

# (#128) . INDICADORES FORMATIVOS EN ECUACIONES ESTRUCTURALES

[REVISIÓN DE ARTÍCULO] La medición de variables latentes con indicadores formativos es un tema muy discutido, tal y como los autores exponen en el comienzo del artículo.

**Un indicador formativo se diferencia de uno reflectivo en el que el primero afecta a la variable latente, mientras que en el segundo es la variable latente la que produce un efecto en el indicador. Están, por tanto, causalmente relacionados con la variable latente de manera opuesta.**

La correcta especificación de un modelo usando indicadores reflectivos o formativos es esencial. La estimación del modelo y su interpretación cambian ostensiblemente según se haya hecho la modelización de una forma o de otra. Por eso este tema es tan importante en cualquier tipo de planteamiento donde se empleen variables latentes.

El objetivo de este artículo es revisar las principales críticas hacia este tipo de indicadores y tratar de rebatirlas.

## **Distinción entre indicadores formativos y reflectivos**

Una forma de hacerlo es a través de experimentos mentales. Por ejemplo, hacerse la sencilla pregunta de: **¿es el indicador el que causa variaciones en la variable latente o al revés? En el primer caso estaríamos hablando de indicadores formativos y en el segundo de reflectivos.**

También abogan por una comparación empírica de varias conceptualizaciones competitivas, para escoger aquella con mejor ajuste, pero en este caso se podría dar un ajuste con

una mala especificación, por lo que resulta mucho más adecuado establecer a priori el tipo de medición antes de cualquier análisis.

Recordemos que una variable latente, en definición de uno de los autores (Bollen), es una variable en la que no hay una realización muestral en al menos alguna observación en la muestra. Esto quiere decir que, de forma habitual, las variables latentes no tienen valores observables en la muestra.

**Según los autores, los indicadores formativos pueden ser o no la única combinación posible de indicadores para formar el concepto, o lo que es lo mismo, el concepto se define a través de los indicadores que lo forman, ya sea de forma perfecta (sin término de error), o de manera aproximada (con término de error). Este segundo caso es el más habitual para Bollen y Diamantopoulos, y se resume matemáticamente así:**

Variable latente =  $\Lambda_1 x_1 + \Lambda_2 x_2 + \dots + \Lambda_n x_n + \text{Error formativo}$

En el caso de 3 indicadores la ecuación formativa quedaría así:

Variable latente =  $\Lambda_1 x_1 + \Lambda_2 x_2 + \Lambda_3 x_3 + \text{Error formativo}$

Sin embargo la ecuación reflectiva sería así:

$x_1 = \Lambda_1 \text{Variable latente} + \text{Error}_1$

$x_2 = \Lambda_2 \text{Variable latente} + \text{Error}_2$

$x_3 = \Lambda_3 \text{Variable latente} + \text{Error}_3$

**Como puede comprobarse, la especificación es distinta, y las implicaciones causales también. Es muy importante recalcar que las variaciones en la variable latente son determinadas de manera diferente; para los indicadores formativos, ellos**

**mismos pueden contribuir a la variación de la variable latente, mientras que para los reflectivos esa variación tiene que ser externa a los indicadores.**

En un planteamiento formativo los pesos de los indicadores pueden fijarse a priori según la teoría o estimarse, y permite la flexibilidad de no tener que admitir que los indicadores puedan ser intercambiables o que tengan que estar todos fuertemente correlacionados. Si dudamos en que ello sea así, es que probablemente estemos ante un caso de indicadores formativos.

Los autores también distinguen entre variables latentes donde los indicadores tienen unidad conceptual, y variables compuestas donde los indicadores suman sus efectos para obtener un índice. Esta divergencia de planteamiento es fundamental para ellos.

