



Educación XX1

ISSN: 1139-613X

educacionxx1@edu.uned.es

Universidad Nacional de Educación a

Distancia

España

Blanco-Blanco, Ángeles; Asensio Muñoz, Inmaculada; Carpintero Molina, Elvira; Ruiz de Miguel, Covadonga; Expósito Casas, Eva

APLICACIONES DE LA SEGMENTACIÓN JERÁRQUICA EN MEDICIÓN Y  
EVALUACIÓN DE PROGRAMAS EDUCATIVOS. EJEMPLOS CON UN PROGRAMA DE  
EDUCACIÓN FINANCIERA

Educación XX1, vol. 20, núm. 2, 2017, pp. 235-257

Universidad Nacional de Educación a Distancia

Madrid, España

Disponible en: <http://www.redalyc.org/articulo.oa?id=70651145010>

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en redalyc.org

redalyc.org

Sistema de Información Científica

Red de Revistas Científicas de América Latina, el Caribe, España y Portugal

Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

# 10

## **APLICACIONES DE LA SEGMENTACIÓN JERÁRQUICA EN MEDICIÓN Y EVALUACIÓN DE PROGRAMAS EDUCATIVOS. EJEMPLOS CON UN PROGRAMA DE EDUCACIÓN FINANCIERA<sup>1</sup>**

**(APPLICATIONS OF HIERARCHICAL SEGMENTATION IN MEASUREMENT  
AND EVALUATION OF EDUCATIONAL PROGRAMS. EXAMPLES  
WITH A FINANCIAL EDUCATION PROGRAM)**

Ángeles Blanco-Blanco  
Inmaculada Asensio Muñoz  
Elvira Carpintero Molina  
Covadonga Ruiz de Miguel  
*Universidad Complutense de Madrid*  
Eva Expósito Casas  
*UNED*

*DOI: 10.5944/educXX1.14464*

### **Cómo referenciar este artículo/How to reference this article:**

Blanco-Blanco, Á; Asensio Muñoz, I.; Carpintero Molina, E.; Ruiz de Miguel, C. y Expósito Casas, E. (2017). Aplicaciones de la segmentación jerárquica en medición y evaluación de programas educativos. Ejemplos con un programa de educación financiera. *Educación XX1*, 20(2), 235-257, doi: 10.5944/educXX1.14464

Blanco-Blanco, Á; Asensio Muñoz, I.; Carpintero Molina, E.; Ruiz de Miguel, C. & Expósito Casas, E. (2017). Aplicaciones de la segmentación jerárquica en medición y evaluación de programas educativos. Ejemplos con un programa de educación financiera. [Applications of hierarchical segmentation in measurement and evaluation of educational programs. Examples with a financial education program]. *Educación XX1*, 20(2), 235-257, doi: 10.5944/educXX1.14464

## **RESUMEN**

En este trabajo se propone el uso en medición y evaluación educativa de las técnicas de segmentación jerárquica o árboles de decisión, utilizados ya extensamente en otros campos, fundamentalmente en marketing, como técnica de estudio de mercados. Mediante *IBM SPSS Statistics 21* se analizan los datos de una encuesta aplicada a 648 profesores de Educación Secundaria para la evaluación de un programa de educación financiera (EF) y se ilustran diferentes posibles aplicaciones metodológicas de los árboles de clasificación y regresión (CART o C&RT: Classification And Regression Trees) para el estudio de la validez basada en evidencias sobre el contenido y la estructura interna de las medidas procedentes del cuestionario de evaluación empleado. Por otra parte, en una

línea más similar a la que se ha seguido para la evaluación de programas en sociología, medicina o psicología, se llegan a identificar claramente las características de los diferentes segmentos de población, entre los que se destacan, por su valor informativo en evaluación de programas educativos, los grupos extremos, constituidos por los usuarios más «entusiastas» y los más «críticos» con el programa. A partir de los resultados obtenidos, se discuten las posibilidades metodológicas de esta técnica tanto en medición como en evaluación de programas. Con el estudio empírico presentado se argumenta la propuesta de aplicación de las técnicas arborescentes, junto con otras aproximaciones metodológicas más tradicionales, como son el análisis de regresión lineal y el análisis factorial exploratorio y confirmatorio, dado que esta metodología se ajusta mejor a las características de los datos que se suelen manejar en investigación y evaluación educativa y permite una interpretación más clara e integrada de resultados.

## **PALABRAS CLAVE**

Análisis estadístico; evaluación de programas; cuestionario; validez.

## **ABSTRACT**

This paper suggests the utilization of hierarchical segmentation or decision trees in educational measurement and evaluation extensively utilized in other areas, mainly in marketing, as a market survey technique. By using *IBM SPSS Statistics 21*, data from a survey answered by 648 Secondary Education teachers are analysed in order to evaluate a program of financial education, and the different methodological applications of the Classification and Regression Trees (CART or C&RT) are shown in the study of validity based on evidence about content and internal structure of the measures from the questionnaire of evaluation which has been used. Moreover, in a more similar trend than the one followed for evaluation in sociology, medicine or psychology programs, the characteristics of the different segments of the population are clearly identified. Among these segments, notable for their informative value in the evaluation of educative programmes, are the extreme groups consisting of the users that are most «enthusiastic» and «critical» with the program. From the achieved results, the methodological possibilities of this measurement and evaluation technique are discussed. With the presented empirical study, the proposal for application of decision trees techniques is justified, together with the most traditional approaches, such as lineal regression analysis and the exploratory and confirmatory factorial analysis. This is because this methodology fits more appropriately to the characteristics of the normally-used data in research and educational evaluations. Also it allows a clearer and more integrated interpretation of the results.

## **KEYWORDS**

Statistical Analysis; program evaluation; questionnaire; validity.

## INTRODUCCIÓN

En investigación de mercados, el término *segmentación* tiene una doble acepción, ya que se refiere tanto a la estrategia consistente en la utilización de segmentos para darles un tratamiento diferencial según la perspectiva de la empresa, como al proceso estadístico o algoritmo que lleva a la delimitación e identificación de tales segmentos entendidos estos como subgrupos homogéneos de una población definida, considerando determinadas características de los sujetos estudiados y de acuerdo con ciertas reglas (Luque, 2012; Pérez, 2011). En sentido amplio, dentro de los modelos de segmentación se suelen incluir diferentes procedimientos de discriminación o de perfilado de segmentos. Entre los modelos *a priori* están el análisis loglineal, la regresión logística o el análisis discriminante, y entre los modelos *post hoc*, según su objetivo sea más descriptivo o más predictivo, se encuentran habitualmente planteamientos metodológicos derivados del análisis factorial (Q análisis), el análisis *cluster*, las redes neuronales artificiales, la inducción de reglas o las técnicas de clasificación arborescentes, también denominadas de segmentación jerárquica, entre las que cabe destacar los métodos CHAID, CART, QUEST (Picón, Varela y Lévy, 2004). Bajo la denominación específica de *análisis de segmentación* mediante árboles de clasificación (Escobar, 2007) o *segmentación jerárquica* (Luque, 2012) se agrupan un conjunto de técnicas estadísticas que pretenden dividir las poblaciones o muestras objeto de estudio, siguiendo un proceso secuencial descendente, para la identificación de subgrupos homogéneos en una variable de interés para el investigador (variable dependiente, respuesta o criterio), atendiendo a las características de los sujetos estudiados en una selección de las variables independientes, predictoras, pronosticadoras o explicativas más importantes al efecto. En segmentación de mercados se utilizan como criterios para fragmentar las muestras todas las variables que permitan identificar y definir diferentes tipos de consumidores, por lo que se suelen incluir en los análisis datos geográficos, demográficos y socioeconómicos, junto con datos comportamentales e indicadores de personalidad, estilos de vida o valores, además de variables motivacionales, actitudinales, de sensibilidad y de opinión.

