MEDWAVE
Revista Médica Revisada por Pares

REVISIONES

# Validación de cuestionarios para la medición de variables en salud: conceptos fundamentales

Natalia Riva[a], Diego Grandi[a], Benjamín Cruzat[a], Ruben Alvarado[b*]

*[a]Estudiante de Medicina, Escuela de Medicina, Facultad de Medicina, Universidad de Valparaíso. Valparaíso, Chile; [b]Departamento de Salud Pública, Escuela de Medicina, Facultad de Medicina, Universidad de Valparaíso. Valparaíso, Chile*

**RESUMEN**

Dentro de la práctica clínica, así como en la salud poblacional, es habitual utilizar cuestionarios que permiten evaluar condiciones o variables que no son directamente observables. No obstante, la construcción y validación de estos instrumentos o cuestionarios suele ser poco conocida. El objetivo de esta revisión narrativa es sintetizar de manera general el proceso de construcción y validación de estos cuestionarios, para así tener una mejor comprensión de este proceso, de los aspectos que se evalúan y de la mejor forma de utilizarlos. La validación de cuestionarios corresponde a un proceso de análisis de este, cuya finalidad es medir una variable latente o constructo, así como sus dimensiones, las que no pueden ser observadas directamente. Una variable latente puede ser inferida a través de un conjunto de atributos específicos que forman parte de ella, como los ítems de un cuestionario y que sí son observables. En este artículo se abordan de manera teórica los conceptos fundamentales de validación de cuestionarios o test, variables latentes o constructos, estudio de la confiabilidad y de la validez, así como los factores que afectan a estas dos últimas características, a través de una revisión narrativa. En el texto, se presentan ejemplos sobre estos conceptos.

**KEYWORDS** Reliability and Validity, Surveys and Questionnaire, Epidemiology

## INTRODUCCIÓN

La validación de cuestionarios corresponde a un proceso de análisis de este, cuya finalidad es medir un constructo o variable latente, así como posibles dimensiones de esta, que no pueden ser observadas directamente. El cuestionario está compuesto por varias preguntas o ítems, a través de los cuales se busca hacer la medición. La respuesta a los ítems es la forma en que medimos una variable que no puede ser observada directamente. Como señalan Carmines y Zeller, es "el proceso de vincular conceptos abstractos con indicadores empíricos, el cual se realiza mediante un plan explícito y organizado para calificar los datos disponibles, en términos del concepto que el investigador tenga en mente" [1].

Un aspecto importante y distintivo del campo de la psicometría, es la forma en que se entiende el concepto de medición. En este caso se trata de una acción asociativa (de un valor numérico con una respuesta determinada), y no sólo un acto de asignar valores numéricos a elementos específicos.

Esta revisión de la literatura es la treceava entrega de una serie metodológica de revisiones narrativas acerca de tópicos generales en bioestadística y epidemiología clínica, las que exploran y resumen en un lenguaje amigable, artículos publicados disponibles en las principales bases de datos y textos de consulta especializados. La serie está orientada a la formación de estudiantes de pre y posgrado. Es realizada por la Cátedra de Medicina Basada en Evidencia de la Escuela de Medicina de la Universidad de Valparaíso, Chile, en colaboración con el Departamento de Investigación del Instituto Universitario del Hospital Italiano de Buenos Aires, Argentina, y el Centro Evidencia UC de la Pontificia Universidad Católica de Chile. El objetivo de este manuscrito es describir el proceso realizado para la validación de instrumentos y los elementos que lo conforman.

Ahora bien, en cuanto a la definición de un constructo o variable latente [2] se refiere a una condición que no es directamente observable, pero que puede ser inferida a través de un conjunto de atributos específicos que forman parte de ella, los cuales sí son observables. Ejemplos de constructos son la autoestima, el apoyo social, la satisfacción con los servicios entregados o la calidad de vida. Dentro de la medicina y la salud pública, hay muchas condiciones que corresponden a variables latentes o constructos que requieren ser investigados.

