

Implicaciones del modelo estructural en la validación de instrumentos clínicos: Modelo reflectivo vs Modelo formativo

Agles Cruz-Peralta^{a,†}, María Luisa Peralta-Pedrero^{b,*§}, Martha Alejandra Morales Sánchez^{c,Δ}

Facultad de Medicina



Resumen

Antecedentes: Los cuestionarios, escalas, índices y criterios de clasificación se desarrollan y someten a validación, para medir de variables subjetivas de manera estandarizada.

Bases teóricas de los modelos estructurales: Cuando las preguntas de un cuestionario se originan a partir del atributo, corresponde al modelo reflectivo que ha sido ampliamente utilizado en psicometría debido a la naturaleza de las variables de interés. En Ciencias de la Salud las variables de interés son diferentes, cuando las preguntas de los cuestionarios nacen de elementos constitutivos y son independientes entre sí, pertenecen a un modelo formativo.

Implicaciones del modelo estructural en el diseño y validación de instrumentos de medición en salud: Cuando se usa el modelo reflectivo se crean un gran número de

preguntas, y su reducción se realizará mediante análisis estadístico. Se debe determinar la fiabilidad de la prueba (consistencia interna) y la validez estructural.

Con el modelo formativo se elabora un número reducido de ítems asegurando la validez de contenido con evidencia científica o, en su defecto, por consenso formal de expertos. Para refinarlos, se utiliza el nivel y la calidad de la evidencia, las características de la población objetivo y el propósito del instrumento, los métodos estadísticos y el juicio del investigador y/o el consenso formal de los expertos. Si existe un estándar de oro se realiza un estudio de pruebas diagnósticas, en caso contrario la validez de constructo.

Diferencias del modelo reflectivo versus formativo en la fase de desarrollo del instrumento: Naturaleza del constructo, cronología del fenómeno, relación entre los indicadores o ítems y el constructo, direccionalidad del

^a Facultad de Odontología, Universidad Nacional Autónoma de México, Cd. Mx., México.

^b Facultad de Medicina, Universidad Nacional Autónoma de México, Cd. Mx., México.

^c Centro Dermatológico Dr. Ladislao de la Pascua, Cd. Mx., México. ORCID ID:

[†] <https://orcid.org/0000-0001-9535-6998>

[§] <https://orcid.org/0000-0003-0301-1955>

^Δ <https://orcid.org/0000-0002-8371-3916>

Recibido: 10-abril-2024. Aceptado: 17-junio-2024.

* Autor para correspondencia: María Luisa Peralta-Pedrero. Escolar 411A, Copilco Universidad, Coyoacán, 04360 Ciudad de México, CDMX. Facultad de Medicina UNAM. Teléfono: 55 5622 5565. Correo electrónico: luisa.peraltap@gmail.com

Este es un artículo Open Access bajo la licencia CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

vínculo entre indicadores y constructo, número de indicadores identificados, ítems a elaborar y posible método de reducción, impacto teórico de la eliminación de ítems o indicadores sobre la idoneidad de la medición del constructo.

Palabras clave: Cuestionarios; escalas; validez; modelo reflectivo; modelo formativo.

Este es un artículo Open Access bajo la licencia CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Implications of the structural model in the validation of clinical instruments: Reflective model vs. formative model

Abstract

Background: Questionnaires, scales, indices and classification criteria are developed and subjected to validation to measure subjective variables in a standardized manner.

Theoretical bases of structural models: When the questions in a questionnaire originate from the attribute, it corresponds to the reflective model that has been widely used in psychometrics due to the nature of the variables of interest. In Health Sciences the variables of interest are different, when the questions in the questionnaires arise from constitutive elements and are independent of each other, they belong to a formative model.

Implications of the structural model in the design and validation of health measurement instruments: When using the reflective model, many questions are created, and their reduction will be carried out through statistical analysis. The reliability of the test (internal consistency) and structural validity must be determined.

With the formative model, a reduced number of items are developed, ensuring content validity with scientific evidence or, failing that, by formal consensus of experts. To refine them, the level and quality of evidence, the characteristics of the target population and the purpose of the instrument, statistical methods, and the researcher's judgment and/or formal expert consensus are used. If there is a gold standard, a study of diagnostic tests is carried out, otherwise the construct validity.