Escobar (2007) ofrece una magnífica y didáctica síntesis de la evolución de estas técnicas, desde el antecedente que supuso por parte de Belson (1959) el desarrollo de un procedimiento para la selección de predictores basado en el tamaño de los residuos de las tablas de contingencia, hasta los *métodos de conjunto* (*ensemble methods*) (Berk, 2006).

Las técnicas de segmentación jerárquica están demostrando su utilidad en diferentes campos y para diferentes propósitos, de modo que se vienen aplicando de manera cada vez más frecuente en investigación

sociológica, psicológica o médica (Ang y Goh, 2013; Escobar, 2007; Gervilla *et al.*, 2009; Gervilla y Palmer, 2009; Richard's, Solanas, Ledesma, Introzzi y López, 2008; Schiattino y Silva, 2008, entre otros) y su empleo, bien como técnica única, bien combinada con otras técnicas de minería de datos, es una tendencia al alza en el ámbito educativo (Delen, 2012; Horner, Fireman y Wang, 2010; Kovacic, 2012; Márquez-Vera, Romero y Ventura, 2013; Tiwari, Singh y Vimal, 2013). Se trata de trabajos centrados, principalmente, en la predicción y el análisis de factores asociados a diferentes variables dependientes, aunque la utilización de la técnica para el análisis exploratorio de los datos no supone el único universo de aplicación posible. En este sentido, resulta de especial interés la aportación de Brandmaier, von Oertzen, McArdle y Lindenberger (2013), quienes proponen el uso combinado de los modelos de ecuaciones estructurales y los árboles de decisión «*SEM Trees*» con fines exploratorios y confirmatorios. En dicha propuesta, se destacan las ventajas de los árboles de decisión frente a los estadísticos tradicionales, al permitir revelar interacciones discontinuas así como relaciones no lineales entre las covariables y el *outcome* (Brandmaier *et al.*, 2013).

Nuestro planteamiento, en línea con la postura que mantienen Liza-soain y Joaristi, (2000) y Lizasoain *et al* (2003), es que la segmentación jerárquica es muy útil en evaluación de programas educativos, ámbito en el cual alcanzan especial relevancia las principales ventajas de esta metodología, entre las que destacan su *versatilidad*, su *flexibilidad*, su *fácil interpretación* y su *relativa novedad o efervescencia creativa* (Picón *et al.*, 2004).

Nuestro objetivo es ir más allá en el uso que con más frecuencia se ha hecho de la técnica: proponemos e ilustramos la utilización de los árboles de decisión como una aproximación metodológica apta para la validación de las medidas en las que se apoya la evaluación educativa. Como es sabido, comparabilidad, equidad, fiabilidad y validez son conceptos fundamentales a la hora de determinar la calidad técnica de una medida. Desde la aportación clave de Stevens (1951), que definió *la medida como la asignación de números, símbolos o valores a las propiedades de los objetos de acuerdo a ciertas reglas*, en el campo de las Ciencias Sociales, la medición se concibe básicamente como *el proceso consistente en vincular conceptos abstractos (constructos) con indicadores empíricos* (Carmines y Zeller, 1991), que se obtienen a través de la aplicación de instrumentos (tests fundamentalmente, pero también cuestionarios, escalas...) y que deben reunir ciertos requisitos que aseguren que realmente estamos midiendo lo que pretendemos. En este sentido, la validez de las medidas se plantea como una cuestión esencial, pero es un concepto complejo que ha ido redefiniéndose, de manera cada vez más precisa, a lo largo de los últimos años, desde las primeras recomendaciones técnicas de la APA de 1954. En la actualidad, como indican Sireci

y Sukin (2013), *validez* es un término comprehensivo que hace referencia no a las características del instrumento de medida, sino a la conveniencia de su uso y a la exactitud de las inferencias hechas sobre la base de los resultados del mismo. La última edición de las *Normas para la Medición Educativa y Psicológica (Standards for educational and psychological testing)*, realizadas en colaboración por la *American Educational Research Association (AERA)*, la *American Psychological Association (APA)* y el *National Council on Measurement in Education (NCME)* (2014) representa la aproximación más actual al concepto, tras las publicadas en 1966, 1974, 1985 y 1999. En ellas se aclara que la validez se refiere al grado en el cual la evidencia y la teoría soportan las interpretaciones de las medidas, o puntuaciones en un test, de acuerdo a los usos para los que este se ha diseñado y es, por lo tanto, la consideración más fundamental que hay que tener en cuenta en el desarrollo y evaluación de instrumentos de medida.

Desde esta perspectiva, el proceso de validación implica la acumulación de evidencias relevantes para proporcionar una base científica sólida a las interpretaciones de las puntuaciones. Aunque ya nadie discute que el concepto de *validez* es único, la aproximación metodológica se sigue planteando desde la pluralidad, puesto que las evidencias obtenidas pueden ser de diversa índole. Así, por ejemplo, entre las técnicas estadísticas más utilizadas para reunir *evidencias de validez basada en el contenido* están los diferentes coeficientes de acuerdo inter-jueces; para el estudio de la *evidencia de validez basada en la relación con otras variables* la propuesta metodológica más generalizada es el análisis de regresión lineal; y para el estudio de la *evidencia de validez basada en la estructura interna* es prácticamente unánime la propuesta del análisis factorial exploratorio y confirmatorio (AFE y AFC) (Bollen, 1989) y, con menos frecuencia, las matrices multirrasgo-multimétodo, que aportan una información conjunta acerca de la fiabilidad y la validez convergente y discriminante simultáneamente (Campbell y Fiske, 1959).

A pesar de que este es el contexto metodológico que representa la tendencia predominante en los trabajos que se publican sobre validación de medidas útiles para la investigación y evaluación educativas, en este trabajo se pretende ilustrar cómo mediante una única técnica se pueden llegar a reunir diferentes tipos de evidencias que apoyen los argumentos de validez de las medidas, para lo que se utilizan los datos obtenidos en un cuestionario con el que se pretende evaluar la calidad percibida de un programa educativo. En el presente artículo se explica, a través de diferentes ejemplos, cómo, con las técnicas de segmentación jerárquica a que nos hemos referido anteriormente, se pueden reunir evidencias de validez que permitan responder a diferentes tipos de cuestiones clave en un proceso de validación, tales como: ¿qué elementos constitutivos del programa pesan más en la valora-

ción, cuando unos expertos evalúan de manera general la calidad de dicho programa?, ¿qué componentes hay en una puntuación global de valoración de un programa obtenida a través de un cuestionario al que han respondido usuarios internos del mismo?, ¿la puntuación obtenida como dato general en la evaluación de un programa se relaciona debidamente con aquello que se pretende pronosticar en último extremo y que en nuestro caso es la calidad percibida del mismo?... Para dar respuesta a estas preguntas, la técnica que se propone en este artículo es siempre la misma, con lo que, aplicada a estudios de validación, se añade la *unicidad* a las otras ventajas del análisis de segmentación jerárquica o árboles de decisión o clasificación que hemos presentado más arriba cuando argumentamos su utilidad en estudios de evaluación de programas. Nuestro planteamiento, poniendo negro sobre blanco la posibilidad de reunir evidencias de distinto tipo que apoyen la validez desde un único enfoque metodológico, pretende contribuir a reforzar la unicidad de un concepto que todavía, en la mente de muchos investigadores y evaluadores educativos, aparece dividido según la antigua concepción trinitaria que diferenciaba la validez de contenido, la de criterio y la de constructo. Lo que en procesos de validación permitirá reunir diferentes evidencias que apoyen la validez de las medidas será el conjunto de variables seleccionadas para cada análisis, que cambiará de acuerdo a los argumentos interpretativos en los que se basa la medida.