### **Crítica 1. Un constructo medido con indicadores formativos no existe separadamente de sus indicadores**

Es decir, si se cambian los indicadores se cambia el constructo. Pero los autores argumentan que esto es así cuando se forman índices o variables compuestas, pero no cuando la variable latente está formada por indicadores y existe un término de error. **Al hacer una definición a priori del constructo, al igual que sucede si fuera una variable latente reflectiva, es ese constructo el que justifica la selección de indicadores.**

### **Crítica 2. Los indicadores formativos son causas, más que medidas**

Para mí esta es la parte más controvertida, ya que los autores **no justifican el trasfondo de la crítica. Un indicador formativo puede en realidad convertirse en una variable latente y un indicador reflectivo.**

Por ejemplo, si definimos la variable latente "Calidad de un

servicio" con dos indicadores formativos (elementos tangibles, elementos intangibles), claramente podemos re-especificar esos indicadores como dos variables latentes diferentes con un indicador reflectivo cada una. Así, tanto los elementos tangibles como los intangibles afectarían a la calidad del servicio, es decir, serían causas, mientras que los indicadores reflectivos de esas causas serían las medidas.

Bollen y Diamantopoulus no hacen referencia a este simple razonamiento, que es en mi opinión la base de la crítica a los indicadores formativos.

### **Crítica 3. Los indicadores formativos implican múltiples dimensiones de un constructo**

En realidad no implican múltiples dimensiones como indican los autores, pero la crítica se refiere más al hecho comentado en el apartado anterior, en que los indicadores de un constructo pueden re-especificarse como múltiples dimensiones con indicadores reflectivos cada una de ellas.

Esto trae también **la discusión del concepto de dimensionalidad**, ya que esas nuevas variables latentes pueden ser consideradas causas de la variable latente primaria (como en el caso del ejemplo comentado de la calidad del servicio), con un error de determinación asociado, lo que bajo mi punto de vista conllevaría la desaparición de la dimensionalidad, y la aceptación de que son constructos diferentes que causan variación en la variable latente principal.

### **Crítica 4. Los indicadores formativos se asumen que están libres de error de medida**

Los autores se escudan en que existe un error asociado a esa determinación de la variable latente, como en un modelo de regresión habitual:

Variable latente =  $\text{Lambda}_1 \cdot x_1 + \text{Lambda}_2 \cdot x_2 + \dots + \text{Lambda}_n \cdot x_n + \text{Error formativo}$

Ese "Error formativo" es diferente del error de medida en los indicadores observables  $x_1 \dots x_n$ , y es ahí de nuevo donde una re-especificación con un modelo de variables latentes reflectivas solucionaría el problema.

### **Crítica 5. Los indicadores formativos están sujetos a confundido interpretativo**

El confundido interpretativo no es más que la consecuencia de establecer un modelo causal amplio entre varias variables latentes, de tal manera que la especificación de la medición en otras variables o relaciones entre ellas puede condicionar la propia relación entre un constructo y sus indicadores observables.

Los autores advierten que este es un problema tanto para indicadores reflectivos como formativos. Ante esta situación, en mi opinión, hay que construir modelos causales con amplitud de miras, fijando parte las mediciones de cada variable (tipo indicador *gold standar* en la terminología de Leslie A. Hayduk).

### **Crítica 6. Los indicadores formativos no pasan las restricciones de proporcionalidad**

Realmente los test de restricciones de proporcionalidad son una forma de ver si esos indicadores son adecuados o no, más que una limitación de los mismos.

### **Crítica 7. Los coeficientes de los indicadores formativos deben fijarse a priori**

No es necesario según su conceptualización de indicadores formativos. Además, el hecho de fijar indicadores puede ser parte también de una forma de testar mediciones reflectivas, como el caso del *gold standard* de Hayduk.

### **Conclusión/Comentarios**

Creo que los autores hacen un buen trabajo profundizando en

las críticas sobre los indicadores formativos, **distinguiendo claramente entre variables que son latentes e índices creados artificialmente**. En el segundo caso no se estaría hablando de una variable latente en el sentido estricto de la palabra, y por tanto tampoco de indicadores formativos en la conceptualización entendida como opuesta a los reflectivos.

