto superior más a la izquierda. Dado que el índice de Youden predomina ante la discrepancia, ante un punto de corte de 11, para detectar la depresión el test PHQ-9 presenta una Sensibilidad de 78,6% y una especificidad de 72,1%.

*Tabla 27. Puntos de corte para PHQ-9. Población general*

| Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
|---|---|---|---|---|
| -1 | 1,000 | 1,000 | 0,000 | 1,000 |
| 0,5 | 1,000 | 0,874 | 0,126 | 0,874 |
| 1,5 | 1,000 | 0,817 | 0,183 | 0,817 |
| 2,5 | 1,000 | 0,752 | 0,248 | 0,752 |
| 3,5 | 0,993 | 0,696 | 0,297 | 0,696 |
| 4,5 | 0,982 | 0,632 | 0,350 | 0,632 |
| 5,5 | 0,948 | 0,572 | 0,376 | 0,575 |
| 6,5 | 0,926 | 0,506 | 0,420 | 0,512 |
| 7,5 | 0,882 | 0,441 | 0,441 | 0,457 |
| 8,5 | 0,849 | 0,385 | 0,464 | 0,414 |
| 9,5 | 0,815 | 0,337 | 0,479 | 0,384 |
| 10,5 | 0,786 | 0,279 | *0,507* | 0,352 |
| 11,5 | 0,745 | 0,239 | 0,506 | *0,349* |
| 12,5 | 0,683 | 0,209 | 0,474 | 0,380 |
| 13,5 | 0,657 | 0,177 | 0,479 | 0,386 |
| 14,5 | 0,627 | 0,154 | 0,474 | 0,403 |
| 15,5 | 0,587 | 0,131 | 0,455 | 0,434 |
| 16,5 | 0,546 | 0,109 | 0,437 | 0,467 |
| 17,5 | 0,487 | 0,097 | 0,391 | 0,522 |
| 18,5 | 0,428 | 0,081 | 0,347 | 0,578 |

| | | | | |
|---|---|---|---|---|
| 19,5 | 0,369 | 0,066 | 0,303 | 0,634 |
| 20,5 | 0,321 | 0,051 | 0,271 | 0,681 |
| 21,5 | 0,280 | 0,046 | 0,234 | 0,721 |
| 22,5 | 0,236 | 0,037 | 0,199 | 0,765 |
| 23,5 | 0,166 | 0,026 | 0,140 | 0,834 |
| 24,5 | 0,125 | 0,013 | 0,112 | 0,875 |
| 25,5 | 0,089 | 0,006 | 0,083 | 0,911 |
| 26,5 | 0,059 | 0,003 | 0,056 | 0,941 |

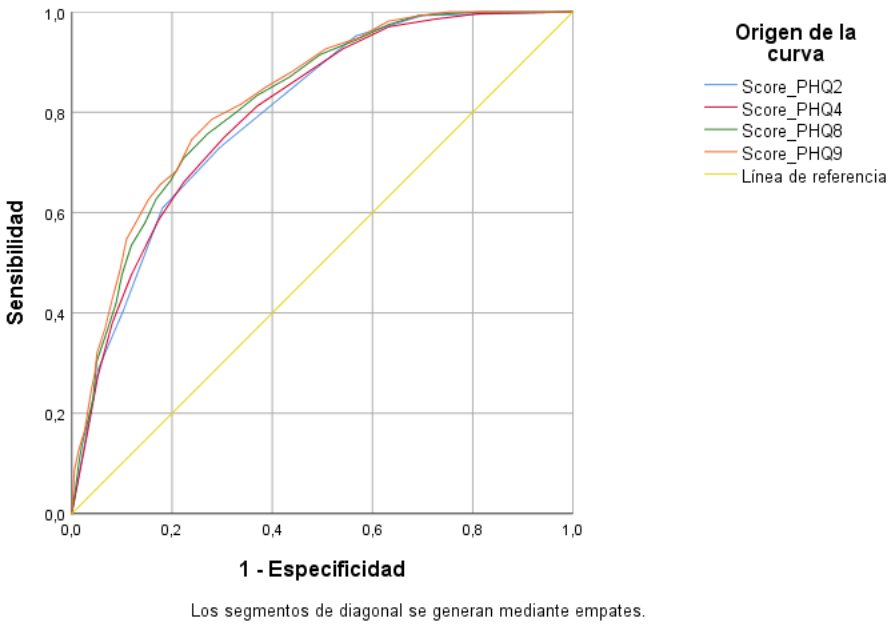
Para el test PHQ-4 también se obtuvieron valores discrepantes entre el índice de Youden y el del punto superior más a la izquierda. De igual forma, y ante la discrepancia, se selecciona el valor de 6 como punto de corte cuyas características son un 78,7% de sensibilidad y 68,2% de especificidad.

*Tabla 28. Puntos de corte para PHQ-4. Población general*

| Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
|---|---|---|---|---|
| -1,0 | 1,000 | 1,000 | 0,000 | 1,000 |
| 0,5 | 0,995 | 0,820 | 0,175 | 0,820 |
| 1,5 | 0,986 | 0,738 | 0,248 | 0,738 |
| 2,5 | 0,973 | 0,650 | 0,323 | 0,651 |
| 3,5 | 0,941 | 0,558 | 0,383 | 0,561 |
| 4,5 | 0,842 | 0,386 | 0,456 | 0,417 |
| 5,5 | 0,787 | 0,318 | *0,469* | 0,383 |
| 6,5 | 0,701 | 0,237 | 0,464 | *0,382* |
| 7,5 | 0,629 | 0,186 | 0,443 | 0,415 |
| 8,5 | 0,520 | 0,128 | 0,392 | 0,497 |
| 9,5 | 0,430 | 0,085 | 0,345 | 0,576 |
| 10,5 | 0,326 | 0,051 | 0,275 | 0,676 |
| 11,5 | 0,195 | 0,031 | 0,164 | 0,806 |

Dado que estos valores de los puntos de corte describen la curva ROC estas pueden representarse en la siguiente figura (Figura 18. Curva ROC de las variantes del PHQ-9 en función del criterio Depresión) representando de forma conjunta la capacidad discriminante del test.

*Figura 18. Curva ROC de las variantes del PHQ-9 en función del criterio Depresión*




### 6.5.2. Población según sexo

El test PHQ-2 no presentó diferencias en la detección del punto de corte óptimo para detectar depresión en función del sexo, situándose este valor tanto en hombres como en mujeres, en 3 (Tabla 29. Puntos de corte para PHQ-2. Valores en función del sexo). De esta forma y para varones el test PHQ-2 presenta una capacidad diagnóstica con valores de 75,0% de sensibilidad y 72,2% de especificidad. Mientras que en mujeres se sitúa en 72,5% de sensibilidad y 69,9% de especificidad.

*Tabla 29. Puntos de corte para PHQ-2. Valores en función del sexo*

| Sexo | Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
|---|---|---|---|---|---|
| Varón | -1 | 1,000 | 1,000 | 0,000 | 1,000 |
| | 0,5 | 1,000 | 0,604 | 0,396 | 0,604 |
| | 1,5 | 0,933 | 0,471 | 0,462 | 0,476 |
| | 2,5 | 0,750 | 0,278 | *0,472* | *0,374* |
| | 3,5 | 0,600 | 0,155 | 0,445 | 0,429 |
| | 4,5 | 0,383 | 0,086 | 0,297 | 0,623 |
| | 5,5 | 0,350 | 0,032 | 0,318 | 0,651 |
| Mujer | -1 | 1,000 | 1,000 | 0,000 | 1,000 |
| | 0,5 | 0,991 | 0,724 | 0,266 | 0,724 |
| | 1,5 | 0,957 | 0,592 | 0,365 | 0,594 |
| | 2,5 | 0,725 | 0,301 | *0,424* | *0,408* |

| | | | | |
|---|---|---|---|---|
| 3,5 | 0,611 | 0,188 | 0,424 | 0,431 |
| 4,5 | 0,412 | 0,108 | 0,304 | 0,598 |
| 5,5 | 0,270 | 0,058 | 0,212 | 0,732 |

Al analizar las propiedades diagnósticas dl test PHQ-8 considerando el sexo, se aprecia que presenta puntos de corte diferentes a los obtenidos para población general. De esta forma y para varones el valor óptimo de diferenciación presencia y ausencia de depresión es la puntuación 9 (sensibilidad 76,6% y especificidad 74,9%). Para mujeres el punto de corte óptimo se sitúa en 12 con una sensibilidad de 73,9% y una especificidad de 75,7% (Tabla 30. Puntos de corte para PHQ-8. Valores en función del sexo).

*Tabla 30. Puntos de corte para PHQ-8. Valores en función del sexo*

| Varón | | | | | Mujer | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left | Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
| 0,5 | 1,000 | 0,754 | 0,246 | 0,754 | 0,5 | 1,000 | 0,905 | 0,095 | 0,905 |
| 1,5 | 1,000 | 0,690 | 0,310 | 0,690 | 1,5 | 1,000 | 0,848 | 0,152 | 0,848 |
| 2,5 | 0,983 | 0,626 | 0,357 | 0,626 | 2,5 | 1,000 | 0,784 | 0,216 | 0,784 |
| 3,5 | 0,983 | 0,567 | 0,416 | 0,567 | 3,5 | 0,995 | 0,724 | 0,271 | 0,724 |
| 4,5 | 0,933 | 0,513 | 0,420 | 0,517 | 4,5 | 0,986 | 0,662 | 0,324 | 0,662 |
| 5,5 | 0,933 | 0,444 | 0,489 | 0,449 | 5,5 | 0,948 | 0,602 | 0,346 | 0,604 |
| 6,5 | 0,883 | 0,369 | 0,514 | 0,387 | 6,5 | 0,924 | 0,530 | 0,394 | 0,535 |
| 7,5 | 0,833 | 0,321 | 0,512 | 0,362 | 7,5 | 0,882 | 0,466 | 0,416 | 0,481 |
| 8,5 | 0,767 | 0,251 | *0,516* | *0,342* | 8,5 | 0,853 | 0,402 | 0,451 | 0,428 |
| 9,5 | 0,717 | 0,203 | 0,514 | 0,348 | 9,5 | 0,820 | 0,357 | 0,463 | 0,400 |
| 10,5 | 0,650 | 0,166 | 0,484 | 0,387 | 10,5 | 0,787 | 0,298 | 0,489 | 0,366 |
| 11,5 | 0,600 | 0,150 | 0,450 | 0,427 | 11,5 | 0,739 | 0,243 | *0,496* | *0,357* |
| 12,5 | 0,533 | 0,123 | 0,410 | 0,483 | 12,5 | 0,701 | 0,217 | 0,484 | 0,369 |
| 13,5 | 0,500 | 0,112 | 0,388 | 0,512 | 13,5 | 0,664 | 0,183 | 0,481 | 0,383 |
| 14,5 | 0,450 | 0,091 | 0,359 | 0,557 | 14,5 | 0,616 | 0,161 | 0,455 | 0,416 |
| 15,5 | 0,417 | 0,064 | 0,353 | 0,587 | 15,5 | 0,569 | 0,134 | 0,435 | 0,451 |
| 16,5 | 0,367 | 0,053 | 0,314 | 0,635 | 16,5 | 0,507 | 0,112 | 0,395 | 0,506 |
| 17,5 | 0,333 | 0,043 | 0,290 | 0,668 | 17,5 | 0,445 | 0,101 | 0,344 | 0,564 |
| 18,5 | 0,317 | 0,021 | 0,296 | 0,683 | 18,5 | 0,379 | 0,084 | 0,295 | 0,627 |
| 19,5 | 0,250 | 0,021 | 0,229 | 0,750 | 19,5 | 0,318 | 0,057 | 0,261 | 0,684 |
| 20,5 | 0,217 | 0,021 | 0,196 | 0,783 | 20,5 | 0,232 | 0,048 | 0,184 | 0,769 |
| 21,5 | 0,217 | 0,021 | 0,196 | 0,783 | 21,5 | 0,171 | 0,034 | 0,137 | 0,830 |
| 22,0 | 0,183 | 0,011 | 0,172 | 0,817 | 22,0 | 0,171 | 0,034 | 0,137 | 0,830 |
| 22,5 | 0,183 | 0,011 | 0,172 | 0,817 | 22,5 | 0,123 | 0,023 | 0,100 | 0,877 |
| 23,5 | 0,133 | 0,005 | 0,128 | 0,867 | 23,5 | 0,095 | 0,017 | 0,078 | 0,905 |

El test PHQ-9 también presenta diferencias en cuanto a sus puntos de corte en función del sexo, de esta forma y con una capacidad de detectar depresión de 71,6% en sensibilidad y

de 83,4% de especificidad, su punto de corte óptimo se sitúa en 11. Mientras que para mujeres su punto de corte se sitúa en 12 con una sensibilidad de 73,3% y especificidad de 73,9% (Tabla 31. Puntos de corte para PHQ-9. Valores en función del sexo).

*Tabla 31. Puntos de corte para PHQ-9. Valores en función del sexo*

| Varón | | | | | Mujer | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left | Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
| 0,5 | 1 | 0,759 | 0,241 | 0,759 | 0,5 | 1 | 0,905 | 0,095 | 0,905 |
| 1,5 | 1 | 0,695 | 0,305 | 0,695 | 1,5 | 1 | 0,849 | 0,151 | 0,849 |
| 2,5 | 1 | 0,631 | 0,369 | 0,631 | 2,5 | 1 | 0,784 | 0,216 | 0,784 |
| 3,5 | 0,983 | 0,583 | 0,400 | 0,583 | 3,5 | 0,995 | 0,726 | 0,269 | 0,726 |
| 4,5 | 0,967 | 0,519 | 0,448 | 0,520 | 4,5 | 0,986 | 0,662 | 0,324 | 0,662 |
| 5,5 | 0,933 | 0,444 | 0,489 | 0,449 | 5,5 | 0,953 | 0,607 | 0,346 | 0,609 |
| 6,5 | 0,9 | 0,38 | 0,520 | 0,393 | 6,5 | 0,934 | 0,54 | 0,394 | 0,544 |
| 7,5 | 0,85 | 0,332 | 0,518 | 0,364 | 7,5 | 0,891 | 0,47 | 0,421 | 0,482 |
| 8,5 | 0,767 | 0,278 | 0,489 | 0,363 | 8,5 | 0,872 | 0,413 | 0,459 | 0,432 |
| 9,5 | 0,733 | 0,241 | 0,492 | 0,360 | 9,5 | 0,839 | 0,362 | 0,477 | 0,396 |
| 10,5 | 0,717 | 0,166 | ***0,551*** | ***0,328*** | 10,5 | 0,806 | 0,31 | 0,496 | 0,366 |
| 11,5 | 0,65 | 0,155 | 0,495 | 0,383 | 11,5 | 0,773 | 0,261 | ***0,512*** | ***0,346*** |
| 12,5 | 0,55 | 0,134 | 0,416 | 0,470 | 12,5 | 0,72 | 0,229 | 0,491 | 0,362 |
| 13,5 | 0,533 | 0,118 | 0,415 | 0,482 | 13,5 | 0,692 | 0,193 | 0,499 | 0,363 |
| 14,5 | 0,5 | 0,102 | 0,398 | 0,510 | 14,5 | 0,664 | 0,168 | 0,496 | 0,376 |
| 15,5 | 0,45 | 0,086 | 0,364 | 0,557 | 15,5 | 0,626 | 0,143 | 0,483 | 0,400 |
| 16,5 | 0,433 | 0,064 | 0,369 | 0,571 | 16,5 | 0,578 | 0,121 | 0,457 | 0,439 |
| 17,5 | 0,4 | 0,048 | 0,352 | 0,602 | 17,5 | 0,512 | 0,109 | 0,403 | 0,500 |
| 18,5 | 0,367 | 0,037 | 0,330 | 0,634 | 18,5 | 0,445 | 0,092 | 0,353 | 0,563 |
| 19,5 | 0,317 | 0,032 | 0,285 | 0,684 | 19,5 | 0,384 | 0,075 | 0,309 | 0,621 |
| 20,5 | 0,3 | 0,021 | 0,279 | 0,700 | 20,5 | 0,327 | 0,058 | 0,269 | 0,675 |
| 21,5 | 0,25 | 0,021 | 0,229 | 0,750 | 21,5 | 0,289 | 0,053 | 0,236 | 0,713 |
| 22,5 | 0,233 | 0,016 | 0,217 | 0,767 | 22,5 | 0,237 | 0,043 | 0,194 | 0,764 |
| 23,5 | 0,2 | 0,016 | 0,184 | 0,800 | 23,5 | 0,156 | 0,028 | 0,128 | 0,844 |
| 24,5 | 0,167 | 0,011 | 0,156 | 0,833 | 24,5 | 0,114 | 0,014 | 0,100 | 0,886 |
| 25,5 | 0,15 | 0 | 0,150 | 0,850 | 25,5 | 0,071 | 0,007 | 0,064 | 0,929 |
| 26,5 | 0,1 | 0 | 0,100 | 0,900 | 26,5 | 0,047 | 0,004 | 0,043 | 0,953 |

El PHQ-4 presentó valores diferentes en función del sexo, siendo 1 punto más para las mujeres que para los hombres. De esta forma y con un punto de corte situado en 6,5 la sensibilidad para detectar depresión se sitúa en 70,8% y su especificidad en 73,4%. En cambio, para las mujeres, su punto de corte se sitúa en 5,5 con una sensibilidad y especificidad de 80,9% y 66,8%, respectivamente.