* **Autor de correspondencia** ruben.alvarado@uv.cl
**Citación** Riva N, Grandi D, Cruzat B, Alvarado R. Validación de cuestionarios para la medición de variables en salud: conceptos fundamentales. Medwave 2024;24(01):e2746
**DOI** 10.5867/medwave.2024.01.2746
**Fecha de envío** Jul 11, 2023, **Fecha de aceptación** Dec 19, 2023, **Fecha de publicación** Jan 23, 2024
**Correspondencia a** Departamento de Salud Pública, Escuela de Medicina, Facultad de Medicina, Universidad de Valparaíso Av. Angamos 655, Viña del Mar, Región de Valparaíso, Chile

**IDEAS CLAVE**

- Las variables latentes no pueden ser observadas directamente y deben ser medidas a través de cuestionarios que demuestren tener una buena validez y confiabilidad.
- La confiabilidad de un cuestionario se refiere al grado con que realiza la medición sin errores.
- La validez de un cuestionario corresponde al grado en que mide la variable latente, y se suele evidenciar en tres aspectos diferentes: la concordancia conceptual de su contenido, la concordancia contra un criterio externo y la concordancia con el constructo teórico.

En consecuencia, podemos medir un constructo mediante las respuestas que dan las personas a los ítems de un cuestionario.

Por ejemplo, en la evaluación del desarrollo psicomotor de un lactante podemos considerar como atributo no observable las habilidades espaciales de este, y como atributo observable a la acción de armar torres de bloques durante una evaluación de desarrollo psicomotor, identificando así las habilidades visoespaciales según la altura de bloques que consiga armar.

## ASPECTOS METODOLÓGICOS DE VALIDACIÓN DE INSTRUMENTOS DE MEDICIÓN

### Requisitos de una buena medición

Para poder realizar una buena medición habrá que seleccionar el cuestionario o test adecuado, el cual debe representar de manera más fidedigna la variable que está siendo estudiada. Por ello este tendrá que cumplir con tres criterios:

*Confiabilidad*

El grado en que el cuestionario produce resultados consistentes en sus diferentes mediciones, tanto por sus diversos ítems, por encuestadores distintos o por su aplicación en diferentes momentos en el tiempo.

*Validez*

El grado en que el cuestionario mide la variable que pretendemos evaluar.

*Objetividad*

Se refiere al grado en que el cuestionario permite valorar las características de un objeto de la realidad tal cual es, evitando lo más posible los aspectos subjetivos de quien lo administra, clasifica e interpreta, para así reducir los sesgos que esto puede conllevar.

## CONFIABILIDAD O FIABILIDAD: CONCEPTO Y DEFINICIÓN

El concepto de confiabilidad, consistencia o fiabilidad de un test está relacionado con los errores de medida aleatorios presentes en las puntuaciones obtenidas a partir de su aplicación . En otras palabras, será la capacidad de realizar una medición libre de errores [3].

Para esto debemos introducir el concepto de consistencia interna [4], siendo el grado en que cada una de las partes de las que se compone el instrumento es equivalente al resto [2]. Por lo tanto, dentro de este concepto tendremos tres dimensiones dentro de la fiabilidad que pueden ser estudiadas en un test o cuestionario:

1. Consistencia de su conjunto de ítems.
2. Consistencia a través del tiempo, estabilidad o mediación.
3. Consistencia en aplicación por diferentes encuestadores.

Hay dos elementos que debemos considerar al evaluar la confiabilidad. En primer lugar, la selección y uso del mejor cuestionario o test para el tópico que estamos estudiando. En segundo lugar, el nivel de error que pudiera tener la medición dada por dicho cuestionario. Así, nuestro objetivo será la construcción y/o uso de test que nos permita reducir significativamente el error, teniendo en cuenta que al realizar una medición estaremos siempre en presencia de variaciones dadas por el azar, las cuales no pueden eliminarse.

### Confiabilidad: fiabilidad intra-test o consistencia interna

Tras la definición mencionada anteriormente del concepto, este se puede calcular mediante diversas pruebas estadísticas, dependiendo de las características de las respuestas a sus ítems. Una de las pruebas más usadas es el α de Cronbach, que expresa el grado de co-varianza de los ítems dentro de un test o en parte de este. Es la prueba adecuada para los test cuyos ítems tienen varias alternativas de respuesta.

La fórmula de α de Cronbach [5] se presenta en la Figura 1.

