

Más allá de la suma unitaria: una propuesta de corrección alternativa para instrumentos psicométricos

Armando González-Sánchez 

Facultad de Psicología. Universidad Pontificia de Salamanca (España)

División de Tráfico y de la Seguridad del Consejo General de la Psicología (España)

Sociedad Española de Patología Dual (España)

INFORMACIÓN

Recibido: 03/07/2026

Aceptado: 13/07/2026

Palabras clave:

PHQ-9, Ítems,
evaluación
clínica,
psicometría

RESUMEN

Antecedentes: La medición en salud mental suele utilizar sumas unitarias (como las escalas Likert), asumiendo que todos los ítems tienen la misma equidistancia y peso explicativo. Esto puede ignorar que síntomas diferentes impacten de forma distinta en constructos complejos, limitando la precisión diagnóstica. **Método:** En este artículo de reflexión metodológica se proponen tres métodos alternativos de corrección ponderada utilizando el cuestionario de depresión PHQ-9 como modelo metodológico: a) Enfoque de Bartlett: basado en el análisis de componentes principales, asignando multiplicadores lineales; b) Modelo de Crédito Parcial (PMC): Basado en la Teoría de Respuesta al Ítem, utilizando parámetros no lineales y umbrales de categoría; c) Interpretación clínica: basada en los criterios diagnósticos categoriales del DSM-5. **Resultados:** El enfoque de Bartlett corrige el sesgo unitario, aunque su eficacia depende de la muestra de calibración. El enfoque PMC dota a la prueba de independencia muestral, acotando el techo psicométrico real del paciente. El modelo clínico aporta valor predictivo y activa alarmas clínicas necesarias. **Conclusiones:** Prescindir de la suma clásica de puntos por sistemas de ponderación puede mejorar la exactitud explicativa de las pruebas. Aunque requiera recalibración y decimales, las capacidades computacionales actuales pueden justificar este avance metodológico.

Beyond the unit score: a proposal for an alternative correction method for psychometric instruments.

ABSTRACT

Keywords,

PHQ-9, items,
clinical
assessment,
psychometrics

Background: Measurement in mental health often uses unit sums (such as Likert scales), assuming that all items have the same equidistance and explanatory weight. This may fail to account for the fact that different symptoms have varying impacts on complex constructs, thereby limiting diagnostic accuracy. **Method:** In this methodological reflection article, three alternative weighted correction methods are proposed, using the PHQ-9 depression questionnaire as a methodological model: a) Bartlett's approach: based on principal component analysis, assigning linear multipliers; b) Partial Credit Model (PMC): based on Item Response Theory, using non-linear parameters and category thresholds; c) Clinical interpretation: based on the categorical diagnostic criteria of the DSM-5. **Results:** The Bartlett approach corrects unitary bias, although its effectiveness depends on the calibration sample. The PMC approach provides the test with sample independence, limiting the patient's actual psychometric ceiling. The clinical model provides predictive value and triggers necessary clinical alerts. **Conclusion:** Moving away from the traditional sum of points in favour of weighting systems may improve the explanatory accuracy of the tests. Although it requires recalibration and the use of decimals, current computational capabilities may justify this methodological advance.

Citar como: González-Sánchez, A. (2026). Más allá de la suma unitaria: una propuesta de corrección alternativa para instrumentos psicométricos. *Prolepsis. Revista de Investigación en Psicología*, 21 (1), 8-19. <https://doi.org/10.67799/prolepsis.2026.13.1.eij>

Autor y e-mail de correspondencia: Armando González Sánchez agonzalezsa@upsa.es

Este artículo está publicado bajo Licencia Creative Commons 4.0 CC-BY-NC-ND

Uso de los instrumentos psicométricos

El objetivo del uso de instrumentos psicométricos en salud mental, ya sean entrevistas estructuradas, autoinformes, pruebas de ejecución, test manipulativos, escalas de evaluación u otros procedimientos de medición, es detectar, describir y cuantificar de la manera más precisa posible la presencia, intensidad o evolución de los fenómenos psicológicos y psicopatológicos que se pretenden evaluar. Sin embargo, la adecuada medición de constructos abstractos, como los que se pretenden medir en el desempeño de la práctica de la profesión relacionada con la salud mental, exige considerar una amplia variedad de factores que pueden influir en la calidad y precisión de los resultados obtenidos.