*Tabla 32. Puntos de corte para PHQ-4. Valores en función del sexo*

| Sexo | Puntuaciones | Sensibilidad | 1- especificidad | Youden | Closest top left |
|---|---|---|---|---|---|
| Varón | 0,5 | 1,000 | 0,729 | 0,271 | 0,729 |
| | 1,5 | 0,958 | 0,623 | 0,335 | 0,624 |
| | 2,5 | 0,938 | 0,543 | 0,395 | 0,547 |
| | 3,5 | 0,896 | 0,442 | 0,454 | 0,454 |
| | 4,5 | 0,792 | 0,307 | 0,485 | 0,371 |
| | 5,5 | 0,708 | 0,266 | 0,442 | 0,395 |
| | 6,5 | 0,667 | 0,186 | *0,481* | *0,381* |
| | 7,5 | 0,583 | 0,166 | 0,417 | 0,449 |
| | 8,5 | 0,479 | 0,106 | 0,373 | 0,532 |
| | 9,5 | 0,375 | 0,060 | 0,315 | 0,628 |
| | 10,5 | 0,333 | 0,035 | 0,298 | 0,668 |
| | 11,5 | 0,250 | 0,015 | 0,235 | 0,750 |
| Mujer | 0,5 | 0,994 | 0,845 | 0,149 | 0,845 |
| | 1,5 | 0,994 | 0,768 | 0,226 | 0,768 |
| | 2,5 | 0,983 | 0,679 | 0,304 | 0,679 |
| | 3,5 | 0,954 | 0,589 | 0,365 | 0,591 |
| | 4,5 | 0,855 | 0,407 | 0,448 | 0,432 |
| | 5,5 | 0,809 | 0,332 | *0,477* | *0,383* |
| | 6,5 | 0,711 | 0,251 | 0,460 | 0,383 |
| | 7,5 | 0,642 | 0,191 | 0,451 | 0,406 |
| | 8,5 | 0,532 | 0,133 | 0,399 | 0,487 |
| | 9,5 | 0,445 | 0,092 | 0,353 | 0,563 |
| | 10,5 | 0,324 | 0,055 | 0,269 | 0,678 |
| | 11,5 | 0,179 | 0,035 | 0,144 | 0,822 |

## 6.6. Puntos de corte

Solo se ha estudiado el positivo de depresión frente a su negativo, esto implica que solo se estudia un punto de corte y no los siguientes estadios de la depresión que implica este test. Por ello se han mantenido los rangos de las puntuaciones que implican los diversos estadios que mide el test. Teniendo en cuenta las puntuaciones para población general y población diferenciada por el sexo de pertenencia se ha elaborado una tabla (Tabla 33. Puntos de corte del test PHQ-9 originales y nuevos para población general y diferenciado por sexo) que especifica las puntuaciones para cada grupo poblacional.

*Tabla 33. Puntos de corte del test PHQ-9 originales y nuevos para población general y diferenciado por sexo*

| Grupo | Versión de PHQ | | | |
|---|---|---|---|---|
| | 2 | 4 | 8 | 9 |
| Original | 2,5 | 5,5 | 9,5 | 9,5 |
| General Costa Rica | 2,5 | 5,5 | 10,5 | 10,5 |
| Varón | 2,5 | 4,5 | 8,5 | 10,5 |
| Mujer | 2,5 | 5,5 | 11,5 | 11,5 |

Como puede apreciarse en la tabla, el test PHQ-2 permanece inalterado en cuanto a sus puntos de corte se refiere, tanto para la población general, como a la población diferenciada por sexo.

En el test PHQ-4 disminuye el punto de corte en los varones 1 punto, situándose en 4,5 como diferenciador del continuo depresión-ansiedad con su ausencia. Lo que implica que a partir de 5 se considera presencia del continuo depresión-ansiedad y una puntuación menor que 5 se considera ausencia de depresión-ansiedad. Esta se trata de la única diferencia realizada con respecto a sus puntos de corte, ya que las puntuaciones de las mujeres no se muestran afectadas. Conviene recordar que el test PHQ-4 también puede corregirse aisladamente disgregando el PHQ-2 del GAD-2.

En cuanto al test PHQ-8, sus puntos de corte originales coinciden con los del PHQ-9, pese a tener un rango menor por carecer del noveno ítem. Para la población general de Costa Rica, este punto de corte aumenta en 1 punto, situándose en 10,5. En cambio, al introducir la variable sexo, para varones este punto baja a 8,5; mientras que para mujeres se eleva a 11,5.

Para adecuar las puntuaciones del test PHQ-9, al igual que para el PHQ-8 se elevó un punto, situándose en 10,5 para detectar depresión en población general. No obstante, para detectar depresión en mujeres se elevó un punto, situándose en 11,5. El punto de corte de los varones permaneció inalterado.

Partiendo de estos nuevos puntos de corte se elaboró una tabla para adecuar las nuevas correcciones y facilitar su interpretación. Para elaborar esta tabla se tuvo en cuenta el rango de las puntuaciones, ya que al establecer nuevos puntos de corte, los rangos se ven alterados y es más fácil (o difícil) que la población obtenga la categoría asociada a ese rango porque este se ha agrandado o reducido. Como puede observarse en la tabla (Tabla 34. Puntos de corte propuestos para los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 tanto para población general como población en función del sexo) la distancia de los rangos de las puntuaciones no se vio alterada.

*Tabla 34. Puntos de corte propuestos para los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 tanto para población general como población en función del sexo*

| Grupo | Versión | Ausencia de trastorno depresivo mayor | Trastorno depresivo mayor | Mínima | Leve | Moderada | Moderadamente severa | Severa |
|---|---|---|---|---|---|---|---|---|
| Original | PHQ-2 | 0-2 | 3-6 | - | - | - | - | - |
| | PHQ-4 | - | - | 0-2 | 3-5 | 6-8 | - | 9-12 |
| | PHQ-8 | - | - | 0-4 | 5-9 | 10-14 | 15-19 | 20-24 |
| | PHQ-9 | - | - | 0-4 | 5-9 | 10-14 | 15-19 | 20-27 |
| Población general | PHQ-2 | 0-2 | 3-6 | - | - | - | - | - |
| | PHQ-4 | - | - | 0-2 | 3-5 | 6-8 | - | 9-12 |
| | PHQ-8 | - | - | 0-5 | 6-10 | 11-15 | 16-20 | 21-24 |
| | PHQ-9 | - | - | 0-5 | 6-10 | 11-15 | 16-20 | 21-27 |
| Masculino | PHQ-2 | 0-2 | 3-6 | - | - | - | - | - |
| | PHQ-4 | - | - | 0-1 | 2-4 | 5-7 | - | 8-12 |
| | PHQ-8 | - | - | 0-3 | 4-8 | 9-13 | 14-18 | 19-24 |
| | PHQ-9 | - | - | 0-5 | 6-10 | 11-15 | 16-20 | 21-27 |
| Femenino | PHQ-2 | 0-2 | 3-6 | - | - | - | - | - |
| | PHQ-4 | - | - | 0-2 | 3-5 | 6-8 | - | 9-12 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| PHQ-8 | - | - | 0-6 | 7-11 | 12-16 | 17-21 | 22-24 |
| PHQ-9 | - | - | 0-6 | 7-11 | 12-16 | 17-21 | 22-27 |

Nota. Los test PHQ-2, PHQ-8 y PHQ-9 se refieren a la depresión y por lo tanto sus estadios también. El test PHQ-4 se refiere al continuo depresión-ansiedad y por lo tanto también sus estadios.

### 6.6.1. Medición de Invarianza: Análisis Factorial Confirmatorio Multi Grupo (MGCFA)

Tras realizar el modelo identificando los sujetos a los grupos *a priori* seleccionados, se evaluó el ajuste para el modelo hacia la población en función de su sexo, obteniendo un $Chi^2$ altamente significativo ($Chi^2$=414,62; gl=54; p<0,001). Posteriormente se evaluó el ajuste de las diferencias ($\Delta\chi^2$), obteniendo diferencias entre los grupos ($Chi^2$=27,904; gl=27; p<0,001), comprobando así que aunque el modelo es válido tanto para varones, como para mujeres, es necesario un ajuste adicional como pueden serlo imponer puntos de corte diferenciados entre sexos. Esto implica que el modelo es válido para cada una de las submuestras, pero las submuestras difieren entre sí. Esto da indicios a pensar que aunque pueda medirse la depresión de forma adecuada con este test, la vivencia de la depresión no es la misma en cada submuestra.

### 6.6.2. Comparación de puntos de corte

Para conocer los cambios que los nuevos puntos de corte implican con respecto a los puntos de corte propuestos por defecto en el test original, se ofrece una comparación de puntos de corte en la siguiente tabla (Tabla 35. Visualización de los puntos de corte de las versiones de los test PHQ).

Puede observarse cómo los puntos de corte del PHQ-2 no sufren alteraciones. En el PHQ-4 para población general y mujeres no sufre modificaciones, pero sí para varones, distribuyéndose las puntuaciones de opciones de respuesta del test entre su siguiente opción de respuesta. Para los test PHQ-8 y PHQ-9 sí han sufrido cambios.

Para ayudar a la interpretación de los puntos de corte respetando los rangos se ofrece la siguiente tabla.

*Tabla 35. Visualización de los puntos de corte de las versiones de los test PHQ*

| | Test | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PHQ-2 | | | | PHQ-4 | | | | PHQ-8 | | | | PHQ-9 | | | |
| Puntuación | Original | General | Varón | Mujer | Original | General | Varón | Mujer | Original | General | Varón | Mujer | Original | General | Varón | Mujer |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| 5 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |
| 6 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 |
| 7 | | | | | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 8 | | | | | 3 | 3 | 5 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 9 | | | | | 5 | 5 | 5 | 5 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 |
| 10 | | | | | 5 | 5 | 5 | 5 | 3 | 2 | 3 | 2 | 3 | 2 | 2 | 2 |
| 11 | | | | | 5 | 5 | 5 | 5 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 2 |
| 12 | | | | | 5 | 5 | 5 | 5 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 13 | | | | | | | | | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 14 | | | | | | | | | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 |
| 15 | | | | | | | | | 4 | 3 | 4 | 3 | 4 | 3 | 3 | 3 |
| 16 | | | | | | | | | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 3 |
| 17 | | | | | | | | | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 18 | | | | | | | | | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 19 | | | | | | | | | 4 | 4 | 5 | 4 | 4 | 4 | 4 | 4 |
| 20 | | | | | | | | | 5 | 4 | 5 | 4 | 5 | 4 | 4 | 4 |
| 21 | | | | | | | | | 5 | 5 | 5 | 4 | 5 | 5 | 5 | 4 |
| 22 | | | | | | | | | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 23 | | | | | | | | | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 24 | | | | | | | | | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 25 | | | | | | | | | | | | | 5 | 5 | 5 | 5 |
| 26 | | | | | | | | | | | | | 5 | 5 | 5 | 5 |
| 27 | | | | | | | | | | | | | 5 | 5 | 5 | 5 |

Nota. Para el test PHQ-2, el 1 se refiere a Ausencia de trastorno depresivo mayor y el 2 a su presencia. Para el resto de test 1: Mínima; 2: Leve; 3: Moderada; 4: Moderadamente severa y 5: Severa.

Con los nuevos puntos de corte generados se comparó la distribución de frecuencias del corte original de los test con respecto a los nuevos cortes.

*Tabla 36. Comparación de las frecuencias de las puntuaciones en función de los diferentes puntos de corte.*

| Versión PHQ | Tipo de población | Estadístico | Puntuaciones | | | | |
|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 |
| 2 | General[a] | Frec. | 700 | 462 | | | |
| | | % | 60,2 | 39,8 | | | |
| 4 | Original[b] | Frec. | 335 | 354 | 238 | | 235 |
| | | % | 28,8 | 30,5 | 20,5 | | 20,2 |
| | Varón | Frec. | 250 | 363 | 235 | | 314 |
| | | % | 21,5 | 31,2 | 20,2 | | 27,0 |
| 8 | Original | Frec. | 336 | 321 | 218 | 161 | 126 |
| | | % | 29 | 28 | 19 | 14 | 11 |
| | General | Frec. | 399 | 317 | 195 | 151 | 100 |
| | | % | 34,3 | 27,3 | 16,8 | 13,0 | 8,6 |
| | Varón | Frec. | 277 | 329 | 236 | 158 | 162 |
| | | % | 23,8 | 28,3 | 20,3 | 13,6 | 13,9 |
| | Mujer | Frec. | 472 | 299 | 173 | 145 | 73 |
| | | % | 40,6 | 25,7 | 14,9 | 12,5 | 6,3 |
| 9 | Original | Frec. | 333 | 308 | 214 | 148 | 159 |
| | | % | 28,7 | 26,5 | 18,4 | 12,7 | 13,7 |
| | Varón[c] | Frec. | 395 | 305 | 186 | 144 | 132 |
| | | % | 34,0 | 26,2 | 16,0 | 12,4 | 11,4 |
| | Mujer | Frec. | 460 | 287 | 170 | 128 | 117 |
| | | % | 39,6 | 24,7 | 14,6 | 11,0 | 10,1 |

[a]: En el PHQ2 los puntos de corte coinciden para las versiones original, población general, varones y mujeres. [b]: En el PHQ4 los puntos de corte coinciden para las versiones original, población general y mujeres. [c]: en el PHQ-9 los puntos de corte coindicen para las versiones de varón y población general. Frec.: Frecuencias. %: Porcentajes. NS: No Significativo.

Para el PHQ-2 las puntuaciones no variaron.

Utilizando los nuevos puntos de corte del test PHQ-4, los varones vieron alteradas sus clasificaciones, aumentando la severidad del continuo depresión-ansiedad pasando de un 71,2% de positivos a un 78,5% y presentando un p-valor de <0,001.

En cuanto al PHQ-8, el porcentaje de varones clasificados como depresivos aumentó, pasando del 71% al 76,2%. Con estos nuevos puntos de corte el porcentaje de varones que obtuvieron la mínima puntuación aumentó un 5,3% (62 sujetos); en mujeres lo hizo en un

10,9%. En cambio, tratándose del resto de puntuaciones, los porcentajes de personas clasificadas como depresivas y sus estadios más severos, se vieron reducidos. Al comparar la versión del PHQ-8 original resulta un p-valor de 0,04 en comparación con la versión nueva general; 0,026 comparado con la versión de varones y <0,001 comparado con la versión de mujeres.

No obstante, y dado que originalmente se aplica la misma forma de corregir el PHQ-8 y PHQ-9, el cambio en la detección de depresión en el PHQ-9 no aumenta, sino que disminuye desde un 71,3% a un 66% en varones, lo que suponen unas diferencias estadísticamente significativas al 0,05 (p-valor= 0,044) y en mujeres un cambio a 60,4% de positivos, lo que suponen diferencias altamente significativas en mujeres (p-valor<0,001).

A continuación se presentan las comparaciones de los puntos de corte por cada una de sus puntuaciones. Con esta tabla (Tabla 37. Comparación entre puntos de corte: puntuaciones originales y nuevas.) puede conocerse hacia qué casillas ha migrado las puntuaciones. Así y puede observarse cómo en el PHQ-2 no se sufren alteraciones, en el PHQ-4 sus nuevos puntos de corte han clasificado a los varones como más ansiosos y depresivos, quedando sin alteraciones a población general y mujeres. Con el PHQ-8, los varones se han visto incrementadas sus clasificaciones como más depresivos, en cambio para mujeres se han clasificado como menos depresivas gracias a estos nuevos puntos de corte. Con respecto al PHQ-9 menos varones, mujeres y población general han sido clasificados como positivos.