Differences of the reflective versus formative model in the instrument development phase: Nature of the construct, chronology of the phenomenon, relationship between the indicators or items and the construct, directionality of the link between indicators and construct, number of indicators identified, items to be developed and possible reduction method, theoretical impact of the elimination of items or indicators on the appropriateness of construct measurement.

Keywords: Questionnaires; scales; validity; reflective model; formative model.

This is an Open Access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

ANTECEDENTES

Los cuestionarios, escalas, índices y criterios de clasificación son instrumentos que se desarrollan y someten a validación, con la finalidad de realizar mediciones estandarizadas de variables no observables, como la gravedad, actividad o progresión de las enfermedades; a este tipo de variables se les denominan constructos o atributos. Estos instrumentos de medición de la salud se pueden clasificar según la función clínica para la que fueron desarrollados¹, por ejemplo, el Nutritional Risk Screening 2002 (NRS-2002)² con capacidad predictiva, el MiniNutritional Assessment (MNA) para detección. En ocasiones, un

instrumento puede cubrir más de una función, como es el caso del MNA, que también puede utilizarse como predictor de complicaciones³. Los criterios CASPAR (Classification Criteria for Psoriatic Arthritis) son un ejemplo de instrumento de diagnóstico⁴ y el SAPS 3 (Simplified Acute Physiology Score) como instrumento de pronóstico⁵. Otro aspecto importante a considerar, es el contexto para el cual se creó el instrumento, por ejemplo, el MNA fue inicialmente validado para aplicar a personas ancianas de la comunidad, pero no para ancianos hospitalizados¹.

Dependiendo de la fuente de información, se clasifican en resultados informados por el paciente

(patient-reported outcome: PRO), resultado informado por el médico (Clinician-Reported Outcome: ClinRo), resultado informado por el observador (observer-reported outcome: ObsRo) y resultado de desempeño (performance outcome: PerfO)⁶.

A finales del siglo XX, Feinstein introdujo el concepto de clinimetría como la disciplina que rige la medición válida y confiable de los fenómenos clínicos, lo que precedió a un incremento exponencial de nuevas escalas y cuestionarios en las ciencias de la salud⁷. Mientras que la psicometría se enfoca en variables que intervienen en los procesos psicológicos y capacidades cognitivas, en la clinimetría existe una gran heterogeneidad en la naturaleza de las variables que deben medirse, como signos vitales, medidas de desempeño, medidas de cambio en los estados clínicos, satisfacción, malestar físico, respuesta a tratamiento, medidas antropométricas, marcadores bioquímicos, etc. Sin embargo, la clinimetría utiliza las mismas técnicas estadísticas y teorías creadas por grandes psicometristas como Charles E. Spearman y Lee J. Cronbach; lo que hace indispensable un análisis cuidadoso de la idoneidad de estos métodos dependiendo de los tipos de instrumentos, su contexto de aplicación y las variables que los componen.

BASES TEÓRICAS QUE EXPLICAN LOS MODELOS ESTRUCTURALES

Existen dos modelos de medición, el modelo reflexivo y el modelo formativo⁸⁻¹⁰. La teoría clásica de los tests (classical test theory: CTT) fue introducida por Spearman en 1904 y establece que la puntuación empírica (puntuación obtenida al aplicar un instrumento) es el resultado de la suma de la puntuación verdadera más un término de error de medición. Entonces la puntuación verdadera expresa la magnitud de la variable no observable que pretendemos medir, también llamada atributo; por tanto, si usando técnicas estadísticas podemos estimar el error y descontarlo de la puntuación observada, podremos conocer la puntuación real que nos informará de la magnitud del atributo que nos interesa y que es imposible medir directamente¹¹. Esta teoría evalúa toda la prueba y se representa en términos matemáticos con la siguiente fórmula:

$$X_i = V_i + \epsilon_i$$

Donde puntuación observada (X_i) = puntuación verdadera (V_i) + error de medición (ϵ_i).

Esta teoría es la más utilizada, se basa en un modelo lineal, que explica la variación de las puntuaciones observadas o puntuaciones totales de un instrumento como la suma de dos tipos de variación:

- a) Variación no compartida.
- b) Variación compartida o varianza común.