En primer lugar, deben garantizarse condiciones ambientales adecuadas durante la evaluación. Aspectos físicos como una iluminación apropiada, temperatura y ventilación agradables, ausencia de ruido, distracciones e interrupciones, así como la disponibilidad de un espacio cómodo y privado, ayudan a reducir fuentes de error ajenas al fenómeno que se desea medir. Asimismo, deben considerarse factores contextuales relacionados con la percepción de confidencialidad, seguridad y confianza por parte del evaluador, ya que estos elementos favorecen respuestas más sinceras y una mayor colaboración durante el proceso de evaluación.

Por otra parte, conviene valorar las condiciones propias de la persona evaluada. El estado físico puede afectar al rendimiento en una prueba, por lo que, comúnmente, resulta deseable que el sujeto se encuentre suficientemente descansado, adecuadamente alimentado y libre de los efectos de sustancias psicoactivas, medicamentos o condiciones médicas que puedan alterar sus capacidades cognitivas o emocionales. Del mismo modo, deben considerarse si el sujeto se encuentra en un estado emocional suficientemente bueno como para poder atender instrucciones; un estado cognitivo suficiente como para aplicar adecuadamente las pruebas, así como atender a la capacidad intelectual. Además, conviene prestar atención a variables motivacionales relacionadas con el interés por participar, el esfuerzo invertido en la tarea y la sinceridad de las respuestas. La presencia de simulación, disimulación o respuestas socialmente deseables pueden comprometer la validez de las conclusiones obtenidas.

El evaluador también constituye una posible fuente de error: unas instrucciones mal dadas, una sugerencia de que los errores penalizan, o incluso un gesto puede interpretarse por el evaluado, de una forma no contemplada, cambiando quizá inintencionalmente los resultados de la prueba. La correcta administración e interpretación de los instrumentos requiere formación especializada en evaluación psicológica y psicometría, así como el conocimiento de las características, alcances y limitaciones de las herramientas empleadas. Además, el profesional debe actuar con neutralidad, objetividad, rigor metodológico y siguiendo los principios éticos de la profesión, evitando influir directa o indirectamente en las respuestas del evaluado y garantizando el respeto a su dignidad, privacidad y autonomía mediante procedimientos como el consentimiento informado y la protección de la confidencialidad de los datos.

Respecto al instrumento de medición, este debe poseer propiedades psicométricas sólidas. Debe demostrar validez, es decir, medir efectivamente el constructo para el cual fue diseñado; confiabilidad o fiabilidad, produciendo resultados consistentes y estables en condiciones similares; sensibilidad suficiente para detectar diferencias entre individuos o cambios a lo largo del tiempo; y especificidad adecuada para discriminar correctamente entre quienes presentan o no la característica evaluada, minimizando errores de clasificación. Asimismo, debe contar con baremos actualizados y representativos que permitan interpretar adecuadamente los resultados en función de las características de la población de referencia.

No obstante, la calidad psicométrica de una prueba no depende únicamente de estas propiedades. También influye que el instrumento haya sido adaptado y validado para la población específica en la que se utilice. Las diferencias culturales, lingüísticas, geográficas, educativas y sociales pueden modificar la manera en que las personas comprenden los ítems y responden a ellos, incluso cuando comparten el mismo idioma (Roncero, 2015). Por ello, resulta necesario comprobar que las propiedades psicométricas se mantienen en diferentes contextos y grupos poblacionales, evitando sesgos asociados al sexo, la edad, el nivel educativo, la pertenencia étnica o el contexto sociocultural. En algunos casos, estas diferencias pueden requerir ajustes en las puntuaciones (González Sánchez, 2023), modificaciones en los criterios de interpretación, el desarrollo de versiones específicas para determinados grupos o incluso ponderaciones diferenciadas debido al Funcionamiento Diferencial de los Ítems (DIF) (González-Sánchez & Vicente Villardón, 2015).