*Tabla 37. Comparación entre puntos de corte: puntuaciones originales y nuevas.*

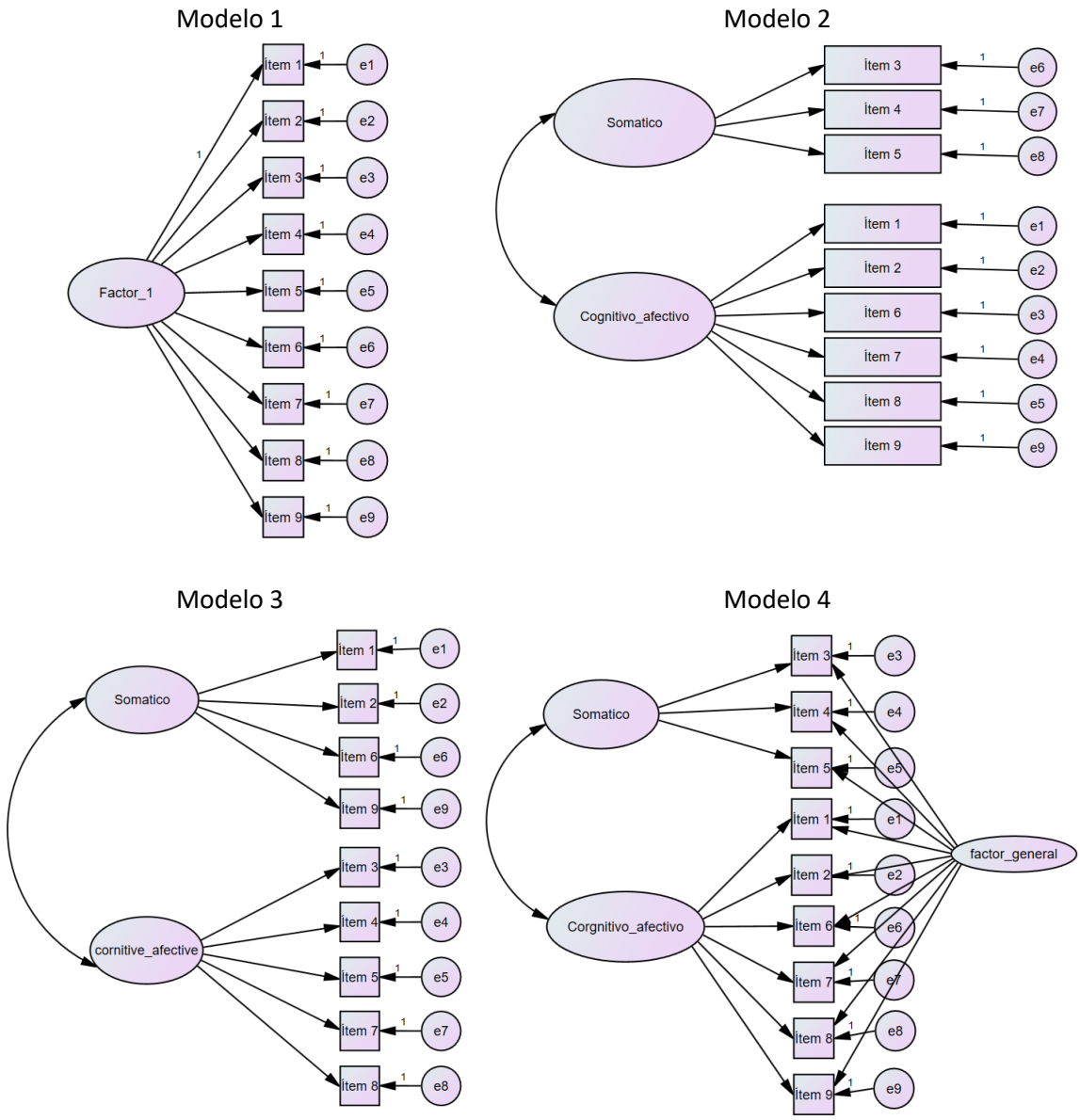
| | Versión nueva de PHQ | Corte | PHQ-2 | | PHQ-4 | | | | PHQ-8 | | | | | PHQ-9 | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | TDM (-) | TDM (+) | Mínima | Leve | Moderada | Severa | Mínima | Leve | Moderada | Moderada-mente severa | Severa | Mínima | Leve | Moderada | Moderada-mente severa | Severa | |
| Varones | PHQ-2 | TDM (-) | 700 | 0 | | | | | | | | | | | | | | | 700 |
| | | TDM (+) | 0 | 462 | | | | | | | | | | | | | | | 462 |
| | PHQ-4 | Mínima | | | 250 | 0 | 0 | 0 | | | | | | | | | | | 250 |
| | | Leve | | | 85 | 278 | 0 | 0 | | | | | | | | | | | 363 |
| | | Moderada | | | 0 | 76 | 159 | 0 | | | | | | | | | | | 235 |
| | | Severa | | | 0 | 0 | 79 | 235 | | | | | | | | | | | 314 |
| | PHQ-8 | Mínima | | | | | | | 277 | 0 | 0 | 0 | 0 | | | | | | 227 |
| | | Leve | | | | | | | 59 | 270 | 0 | 0 | 0 | | | | | | 329 |
| | | Moderada | | | | | | | 0 | 51 | 185 | 0 | 0 | | | | | | 236 |
| | | Moderadamente severa | | | | | | | 0 | 0 | 33 | 125 | 0 | | | | | | 158 |
| | | Severa | | | | | | | 0 | 0 | 0 | 36 | 126 | | | | | | 162 |
| | PHQ-9 | Mínima | | | | | | | | | | | | 333 | 62 | 0 | 0 | 0 | 395 |
| | | Leve | | | | | | | | | | | | 0 | 246 | 59 | 0 | 0 | 305 |
| | | Moderada | | | | | | | | | | | | 0 | 0 | 155 | 31 | 0 | 186 |
| | | Moderadamente severa | | | | | | | | | | | | 0 | 0 | 0 | 117 | 27 | 144 |
| | | Severa | | | | | | | | | | | | 0 | 0 | 0 | 0 | 132 | 132 |
| Mujeres | PHQ-2 | TDM (-) | 700 | 0 | | | | | | | | | | | | | | | 700 |
| | | TDM (+) | 0 | 462 | | | | | | | | | | | | | | | 462 |
| | PHQ-4 | Mínima | | | 335 | 0 | 0 | 0 | | | | | | | | | | | 335 |
| | | Leve | | | 0 | 354 | 0 | 0 | | | | | | | | | | | 354 |
| | | Moderada | | | 0 | 0 | 238 | 0 | | | | | | | | | | | 238 |
| | | Severa | | | 0 | 0 | 0 | 235 | | | | | | | | | | | 235 |
| | PHQ-8 | Mínima | | | | | | | 336 | 136 | 0 | 0 | 0 | | | | | | 472 |
| | | Leve | | | | | | | 0 | 185 | 114 | 0 | 0 | | | | | | 299 |
| | | Moderada | | | | | | | 0 | 0 | 104 | 69 | 0 | | | | | | 173 |
| | | Moderadamente severa | | | | | | | 0 | 0 | 0 | 92 | 53 | | | | | | 145 |
| | | Severa | | | | | | | 0 | 0 | 0 | 0 | 73 | | | | | | 73 |
| | PHQ-9 | Mínima | | | | | | | | | | | | 333 | 127 | 0 | 0 | 0 | 460 |
| | | Leve | | | | | | | | | | | | 0 | 181 | 106 | 0 | 0 | 287 |
| | | Moderada | | | | | | | | | | | | 0 | 0 | 108 | 62 | 0 | 170 |
| | | Moderadamente severa | | | | | | | | | | | | 0 | 0 | 0 | 86 | 42 | 128 |
| | | Severa | | | | | | | | | | | | 0 | 0 | 0 | 0 | 117 | 117 |
| Población general | PHQ-2 | TDM (-) | 700 | 0 | | | | | | | | | | | | | | | 700 |
| | | TDM (+) | 0 | 462 | | | | | | | | | | | | | | | 462 |
| | PHQ-4 | Mínima | | | 335 | 0 | 0 | 0 | | | | | | | | | | | 335 |
| | | Leve | | | 0 | 354 | 0 | 0 | | | | | | | | | | | 354 |
| | | Moderada | | | 0 | 0 | 238 | 0 | | | | | | | | | | | 238 |
| | | Severa | | | 0 | 0 | 0 | 235 | | | | | | | | | | | 235 |
| | PHQ-8 | Mínima | | | | | | | 336 | 63 | 0 | 0 | 0 | | | | | | 339 |
| | | Leve | | | | | | | 0 | 258 | 59 | 0 | 0 | | | | | | 317 |
| | | Moderada | | | | | | | 0 | 0 | 159 | 36 | 0 | | | | | | 195 |
| | | Moderadamente severa | | | | | | | 0 | 0 | 0 | 125 | 26 | | | | | | 151 |
| | | Severa | | | | | | | 0 | 0 | 0 | 0 | 100 | | | | | | 100 |
| | PHQ-9 | Mínima | | | | | | | | | | | | 333 | 62 | 0 | 0 | 0 | 395 |
| | | Leve | | | | | | | | | | | | 0 | 246 | 59 | 0 | 0 | 305 |
| | | Moderada | | | | | | | | | | | | 0 | 0 | 155 | 31 | 0 | 186 |
| | | Moderadamente severa | | | | | | | | | | | | 0 | 0 | 0 | 117 | 27 | 144 |
| | | Severa | | | | | | | | | | | | 0 | 0 | 0 | 0 | 132 | 132 |
| | | Total | 700 | 462 | 335 | 354 | 238 | 235 | 336 | 321 | 218 | 161 | 126 | 333 | 308 | 214 | 148 | 159 | 1162 |

## 6.7. Modelo de ecuaciones estructurales

Se evaluaron 4 modelos de ecuaciones estructurales encontradas en la literatura.

Para comprobar que el modelo teórico se ajustaba a la realidad, primero se estudiaron los 4 modelos en los que el test PHQ-9 ha demostrado estructurarse en otras culturas e idiomas.

*Figura 19. Modelos teóricos de estructuras factoriales*



Tras comprobar el ajuste de los modelos a los datos de la muestra, se pudo comprobar que, de entre ellos, el único modelo válido era el primero de estructura factorial única comprobada previamente tras realizar análisis factorial. De esta forma un ajuste mediante $Chi^2$ altamente significativo demuestra el ajuste a este modelo unifactorial, mientras que los demás modelos no demuestran ajuste.

*Tabla 38. Índices de ajuste de los modelos factoriales del PHQ-9*

| | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 |
|---|---|---|---|---|
| $Chi^2$ | 386,716 | ND | ND | ND |
| gl | 27 | ND | ND | ND |
| p-valor | <0,001 | ND | ND | ND |
| TLI | 0,931 | 0 | 0 | 0 |
| CFI | 0,948 | 0 | 0 | 0 |
| RMSEA | 0,107 | 0,407 | 0,407 | 0,407 |

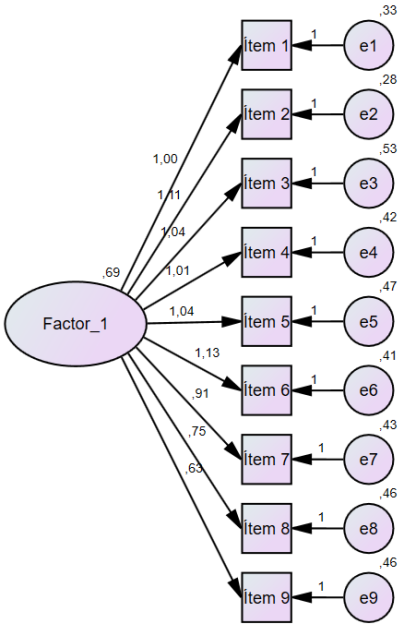
ND: No determinado

Al realizar el análisis el programa mostraba que el modelo no se correspondía con los datos, obteniéndose un mensaje de error

Entonces y tras comprobarse que el único modelo que ofrecía resultados adecuados era el modelo 1, se realiza el Modelo de Ecuaciones Estructurales con sus índices entre variables latentes.

*Figura 20*

*Modelo resultante de ecuaciones estructurales*




Tras ello, los índices que pueden aportar más información al modelo se detallan en la tabla (Tabla 39 Índices de ajuste del modelo de ecuaciones estructurales del PHQ-9.). Teniendo en cuenta toda la muestra, con los resultados obtenidos puede afirmarse que el modelo teórico propuesto (modelo 1) se ajusta al modelo real, ya que los índices obtenidos así lo indican: $Chi^2$=386,716; gl=27; p<0,001.

*Tabla 39* Índices de ajuste del modelo de ecuaciones estructurales del PHQ-9.

| | Parámetro | Por defecto | Saturado | Independiente |
|---|---|---|---|---|
| CMIN | NPAR | 18 | 45 | 9 |
| | CMIN | 386,716 | 0 | 6964,263 |
| | DF | 27 | 0 | 36 |
| | P | 0 | - | 0 |
| | CMIN/DF | 14,323 | - | 193,452 |
| RMR, GFI | RMR | 0,038 | 0 | 0,581 |
| | GFI | 0,927 | 1 | 0,261 |
| | AGFI | 0,878 | - | 0,077 |
| | PGFI | 0,556 | - | 0,209 |
| Comparaciones línea de base | NFI | 0,944 | 1 | 0 |
| | RFI | 0,926 | - | 0 |
| | IFI | 0,948 | 1 | 0 |
| | TLI | 0,931 | - | 0 |
| | CFI | 0,948 | 1 | 0 |
| Medidas Parsimoniosas ajustadas | PRATIO | 0,75 | 0 | 1 |
| | PNFI | 0,708 | 0 | 0 |
| | PCFI | 0,711 | 0 | 0 |
| NCP | NCP | 359,716 | 0 | 6928,263 |
| | LO 90 | 299,838 | 0 | 6657,238 |
| | HI 90 | 427,036 | 0 | 7205,574 |
| FMIN | FMIN | 0,333 | 0 | 5,999 |
| | F0 | 0,31 | 0 | 5,967 |
| | LO 90 | 0,258 | 0 | 5,734 |
| | HI 90 | 0,368 | 0 | 6,206 |
| AIC | AIC | 422,716 | 90 | 6982,263 |
| | BCC | 423,029 | 90,782 | 6982,419 |
| | BIC | 513,758 | 317,605 | 7027,784 |
| | CAIC | 531,758 | 362,605 | 7036,784 |
| ECVI | ECVI | 0,364 | 0,078 | 6,014 |
| | LO 90 | 0,313 | 0,078 | 5,781 |
| | HI 90 | 0,422 | 0,078 | 6,253 |
| | MECVI | 0,364 | 0,078 | 6,014 |
| RMSEA | RMSEA | 0,107 | - | 0,407 |
| | LO 90 | 0,098 | - | 0,399 |
| | HI 90 | 0,117 | - | 0,415 |
| | PCLOSE | 0 | - | 0 |
| HOELTER | Al 0.05 | 121 | - | 9 |
| | Al 0.01 | 141 | - | 10 |

# Capítulo 7

# Discusión

*Nunca discutas con un superior. Corres el riesgo de tener razón.*

Marco Aurelio

# 7.-Discusión

Discusión general

El test de depresión PHQ-9 representa el concepto de depresión recogido en el DSM-IV (Bell, 1994) y al ser un extracto o una forma de aplicar los criterios establecidos de depresión al mundo real, en una franja temporal de dos semanas representa adecuadamente el concepto de trastorno depresivo mayor. No obstante, en el DSM se considera diagnóstico el diferencial con otras patologías, lo cual no viene contemplado en el test. Tampoco se consideran otras formas de depresión, ya que las preguntas, al hacer referencia al Trastorno Depresivo Mayor, concreta temporalmente a dos semanas como espacio de tiempo en el que deben trascurrir los eventos sobre los que se pregunta, ignorando así aquellos que se encuentren fiera de ese espacio de tiempo. Entonces no quiere decir que las personas detectadas con depresión sean aquellas que aparecen en los porcentajes como positivas (habiendo superado su respectivo punto de corte en función de su grupo); sino que de estas habría que descartar aquellas cuyos síntomas se explican mejor por otras causas. Además habría que añadir aquellas personas cuyos síntomas no son suficientemente evidentes, pero sí son continuados en el tiempo, como el caso de la distimia; o aquellas personas cuyo diagnóstico no es el de Trastorno Depresivo Mayor, pero sí otras formas de trastorno depresivo. Recordemos que en DSM-V (Asociación Americana de Psiquiatría, 2014). Por último siempre cabe resaltar que depresión y salud mental no son sinónimos, sino que la depresión forma parte de un extenso bagaje de patologías psicológicas y que indicar ausencia de depresión no implica indicar un adecuado estado de salud mental.

Se encontraron similitudes con otras validaciones de este test, conformándose la idea de que el trastorno depresivo mayor queda bien recogido con el test. No obstante, la cuestión del sexo queda en entredicho, aunque se comentará a continuación.

La validación del instrumento ha sido posible gracias a las técnicas estadísticas, siempre comprendiendo su utilidad, ya que el análisis factorial introdujo la dimensionalidad, induciendo la existencia o no de otros constructos paralelos. Gracias a la curva ROC se pudo conocer puntos de corte óptimos tanto para población general como en función de la variable sexo. Aunque se complicó la interpretabilidad, el poder disponer de puntos de corte adaptados por sexo implica una mayor explicación de las circunstancias de los individuos. Al ser su aplicación en evaluación de poblaciones, implica un salto cualitativo en poder explicar de materia notable las características de los sujetos.

Queda demostrada la capacidad diagnóstica de este test (Levis et al., 2019), tanto en población general como en función del sexo. Así el Trastorno Depresivo Mayor puede medirse adecuadamente con este test en diferentes culturas (Galenkamp et al., 2017; Ong et al., 2022; Zhou et al., 2020), en diferentes contextos como el educativo (Keum et al., 2018), en urgencias (Nallusamy et al., 2016), en contextos de emergencia (Ren et al., 2020) y en diferentes grupos etarios (Ong et al., 2022).

### Sexo

En la muestra el porcentaje de mujeres participantes es superior que el de varones, lo cual es esperable de acuerdo con otras investigaciones (GBD 2019 Mental Disorders Collaborators, 2022). Una de las posibles explicaciones de por qué se ha detectado más psicopatología en mujeres que en hombres es que la misma muestra se autoclasifique. Esto quiere decir que el hecho de contestar un cuestionario en línea como en el que se utilizó para esta investigación haga de llamamiento a las personas que deseen contestar, ya sea porque su forma de conocerse es a través de la exploración, que la propia exploración sea terapéutica, que deseen expresar su malestar u otros. Si bien el tipo de muestreo y de participación posee equivalencias de tasas de respuesta con otros estudios (Steel et al., 2014), siempre ha de

considerarse el tipo de muestreo como una limitación explicativa. No obstante, en investigaciones cuyas exploraciones son presenciales y, por alguna razón la tasa de voluntariedad en los varones haya aumentado (como puede serlo los beneficios de colaborar en investigación) la tasa de prevalencia de depresión entre hombres y mujeres se mantiene (Parmar et al., 2016). Por esta razón la consideración de autoclasificación toma menos peso como posible razón a tanto encontrar más participantes mujeres como para encontrar más prevalencia del factor entre este sexo.