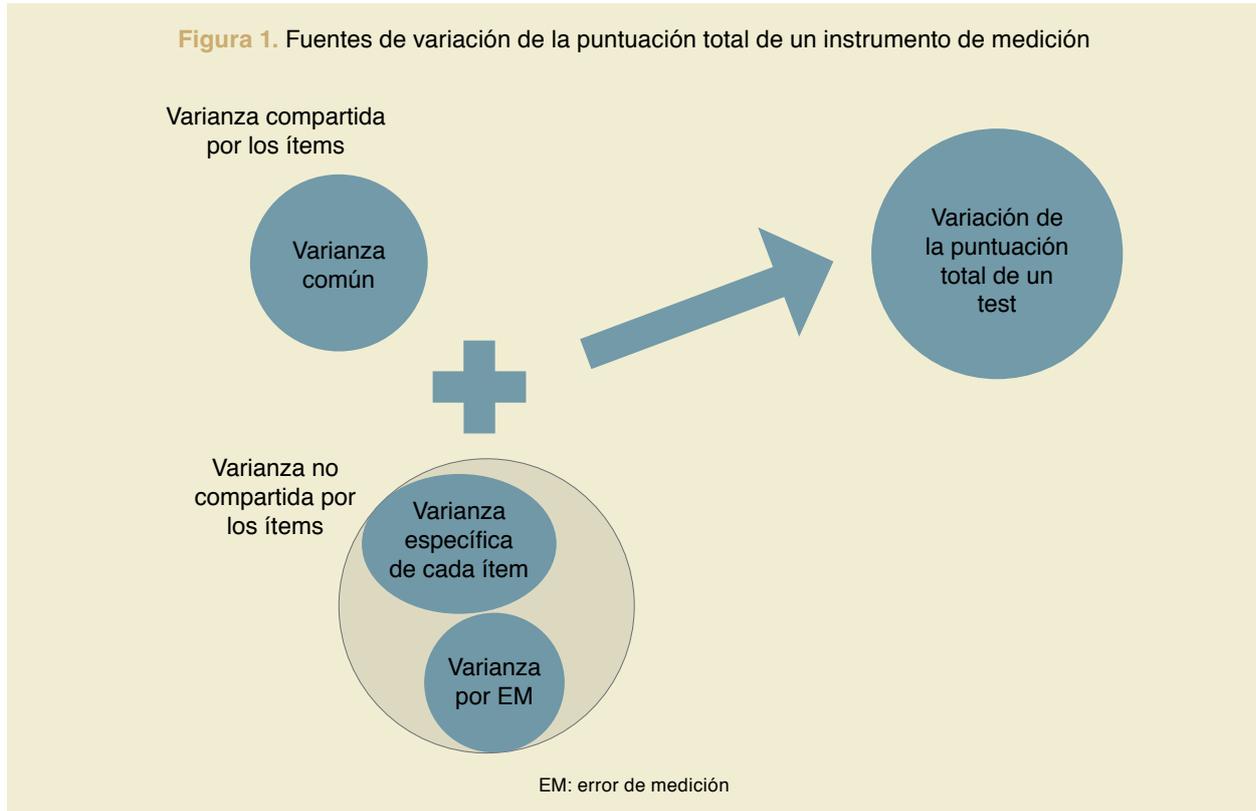
La variación no compartida proviene de dos fuentes:

El error de medición: son variaciones en las puntuaciones, que son independientes del atributo que nos interesa medir y pueden provenir del encuestado o del contexto de aplicación. Para tratar de disminuir este error se utiliza la estandarización, haciendo que la aplicación de los cuestionarios sea en las mismas condiciones para todos los participantes y un tamaño de muestra apropiado que proporcione precisión. Sin embargo, siempre habrá algunas situaciones que no se podrán controlar, por ejemplo, si el encuestado tuvo un conflicto familiar antes de asistir a la medición, si al momento de responder la prueba tiene algún malestar físico o mental, si hace calor o frío en el lugar donde se encuentra, si escucha algún ruido inesperado, etc.^{11,12}

La variación propia del ítem: es una variación que aporta el ítem por la forma en que es percibido por el encuestado y que no expresa el atributo de interés. Esta fuente de variación se intenta controlar en la fase de desarrollo del instrumento, por ejemplo, con entrevistas cognitivas y ajustes posteriores hasta que la percepción sea lo más uniforme posible¹¹.