Dado que se trata de una covariación, la mejor consistencia interna se obtiene con valores que se acerquen al 1. En general, los rangos aceptados son los siguientes:

1. Valores menores a 0,5 puntos son inaceptables.
2. Valores entre 0,5 y 0,6 puntos tienen una consistencia pobre.
3. Valores entre 0,6 y 0,7 puntos tienen una consistencia cuestionable.
4. Valores entre 0,7 y 0,8 puntos tienen una consistencia aceptable.
5. Valores entre 0,8 y 0,9 puntos tienen una buena consistencia.
6. Valores mayores a 0,9 puntos tienen una consistencia excelente [6].

**Figura 1.** Fórmula de α de Cronbach para determinar fiabilidad intra-test.

$$\alpha = \frac{K}{K-1}\left(\frac{\sum_{i=1}^{K}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

*α:* Alfa de Cronbach.
$\sigma_{Y_i}^2$ : varianza del ítem *i*.
$\sigma_x^2$ : varianza de los valores totales observados.
$K$: número de preguntas o ítems considerados.
Fuente: preparado por los autores a partir de fuente [3].

Específicamente, para las evaluaciones psicológicas se tiene un consenso sobre los valores entre 0,65 y 0,8; los cuales se consideran como adecuados [7].

Por ejemplo, en la validación del instrumento de Warwick-Edimburgo para bienestar mental en población chilena, realizado por Carvajal *et al.* [8], se utilizó el α de Cronbach para evaluar la consistencia interna de los 14 ítems, con un valor de 0,875 (que es bueno) y sin la necesidad de eliminar algún ítem para mejorar este indicador.

### Confiabilidad: fiabilidad test-retest

Este método se basa en la existencia de una estabilidad temporal del constructo que está siendo evaluado a través del cuestionario. Si se realizan dos mediciones separadas por un tiempo prudente, los resultados en ambos test no deberían variar de forma significativa. Nos permite utilizar el mismo test en ambas ocasiones [3], dándonos la ventaja de no tener que construir un cuestionario diferente para medir el mismo constructo [2].

Se puede calcular utilizando algún coeficiente de correlación entre la primera y segunda medición (por ejemplo, r de Pearson o Rho de Spearman) [9], donde un valor cercano a 1 indicaría que existe una correlación positiva entre los resultados de ambas aplicaciones. Es decir, mide con precisión y un valor cercano a 0 señalaría que no hay correlación entre estas dos aplicaciones en el tiempo.

Uno de los factores importantes a considerar es el lapso de tiempo que transcurre entre la primera y la segunda aplicación del test, ya que este debe ser concordante con un tiempo en el que se espera que el constructo no varíe o no lo haga de manera importante. Otra condición relevante es que no exista una alteración en el número de los sujetos evaluados, ya que pudiera introducirse un sesgo de selección, disminuyendo la confiabilidad de la medición.

En el mismo ejemplo previo de Carvajal *et a*l. [8], la evaluación test-retest se hizo con dos semanas de diferencia y en una sub-muestra de 50 personas (22,7% de la muestra total), con un buen valor para la correlación de Rho de Spearman (0,556) y p > 0,001).

### Confiabilidad: fiabilidad inter-jueces o κ de Cohen

Dentro del contexto clínico se da el caso de que múltiples entidades evalúan al mismo sujeto, por ello debe existir un nivel de concordancia en las clasificaciones a partir de dos o más administradores del test [2].

Al igual que en los otros análisis de fiabilidad que describimos previamente, existen varias pruebas que pueden ser utilizadas, lo que depende del tipo de respuesta a los ítems. Una de las pruebas más usadas es el coeficiente kappa (κ), que es útil frente a ítems con escalas nominales. Por ejemplo tenemos dos médicos, uno especialista y otro no, que tendrán que determinar si un grupo de pacientes posee fracción de eyección cardiaca preservada o reducida (clasificación nominal). Como este examen dependerá de la experiencia del realizador, podemos tener una diferencia con los resultados obtenidos, por lo que para medir la confiabilidad de esta evaluación se puede aplicar el coeficiente de κ. Cuando se está frente a ítems con escalas ordinales o de intervalo, se puede utilizar el coeficiente κ modificado o el coeficiente de correlación intraclase. Utilizando el mismo ejemplo anterior, si quisiéramos evaluar qué porcentaje de fracción de eyección reducida (clasificación ordinal) tendrían los pacientes evaluados, también podríamos tener diferencias con los evaluadores. En este caso, sería sensato evaluarlo mediante el coeficiente de κ modificado.

