anales de psicología, 2012, vol. 28, nº 3 (octubre), 1011-1020
http://dx.doi.org/10.6018/analesps.28.3.156211

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

La metodología Delphi como técnica de estudio de la validez de contenido

Beatriz Gil-Gómez de Liaño[1*] y David Pascual-Ezama[2]

[1] *Universidad Autónoma de Madrid (España)*
[2] *Universidad Complutense Madrid (España)*

Resumen: Encontrar procedimientos que nos garanticen altos niveles de validez es uno de los grandes retos de la metodología en psicología hoy en día. En el presente estudio presentamos la metodología Delphi como técnica para el estudio de la validez de contenido. Esta metodología presenta una ventaja esencial respecto a los métodos de expertos tradicionales: su aplicación es flexible y dinámica permitiendo una mayor retroalimentación entre los expertos participantes. Se aplicó la metodología Delphi en la elaboración de un cuestionario sobre el comportamiento de inversores individuales en bolsa (Pascual-Ezama, et al., 2010) y se calcularon los *índices de congruencia* (Rovinelli y Hambleton, 1977) a lo largo de las distintas fases del Delphi. Los resultados ponen de manifiesto que la metodología Delphi puede ser una buena herramienta que nos permita garantizar altos niveles de validez de contenido en un cuestionario. Se discuten también las ventajas e inconvenientes de los métodos de expertos habituales respecto a la metodología Delphi, y cómo el presente estudio aúna metodologías esencialmente cuantitativas con metodologías más cualitativas como el método Delphi. Este nexo entre métodos tan dispares es, sin duda, uno de los campos a explorar y desarrollar en la metodología actualmente.
Palabras clave: Validez de Contenido, Metodología Delphi, Métodos de Expertos, Metodologías Cuantitativas y Cualitativas.

Title: The Delphi Method as a technique to study Validity of Content.
Abstract: Finding procedures that may guarantee high levels of validity is one of the big challenges in methods in psychology nowadays. In the present study, we show the Delphi method as a technique to study the validity of content. This method shows an important advantage regarding to traditional expert methods: it is flexible and dynamic in its application, allowing the presence of feedback between the expert participants. We applied the Delphi method during the development of a questionnaire of individual investors' behavior in the stock exchange (Pascual-Ezama, et al., 2010), also calculating *congruency indexes* (Rovinelli y Hambleton, 1977) along the different stages of the Delphi application. The results show that the Delphi method may be a good tool in order to guarantee high levels of validity of content in a questionnaire. Advantages as well as disadvantages of the traditional expert methods and the Delphi method, and how the present study joins quantitative and qualitative methods like the Delphi, are discussed. This link between those different types of methods is, with no doubt, one of the fields to explore and develop in methodology nowadays.
Key words: Validity of Content, Delphi Method, Expert Methods, Quantitative and Qualitative Methods.

Introducción

Dentro del estudio de las propiedades psicométricas de los cuestionarios, los conceptos más importantes que se han tratado de desarrollar han sido esencialmente los conceptos de fiabilidad y validez. Sin embargo, así como el estudio de la fiabilidad ha sido ampliamente desarrollado (p.ej., Furr y Bacharach, 2008; Muñiz, 1998; Spearman, 1904, 1907), la validez es uno de los grandes retos de la metodología en psicología. Probablemente uno de los problemas que entraña el estudio de la validez en psicología en general y particularmente en el estudio de las propiedades psicométricas de los cuestionarios ha sido la variedad de acepciones que existen respecto al término de validez, y cómo éstas han ido cambiando a lo largo del tiempo (p.ej., Anastasi y Urbina, 1997; Cronbach, 1988). Podemos encontrar descritos en la literatura conceptos tales como *validez aparente*, *validez de contenido*, *validez de constructo*, *validez discriminante*, *validez orientada a un criterio*… (p.ej., Leon y Montero, 2003). Todas ellas hacen alusión al concepto general de validez recogido por tres importantes asociaciones de profesionales americanas (AERA, APA y NCME) en los *Standards for Educational and Psychological Testing* en su edición de 1999, que define la validez como el grado en que la teoría y los datos disponibles apoyan la interpretación de las puntuaciones de un test para un uso concreto. Es decir, la validez debe garantizar que cualquier estudio llevado a cabo en psicología o cualquier cuestionario nos dé garantías de que mide lo que realmente dice que está midiendo y que sirve para los propósitos de la investigación para los que fue elaborado. De esta manera, se definen las distintas acepciones al término de validez haciendo alusión a la manera en que los psicólogos tratan de dar respuesta a dicha pregunta; desde, por ejemplo, su funcionalidad (validez aparente), hasta la posibilidad de establecer relaciones con aquello que la teoría especifica que debe relacionarse (validez de constructo), pasando por la evaluación de si estoy incluyendo todos aquellos elementos representativos de todos los aspectos que se pretenden medir (validez de contenido), etc.