Ambas fuentes de variación son inseparables entre sí y también independientes del atributo por lo que afectan la medición “exacta o real” disminuyendo la validez de los datos (**figura 1**).

Por otro lado, la parte de la variación de la puntuación total, que corresponde a la varianza común, es una variación compartida por todos los ítems. Esta varianza común se explica porque todos los ítems del instrumento tienen el mismo origen, que es el atributo que nos interesa medir, por ejemplo, para desarrollar un instrumento que mida la satis-

Figura 1. Fuentes de variación de la puntuación total de un instrumento de medición

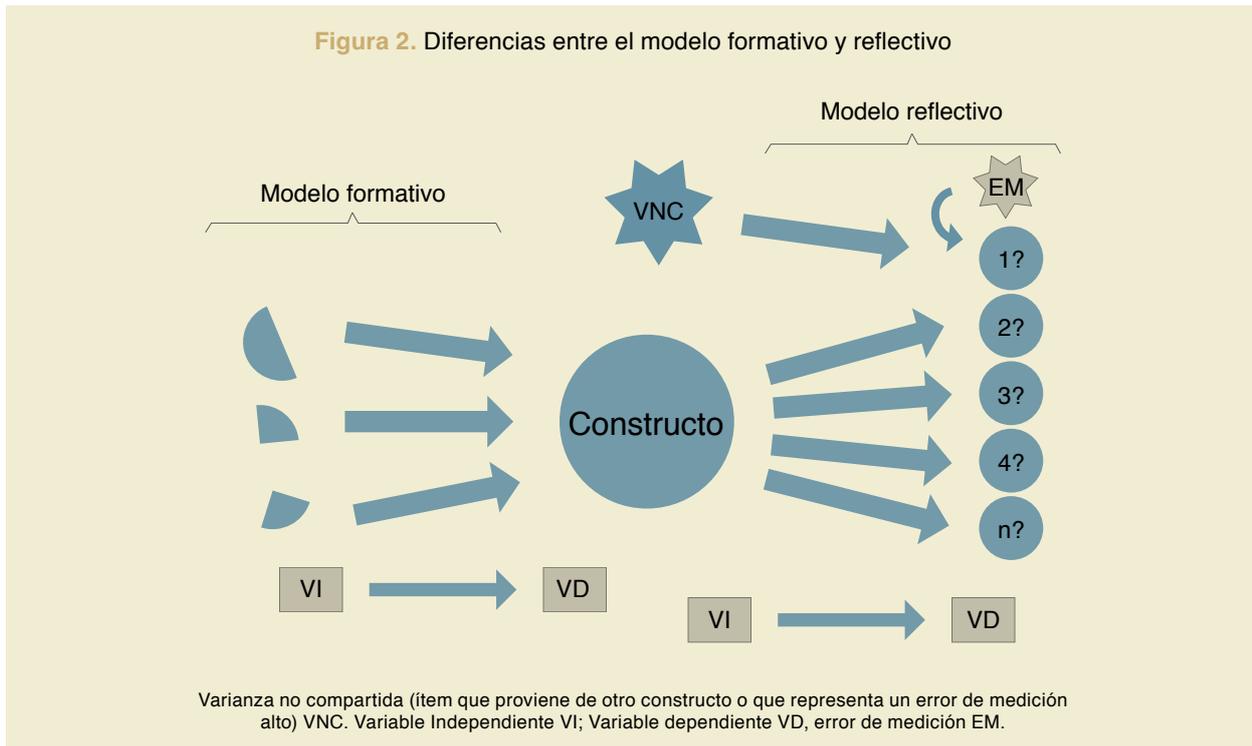
facción del paciente, es necesario elaborar un conjunto de preguntas o declaraciones que “reflejen” esa satisfacción (manifestaciones observables del estar satisfecho); por supuesto, el límite de preguntas será infinito⁸⁻¹⁰.