A modo de ejemplo, la fórmula para calcular el coeficiente κ de Cohen [10] se puede evidenciar en la Figura 2.

Los valores cercanos a 1 indicarán que existe una elevada consistencia entre los resultados obtenidos por diferentes encuestadores. En cambio, los valores cercanos a 0 indican que hay una baja consistencia entre estos. Dicho de otra forma, significa que los resultados obtenidos por diferentes encuestadores están más determinados por otros factores que por el cuestionario en sí [10].

## VALIDEZ: CONCEPTO Y DEFINICIÓN

El grado de validez de un cuestionario o test es una de sus características fundamentales. Como ya se mencionó anteriormente, la validez hace referencia a que el test "mida lo que se desea medir", es decir que evalúe esa variable y no algo distinto [11,12]. De esta manera, un cuestionario que pretenda medir el nivel de liderazgo mida aquello y no el nivel de autonomía, por ejemplo.

Para establecer que se cuenta con un nivel adecuado de validez de un test o cuestionario, se busca tener evidencia de tres aspectos de esta validez: la validez de contenido, la validez de criterio y la validez de constructo [10,13].

### Validez: en relación al contenido

Una primera dimensión de la validez corresponde al contenido. La validez de contenido es evaluada por el conjunto de ítems que componen el cuestionario. Se espera que estos sean una muestra representativa del constructo o variable latente que está siendo estudiada. Es decir, que el conjunto de ítems está incluyendo todos los aspectos que involucra el

**Figura 2.** Cálculo de fiabilidad inter-jueces en escalas nominales a través de κ de Cohen.

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

*κ:* Kappa de Cohen.
$P_0$ : proporción de individuos clasificados de manera consistente.
$P_e$: proporción de individuos clasificados por efecto del azar.
Fuente: preparado por los autores a partir de fuente [4].

concepto de ese constructo. Y, por otra parte, que no existan ítems que estén evaluando aspectos que no estén incluidos en ese concepto.

Si deseamos medir conocimientos respecto de un tema, debemos considerar incluir todos los aspectos clave de lo medido [14]. Por ejemplo, al medir la presencia de un trastorno depresivo, un cuestionario que pregunte solo por características del estado afectivo (desánimo, apatía, etc.), pero no por los síntomas cognitivos (pesimismo, baja autoestima, etc.) o físicos (cambios en el apetito o el sueño), no tendría suficiente validez de contenido. Por otra parte, si incorpora síntomas que no son propios de los trastornos depresivos, pero que con frecuencia están asociados (como es el abuso de alcohol o las crisis de pánico), también se reduciría su validez de contenido.

Un cuestionario con buena validez de contenido debe medir adecuadamente todas y cada una de las principales dimensiones que forman parte de la variable en estudio [15].

**Validez: en relación al criterio**

La validez de criterio se obtiene cuando comparamos los resultados del cuestionario que estamos probando con los de otro test que mide lo mismo o un constructo afín, y que ya ha demostrado previamente tener una buena validez [16].

Si el criterio con el que comparamos nuestro test se puede aplicar y evaluar en el presente, hablamos de "validez concurrente".

Por ejemplo, Silva *et al*. (2012) adaptaron y validaron la herramienta *Eating Disorders Diagnostic Scale* (EDDS) [17] al español, un cuestionario diseñado para el diagnóstico de trastornos de la conducta alimentaria. Seleccionando a un grupo de pacientes y a un grupo control, aplicó la EDDS y la entrevista psiquiátrica estructurada (CIDI, por su sigla en inglés *Composite International Diagnostic Interview*), que permite evaluar la presencia de desórdenes mentales en la persona encuestada, de acuerdo a las clasificaciones del Manual de Diagnóstico y Estadísticas para Enfermedades Psiquiátricas, IV edición (DSM-IV, *Diagnostic and Statistical Manual of Mental Disorders*) y la Clasificación Internacional de Enfermedades, 10° edición (CIE-10). Usando el CIDI como estándar de oro, crearon tablas de contingencia y encontraron una correlación desde moderada a alta entre los resultados de ambas pruebas para diagnosticar la presencia de un trastorno de la conducta alimentaria.