La Validez de Contenido y los Métodos de Expertos

Concretamente, en el caso de la *validez de contenido* (entendida como el grado en el que se incluyen todos los elementos representativos de lo que se pretende medir) se han propuesto diversas formas para evaluar si un estudio o un cuestionario tienen alta validez de contenido. Aunque hay varios procedimientos para valorar la validez de contenido de un cuestionario, como la aplicación de técnicas de escalamiento multidimensional, análisis cluster (Deville, 1996; Oltman, Stricker y Barrows, 1990) análisis factorial (Dorans y Lawrence, 1987), modelos de ecuaciones estructurales (Ding y Hershberger, 2002), etc; el procedimiento más habitual se basa en el trabajo con jueces expertos en el tema que se pretende estudiar (Sireci, 1998). Se han desarrollado diversos índices basados en el trabajo con expertos para valorar la validez de contenido. Uno de los más utilizados es el *índice de congruencia* propuesto por Rovinelli y Hambleton (1977). Cada juez valora si el contenido de cada ítem incluido en el cuestionario refleja los objetivos especificados; "1"

[1] *** Dirección para correspondencia [Correspondence address]**: Beatriz Gil-Gómez de Liaño. Dpto. Psicología Social y Metodología (Facultad de Psicología). C/ Ivan Pavlov, 6. Carretera de Colmenar, km 15 (Campus de Cantoblanco, UAM). 28049 Madrid (España).
E-mail: bgil.gomezdelianno@uam.es

si el contenido está claramente especificado, "-1" si cree que no lo mide y "0" si existen dudas sobre si lo mide o no. De esta manera se elabora una tabla con los expertos y las valoraciones que los mismos hacen a cada uno de los ítems del cuestionario y se calcula un índice por cada ítem a partir de una fórmula matemática (Rovinelli y Hambleton, 1977). También se utilizan para la valoración de la representatividad de los ítems preguntas con formato de respuesta tipo Likert de 5 puntos, donde la tarea del juez es evaluar la relevancia de cada ítem para medir el objetivo propuesto; desde "nada relevante" hasta "totalmente relevante". La media y/o la mediana de las valoraciones de cada ítem son utilizadas como el *índice de relevancia* del ítem (Hambleton, 1980, 1984). Para una revisión sobre el tema, ver Abad, Olea, Ponsoda y García (2011).

Sin embargo, este tipo de procedimientos conlleva una serie de problemas. Por un lado, se trata de un proceso "estático" que no permite a los expertos ningún tipo de argumentación sobre la puntuación dada a un determinado ítem. Es decir, se elabora el cuestionario y se pasa la tabla cruzada de ítems y jueces para que cada uno de los expertos rellene con un dato numérico si considera que el ítem es representativo del objetivo definido. A continuación, se calculan los índices de cada ítem y se establece un criterio de inclusión; por ejemplo, si el ítem tiene un índice de congruencia inferior a 0,75 será eliminado del cuestionario. Finalmente, se eliminan aquellos ítems que no cumplen el criterio y se mantienen los que sí lo cumplen, conformándose el cuestionario final con los ítems que cumplen los criterios exigidos. Es decir, se trata de un proceso de elaboración del cuestionario que se detiene en ese momento; los expertos no reciben *feedback* sobre lo que los otros expertos han considerado, ni pueden especificar criterios sobre porqué consideran que un determinado ítem no mide lo que los objetivos marcan. Con este tipo de procedimientos no existe la posibilidad de cambiar alguna respuesta tras valorar otras posibles argumentaciones. Se trata de un proceso "estático" que termina en el momento en que se calculan los índices y se decide la inclusión o no de los distintos ítems propuestos.

Por otro lado, con los índices de congruencia y relevancia, al informar a los jueces sobre lo que se supone que mide el test, estamos restringiendo sus evaluaciones a las dimensiones propuestas y, por tanto, influenciando sus percepciones sobre lo que mide el ítem. Aunque se han propuesto métodos de escalamiento multidimensional y análisis de conglomerados para tratar de solucionar este problema (Sireci y Geisinger, 1992, 1995), seguimos restringiendo el proceso a un mero análisis cuantitativo de la validez de contenido, que termina, como hemos señalado anteriormente, en la obtención de dichos indicadores y en la decisión de incluir o excluir el ítem del cuestionario.

La Metodología Delphi

Sin embargo, los métodos de expertos no están limitados en este sentido. Podrían utilizarse de una manera más flexible dentro de un proceso dinámico de cambio, retroalimentación y toma de decisiones a lo largo de la elaboración del cuestionario, que enriquezca la confección del mismo con el objetivo de garantizar una mayor validez de contenido. De hecho, la *metodología Delphi* es una técnica enmarcada dentro de los métodos de expertos que se utiliza para obtener la opinión más consensuada posible de un grupo de personas, consideradas expertos, en relación con un determinado objetivo de investigación (Leon y Montero, 2003). Se trata de uno de los métodos de expertos más conocidos y contrastados (Landeta, 1999) que ha demostrado una gran eficacia siempre que se ha utilizado (Christie y Barela, 2005; Okoli y Pawlowski, 2004). Aunque el objetivo original de la metodología Delphi fue el de realizar pronósticos sobre hechos futuros, se ha utilizado en multitud de ámbitos en psicología (p.ej., Hernández, 1996; Leon, 2001). Su origen surge en 1950 cuando la Rand Corporation realizó un estudio para la fuerza aérea de EE.UU que llamó "Proyecto Delphi". De ahí derivó la metodología su nombre. El objetivo del estudio fue obtener el mayor consenso posible de un grupo de expertos sobre una serie de cuestiones presentadas a través de cuestionarios intensivos, a los cuales se les intercalaba una retroalimentación controlada.