Como todas las preguntas se originan a partir del atributo o en otras palabras son un reflejo del atributo, este modelo se conoce como “Constructo Latente Común” o “Modelo Reflectivo”. Este origen común puede comprobarse mediante estadística, cuando se obtiene una alta correlación entre los ítems (interrelación). Y con esto ya hablamos de una propiedad de medición muy importante que es la confiabilidad de manera particular la “consistencia interna” o “fiabilidad”, que mide el grado de homogeneidad de una escala (**figura 2**)^{12,13}. Así, si todos los ítems tienen el atributo como origen común y lo demostramos con una alta interrelación de los ítems, se puede confiar en que la escala es un adecuado “reflejo del constructo” adecuado y por tanto es válida. Entonces, en psicometría, la consistencia interna es una medida indirecta de validez,

ya que no existe un estándar de oro para validar el instrumento. Esta propiedad, al menos muestra que todos los ítems miden lo mismo, por lo que es probable que sea válida (**figura 2**).

Volviendo a la construcción de instrumentos de este tipo, un método para afinar esta gran cantidad de preguntas es precisamente la interrelación de los ítems y se utiliza el estadístico α de Cronbach o K-20¹⁴. Como se puede observar en la **figura 2**, la exclusión o inclusión de cada pregunta afecta el grado de homogeneidad de la prueba, de tal manera que, si alguna pregunta lo disminuye, puede ser por lo siguiente:

- Esa pregunta puede ser que no refleje el constructo que se pretende medir, de tal manera que puede provenir de algún constructo relacionado y afectará negativamente la medición.
- Esa pregunta puede tener una alta variación no compartida, lo que significa, que débilmente refleja el constructo de interés porque la variación se contamina con la alta variación de error y propia de ítem.

Figura 2. Diferencias entre el modelo formativo y reflectivo

En ambos casos es mejor eliminarla (**figura 2**).

La validez es una cuestión de grado, por lo tanto, cuantas más pruebas de validez se proporcionen, más certeza habrá de que el instrumento realmente mida lo que pretende medir^{12,13}.

Cuando los instrumentos son multidimensionales, es necesario evaluar la validez estructural, la cual se determina con técnicas de “interdependencia” multivariadas como el análisis factorial. Esta técnica se utiliza para “reducir ítems” en función de la varianza común, lo que permite identificar el conjunto de ítems que mejor refleja el constructo¹⁴.

El CTT y el modelo reflectivo han sido ampliamente utilizados en psicometría debido a la naturaleza de las variables¹¹. En Ciencias de la Salud también necesitamos medir atributos no observables, pero las variables de interés para la clinimetría son diferentes. Por ejemplo, en el caso de pacientes con psoriasis, hace tiempo se observó que desarrollaban una enfermedad inflamatoria articular, luego de varios reportes se vislumbró que se trataba de una entidad nosológica que hasta el momento no había sido descrita (constructo: artritis psoriásica) y así surgieron los diferentes criterios de clasificación:

Criterios de Moll y Wright, Bennett, Gladman et al., Vasey y Espinoza, los del Grupo Europeo de Estudio de Espondilo artropatías (ESSG), McGonagle y los de Fournie. Finalmente, se desarrollaron los criterios de clasificación de la Artritis Psoriásica (CASPAR), (**tabla 1**), que demostraron una sensibi-

Tabla 1. Características clínicas que demostraron presentarse con mayor frecuencia en pacientes con artritis psoriásica (APso) comparado con controles

	Diferencia en puntos porcentuales más frecuente en pacientes con APso
Psoriasis	84%
Psoriasis actual	80%
Psoriasis previa	5%
Antecedente familiar de psoriasis	10%
Psoriasis en uñas	37%
Factor reumatoide negativo	49%
Dactilitis previa o actual	27%
Evidencia radiográfica de nuevo hueso	1%

Fuente: Chandran V, Schentag CT, Gladman DD. Sensitivity and specificity of the CASPAR criteria for psoriatic arthritis in a family medicine clinic setting. *J Rheumatol.* 2008;35(10):2069-70.