Ahora, cuando el criterio con el que comparamos nuestro test se aplica en el futuro, estamos hablando de "validez predictiva". La validez predictiva se evalúa con pruebas estadísticas que dependen del tipo de variables con que se está trabajando. Podría ser una prueba t cuando se comparan promedios [18], una correlación de Pearson o Spearman cuando se trata de variables continuas [19], o alguna prueba de concordancia cuando se trata de dos variables dicotómicas [19,20].

Un ejemplo para este tipo de validez es el estudio de Larzelere *et al*. (1996), donde se evaluó la validez predictiva del cuestionario *Suicide Probability Scale* [21]. En este trabajo se analizaron los intentos de suicidio, verbalización suicida y comportamientos autolesivos que se realizaban posterior a la aplicación de este cuestionario. la cual se efectuaba al momento de la admisión en un centro de acogida de menores en riesgo.

**Validez: en relación al constructo**

La validez de constructo se refiere a qué tan bien un test mide y representa el concepto teórico (constructo o variable latente) que está siendo evaluado. Se trata de contrastar la teoría con la evidencia empírica que se obtiene al aplicar el test a un conjunto de sujetos. Por ejemplo, si la teoría señala que el constructo en estudio tiene dos dimensiones y el cuestionario utilizado tiene ítems que miden estas dos dimensiones, se esperaría que exista una alta correlación entre los ítems dentro de cada dimensión, y al mismo tiempo que haya una baja correlación entre los ítems que miden las dos dimensiones diferentes [22].

Para esto, la prueba más utilizada es el análisis factorial [23]. Esta prueba busca identificar patrones de asociación entre variables o ítems, de forma de reunir en “factores” aquellos que se encuentran más correlacionados entre sí. De esta forma, un cuestionario muestra una buena validez de constructo cuando en el análisis factorial los ítems que conforman una dimensión teórica se reúnen dentro de un mismo factor.

Por ejemplo, Alvarado *et al*. (2015), evaluaron el cuestionario de Edimburgo [24] para identificar un trastorno depresivo durante el embarazo. En el análisis factorial exploratorio encontraron que los 12 ítems que forman este test se reunían en un sólo factor, tal como se esperaba teóricamente.

## FACTORES QUE AFECTAN LA CONFIABILIDAD Y LA VALIDEZ

Tal como se mencionó anteriormente, el "error aleatorio" y el "sesgo de medición" son elementos en estrecha relación con la confiabilidad y validez [25]. Por su parte, el sesgo corresponde a una tendencia sistemática a subestimar o sobrestimar el estimador de interés [25], originado por una deficiencia en el diseño o en la ejecución de un estudio [26], afectando en forma negativa la validez del cuestionario. Por otra parte, el error aleatorio corresponde a las variaciones explicadas por el azar [27] y no puede eliminarse (aunque sí minimizarse), y

**Figura 3.** Infografía de validación de instrumentos de medición de variables latentes o constructos.


Fuente: diseñado por los autores.

afecta negativamente la fiabilidad del test. Existen tres factores principales [28] asociados al grado de error aleatorio y cómo este pudiera afectar en nuestra medición:

1. Grado de variabilidad individual e interindividual.
2. El tamaño de la muestra.
3. La magnitud de las diferencias encontradas.

Teniendo en cuenta lo anterior, se pueden tomar medidas para disminuir el error aleatorio, aumentando el tamaño de la muestra o estableciendo *targets* de medición con diferencias de magnitud mayores. Un ejemplo de error aleatorio en las mediciones se puede ver en la estimación del peso fetal a través de ecografía por método biométrico, en donde se han

reportado diferencias de ±30 gramos respecto al peso del feto [29].

También existen otros factores que afectan la validez y confiabilidad final de un cuestionario y que se mencionan a continuación, organizados según la etapa en que pueden aparecer dentro del proceso de construcción y validación del test:

*Construcción del cuestionario*

En esta etapa se debe intentar que el test sea construido de manera idónea para lo que buscamos medir. La **improvisación** [30], escasa búsqueda de información en el tema de interés y la poca experiencia en la creación de cuestionarios de recolección propicia una construcción deficiente, pudiendo aumentar el error aleatorio y sistemático. Otro punto importante a mencionar es el **uso de cuestionarios validados en el extranjero** [30], los cuales sin importar que estén traducidos y adaptados al idioma local [31], **no se encuentren contextualizados** para la investigación en nuestro grupo de interés. Por otra parte, el proceso de **estandarización** permite disminuir el sesgo de medición del estudio, al utilizar en todos los participantes los mismos cuestionarios y maneras de medir la variable de interés. Finalmente, se deben mencionar **aspectos mecánicos de la construcción** [30] del test. Por ejemplo, que las instrucciones no sean claras o se usen muchos tecnicismos que impidan comprenderlas.