Tabla 1. Características Principales del Método Delphi (Landeta, 1999)

Anonimato
Durante un Delphi, cada experto desconoce la identidad de los demás integrantes del panel. Esto tiene una serie de aspectos positivos: • *Impide la posibilidad de que un miembro del grupo sea influenciado por la reputación de otro de los miembros o por el peso que supone oponerse a la mayoría. La única influencia posible es la de la congruencia de los argumentos.* • *Permite que un miembro pueda cambiar sus opiniones sin que eso suponga una pérdida de imagen.* • *El experto puede defender sus argumentos con la tranquilidad que da saber que en caso de que sean erróneos, su equivocación no va a ser conocida por los otros expertos.*
Iteración y realimentación controlada
La iteración se consigue al presentar varias veces el mismo cuestionario. Como además se van presentando los resultados obtenidos con los cuestionarios anteriores, se consigue que los expertos vayan conociendo los distintos puntos de vista y puedan ir modificando su opinión si los argumentos presentados les parecen más apropiados que los suyos. Tanto las posturas minoritarias como las mayoritarias tienen presencia en los resultados finales.
Respuesta del grupo
La información que se presenta a los expertos no es sólo el punto de vista de la mayoría, sino que se presentan todas las opiniones indicando el grado de acuerdo que se ha obtenido.
Heterogeneidad
Pueden participar expertos de diferentes ramas de actividad sobre las mismas bases o "reglas de juego".

Según Linstone y Turoff (1975), la metodología Delphi se define como "*un método para estructurar el proceso de comunicación grupal, de modo que ésta sea efectiva para permitir a un grupo de individuos, como un todo, tratar con problemas complejos*". El desarrollo de la metodología sigue un programa, cuidadosamente elaborado, constituido por una secuencia de preguntas a las que se contesta de forma individual y sobre las que los participantes reciben sistemáticamente una retroalimentación. Dicha retroalimentación, elaborada a partir de las respuestas obtenidas, orienta la siguiente batería de preguntas (Linstone y Turoff, 1975). En la Tabla 1 podemos ver un resumen de las principales características del método Delphi (Landeta, 1999).

Objetivos del Presente Estudio

El presente estudio trata de garantizar cualidades propias de las metodologías cuantitativas, concretamente la validez de contenido en la elaboración de cuestionarios, a través de la utilización de una metodología esencialmente cualitativa: la metodología Delphi. Aunque los métodos de expertos son habituales en el estudio de la validez de contenido (tal y como hemos visto anteriormente), e incluso en el contexto médico ya se ha empleado la metodología Delphi en la construcción de cuestionarios (p.ej. Kröger et al., 2007; Wenger, et al., 2001; excepcionalmente en un contexto más psicológico: Rodríguez-Carballeira et al., 2010), hasta donde llega nuestro conocimiento no se ha planteado todavía la metodología Delphi como herramienta para garantizar y valorar la validez de contenido en el proceso de elaboración de cuestionarios en psicología. Esto nos garantizaría un proceso dinámico de retroalimentación y toma de decisiones cambiante a lo largo de las distintas etapas de la implantación del Delphi, donde todos los expertos opinan para tratar de llegar a un consenso final en la elaboración del cuestionario, cosa que no ocurriría según las técnicas de expertos habitualmente utilizadas. De esta manera, esperamos que el test definitivo obtenido a través del consenso potenciado por la metodología Delphi pudiera generar un cuestionario más manejable en su longitud y garantizando una alta validez de contenido. Por tanto, hemos aplicado una adaptación de la metodología Delphi en el proceso de elaboración y validación de un cuestionario sobre variables que afectan al comportamiento de los inversores individuales en los mercados bursátiles ININBE (Pascual-Ezama, San Martín, Gil-Gómez de Liaño y Scandroglio, 2010). También hemos aplicado métodos tradicionales de expertos utilizados en la elaboración de cuestionarios (concretamente hemos calculado el índice de congruencia) basados en metodologías más cuantitativas, para estudiar su evolución a lo largo de las distintas etapas del Delphi.

Método

Participantes

Se seleccionaron inicialmente 160 expertos para participar en el proceso de elaboración del cuestionario. Contactamos con profesionales tanto del área financiera como más específicamente del sector de inversión bursátil, así como profesores universitarios especialistas en dichas áreas de trabajo. De todos ellos, 26 aceptaron participar en el mismo y 17 de ellos participaron durante todo el proceso. De estos últimos, el 29% pertenecen al ámbito académico y un 71% al empresarial. Fueron fundamentalmente hombres (71%), españoles (59%), que trabajan en España (76%) y con estudios de administración y dirección de empresas (41%), licenciados en ingeniería (29%) y licenciados en derecho (12%) esencialmente. De todos ellos, el 59% son expertos en el campo de inversión, mientras que el 41% restante son expertos en el sector financiero.

Una vez elaborado el cuestionario, contestaron voluntariamente al mismo 257 inversores individuales, con una edad media de 35 años, 40% mujeres y 60% varones. El grado de formación de la muestra abarcó desde niveles básicos universitarios a formación de máster MBA. La experiencia media como inversores fue de más de 8 años, con una cantidad media de dinero invertido de más de 20.000 euros y un tiempo medio de inversión aproximado de 15 meses. Estos resultados van en consonancia con el perfil de inversores encontrados por Perera y Toharia (2006) sobre una muestra de inversores en los mercados españoles.