lidad del 91.4% y una especificidad del 98.7%¹⁶. En este caso, el instrumento CASPAR fue desarrollado con base en evidencia científica que demostró que cada variable observable y no observable (signos, síntomas y pruebas de laboratorio) contribuyen de forma independiente a identificar la presencia de la enfermedad, contando además con el peso de cada uno ponderando cada ítem; esto se obtuvo con técnicas de dependencia multivariada como la regresión logística o lineal. Este es un claro ejemplo de un “Constructo Latente Agregado” que nace de elementos constitutivos (ítems o indicadores) que tienen diferente origen y son independientes entre sí, por esta razón, en los instrumentos de medición que cuentan con un modelo formativo, no se justifica utilizar el alfa de Cronbach o análisis factorial para reducir ítems o demostrar homogeneidad. Reforzando este punto, recordemos que, en el caso de las regresiones, para que los modelos sean estables y replicables, requieren que las variables independientes (ítems) no tengan altas correlaciones entre sí (descartar colinealidad); además, estas técnicas buscan determinar el peso de cada indicador de forma independiente. Entonces, los instrumentos con un modelo de medición formativo deben demostrar inicialmente validez de contenido y posteriormente, si existe un estándar de oro, validez de criterio o en su defecto validez de constructo.

IMPLICACIONES DEL MODELO ESTRUCTURAL EN EL DISEÑO Y VALIDACIÓN DE INSTRUMENTOS DE MEDICIÓN EN SALUD

Comparando el modelo de “Constructo Latente Común” o modelo reflectivo con el modelo de “Constructo Latente Agregado” o modelo formativo, en relación con las propiedades de medición y método para crear un instrumento, se puede concluir lo siguiente:

Modelo reflectivo

Desarrollar un gran número de preguntas, no existe un número universalmente aceptado, pero parece razonable de 3 a 10 por cada una que se pretende que quede al final del estudio. Su reducción se realizará mediante análisis estadístico. Se debe determinar la fiabilidad de la prueba (consistencia interna) y la validez estructural.

Modelo formativo

Elaborar un número reducido de ítems a partir de un número generalmente reducido de indicadores que hayan mostrado asociación independiente con el constructo de interés, asegurando la validez de contenido con evidencia científica o en su defecto por consenso formal de expertos. Para refinarlos se utiliza el nivel y la calidad de la evidencia, las características de la población objetivo y el propósito del instrumento, los métodos estadísticos y el juicio del investigador y/o el consenso formal de los expertos¹. Finalmente, si existe un estándar de oro, la validez de criterios externos (estudio de pruebas diagnósticas o estudio de proceso) y en caso contrario la validez de constructo. Para este tipo de instrumentos no se justifica utilizar el análisis factorial ya que eliminar un indicador de esta forma puede dañar seriamente la validez de contenido¹⁵. En cuanto a la consistencia interna, no existe un patrón preconcebido de interrelación entre los ítems, incluso se puede esperar un alfa de Cronbach o KR 20 moderado a bajo y esto no indica deficiente validez del instrumento ya que, como se explicó, los indicadores suelen ser heterogéneo. Hasta la fecha no existen criterios simples, heterogéneos y universalmente aceptados para evaluar la consistencia interna de los indicadores formativos⁸⁻¹⁰.

Tanto la psicometría como la clinimetría se basan en construcciones teóricas de los fenómenos en estudio, esquemas conceptuales o estructurales, conocimientos previos y enfoques analíticos; sin embargo, lo importante es identificar el modelo de medición formativo versus reflectivo que puede darse en ambas disciplinas.

Es importante hacer especial referencia a la validez de constructo que se puede utilizar en ambos modelos, a través del método de prueba de hipótesis.

Por otro lado, la validez estructural o dimensionalidad, algunos autores la ubican dentro de la validez de constructo, otros la consideran por separado; lo importante es que solo aplica para el modelo reflectante.

Otro aspecto importante, que ha generado confusión, es el tipo de instrumento según la fuente de información y el modelo de medición. Se considera que los instrumentos de resultados informados por el paciente (PRO) siguen un modelo reflectivo

ya que miden las experiencias del paciente, los instrumentos informados por el médico (ClinRO) se consideran un modelo formativo^{14,15}. En la mayoría de los casos esto puede coincidir; sin embargo, es importante no olvidar que la fuente de información no determina el modelo de medición, ya que el modelo de medición dependerá del tipo de ítems y naturaleza del constructo.

Finalmente, no olvidemos que, para desarrollar un instrumento de medición de la salud, en “todos los casos” el primer paso será definir o delimitar con la mayor precisión posible el constructo que interesa medir, la población de estudio y el contexto¹.

