

## Evaluación de programas de formación continua en contextos no estandarizados: complementariedad entre Análisis Factorial y Multinivel para la obtención de evidencias de validez de constructo

Francisco P. Holgado-Tello<sup>1</sup>, Salvador Chacón-MoscOSO<sup>4,2</sup>, Enrique Vila-Abad<sup>1</sup>, Begoña Delgado<sup>3</sup>, Susana Sanduvete-Chaves<sup>4</sup> e Isabel Barbero-García<sup>1</sup>

*1 Metodología de las Ciencias del Comportamiento. Facultad de Psicología. UNED*

*2 Universidad Autónoma de Chile, Chile*

*3 Psicología Evolutiva y de la Educación, UNED*

*4 Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Sevilla*

**Resumen:** La evaluación de programas se aplica frecuentemente en ámbitos de intervención no estandarizados. Esto conlleva, entre otras, las carencias de: a) modelo teórico validado previamente; b) instrumentos de medida estándares; c) fiabilidad de las medidas. En este trabajo, se plantea que el Análisis Factorial con correlaciones policóricas y el Análisis Multinivel puede ser un procedimiento adecuado hacia el logro de la validez de constructo en contextos no estandarizados de evaluación donde, además, las variables suelen ser no cuantitativas y estar anidadas. El estudio empírico se realiza sobre una muestra de 2754 trabajadores de la Universidad de Sevilla que han respondido a una encuesta de satisfacción elaborada ad-hoc sobre la formación recibida en distintos cursos encaminados a capacitarlos para el correcto desempeño de sus funciones. Cabe destacar la complementariedad entre ambas técnicas de análisis para examinar la variabilidad diferencial aportada por variables explicativas de distinto nivel jerárquico en la predicción de la satisfacción percibida.

**Palabras claves:** Validez; evaluación de programas; satisfacción; Análisis Factorial; Análisis Multinivel.

**Title:** Training program evaluation in non-standardized context: Complementarity across Factorial and Multilevel Analysis to obtain construct validity evidences.

**Abstract:** Program evaluation is usually applied to non-standardized intervention contexts. This implies, among others, deficiencies of: a) validated theoretical models; b) non-standard measurement instruments; c) reliable measures. In this work, we show that Factor Analysis with polychoric correlations and Multilevel Analysis could be an adequate procedure to gain construct validity evidence in non-standard evaluative contexts, where the measures are not quantitative and usually are nested. The empirical study is carried out on a sample of 2754 workers of the University of Sevilla. They have completed a satisfaction questionnaire about training courses aimed to prepare them for the correct performance of their jobs. We highlight the complementarities between both analytical techniques to study the differential variability provided by explained variables nested in different hierarchical level to predict the perceived satisfaction.

**Key words:** Validity; program evaluation; satisfaction; Factor Analysis; Multilevel Analysis.

### Introducción

En términos generales, un programa de intervención es un conjunto específico de acciones y recursos diseñados e implantados organizadamente en una situación social para resolver algún problema que atañe a una serie de personas (Mejías y Huaccho, 2011). Un programa de intervención representa un modelo en el que, tras haberse estudiado el objeto de intervención, se han de clarificar las posibles relaciones existentes entre un conjunto de actividades y el logro de unos resultados, considerando unos medios disponibles (tanto humanos como materiales), según el punto de vista de las personas implicadas (ya sean políticos, profesionales, usuarios, o un conjunto de ellos), en el contexto de intervención y dentro del marco de algún planteamiento teórico, a ser posible explícito (Herman, Morris y Fitz-Gibbon, 1987; Rossi y Freeman, 1985; Sarkis, Gonzalez y Adenso 2010).

A pesar de la aparente simplicidad que conlleva la evaluación de un programa de intervención, en la práctica profesional nos encontramos con que los contextos de trabajo son inestables y sometidos a un cambio continuo debido a las interacciones de los sujetos entre sí y con su entorno (Anguera, 2008; Chacón-MoscOSO, Sanduvete-Chaves, Portell-Vidal y Anguera, 2013). El carácter inestable de dicho contexto implica una serie de condicionantes sobre los pro-

gramas de intervención, que podrían resumirse en: a) una delimitación interna y de funcionamiento de los programas fuertemente condicionada por prioridades de orden político que, a su vez, amortiguan las variaciones en el *status quo* y condicionan el uso posterior de los resultados (Campbell, 1982; Cook, 1981; Cook, Leviton y Shadish 1985; House, 1993, 1994; House y Shull, 1988; Palumbo y Nachmias, 1983); b) una gran dispersión y heterogeneidad en la implementación de las intervenciones concretas -tanto a nivel profesional como de los destinatarios del programa- (Yin y Schmeidler, 2009); y c) la incidencia de una gran diversidad de elementos en interacción en los programas de intervención provocando un cambio continuo (Sechrest y Figueredo, 1993).

Estas circunstancias hacen que difícilmente se disponga de modelos teóricos validados o instrumentos de evaluación estandarizados. Dichos instrumentos se suelen diseñar a posteriori, dependiendo de los contextos de intervención particulares y, por lo general, se trata de escalas de clasificación que demandan respuestas categóricas (cualitativas u ordinales) (Anguera, Chacón-MoscOSO y Blanco, 2008; Holgado-Tello y Barbero-García, 2008).

En estos casos, tanto el Análisis Factorial como el modelo de regresión, usualmente utilizados para obtener evidencias de la validez de constructo de los modelos teóricos que justifican el diseño del programa, están condicionados por la métrica y estructura o jerarquización de los datos. A pesar de ello, se suele dar un tratamiento continuo-cuantitativo a datos categóricos, sin tener a su vez en cuenta su posible carác-

**\* Dirección para correspondencia [Correspondence address]:**

Fco. Pablo Holgado Tello. Dpto. Metodología de las CC. del Comportamiento. Facultad de Psicología. UNED. c/ Juan del Rosal, nº 10. 28040 Madrid (España). E-mail: [pfholgado@psi.uned.es](mailto:pfholgado@psi.uned.es)

ter jerárquico (Barbero-García, Vila-Abad y Holgado-Tello, 2011; Ferrando y Anguiano, 2010).