Para conocer la influencia del sexo en la aplicación del test, o dicho de otra forma, conocer si al aplicar el mismo test en la población en función del sexo, implicaría diferencias notables se utiliza la invarianza multigrupo. Esta técnica ha sido utilizada en este mismo test con similares resultados. Implicando que efectivamente el test puede aplicarse en ambos grupos poblacionales (Galenkamp et al., 2017; González-Blanch et al., 2018). El uso de esta técnica quedaría justificada por la diferente vivencia de la depresión en función del sexo (Martin et al., 2013).

### Factores

En este estudio se observó un solo factor, congruente con otras investigaciones (Merz et al., 2011) y se prefiere la estructura unifactorial por ser esta más parsimoniosa (González-Blanch et al., 2018), en cambio, en otros estudios diferencian 2 factores: somático y cognitivo-afectivo (Chilcot et al., 2013). La estructura factorial, así como la invarianza multigrupo son áreas que se estudian al comprobar las propiedades psicométricas de los tes. En concreto el test PHQ-9 ha sido estudiada su estructura en diversos grupos poblacionales como estudiantes estadounidenses (Keum et al., 2018), en centros de atención primaria de Chile (Saldivia et al., 2019) o España (Muñoz-Navarro et al., 2017). Además, que gracias a conocer la estructura del test pueden hacerse comparaciones entre países para conocer si el concepto de depresión

para las diferentes muestras implica otras dimensiones a considerar y, por lo tanto, tratar a la depresión no como un concepto unitario, sino como un constructo de varios componentes o multicomponente, tal y como discuten Zhou y compañía (Zhou et al., 2020).

Al haber obtenido un resultado unidimensional se descartan otras opciones de análisis como las mejoras a las reducciones de dimensiones (González-García et al., 2023) o en formas de optimización en selección de factores (de Pierrefeu et al., 2018).

Versiones abreviadas del PHQ-9

Como ventajas de utilizar el PHQ-2 se destaca que se pueden realizar diagnósticos rápidos y sin perder efectividad, dado que la especificidad y sensibilidad se mantienen similares a los valores que arroja su versión más extensa de PHQ-9. Estos hallazgos son congruentes con otras investigaciones que comparan el PHQ-2 con el PHQ-9 (Richardson et al., 2010).

El PHQ-4 es usado comúnmente por su fácil aplicación. Con solo 4 ítems ya puede conocerse de forma esquemática la situación de salud mental en cuanto a depresión y ansiedad puede tener una persona. A nivel poblacional es una herramienta muy positiva, la cual, a niveles de conclusiones no va más allá de las que pudieran ofrecer las herramientas PHQ-2 y GAD-2 por separado. Todas las investigaciones encontradas (Christodoulaki et al., 2022; Levis et al., 2020; Plummer et al., 2016) indican adecuadas propiedades para identificar la ansiedad y la depresión. Estos cuestionarios son adecuados y válidos tanto en precisión diagnóstica de cada una de las patologías como en sensibilidad y especificidad por separado.

La discusión de PHQ-8 se hará en el apartado de suicidio.

Suicidio

Cualquier puntuación mayor de 0 en el ítem 9 se considera susceptible de iniciar el protocolo de prevención de conducta suicida (España & Fernández, 2010). Un 27,88% de la población presentó pensamientos suicidas en algún grado. El mayor grado de ideación suicida fue seleccionado por el 6,28 de la población, la cual tiene pensamientos suicidas cada día o casi cada día. Estos porcentajes son elevados en comparación con otras investigaciones tanto de población general (Hawton, 2009; Shenassa et al., 2004; Zhang & Jia, 2010), en función de las edades (Borowsky et al., 1999; Brown et al., 1991; Szanto et al., 2001; Yuryev et al., 2010) o en presencia de enfermedades específicas (Eng et al., 2019; J. Fawcett et al., 1990; Hawton et al., 2005; Inskip et al., 1998); pero más común en situaciones de crisis (Chan et al., 2006; Dreyer et al., 2010; Funder, 2014; Sant et al., 2007).

La estimación de ideación suicida se sitúa en un 12.9% (CI95%: 10.3% -15.5%) (Karimi et al., 2021). Esto quiere decir que se ha detectado 2,16 veces más de ideación suicida que en la población general, considerando una muestra de comparación de 61.180 personas en situación de desastre a través de 33 estudios.

Existen tipos de suicidio que responden a términos impulsivos, ante los cuales no concominan con depresión(Conner & Duberstein, 2004; Cooper et al., 2005; Marty et al., 2010; Viglione et al., 2014). Aunque una de las formas de expresión de la depresión es el suicidio, la inclusión del ítem de suicidio en la escala PHQ-9 tiene razón de ser más de manera exploratoria que definitoria de depresión. De todas formas el relato del paciente como indicador de suicidio no debería ser una manera adecuada de exploración de los intentos autolíticos. No obstante, al valorar la depresión y siendo el suicidio una expresión de esta, puede reflejarse en la consistencia interna, así como correlación inter-ítems; lo cual justifica la permanencia del ítem en la escala al tratarse de un constructo unidimensional (Razykov et al., 2012).

El ítem 9 es un candidato a ser eliminado del cuestionario ya que cuando este es eliminado del mismo, su valor Alpha de Cronbach aumenta. Pese a que es un aumento ligero el hecho de que aumente en cualquier valor es un indicador sobre la susceptibilidad de su eliminación. Este hecho sustentaría la existencia del test PHQ-8. Además, el ítem 9 se refiere a los pensamientos de suicidio. En determinados casos puede entenderse como un nivel más elevado de depresión. El PHQ-9 está basado en el DSM-IV, el cual contempla dentro del trastorno depresivo mayor: los pensamientos recurrentes de muerte (y no sólo temor a la muerte), la ideación suicida recurrente sin un plan específico o una tentativa de suicidio o un plan específico para suicidarse; como un indicador de depresión. Por lo tanto se considera que el test posee validez de constructo al incluirse este ítem.

Las diferencias entre el PHQ-8 y 9 vienen explicadas por la consideración del último ítem: el de depresión. La equivalencia en la precisión del diagnóstico entre el PHQ-8 y 9 en este estudio ha resultado muy similar. Lo mismo descubrieron otros investigadores en una revisión sistemática y metaanálisis (Borghero et al., 2018; Calderón et al., 2012; Wu et al., 2020).

Aunque el suicidio pueda incrementarse en situaciones de crisis (Aknin et al., 2022; Yan et al., 2023), este hecho no parece importante en la capacidad diagnóstica del test en cuanto al suicidio, ya que los pensamientos autolíticos, pese a aumentar en número, lo hacen manteniendo el síntoma como tal con independencia del contexto (Joiner, 2011; Klomek, 2020; Maris et al., 1992).

Puntos de corte

Después de determinar los puntos de corte óptimos tanto para población general como para la población en función del sexo, tal y como se han planteado previamente (Snijkers et al., 2021) indica unos índices de sensibilidad y especificidad aceptables, aunque no óptimos. Como forma ideal sería poder establecer estos nuevos puntos de corte para diferentes culturas, aplicaciones o circunstancias. Pero ¿realmente implica una diferencia importante en

los test de cribado? Tal y como debaten Yuying Luo y Laurie Keefe (Luo & Keefer, 2021) esta diferencia afecta a la prevalencia y las diferencias epidemiológicas pueden explicarse por las diferencias en la sensibilidad y especificidad, sobre todo en comparaciones de validación de instrumentos entre culturas.

# CUARTA PARTE: CONCLUSIONES Y PERSPECTIVAS FUTURAS DE INVESTIGACIÓN

# Capítulo 8

# Conclusiones

*Realmente no se tiene una gran cantidad de datos sobre el tema, y sin datos, ¿cómo podemos llegar a conclusiones definitivas?*

Thomas Alva Edison

# 8.-Conclusiones

Según los resultados encontrados se concluye que:

1.- Las adaptaciones realizadas a los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 son suficientes para que los test sean adecuados cultural y lingüísticamente para poder ser aplicados a la población costarricense.

2.- A través del estudio de la invarianza de medida multigrupo de los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 se han descubierto diferencias en la vivencia de la depresión en función del sexo.

3.- Los test de depresión PHQ-9, PHQ-8, PHQ-4 y PHQ-2 poseen una correlación significativa y directa con la Escala de Miedo a la COVID-19 (FCV-19S), Escala de Miedo Social a la COVID-19 (SFCV-19S) test de Estrés (HSAY), efectos emocionales del trabajo (PROQOL) e inversa con Resiliencia (RS14), demostrando así validez externa.

4.- Los test PHQ-9, PHQ-8, PHQ-4 y PHQ-2 presentan adecuadas propiedades psicométricas para poder evaluar depresión en población costarricense, tanto para población general como para población diferenciada por sexo. Los test PHQ-9, PHQ-8 y PHQ-2 presentan estructura unifactorial, mientras que el PHQ-4 presenta estructura bifactorial.

5.- Los puntos de corte propuestos implican diferencias estadísticamente significativas con respecto a los resultados obtenidos con los puntos de corte originales, obteniendo mejoras en la sensibilidad y especificidad.

6.- Los modelos de ecuaciones estructurales permiten conocer la estructura interna del test pudiendo diferenciarse los rasgos y los ítems que los componen.

7.- Los métodos de la curva Característica Operativa del Receptor son adecuados para esclarecer la sensibilidad y especificidad de los test psicométricos.

# Capítulo 9

# Perspectivas futuras de investigación

*El investigador sufre las decepciones, los largos meses pasados en una dirección equivocada, los fracasos. Pero los fracasos son también útiles, porque, bien analizados, pueden conducir al éxito. Y para el investigador no existe alegría comparable a la de un descubrimiento, por pequeño que sea...*

Sir Alexander Fleming

# 9.-Líneas futuras de investigación

Las diferencias sexuales son notables en cuanto a la detección de la depresión. Si bien cabe resaltar que la depresión y la salud mental no son equivalentes, uno implica al otro y viceversa. De esta forma no podemos afirmar que una persona está sana cuando no se le ha detectado depresión. No quiere decir que no exista, sino que no se le ha detectado. Las propuestas para el futuro van en la línea de la detección de la depresión inicialmente no evaluada. Existen cuestionarios que pretenden eliminar la brecha sexual en cuanto a la detección de la depresión (Martin et al., 2013). Estos cuestionarios recogen la depresión de forma diferente a lo contemplado en el DSM-IV, haciendo hincapié en las vivencias de la depresión por sexo, habiendo obtenido esta información mediante métodos adecuados como revisión bibliográfica. Actualmente (Martin et al., 2013) generaron 2 test de equivalencia en depresión con independencia del sexo: el Gender Inclusive Depression Scale y el Masculine Depression Scale. Al ser poco conocidos aún no han sido validados en otros países, por lo que es una adecuada línea de investigación.

# QUINTA PARTE: REFERENCIAS Y ANEXOS

*Aprended a hacer trabajo de peón en la ciencia. Estudiad, confrontad, acumulad hechos. Por muy perfectas que hubiesen sido las alas del ave, jamás le habrían podido permitir elevarse si no se apoyasen en el aire. Los hechos son el aire del hombre de ciencia. Sin ellos, jamás podréis levantar el vuelo. Sin ellos vuestras teorías serán esfuerzos vanos.*

Iván Petróvich Pávlov

# Capítulo 10

# Referencias

## 10.- Referencias


Addis, M. E., & Hoffman, E. (2017). Men's depression and help-seeking through the lenses of gender. In *The psychology of men and masculinities.* (pp. 171–196). American Psychological Association. https://doi.org/10.1037/0000023-007

Ahorsu, D. K., Lin, C. Y., Imani, V., Saffari, M., Griffiths, M. D., & Pakpour, A. H. (2020). The Fear of COVID-19 Scale: Development and Initial Validation. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00270-8

Ahorsu, D. K., Lin, C.-Y., Imani, V., Saffari, M., Griffiths, M. D., & Pakpour, A. H. (2022). The Fear of COVID-19 Scale: Development and Initial Validation. *International Journal of Mental Health and Addiction*, *20*(3), 1537–1545. https://doi.org/10.1007/s11469-020-00270-8

Aknin, L. B., De Neve, J.-E., Dunn, E. W., Fancourt, D. E., Goldberg, E., Helliwell, J. F., Jones, S. P., Karam, E., Layard, R., Lyubomirsky, S., Rzepa, A., Saxena, S., Thornton, E. M., VanderWeele, T. J., Whillans, A. V., Zaki, J., Karadag, O., & Ben Amor, Y. (2022). Mental Health During the First Year of the COVID-19 Pandemic: A Review and Recommendations for Moving Forward. *Perspectives on Psychological Science*, *17*(4), 915–936. https://doi.org/10.1177/17456916211029964

American Psychological Association. (1999). *Standards for psychological and educational testing*.

Amirkhan, J. H. (2012). Stress Overload: A New Approach to the Assessment of Stress. *American Journal of Community Psychology*, *49*(1–2), 55–71. https://doi.org/10.1007/s10464-011-9438-x

Andrés-Olivera, P., García-Aparicio, J., Lozano López, M. T., Benito Sánchez, J. A., Martín, C., Maciá-Casas, A., González-Sánchez, A., Marcos, M., & Roncero, C. (2022). Impact on Sleep Quality, Mood, Anxiety, and Personal Satisfaction of Doctors Assigned to COVID-19 Units. *International Journal of Environmental Research and Public Health*, *19*(2712), 1–15. https://doi.org/10.3390/ijerph19052712

Angell, R. (2019). *Learn to Test for Metric Invariance Using Multi-Group Confirmatory Factor Analysis (MGCFA) in SPSS AMOS With Data From the International Sponsorship Study (2016)*. SAGE Publications, Ltd. https://doi.org/10.4135/9781526469625

Annunziata, M. A., Muzzatti, B., Bidoli, E., Flaiban, C., Bomben, F., Piccinin, M., Gipponi, K. M., Mariutti, G., Busato, S., & Mella, S. (2020). Hospital Anxiety and Depression Scale (HADS)

accuracy in cancer patients. *Supportive Care in Cancer*, *28*(8), 3921–3926. https://doi.org/10.1007/s00520-019-05244-8

Arbuckle, J. L. (2016). *Amos (Version 4.0)* (24.0). IBM SPSS. http://amosdevelopment.com/index.html

Armesto, D. (2011). Pruebas diagnósticas: curvas ROC. *Rev Electron Biomed / Electron J Biomed*, *1*, 77–82.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Asociación Americana de Psiquiatría. (2014). *Manual Diagnóstico y Estadístico de los Trastornos Mentales (DSM-5®)* (Arlington, Ed.; 5ª).

Barón López, F. J. (2022). *Curvas ROC: Elección de puntos de corte y área bajo la curva (AUC)*. https://www.bioestadistica.uma.es/app/roc1/

Bell, C. C. (1994). DSM-IV: Diagnostic and Statistical Manual of Mental Disorders. *JAMA*, *272*(10), 828–829. https://doi.org/10.1001/jama.1994.03520100096046

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*. https://doi.org/10.1177/2059799116672875

Bjelland, I., Dahl, A. A., Haug, T. T., & Neckelmann, D. (2002). The validity of the Hospital Anxiety and Depression Scale. *Journal of Psychosomatic Research*, *52*(2), 69–77. https://doi.org/10.1016/S0022-3999(01)00296-3

Bogduk, N. (2013). *Diagnostic and Statistical Manual of Mental Disorders BT - Encyclopedia of Pain* (G. F. Gebhart & R. F. Schmidt, Eds.; pp. 979–982). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28753-4_1094

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

Borghero, F., Martínez, V., Zitko, P., Vöhringer, P. A., Cavada, G., & Rojas, G. (2018). Screening for depressive episode in adolescents. Validation of the PHQ-9 instrument. [Tamizaje de episodio depresivo en adolescentes. Validación del instrumento PHQ-9]. *Revista Médica de Chile*, *146*(4), 479–486. https://doi.org/10.4067/s0034-98872018000400479

Borowsky, I. W., Resnick, M. D., Ireland, M., & Blum, R. W. (1999). *Suicide Attempts Among American Indian and Alaska Native Youth*. *153*(June), 573–580.

Brantley, P. J., Cocke, T. B., Jones, G. N., & Goreczny, A. J. (1988). The Daily Stress Inventory: Validity and effect of repeated administration. *Journal of Psychopathology and Behavioral Assessment*, *10*(1), 75–81. https://doi.org/10.1007/BF00962987

Brown, L. K., Overholser, J., Spirito, A., & Fritz, G. K. (1991). The correlates of planning in adolescent suicide attempts. *Journal of the American Academy of Child & Adolescent Psychiatry*, *30*(1), 95–99.