Instrumento

A los expertos se les facilitó una lista inicial de 70 ítems, que fue confeccionada con ítems empleados en estudios anteriores entre los que se consideraron los de Baker y Haslem (1973), Clark-Murphy y Soutar (2004), Fundación de Estudios Financieros (2003), Lease, Lewellen y Schlarbaum (1974), Nagy y Obenberger (1994), Potter (1971), Rogers y Grant (1998), Warren, Stevens, y McConkey (1990),

El cuestionario ININBE quedó constituido por tres cuestiones referentes a datos personales y por una escala de 47 ítems (Pascual-Ezama, et al., 2010).

Procedimiento

La metodología utilizada para elaborar la lista de ítems definitiva fue una adaptación de la metodología Delphi. Los expertos que participaron tenían que realizar una revisión de la lista inicial de ítems seleccionando o proponiendo nuevos ítems considerados adecuados y no contemplados en la lista inicial, y excluyendo los que, por el contrario, consideraban que no serían utilizados por los inversores individuales en los mercados bursátiles. Previamente a cada una de las fases se envió a los expertos, junto con la lista inicial de ítems, unos *supuestos básicos iniciales* (Tabla 2) y la *metodología de trabajo*

(Tabla 3) que deberían seguir para el buen funcionamiento del proceso. A lo largo de las distintas fases de la metodología Delphi, se fueron seleccionando los distintos ítems que finalmente integraron el cuestionario final. Con el objetivo de valorar si el cuestionario final elaborado a través de la metodología Delphi hubiera generado el mismo cuestionario que utilizando un método tradicional de expertos, como los anteriormente comentados, se pidió a los expertos que valorasen en la primera fase del Delphi los distintos ítems de la siguiente manera: cada juez debía valorar si el contenido de cada ítem incluido en el cuestionario reflejaba los objetivos especificados; "1" si el contenido está claramente especificado, "-1" si cree que no lo mide y "0" si existen dudas sobre si lo mide o no. Además, con el objetivo de estudiar la evolución de dicho proceso se pidió a los expertos que también valorasen los ítems en las sucesivas fases del Delphi. De esta manera, podemos calcular el índice de congruencia descrito anteriormente y valorar si la inclusión de los ítems finales del test varía substancialmente respecto a la metodología Delphi, y si las decisiones se ven influidas por la retroalimentación y flexibilidad propias del método Delphi a lo largo de las distintas etapas del mismo.

Tabla 2. Supuestos Iniciales de Trabajo Propuestos a los Expertos en la Metodología Delphi

Se envío a cada uno de los expertos un listado con los supuestos básicos iniciales. El listado que se les envío contenía los siguientes supuestos:
Supuesto I
En nuestro caso los inversores ya han tomado la decisión de invertir en renta variable y comprar acciones de las empresas. Por lo tanto, debemos decidir qué ítems son los que podrían influir en la decisión de comprar acciones de una empresa u otra, pero se asumirá que la decisión de comprar acciones en bolsa ya está tomada después de estudiar otras alternativas.
Supuesto II
Partimos de una lista de 70 ítems obtenidos de varios estudios anteriormente realizados. Sería más que recomendable que el número de ítems no pasara de 50, siendo 40 un número ideal.
Supuesto III
En esta primera lista se han incluido ítems que en ningún caso deberían aparecer y otros que se puede considerar que están repetidos. Esto se ha hecho de manera consciente para poder tener una referencia inicial para evaluar la aportación de los expertos.
Supuesto IV
La toma de decisiones se realizará en relación a empresas de un mismo mercado. Como la muestra se hará en España, supongamos que la decisión de invertir será dentro del mercado español de valores, sea mercado continuo o cualquiera de las bolsas nacionales.

Tabla 3. Procedimiento de Trabajo Propuesto a los Expertos en la Metodología Delphi