El problema que se plantea con estos tratamientos es que categorizar supuestas variables continuas mediante escalas ordinales puede ocasionar problemas de estimación al verse *atenuadas* las correlaciones entre las mismas. Por este motivo, tanto si se utiliza el Análisis Factorial Exploratorio (AFE) como el Análisis Factorial Confirmatorio (AFC) para analizar la relación entre dichas variables, es más adecuado partir de la matriz de correlaciones policóricas, basada en la distribución bivariada de los ítems, en lugar de la de Pearson (Holgado-Tello, Chacón-MoscOSO, Barbero-García y Vila-Abad, 2010; Jöreskog, 2001; Lee, Zhang y Edwards, 2012; Yang, Jöreskog y Luo, 2010).

Por otro lado, para evaluar la incidencia diferencial de variables concretas (parámetros de regresión obtenidos en los modelos jerárquicos) sobre la eficacia de un determinado programa, los modelos lineales jerárquicos logran ser de gran utilidad puesto que pueden proporcionar indicios acerca de la contribución relativa de cada una de las dimensiones teóricas propuestas (Kreft y Leeuw, 1998). Si no se contempla el carácter anidado de los datos, los parámetros de regresión estimados a partir de la muestra total de sujetos pueden estar funcionando ineficazmente al ser aplicados a cada uno de los subgrupos que la componen (Du Toit, Du Toit y Cudeck, 1999).

En síntesis, debido al carácter ordinal y anidado de los datos que suelen caracterizar a los programas de intervención implementados en contextos inestables, el Análisis Factorial (tanto exploratorio como confirmatorio) mediante correlaciones policóricas y el Análisis Multinivel se complementan para obtener evidencias de validez de constructo de la operativización adoptada de dichos programas.

El estudio empírico que se presenta se desarrolló en la Unidad de Evaluación del Centro de Formación y Perfeccionamiento del Personal de Administración y Servicios (PAS) de la Universidad de Sevilla, cuyo cometido es planificar y evaluar sistemáticamente el Plan de Formación para el personal. El objetivo principal de este trabajo fue obtener evidencias de validez de constructo sobre la operativización adoptada acerca del concepto *satisfacción con la formación*, en un contexto de intervención inestable y sometido a las demandas tanto de los responsables del Centro de Formación como de sus usuarios. Entendemos que dichas evidencias vienen a partir de que exista coherencia teórica, que los datos se correspondan con el modelo elegido y que las consecuencias de la medida sean útiles para los responsables del Centro de Formación. Para ello, utilizamos conjuntamente Análisis Factorial y Multinivel, dado que cabe esperar que ambas técnicas analíticas aporten información relevante, dadas las características planteadas anteriormente, para el estudio de la validez de variables psicológicas relevantes utilizadas en programas de intervención implementados en contextos inestables. En el campo de la evaluación de la formación, el modelo teórico utilizado para la operativización efectiva de variables es el denominado de los Cuatro Niveles planteado por

Kirkpatrick (1999). Desarrolla un sistema de trabajo que trata de obtener evidencias empíricas sobre: a) el grado de satisfacción de los participantes con el programa o reacción de los mismos; b) los conocimientos, habilidades o actitudes aprendidas mediante la formación; c) si los participantes han cambiado su conducta en el puesto de trabajo basándose en lo aprendido durante la formación; y d) si dichos cambios han afectado positivamente a la organización (Rajeev, Madan y Jayarajan, 2009).

Evaluar la reacción supone valorar el nivel de satisfacción de los participantes con la formación recibida. Para ello, se suelen usar los cuestionarios de satisfacción que habitualmente incluyen aspectos sobre los materiales empleados, los formadores, la metodología, o los contenidos, por ejemplo. El problema con este tipo de evaluación se liga a la expresión que denomina a los cuestionarios de satisfacción como *hojas de felicidad*; es decir, se plantea la escasa sensibilidad de esta medida para detectar aspectos a mejorar, a no ser que se trate de elementos muy salientes (positivos o negativos) del programa objeto de evaluación (Thayer, 1991). Evaluar en este nivel ayuda a recabar información que servirá para mejorar futuras programaciones; pero sobre todo, las respuestas a los cuestionarios de satisfacción servirán para avalar decisiones sobre los programas de formación, en contraposición con las opiniones y sugerencias de unos pocos participantes muy satisfechos, o por el contrario muy descontentos, que simplemente hayan hecho oír su voz.

La evaluación de la reacción, por ser la más fácil de llevar a cabo es la que se realiza con mayor frecuencia, y precisamente por ello es necesario cuestionarse la calidad psicométrica con la que se está realizando esta medición (Ventosa, 1998), así como las implicaciones relativas de cada dimensión evaluada en el proceso de toma de decisiones.

Teniendo en cuenta que la elaboración y validación de instrumentos de evaluación supone la operacionalización del modelo teórico subyacente al programa de intervención, se ha decidido utilizar este trabajo por un doble motivo: en primer lugar, por haber participado desde el principio en la elaboración del instrumento de medida, con el consecuente conocimiento del modelo teórico subyacente; y en segundo lugar, porque, al igual que en otros ámbitos de intervención, se utilizan escalas tipo Likert de satisfacción de los usuarios como medida de evaluación elaboradas ad-hoc.

En lo que concierne al tratamiento de los datos, el contexto está influyendo en las características de los mismos. En nuestro estudio, los datos son ordinales ya que se ha usado una escala Likert para medir la satisfacción; y están anidados en variables de orden superior, como es el área temática de la acción formativa y el tipo de formador. Ante estas circunstancias, las técnicas analíticas utilizadas para la obtención de evidencias de validez de constructo han de contemplar dichos aspectos; y en este sentido, el Análisis Factorial utilizando correlaciones policóricas y el Análisis Multinivel se adecuan a esta casuística, además de complementar mutuamente la información que proporcionan.