Browne, M. W. (1972). OBLIQUE ROTATION TO A PARTIALLY SPECIFIED TARGET. *British Journal of Mathematical and Statistical Psychology*, *25*(2), 207–212. https://doi.org/10.1111/j.2044-8317.1972.tb00492.x

Buttò, S., Suligoi, B., Fanales-Belasio, E., & Raimondo, M. (2010). Laboratory diagnostics for HIV infection. *Annali Dell'Istituto Superiore Di Sanita*, *46*(1), 24–33. https://doi.org/https://doi.org/10.4415/ANN_10_01_04

Cabrera, V., Martín-Aragón, M., Terol, M. del C., Núñez, R., & Pastor, M. de los Á. (2015). La Escala de Ansiedad y Depresión Hospitalaria (HAD) en fibromialgia: Análisis de sensibilidad y especificidad. *Terapia Psicológica*, *33*(3), 181–193. https://doi.org/10.4067/S0718-48082015000300003

Calderón, M., Antonio Gálvez-Buccollini, J., Cueva, G., Ordoñez, C., Bromley, C., & Fiestas, F. (2012). Validation of the peruvian version of the PHQ-9 for diagnosing depression [Validación de la versión peruana del PHQ-9 para el diagnóstico de depresión]. *Rev Peru Med Exp Salud Publica*, *29*(4), 578–585.

Call, J. B., & Shafer, K. (2018). Gendered Manifestations of Depression and Help Seeking Among Men. *American Journal of Men's Health*, *12*(1), 41–51. https://doi.org/10.1177/1557988315623993

Camargo, L., Soto-Añari, M., Caldichoury-Obando, N., Ramos-Henderson, M., Porto, M. F., Salomón, S., Saldías-Solis, C., Gargiulo, P., & López, N. (2022). Validity of the PHQ-2 questionnaire for the detection of depression in Colombian health personnel during Covid-19. [Validez del cuestionario PHQ-2 para la detección de depresión en personal sanitario colombiano durante Covid-19]. *Revista Chilena de Neuro-Psiquiatría*, *60*(3), 281–288. https://doi.org/10.4067/s0717-92272022000300281

Carazo-Vargas, E., Ortega-Moreno, R., Arias, H., González-García, N., González-Sánchez, A., & Villegas, G. (2021). *Mental health and relationships with the environment in times of COVID-19*. https://investiga.uned.ac.cr/wp-content/uploads/2021/01/INFORME-Salud-mental-en-tiempos-de-COVID-19.pdf

Carazo-Vargas, E., Ortega-Moreno, R., Villegas Barahona, G., Arias-LeClaire, H., González-García, N., & González-Sánchez, A. (2021, November 10). Impacto de las situaciones de emergencia en el país. (Secuelas emocionales). *XXXII Jornadas de Psicología y Salud. Intervención Psicológica En Situaciones de Emergencia*.

Carstairs, S. D. (2013). Diagnosis of testicular cancer with a urine pregnancy test in an austere military medical environment. *The American Journal of Emergency Medicine*, *31*(11), 1615. https://doi.org/10.1016/j.ajem.2013.08.010

Cattell, R. B. (1988). The Meaning and Strategic Use of Factor Analysis. In *Handbook of Multivariate Experimental Psychology* (pp. 131–203). Springer US. https://doi.org/10.1007/978-1-4613-0893-5_4

Cénat, J. M., Hébert, M., Karray, A., & Derivois, D. (2018). Psychometric properties of the Resilience Scale – 14 in a sample of college students from France. *L'Encéphale*, *44*(6), 517–522. https://doi.org/10.1016/j.encep.2018.04.002

Chan, S. M. S., Chiu, F. K. H., Lam, C. W. L., Leung, P. Y. V., & Conwell, Y. (2006). Elderly suicide and the 2003 SARS epidemic in Hong Kong. *International Journal of Geriatric Psychiatry*, *21*(2), 113–118. https://doi.org/10.1002/gps.1432

Chen, I. H., Chen, C. Y., Zhao, K. Y., Gamble, J. H., Lin, C. Y., Griffiths, M. D., & Pakpour, A. H. (2022). Psychometric evaluation of fear of COVID-19 Scale (FCV-19S) among Chinese

primary and middle schoolteachers, and their students. *Current Psychology*, *0123456789*. https://doi.org/10.1007/s12144-021-02471-3

Chen, W., Xie, E., Tian, X., & Zhang, G. (2020). Psychometric properties of the Chinese version of the Resilience Scale (RS-14): Preliminary results. *PLOS ONE*, *15*(10), e0241606. https://doi.org/10.1371/journal.pone.0241606

Chilcot, J., Rayner, L., Lee, W., Price, A., Goodwin, L., Monroe, B., Sykes, N., Hansford, P., & Hotopf, M. (2013). The factor structure of the PHQ-9 in palliative care. *Journal of Psychosomatic Research*, *75*(1), 60–64. https://doi.org/10.1016/j.jpsychores.2012.12.012

Christodoulaki, A., Baralou, V., Konstantakopoulos, G., & 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

Cohen, S., Kamarck, T., & Mermelstein, R. (1994). Perceived stress scale. *Measuring Stress: A Guide for Health and Social Scientists*, *10*, 1–2.

Conner, K. R., & Duberstein, P. R. (2004). Predisposing and Precipitating Factors for Suicide Among Alcoholics: Empirical Review and Conceptual Integration. *Alcoholism: Clinical and Experimental Research*, *28*(5), 6S-17S. https://doi.org/10.1097/01.alc.0000127410.84505.2a

Cooper, J., Kapur, N., Webb, R., Lawlor, M., Guthrie, E., Mackway-Jones, K., & Appleby, L. (2005). Suicide after deliberate self-harm: A 4-year cohort study. *American Journal of Psychiatry*, *162*(2), 297–303. https://doi.org/10.1176/appi.ajp.162.2.297

Crespi, C. M., Wong, W. K., & Wu, S. (2011). A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. *Clinical Trials*, *8*(6), 687–698. https://doi.org/10.1177/1740774511423851

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

de Pierrefeu, A., Lofstedt, T., Hadj-Selem, F., Dubois, M., Jardri, R., Fovet, T., Ciuciu, P., Frouin, V., & Duchesnay, E. (2018). Structured Sparse Principal Components Analysis With the TV-Elastic Net Penalty. *IEEE Transactions on Medical Imaging*, *37*(2), 396–407. https://doi.org/10.1109/TMI.2017.2749140

del Valle Benavides, A. R. (2017). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones*. Universidad de Sevilla.

Denby, L., & Mallows, C. (2009). Variations on the histogram. *Journal of Computational and Graphical Statistics*, *18*(1), 21–31. https://doi.org/10.1198/jcgs.2009.0002

Derogatis, L. R. (1987). The Derogatis stress profile (DSP): Quantification of psychological stress. In *Research paradigms in psychosomatic medicine* (Vol. 17, pp. 30–54). Karger Publishers.

Diez-Quevedo, C., Rangil, T., Sánchez-Planell, L., Kroenke, K., & Spitzer, R. L. (2001). Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. *Psychosomatic Medicine*, *63(4)*, 679–686.

Dreyer, L., Kendall, S., Danneskiold-Samsøe, B., Bartels, E. M., & Bliddal, H. (2010). Mortality in a cohort of Danish patients with fibromyalgia: Increased frequency of suicide. *Arthritis and Rheumatism*, *62*(10), 3101–3108. https://doi.org/10.1002/art.27623

Dube, P., Kroenke, K., Bair, M. J., Theobald, D., & Williams, L. S. (2010). The P4 Screener. *The Primary Care Companion to The Journal of Clinical Psychiatry*. https://doi.org/10.4088/PCC.10m00978blu

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. (1951). Intelligence and cultural differences; a study of cultural learning and problem-solving. In *Intelligence and cultural differences; a study of cultural learning and problem-solving.* University of Chicago Press.

Eng, J. (2014, March 19). *ROC analysis: web-based calculator for ROC curves*. http://www.jrocfit.org

Eng, J., Drabwell, L., Stevenson, F., King, M., Osborn, D., & Pitman, A. (2019). Use of Alcohol and Unprescribed Drugs after Suicide Bereavement: Qualitative Study. *International Journal of Environmental Research and Public Health*, *16*(21), 4093. https://doi.org/10.3390/ijerph16214093

España, A., & Fernández, C. (2010). Protocolo de Urgencias Hospitalarias ante conductas suicidas. *Rev Méd de Ja*, *1*(1), 29–32.

Espejo, B., & Checa, I. (2021). The fear of Covid-19 scale (FCV-19s) in Spain: Adaptation and confirmatory evidence of construct and concurrent validity. *Mathematics*, *9*(19). https://doi.org/10.3390/math9192512

Eysenck, H. J. (1952). *The scientific study of personality*. Macmillan.

Faller, K. (2014). Forty Years of Forensic Interviewing of Children Suspected of Sexual Abuse, 1974–2014: Historical Benchmarks. *Social Sciences*, *4*(1), 34–65. https://doi.org/10.3390/socsci4010034

Fawcett, J., Scheftner, W. A., Fogg, L., Clark, D. C., Young, M. A., Hedeker, D., & Gibbons, R. (1990). Time-related predictors of suicide in major affective disorder. *The American Journal of Psychiatry*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decalogue for the Factor Analysis of Test Items. In *Psicothema* (Vol. 34, Issue 1, pp. 7–17). NLM (Medline). https://doi.org/10.7334/psicothema2021.456

Fisher, R. P., & Geiselman, R. E. (2010). The Cognitive Interview method of conducting police interviews: Eliciting extensive information and promoting Therapeutic Jurisprudence. *International Journal of Law and Psychiatry*, *33*(5–6), 321–328. https://doi.org/10.1016/j.ijlp.2010.09.004

Fleiss, J. L. (1999). *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118032923

Flora, D. B., & Curran, P. J. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Freedman, D., & Diaconis, P. (1981). On the Histogram as a Density Estimator: L 2 Theory. In *Z. Wahrscheinlichkeitstheorie verw. Gebiete* (Vol. 57).

Funder, D. C. D. C. (2014). Weighing dispositional and situational factors in accounting for suicide terrorism. *Behavioral and Brain Sciences*, *37*, 367–368. https://doi.org/10.1017/S0140525X13003397

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: the HELIUS study. *BMC Psychiatry*, *17*(1), 349. https://doi.org/10.1186/s12888-017-1506-9

Galindo, P. (1986). An alternative for simultaneous representation: HJ-Biplot. Questiió: Quaderns d'Estadística, Sistemes, Informatica I Investigació Operativa. *Questiio*, *10*, 13–23.

Gamonal-Limcaoco, S., Montero-Mateos, E., Lozano-López, M. T., Maciá-Casas, A., Matías-Fernández, J., & Roncero, C. (2021). Perceived stress in different countries at the beginning of the coronavirus pandemic. *The International Journal of Psychiatry in Medicine*, 00912174211033710. https://doi.org/10.1177/00912174211033710

Garaiman, A., Mihai, C., Dobrota, R., Jordan, S., Maurer, B., Flemming, J., Distler, O., & Becker, M. O. (2021). The Hospital Anxiety and Depression Scale in patients with systemic sclerosis: a psychometric and factor analysis in a monocentric cohort. *Clinical and Experimental Rheumatology*, *39*(4), 34–42. https://doi.org/10.55563/clinexprheumatol/qo1ehz

García, F., Álvarez, M., Bernal, C., Chueca, N., & Guillot, V. (2011). Diagnóstico de laboratorio de la infección por el VIH, del tropismo viral y de las resistencias a los antirretrovirales. *Enfermedades Infecciosas y Microbiologia Clinica*, *29*(4), 297–307. https://doi.org/10.1016/j.eimc.2010.12.006

García-Campayo, J., Zamorano, E., Ruiz, M. A., Pérez-Páramo, M., López-Gómez, V., & Rejas, J. (2012). The assessment of generalized anxiety disorder: psychometric validation of the Spanish version of the self-administered GAD-2 scale in daily medical practice. *Health and Quality of Life Outcomes*, *10*, 1–10. https://doi.org/10.1186/1477-7525-10-114

Garven, S., Wood, J. M., Malpass, R. S., & Shaw, J. S. (1998). More than suggestion: The effect of interviewing techniques from the McMartin Preschool case. *Journal of Applied Psychology*, *83*(3), 347–359. https://doi.org/10.1037/0021-9010.83.3.347

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

González, A., Villegas, G., Carazo, E., Ortega, R., & Arias, H. (2021). Propuesta de un modelo de salud mental considerando dimensiones psicológicas en tiempos de pandemia. *IV Jornadas de Estadística Como Herramienta Científica*.

Gonzalez, V. M. (2008). Recognition of Mental Illness and Suicidality Among Individuals With Serious Mental Illness. *Journal of Nervous & Mental Disease*, *196*(10), 727–734. https://doi.org/10.1097/NMD.0b013e3181879deb

González-Blanch, C., Medrano, L. A., Muñoz-Navarro, R., Ruíz-Rodríguez, P., Moriana, J. A., Limonero, J. T., Schmitz, F., & Cano-Vindel, A. (2018). Factor structure and measurement invariance across various demographic groups and over time for the PHQ-9 in primary care patients in Spain. *PLoS ONE*, *13*(2), 1–16. https://doi.org/10.1371/journal.pone.0193356

González-García, N., Nieto-Librero, A. B., & Galindo-Villardón, P. (2023). CenetBiplot: a new proposal of sparse and orthogonal biplots methods by means of elastic net CSVD. *Advances in Data Analysis and Classification*, *17*(1), 5–19. https://doi.org/10.1007/s11634-021-00468-1

González-Sánchez, A., Ortega-Moreno, R., Villegas-Barahona, G., Carazo-Vargas, E., Arias-LeClaire, H., & Vicente-Galindo, P. (2023). New cut-off points of PHQ-9 and its variants, in Costa Rica: a nationwide observational study. *Scientific Reports*, *13*(1), 14295. https://doi.org/10.1038/s41598-023-41560-0

Good, G. E., & Wood, P. K. (1995). Male Gender Role Conflict, Depression, and Help Seeking: Do College Men Face Double Jeopardy? *Journal of Counseling & Development*, *74*(1), 70–75. https://doi.org/10.1002/j.1556-6676.1995.tb01825.x

Griffiths, M. D. (2001). *Intolerance of Uncertainty and Mental Wellbeing : Serial Mediation by Rumination and Fear of COVID-19*.

Grupo de trabajo e investigación de la sección de psicología jurídica y forense del COPC. (2014). *Guía de buenas prácticas para la evaluación psicológica forense y la práctica pericial*. Col·legi Oficial de Psicologia de Catalunya.

Hamilton, M. (1959). The assessment of Anxiety States by Rating. *British Journal of Medical Psychology*, *32*(1), 50–55. https://doi.org/10.1111/j.2044-8341.1959.tb00467.x

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Harman, H. H. (1976). *Modern Factor Analysis* (3rd ed.). The University of Chicago.

Hawton, K. (2009). van HK. Suicide. *Lancet*, *373*(9672), 1372–1381.

Hawton, K., Sutton, L., Haw, C., Sinclair, J., & Harriss, L. (2005). Suicide and attempted suicide in bipolar disorder: a systematic review of risk factors. *The Journal of Clinical Psychiatry*.

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*, *00*(00), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hemsworth, D., Baregheh, A., Aoun, S., & Kazanjian, A. (2018). A critical enquiry into the psychometric properties of the professional quality of life scale (ProQol-5) instrument. *Applied Nursing Research*, *39*, 81–88. https://doi.org/10.1016/j.apnr.2017.09.006

Hernández, A., Elosua, P., Abad, F. J., Antón, M., Martínez, A., Vallar, F., & Galve, J. L. (2015). *Informe sobre el uso de los test psicométricos en los procesos de selección de personal de las administraciones públicas*.

Herrero, M. J., Blanch, J., Peri, J. M., de Pablo, J., Pintor, L., & Bulbena, A. (2003). A validation study of the hospital anxiety and depression scale (HADS) in a Spanish population. *General Hospital Psychiatry*, *25*(4), 277–283. https://doi.org/10.1016/S0163-8343(03)00043-4

Herrmann, C. (1997). International experiences with the Hospital Anxiety and Depression Scale-A review of validation data and clinical results. *Journal of Psychosomatic Research*, *42*(1), 17–41. https://doi.org/10.1016/S0022-3999(96)00216-4

Hoffmann, A. F., Stover, J. B., & Liporace, M. F. (2013). Polychoric and tetrachoric correlations in exploratory and confirmatory factorial studies. [Correlaciones policóricas y tetracóricas en estudios factoriales exploratorios y confirmatorios]. *Ciencias Psicológicas*, *7*(2).

Hoy, S. (2012). Beyond Men Behaving Badly: A Meta-Ethnography of Men's Perspectives on Psychological Distress and Help Seeking. *International Journal of Men's Health*, *11*(3), 202–226. https://doi.org/10.3149/jmh.1103.202

Hsieh, F., & Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver Operating Characteristic Curve. In *The Annals of Statistics* (Vol. 24, Issue 1).

Huang, Y., & Zhao, N. (2020). Generalized anxiety disorder, depressive symptoms and sleep quality during COVID-19 outbreak in China: a web-based cross-sectional survey. *Psychiatry Research*, *288*(April), 112954. https://doi.org/10.1016/j.psychres.2020.112954

Huarcaya-Victoria, J., Villarreal-Zegarra, D., Podestà, A., & Luna-Cuadros, M. A. (2020). Psychometric Properties of a Spanish Version of the Fear of COVID-19 Scale in General Population of Lima, Peru. *International Journal of Mental Health and Addiction*, *19*. https://doi.org/10.1007/s11469-020-00354-5

IASC Inter-Agency Standing Committee. (2006). *IASC guidelines on mental health and psychosocial support in emergency settings*.