A cada uno de los expertos se le envío el procedimiento de trabajo que decía lo siguiente:
El procedimiento para desarrollar el proceso de consenso será sencillo y no será necesario invertir apenas tiempo por parte de los participantes. La intención de este trabajo de colaboración es que ninguno de los colaboradores tenga que realizar ningún trabajo específico sino que simplemente aporte sus conocimientos. Para el buen funcionamiento del proceso, ninguno de los colaboradores conocerá la identidad del resto, al menos, hasta el final del proceso. La forma de trabajar será sencilla:
Paso I
Se enviará a cada colaborador una lista inicial de ítems tomados de algunos estudios previos realizados en otros países y algunos ítems que se han considerado no estaban reflejados en dichos estudios que podrían tener influencia en la decisión del inversor.
Paso II
Cada experto de manera independiente añadirá o eliminará los ítems que considere oportuno. En caso de que crea que alguno de los ítems iniciales sea repetitivo o simplemente no influya en la decisión de inversión, deberá justificar brevemente el porqué de la exclusión.
Paso III
Cuando cada uno de los colaboradores envíe sus opiniones, se recopilara la información, componiéndose una nueva lista con todos aquellos ítems nuevos y un anexo con los excluidos por alguno de los colaboradores con la pertinente justificación, para que el resto pueda opinar tanto a favor como en contra y tanto sobre las inclusiones como sobre las exclusiones.
Paso IV
El proceso se repetirá durante cuatro semanas hasta intentar llegar a un consenso en el grupo de trabajo. En caso de que, tras las cuatro semanas no se llegara a consenso, realizaremos un análisis del grado de acuerdo.
Paso V
El tiempo de recepción y entrega será el siguiente: • *Se enviará la lista cada lunes a primera hora.* • *Cada colaborador tendrá hasta el viernes antes de las 12:00 de esa misma semana para enviar su respuesta.* • *Durante el fin de semana se realizará el documento con la nueva lista, inclusiones y exclusiones.* • *El nuevo documento será enviado el lunes a primera hora.* • *El proceso se repetirá a lo largo de cuatro semanas.*
Nota final importante
Como todos suponéis, esta es una tarea muy ambigua por la gran diversidad de perfiles inversores que existe y los factores que afectan a unos no son las mismas que influyen sobre otros. Por ese motivo ruego a todos los colaboradores una visión amplia de la situación. Por ejemplo, uno de los ítems iniciales será "ultimas cotizaciones de la acción". Lógicamente un inversor no actúa de igual manera con la cotización de los últimos 3 días que con la de los últimos 300. Si es especulador tomara la de los últimos 3 días y si realiza análisis chartista para invertir a largo plazo tendrá muy en cuenta, además de los últimos días recientes, la del último año o años. Teniendo en cuenta esto se intentará que las variables sean "genéricas", aún a riesgo de que los inversores individuales que completen los test puedan tener distintas interpretaciones de cada ítem.

En la primera fase de la metodología Delphi se pidió a los expertos que seleccionaran aquellos ítems que según ellos eran utilizados por los inversores individuales y propusieran aquellos que consideraban importantes y no estaban inclui-

dos en la lista inicial propuesta, justificando sus decisiones. En una segunda fase, y tras recoger los resultados de la primera, se informó a cada uno de los expertos sobre las respuestas y las justificaciones dadas por el resto de los participantes. Para ello se les mostraron tres listas: una de ítems seleccionados por la mayoría de los expertos (entendiendo por mayoría más del 50% de los expertos); otra con aquellos ítems que la mayoría de los expertos consideraron no utilizados por los inversores individuales y una tercera con ítems propuestos por los expertos que no estaban en la lista inicial. A continuación se les pidió que, en función de estos resultados, volvieran a dar su opinión sobre cuáles creían que eran los ítems más utilizados por los inversores individuales en los mercados bursátiles. En la tercera y última fase, se repitió el procedimiento anterior. Tanto los ítems finalmente incluidos como los excluidos fueron seleccionados por más del 75% de los expertos. Esto fue debido a que el acuerdo de los expertos sobre aquellos ítems que debían permanecer en la lista final y los que no, fue aumentando de una fase a otra, tal y como se puede observar en la Figura 1. Por tanto, de los 70 ítems que constituían la lista inicial fueron seleccionados 41 (37 de los cuales pertenecían a la lista inicial propuesta y otros 4 fueron añadidos por los expertos). El resto de ítems, hasta llegar a los 47 finales, fueron añadidos por los autores, como ítems de control.

Finalmente y una vez aplicado el cuestionario a la muestra de inversores, se calculó la fiabilidad del cuestionario y algunas de las evidencias de validez más habituales como las basadas en contenidos o en la estructura interna (Muñiz, 2003) a través de un Análisis Factorial Exploratorio (AFE). Por último y una vez elaborado el cuestionario final, otros diez expertos (distintos de los anteriores y seleccionados en base a los mismos criterios) fueron consultados con el objetivo de validar el contenido del cuestionario a posteriori (Calvo y Díaz, 2004).

Resultados

Resultados derivados de la aplicación del método Delphi

Fase I

En la primera fase de la metodología Delphi fueron incluidos por parte de los expertos 25 de los 70 ítems de la lista inicial, mientras que los 45 restantes fueron excluidos. Por su parte, los expertos propusieron 31 ítems que no estaban en la lista inicial que se les facilitó y que consideraban debían formar parte de la lista final.

Fase II

Tras el primer *feedback* y ya en la segunda fase, el número de ítems excluidos disminuyó a 40 y el de ítems que debían permanecer en la lista final aumentó a 30, es decir, 5 de los ítems que habían sido excluidos en una primera fase, entraron a formar parte de la lista de ítems. Respecto a los 31 ítems propuestos por los expertos para formar parte de la lista final y que no estaban en la lista inicial, tan sólo 5 de ellos fueron incluidos en la lista en esta segunda fase.

Fase III

Una vez enviado de nuevo el *feedback* a los expertos sobre los resultados de la segunda fase, en la tercera y última fase, el número de ítems que los expertos decidieron que debían permanecer en la lista final, respecto a los 70 ítems iniciales propuestos, fue de 37. A estos 37 ítems hay que añadirles 5 que no estaban en la lista inicial y que fueron propuestos por los expertos e incluidos en la lista definitiva (los mismos 5 propuestos en la segunda fase).

En la Figura 1 se resumen los resultados del proceso a lo largo de las distintas fases del Delphi.