Si admitimos esa buena distinción de los autores, la cuestión que habría que abordar con mayor claridad es la posible re-especificación de los indicadores formativos en variables latentes reflectivas. Bajo mi punto de vista, **esto concordaría mucho más con las visiones sobre medición de otros grandes nombres en la metodología actual (ej. Hayduk, Borsboom), y ayudaría a modelar los errores de medida que cualquier indicador puede tener.**

Ante esta situación, mi visión a día de hoy es que es preferible la re-especificación, y luego seguir una estrategia de modelado similar a la [recomendada por Hayduk en su libro de 1996](#), fijando los indicadores *gold standar*, y reduciendo al mínimo posible los indicadores de cada variable latente.

**Lo que sí es importante es, desde luego, saber distinguir los indicadores reflectivos de los formativos**, y no emplear indicadores reflectivos “a discreción”, que es uno de los errores comunes que algunos investigadores cometen. Una buena reflexión acerca de la naturaleza de las mediciones en el modelo planteado es el primer paso, para después pensar acerca de posibles re-especificaciones y estrategias de análisis.

**Bollen, K. A. & Diamantopoulos, A. (2015). In defense of causal-formative indicators: A minority report. Psychological Methods, doi: 10.1037/met0000056**

Indicadores de calidad de la revista\*

|  | Impact Factor (2014) | Cuartil | Categoría |
|--|----------------------|---------|-----------|
|--|----------------------|---------|-----------|

|                       |      |    |                              |
|-----------------------|------|----|------------------------------|
| Thomson-Reuters (JCR) | 7.34 | Q1 | PSYCHOLOGY-MULTIDISCIPLINARY |
| Scimago (SJR)         | 4.95 | Q1 | PSYCHOLOGY                   |

\* Es simplemente un indicador aproximado para valorar la calidad de la publicación

[Todos los posts relacionados](#)

---

## (#127). AJUSTE EXACTO EN ECUACIONES ESTRUCTURALES

[REVISIÓN DE ARTÍCULO] El profesor **Leslie A. Hayduk** es uno de los mayores expertos en los modelos de ecuaciones estructurales (SEM), una de las metodologías más empleadas en ciencias sociales. En este artículo intenta lo más didácticamente posible exponer postulados fundamentales sobre su correcto uso, algo que, lamentablemente, no se lleva a cabo por la mayoría de investigadores que utilizan este método. De este modo, el autor comenta algunos de los puntos clave para entender que **la única vía posible es la del ajuste exacto mediante el test de la chi-cuadrado**, rebatiendo algunos argumentos falaces.

### 1. Hipótesis nula con cero efectos

En muchos contextos de investigación esto se refiere a la especificación de una hipótesis nula con escaso interés para la investigación, ya que postula cero efecto en la población, lo que puede alejarse del verdadero interés de estudiar la existencia de un efecto sustantivo.

Sin embargo, en SEM la hipótesis nula sí que tiene interés por sí misma porque encapsula afirmaciones teóricas sobre las

estructuras que producen las variables observables. En SEM, la hipótesis nula hace referencia a que la matriz de covarianzas entre las variables implicada por el modelo propuesto es compatible con la matriz observable, dentro del error muestral (de ahí el test estadístico). En otras palabras, testa si la conceptualización que se especifica es consistente con los datos empíricos.

Así, la matriz de covarianzas implicada por el modelo es la muestra, y la de los datos empíricos es la de la población. Esto tiene que quedar muy claro, porque es una de las grandes diferencias con el modo de enfocar las hipótesis nulas en los test estadísticos más habituales: diferencia de medias, correlación, etc.

Una de las grandes ventajas de SEM es que podemos testar coeficientes no nulos, es decir, fijar el valor de determinados parámetros y ver si el modelo se ajusta. Esto sería equivalente a analizar si un tamaño de efecto determinado para una asociación entre variables es compatible con los datos empíricos.