## MÉTODO

### Muestra

Para ilustrar las posibilidades de la alternativa metodológica propuesta, se utilizan como ejemplo los datos recogidos en una muestra de 648 profesores de Educación Secundaria Obligatoria (ESO) participantes voluntarios en una experiencia piloto de Educación Financiera (EF) implementada con estudiantes de 3.º de ESO en toda España. En ella, los profesores aplicaron durante tres meses un programa específico de 9 unidades que contaba con una guía didáctica, donde se describía el contenido de cada unidad y las actividades a realizar en el aula, y con una página *web* de apoyo que recogía actividades de ampliación.

### *Instrumento*

Con el fin de estudiar la calidad del programa a partir del valor percibido del mismo por los docentes, se elaboró un cuestionario *ad hoc*, con preguntas que se agrupan en 5 categorías en función de las variables a las que hacen referencia, como se resume en la Figura 1.

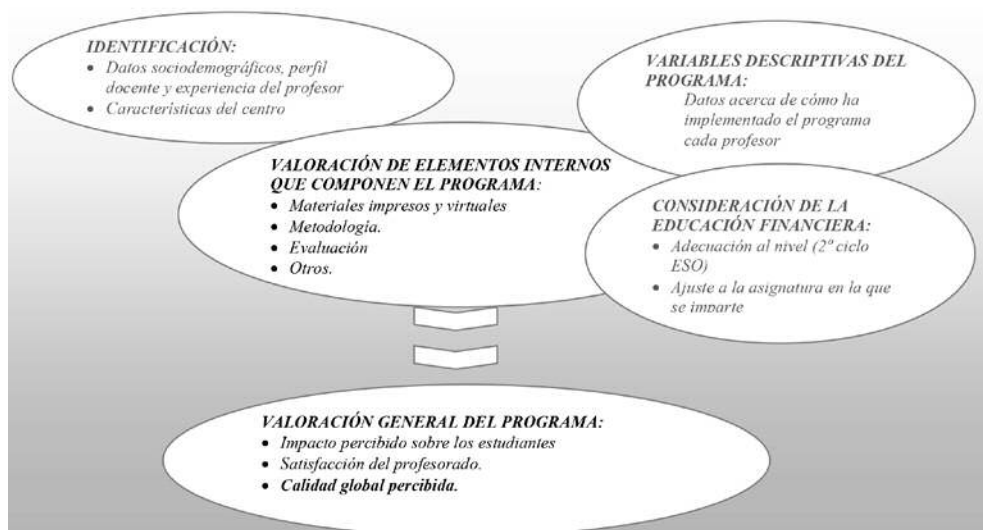


Figura 1. Variables incluidas en el cuestionario elaborado *ad hoc* para la evaluación curricular del programa de educación financiera tomado como ejemplo

A partir del cuestionario se pretendía la evaluación de un programa educativo, lo que supone hacer inferencias acerca de su calidad tomando como base las percepciones de los profesores, considerados como expertos y usuarios internos del mismo. Las puntuaciones obtenidas, en consecuencia, han de ser útiles para medir la calidad global percibida del programa (variable dependiente, respuesta o criterio fundamental del estudio) y para identificar qué aspectos intervienen o pesan más en dicha valoración (variables predictoras o explicativas), como medio de interpretación del significado o significados de la medida conseguida.

### *Procedimiento*

La elección del criterio base y de los predictores depende de la finalidad de cada estudio y varía según las posibles hipótesis de partida que vinculen la variable dependiente con dichos predictores (Picón *et al.*, 2004). En este trabajo se ha utilizado siempre como criterio la valoración de la calidad global percibida, medida con un ítem de valoración en una escala de 0 a 10. Como variables explicativas se consideran diferentes grupos de variables atendiendo a los dos tipos de aplicaciones que aquí se propone ilustrar: evaluación de un programa educativo y validación de la medida mediante la obtención de diferentes tipos de evidencias de



validez que apoyen la interpretación de las puntuaciones obtenidas en el cuestionario empleado.

Para la evaluación del programa la cuestión de fondo es determinar qué variables de identificación y descriptivas del programa discriminan más entre los docentes que lo valoran mejor y peor, cuando evalúan el programa globalmente. Con este ejemplo se ilustra cómo se pueden identificar diferentes perfiles de profesorado, unos más proclives que otros a valorar positiva o negativamente el programa, siguiendo un procedimiento similar al utilizado en estudios de mercado para llegar a la identificación de perfiles de usuarios que se corresponden con segmentos específicos de población. Se expone así cómo el análisis de segmentación jerárquica supone un avance respecto de los enfoques descriptivos al uso en evaluación de programas educativos, que suelen limitarse a caracterizar el programa según la tendencia central de los usuarios sea hacia una valoración más positiva o más negativa y a valorar la representatividad de esa tendencia, teniendo en cuenta la dispersión de los datos o la forma de la distribución global.

Para el estudio de la validez de las medidas, la cuestión es obtener patrones de respuesta indicativos de que los ítems efectivamente miden el constructo «*calidad percibida del programa*». De esta manera, se ilustra, en primer lugar, cómo obtener evidencias de validez, mediante el estudio del peso que tienen los componentes internos del programa en la valoración global otorgada al mismo. Así se puede llegar a discernir, cuando los profesores valoran globalmente el programa, qué elementos del mismo están valorando realmente, lo que podría entenderse como una evidencia de validez de contenido. Por otra parte, es importante, para la correcta interpretación de las puntuaciones obtenidas con el cuestionario, determinar si la *calidad total percibida* tiene componentes de satisfacción y de impacto, como cabe esperar según la teoría, motivo por el cual se introdujeron en el cuestionario dos ítems para evaluar cada uno de estos dos aspectos. En este caso se trata de obtener evidencias que apoyen la relación de la *calidad global percibida* con otras variables muy relacionadas con el concepto. Por último, mediante la inclusión conjunta de los componentes internos del programa educativo y de los indicadores de satisfacción e impacto percibido, se puede llegar a ponderar el mayor o menor peso de cada uno de estos aspectos, de cara a identificar la red nomológica que define el constructo, esto es, su estructura interna, lo que sirve de base para una más completa interpretación de las puntuaciones obtenidas.