## Método

### Participantes

La muestra fue intencional y constituida por 2754 sujetos pertenecientes al PAS, incluyendo trabajadores fijos y eventuales, que habían participado en acciones formativas del Centro de Formación y Perfeccionamiento de la Universidad de Sevilla. Para obtener el mayor número posible de respuestas, se garantizó el anonimato lo que supuso no recabar información sobre otras variables sociodemográficas.

Los datos se obtuvieron a partir de 79 cursos de formación, cada uno de los cuales contemplaba varias ediciones. La duración media de las acciones formativas era de 20 horas y su contenido ampliamente variado (normativas legales, servicios deportivos, bibliotecas, calidad en la gestión, formación de formadores o gestión financiera, por ejemplo). Cada participante aportó un sólo dato en cada curso relacionado con su trabajo.

### Instrumentos

El instrumento utilizado se enmarca dentro del nivel de la evaluación de la reacción de los participantes de acuerdo con el modelo de los Cuatro Niveles de Kirkpatrick (1999). Ha sido elaborado a partir de la revisión de otras pruebas y medidas de satisfacción empleadas por distintas unidades de formación (e.g., Fernández y Ovejero, 1994; Medina, 1996), así como a partir de las aportaciones de los responsables técnicos y políticos del Centro.

De una versión inicial de 72 ítems, se obtuvo un cuestionario final de 12 ítems tipo Likert (1 = totalmente en desacuerdo; 5 = totalmente de acuerdo), que se agrupan en tres dimensiones: a) Objetivos y contenidos (ítems 1-3); b) Metodología y ambiente (ítems 4-9); y c) Utilidad y valoración global (ítems 10-12). El contenido concreto de los ítems puede verse más adelante en la tabla 2. En la muestra utilizada, la escala total obtiene un coeficiente alfa de Cronbach de .888 puntos; y el índice de discriminación medio de los ítems es .674. La dimensión sobre Objetivos y Contenidos obtiene un coeficiente alfa de .78, y una discriminación media de sus ítems de .61; la dimensión de Metodología, un alfa de .78 y una discriminación media igual a .54; y la dimensión sobre Utilidad, un coeficiente alpha de .79 y una discriminación media igual a .63. Todos estos resultados indican que, desde un punto de vista psicométrico, tanto los ítems como las dimensiones propuestas teóricamente se comportan adecuadamente.

La aplicación informática utilizada para el análisis y tratamiento de los datos ha sido el programa LISREL 8.54 (Jöreskog y Sörbom, 2003) y PRELIS 2.30 (Jöreskog y Sörbom, 1999).

### Análisis de datos y variables

En primer lugar, se dividió aleatoriamente la muestra total en dos submuestras de igual tamaño (A y B). La A se utilizó para realizar el AFE, mientras que con la B se ejecutó un AFC.

Una vez estimada la matriz de correlaciones policóricas, se comprobó el supuesto de normalidad bivariada, ya que en caso contrario no estaría justificada su utilización. Para ello, siguiendo a Jöreskog (2001), se tuvo en cuenta el porcentaje de ítems cuyo *root mean square error of approximation (RMSEA)* era superior a .1 para cada par de correlaciones. A continuación, se realizó un AFE mediante PRELIS 2.30 utilizando como método de estimación MINRES, dado el carácter ordinal de los datos y su distribución asimétrica (Jöreskog, 2003); y una modalidad oblicua de rotación (promax), debido a la hipotética asociación entre los factores. Del AFE se excluyó el ítem 12 dado que posteriormente se utiliza como variable criterio. A continuación, en la submuestra B, el modelo resultante se sometió a prueba mediante un AFC usando mínimos cuadrados ponderados como método de estimación (Morata y Holgado-Tello, 2013; Yang et al., 2010).

Finalmente, se ejecutó un Análisis Multinivel utilizando LISREL 2.54. En dicho análisis, el nivel 1 está representado por los sujetos y el nivel 2 por los cursos a los que éstos asisten (79 en total). Como variable de respuesta (dependiente), hemos tomado el ítem 12 de la encuesta, referido a la valoración global de satisfacción con la acción formativa. Como variables predictoras de primer nivel (individuales), hemos considerado la mediana de cada una de las dimensiones de la encuesta de satisfacción; es decir, la mediana de objetivos y contenidos (MDX1-X3); de metodología y ambiente (MDX4-X9); y de utilidad de la acción formativa (MDX10X11).

Como variables predictoras de segundo nivel, o contextuales dadas por los cursos, hemos utilizado la mediana de la satisfacción con el profesorado (SATIPROM) en cada acción formativa; y el tipo de contenido de la acción formativa (COTENCUR), que toma doce valores (calidad en la gestión; modernización de la gestión; promoción; seguridad y salud; formación de formadores; herramientas ofimáticas; actualización de normativas; cursos para personal informático; cursos para personal de servicios deportivos; cursos para personal de biblioteca; cursos para personal de mantenimiento; y cursos para personal de servicios audiovisuales).

Los supuestos del Modelo Multinivel son similares a los de cualquier análisis de regresión, o modelo ANOVA. Es decir, linealidad, normalidad, homocedasticidad e independencia de las observaciones. En este tipo de análisis, las dos últimas presentan ciertas peculiaridades dado que los sujetos pertenecientes al mismo grupo son más similares entre sí, que comparados con los sujetos de otros grupos, y por tanto el supuesto de independencia no se satisface en cada nivel de la jerarquía. Sin embargo, esta cuestión queda superada al incluirse en el análisis la correlación intraclase que precisamente sirve para cuantificar cuál es la variabilidad intra-grupo.

## Resultados

Si nos centramos en la distribución bivariada de los ítems, podemos asumir normalidad bivariada ya que sólo 2 (3%) de las 66 correlaciones tienen *RMSEA* superior a .1 (nivel de significación del 5%) (Joreskog, 2001).

## Análisis Factorial

En la tabla 1, se presenta la matriz de correlaciones policóricas obtenidas en la submuestra A entre todos los ítems que componen la escala, excepto el 12 que se toma como variable criterio al ser de valoración global.

Tabla 1. Matriz de correlaciones policóricas.