## **2. El desajuste del modelo no tiene necesariamente que estar asociado al grado de mala especificación**

Este es otra de las ideas que cuesta asimilar, porque difiere del pensamiento más intuitivo de considerar que cuanto más pobre sea el ajuste del modelo peor va a ser el grado de mala especificación, y que cuanto mejor sea el ajuste el modelo será más correcto.

Esta falacia se derrumba al considerar los modelos equivalentes, que son un tipo de modelos que obtienen el mismo ajuste pero con una especificación causal diferente. También al entender que los modelos saturados (mismo número de ecuaciones que de incógnitas) produce a menudo ajuste perfecto, aunque la especificación obviamente puede ser incorrecta.



El uso del análisis factorial exploratorio es otra vía para ver que modelos causalmente mal especificados pueden pasar el test de la chi-cuadrado, es decir, ajustarse. Incrementando el número de factores puede mejorarse el ajuste.

Como bien indica Hayduk, **como ese tipo de modelos mal especificados pueden indicar ajustes perfectos, una pequeña variación de sus especificaciones causales puede producir pequeños grados de desajuste estadístico, cuando el grado de mala especificación real sigue siendo enorme. De este modo, confiar en índices incrementales de ajuste (CFI, IFI, RMSEA, etc.) no tiene sentido, porque no sabemos si los valores que se aconsejan de esos índices están relacionados con una correcta especificación, y además no hay ningún criterio estadístico para establecer un punto de corte.**

### **3. El uso de los intervalos de confianza**

Muchos investigadores emplean el RMSEA y sus intervalos de confianza como criterio de ajuste. Hayduk utiliza el argumento anterior para criticarlo. **Como no hay una relación directa entre el grado de mala especificación y el grado de ajuste incremental o aproximado que provee el RMSEA, el uso de los intervalos de confianza es innecesario.**

Es muy diferente el caso del test de parámetros o tamaños de efecto, donde sí hay una correspondencia entre el resultado de los intervalos de confianza y el valor del parámetro desde el punto de vista estadístico.

### **4. Potencia estadística**

La potencia es un atributo deseable porque nos permite detectar con más seguridad desviaciones de la hipótesis nula en los test de parámetros, por ejemplo correlaciones, diferencia de medias, etc.

La potencia estadística debe evaluarse en conjunción con el tamaño del efecto encontrado, lo que es un fundamento básico

de la interpretación de los análisis estadísticos, con el fin de evitar la confusión entre efectos estadísticamente significativos y sustancialmente significativos.

Sin embargo, en SEM las cosas son diferentes. Bien es cierto que al aumentar el tamaño de la muestra se pueden detectar más fácilmente modelos que fallan, **pero de nuevo tenemos que admitir que modelos que se ajustan pueden estar mal especificados, y que modelos que fallan trivialmente pueden estar gravemente mal especificados.**

Tamaños de muestra altos son deseables en SEM, al igual que en cualquier metodología. La chi-cuadrado se calcula así:

$$\text{Chi-cuadrado} = N * F_0$$

siendo  $N$  el tamaño de la muestra  $-1$ , y  $F_0$  el resultado de la función de discrepancia o el grado en que los residuos son importantes al comparar la matriz implicada por el modelo y la de datos observados.

Pero en modelos correctamente especificados  $F_0 = df/N$ , siendo  $df$  los grados de libertad, por lo que da igual lo grande que sea  $N$ , ya que se cancela. Es decir, **si el modelo es correcto, aunque tengamos un elevado tamaño muestral no vamos a poder detectar desviaciones de la hipótesis nula.**

## 5. Todos los modelos son falsos

Hayduk hace referencia a este mantra que, aunque no lo especifica en el artículo, está muy extendido en el área de la dinámica de sistemas. Es cierto que los modelos son aproximaciones a la realidad, pero **usando SEM se pueden especificar restricciones causales en aspectos que no conocemos bien de los modelos. Es decir, todo se puede modelar.**

## 6. Pescar los p-valores

Emplear el clásico valor de corte de 0.05 para los p-valores

significa que alrededor de 1 de cada 20 test estadísticos reportarán un resultado significativo cuando el efecto sea realmente cero. Este estándar se admite en ciencia, como la probabilidad de cometer un error Tipo I.