### *Análisis de datos*

Las respuestas dadas por los profesores a las diferentes preguntas planteadas en el cuestionario se analizan mediante una técnica de segmen-

tación jerárquica, concretamente la técnica de clasificación, disponible en *IBM SPSS Statistics 21*, con el método de crecimiento C&RT. Se ha elegido el algoritmo CART (Breiman, Friedman, Olshen y Stone, 1984) porque es un método de árbol binario numérico y exhaustivo (Román y Levy, 2003), de mayor complejidad computacional que otros, pero que da lugar a segmentos efectivos e interpretables. Desde un *punto de vista metodológico* el criterio fundamental de valoración de los resultados obtenidos será el de máxima homogeneidad intragrupos y heterogeneidad entre los grupos. Desde el *punto de vista estratégico* se considerará que los segmentos resultantes sean fácilmente identificables, sustanciales, accesibles, estables, factibles y diferenciados (Picón *et al.*, 2004).

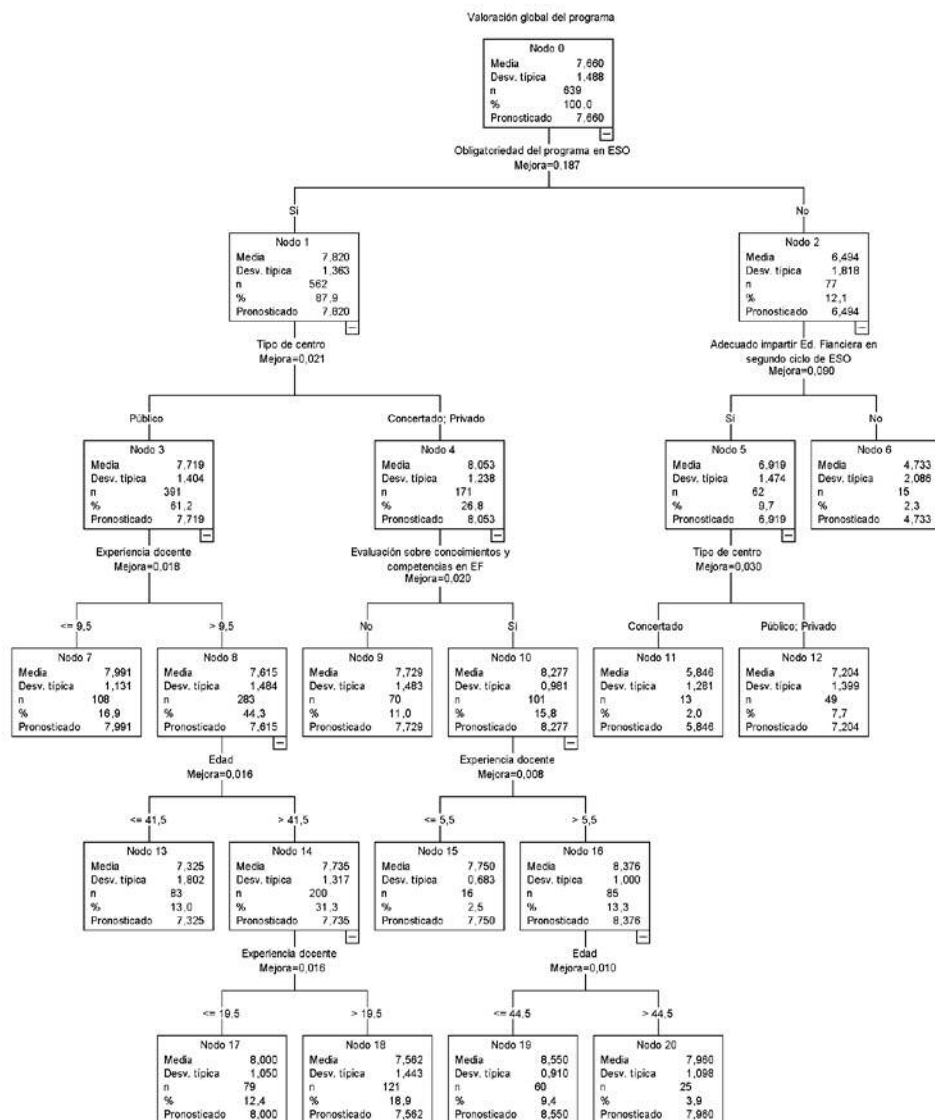
## RESULTADOS

### Ejemplo de aplicación del algoritmo CART en la evaluación de un programa

De cara a la toma de decisiones derivadas de la evaluación, resultan de especial interés los perfiles de los profesores más extremos. Para ello, podemos denominar «*entusiastas*» a los que otorgan al programa las puntuaciones más altas, frente a los profesores «*críticos*», que son los que dan al programa las valoraciones más bajas.

Se han incorporado en el análisis como variables predictoras variables de diferente naturaleza correspondientes a los grupos de variables de identificación, descriptivas del programa y de consideración de la educación financiera y como variable dependiente la calidad global (un ítem de valoración en una escala de 0 a 10). Se ha aplicado el procedimiento CART, reduciendo el tamaño mínimo de los nodos padre e hijo a 50 y 10 respectivamente, dado el tamaño total de la muestra y con el fin de dar posibilidad de que emerjan los grupos más extremos, aunque sean minoritarios. En la gráfica 1 se ofrecen los resultados obtenidos. En ella se observa cómo en el nodo raíz, o nodo cero, los 639 docentes con valores válidos de los 648 encuestados han valorado globalmente el programa con una puntuación media de 7,66 puntos (sobre 10). En el primer nivel de segmentación se generan dos nodos (dos subgrupos o segmentos) en función de si el profesorado considera que el contenido de EF debería ser obligatorio o no en el curriculum de la ESO, que emerge como la principal variable «explicativa». El nodo 1 contiene los 562 docentes que han contestado que sí debería ser obligatorio, cuya valoración media global del programa es de 7,819 y por tanto mayor que 6,494, que es la valoración global media de los 77 docentes que consideran que los contenidos no deberían ser obligatorios (nodo 2). Así, analizando los distintos niveles de segmentación

y siguiendo los nodos del árbol puede observarse que el perfil de los 60 docentes que mejor valoran globalmente el programa (nodo 19), con una media de 8,55 es el siguiente: consideran que los contenidos de referencia deberían ser obligatorios (nivel 1), pertenecen a centros privados o concertados (nivel 2), cuentan con una experiencia docente superior a 5 años (nivel 4), aunque tienen edades inferiores a 44 años y durante la aplicación del programa han llevado a cabo evaluaciones a sus estudiantes (nivel 3). También dan una valoración global al programa de 8 los profesores del nodo 17, que consideran que los contenidos de referencia deberían ser obligatorios (nivel 1), pero son de centros públicos (nivel 2) con experiencia docente entre 10 y 20 años aproximadamente y, a diferencia de los de centros concertados y privados, de mayor edad. Por otra parte, se obtiene el perfil de los 15 profesores que peor valoran globalmente el programa (nodo 6), con una media de 4,733: son los docentes que consideran que los contenidos de referencia no deberían ser obligatorios y además no deberían impartirse en el segundo ciclo de la ESO. Con las reglas de parada definidas, el resto de variables independientes incluidas en el análisis resultan irrelevantes para discriminar entre los usuarios *más entusiastas* y los *más críticos* con el programa.