	Obj1	Obj2	Obj3	Met4	Met5	Met6	Met7	Met8	Met9	Val10	Val11
Obj1	---										
Obj2	.652	---									
Obj3	.574	.519	---								
Met4	.705	.636	.538	---							
Met5	.556	.518	.434	.681	---						
Met6	.333	.262	.326	.392	.502	---					
Met7	.625	.534	.419	.677	.666	.500	---				
Met8	.519	.491	.345	.570	.439	.303	.576	---			
Met9	.300	.235	.151	.371	.376	.152	.378	.474	---		
Val10	.492	.380	.330	.479	.420	.316	.503	.421	.338	---	
Val11	.503	.448	.312	.537	.482	.413	.558	.471	.386	.665	---

Una vez factorizada la matriz anterior, se obtiene una estructura trifactorial. De ellos, el primer factor explica un 44.71% de la varianza, el segundo un 9.83%, y el tercero un 8.52%. A continuación, en la Tabla 2 se presentan las saturaciones factoriales de la solución rotada de cada uno de los

ítems. A pesar de que se podría considerar que la escala es unidimensional debido al pequeño porcentaje de varianza explicada por los dos últimos factores, se consideraron tres factores por cuestiones teóricas (previamente el instrumento se había diseñado contemplando tres dimensiones).

Tabla 2. Medias (*M*) y desviaciones típicas (*DT*). Pesos factoriales de la solución rotada (*AFE*). Solución completamente estandarizada (*AFC*) entre paréntesis y en cursiva.

	Ítem	<i>M(DT)</i>	<i>F1</i>	<i>F2</i>	<i>F3</i>
Objetivos y contenidos	1. A su juicio se han alcanzado los objetivos previstos (OBJ1)	<b>3.88(.95)</b>	<b>.799 (.86)</b>	-.027	.123
	2. Los temas se han tratado con la profundidad que permitía la duración del curso (OBJ2)	<b>3.93(1.04)</b>	<b>.724 (.81)</b>	-.054	.116
	3. La duración del curso ha sido adecuada a los objetivos y contenidos (OBJ3)	<b>3.18(1.27)</b>	<b>.660 (.68)</b>	.128	-.119
Metodología y ambiente	4. La metodología se ha adecuado a los objetivos y contenidos (MET4)	<b>4.01(.95)</b>	<b>.599 (.65)</b>	.171	.201
	5. La metodología ha permitido la participación activa de los asistentes (MET5)	4.12(1.03)	.276	<b>.509 (.86)</b>	.137
	6. Esta acción formativa me ha facilitado compartir experiencias profesionales con otros compañeros (MET6)	3.66(1.22)	-.069	<b>.706 (.66)</b>	.052
	7. La acción formativa ha sido realista y práctica (MET7)	3.99(1.04)	.258	<b>.392 (.89)</b>	.324
	8. La documentación entregada ha sido de calidad (MET8)	4.11(1.03)	.293	-.019	<b>.492 (.55)</b>
	9. Las condiciones ambientales han sido adecuadas para facilitar el proceso formativo (MET9)	4.16(1.04)	-.002	-.035	<b>.568 (.58)</b>
Utilidad y valoración global	10. La acción formativa recibida es útil para mi puesto de trabajo (VAL10)	4.25(1.00)	.049	.041	<b>.657 (.78)</b>
	11. La acción formativa recibida es útil para mi formación personal (VAL11)	4.33(.90)	-.028	.132	<b>.740 (.86)</b>
<b>F2</b>			.581	---	
<b>F3</b>			.610	.550	---

En el primer factor, los ítems que presentan mayor peso factorial se refieren a los objetivos y contenidos, siendo el ítem 1 (*a su juicio se han alcanzado los objetivos previstos*) el de mayor saturación (.799). El segundo factor está principalmente compuesto por los elementos referidos a la metodología y ambiente; en este caso el ítem 6 (*esta acción formativa me ha facilitado compartir experiencias profesionales con otros compañeros*) es el que más aporta al factor (.706). Por último, en el tercer factor aparecen ítem relacionados con la utilidad y la valoración

global siendo el 11 (*la acción formativa recibida es útil para mi formación personal*) el más representativo (.704).

En total, resultaron tres factores que guardan una estructura similar a la propuesta en la encuesta de satisfacción. La matriz de correlaciones entre los factores aparece al final de la Tabla 2.

En la submuestra B, el modelo trifactorial obtenido en la submuestra A se sometió a un *AFC*. Los índices globales de ajuste obtenidos son  $\chi^2_{(df=41)} = 232.54$  ( $p < .0001$ ); *RMSEA*

= .06; *GFI* = .98; *AGFI* = .98; *CFI* = .94; y *NNFI* = .92. Dado que los índices globales de ajuste presentan valores adecuados, podríamos concluir que el modelo representa razonablemente bien la forma en que se comportan los datos. En la Tabla 2, bajo las saturaciones de la solución obtenida en el *AFE*, se presenta la solución completamente estandarizada de los parámetros lambda (en cursiva y entre paréntesis).

En síntesis, se observa que la respuesta de los sujetos se ajusta, aunque con matices, al planteamiento utilizado en la elaboración del instrumento. Dicho planteamiento, como ya hemos apuntado, vino dado por una exhaustiva revisión de otras encuestas implementadas en distintas instituciones, por las aportaciones realizadas por responsables del Centro de Formación, y por el examen de los aspectos teóricos relacionados con las dimensiones implicadas en los procesos de formación (Kirkpatrick, 1999; Passmore y Velez, 2012; Pineda, 2010).

La convergencia de ambas soluciones factoriales supone un buen indicador de la coherencia teoría-dato, que es considerada una de las piedras angulares de la validación de constructo (Gadermann, Guhn y Zumbo, 2011). Además, el uso de la validación cruzada, que evita el efecto de la capitalización del azar, permite optimizar la generalización del modelo bajo estudio (Wilkins, 2010).

### Análisis Multinivel

Por otro lado, al considerar la estructura anidada que conforman los datos, es decir, características de sujetos y a su vez éstos dentro de acciones formativas particulares, se pueden complementar los resultados del Análisis Factorial utilizando correlaciones policóricas con los del Análisis Multinivel (Kreft y Leeuw, 1998). Se pretende determinar hasta qué punto están influyendo en los niveles de satisfacción tanto las características individuales de los sujetos, como las características que definen las acciones formativas a los que éstos asisten (en nuestro caso, y de acuerdo con las demandas del Centro de Formación, tipo de formador, y área temática del curso).