*Aplicación del cuestionario*

En este punto se debe tener en cuenta que el **test debe ser adecuado** [30] para la persona a quien se le está aplicando (por ejemplo, el uso de materiales de lectura en sujetos que no sepan leer, o no considerar diferencias importantes basadas en el género). Además, las **condiciones de aplicación** deben ser las adecuadas, tanto a nivel contextual como propios de la persona (por ejemplo, propiciando un espacio tranquilo que permita la concentración, sin presiones de por medio). Otro elemento a destacar en este sentido es el **estilo personal propio de los participantes** [30,32], direccionando las respuestas hacia aquello socialmente aceptable y omitiendo aquello indeseable. Esto se conoce como sesgo de aceptabilidad social o sesgo por complacencia social. Por ejemplo, una persona responde que dentro de las comidas de su hijo no se encuentra habitualmente el consumo de bebidas azucaradas, tendiendo a responder lo que se supone como deseable a nivel social.

## CONSIDERACIONES FINALES

La validación de cuestionarios es un proceso clave en la medición de variables latentes o sus dimensiones, las cuales no pueden ser observadas directamente. Para garantizar que una medición sea lo más precisa posible, es necesario que los cuestionarios posean una alta confiabilidad y una buena validez, ambas características están relacionadas con errores y sesgos de medición. La confiabilidad hace referencia al grado en que un cuestionario genera resultados precisos y consistentes, mientras que la validez se relaciona con el grado en que el test mide la variable que pretende evaluar.

La confiabilidad puede ser evaluada mediante pruebas de consistencia interna, test-retest y de fiabilidad inter-jueces. Por otro lado, la validez se compone de tres dimensiones: contenido, criterio y constructo. La validez de contenido hace referencia a la representatividad y pertinencia de los ítems del cuestionario, mientras que la validez de criterio implica comparar los resultados del cuestionario con otros o con criterios ya validados. Por último, la validez de constructo se relaciona con la evidencia de que el cuestionario mide el constructo teórico esperado.

Finalmente, durante la investigación de variables latentes se debe tener en cuenta la existencia de factores que puedan afectar la validez y confiabilidad durante la construcción, elección y aplicación de los cuestionarios. Al seguir estos principios metodológicos, se asegura una mayor precisión de los resultados, pudiendo realizar análisis y conclusiones con bajos índices de error (Figura 3).

Un buen ejemplo de la utilidad de los cuestionarios es la incorporación de este tipo de instrumentos para mejorar la detección de cuadros clínicos que habitualmente pasan desapercibidos, como sucede con los cuadros depresivos en el embarazo. Alvarado *et al* [24] validaron un cuestionario que mostró tener buenos indicadores de validez y fiabilidad para la detección de este problema, por lo cual actualmente se incorpora como una herramienta en los controles del embarazo.

**Autoría** NR: conceptualización, redacción del manuscrito original, revisión y edición, visualización. DG: conceptualización, redacción del manuscrito original, revisión y edición, visualización. BC: conceptualización, redacción del manuscrito original, revisión y edición, visualización. RA: conceptualización, metodología, redacción del manuscrito original, revisión y edición, visualización, supervisión.

**Conflictos de intereses** Los autores completaron la declaración de conflicto de intereses del ICMJE y declararon que no recibieron fondos para completar este artículo; no tienen relaciones financieras con organizaciones que puedan tener interés en el artículo publicado en los últimos tres años; y no tienen otras relaciones o actividades que puedan influir en la publicación del artículo.

**Financiamiento** Los autores declaran que no contaron con fuentes de financiamiento externo para la publicación de este artículo.