Además, una herramienta de evaluación debe poseer validez ecológica y utilidad práctica, de modo que los resultados obtenidos reflejen de manera razonable el funcionamiento real de la persona en su entorno cotidiano y puedan emplearse eficazmente para la toma de decisiones clínicas, educativas o de investigación. Una solución a la validez ecológica se propone en el EMA (Evaluación Ecológica Momentánea), la cual consiste en una monitorización, a tiempos aleatorios, del rasgo (quizá con una alarma aleatoria para solicitar la cumplimentación de un test de autoinforme). Además, el instrumento debe ser factible de aplicar, corregir e interpretar en tiempos razonables por profesionales adecuadamente capacitados. Una prueba debe ser útil desde un punto de vista práctico y, si se puede, también clínico.

En la actualidad, el foco de atención para conocer la excelencia de un test se ha puesto en la calidad de sus ítems: cuanto más explicativo del rasgo sea, más asentado estará su puesto en la herramienta, pero ¿por qué no darle un peso mayor al ítem que más explica? ¿Por qué, en un test de depresión, el ítem de suicidio pesa lo mismo que el ítem de cansancio?

Actualmente la forma de evaluar rasgos psicológicos, en cuanto a las pruebas en las que el propio sujeto debe escoger entre unas opciones preestablecidas, está bastante extendida la técnica de medición de actitudes propuesta por Likert (Likert, 1930), que asume que, para un mismo ítem, puede establecerse un rango de niveles para que

evaluado elija entre ellas. Esta forma de trabajar toma el supuesto que el rango de cada segmento es equidistante, capturando eficientemente la variación en la intensidad del rasgo. Likert no ideó su escala de la nada, sino que parte del método Sigma, el cual se basa en la frecuencia relativa de respuestas de cada categoría, calculando sus puntuaciones Z y no cuestionando la equidistancia de los ítems que intentan explicar el constructo. Likert trabajó con el método Sigma y demostró que se podían obtenerse resultados casi idénticos al permitir que el evaluado estableciese por sí mismo esos rangos de cada ítem, lo que se entiende por equidistancia subjetiva. Debido a la simplicidad del método y que sus resultados son los mismos, actualmente se ha posicionado como un método ampliamente extendido en la evaluación en salud mental.

La equidistancia subjetiva de los ítems al comparar ambos métodos no se pone en duda, especialmente tras la revisión de la evidencia realizada por Likert. No obstante, actualmente los investigadores, clínicos y usuarios de las escalas basan el puntaje final en la suma de cada respuesta del evaluado; asumiendo un valor unitario por cada nivel del ítem e ignorando por completo cómo cada uno de esos ítems miden el rasgo. A menudo las escalas tipo Likert se denominan escalas sumativas. Otras alternativas a Likert utilizan la suma de puntos unitarios: como el diferencial semántico (Osgood et al., 1957), que a usa conceptos opuestos separados por segmentos numéricos equidistantes; también está la escala análoga-visual (Aitken, 1969), la cual para corregirla, a cada medida se le asocia un número natural, con independencia de que sus medias se sumen para cada rasgo; por otro lado, la Escala de Guttman (Guttman, 1944) al consistir en una ordenación jerárquica, el evaluador obtiene la puntuación del sujeto, a través de un punto marcado por la distancia del rasgo entre ítems de diferentes niveles de intensidad, pero sin establecer el peso de cada ítem. En cambio, la escala de intervalos aparentemente iguales de Thurstone (Thurstone, 1928) propone ponderación de los ítems mediante prueba de jueces, aunque estas ponderaciones tengan base en la opinión de expertos y no a través de haber estudiado la muestra de forma estadística.