El objetivo de este análisis es estudiar si las dimensiones planteadas en la encuesta, de las que se obtuvieron evidencias de validez factorial en el análisis previo, pueden predecir, y en qué grado, la satisfacción global de los asistentes con las acciones formativas (nivel 1). Por otro lado, se examinará si existen diferencias entre los cursos en cuanto a dichas dimensiones, y si fuera el caso, intentar explicarlas a través de los predictores de segundo nivel (nivel 2). La pregunta a la que intentamos dar respuesta es si existe la misma relación entre la satisfacción global y los ítems a través de los cursos y, si es así, analizar si los predictores de cambio operativizados (tipo de profesor y contenido de los cursos) se relacionan significativamente con los efectos. Ello permitirá mejorar el proceso de toma de decisiones como una consecuencia deseable de la medida de la satisfacción.

En un principio, se pudo haber optado por incluir en el modelo tantas variables predictoras de primer nivel como ítems de satisfacción; y a pesar de que este modelo hubiera parecido más realista, sin embargo hubiera perdido parsimonia, ya que la selección de un reducido número de variables explicativas es esencial en el éxito de la modelización multinivel (Kreft y Leeuw, 1998). Esta decisión está avalada por el Análisis Factorial realizado anteriormente, donde se obtuvieron evidencias de validez de constructo sobre el modelado teórico en tres dimensiones.

El modelo multinivel más simple, conocido también como modelo nulo o incondicional, es equivalente al análisis de la varianza de un factor, y se caracteriza porque sólo varía aleatoriamente la ordenada en el origen. A pesar de no ser un modelo interesante por sí mismo, sin embargo es de gran utilidad porque proporciona una importante información sobre la variabilidad de los resultados en cada uno de los niveles de la jerarquía. Además, supone el referente respecto al que comparar el ajuste de los modelos más complejos que se vayan probando. Dicho ajuste progresivo se mide a través de la diferencia en el valor de la función de máxima verosimilitud conocida como *-2 log-verosimilitud* que, para modelos basados en los mismos datos, tiene una distribución de *Chi Cuadrado*, siendo los grados de libertad igual a la diferencia en el número de parámetros (coeficientes de regresión en este caso) estimados en los sucesivos modelos (Jöreskog, Sörbom, Du Toit y Du Toit, 1999).

Los resultados del modelo nulo sobre nuestros datos se presentan en la tabla 3. Cuya función de verosimilitud obtuvo un valor de 6616.143.

**Tabla 3.** Significación de parámetros del modelo nulo.

<i>Coefficientes</i>	<i>Parámetro</i>	<i>Error típico</i>	<i>Valor de <math>\chi^2</math></i>	<i>p &gt;  <math>\chi^2</math> </i>
<b>Parte fija</b>				
Constante	$\beta = 4.1813$	.0613	68.2336	.0000
<b>Parte aleatoria</b>				
<b>Nivel 2,</b>	$\tau = .1501$	.0366	4.0962	.00004
Const./Const.				
<b>Nivel 1,</b>	$\tau = .6499$	.0178	36.4599	.0000
Const./Const.				

En estos resultados, la constante representa la ordenada en el origen y ha sido especificada como fija en el nivel 1, así como aleatoria en los niveles 1 y 2; es decir, puede variar aleatoriamente tanto entre sujetos como entre cursos. Ello nos permite observar que la variación entre los sujetos (nivel 1) es mayor que entre los cursos (nivel 2). En este sentido, la también llamada correlación intraclase ( $\rho = \Phi_{(2)} / \Phi_{(2)} + \Phi_{(1)}$ ) informa que la proporción de variación entre grupos es de 18.76 ( $\rho = 0.1501 / (0.1501 + 0.6499)$ ). Lo cual quiere decir que el 18.76% de la variabilidad en los datos está relacionada con las diferencias entre los grupos que conforman los cursos (nivel 2).

A partir de aquí, se fueron probando distintos modelos, introduciendo sucesivamente las variables predictoras del nivel 1 y sus interacciones; posteriormente, se dejaron libres

las predictoras de nivel 1 en el 2; y por último, se fueron incorporando los predictores de nivel 2 (SATIPROM y CONTECUR) para intentar explicar la variabilidad entre los cursos.

El modelo que mejor ajustó se muestra en la Tabla 4 ( $-2 \log\text{-verosimilitud} = 3747.799$ ).

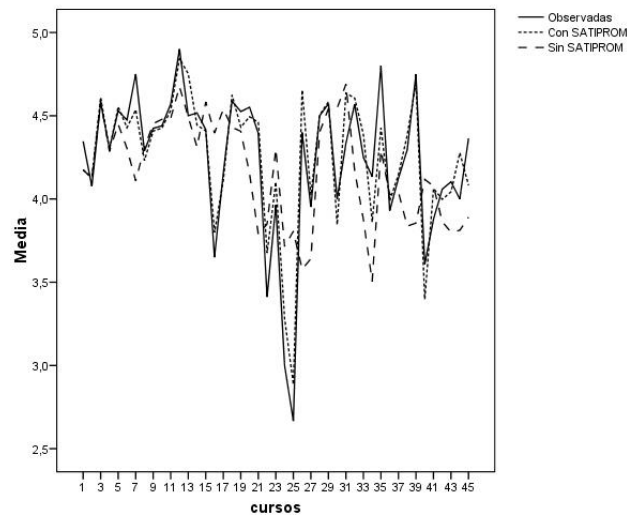
**Tabla 4.** Significación de parámetros del modelo.