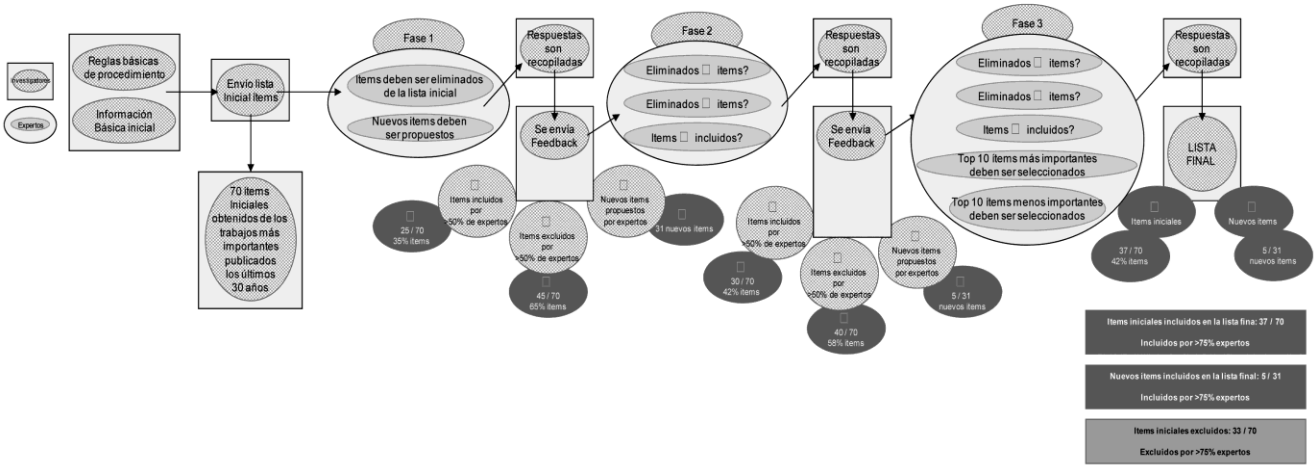

Figura 1. Metodología Delphi. Resultado de las fases de trabajo

Índices de Congruencia, I_{ik}

Por otro lado, calculamos el índice de congruencia en las tres fases del proceso de la aplicación de la metodología Delphi. Si tenemos en cuenta el criterio de incluir aquellos ítems con un índice de congruencia I_{ik}>.50 (Osterlind, 1989), de los ítems finalmente incluidos en el cuestionario, todos ellos cumplieron dicho criterio a lo largo de las tres fases de la aplicación del Delphi. Además, el índice I_{ik} medio de estos ítems finalmente seleccionados aumentó significativamente en las tres fases [$F(2,72) = 10.27$; $p < .001$; $\eta^2 = .22$]. La media del I_{ik} para la primera fase fue de .79; para la segunda .89 y para la tercera .86. El aumento fue significativo entre la fase primera y las fases segunda y tercera ($p < .05$ en ambos casos), mientras que no hubo diferencias entre la segunda y la tercera fase ($p = .14$). Sin embargo, de los ítems que finalmente se excluyeron según la metodología Delphi, cinco de ellos habrían sido incluidos también si hubiéramos utilizado el índice de congruencia en la primera fase del Delphi. Todos los ítems excluidos finalmente (incluyendo estos cinco ítems que sí hubiéramos seleccionado si sólo hubiéramos tenido en cuenta los índices de congruencia en la primera fase del proceso), también fueron valorados en menor medida en las fases segunda y tercera del Delphi. De hecho, la media del I_{ik} en la primera fase para los ítems excluidos fue .006, en la segunda -.56 y en la tercera -.48. Al igual que en el grupo de ítems incluidos, el índice medio disminuyó significativamente entre las tres fases [$F(2,52) = 74.05$; $p < .001$; $\eta^2 = .74$], siendo esta bajada significativa entre la primera y las otras dos fases ($p < .001$ en ambos casos). Nuevamente, no hubo diferencias significativas entre las fases segunda y tercera ($p = .31$) para los ítems excluidos.

Estudio psicométrico del cuestionario

Fiabilidad

El coeficiente Alfa del cuestionario no aumentó substancialmente al eliminar cualquiera de los 47 ítems que la conformaban. Por otro lado, siendo los índices de homogeneidad índices de correlación, y éstos significativamente distintos de cero cuando $|r|\sqrt{n} \geq 1.96$ (Pardo y San Martín, 2004), no encontramos en nuestro estudio ningún ítem que debiera ser eliminado. Por tanto, el cuestionario final tras el análisis inicial de ítems quedó compuesto por los 47 ítems considerados inicialmente. Finalmente, obtuvimos un índice de consistencia interna bastante alto tanto para el test global (alpha de Cronbach = .934), como para cada una de las dimensiones encontradas en el AFE (Pascual-Ezama et al., 2010).

Evidencias de validez basada en la estructura interna del cuestionario: AFE

La validez de los resultados de un AFE viene condicionada por obtener valores significativos – próximos a la unidad – del índice de «adecuación muestral» de Kaiser-Meyer-Olkin (KMO) y el test de esfericidad de Bartlett ($p < .05$) que se obtienen de la matriz de correlaciones. En nuestro caso, se obtuvo un índice KMO de .786, que corresponde a un valor medio según el baremo de Kaiser (Kaiser, 1974), y una $p \leq .001$ en la prueba de esfericidad de Bartlett. Todo ello nos permitió concluir que las correlaciones entre las variables son significativas y justifican el análisis factorial (Ochoa, Repáraz y Polaino-Lorente, 1997). Por otro lado, el estudio de las comunalidades nos muestra que todos los ítems presentan pesos superiores a .30, por lo que parece justificado llevar a cabo dicho análisis factorial.