Por su parte la Teoría de Respuesta al Ítem (Lord, 2012) no se centra en el valor diferenciado de cada ítem. Aunque estudia su índice de dificultad, igualmente parte del supuesto de que un ítem vale lo mismo que otro, ya que, aunque estudia los ítems por separado y evalúa su calidad, sigue usando la suma de puntos unitarios para calcular un puntaje, aunque el aporte del ítem a la explicación del factor sea diferente. Por otra parte, Masters (Masters, 1982) demostró la quimera de equidistancia entre niveles de un mismo ítem. Propuso el Modelo de Crédito Parcial (PCM por sus siglas en inglés), y para entenderlo conviene aclarar lo que entiende por umbral, que es la decisión subjetiva del sujeto de decidir entre cada uno de los niveles u opciones de respuesta de un ítem, separándolos en decisiones dicotómicas entre esa parte del nivel y la siguiente. De esta forma obtiene tantos umbrales como comparaciones por pares, lo que es lo mismo a opciones de respuesta menos uno. Estos umbrales se visualizan en la Curva Característica de las Categorías (CCC) evidenciando distancias diferentes entre cada una de las categorías de respuesta o niveles del ítem. Aunque el PCM transforma las respuestas ordinales en una escala de intervalo, realmente se centra en las diferencias entre opciones de respuesta en lugar de los pesos de los ítems.

Spearman (Spearman, 1927) defendía la idea de que no todos los ítems valen lo mismo, al teorizar la correlación entre el peso factorial y el factor latente teniendo en cuenta la proporción de saturación del factor en el ítem y de su variabilidad. La idea de la asignación de pesos factoriales a los ítems es básicamente la definición de Análisis Factorial más formalmente detallada por Bartlett (Bartlett, 1937).

También existen los criterios clínicos, los cuales implican la opinión de expertos debido a que deben fijarse en los aspectos que pueden o no encender las alarmas clínicas. Alternativamente y de forma más formal, los Modelos de Diagnóstico Cognitivo (CDM) implican analizar no el test en su conjunto, sino por cada uno de sus síntomas de forma aislada (Rupp et al., 2010). A través de los modelos de Orden Superior (de la Torre, 2011) se asume la existencia de un rasgo latente continuo de carácter general que rige la probabilidad de presentar los síntomas específicos. Posteriormente y bajo esta idea han surgido adaptaciones que han desarrollado metodologías para la derivación de índices continuos de gravedad o criticidad global (y no de forma aislada de cada uno de sus componentes) (Zhan & He, 2021).

El objetivo de este manuscrito es proponer formas alternativas de corrección de test psicométricos, utilizando el cuestionario de depresión PHQ-9 como modelo metodológico.

Para facilitar la explicación de la aplicación de diferentes formas de corrección ponderada por los pesos de los ítems, supongamos un test de depresión como lo es el PHQ-9, el cual es sólido y robusto en su medición del constructo (González Sánchez, 2023; González-Sánchez et al., 2023), cuyos ítems tienen opciones de respuesta que van de 0 a 3. Esta escala sirve para evaluar el Trastorno Depresivo Mayor (TDM) y básicamente convierte en ítems los nueve preceptos del DSM-5 para la detección del TDM. Siempre hay que tener en cuenta el diagnóstico diferencial, pero a efectos de la explicación psicométrica se obvia este hecho. De los nueve ítems que considera el test, intentando cubrir las características de la depresión, el ítem más saliente (el que tiene un índice de dificultad más bajo) es el de baja energía; este ítem se erige como el ítem que es más fácil superar, el que primero aparece cuando la depresión aflora: el ítem más sensible. En cambio, el ítem más difícil es el de suicidio. Existe controversia sobre si el suicidio debiera formar parte del constructo de la depresión o quedarse como una entidad aislada y por tanto un factor aparte. Sea como fuere, no creo que sea discutible la valoración clínica de estos dos ítems: a un clínico la información que le da el cansancio no le hace saltar las alarmas, pero, aunque tenga un valor de uno en el test debido a haber marcado el ítem de suicidio, sí que lo hace. Si este fuese el caso, el clínico debería iniciar protocolos específicos ante el suicidio.