Coefficientes	Parámetro	Error típico	Valor de $z$ , $p >  z $	
<b>Parte fija</b>				
Constante	$\beta = -.4955$	.1789	-2.7857	.0053
MDX1-X3	$\beta = .5365$	.0553	9.6959	.0000
MDX4-X9	$\beta = .6292$	.0445	14.1388	.0000
MDX10X11	$\beta = .3249$	.0256	12.6785	.0000
MDX1-X3 *MDX4-X9	$\beta = -.0802$	.0121	-6.6473	.0000
SATIPROM	$\beta = .0016$	.0016	1.0385	.2990
<b>Parte aleatoria, Nivel 2</b>				
Const./Const.	$\tau = .0885$	.0543	1.6306	.1030
MDX1-X3/Const.	$\tau = -.0034$	.0109	-3.138	.7537
MDX1-X3/MDX1-X3	$\tau = .0096$	.0041	2.3468	.0189
MDX10X11/Const.	$\tau = -.0135$	.0131	-1.0266	.3046
MDX10X11/MDX1-X3	$\tau = -.0083$	.0038	-2.1621	.0306
MDX10X11/MDX10X11	$\tau = .0103$	.0048	2.1300	.0033
<b>Parte aleatoria, Nivel 1</b>				
Const./Const.	$\tau = .2781$	.0082	33.7885	.0000

El ajuste global del modelo ha mejorado significativamente con respecto al modelo anterior, en el que no se incluyó ninguna variable de segundo nivel; es decir, de un valor para la función de máxima verosimilitud de 3850.564 pasa a 3747.799 para un grado de libertad que viene dado por el cálculo del coeficiente de SATIPROM (satisfacción con el profesorado en cada acción formativa). Aunque el coeficiente para SATIPROM no haya resultado significativo, sin embargo su inclusión ha supuesto una mejora sustancial en el ajuste del modelo. En este sentido, hemos de tener en cuenta que la interpretación de los efectos necesita basarse en el ajuste global del modelo antes que en el test de los coeficientes simples. Si el ajuste del modelo mejora significativamente, entonces podemos considerar la significación individual del nuevo coeficiente incluido (Kreft y Leeuw, 1998, p. 66). Este modelo final llega a explicar hasta el 58% de la varianza total. A título ilustrativo, mediante la Figura 1 se muestran comparativamente los valores observados y pronosticados para ambos modelos, donde gráficamente se aprecian las diferencias existentes entre las puntuaciones pronosticadas por ambos modelos con respecto a las observadas.

A la vista de los resultados, encontramos que la satisfacción global con las acciones formativas depende en parte de cada una de las dimensiones que componen la encuesta (MDX1-X3; MDX4-X9; MDX10X11), así como de la interacción entre objetivos–contenidos con metodología–ambiente (MDX1-X3\*MDX4-X9) (ver Tabla 4). A su vez, cuando se permite que objetivos–contenidos y la utilidad

puedan variar aleatoriamente entre las distintas acciones formativas, encontramos que dichas dimensiones así como su interacción tienen un efecto distinto a través de los cursos utilizando un nivel de confianza del 95%. Dicho efecto es explicado sólo parcialmente por la satisfacción media con el profesor en cada acción formativa.



**Figura 1.** Comparativa entre puntuaciones observadas en el ítem 12 y predichas por los modelos con SATIPROM y sin SATIPROM.

A título ilustrativo, en la Figura 2 se muestra el esquema del planteamiento multinivel resultante donde se reflejan exclusivamente los efectos principales más relevantes de cara a la interpretación sustantiva de los datos. Los círculos representan las variables de nivel 1 y los rectángulos, las de nivel 2. Las flechas siempre fluyen desde las variables explicativas (MDX1-X3, MDX4-X9 y MDX10X11) hacia la variable de respuesta (satisfacción global), pero siempre en el nivel 1. Dicha relación es interceptada por la variable de nivel 2 (SATIPROM) que puede explicar, muy parcialmente, la influencia diferencial de las variables predictoras objetivos–contenidos y utilidad en los diferentes cursos.

En síntesis, tras haber obtenido evidencias de validez de constructo de la encuesta de satisfacción mediante el Análisis Factorial adaptado a datos ordinales, en el Análisis Multinivel hemos pretendido determinar hasta qué punto las dimensiones planteadas (Objetivos, Metodología, y Utilidad) están sujetas a un funcionamiento diferencial de acuerdo con la jerarquía de los datos. En este sentido hemos encontrado que, tanto los objetivos como la utilidad, varían aleatoriamente entre los distintos cursos, lo cual es explicado sólo parcialmente por las características del formador.

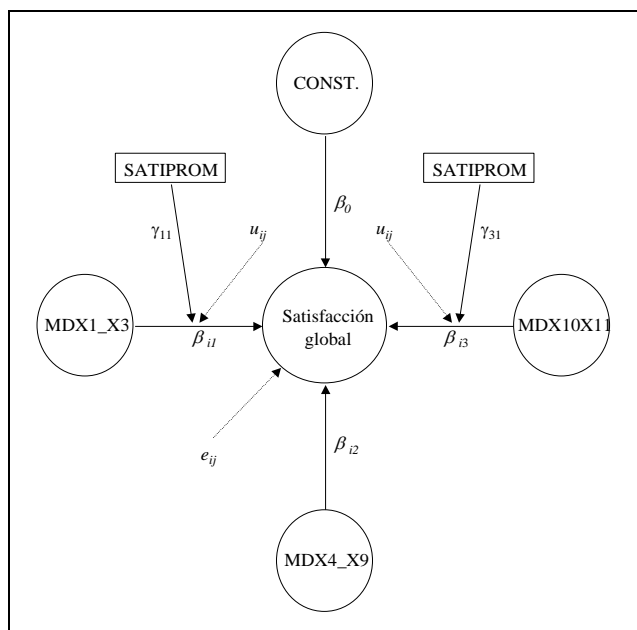


Figura 2. Esquema del modelo multinivel.

## Discusión y conclusiones

La obtención de evidencias de validez de constructo es un aspecto central en cualquier proceso de medición de lo psicológico como es el caso de los programas de evaluación-intervención, en tanto que *facilita* la coherencia teórica, la correspondencia datos-teoría y la utilidad de las consecuencias. Sin embargo, el proceso de obtención de dichas evidencias no es ajeno a la inestabilidad que caracteriza a los contextos de intervención, lo que suele determinar una métrica categórica en los datos que sirven para medir las variables relevantes.