El AFE generó seis factores que, en su conjunto, explican aproximadamente el 54% de la varianza total. A pesar de que si para la extracción de dichos factores hubiéramos utilizado criterios estadísticos como el criterio de Kaiser ($\lambda \geq 1$) el resultado serían 12 factores, tan sólo los seis primeros muestran un sustento teórico lo suficientemente robusto como para mantenerlos. En la Tabla 4 podemos ver el porcentaje de varianza explicada por cada uno de los factores y en la Tabla 5 podemos ver los pesos de cada uno de los ítems en su factor. Para una mayor explicación teórica sobre cada uno de los factores remitimos al lector a Pascual-Ezama et al. (2010). Lo que nos interesa en este caso es comprobar si el resultado obtenido tras el análisis factorial tiene coherencia teórica con los resultados de estudios anteriores. Pues bien, las cuatro dimensiones encontradas sistemáticamente en la literatura y que entendíamos que deberían aparecer en nuestro trabajo para evidenciar validez en la estructura interna del cuestionario, son cuatro de las seis dimensiones encontradas al realizar el AFE. Por lo tanto, la relación que mantienen los ítems del cuestionario entre sí, parece conformar dimensiones consistentes con el marco conceptual en el que está basado el cuestionario.

Tabla 4. Varianza Total Explicada

Componente	Autovalores iniciales		
	Total	% de la varianza	% acumulado
1	11.94	25.41	25.41
2	3.99	8.50	33.91
3	3.03	6.44	40.36
4	2.33	4.96	45.32
5	2.09	4.45	49.77
6	1.93	4.11	53.89
7	1.53	3.27	57.16
8	1.33	2.83	60.00
9	1.25	2.66	62.66
10	1.20	2.56	65.22
11	1.13	2.40	67.63
12	1.11	2.36	69.99

Tabla 5. Matriz de Factores Rotados.

	Factores					
Items	I	II	III	IV	V	VI
Ratios comparables entre sectores y compañías	.861					
Análisis fundamental comparado entre compañías	.821					
Análisis por fundamentales	.561					
Situación de la empresa dentro del sector	.555					
Expectativas de crecimiento continuado	.550					
Uso de métodos de valoración	.542			.596		
Percepción interna sobre la economía		.724				
Indicadores económicos actuales		.597				
Endeudamiento de la empresa		.589				
Datos financieros relevantes (*Cash Flows, VAN, TIR*, etc..)		.578				
Balance y cuenta de *PyG*		.547				
Información publicada en la página *WEB*		.502				
Recomendaciones de las casas de *brokers*			.830			
Recomendaciones de expertos bursátiles			.793			
Recomendaciones de asesores en inversiones			.786			
Información prensa económica			.537			
Rendimiento en el pasado de los títulos				.762		
Rendimiento pasado de la cartera de inversión				.750		
Cambios significativos en los órganos de gestión					.658	
Participaciones institucionales					.564	
Planes de inversiones mediante endeudamiento					.551	
Variaciones bruscas de la cotización por hechos puntuales					.510	
Volumen de negociación						.756
Liquidez de los títulos						.645
Beneficios recurrentes u operativos esperados.						.542
Plazo de inversión						.528
Porcentaje de *free float*						.529

Evidencias de validez medidas a través de la evaluación del cuestionario *a posteriori* por parte de otros expertos

Tal y como podemos observar en la Tabla 6, los diez expertos que participaron para validar el contenido del cuestionario una vez elaborado (como ya hemos comentado anteriormente, distintos de los que participaron en el proceso de elaboración del mismo a través de la metodología Delphi) estuvieron de acuerdo en que el cuestionario es un instrumento válido y su denominación se ajusta al contenido en términos generales: nueve de ellos opinan que el número de ítems que componen el cuestionario es suficiente y ocho de ellos que están incluidos todos los elementos representativos de los aspectos que se pretenden medir, es decir, aquellos factores que afectan a la conducta de los inversores individuales al invertir en los mercados bursátiles.

Tabla 6. Validez del Contenido por Juicio de Expertos tras la Selección Final de Ítems del Cuestionario

Pregunta	**SI**	**NO**
Su denominación se ajusta al contenido	10	0
Se explora el universo completo del constructo	8	2
El número de ítems es suficiente	9	1
Es en apariencia un instrumento valido	10	0
Total	37 (92.5%)	3 (7.5%)

Discusión general y conclusiones

El objetivo principal del presente trabajo ha sido tratar de garantizar una alta validez de contenido en la elaboración de un cuestionario, basándonos esencialmente en un proceso dinámico de retroalimentación y toma de decisiones cambiante, como es el método Delphi.