Gráfica 1. Árbol de segmentación para el análisis de la calidad global del programa con variables de identificación, descriptivas del programa y de consideración de la EF

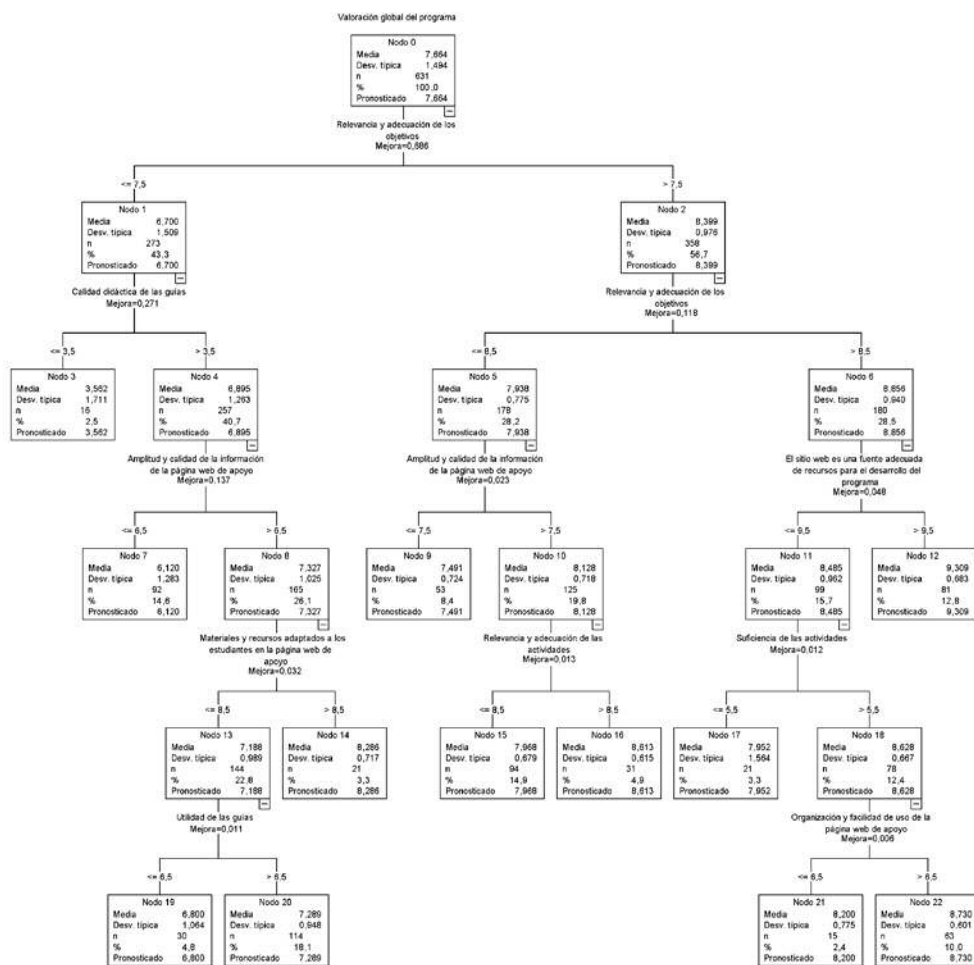
## Ejemplos de aplicación del algoritmo CART para el estudio de la validez de las medidas

Por otra parte, se ha utilizado esta técnica estadística para estudiar la relación de la calidad global percibida con la valoración dada a los elemen-

tos que definen el programa, como son los objetivos, la metodología, los contenidos, los materiales impresos, las actividades o los recursos electrónicos. En este caso todas las variables incluidas en el análisis como explicativas son también cuantitativas (ítems en los que el profesor da una valoración de 0 a 10 a cada aspecto evaluado). Como se aprecia en la gráfica 2, los nodos más extremos son el 12 y el 3 con medias en la valoración global del programa de 9,309 y 3,562 respectivamente. Las características del programa que más peso tienen en la valoración global del mismo son:

- la relevancia y adecuación de los objetivos, que emerge en el primer nivel de segmentación, dando lugar a la formación de dos grupos según la valoración de este aspecto sea menor o igual que 7,5 (nodo 1) o mayor que 7,5 (nodo 2).
- la calidad didáctica de las guías: los profesores que valoran con 3.5 o menos este aspecto, si habían otorgado un 7,5 o menos a la relevancia de los objetivos, constituyen un grupo homogéneo de  $n=16$  que se caracteriza por ser el más crítico con el programa, dando al mismo una valoración global de 3.562 (nodo 3).
- los profesores que otorgan valores superiores a 9.5 puntos a la adecuación de los recursos del sitio *web* para el desarrollo del programa y que, al mismo tiempo, otorgan una puntuación superior a 8,5 puntos a la relevancia y adecuación de los objetivos, son los docentes más entusiastas con el programa ( $n = 81$ ), conformando el nodo 12 con una valoración global media de 9,309.

Otras variables que discriminan son la amplitud y calidad de la información de la página *web* de apoyo, la valoración de la página *web* como una fuente adecuada de recursos para el desarrollo del programa, la adaptación a los estudiantes de los materiales y recursos de la página *web*, la relevancia y adecuación de las actividades, la suficiencia de las actividades, la utilidad de las guías y por último la organización y facilidad de uso de la página *web* de apoyo. No aparecen en el árbol aportado por la solución CART la adecuación de los contenidos, la variedad de las actividades que se plantean en el manual y la suficiencia del material que contienen las guías para impartir el programa. Sin embargo, de las cinco cuestiones incluidas en el cuestionario para obtener la valoración de la página *web* del programa discriminan entre los que valoran mejor y peor globalmente el programa cuatro, siendo de ellas la menos importante la organización y accesibilidad del sitio *web*. La utilidad de los materiales y recursos para preparar los contenidos y actividades en clase resulta ser la única variable no relevante según la solución CART, puesto que no discrimina en ningún caso.

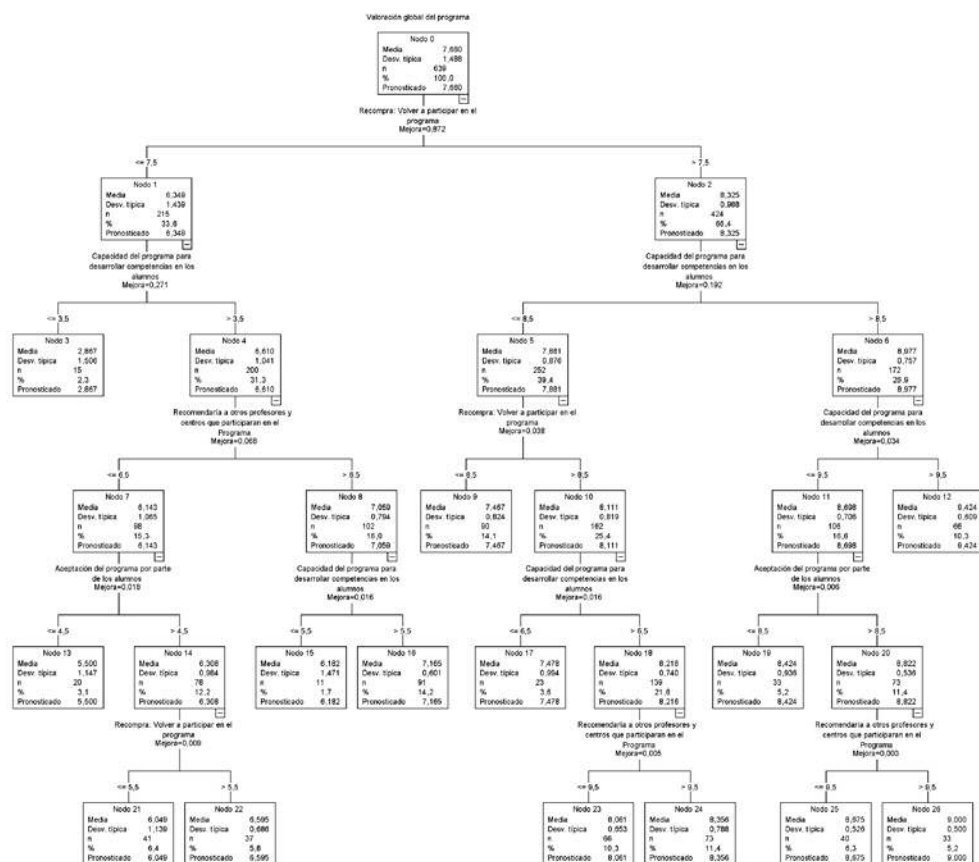


Gráfica 2. Árbol de segmentación para explorar qué elementos del programa intervienen en la valoración global de la calidad del mismo