En nuestro caso, y a modo de ejemplo típico, hemos pretendido medir la satisfacción con un programa de formación. Para ello, se elaboró un instrumento a partir de una revisión teórica sobre formación, y desde el examen de otros instrumentos que ya se estaban empleando en este ámbito de intervención. Además, se contó con las aportaciones de los responsables técnicos y políticos del Centro de Formación de la Universidad de Sevilla con el propósito de que dicha medida tuviera consecuencias prácticas. Al igual que en muchas ocasiones de medida de variables psicológicas, los datos se caracterizan por ser ordinales y estar anidados en variables de orden superior; por tanto, dicha casuística ha de ser considerada para la obtención de evidencias de validez de constructo.

Se ha intentado describir el valor añadido que en estos contextos evaluativos puede aportar el Análisis Factorial y el Multinivel. Permiten explicitar aspectos sobre los que incidir de manera más efectiva para potenciar el proceso de mejora continua al realizar análisis basados en el conjunto de los datos y no en datos estudiados parcialmente, y a su vez teniendo en consideración la estructura jerárquica de los mismos.

Así, el Análisis Factorial se puede emplear para el estudio de validez, en el sentido de intentar operativizar los procesos y relaciones entre los distintos factores que están incidiendo en la consecución final de los resultados. De esta forma, se puede intentar representar las hipótesis del programa a evaluar sobre el modelo estructural de relaciones entre las variables latentes y modelos de medida (de esas variables latentes) respecto a indicadores observables. En este sentido, desde el Análisis Factorial se ha esbozado un determinado modelo de medida a partir de los ítems utilizados para medir la satisfacción. El modelo obtenido en el análisis ha coincidido, en gran parte, con el planteamiento asumido por el Centro de Formación, que vino dado originariamente a partir del examen de experiencias anteriores sobre formación en distintas instituciones, así como por la explicitación de las demandas de información de los responsables del Centro y de cargos de responsabilidad de la Universidad.

Con respecto al modelo multinivel, encontramos que puede ser de gran utilidad para evaluar la contribución relativa de determinadas dimensiones; en este caso, las planteadas en la encuesta y de las que obtuvimos evidencias de validez de constructo mediante el Análisis Factorial (Bryk y Raudenbush, 1992; Du Toit, Du Toit y Cudeck, 1999; Kreft y Leeuw, 1998). De hecho, la variabilidad debida al contexto, aunque no era elevada, suponía un 18.76%, lo que podría motivar la búsqueda de predictores de segundo nivel.

Puede considerarse que los resultados obtenidos con el Análisis Multinivel presentan evidencias a favor de que los objetivos–contenidos, la metodología–ambiente y utilidad, en primer lugar son efectivamente dimensiones subyacentes a la conceptualización que se ha desarrollado del constructo satisfacción; y en segundo lugar, de que sirven para predecir parcialmente la satisfacción global de los participantes con las acciones formativas. Por otro lado, la metodología–ambiente no varía entre las acciones formativas, lo cual puede ser un indicio de que el proceso formativo ha sido adecuadamente estandarizado por el Centro de Formación a través de las distintas acciones formativas; sin embargo, los objetivos–contenidos así como la utilidad varían entre acciones formativas. Esta variación, puede ser explicada sólo parcialmente a través de la satisfacción global con el formador, ya que la otra variable de segundo nivel estudiada (área temática) no resultó significativa. Por ello, sería necesario seguir indagando sobre las posibles fuentes de variación de segundo nivel.

A partir de los análisis desarrollados, se ha pretendido plantear la complementariedad entre el Análisis Factorial y la modelización jerárquica para la toma de decisiones dentro de la organización. En nuestro caso concreto, además de la dimensionalidad del constructo, se observó que la calidad del profesorado modelaba los parámetros de la dimensiones sobre la satisfacción global. Se ha mostrado cómo ambas técnicas posibilitan el estudio de la distinta dimensionalidad de los datos dependiendo del nivel jerárquico de las variables, contribuyendo además en el aporte de evidencias empíricas sobre la contribución relativa de cada una de las dimensiones

planteadas en la delimitación del constructo de interés, en este caso, satisfacción.

**Agradecimientos:** Este estudio forma parte de los resultados obtenidos en el proyecto de investigación PSI2011-29587 (Ministerio de Innovación y Ciencia) y en el proyecto de título Calidad metodológica y eficacia desde la evidencia (Fondo Nacional de Desarrollo

Científico y Tecnológico –FONDECYT– de Chile). Los datos se obtuvieron a través de la colaboración con el Centro de Formación y Perfeccionamiento del Personal de Administración y Servicios de la Universidad de Sevilla. Más información sobre esta organización pública, así como el plan de formación anual para sus empleados, puede consultarse en [www.forpas.us.es](http://www.forpas.us.es).