Inagaki, M., Ohtsuki, T., Yonemoto, N., Kawashima, Y., Saitoh, A., Oikawa, Y., Kurosawa, M., Muramatsu, K., Furukawa, T. A., & Yamada, M. (2013). Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: A cross-sectional study. *General Hospital Psychiatry*, *35*(6), 592–597. https://doi.org/10.1016/j.genhosppsych.2013.08.001

Inskip, H., Harris, C., & Barraclough, B. (1998). Lifetime risk of suicide for affective disorder, alcoholism and schizophrenia. *The British Journal of Psychiatry*, *172*(1), 35–37.

Iversen, M. M., Norekvål, T. M., Oterhals, K., Fadnes, L. T., Mæland, S., Pakpour, A. H., & Breivik, K. (2021). Psychometric Properties of the Norwegian Version of the Fear of COVID-19 Scale. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00454-2

Jensen, A. (1969). How Much Can We Boost IQ and Scholastic Achievement? *Harbard Educational Review*, *39*(1), 1–124.

Johnson, J. G., Harris, E. S., Spitzer, R. L., & Williams, J. B. W. (2002). The patient health questionnaire for adolescents. *Journal of Adolescent Health*, *30*(3), 196–204. https://doi.org/10.1016/S1054-139X(01)00333-0

Johnson, S. U., Ulvenes, P. G., Øktedalen, T., & Hoffart, A. (2019). Psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. *Frontiers in Psychology*, *10*(JULY), 1–8. https://doi.org/10.3389/fpsyg.2019.01713

Joiner, T. (2011). *Myths about suicide*. Harvard University Press.

Jubal, J. S., Luque, J. C. G., & González, F. J. Á. (2007). Protocolo de exploración médico-psicológica para centros de reconocimiento de conductores. Guía para la historia clínica. In *Ministerio de Sanidad y Consumo*.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Karimi, A., Bazyar, J., Malekyan, L., & Daliri, S. (2021). Prevalence of Suicidal Ideation and Suicide Attempts after Disaster and Mass Casualty Incidents in the World: A Systematic Review and Meta-Analysis. *Iranian Journal of Psychiatry*. https://doi.org/10.18502/ijps.v17i1.8054

Karni-Visel, Y., Hershkowitz, I., Lamb, M. E., & Blasbalg, U. (2019). Facilitating the Expression of Emotions by Alleged Victims of Child Abuse During Investigative Interviews Using the Revised NICHD Protocol. *Child Maltreatment*, *24*(3), 310–318. https://doi.org/10.1177/1077559519831382

Kempen, G. I. J. M., Yardley, L., van Haastregt, J. C. M., Zijlstra, G. A. R., Beyer, N., Hauer, K., Todd, C., GIJ, K., Yardley, L., JCM, van H., GAR, Z., Beyer, N., Hauer, K., & Todd, C. (2008). The Short FES-I: a shortened version of the Falls Efficacy Scale-International to assess fear of falling. *Age & Ageing*, *37*(1), 45–50. https://doi.org/10.1093/ageing/afm157

Kennedy, J. J. (1970). The Eta Coefficient in Complex Anova Designs. *Educational and Psychological Measurement*, *30*(4), 885–889. https://doi.org/10.1177/001316447003000409

Keum, B. T., Miller, M. J., & Inkelas, K. K. (2018). Testing the factor structure and measurement invariance of the PHQ-9 across racially diverse U.S. college students. *Psychological Assessment*, *30*(8), 1096–1106. https://doi.org/10.1037/pas0000550

Klomek, A. B. (2020). Suicide prevention during the COVID-19 outbreak. *The Lancet Psychiatry*, *7*(5), 390. https://doi.org/10.1016/S2215-0366(20)30142-5

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krijnen, W. P. (1996). Algorithms for unweighted least-squares factor analysis. *Computational Statistics & Data Analysis*, *21*(2), 133–147. https://doi.org/10.1016/0167-9473(95)00011-9

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 Validity of a Brief Depression Severity Measure. *J Gen Intern Med*, *16*, 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2002). *The PHQ-15: Validity of a New Measure for Evaluating the Severity of Somatic Symptoms*.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The Patient Health Questionnaire-2. *Medical Care*, *41*(11), 1284–1292. https://doi.org/10.1097/01.MLR.0000093487.78664.3C

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & 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

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry*, *32*(4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Lamela, D., Soreira, C., Matos, P., & Morais, A. (2020). Systematic review of the factor structure and measurement invariance of the patient health questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. *Journal of Affective Disorders*, *276*, 220–233. https://doi.org/10.1016/j.jad.2020.06.066

Lang, B. A., & Harrington, R. (2020). *Validation of the Fear of COVID-19 Scale in a US College Sample*.

Levant, R. F., Allen, P. A., & Lien, M.-C. (2014). Alexithymia in men: How and when do emotional processing deficiencies occur? *Psychology of Men & Masculinity*, *15*(3), 324–334. https://doi.org/10.1037/a0033860

Levant, R., Wimer, D., & Williams, C. (2011). Evaluation of the Health Behavior Inventory – 20 (HBI – 20) and its relationships to masculinity and attitudes towards seeking psychological help. *Psychology of Men & Masculinity*, *12*, 26–41.

Levine, T. R., & Hullett, C. R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research*, *28*(4), 612–625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*, l1476. https://doi.org/10.1136/bmj.l1476

Levis, B., Sun, Y., He, C., Wu, Y., Krishnan, A., Bhandari, P. M., Neupane, D., Imran, M., Brehaut, E., Negeri, Z., Fischer, F. H., Benedetti, A., & Thombs, B. D. (2020). Accuracy of the PHQ-2

Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression. *JAMA*, *323*(22), 2290. https://doi.org/10.1001/jama.2020.6504

Li, W., Yang, Y., Liu, Z.-H., Zhao, Y.-J., Zhang, Q., Zhang, L., Cheung, T., & Xiang, Y.-T. (2020). Progression of Mental Health Services during the COVID-19 Outbreak in China. *International Journal of Biological Sciences*, *16*(10), 1732–1738. https://doi.org/10.7150/ijbs.45120

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLOS ONE*, *14*(7), e0219854. https://doi.org/10.1371/journal.pone.0219854

Liu, S., Yang, L., Zhang, C., Xiang, Y. T., Liu, Z., Hu, S., & Zhang, B. (2020). Online mental health services in China during the COVID-19 outbreak. *The Lancet Psychiatry*, *7*(4), e17–e18. https://doi.org/10.1016/S2215-0366(20)30077-8

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493–504. https://doi.org/10.1037/h0058543

López-Aguado, M., & Gutiérrez-Provecho, L. (2019). Cómo realizar e interpretar un análisis factorial exploratorio utilizando SPSS. *REIRE. Revista d Innovació i Recerca En Educació*, *12*(2), 1–14. https://doi.org/10.1344/reire2019.12.227057

Lorenzo-Seva, U. (2021). SOLOMON: a method for splitting a sample into equivalent subsamples in factor analysis. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01750-y

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91. https://doi.org/10.3758/BF03192753

Lorenzo-Seva, U., & Ferrando, P. J. (2020). Unrestricted factor analysis of multidimensional test items based on an objectively refined target matrix. *Behavior Research Methods*, *52*(1), 116–130. https://doi.org/10.3758/s13428-019-01209-1

Lorenzo-Seva, U., & Ferrando, P. J. (2021). MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis. *Methodology*, *17*(4), 296–306. https://doi.org/10.5964/meth.7185

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull Method for Selecting the Number of Common Factors. *Multivariate Behavioral Research*, *46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

Löwe, B., Spitzer, R. L., Williams, J. B. W., Mussell, M., Schellberg, D., & Kroenke, K. (2008). Depression, anxiety and somatization in primary care: syndrome overlap and functional impairment. *General Hospital Psychiatry*, *30*(3), 191–199. https://doi.org/10.1016/j.genhosppsych.2008.01.001

Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004). Monitoring Depression Treatment Outcomes With the Patient Health Questionnaire-9. *Medical Care*, *42*(12), 1194–1201. https://doi.org/10.1097/00005650-200412000-00006

Löwe, B., Wahl, I., Rose, M., Spitzer, C., Glaesmer, H., Wingenfeld, K., Schneider, A., & 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

Luo, Y., & Keefer, L. (2021). Role of psychological questionnaires in clinical practice and research within functional gastrointestinal disorders. *Neurogastroenterology & Motility*, *33*(12). https://doi.org/10.1111/nmo.14297

Lusted, L. B. (1971). Signal Detectability and Medical Decision-Making. *Science*, *171*(3977), 1217–1219. https://doi.org/10.1126/science.171.3977.1217

MacCallum, R., Roznowski, M., & Necowitz, L. (1992). Model Modifications in Covariance Structure Analysis: The problem of Capitalization on Chance. *Quantitative Methods in Psychology*, *111*(3), 490–504.

Magano, J., Vidal, D. G., e Sousa, H. F. P., Pimenta Dinis, M. A., & Leite, Â. (2021). Validation and psychometric properties of the portuguese version of the coronavirus anxiety scale (Cas) and fear of covid-19 scale (fcv-19s) and associations with travel, tourism and hospitality. *International Journal of Environmental Research and Public Health*, *18*(2), 1–14. https://doi.org/10.3390/ijerph18020427

Mahmood, Q. K., Jafree, S. R., & Qureshi, W. A. (2020). *The Psychometric Validation of FCV19S in Urdu and Socio-Demographic Association with Fear in the People of the Khyber Pakhtunkhwa ( KPK ) Province in Pakistan*.

Malik, S., Ullah, I., Irfan, M., Ahorsu, D. K., Lin, C.-Y., Pakpour, A. H., Griffiths, M. D., Rehman, I. U., & Minhas, R. (2021). Fear of COVID-19 and workplace phobia among Pakistani doctors: A survey study. *BMC Public Health*, *21*(1), 833. https://doi.org/10.1186/s12889-021-10873-y

Maris, R. W., Berman, A. L., Maltsberger, J. T., & Yufit, R. I. (1992). Assessment and prediction of suicide. *This Volume Is Based on a Workshop Entitled" Assessment and Prediction of Suicide" Held at the 1990 Annual Meeting of the American Association of Suicidology (ASS) in New Orleans, Louisiana.*

Martin, L. A., Neighbors, H. W., & Griffith, D. M. (2013). The Experience of Symptoms of Depression in Men vs Women. *JAMA Psychiatry*, *70*(10), 1100. https://doi.org/10.1001/jamapsychiatry.2013.1985

Martínez-Lorca, M., Martínez-Lorca, A., Criado-Álvarez, J. J., Armesilla, M. D. C., & Latorre, J. M. (2020). The fear of COVID-19 scale: Validation in spanish university students. *Psychiatry Research*, *293*, 113350. https://doi.org/10.1016/j.psychres.2020.113350

Marty, M. A., Segal, D. L., & Coolidge, F. L. (2010). Relationships among dispositional coping strategies, suicidal ideation, and protective factors against suicide in older adults. *Aging and Mental Health*, *14*(8), 1015–1023.

Matza, L. S., Morlock, R., Sexton, C., Malley, K., & Feltner, D. (2010). Identifying HAM-A cutoffs for mild, moderate, and severe generalized anxiety disorder. *International Journal of Methods in Psychiatric Research*, *19*(4), 223–232. https://doi.org/10.1002/mpr.323

McArdle, R., Chisolm, T. H., Abrams, H. B., Wilson, R. H., & Doyle, P. J. (2005). The WHO-DAS II: Measuring Outcomes of Hearing Aid Intervention for Adults. *Trends in Amplification*, *9*(3), 127–143. https://doi.org/10.1177/108471380500900304

McCarron, R. M., Vanderlip, E. R., & Rado, J. (2016). Depression. *Annals of Internal Medicine*, *165*(7), ITC49. https://doi.org/10.7326/AITC201610040

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of Fit in Item Response Models. *Multivariate Behavioral Research*, *30*(1), 23–40. https://doi.org/10.1207/s15327906mbr3001_2

Mergl, R., Koburger, N., Heinrichs, K., Székely, A., Tóth, M. D., Coyne, J., Quintão, S., Arensman, E., Coffey, C., Maxwell, M., Värnik, A., van Audenhove, C., McDaid, D., Sarchiapone, M., Schmidtke, A., Genz, A., Gusmão, R., & Hegerl, U. (2015). What Are Reasons for the Large Gender Differences in the Lethality of Suicidal Acts? An Epidemiological Analysis in Four European Countries. *PLOS ONE*, *10*(7), e0129062. https://doi.org/10.1371/journal.pone.0129062

Merz, E. L., Malcarne, V. L., Roesch, S. C., Riley, N., & Sadler, G. R. (2011). A multigroup confirmatory factor analysis of the Patient Health Questionnaire-9 among English- and Spanish-speaking Latinas. *Cultural Diversity and Ethnic Minority Psychology*, *17*(3), 309–316. https://doi.org/10.1037/a0023883

Meyer, E. P., Kaiser, H. F., Cerny, B. A., & Green, B. F. (1977). MSA for a special spearman matrix. *Psychometrika*, *42*(1), 153–156. https://doi.org/10.1007/BF02293753

Midorikawa, H., Aiba, M., Lebowitz, A., Taguchi, T., Shiratori, Y., Ogawa, T., Takahashi, A., Takahashi, S., Nemoto, K., Arai, T., & Tachikawa, H. (2021). Confirming validity of the Fear of COVID-19 Scale in Japanese with a nationwide largescale sample. In *PLoS ONE* (Vol. 16, Issue 2 February). https://doi.org/10.1371/journal.pone.0246840

Oficialización de la Norma de atención integral de la salud mental y de abordaje psicosocial en situaciones de emergencias y desastres en los escenarios de servicios de salud y en la comunidad N° 41599 - S, Pub. L. No. N° 41599-S (2019). http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=88556&nValor3=115853&strTipM=TC

Ministerio de Sanidad Consumo y Bienestar Social. (2020). *CIE • 10 • ES. Clasificación Internacional de Enfermedades - 10.ª Revisión Modificación Clínica*. http://eciemaps.mcsb.gob.es/ecieMaps/errata/errata.html

Mirabet, E., Ozcoidi, M., Sanz, R., Patricia, P., Valdés, E., Gil, S., & Justo, S. (2022). *Protocolo de exploración médico-psicológica para Centros de Reconocimiento de Conductores. Actualización 2022.* (& Dirección General de Tráfico. Ministerio del Interior & Dirección General de Salud Pública. Ministerio de Sanidad, Eds.; 3ª Edición). https://www.sanidad.gob.es/en/profesionales/saludPublica/prevPromocion/Prevencion/SeguridadVial/docs/Centros_reconocimiento_conductores.pdf

Mitchell, A. J., Yadegarfar, M., Gill, J., & Stubbs, B. (2016). Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. *BJPsych Open*, *2*(2), 127–138. https://doi.org/10.1192/bjpo.bp.115.001685

Mohsen, F., Bakkar, B., Alsrouji, S. K., Abbas, E., Najjar, A., Marrawi, M., & Latifeh, Y. (2022). Fear among Syrians: A Proposed Cutoff Score for the Arabic Fear of COVID-19 Scale. *PLoS ONE*, *17*(3 March). https://doi.org/10.1371/journal.pone.0264257

Molina Arias, M. (2013). Características de las pruebas diagnósticas. *Pediatría Atención Primaria*, *15*(58), 169–173. https://doi.org/10.4321/S1139-76322013000200013

Molinero, L. M. (2003). Elección de los puntos de corte para convertir una variable cuantitativa en cualitativa. *Asociación de La Sociedad Española de Hipertensión*. http://www.seh-lelha.org/stat1.htm

Morrison, & James. (2014). *DSM-5 ® Guía para el diagnóstico clínico*.

Moses, L. E., Shapiro, D., & Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine*, *12*(14), 1293–1316. https://doi.org/10.1002/sim.4780121403

Muller, A. E., Himmels, J. P. W., & van de Velde, S. (2021). Instruments to measure fear of COVID-19: a diagnostic systematic review. *BMC Medical Research Methodology*, *21*(1), 1–14. https://doi.org/10.1186/s12874-021-01262-5

Muñoz-Navarro, R., Cano-Vindel, A., Medrano, L. A., Schmitz, F., Ruiz-Rodríguez, P., Abellán-Maeso, C., Font-Payeras, M. A., & Hermosilla-Pasamar, A. M. (2017). Utility of the PHQ-9 to identify major depressive disorder in adult patients in Spanish primary care centres. *BMC Psychiatry*, *17*(1). https://doi.org/10.1186/s12888-017-1450-8

Muraki, E. (1992). A GENERALIZED PARTIAL CREDIT MODEL: APPLICATION OF AN EM ALGORITHM. *ETS Research Report Series*, *1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muthén, B. O. (1993). Testing Structural Equation Models. In K. A. Bollen & J. S. Long (Eds.), *Goodness of Fit with Categorical and Other Non-Nor-mal Variables* (pp. 205–243). Sage.

Nallusamy, V., Afgarshe, M., & Shlosser, H. (2016). Reliability and validity of Somali version of the PHQ-9 in primary care practice. *The International Journal of Psychiatry in Medicine*, *51*(6), 508–520. https://doi.org/10.1177/0091217417696732

National Institute of Mental Health. (2022). *Major depression*. https://www.nimh.nih.gov/health/statistics/major-depression

Nikopoulou, V. A., Holeva, V., Parlapani, E., Karamouzi, P., Voitsidis, P., Porfyri, G. N., Blekas, A., Papigkioti, K., Patsiala, S., & Diakogiannis, I. (2022). Mental Health Screening for COVID-19: a Proposed Cutoff Score for the Greek Version of the Fear of COVID-19 Scale (FCV-19S). *International Journal of Mental Health and Addiction*, *20*(2), 907–920. https://doi.org/10.1007/s11469-020-00414-w

Nunnally, J. C. (1970). *Introducción a la medición psicológica*. Paidós.