Con este ejemplo quiero poner en evidencia que, tratándose de suma de puntos, en el supuesto de que tengamos dos sujetos y ambos hayan puntuado con uno sobre 27 en el test, no da la información que estamos buscando, ya que pueden dar esa puntuación tanto marcando uno en cansancio o uno en suicidio; aunque clínicamente no tengan el mismo peso estos ítems. La TRI o Guttman solventan este problema al dibujarse un patrón en las respuestas: no tiene sentido que un sujeto haya seleccionado uno en suicidio si no ha

seleccionado también otros menos extremos. Esta forma de corregir ignora la variabilidad humana y cristaliza el rasgo, ignorando los procesos personales y cómo los desarrolla el individuo. No siempre los ítems están ordenados jerárquicamente, especialmente en entidades con partes del rasgo que pueden aflorar paralelamente, y por tanto el superar un ítem no implica haber superado todos los anteriores. Aunque a la hora de la verdad en la suma de puntos no tomamos en cuenta este patrón de respuestas.

Hablando de la parte del suicidio en la depresión, tampoco se tendría en cuenta el suicidio dentro de la depresión si se considera que el suicidio se posiciona como una entidad a parte de la depresión, en este caso tendríamos dos factores y no uno. Este hecho podría allanar el terreno para considerar este ejemplo de test concreto en un exponente de modelo jerárquico. Sea como fuere puede ser entendido como una laguna o deficiencia considerar los tramos de los rasgos como unidades métricas si cada uno de los ítems aporta una explicación diferente al factor que desea medir. Dicho de otra forma (y dentro del contexto de considerar depresión y suicidio como una única entidad, unifactorial): no tiene sentido considerar ausencia de depresión si se ha marcado un uno en suicidio.

Por ello se propone una alternativa: dar pesos a cada ítem. Para llevar a cabo esta ponderación, una propuesta rápida y simple radica en el peso factorial que permite explicar el factor por parte del ítem. Aunque hay alternativas, es básicamente la definición de factor y análisis factorial: la explicación de la variabilidad del factor por parte del ítem.

En la actualidad, con las facilidades de cómputo, cobra especial sentido no corregir los test por suma unitaria de puntos. Defendiendo esta postura, sí que su simplicidad interpretativa facilita su uso, en cambio, el peso de cada ítem a cada factor ayuda a explicar de forma minuciosa su aporte a la explicación del rasgo medido.

Como dificultades se encuentra que es difícil encontrar un estudio de análisis factorial lo suficientemente robusto como para que exista una correspondencia única de cada muestra. Dicho de otra forma: la consistencia entre muestras varía necesariamente y consecuentemente el comportamiento del test ante ellas; pero cuanto más robusto sea el test, mejor detectará el rasgo en diferentes muestras.

La psicometría se basa en la mejora de los métodos de medición. Esta propuesta intenta hacer eso mismo: utilizar herramientas conocidas de una forma que antes no había tenido cabida. Algo simple: ponderar cada ítem con su explicación al factor y de ahí hacer una suma de puntos ponderada.

Una de las implicaciones de utilizar este método lleva a una recalibración de los test: una baremación usando la misma inspiración que está llevando hasta ahora, pero un ligero aumento de la complejidad de la interpretación, ya que, necesariamente habrá decimales en los puntajes finales.