## Referencias

- Anguera, M. T. (2008). Evaluación de programas desde la metodología cualitativa. *Acción Psicológica*, 2(5), 87-101.
- Anguera, M. T., Chacón-Moscoso, S. y Blanco, A. (2008). *Evaluación de programas sociales y sanitarios: Un abordaje metodológico* (pp. 173-224). Madrid: Síntesis.
- Barbero-García, M. I., Vila-Abad, E. y Holgado-Tello, F. P. (2011). Introducción básica al Análisis Factorial. Madrid: UNED.
- Bryk, A. y Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage Publications.
- Campbell, D. T. (1982). *Can we be scientific in applied social science?* Paper presented at the Annual Meeting of the Educational Research Association. (Reimpreso en R. F. Conner, D. G. Altman y C. Jackson (1984) *Evaluation Studies Review Annual*, 9, 26-48.
- Chacón-Moscoso, S., Sanduvete-Chaves, S., Portell-Vidal, M. & Anguera, M. T. (2013). Reporting a program evaluation: Needs, program plan, intervention, and decisions. *International Journal of Clinical and Health Psychology*, 13(1), 58-66.
- Cook, T. D. (1981). Dilemmas in evaluation of social programs. En M. B. Brewer y B. E. Collins (Eds.), *Scientific inquiry and the social sciences. A volume in honor of Donald T. Campbell* (pp. 257-287). San Francisco: Jossey-Bass.
- Cook, T. D., Leviton, L. C. y Shadish, W. R. (1985). Program evaluation. En G. Lindzey y E. Aronson (Eds.), *Handbook of Social Psychology*. (pp. 699-777). Nueva York: Random House.
- Du Toit, S., Du Toit, M. y Cudeck, R. (1999). *Introduction to the analysis of multilevel models with LISREL 8.30*. Chicago: Scientific Software International.
- Fernández, J. A., y Ovejero, A. (1994). Satisfacción laboral en un centro hospitalario: Un análisis del cuestionario de Porter. *Psicología del Trabajo y de las Organizaciones*, 10, 39-61.
- Ferrando, P. y Anguiano, C. (2010). El Análisis Factorial como técnica de investigación en Psicología. *Papeles del Psicólogo*, 31(1), 18-33.
- Gademann, A. M., Guhn, M. y Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the Satisfaction with Life Scale adapted for children: A focus on cognitive processes. *Social Indicators Research*, 100(1), 37-60.
- Herman, J. L., Morris, L. L. y Fitz-Gibbon, C. T. (1987). *Program evaluation kit (2ª edición)* Beverly Hills: Sage Publications. (9 Vols.).
- Holgado-Tello, F. P. y Barbero-García, I. (2008). Innovaciones metodológicas en evaluación psicológica: perspectivas de futuro. *Acción Psicológica*, 5, 5-6.
- Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I. y Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, the International Journal of Methodology*, 44(1), 153-166.
- House, E. R. (1993). *Professional Evaluation. Social impact and political consequences*. Londres: Sage.
- House, E. R. (1994). *Evaluación, ética y poder*. Madrid: Morata.
- House, E. R. y Shull, R. D. (1988). *Rush to Policy. Using Analytic Techniques in Public Sector Decision-Making*. New Brunswick: Transaction Publishers.
- Jöreskog, K. G. (2001). Analysis of ordinal variables 2: Cross-Sectional Data. Documentación del Workshop *Structural equation modelling with LISREL 8.51*. Jena: Friedrich-Schiller-Universität Jena.
- Jöreskog, K. G. (2003). *Factor Analysis by MINRES. To the memory of Harry Harman and Henry Kaiser*. Recuperado de <http://www.ssicentral.com/lisrel/techdocs/minres.pdf>
- Jöreskog, K. G. y Sörbom, D. (1999). *PRELIS 2.30*. Mooresville, IN: Scientific Software International.
- Jöreskog, K. G. y Sörbom, D. (2003). *LISREL 8.54*. Chicago: Scientific Software International.
- Jöreskog, K. G., Sörbom, D., Du Toit, S. y Du Toit, M. (1999). *LISREL 8: New statistical features*. Chicago: Scientific Software International.
- Kirkpatrick, D. (1999). *Evaluación de acciones formativas: Los cuatro niveles*. Barcelona: Training Club y Epise.
- Kreft, I. y Leeuw, J. (1998). *Introducing multilevel modeling*. Londres: Sage Publications.
- Lee, C., Zhang, G. y Edwards, M. C. (2012). Ordinary least squares estimation of parameters in exploratory factor analysis with ordinal data. *Multivariate Behavioral Research*, 47, 314-339.
- Medina, M. (1996). Evaluation of the quality of assistance in social services. *Intervención Psicosocial*, 14, 23-42.
- Mejías, S. y Huaccho, L. (2011). Macroergonomics intervention programs: recommendations for their design and implementation. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 21(3), 227-243.
- Morata, M. A. y Holgado-Tello, F. P. (2013). Construct validity of Likert scales through Confirmatory Factor Analysis: a simulation study comparing different methods of estimation based on Pearson and polychoric Correlations. *International Journal of Social Science Studies*, 1(1), 54-61.
- Palumbo, D. J. y Nachmias, D. (1983). The preconditions for successful evaluation: is there an ideal paradigm? *Policy Sciences*, 16, 67-79. (Reimpreso en R. F. Conner, D. G. Altman y C. Jackson (Eds.), *Evaluation Studies Review Annual*, 9, 102-114).
- Passmore, J. y Velez, M. (2012). SOAP-M: A training evaluation model for HR. *Industrial and Commercial Training*, 44(6), 315-325.
- Pineda, P. (2010). Evaluation of training in organisations: A proposal for an integrated model. *Journal of European Industrial Training*, 34(7), 673-693.
- Rajeev, P., Madan, M. S. y Jayarajan, K. (2009). Revisiting Kirkpatrick's model - an evaluation of an academic training course. *Current Science*, 96(2), 272-276.
- Rossi, P. H. y Freeman, H. E. (1985). *Evaluation: A systematic approach*. Londres: Sage Publications.
- Sarkis, J., Gonzalez, P. y Adenso, B. (2010). Stakeholders pressure and the adoption of environmental practices: The mediating effect of training. *Journal of Operations Management*, 28, 163-176.
- Sechrest, L. y Figueredo, A. J. (1993). Program evaluation. *Annual Review of Psychology*, 44, 645-674.
- Thayer, P. W. (1991). A historical perspective on training. En I. L. Goldstein and Associates (Eds.), *Training and development in organizations* (pp. 457-468). San Francisco: Jossey-Bass.
- Ventosa, P. (1998). *Desde la evaluación de la formación al rendimiento de la inversión*. Barcelona: Epise.
- Wilkins, J. L. M. (2010). Modeling quantitative literacy. *Educational and Psychological Measurement*, 70(2), 267-290.
- Yang, F., Jöreskog, K. G. y Luo, H. (2010). Confirmatory Factor Analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(3), 392-423.
- Yin, H. y Schmeidler, P. J. (2009). Why do standardized ISO 14001 environmental management systems lead to heterogeneous environmental outcomes? *Business Strategy and the Environment*, 18, 469-486.

(Artículo recibido: 09-04-2013; revisado: 09-07-2013; aceptado: 24-09-2013)