Olejnik, S., & Algina, J. (2003). Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Ong, C. W., Pierce, B. G., Klein, K. P., Hudson, C. C., Beard, C., & Björgvinsson, T. (2022). Longitudinal Measurement Invariance of the PHQ-9 and GAD-7. *Assessment*, *29*(8), 1901–1916. https://doi.org/10.1177/10731911211035833

Pakpour, A. H., Griffiths, M. D., & Lin, C. Y. (2020). Assessing Psychological Response to the COVID-19: The Fear of COVID-19 Scale and the COVID Stress Scales. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00334-9

Pang, N. T. P., Kamu, A., Hambali, N. L. B., Mun, H. C., Kassim, M. A., Mohamed, N. H., Ayu, F., Rahim, S. S. S. A., Omar, A., & Jeffree, M. S. (2022). Malay Version of the Fear of COVID-19 Scale: Validity and Reliability. *International Journal of Mental Health and Addiction*, *20*(1), 263–272. https://doi.org/10.1007/s11469-020-00355-4

Parmar, D., Stavropoulou, C., & Ioannidis, J. P. A. (2016). Health outcomes during the 2008 financial crisis in Europe: systematic literature review. *BMJ*, i4588. https://doi.org/10.1136/bmj.i4588

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, *163*(7), 670–675. https://doi.org/10.1093/aje/kwj063

Perz, C. A., Lang, B. A., & Harrington, R. (2022). Validation of the Fear of COVID-19 Scale in a US College Sample. *International Journal of Mental Health and Addiction*, *20*(1), 273–283. https://doi.org/10.1007/s11469-020-00356-3

Piqueras, J. A., Gomez-Gomez, M., Marzo, J. C., Gomez-Mir, P., Falco, R., Valenzuela, B., Falcó, R., Lopez-Nuñez, A., Martínez-González, A. E., Marzo, J. C., Mateu, O., & Moreno-Amador, B. (2021). Validation of the Spanish Version of Fear of COVID-19 Scale: its Association with Acute Stress and Coping. *International Journal of Mental Health and Addiction*, *0123456789*. https://doi.org/10.1007/s11469-021-00615-x

Piqueras, J., Gomez-Gomez, M., Marzo, J. C., Gomez-Mir, P., Falco, R., & Valenzuela, B. (2020). *Validation of the Spanish version of Fear of COVID-19 Scale Its association with acute stress and coping*. https://doi.org/https://doi.org/10.21203/rs.3.rs-75063/v1 License:

Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016). Screening for anxiety disorders with the GAD-7 and GAD-2: A systematic review and diagnostic metaanalysis. *General Hospital Psychiatry*, *39*, 24–31. https://doi.org/10.1016/j.genhosppsych.2015.11.005

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *General Psychiatry*, *33*, 100213. https://doi.org/10.1136/gpsych-2020-100213

Ramón, J., López, J., & Ramos, F. Á. (2018). *Evaluación Psicológica Forense de los abusos y maltratos a niños, niñas y adolescentes. Guía de buenas prácticas*. Asociación de Psicólogos Forenses de la administración de justicia.

Razykov, I., Ziegelstein, R. C., Whooley, M. A., & Thombs, B. D. (2012). The PHQ-9 versus the PHQ-8 — Is item 9 useful for assessing suicide risk in coronary artery disease patients? Data from the Heart and Soul Study. *Journal of Psychosomatic Research*, *73*(3), 163–168. https://doi.org/10.1016/j.jpsychores.2012.06.001

Ren, X., Huang, W., Pan, H., Huang, T., Wang, X., & Ma, Y. (2020). Mental Health During the Covid-19 Outbreak in China: a Meta-Analysis. *Psychiatric Quarterly*, *91*(4), 1033–1045. https://doi.org/10.1007/s11126-020-09796-5

Reznik, A., Gritsenko, V., Konstantinov, V., Khamenka, N., & Israelowitz, R. (2020). COVID-19 Fear in Eastern Europe: Validation of the Fear of COVID-19 Scale. In *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00283-3

Reznik, A., Gritsenko, V., Konstantinov, V., Khamenka, N., & Israelowitz, R. (2021). COVID-19 Fear in Eastern Europe: Validation of the Fear of COVID-19 Scale. *International Journal of Mental Health and Addiction*, *19*(5), 1903–1908. https://doi.org/10.1007/s11469-020-00283-3

Richardson, L. P., Rockhill, C., Russo, J. E., Grossman, D. C., Richards, J., McCarty, C., McCauley, E., & Katon, W. (2010). Evaluation of the PHQ-2 as a brief screen for detecting major depression among adolescents. *Pediatrics*, *125*(5). https://doi.org/10.1542/peds.2009-2712

Rick H. Hoyle. (2023). *Handbook of Structural Equation Modeling. Concepts, Issues, and Applications* (Rick H. Hoyle, Ed.; Second Edition). SAGE Publications, Inc.

Rivas-Ruiz, R., Castelán-Martínez, O. D., Pérez, M., & Talavera, J. O. (2013). Clinical research XVII. χ2 test, from the expected to the observed. *Rev Med Inst Mex Seguro Soc*, *51*(5), 552–557.

Robles-bello, A. (2014). *Escala de Resiliencia 14 ítems (RS-14): Propiedades Psicométricas de la Versión en Español*. *2*(January 2016), 103–113.

Roncero, C., González-Sánchez, A., Pérez-Laureano, Á., Ortiz-Fune, C., Díaz-Trejo, S., Bersabé-Pérez, M., Braquehais, M. D., Pérez-Rodríguez, J., Maderuelo-Fernández, J. Á., & Benito-Sánchez, J. A. (2022). The challenge of community mental health interventions with patients, relatives, and health professionals during the COVID-19 pandemic: a real-world 9-month follow-up study. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-25297-w

Sakib, N., Bhuiyan, A. K. M. I., Hossain, S., al Mamun, F., Hosen, I., Abdullah, A. H., Sarker, M. A., Mohiuddin, M. S., Rayhan, I., Hossain, M., Sikder, M. T., Gozal, D., Muhit, M., Islam, S. M. S., Griffiths, M. D., Pakpour, A. H., & Mamun, M. A. (2020). Psychometric Validation of the Bangla Fear of COVID-19 Scale: Confirmatory Factor Analysis and Rasch Analysis. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00289-x

Sakib, N., Bhuiyan, A. K. M. I., Hossain, S., & Mamun, F. al. (2020). *Psychometric Validation of the Bangla Fear of COVID-19 Scale : Confirmatory Factor Analysis and Rasch Analysis*.

Saldivia, S., Aslan, J., Cova, F., Vicente, B., Inostroza, C., & Rincón, P. (2019). Propiedades psicométricas del PHQ-9 (Patient Health Questionnaire) en centros de atención primaria de Chile. *Revista Médica de Chile*, *147*(1), 53–60. https://doi.org/10.4067/S0034-98872019000100053

Salk, R. H., Hyde, J. S., & Abramson, L. Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological Bulletin*, *143*(8), 783–822. https://doi.org/10.1037/bul0000102

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97. https://doi.org/10.1007/BF03372160

Sant, O., Sapienza, L., Nazionale, C., Superiore, I., Regionale, R., Maria, O. S., & Sapienza, L. (2007). *Depression , hopelessness and suicide risk among patients suffering from epilepsy*. *c*, 425–429.

Saywitz, K. J., Geiselman, R. E., & Bornstein, G. K. (1992). Effects of cognitive interviewing and practice on children's recall performance. *Journal of Applied Psychology*, *77*(5), 744–756. https://doi.org/10.1037/0021-9010.77.5.744

Scientific Software International. (2023). *Lisrel v.8.50*. Mooresville, Ind. https://ssicentral.com/index.php/products/lisrel/lisrel-licenses/

Scoppetta, O., Cassiani-Miranda, C. A., Arocha-Díaz, K. N., Cabanzo-Arenas, D. F., & Campo-Arias, A. (2021). Validity of the patient health questionnaire-2 (PHQ-2) for the detection of depression in primary care in Colombia. *Journal of Affective Disorders*, *278*(55), 576–582. https://doi.org/10.1016/j.jad.2020.09.096

Shenassa, E. D., Rogers, M. L., Spalding, K. L., & Roberts, M. B. (2004). Safer storage of firearms at home and risk of suicide: a study of protective factors in a nationally representative sample. *Journal of Epidemiology & Community Health*, *58*(10), 841–848.

Shi, P., Yang, A., Zhao, Q., Chen, Z., Ren, X., & Dai, Q. (2021). A Hypothesis of Gender Differences in Self-Reporting Symptom of Depression: Implications to Solve Under-Diagnosis and Under-Treatment of Depression in Males. *Frontiers in Psychiatry*, *12*. https://doi.org/10.3389/fpsyt.2021.589687

Shirkey, E. C., & Dziuban, C. D. (1976). A Note on Some Sampling Characteristics of the Measure of Sampling Adequacy (MSA). *Multivariate Behavioral Research*, *11*(1), 125–128. https://doi.org/10.1207/s15327906mbr1101_9

Snijkers, J. T. W., Oever, W., Weerts, Z. Z. R. M., Vork, L., Mujagic, Z., Leue, C., Hesselink, M. A. M., Kruimel, J. W., Muris, J. W. M., Bogie, R. M. M., Masclee, A. A. M., Jonkers, D. M. A. E., & Keszthelyi, D. (2021). Examining the optimal cutoff values of HADS, PHQ-9 and GAD-7 as screening instruments for depression and anxiety in irritable bowel syndrome. *Neurogastroenterology & Motility*, *33*(12). https://doi.org/10.1111/nmo.14161

Soares, F. R., Afonso, R. M., Martins, A. P., Pakpour, A. H., & Rosa, C. P. (2021). The fear of the COVID-19 Scale: validation in the Portuguese general population. *Death Studies*, 1203. https://doi.org/10.1080/07481187.2021.1889722

Soper, D. S. (2022). *A-priori Sample Size Calculator for Structural Equation Models*. https://www.danielsoper.com/statcalc

Soraci, P., Ferrari, A., Abbiati, F. A., del Fante, E., de Pace, R., Urso, A., & Griffiths, M. D. (2020). Validation and Psychometric Evaluation of the Italian Version of the Fear of COVID-19 Scale. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00277-1

Soto-Briseño, A. I., Gómez-Díaz, R. A., Valdez-González, A. L., Saldaña-Espinoza, R. C., Favila Bojórquez, J. J., & Wacher, N. H. (2021). Fear of covid-19 scale: Validation in spanish in

the mexican general population. *Gaceta Medica de Mexico*, *157*(6), 586–593. https://doi.org/10.24875/GMM.21000147

Spitzer, R. L. (1999). Validation and Utility of a Self-report Version of PRIME-MD<SUBTITLE>The PHQ Primary Care Study</SUBTITLE> *JAMA*, *282*(18), 1737. https://doi.org/10.1001/jama.282.18.1737

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Spitzer, R. L., Williams, J. B. W., Johnson, J. G., Kroenke, K., Linzer, M., Degruy, F. V., Brody, D., & Hahn, S. R. (1994). Utility of a New Procedure for Diagnosing Mental Disorders in Primary Care: The PRIME-MD 1000 Study. *JAMA: The Journal of the American Medical Association*, *272*(22), 1749–1756. https://doi.org/10.1001/jama.1994.03520220043029

Spitzer, R. L., Williams, J. B. W., Kroenke, K., Hornyak, R., & McMurray, J. (2000). Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: The PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *American Journal of Obstetrics and Gynecology*, *183*(3), 759–769. https://doi.org/10.1067/mob.2000.106580

SPSS Corp., I. (2017). *IBM SPSS Statistics for Windows v.25. Armonk, NY: IBM Corp.*

Stănculescu, E. (2021). Fear of COVID-19 in Romania: Validation of the Romanian Version of the Fear of COVID-19 Scale Using Graded Response Model Analysis. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00428-4

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology*, *43*(2), 476–493. https://doi.org/10.1093/ije/dyu038

Swetlitz, N. (2021). Depression's Problem With Men. *AMA Journal of Ethics*, *23*(7), 586–589. https://doi.org/10.1001/amajethics.2021.586

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, *240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better Decisions through Science. *Scientific American*, *283*(4), 82–87. https://doi.org/10.1038/scientificamerican1000-82

Szanto, K., Prigerson, H. G., & Reynolds, C. F. (2001). Suicide in the elderly. *Clinical Neuroscience Research*, *1*(5), 366–376. https://doi.org/10.1016/S1566-2772(01)00039-1

Tanzer, N. K. (1995). Cross-Cultural Bias in Likert-Type Inventories: Perfect Matching Factor Structures and Still Biased? *European Journal of Psychological Assessment*, *11*(3), 194–201. https://doi.org/10.1027/1015-5759.11.3.194

Tennant, A., & Pallant, J. F. (2012). Transactions of the Rasch Measurement SIG. *American Educational Research Association*, *25*(4), 1339–1350. www.leeds.ac.uk/medicine/rehabmed/psychometric

Teymoori, A., Gorbunova, A., Haghish, F. E., Real, R., Zeldovich, M., Wu, Y. J., Polinder, S., Asendorf, T., Menon, D., & Steinbüchel, N. v. (2020). Factorial structure and validity of depression (Phq-9) and anxiety (gad-7) scales after traumatic brain injury. *Journal of Clinical Medicine*, *9*(3). https://doi.org/10.3390/jcm9030873

Thomson, I. (2013, September 27). 30 years on: The day a computer glitch nearly caused World War III. *The Register*. https://www.theregister.com/2013/09/27/30_years_on_the_day_a_computer_glitch_nearly_caused_world_war_iii/

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353

Tornimnbeni, S., Edgardo, P., & Olaz, F. (2008). *Introducción a la Psicometría*. Paidós.

Tzur Bitan, D., Grossman-Giron, A., Bloch, Y., Mayer, Y., Shiffman, N., & Mendlovic, S. (2020). Fear of COVID-19 scale: Psychometric characteristics, reliability and validity in the Israeli population. *Psychiatry Research*, *289*(May). https://doi.org/10.1016/j.psychres.2020.113100

Tzur, D., Grossman-giron, A., Bloch, Y., Mayer, Y., & Shi, N. (2020). *Fear of COVID-19 scale : Psychometric characteristics , reliability and validity in the Israeli population*. *289*(May). https://doi.org/10.1016/j.psychres.2020.113100

van den Brink, W. (1982). Binomial test models for domain-referenced testing. *Evaluation in Education*, *5*(2), 165–176. https://doi.org/10.1016/0191-765X(82)90016-7

Vicente-Villardón, J. L. (2016). *Multbiplot: A package for Multivariate Analysis Using Biplots* (180312). Departamento de Estadística. Universidad de Salamanca. http://biplot.usal.es/multbiplot/

Viglione, D., Giromini, L., Gustafson, M. L., & Meyer, G. J. (2014). Developing Continuous Variable Composites for Rorschach Measures of Thought Problems, Vigilance, and Suicide Risk. *Assessment*, *21*(1), 42–49. https://doi.org/10.1177/1073191112446963

Villegas, G., Arias-LeClaire, H., González-García, N., González-Sánchez, A., León, I., & Avarado, F. (2020). *Siguiendo el COVID-19. Percepción de riesgo de contagio en la población costarricense.*

Vozmediano, L., San Juan, C., & Vergara, A. (2008). Problemas de medición del miedo al delito: algunas respuestas teóricas y técnicas. *Revista Electrónica de Ciencia Penal y Criminología*, *07*(10), 8.

Wagnild, G. (2022). *The Resilient Center*. https://www.resiliencecenter.com/products/resilience-scales-and-tools-for-research/the-rs14/

Wagnild, G. M., & Collins, J. A. (2009). Assessing Resilience. *JOURNAL OF PSYCHOSOCIAL NURSING AND MENTAL HEALTH SERVICES*, *47*(12), 28–33. https://doi.org/10.3928/02793695-20091103-01

Wakashima, K., Asai, K., Kobayashi, D., Koiwa, K., Kamoshida, S., & Sakuraba, M. (2020). The Japanese version of the Fear of COVID-19 scale: Reliability, validity, and relation to coping behavior. *PLoS ONE*, *15*(11 November). https://doi.org/10.1371/journal.pone.0241958

Wicke, F. S., Krakau, L., Löwe, B., Beutel, M. E., & 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

Wilson, S., & Durbin, C. E. (2010). Effects of paternal depression on fathers' parenting behaviors: A meta-analytic review. *Clinical Psychology Review*, *30*(2), 167–180. https://doi.org/10.1016/j.cpr.2009.10.007

Winter, T., Riordan, B. C., Pakpour, A. H., Griffiths, M. D., Mason, A., Poulgrain, J. W., & Scarf, D. (2020). *Evaluation of the English Version of the Fear of COVID-19 Scale and Its Relationship with Behavior Change and Political Beliefs*.

Wu, Y., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., Rice, D. B., Boruff, J., Cuijpers, P., Gilbody, S., Ioannidis, J. P. A., Kloda, L. A., McMillan, D., Patten, S. B., Shrier, I., Ziegelstein, R. C., Akena, D. H., Arroll, B., Ayalon, L., … Thombs, B. D. (2020). Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychological Medicine*, *50*(8), 1368–1380. https://doi.org/10.1017/S0033291719001314

Yahya, A. S., Khawaja, S., & Chukwuma, J. (2020). The Impact of COVID-19 in Psychiatry. *The Primary Care Companion for CNS Disorders*, *22*(2).