Formas alternativas de corrección

A continuación, se mostrarán diferentes formas de corregir un mismo test. Tomando la aproximación de Bartlett, de PMC y de forma clínica. Los puntos de corte para la detección

del rasgo se mantendrían en los dos primeros, ya que, para establecer gravedad o curso, se toman en cuenta otros aspectos que no evalúa el PHQ-9. Este manuscrito se complementa con un archivo Excel que detalla los pasos aplicados en cada tipo de corrección (véase material suplementario).

Corrección mediante el enfoque de Bartlett

Parte del Análisis de Componentes Principales (ACP) y pondera el aporte de cada ítem a su factor mediante la matriz de coeficientes de puntuación factorial de Bartlett. Este enfoque asume que los saltos entre opciones de respuesta tienen el mismo recorrido: es lineal y continuo. Dicho de otra forma: los segmentos en los que se subdivide las opciones de respuesta de los ítems mantienen idéntica distancia. Pasar de la opción 0 a la 1 es la misma que de la 1 a la 2 e igual que de la 2 a la 3: son métricamente idénticos y constantes. El peso asignado al ítem actúa como un multiplicador fijo sobre la puntuación (posteriormente estandarizada) que el sujeto haya elegido.

Constituye una ponderación que corrige el sesgo de otorgar los mismos pesos a todos los ítems con independencia de su aporte real al constructo. No obstante, depende de la media y desviación típica de su grupo, por lo que la corrección queda supeditada al grupo de calibración, restringiendo su utilidad como un indicador absoluto de criterio.

A continuación, se detallan los pasos a seguir para operar mediante esta técnica:

- Paso 1 (Extracción de Datos de Partida): A partir de los resultados del ACP para los ítems se extrajeron tres estadísticos de la muestra: el coeficiente de puntuación de componente de Bartlett, la media y su desviación típica.
- Paso 2 (Tipificación de Extremos de Respuesta): Se calculan las puntuaciones estándar teóricas para las respuestas extremas de cada ítem. Se determinó la puntuación Z mínima (el escenario en el que el paciente responde cero) y la puntuación Z máxima (cuando el paciente responde siempre tres), aplicando la fórmula de tipificación tradicional basada en la media y la desviación típica del grupo de calibración: puntuación menos la media partido la desviación típica.
- Paso 3 (Cálculo de Límites Factoriales): Multiplicando el coeficiente de Bartlett por las puntuaciones Z mínimas y máximas de cada ítem, se obtienen las matrices de aportes límite individuales. El sumatorio vertical de estos aportes (uno por cada ítem) da como resultado el rango factorial del test completo (el suelo y el techo factorial del constructo).
- Paso 4 (Determinación de Pesos Finales y Ajuste de Escala): Para que los datos finales tengan puntuaciones interpretables (el máximo y mínimo del PHQ-9 son cero y 27), se calculó un multiplicador escalar que divide el rango objetivo del instrumento entre el rango factorial obtenido. Este multiplicador se aplica a la relación entre el coeficiente y la desviación típica de cada reactivo para obtener los Pesos Finales (V_i). Finalmente, y para que la puntuación mínima siempre sea cero, se calcula una constante de ajuste global que neutraliza el desplazamiento

provocado por las medias de la muestra.

Corrección mediante el enfoque del Modelo de Crédito Parcial

Parte de la TRI teniendo como base los parámetros de discriminación y umbrales de categoría, calculados mediante ecuaciones de estimación logística. Asume que las distancias entre opciones de respuesta no son lineales ni equidistantes, permitiendo dar un peso diferencial a cada ítem y a cada opción de respuesta. Supone un cambio con respecto al enfoque de Barttlet dado que concede libertad métrica absoluta: el peso del ítem solo está limitado por la puntuación máxima del test. Mediante el reescalamiento Min-Max Global, los pesos se transforman en valores acumulados, estrictamente crecientes y positivos, blindando al sistema frente a distorsiones y acotando el resultado final en el rango universal de los extremos de la puntuación del test.