Hayduk sostiene que los investigadores tienen que ser cautos a la hora de re-especificar sus modelos en base a las mejoras potenciales de ajuste. Esas modificaciones tienen siempre que estar ligadas por la teoría subyacente. **Recordemos que SEM es un método para contrastar teorías, por lo que si se acometen nuevas especificaciones en aras de buscar un mejor ajuste de los datos, estas deben testarse con nuevos datos, es decir, replicarse.**

Otro elemento importante en relación al p-valor es el punto de corte. **Santificar el 0.05 no tienen ningún sentido científico**, ya que es una mera convención. Los autores deben hacer el mismo esfuerzo por evaluar sus modelos con p-valores de 0.04, 0.06, 0.08 o 0.03. La idea es siempre la misma, realizar una evaluación crítica del modelo propuesto, que evidentemente tendrá connotaciones diferentes cuando no existe el ajuste por la chi-cuadrado. Cuando se está en la frontera del ajuste, ya esté por encima o por debajo de 0.05, la interpretación debe ser similar.

## **7. Asunciones estadísticas**

**El test de la chi-cuadrado es sensible a las desviaciones de normalidad multivariante.** Pero existen correcciones, como el test escalado de Satorra-Bentler, que pueden permitir seguir confiando en el test en situaciones de desviación de normalidad. No obstante, Hayduk no comenta casos de desviaciones importantes o de uso de datos categóricos, donde existen otros procedimientos de estimación que requieren de tamaños de muestra necesariamente elevados.

## **8. Sesgo editorial y honestidad de los investigadores**

Hay un gran problema en la ciencia actual relacionado con la

**creencia de que sólo los resultados que encuentran efectos tienen interés científico.** En el caso de parámetros sería encontrar un efecto significativo, y en el caso de ecuaciones estructurales sería que el modelo se ajustara. Esta tremenda tontería hace que haya un sesgo de publicaciones y que un número considerable de investigadores no sean honestos con sus datos y modelos, cometan fraude y manipulen.

**Los modelos teóricamente bien contruidos, aunque no se ajusten, son de interés científico, y los editores y revisores no deberían desdeñarlos.** Así la ciencia avanzaría de una manera mucho más rápida. Es mucho más informativo y útil testar una teoría contra los datos empíricos y que no sea consistente, que artificialmente forzar el ajuste de la teoría a los datos para que se cumplan los objetivos de la investigación. Esto hace que se generen cientos y cientos de artículos todos los años sin ningún interés científico, ni utilidad práctica alguna, embolicando mucho más todavía los marcos teóricos para las investigaciones subsiguientes.

Hayduk ilustra con una anécdota bien conocida entre los que seguimos este mundo de la ecuaciones estructurales el porqué algunos “padres” de los programas informáticos más usados introdujeron los índices de ajuste aproximado, Karl Joreskog y Dag Sorbom:

*“Another one [anecdote] is from about 1985 when we gave a LISREL workshop. At that time, when use of the new methodology was not widespread at all, there were many LISREL applications producing large or huge chi-squares. We had just added GFI and AGFI to the program. In his lecture Karl would say that the chi-square is really all you need. One participant then asked “Why have you then added GFI?” Whereupon Karl answered “Well, users threaten us saying they would stop using LISREL if it always produces such large chi-squares. So we had to invent something to make people happy. GFI serves that purpose.”*”

## Conclusión/Comentarios

Leslie A. Hayduk lleva toda su vida investigando en el ámbito de los modelos de ecuaciones estructurales. Su libros de 1987 y 1996 son, a día de hoy, todavía recomendables. Fue un adelantado a su tiempo hablando de tópicos hace 20 o 30 años que hoy se han mostrado como válidos.