Como aproximación a lo que entienden los profesores por calidad global, se ha realizado un análisis de lo que entienden los profesores como «calidad global percibida». En este análisis de segmentación se ha considerado como variable dependiente la valoración dada en este ítem y como variables independientes los indicadores incluidos en el cuestionario para evaluar la satisfacción y el impacto percibido del programa. Concretamente la satisfacción fue evaluada con dos ítems relativos al deseo manifestado por los profesores de volver a participar en el programa y a su posible recomendación a otros profesores y centros. El impacto se evaluó con dos ítems sobre el interés que consideran que el programa tiene para los alumnos y sobre la capacidad del programa para desarrollar compe-

tencias en los alumnos. Estos cuatro ítems también tenían una escala de respuesta de 0 a 10.

La pregunta de fondo en este estudio es ¿qué valoran los profesores cuando valoran globalmente el programa?, ¿están teniendo en cuenta solo su impacto en los estudiantes o solo la satisfacción que les ha producido impartirlo?, ¿se mezclan factores de impacto y de satisfacción, como por otra parte cabe suponer, si hacen una valoración de la calidad global del programa? Se trata en suma de desentrañar qué hay detrás del concepto de *calidad*. El resultado obtenido se ofrece en la gráfica 3. En ella se observa que los nodos más extremos son el 12 y el 3, lo que quiere decir que lo que mueve al profesorado a valorar peor o mejor el programa de forma global es principalmente un factor de satisfacción y en segundo lugar un factor de impacto, de modo que los más críticos no se ven participando de nuevo en la experiencia y además creen que los alumnos no han desarrollado las competencias pretendidas (los 15 profesores que constituyen el nodo 3 dan una valoración global media al programa de 2,867 sobre 10). Por otro lado, los profesores que volverían a impartir el programa y que creen que sus alumnos han desarrollado las competencias pretendidas (otorgando a este aspecto una puntuación superior a 9,5 sobre 10) y que constituyen un grupo homogéneo de 66 docentes, son los que globalmente consideran mejor el programa, dándole un valor medio de 9,424. Aunque los otros indicadores de impacto y satisfacción incluidos en el cuestionario no discriminan entre los dos grupos más extremos, sí entran a formar parte del árbol y discriminan, entre los nodos 26 y 13 que son los segundos segmentos más extremos, con medias de 9 y 5,5 respectivamente, por lo que cabe concluir que todas las variables incluidas en el análisis son relevantes. Estos resultados se pueden interpretar para extraer conclusiones acerca de la relación entre la calidad percibida, la satisfacción que ha producido al profesorado participar en el programa y el impacto que consideran que este ha tenido en los estudiantes, de modo que la evaluación global que los docentes hacen del programa que han impartido no es sesgada y tiene componentes tanto de impacto como de satisfacción.



Gráfica 3. Árbol de segmentación para el pronóstico de valoración global del programa a partir de indicadores de satisfacción e impacto

Los resultados obtenidos pueden considerarse aproximaciones parciales a la evidencia de validez en función de la estructura interna, en tanto que vienen a definir los componentes que intervienen en la definición de lo que los profesores entienden por calidad del programa, lo que por otra parte está en línea con las definiciones que de este término se han dado en la literatura, en la que junto con otras acepciones en las que la calidad se puede entender como «excelencia», como «valor» o como «ajuste a unas especificaciones concretas», también puede entenderse como «satisfacción de las expectativas de los usuarios» (Reeves y Bednar, 1994). Como se muestra en la tabla 1, cuando se incluyen todas las variables de valoración en el análisis conjuntamente, los indicadores de «satisfacción» se posicionan como las dos primeras variables en una serie ordenada según la importancia en la explicación de la valoración global otorgada al programa. Los indicadores de impacto aparecen en tercer y sexto lugar, porque dos de los elementos



constitutivos del mismo, como son la relevancia de los objetivos y de las actividades, pasan a ocupar el cuarto y quinto lugar.

Tabla 1  
*Importancia de las variables independientes*

Variable independiente	Importancia	Importancia normalizada
Intención de volver a participar en el programa	1.019	100.0 %
Recomendación a otros profesores y centros que participaran en el programa	1.006	98.8 %
Utilidad del programa para desarrollar competencias en Educación Financiera	.975	95.7 %
Relevancia y adecuación de objetivos	.794	77.9 %
Relevancia y adecuación de actividades planteadas	.776	76.2 %
Aceptación del programa por parte de los alumnos	.775	76.0 %
Calidad didáctica de las guías	.767	75.2 %
Utilidad de las guías	.758	74.4 %
Adecuación del sitio web como fuente de recursos	.700	68.7 %
Coherencia y relación entre objetivos, contenidos y actividades	.671	65.9 %
Adaptación de contenidos a los estudiantes	.626	61.4 %
Utilidad de materiales y recursos para preparar actividades	.624	61.3 %
Amplitud y calidad de información	.610	59.9 %
Ajuste de materiales y recursos a los alumnos	.594	58.3 %
Variedad de actividades del manual	.570	56.0 %
Suficiencia del material de las Guías para impartir el programa	.563	55.2 %
Organización del sitio web	.525	51.5 %
Suficiencia de actividades	.193	18.9 %
Métodos de crecimiento: CRT		
Variable dependiente: Valoración global del programa		

## CONCLUSIONES

En este trabajo se ha ilustrado cómo la aplicación de la segmentación jerárquica a las respuestas dadas por profesores en un cuestionario para la evaluación de una experiencia piloto, aporta información relevante no solo para la evaluación en sí misma del programa en cuestión, sino también en cuanto a la calidad de los datos en los que se basa dicha evaluación y, por tanto, a la toma de decisiones.

El estudio realizado pone de manifiesto, en primer lugar, la utilidad de la técnica para la segmentación de poblaciones de usuarios de servicios y programas educativos. La detección de grupos de población de interés, junto con el hecho de que las evaluaciones de programas educativos se apoyen frecuentemente en datos recogidos mediante cuestionarios, que muchas veces no reúnen las características métricas que se le requiere a los datos por parte de técnicas más exigentes, son factores que avalan la utilización de la segmentación jerárquica como herramienta de análisis en evaluación de programas educativos.