No depende de la estandarización de su muestra de referencia, por lo que el rasgo puede medirse con independencia del nivel de rasgo de la muestra de la que se obtuvo, pudieron establecerse puntos de corte diferenciales y hacer comparaciones más precisas. Al liberar al test de las restricciones de la muestra, la puntuación obtenida refleja la gravedad intrínseca de la sintomatología del paciente.

Pasos para el cálculo mediante la corrección del PMC:

- Paso 1 (extracción de parámetros logísticos): Primero se realiza una calibración no lineal del PHQ-9 mediante ecuaciones de estimación politómica (por ejemplo, mediante *lavaan* usando R (R Core Team, 2026) (véase anexo, Tabla I)). De este análisis se obtienen los parámetros de discriminación ($[a]_{ij}$), que miden la potencia de cada pregunta para distinguir la gravedad del rasgo, y los tres umbrales de categoría (conocidos en inglés como *thresholds*) (b_{ij}), que marcan el nivel de depresión necesarios para pasar de un nivel de respuesta al siguiente.
- Paso 2 (modelado de pesos acumulativos por nivel): A partir de la ecuación del PMC, se pasan de los umbrales lógicos aislados a umbrales de tipo acumulativo por opción de respuesta. Para evitar puntuaciones latentes negativas provocadas por el centrado poblacional del programa, se ejecuta una transformación lineal que desplaza el origen de la escala. Esto genera una matriz de pesos independientes, estrictamente crecientes y positivos, explicitando cada nivel del ítem en el test.
- Paso 3 (establecimiento del techo psicométrico): Se suman los pesos acumulados correspondientes al nivel máximo (respuesta tres en el test de todos ítems). Este sumatorio fija el límite superior del rasgo o "Techo Latente Absoluto" de la sintomatología del paciente.
- Paso 4 (reescalado proporcional directo): En lugar de utilizar constantes de sustracción poblacionales, se evalúa la respuesta del paciente, se extrae el peso positivo condicional que le corresponde a su nivel y se realiza un sumatorio directo. Esta puntuación se divide por el Techo Latente Absoluto y se multiplica

por la constante 27 (que es la puntuación máxima teórica del test). Al ser una proporción pura del rasgo, se dota de independencia de la muestra y se acotan las puntuaciones máximas y mínimas.

Corrección mediante interpretación clínica

Concretamente el PHQ-9 está basado en el DSM-IV y este no ha cambiado en su definición de depresión en el DSM5. Este tipo de corrección no implica sustituir los cálculos anteriores, sino darle un significado explicativo y predictivo; así como dotar de alertas clínicas.

Siguiendo las indicaciones del DSM5 en cuanto al TDM, su corrección como presencia o ausencia se marca en el mismo: cinco o más de los nueve ítems, en entre los que, al menos, deben estar presentes el de estado de ánimo y el de pérdida de interés o placer. Además, cobra relevancia el décimo e ignorado ítem del PHQ-9, que se refiere del malestar clínicamente significativo en las áreas social, laboral u otras áreas importantes. Desgraciadamente se marcan unos criterios diferenciales que implicarían evaluaciones diferentes: las utilizadas para descartar otros trastornos clínicos con los que pueden confundirse, respuestas normales del sujeto o acontecimientos inducidos o transitorios. Igualmente, para evaluar la gravedad o el curso del TDM se requiere de otras herramientas.

Para la corrección del test mediante criterios del DSM5 se tuvieron en cuenta sus condicionantes: al menos superar en un punto los ítems uno o dos (cualquiera de los dos), haber superado al menos cinco de los ítems (con puntuaciones mayores a cero) y haber superado el ítem de malestar significativo en las áreas importantes de la vida.

Conflicto de Interés

Los autores declaran no tener ningún conflicto de interés.