Tiene un carácter ciertamente agrio en ocasiones, seguirlo a través de SEMNET es un verdadero desafío porque las luchas dialécticas son brutales. No obstante, una revisión calmada y crítica de toda su obra es esclarecedora. Y obviamente no está sólo en esos postulados, existen otros bien reputados investigadores que lo secundan.

**La defensa del test de la chi-cuadrado es un elemento fundamental en SEM. Los investigadores que no respetan la evidencia estadística de este test están realizando una mala praxis, porque los índices incrementales o aproximados no están relacionados necesariamente con el grado de mala especificación, y porque los valores de corte son arbitrarios y han cambiado a lo largo de los años.** Obviamente el test de la chi-cuadrado no es perfecto, porque no es capaz de discriminar entre modelos equivalentes, pero es ahí donde la teoría sustantiva se convierte en fundamental.

Lo que Hayduk defiende es **ser honesto con los datos empíricos y con el proceso de modelización, reportar cuando hay inconsistencias entre los datos observables y el modelo propuesto, y profundizar en los modelos para entender las causas por las que fallan.** Propone un uso sincero y responsable de SEM, admitiendo las debilidades que tiene, pero instando a los investigadores, revisores y editores a que abracen de una vez el paradigma del ajuste exacto, como la única vía válida para evaluar modelos.

Al fin y al cabo no es más que reclamar que los investigadores mimen todo el proceso de investigación, desde el planteamiento

inicial, la recogida de datos, hasta la forma de tratar esos datos estadísticamente, y no se dejen llevar por esa equivocada tendencia de buscar la significación estadística a toda costa, o la aproximación al ajuste del modelo empleando índices que fueron creados de manera espuria para que los investigadores manejaran el software comercial.

Bajo mi experiencia personal, es posible obtener modelos que se ajusten vía chi-cuadrado (tengo publicados artículos así), y creo que ese es el camino correcto. Al comienzo de mi carrera como investigador publiqué un par de artículos empleando el ajuste aproximado, debido a los equivocados consejos de algunos maestros que tuve. Pero en cuanto me interesé personalmente por este tema, investigué por mi cuenta, y leí la bibliografía más relevante, me di cuenta de que confiar en índices como CFI, TLI, RMSEA, IFI, GFI, etc. es una lotería, algo que incrementa ostensiblemente el riesgo de que mis investigaciones no sirven para absolutamente nada. **La chi-cuadrado no es perfecta, de nuevo hay que recalcarlo, pero es el único paso consistente para intentar plantear modelos útiles y tratar de entender la complejidad de los fenómenos bajo estudio.**

Si no se está dispuesto a ello, creo que lo más adecuado es no emplear esta metodología, y explorar con otro tipo de métodos no paramétricos, tratando de considerar relaciones complejas y no lineales, y yendo paso a paso y caminando despacio. Pero si se usa SEM, hay que hacerlo de manera honesta. Si no se hace así, se estará perdiendo el tiempo, el dinero, y la reputación.

Hayduk, L. A. (2014). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. BMC Medical Research Methodology, 14, 124.

Indicadores de calidad de la revista\*

|                       | Impact Factor (2014) | Cuartil   | Categoría                       |
|-----------------------|----------------------|-----------|---------------------------------|
| Thomson-Reuters (JCR) | <b>2.27</b>          | <b>Q2</b> | HEALTH CARE SCIENCES & SERVICES |
| Scimago (SJR)         | <b>1.28</b>          | <b>Q2</b> | EPIDEMIOLOGY                    |

*\* Es simplemente un indicador aproximado para valorar la calidad de la publicación*

**Todos los posts relacionados**