**Idioma del envío** Español.

**Origen y revisión por pares** Con revisión externa por cuatro pares revisores, a doble ciego.

## REFERENCIAS

1. Carmines EG, Zeller RA. In: Reliability and Validity Assessment [Internet]. SAGE Publications; 1979. https://play.google.com/store/books/details?id=o5x1AwAAQBAJ
2. Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. Psychol Rev. 2003;110: 203–219. http://dx.doi.org/10.1037/0033-295X.110.2.203 https://doi.org/10.1037/0033-295X.110.2.203

3. Meneses J, Barrios M, Lozano LM, Bonillo A, Turbany J, Cosculluela A, et al. Psicometría. Editorial UOC. 2014. https://play.google.com/store/books/details?id=2JxuBAAAQBAJ
4. Manterola C, Grande L, Otzen T, García N, Salazar P, Quiroz G. Reliability, precision or reproducibility of the measurements. Methods of assessment, utility and applications in clinical practice. Rev Chilena Infectol. 2018;35: 680–8. http://dx.doi.org/10.4067/S0716-10182018000600680 https://doi.org/10.4067/S0716-10182018000600680
5. Sijtsma K. In: On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha [Internet]. http://dx.doi.org/10.1007/s11336-008-9101-0https://doi.org/10.1007/s11336-008-9101-0
6. Toro R, Peña-Sarmiento M, Avendaño-Prieto BL, Mejía-Vélez S, Bernal-Torres A. Análisis Empírico del Coeficiente Alfa de Cronbach según Opciones de Respuesta, Muestra y Observaciones Atípicas. REV IBEROAM DIAGN EV. 2022;63: 17. https://www.redalyc.org/journal/4596/459671926003/459671926003.pdf https://doi.org/10.21865/RIDEP63.2.02
7. Komorita SS, Graham WK. Number of Scale Points and the Reliability of Scales. Educational and Psychological Measurement. 1965;25: 987–995. https://journals.sagepub.com/doi/10.1177/001316446502500404 https://doi.org/10.1177/001316446502500404
8. Carvajal D, Aboaja A, Alvarado R. Validación de la Escala de bienestar mental de Wareick-Edinburgo, en Chile. Rev Salud Pública (Córdoba). 2015;19: 13–21. https://revistas.unc.edu.ar/index.php/RSD/article/view/11822
9. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018;18: 91–93. http://dx.doi.org/10.1016/j.tjem.2018.08.001 https://doi.org/10.1016/j.tjem.2018.08.001
10. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med [Internet]. 2012;22: 276–82. http://dx.doi.org/10.1016/j.jocd.2012.03.005 https://doi.org/10.1016/j.jocd.2012.03.005
11. Sage Journals. 3 Oct 2023. https://journals.sagepub.com/action/cookieAbsent
12. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. West J Nurs Res. 2003;25: 508–18. http://dx.doi.org/10.1177/0193945903252998 https://doi.org/10.1177/0193945903252998
13. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995;50: 741–749. https://psycnet.apa.org/fulltext/1996-10004-001.pdf https://doi.org/10.1037/0003-066X.50.9.741
14. American Educational Research Association. Standards for Educational and Psychological Testing. 2014. https://play.google.com/store/books/details?id=clI_mAEACAAJ
15. Schuwirth LWT, van der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33: 783–97. http://dx.doi.org/10.3109/0142159X.2011.611022 https://doi.org/10.3109/0142159X.2011.611022
16. Kanya L, Sanghera S, Lewin A, Fox-Rushby J. The criterion validity of willingness to pay methods: A systematic review and meta-analysis of the evidence. Soc Sci Med. 2019;232: 238–261. http://dx.doi.org/10.1016/j.socscimed.2019.04.015 https://doi.org/10.1016/j.socscimed.2019.04.015
17. Silva JR, Behar R, Cordella P, Ortiz M, Jaramillo K, Alvarado R, et al. Validation of the Spanish version of the Eating Disorders Diagnostic Scale. Rev Med Chil. 2012;140: 1562–70. http://dx.doi.org/10.4067/S0034-98872012001200007 https://doi.org/10.4067/S0034-98872012001200007
18. Wadhwa RR, Test R. T Test. In: StatPearls [Internet] [Internet]. 2023. https://pubmed.ncbi.nlm.nih.gov/31971709/
19. Beath A, Jones MP. Guided by the research design: choosing the right statistical test. Med J Aust. 2018;208: 163–165. https://pubmed.ncbi.nlm.nih.gov/29490219/ https://doi.org/10.5694/mja17.00422
20. Cortés-Reyes É, Rubio-Romero JA, Gaitán-Duarte H. Métodos estadísticos de evaluación de la concordancia y la reproducibilidad de pruebas diagnósticas. Rev Colomb Obstet Ginecol. 2010;61: 247–255. http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0034-74342010000300009&lng=en&nrm=iso&tlng=es https://doi.org/10.18597/rcog.271
21. Larzelere RE, Smith GL, Batenhorst LM, Kelly DB. Predictive validity of the suicide probability scale among adolescents in group home treatment. J Am Acad Child Adolesc Psychiatry. 1996;35: 166–72. https://pubmed.ncbi.nlm.nih.gov/8720626/ https://doi.org/10.1097/00004583-199602000-00009
22. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955;52: 281–302. https://doi.org/10.1037/h0040957
23. Cuadras C. In: Nuevos métodos de análisis multivariante [Internet]. CMC Editions; 2014. https://gc.scalahed.com/recursos/files/r161r/w24899w/Semana5/METODOS_S5.pdf
24. Alvarado R, Jadresic E, Guajardo V, Rojas G. First validation of A Spanish-translated version of the Edinburgh postnatal depression scale (EPDS) for use in pregnant women. A Chilean study. Arch Womens Ment Health. 2015;18: 607–12. http://dx.doi.org/10.1007/s00737-014-0466-z https://doi.org/10.1007/s00737-014-0466-z
25. Gempp Fuentealba R. El error estándar de medida y la puntuación verdadera de los tests psicológicos: Algunas recomendaciones prácticas. Terapia [Internet]. 2006;24: 117–29. https://www.redalyc.org/articulo.oa?id=78524201
26. Barraza F, Arancibia M, Madrid E, Papuzinski C. General concepts in biostatistics and clinical epidemiology: Random error and systematic error. Medwave. 2019;19. https://doi.org/10.5867/medwave.2019.07.7687
27. Vetter TR, Mascha EJ. Bias, Confounding, and Interaction: Lions and Tigers, and Bears, Oh My! Anesth Analg. 2017;125: 1042–1048. https://journals.lww.com/anesthesia-analgesia/abstract/2017/09000/bias,_confounding,_and_interaction__lions_and.46.aspx https://doi.org/10.1213/ANE.0000000000002332