CLAVES PARA DIFERENCIAR EL MODELO REFLECTIVO VERSUS FORMATIVO EN LA FASE DE DESARROLLO DEL INSTRUMENTO⁸⁻¹⁰

a) Naturaleza del constructo

¿Qué nos interesa medir? Cuando el constructo es una entidad nosológica y la finalidad es establecer un diagnóstico, pronóstico o categorizar grupos, por ejemplo; será un punto a favor del modelo formativo. En cambio, puede ser o no una entidad nosológica, pero se pretende saber, ¿cuáles son sus manifestaciones, experiencias u opiniones en la población que la padece?, en diferentes contextos desde un determinado enfoque; apuntan a favor del modelo reflectivo.

b) Cronología del fenómeno

Partiendo de la representación de un mapa conceptual del fenómeno que interesa investigar: Si los indicadores se encuentran antes del constructo que se pretende medir; apuntan a favor del modelo formativo (indicadores causales). Si los indicadores se encuentran posteriores al constructo; apunta a favor del modelo reflectivo (indicadores de efecto).

c) Relación entre los indicadores o ítems y el constructo

Los indicadores o ítems son muy similares, se refieren al mismo tema, tienen un origen común y se espera una alta “intercorrelación”. Punto a favor del modelo reflectante. A diferencia, cuando los indicadores colaboran de forma independiente en la ocurrencia del constructo, generalmente tienen orígenes diferentes, son diferentes entre sí y tienen una

relación de dependencia con el constructo. Punto a favor del modelo formativo.

d) Direccionalidad del vínculo entre indicadores y constructo

Si modificar los indicadores modifica el constructo, apuntar a favor del modelo formativo, si modificar el constructo modifica los indicadores, señalar a favor del modelo reflectivo. Es decir, cuando los indicadores funcionan como variables independientes (indicadores causales) y el constructo como dependiente, corresponde a un modelo formativo. Cuando es al revés, el constructo funciona como variable independiente y los indicadores como variable dependiente (indicadores de efecto) corresponden a un modelo reflectivo.

e) Número de indicadores identificados, ítems a elaborar y posible método de reducción

Hay un número casi infinito de indicadores o preguntas que podrían plantearse, cada experto puede desarrollar un subconjunto de ellos, manteniendo su utilidad porque se basan en constructos teóricos válidos (“diferentes formas de plantear un mismo tema”). Punto a favor del modelo reflectante.

Cuando se identifica un número relativamente pequeño de indicadores basados en la evidencia que inciden en la ocurrencia del constructo, a partir de estos indicadores se pueden elaborar 2 o 3 formas de preguntar o aseverar según la estructura que se requiere del instrumento, con el único fin de que sean entendibles, pero manteniendo el indicador. Punto a favor del modelo formativo.

f) Impacto teórico de la eliminación de ítems o indicadores sobre la idoneidad de la medición del constructo

En caso de que, del total de preguntas o indicadores candidatos, cualquier “subconjunto de ellos” podría incluirse en el instrumento sin afectar la medición del constructo desde el punto de vista teórico. Punto a favor del modelo reflectante.

En contraste, cuando del total de preguntas candidatas solo se pueden eliminar aquellas que no garanticen un significado estándar para los respondientes, pero debe quedar una por indicador identificado

como necesario. En otras palabras, si algún indicador no está representado por al menos un ítem, la validez de contenido del instrumento puede verse afectada. Punto a favor del modelo formativo.

CONTRIBUCIÓN INDIVIDUAL

- ACP: Coordinación del equipo, elaboración y discusión del manuscrito, aprobación final del documento.
- MLPP: Asesoría, aclaración de conceptos, elaboración y discusión del manuscrito y aprobación final del documento.
- MAMS: Elaboración y discusión del manuscrito aprobación final del documento y búsqueda y recuperación de documentos.

PRESENTACIONES PREVIAS

Ninguna.

FINANCIAMIENTO

Ninguno.

CONFLICTOS DE INTERESES

Ninguno. 