Si se parte de que la consideración de la opinión del profesorado ante la puesta en marcha de una innovación o la implementación de una reforma se presenta como una variable de interés, porque de su implicación profesional puede depender en último extremo el éxito del cambio (Pareja, Maciá, y Serrano, 2013), los resultados del análisis de segmentación con una experiencia piloto pueden llegar a ser muy informativos, al permitir conocer las características de los profesores más entusiastas o críticos, para anticiparse a atender las demandas y proponer las soluciones que, en cada caso, fueran convenientes. Por supuesto que una opción metodológica clásica para la identificación de subgrupos homogéneos, exhaustivos y mutuamente excluyentes es el análisis *cluster* (Escobar, 2007), pero la alternativa que aquí se defiende aporta a nuestro juicio dos ventajas: que el modo de clasificación se hace de manera descendente, esto es, de los grupos a los casos, y siempre con una variable criterio de referencia.

Por otra parte, la propuesta de aplicación de estos métodos como herramienta de validación de medidas es atrevida en tanto que novedosa, pero, a juzgar por los resultados obtenidos del estudio empírico aquí presentado, se trata de una propuesta al menos prometedora. El uso de las técnicas arborescentes se ha ido haciendo cada vez más extenso y lo que inicialmente pretendía servir solo para la detección automática de la interacción (AID), ha demostrado su utilidad para la segmentación de poblaciones, para la identificación de perfiles, para la clasificación de sujetos, para la selección de variables explicativas relevantes o para la predicción o pronóstico mediante el reconocimiento de pautas (Escobar, 2007). Nuestra apuesta fundamental es que, entre las aplicaciones del análisis de segmentación jerárquico, se han de incluir aquellas que se refieren a cuestiones métricas.

En las Normas para la Medición Educativa y Psicológica, que suponen la principal guía de referencia conceptual y metodológica al respecto, la cuestión de la validez se ha centrado tradicionalmente en torno al uso de tests para la medida de constructos pedagógicos o psicológicos más o menos complejos. En este artículo la afrontamos desde el uso de cuestionarios para la obtención de información cuantitativa que permita una aproximación

empírica a la calidad percibida de un programa o servicio educativo, con garantías de que la interpretación que se hace de las puntuaciones obtenidas en el mismo se basa en evidencias de validez relevantes. La realidad actual es que en investigación y evaluación educativas, cuando se trata de datos procedentes de cuestionarios o escalas, no hay pruebas de validez en sentido estricto. Lo que hay son diferentes aproximaciones metodológicas, correlacionales o basadas en las diferencias de medias, muchas veces con requisitos y supuestos difícilmente asumibles, que apoyan, desde una perspectiva empírica, la utilidad del instrumento y aportan información acerca del significado de lo que medimos (Morales, Urosa y Blanco, 2003). Una ventaja del análisis de segmentación jerárquica está en que, al ser un procedimiento no lineal, que se basa en una lógica booleana, no impone restricciones fuertes a los datos. Así, respecto a la socorrida regresión lineal, técnica con la que comparte el hecho de que en función de una serie de variables independientes puede predecirse el comportamiento en una variable dependiente, se diferencia en que no lo hace mediante funciones lineales o suma ponderada de los valores en los predictores, sino mediante el empleo de procedimientos lógicos basados en la igualdad /desigualdad, conectores y operadores lógicos (Escobar, 2007).

En los procesos de validación se recomienda partir de una matriz de especificaciones y de una red nomológica que represente la esencia del constructo, para la mejor explicación y comprensión de aquello que se mide. Los árboles de decisión se presentan como herramientas de validación en las que se pueden integrar los diferentes elementos a considerar, de manera que el tronco representa aquello que se mide y la ramas aquello que interviene en la explicación o predicción del constructo. Así, en la misma gráfica o tabla de resultados se pueden incorporar tanto los elementos internos como los externos en referencia con la variable objeto de estudio, en una red de relaciones tan compleja y parsimoniosa al mismo tiempo como la que podría surgir de la aplicación de un modelo de estructuras de covarianza, representada mediante el clásico diagrama causal. Los árboles de decisión aquí utilizados no permiten por sí solos concluir en términos de causalidad, pero representan un gran avance en la comprensión holística de aquello que se pretende medir, en el que el desarrollo metodológico propuesto por Brandmaier *et al.*, (2013) puede representar un paso todavía más significativo a explorar.

Como indican Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006) la validez es el concepto más importante, pero el peor comprendido en psicometría. En educación la situación no es mejor. No obstante, si hay que considerar la validez desde una perspectiva unitaria como un concepto en el que se integran o hacia el que convergen las diferentes fuentes de evidencia que apoyan científicamente la interpretación que se hace de

las puntuaciones recogidas con un instrumento de medida, las técnicas de segmentación jerárquica pueden representar el correlato metodológico que hace posible la plasmación de esta convergencia en un solo tipo de análisis, respetuoso al mismo tiempo con las posibles limitaciones de unos datos, que muchas veces no soportan las exigencias de los modelos matemáticos más frecuentemente empleados. Desde aquí pretendemos poner en valor la idoneidad de esta técnica y plantear la necesidad de estudiar con profundidad sus posibilidades para ser incorporada de manera sistemática como estrategia de validación, especialmente cuando se trata del uso de cuestionarios de evaluación y de conceptos complejos, como es el de «calidad global percibida» de un programa o servicio. Los ejemplos ofrecidos en este artículo se refieren a un programa educativo, pero la aplicación de la técnica como herramienta de análisis de datos en procesos de evaluación y de validación puede hacerse extensiva, en ciencias sociales y de la salud, a cualquier tipo de programa, producto o servicio que se pretenda evaluar a partir de la percepción de los usuarios internos o externos del mismo.

## NOTAS

- 1 En este trabajo se han utilizado parcialmente los datos obtenidos en la «Evaluación del Programa de Educación Financiera en la Enseñanza Secundaria Obligatoria». Dicho programa fue impulsado por la CNMV y el Banco de España de forma experimental durante el curso académico 2012-2013 y su evaluación fue coordinada por la Fundación Europea Sociedad y Educación, con la dirección y ejecución científico-técnica del grupo complutense de investigación: Medida y Evaluación de Sistemas Educativos (MESE).

## REFERENCIAS BIBLIOGRÁFICAS

- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing*. Washington: APA.
- Ang, R. P. & Goh, D. H. (2013). Predicting Juvenile Offending: A Comparison of Data Mining Methods. *International Journal of Offender Therapy and Comparative Criminology*, 57, 191-207.
- Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Applied Statistics*, 8(2), 195-202.
- Berk, R. B. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods and Research*, 34(3), 263-295.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J. y Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, 18(1), 71-86.
- Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. Belmont CA: Wadsworth.
- Campbell, D. T. & Fiske, A. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Carmines, E. & Zeller, R. (1991). *Reliability and validity assessment*. Newbury Park: Sage Publications.
- Delen, D. (2012). Predicting Student Attrition with Data Mining Methods. *Journal of College Student Retention: Research, Theory and Practice*, 13(1), 17-35.
- Escobar, M. (2007). *El análisis de segmentación: Técnicas y aplicaciones de los árboles de clasificación*. Madrid: CIS.
- Gervilla, E. y Palmer, A. (2009). Predicción del consumo de cocaína en adolescentes mediante árboles de decisión. *Revista de Investigación en Educación*, 6, 7-13.
- Gervilla, E., Jiménez López, R., Montaña, J. J., Sesé, A., Cajal, B. y Palmer, A. (2009). La metodología del Data Mining. Una aplicación al consumo de alcohol en adolescentes. *Adicciones*, 21(1), 65-80.
- Horner, S. B., Fireman, G. D. & Wang, E. W. (2010). The Relation of Student Behavior, Peer Status, Race, and Gender to Decisions about School Discipline Using CHAID Decision Trees and Regression Modeling. *Journal of School Psychology*, 48(2), 135-161.
- Kovacic, Z. J. (2012). Predicting student success by mining enrolment data. *Research in Higher Education Journal*, 15, 1-20.
- Lizasoain, L. y Joaristi, L. (2000). El análisis de datos en la evaluación de programas educativos. *Revista de Investigación Educativa*, 18(2), 357-379.
- Lizasoain, L., Joaristi, L., Santiago, C., Lukas, J. F., Moyano, N., Sedano, M. y Munárriz, B. (2003). El uso de las técnicas de segmentación en la evaluación del rendimiento en lenguas. Un estudio en la Comunidad Autónoma Vasca. *Revista de Investigación Educativa*, 21(1), 93-111.
- Luque, T. (Coord.) (2012). *Técnicas de análisis de datos en investigación de mercados*. Madrid: Pirámide (2.ª ed.).
- Márquez-Vera, C., Romero, C. & Ventura, S. (2013). Predicting School Failure and