Yamamoto-Furusho, J. K., Sarmiento-Aguilar, A., García-Alanis, M., Gómez-García, L. E., Toledo-Mauriño, J., Olivares-Guzmán, L., & Fresán-Orellana, A. (2018). Escala de Ansiedad y Depresión Hospitalaria (HADS): Validación en pacientes mexicanos con enfermedad inflamatoria intestinal. *Gastroenterología y Hepatología*, *41*(8), 477–482. https://doi.org/10.1016/j.gastrohep.2018.05.009

Yan, Y., Hou, J., Li, Q., & Yu, N. X. (2023). Suicide before and during the COVID-19 Pandemic: A Systematic Review with Meta-Analysis. *International Journal of Environmental Research and Public Health*, *20*(4), 3346. https://doi.org/10.3390/ijerph20043346

Yerushalmy, J. (1896). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports*, *62*(40), 1432–1449.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Yuryev, A., Leppik, L., Tooding, L.-M., Sisask, M., Värnik, P., Wu, J., & Värnik, A. (2010). Social inclusion affects elderly suicide mortality. *International Psychogeriatrics*, *22*(8), 1337–1343.

Zelviene, P., Jovarauskaite, L., & Truskauskaite-Kuneviciene, I. (2021). The Psychometric Properties of the Resilience Scale (RS-14) in Lithuanian Adolescents. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.667285

Zemła, A. J., Nowicka-Sauer, K., Jarmoszewicz, K., Wera, K., Batkiewicz, S., & Pietrzykowska, M. (2019). Measures of preoperative anxiety. *Anestezjologia Intensywna Terapia*, *51*(1), 64–69. https://doi.org/10.5603/AIT.2019.0013

Zhang, J., & Jia, C. X. (2010). Attitudes toward suicide: The effect of suicide death in the family. *Omega: Journal of Death and Dying*, *60*(4), 365–382. https://doi.org/10.2190/OM.60.4.d

Zhou, Y., Xu, J., & Rief, W. (2020). Are comparisons of mental disorders between Chinese and German students possible? An examination of measurement invariance for the PHQ-15, PHQ-9 and GAD-7. *BMC Psychiatry*, *20*(1), 480. https://doi.org/10.1186/s12888-020-02859-8

Zou, K. H., Hall, W. J., & Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, *16*(19), 2143–2156. https://doi.org/10.1002/(SICI)1097-0258(19971015)16:19<2143::AID-SIM655>3.0.CO;2-3

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation*, *115*(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

# Capítulo 11

## Anexos

# 11.- Anexos

## Anexo I. PHQ-9

Durante las últimas 2 semanas, ¿con qué frecuencia ha sentido molestias por los siguientes problemas? Siendo (0) Para nada, (1) Varios días, (2) Más de la mitad de los días y (3) Casi todos los días.

| Nº | Ítem | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 1 | Poco interés o placer en hacer las cosas | | | | |
| 2 | Sentirse con depresión, sin ánimo o sin esperanzas | | | | |
| 3 | Problemas para iniciar o mantener el sueño o dormir demasiado | | | | |
| 4 | Sentir cansancio o con poca energía | | | | |
| 5 | Sentir poco apetito o comer en exceso | | | | |
| 6 | Sentirse mal acerca de sí mismo o tener un sentimiento de fracaso o de abandono propio o de la familia | | | | |
| 7 | Dificultad para concentrarse en diferentes actividades tales como leer el periódico o ver televisión | | | | |
| 8 | Moverse o hablar tan despacio que otras personas lo han notado o bien, por el contrario, estar con tanta inquietud o intranquilidad, que se mueve mucho más de lo normal | | | | |
| 9 | Pensamientos de deseo de muerte o que quisiera lastimarse de alguna manera | | | | |

Si ha marcado alguno de los problemas de este cuestionario, ¿hasta qué punto estos problemas le han creado dificultades para hacer su trabajo, ocuparse de la casa o relacionarse con los demás? Conteste con (0) Ninguna dificultad, (1) Un poco de dificultad, (2) Mucha dificultad y (3) Extremada dificultad.

| Nº | Ítem | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 10 | Si usted tuvo molestias por alguno de los problemas mencionados, ¿cuánta dificultad le causaron estos problemas para **hacer su trabajo**? | | | | |
| 11 | Si usted tuvo molestias por alguno de los problemas mencionados, ¿cuánta dificultad le causaron estos problemas para **encargarse de las tareas del hogar**? | | | | |
| 12 | Si usted tuvo molestias por alguno de los problemas mencionados, ¿cuánta dificultad le causaron estos problemas para **relacionarse con los demás**? | | | | |

Corrección:

| | |
|---|---|
| (0-4) | Depresión no detectada |
| (5-9) | Síntomas mínimos |
| (10-14) | Depresión menor, distimia, depresión mayor leve |
| (15-19) | Depresión mayor, de moderada a severa |
| (≥20) | Depresión mayor severa |

## Anexo II. PHQ-4

Durante las últimas 2 semanas, ¿con qué frecuencia ha sentido molestias por los siguientes problemas? Siendo (0) Para nada, (1) Varios días, (2) Más de la mitad de los días y (3) Casi todos los días.

| Nº | Factor | Ítem | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|
| 1 | Depresión | Poco interés o placer en hacer las cosas | | | | |
| 2 | Depresión | Sentirse con depresión, sin ánimo o sin esperanzas | | | | |
| 3 | Ansiedad | Sentir nervios, angustia o mucha tensión | | | | |
| 4 | Ansiedad | No poder dejar de preocuparse o no poder controlar la preocupación | | | | |

Corrección:

Cuanto más elevada sea la puntuación, más probabilidad existe de que haya un trastorno depresivo o de ansiedad. Se corrigen ambos test por separado. Puntuaciones mayores a 3 se consideran como caso.

## Anexo III. GAD-7

Durante el último mes, ¿con qué frecuencia ha sentido molestias por los siguientes problemas? Las opciones de respuesta son: (0) Para nada, (1) Varios días, (2) Más de la mitad de los días y (3) Casi todos los días.

| Nº | Ítem | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 1 | Sentirse nervioso/a, angustiado/a o muy tenso/a | | | | |
| 2 | No poder dejar de preocuparse o no poder controlar la preocupación | | | | |
| 3 | Preocuparse demasiado por diferentes cosas | | | | |
| 4 | Dificultad para relajarse | | | | |
| 5 | Se ha sentido tan inquieto/a que no ha podido quedarse quieto/a | | | | |
| 6 | Molestarse o enojarse fácilmente | | | | |
| 7 | Sentir miedo como si algo terrible pudiera pasar | | | | |

Corrección:

| | |
|---|---|
| (0-4) | Mínimo |
| (5-9) | Leve |
| (10-14) | Moderado |
| (15-21) | Severo |

## Anexo IV. Escala de Miedo al Covid-19

Seleccione su grado de acuerdo o desacuerdo con las siguientes afirmaciones siendo (1) Muy en desacuerdo, (2) En desacuerdo, (3) Ni de acuerdo ni en desacuerdo, (4) De acuerdo y (5) Muy de acuerdo.

| Nº | Ítem | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | Tengo miedo del coronavirus-19. | | | | | |
| 2 | Me siento incómodo al pensar sobre el coronavirus-19. | | | | | |
| 3 | Me sudan las manos al pensar en el coronavirus-19. | | | | | |
| 4 | Tengo miedo de perder la vida por el coronavirus-19. | | | | | |
| 5 | Cuando veo noticias o escucho historias sobre el coronavirus-19 en redes sociales, me pongo nervioso o ansioso | | | | | |
| 6 | No puedo dormir por preocupación de contagiarme de coronavirus-19. | | | | | |
| 7 | Mi corazón se acelera o palpita cuando pienso sobre contagiarme de coronavirus-19 | | | | | |

Corrección:

Cuanto más elevada sea la puntuación, más probabilidad existe de poseer miedo a la COVID-19.

## Anexo V. Escala de Resiliencia RS-14

Seleccione el grado de acuerdo o desacuerdo que mejor indique sus sentimientos en las siguientes afirmaciones en una escala de 1 a 7, siendo 1 en Desacuerdo y 7 De acuerdo.

| Nº | Ítem | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 1 | Normalmente me las arreglo de una manera u otra | | | | | | | |
| 2 | Me siento orgulloso/a de las cosas que he logrado | | | | | | | |
| 3 | En general, me tomo las cosas con calma | | | | | | | |
| 4 | Soy una persona con una adecuada autoestima | | | | | | | |
| 5 | Siento que puedo manejar muchas situaciones a la vez | | | | | | | |
| 6 | Soy resuelto/a y decidido/a | | | | | | | |
| 7 | No me asusta sufrir dificultades porque ya las he experimentado en el pasado | | | | | | | |
| 8 | Soy una persona disciplinada | | | | | | | |
| 9 | Pongo interés en las cosas | | | | | | | |
| 10 | Generalmente puedo encontrar algo sobre lo que reírme | | | | | | | |
| 11 | La seguridad en mí mismo/a me ayuda en los momentos difíciles | | | | | | | |
| 12 | En una emergencia soy alguien en quien la gente puede confiar | | | | | | | |
| 13 | Mi vida tiene sentido | | | | | | | |
| 14 | Cuando estoy en una situación difícil, por lo general, puedo encontrar una salida | | | | | | | |

Corrección:

| | |
|---|---|
| (14-30) | Muy baja |
| (31-48) | Baja |
| (49-63) | Normal |
| (64-81) | Alta resiliencia |
| (82-92) | Muy alta resiliencia |

## Anexo VI. HSAY

En los últimos 30 días, ¿con qué frecuencia se identifica con cada una de las siguientes afirmaciones? Las opciones de respuesta son Nunca (1); Raramente (2); A veces (3); Muy a menudo (4); Siempre (5).

| Nº | Ítem | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | Me siento cansado/a. | | | | | |
| 2 | Me resulta muy difícil relajarme o "desconectar". | | | | | |
| 3 | Me resulta difícil tomar decisiones. | | | | | |
| 4 | Mi corazón se acelera y respiro rápidamente. | | | | | |
| 5 | Tengo problemas para pensar con claridad. | | | | | |
| 6 | Como demasiado o muy poco. | | | | | |
| 7 | Tengo dolores de cabeza. | | | | | |
| 8 | Me siento emocionalmente insensible. | | | | | |
| 9 | Pienso en mis problemas una y otra vez durante el día. | | | | | |
| 10 | Tengo problemas para dormir (por ejemplo: problemas para conciliar el sueño, problemas para permanecer dormido, problemas para despertarme, pesadillas, etc.). | | | | | |
| 11 | Me cuesta ser optimista. | | | | | |
| 12 | Tomo riesgos innecesarios o me involucro en comportamientos peligrosos para la salud y/o la seguridad. | | | | | |
| 13 | Tengo dolor de espalda y cuello u otro dolor crónico relacionado con la tensión. | | | | | |
| 14 | Consumo cafeína o nicotina más de lo habitual. | | | | | |
| 15 | Me siento agobiado/a e impotente. | | | | | |
| 16 | Tengo hábitos nerviosos (por ejemplo: morderme las uñas, rechinar los dientes, estar inquieto, caminar de un lado a otro, etc.). | | | | | |
| 17 | Me olvido de las pequeñas cosas (por ejemplo: dónde dejo mis llaves, nombres de personas, detalles hablados durante la última reunión de trabajo). | | | | | |
| 18 | Tengo problemas de estómago (por ejemplo, náuseas, vómitos, diarrea, estreñimiento, gases). | | | | | |
| 19 | Estoy irritable y me molesto fácilmente. | | | | | |
| 20 | Tengo cambios de humor y me siento demasiado emocional. | | | | | |
| 21 | Me cuesta concentrarme. | | | | | |
| 22 | Me cuesta sentir que la vida tiene sentido. | | | | | |
| 23 | Soy reservado/a y me siento distante y aislado de otras personas. | | | | | |
| 24 | Tomo alcohol y/u otras drogas para tratar de ayudarme a salir adelante. | | | | | |
| 25 | Mi rendimiento ha disminuido y tengo problemas para terminar las tareas. | | | | | |

### Propuesta de corrección:

Una puntuación en este rango sugiere que:

| | |
|---|---|
| (0-25) | Probablemente su estrés se encuentre en niveles adecuados. |
| (26-50) | Puede estar experimentando un grado de estrés de bajo a moderado. |
| (51-75) | Puede estar experimentando un grado de estrés de moderado a alto. |
| (76-100) | Puede estar experimentando un grado muy alto de estrés. |

## Anexo VII. PROQOL

Ayudar a otras personas le pone en contacto directo con sus vidas. Como usted habrá comprobado, su compasión o empatía hacia quienes ayuda puede afectar de formas tanto positivas como negativas. Seguidamente se presentan afirmaciones sobre sus experiencias tanto positivas como negativas. Considere cada uno de las siguientes preguntas de acuerdo con su situación de los últimos 30 días. Marque en cada frase con la mayor sinceridad posible de forma que refleje su experiencia más frecuente. Siendo (1) Nunca, (2) Raramente, (3) A veces, (4) Muy a menudo y (5) Siempre.

| Nº | Ítem | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | Estoy feliz | | | | | |
| 2 | Estoy preocupado por una o más personas a las que he ayudado o ayudo | | | | | |
| 3 | Siento satisfacción por poder ayudar a la gente | | | | | |
| 4 | Siento que tengo vínculos con otras personas | | | | | |
| 5 | Me sobresaltan los sonidos inesperados | | | | | |
| 6 | Me siento fortalecido/a después de compartir con las personas a las que he ayudado | | | | | |
| 7 | Encuentro difícil separar mi vida personal de mi vida laboral | | | | | |
| 8 | La pérdida de sueño por las experiencias traumáticas de las personas a quienes he ayudado interfiere en mi vida diaria | | | | | |
| 9 | Creo que me afectan negativamente las experiencias traumáticas de aquellas personas a quienes ayudo o he ayudado | | | | | |
| 10 | Me siento atrapado/a por mis compromisos | | | | | |
| 11 | Debido a mis compromisos tengo la sensación de estar al límite en varias cosas | | | | | |
| 12 | Me gusta ayudar a la gente | | | | | |
| 13 | Me siento deprimido/a como resultado de mis compromisos | | | | | |
| 14 | Me siento como si fuera yo quien experimenta el trauma de alguien a quien he ayudado | | | | | |
| 15 | Tengo creencias (religiosas, espirituales u otras) que me apoyan | | | | | |
| 16 | Estoy satisfecho/a por cómo soy capaz de mantenerme actualizado/a (en mis conocimientos, habilidades, técnicas…) | | | | | |
| 17 | Soy la persona que siempre he querido ser | | | | | |
| 18 | Me dedico a cosas que me satisfacen | | | | | |
| 19 | Me siento agotado/a | | | | | |
| 20 | Tengo pensamientos de satisfacción acerca de las personas a quienes he ayudado y sobre cómo he podido ayudarles | | | | | |
| 21 | Me siento abrumado/a por la cantidad y tipo de actividades que tengo que afrontar | | | | | |
| 22 | Creo que puedo hacer cambiar las cosas | | | | | |
| 23 | Evito ciertas actividades o situaciones porque me recuerdan a las experiencias espantosas de la gente que he ayudado | | | | | |
| 24 | Planeo continuar con mi situación actual por muchos años | | | | | |
| 25 | Tengo pensamientos molestos, repentinos o indeseados | | | | | |
| 26 | Me siento “estancado/a” (sin saber qué hacer) por cómo funcionan las cosas | | | | | |
| 27 | Considero que soy un bueno/a en las actividades que realizo | | | | | |
| 28 | No puedo recordar determinados acontecimientos relacionadas con personas que experimentaron mucho trauma | | | | | |
| 29 | Soy una persona demasiado sensible | | | | | |
| 30 | Estoy feliz por mi situación actual | | | | | |

Se corrigen las tres dimensiones por separado:

- Satisfacción por compasión. Puntuaciones altas describirían buena satisfacción en el trabajo, en cambio, puntuaciones menores de 23 se asocian a problemas laborales o de otra índole.

- *Burnout*. Puntuaciones altas (mayor de 41) indicarían un ambiente sin apoyos y gran carga laboral, mientras que puntuaciones bajas (menor de 23) están asociadas a mayor efectividad laboral.
- Estrés secundario traumático.

Según señalan los autores puntuaciones asociadas al cuartil 3, son las mayores de 17 e indicarían la necesidad de preguntarse cómo se siente ante el trabajo o lo que este le rodea.

## Anexo VIII. SFCV-19S

Desde el momento que inició la propagación de la COVID-19, indique la frecuencia con que:

Las opciones de respuesta son: Nunca (0); Varios días (1); Más de la mitad de los días (2); Casi todos los días (3).

| Nº | Ítem | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 1 | Le preocupa que exista algún caso de contagio en su calle o su vecindario. | | | | |
| 2 | Le preocupa que exista algún caso de contagio en su familia o seres queridos. | | | | |
| 3 | Le preocupa que exista algún caso de contagio en su círculo de amistades o personas conocidas (incluyendo entorno académico/laboral). | | | | |

Corrección:

Cuanto más elevada sea la puntuación, más probabilidad existe de poseer miedo social a la COVID-19.