Descargo de responsabilidad

Se utilizó el programa Ollama con su modelo de IA nomotron-3-super:cloud como apoyo en limpieza y ejecución en el programa estadístico R.

Fuentes de Financiación

El autor declara que no recibió financiación para la realización de este estudio

Referencias

- Aitken, R. C. B. (1969). A Growing Edge of Measurement of Feelings. *Proceedings of the Royal Society of Medicine*, 62(10), 989–993. <https://doi.org/10.1177/003591576906201005>
- Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology. General Section*, 28(1), 97–104. <https://doi.org/10.1111/j.2044-8295.1937.tb00863.x>

- González-Sánchez, A. (2023). *Mejoras psicométricas en la evaluación de la salud pública: depresión en Costa Rica* [Universidad de Salamanca]. <https://doi.org/10.14201/gredos.158237>
- 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>
- González-Sánchez, A., & Vicente Villardón, J. L. (2015). *Detección del funcionamiento diferencial de los ítems en el test MMSE en la población española* [Universidad de Salamanca]. <http://hdl.handle.net/10366/128281>
- Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, 9(2), 139. <https://doi.org/10.2307/2086306>
- Likert, R. (1930). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1–55.
- Lord, F. M. (2012). *Applications of Item Response Theory To Practical Testing Problems*. Routledge. <https://doi.org/10.4324/9780203056615>
- Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.
- R Core Team. (2026). *R: A language and environment for statistical computing* (4.6.1). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Roncero, C. (2015). La validación de instrumentos psicométricos: un asunto capital en la salud mental. *Salud Mental*, 38(4), 235–236. <https://doi.org/10.17711/SM.0185-3325.2015.032>
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. Macmillan.

Anexos

Tabla 1

Relación de comandos R, sus usos y sus librerías.

Comando/Función	Uso Específico	Librería	Notas de Compatibilidad
library(mirt)	Intento inicial para modelo PCM (falló por errores de versión en itemtype y extracción de parámetros)	mirt	Problemas: Unknown itemtype: pcm, errores en coef(..., IRTpars=TRUE)
library(eRm)	Intento intermedio (falló por ausencia de thresh() y estructura inesperada de \$betas)	eRm	Problemas: Error en thresh(pcm_model): no se pudo encontrar la función "thresh"
library(lavaan)	Solución final exitosa	lavaan	Funcionó con cfa() + estimator="WLSMV" + parameterEstimates()
as.data.frame(lapply(..., factor(..., levels=0:3, ordered=TRUE)))	Conversión de variables numéricas a factores ordinales (requisito para WLSMV en lavaan)	base R	Esencial para manejar correctamente datos de escala Likert (0-3)
cfa(modelo, data=phq_ordinal, ordered=..., estimator="WLSMV")	Estimación de modelo factorial confirmatorio unidimensional con estimador robusto para datos ordinales	lavaan	WLSMV es el oro estándar para variables ordinales como el PHQ-9
parameterEstimates(fit)	Extracción universal y version-independent de todos los parámetros del model	lavaan	Solucionó errores de lavInspect() (ej: "lambda" desconocido); funciona en todas las versiones de lavaan
pe[pe\$op == "=~", ...]	Filtra cargas factoriales (λ) de parameterEstimates()	lavaan	op == "=~" = relaciones latente ← ítem (análogas a discriminación)
`pe[pe\$op == " ", ...]`	Filtra umbrales (τ) de parameterEstimates()	lavaan	op == " " = umbrales de categoría (ítem \ nivel)
fitMeasures(fit, c(...))	Extracción robusta de índices de ajuste (evita problemas de slots en objetos lavaan)	lavaan	Alternativa segura a fit@test[[1]]\$stat etc. (que falló en su versión)
reliability(fit)\$omega.total	Cálculo de confiabilidad omega (equivalente moderno a alfa de Cronbach)	lavaan	Más adecuado para modelos factoriales que alfa de Cronbach tradicional