28. Bowling A. Research Methods In Health: Investigating Health And Health Services. McGraw-Hill Education (UK); 2014. https://play.google.com/store/books/details?id=6lOLBgAAQBAJ
29. Ferreiro RM, Valdés Amador L. Eficacia de distintas fórmulas ecográficas en la estimación del peso fetal a término. Rev Cubana Obstet Ginecol. 2010. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0138-600X2010000400003&lng=es&nrm=iso&tlng=es
30. Hernández Sampieri R, Fernández Collado C, Bautista Lucio P. Metodología de la investigación. McGraw-Hill. 2014.
31. Ramada-Rodilla JM, Serra-Pujadas C, Delclós-Clanchet GL. Cross-cultural adaptation and health questionnaires validation: revision and methodological recommendations. Salud Publica Mex. 2013;55: 57–66. http://dx.doi.org/10.1590/s0036-36342013000100009 https://doi.org/10.1590/s0036-36342013000100009
32. Foddy W. Constructing Questions for Interviews and Questionnaires. Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research. Cambridge University Press; 1993. https://www.cambridge.org/core/product/identifier/9780511518201/type/book https://doi.org/10.1017/CBO9780511518201

# Validation of questionnaires for the measurement of health variables: Fundamental concepts

**ABSTRACT**

In clinical practice and population health, it is common to use questionnaires to assess conditions or variables that are not directly observable. However, the construction and validation of these instruments or questionnaires are often poorly understood. This narrative review aims to summarise in a general way the process of construction and validation of these questionnaires in order to have a better understanding of this process, the aspects that are evaluated, and the best way to use them. The validation of questionnaires corresponds to a process of analysis of the questionnaire, aiming to measure a latent variable and its dimensions, which cannot be observed directly. A latent variable can be inferred through a set of specific attributes that are part of it, such as the items of a questionnaire, which are observable. Through a narrative review, this article addresses the fundamental concepts of questionnaire or test validation, latent variables or constructs, reliability and validity studies, and the factors that theoretically affect the latter two characteristics. Examples of these concepts are presented in the text