REFERENCIAS

1. Cruz-Avelar A, Sinaí Cruz-Peralta E. Metodología para la construcción de instrumentos de medición en salud. *Alergia, Asma e Inmunología pediátricas*. 2017;26:100-105.
2. Hersberger L, Bargetzi L, Bargetzi A, Tribolet P, Fehr R, Baechli V, Geiser M, Deiss M, Gomes F, Kutz A, Kägi-Braun N, Hoess C, Pavlicek V, Schmid S, Bilz S, Sigrist S, Brändle M, Benz C, Henzen C, Nigg M, Thomann R, Brand C, Rutishauser J, Aujesky D, Rodondi N, Donzè J, Stanga Z, Mueller B, Schuetz P. Nutritional risk screening (NRS 2002) is a strong and modifiable predictor risk score for short-term and long-term clinical outcomes: secondary analysis of a prospective randomised trial. *Clin Nutr*. 2020;39(9):2720-2729. doi: 10.1016/j.clnu.2019.11.041. Epub 2019 Dec 14. PMID: 31882232.
3. Rubenstein LZ, Harker JO, Salvà A, Guigoz Y, Vellas B. Screening for undernutrition in geriatric practice: developing the short-form mini-nutritional assessment (MNA-SF). *J Gerontol A Biol Sci Med Sci*. 2001;56(6):M366-72. doi: 10.1093/gerona/56.6.m366. PMID: 11382797.
4. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P, Mielants H; CASPAR Study Group. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. *Arthritis Rheum*. 2006;54(8):2665-73. doi: 10.1002/art.21972. PMID: 16871531.
5. Van der Merwe E, Kapp J, Pazi S, Aylward R, Van Niekerk M, Mrara B, Freercks R. The SAPS 3 score as a predictor of hospital mortality in a South African tertiary intensive care unit: A prospective cohort study. *PLoS One*. 2020;15(5):e0233317. doi: 10.1371/journal.pone.0233317. PMID: 32437390; PMCID: PMC7241826.
6. Walton MK, Powers JH 3rd, Hobart J, Patrick D, Marquis P, Vamvakas S, Isaac M, Molsen E, Cano S, Burke LB; International Society for Pharmacoeconomics and Outcomes Research Task Force for Clinical Outcomes Assessment. Clinical Outcome Assessments: Conceptual Foundation-Report of the ISPOR Clinical Outcomes Assessment - Emerging Good Practices for Outcomes Research Task Force. *Value Health*. 2015;18(6):741-52. doi: 10.1016/j.jval.2015.08.006. Epub 2015 Aug 24. PMID: 26409600; PMCID: PMC4610138.
7. Feinstein AR, Clinimetrics. 1st ed. Massachusetts: Yale University Press; 1987.
8. Coltman T, Devlin TM, Midgley DF, Venaik S. Formative versus reflective measurement models: Two applications of formative measurement. *J Bus Res*. 2008;61(12):1250-1262. doi:10.1016/j.busres.2008.01.013
9. Fleuren BPI, van Amelsvoort LGPM, Zijlstra FRH, de Grip A, Kant I. Handling the reflective-formative measurement conundrum: a practical illustration based on sustainable employability. *J Clin Epidemiol*. 2018;103:71-81. doi: 10.1016/j.jclinepi.2018.07.007. Epub 2018 Jul 19. PMID: 30031210.
10. Van Amelsvoort LGPM, Fleuren BPI, Kant I. Improving measurement models in clinical epidemiology: time to move beyond the inherent assumption of an underlying reflective measurement model. *J Clin Epidemiol*. 2020;118:119-123. doi: 10.1016/j.jclinepi.2019.11.003. Epub 2019 Nov 7. PMID: 31706961.
11. Magnusson D. Análisis de ítems. En: Magnusson D, editor. *Teoría de tests: psicometría diferencial, psicología aplicada, orientación vocacional*. 2ª ed. México: Trillas; 1990. p 237-267.
12. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007 Jan;60(1):34-42. doi: 10.1016/j.jclinepi.2006.03.012. Epub 2006 Aug 24. PMID: 17161752.
13. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. Cambridge University Press; 2011. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1147-1157. doi: 10.1007/s11136-018-1798-3. Epub 2018 Feb 12. PMID: 29435801; PMCID: PMC5891568.
14. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. 2018;27(5):1159-1170.
15. Chandran V, Schentag CT, Gladman DD. Sensitivity and specificity of the CASPAR criteria for psoriatic arthritis in a family medicine clinic setting. *J Rheumatol*. 2008;35(10):2069-70.