- Dropout by Using Data Mining Techniques. *IEEE Journal of Latin-american Learning Technologies*, 8(1), 7-14.
- Martínez Arias, M. R., Hernández Lloreda, M. J. y Hernández Lloreda, M. V (2006). *Psicometría*. Madrid: Alianza.
- Morales, P., Urosa, B. y Blanco, A. (2003). *Construcción de escalas de actitudes tipo Likert*. Madrid: La Muralla.
- Pareja, J. A., Maciá, J. y Serrano, R. (2013). Rasgos fundamentales para que las innovaciones en educación sean exitosas: algunas experiencias que lo corroboran. *Revista científica electrónica de Educación y Comunicación en la Sociedad del Conocimiento*, 13(2), 324-356.
- Pérez, C. (2011). *Técnicas de segmentación: Conceptos, herramientas y aplicaciones*. Madrid: Garceta.
- Picón, E., Varela, J. y Lévy, J. P. (2004). *Segmentación de mercados. Aspectos estratégicos y metodológicos*. Madrid: Pearson Educación.
- Reeves, C. A. & Bernar, D. A. (1994). Defining quality: alternatives and Implications. *Academy of Management Review*, 19 (3), 419-445
- Román, M. V. y Lévy, J.-P. (2003) Clasificación y segmentación jerárquica. En J.-P. Lévy y J. Valera (Dir.). *Análisis Multivariable para las Ciencias Sociales* (pp. 567-630). Madrid: Pearson-Prentice Hall.
- Richard's, M. M., Solanas, A., Ledesma, R. D., Introzzi, I. M. y López Ramón, M. F. (2008). Técnicas estadísticas de clasificación: un estudio comparativo y aplicado. *Psicothema*, 20(4), 863-871.
- Schiattino, I. y Silva, C. (2008). *Árboles de clasificación y regresión: Modelos Cart*. *Ciencia y Trabajo*, 10(30), 161-166.
- Sireci, G. & Sukin, T. (2013). Test validity. En K. F. Geisinger (Ed.) *APA Handbook of Testing and Assessment in Psychology* (pp. 61-84). Washington: American Psychological Association.
- Stevens, S. S. (Ed.) (1951). *Handbook of Experimental Psychology*. Nueva York: Wiley.
- Tiwari, M., Singh, R. & Vimal, N. (2013). An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education, *International Journal of Computer Science and Mobile Computing*, 2(2), 53-57.

## PERFIL ACADÉMICO Y PROFESIONAL DE LAS AUTORAS

Ángeles Blanco Blanco. Profesora Contratada Doctora en el Departamento de Métodos de Investigación y Diagnóstico en Educación de la Universidad Complutense de Madrid desde 2009. Acreditada para el acceso al cuerpo de Profesores Titulares de Universidad desde 2012. *Visiting Scholar* en la Universidad de Maryland (2013). Miembro del grupo de investigación complutense *Medida y Evaluación de Sistemas Educativos* y del grupo de Investigación consolidado de la UNED *Sistemas de Orientación Psicopedagógica y competencias de los orientadores* desde 2007.

Inmaculada Asensio Muñoz. Doctora en Filosofía y Ciencias de la Educación por la UCM, con premio extraordinario de Doctorado. Es Profesora Titular de Universidad en el Departamento de Métodos de Investigación y Diagnóstico en Educación de la Facultad de Educación (UCM). Imparte docencia en asignaturas de Métodos de Investigación Educativa y Estadística Aplicada a las CC Sociales.

Elvira Carpintero Molina. Profesora Contratada Doctora en el departamento de Métodos de Investigación y Diagnóstico en Educación en la Facultad de Educación de la Universidad Complutense de Madrid. Miembro del grupo de investigación complutense *Medida y Evaluación de Sistemas Educativos* y del grupo de investigación *Pedagogía Adaptativa*. Ha participado en diversos proyectos de investigación (I+D) relacionados con la evaluación del sistema educativo.

Covadonga Ruiz de Miguel. Doctora en Filosofía y Ciencias de la Educación por la UCM, con premio extraordinario de Doctorado. Es Profesora Titular de Universidad en el Departamento de Métodos de Investigación y Diagnóstico en Educación de la Facultad de Educación (UCM). Imparte docencia en asignaturas de Métodos de Investigación y Estadística Aplicada a las CC Sociales en los Grados de Infantil y Pedagogía y es Coordinadora del Grado en Pedagogía de la UCM.

Eva Expósito Casas. Profesora Ayudante Doctora en el departamento de Métodos de Investigación y Diagnóstico en Educación II de la Universidad Nacional de Educación a Distancia desde 2011. Es miembro del grupo de investigación complutense *Medida y Evaluación de Sistemas Educativos* (Grupo MESE) desde el año 2007 y miembro del Grupo de Investigación en *Sistemas de Orientación Psicopedagógica y competencias de los orientadores* (GRISOP) desde 2011.

Dirección de las autoras: Ángeles Blanco Blanco  
Inmaculada Asensio Muñoz  
Elvira Carpintero Molina  
Covadonga Ruiz de Miguel  
Departamento de Métodos de Investigación  
y Diagnóstico en Educación  
Facultad de Educación  
Universidad Complutense de Madrid  
Avda. Rector Royo Villanova, s/n  
28040 Madrid  
E-mail: [ablancob@ucm.es](mailto:ablancob@ucm.es)  
[macu@edu.ucm.es](mailto:macu@edu.ucm.es)  
[ecarpintero@edu.ucm.es](mailto:ecarpintero@edu.ucm.es)  
[covarui@ucm.es](mailto:covarui@ucm.es)

Eva Expósito Casas  
Departamento de Métodos de Investigación  
y Diagnóstico en Educación II  
Universidad Nacional de Educación  
a Distancia  
Juan del Rosal, 14  
28040 Madrid  
E-mail: [evaexpositocasas@edu.uned.es](mailto:evaexpositocasas@edu.uned.es)

Fecha Recepción del Artículo: 28. Abril. 2015  
Fecha modificación Artículo: 22. Junio. 2015  
Fecha Aceptación del Artículo: 24. Junio. 2015  
Fecha Revisión para Publicación: 09. Marzo. 2017