Por un lado, podemos afirmar que la validez de contenido es alta. Como hemos visto, las decisiones de los expertos se han ido afianzando cada vez más a lo largo de las distintas fases del Delphi (sobre todo entre las dos primeras fases), lo que podemos ver reflejado en el aumento significativo de los índices de congruencia para los ítems incluidos y la disminución significativa de los mismos para los ítems excluidos. Es decir, no sólo se ve claro que la metodología Delphi está funcionando correctamente (pues al fin y al cabo para lo que fue diseñada fue para tomar decisiones y llegar a un consenso y a un acuerdo común en un grupo respecto a una determinada decisión), sino que parece que sirve para aumentar la seguridad en las decisiones tomadas respecto a los objetivos propuestos (la selección de los ítems más relevantes en la elaboración del cuestionario). Por otro lado, parecen existir evidencias de una alta validez tanto del contenido del cuestionario como de la estructura interna del mismo. Respecto a la validez de contenido, se alcanzó un 90% de acuerdo entre los expertos que evaluaron el cuestionario *a posteriori*. Si bien es cierto que uno de los expertos considera el número de

ítems insuficiente y dos opinan que no se incluyen todos los aspectos que se pretenden medir, consideramos que es un índice de acuerdo muy alto, sobre todo si tenemos en cuenta la complejidad de la conducta que tratamos de estudiar. Además, el grado de acuerdo de los expertos en la elaboración del cuestionario mediante la metodología Delphi, fue también muy alto, tal y como acabamos de señalar.

Respecto a la validez basada en la estructura interna del cuestionario, obtenemos los cuatro factores esperados teóricamente *a priori*, entre los seis primeros factores encontrados en el AFE. Estos seis primeros factores explican aproximadamente un 53% de la varianza, si bien es cierto que el número de factores total obtenido si tuviéramos en cuenta la regla de Kaiser sería elevado: doce factores que explicarían un 70% de la varianza total. En resumen, parece que están incluidos todos aquellos elementos encontrados en otros trabajos relacionados con el tema (para un estudio más detallado de los distintos factores a nivel teórico remitimos al lector a Pascual-Ezama et al., 2010).

Además, las propiedades psicométricas del cuestionario también han sido muy altas. La *fiabilidad* encontrada al realizar el análisis inicial de los ítems podría reflejar la eficacia de la adaptación del método Delphi (Christie y Barela, 2005; Okoli y Pawlowski, 2004), ya que ninguno de los ítems del cuestionario tuvo que ser eliminado (Pascual-Ezama et al., 2010). En este sentido, los trabajos anteriores no han empleado, o al menos no lo especifican, ningún método concreto a la hora de elaborar el instrumento o seleccionar los ítems de los cuestionarios utilizados (Clark-Murphy y Soutar, 2004; Nagy y Obenberger, 1994; Potter, 1971). La aplicación de la metodología Delphi parece haber ayudado a obtener una buena consistencia interna (alpha de Cronbach = .934).

Por último, cabe destacar, como hemos visto en los resultados, que los ítems incluidos finalmente en el cuestionario tienen índices de congruencia iguales o superiores a I_{ik}>.50 (Osterlind, 1989). Es decir, ninguno de estos ítems fue eliminado en posteriores fases del Delphi, sino todo lo contrario, aumentaron sus I_{ik} significativamente de unas fases a otras. Esto parece reflejar la efectividad tanto de los índices de congruencia utilizados tradicionalmente (Sireci, 1998), como de la metodología Delphi que acabamos de presentar, como herramienta para la elaboración del cuestionario. Ocurre lo mismo con los ítems eliminados, tal y como hemos comentado anteriormente. Sin embargo, como podemos comprobar en los resultados, de los ítems que finalmente se excluyeron según la metodología Delphi, cinco de ellos habrían sido incluidos también si hubiéramos utilizado el índice de congruencia en la primera fase del Delphi. Es decir, finalmente hubiéramos utilizado un cuestionario más largo. Aunque el principal inconveniente de esta metodología es el tiempo invertido en llevarla a cabo, merece la pena dedicar este tiempo durante el proceso de elaboración del cuestionario para garantizar una alta validez de contenido con un número suficiente de ítems, adaptado a las necesidades y el tiempo del grupo de participantes donde será finalmente administrado. Merece la pena, por tanto, dedicar algo más de tiempo en la elaboración del cuestionario para garantizar una herramienta no sólo válida sino también práctica en su aplicación. Además, teniendo en cuenta los datos del estudio, parece que con sólo dos fases de aplicación hubiera sido suficiente para llegar a las mismas conclusiones e incluir los mismos ítems, aunque en la aplicación de la metodología Delphi tradicional se especifique la necesidad de un número mínimo de tres o cuatro fases de aplicación (Linstone y Turoff, 1975).

Por tanto, teniendo en cuenta los resultados que acabamos de comentar, podemos afirmar que el objetivo de elaborar un cuestionario manejable y práctico en cuanto a su longitud y con alta validez de contenido a través del uso de la metodología Delphi ha sido llevado a cabo satisfactoriamente. La metodología Delphi parece ser un buen procedimiento enmarcado dentro de los métodos de expertos para garantizar una alta validez de contenido (los resultados de los índices de congruencia han sido coherentes con la decisión de inclusión/exclusión final de ítems a lo largo de la aplicación del Delphi), que además ha complementado el proceso habitual añadiendo flexibilidad y retroalimentación al mismo; en comparación con un método de expertos tradicional y como acabamos de comprobar, se habrían incluido un mayor número de ítems en el procedimiento tradicional respecto al Delphi utilizado.

Otro de los objetivos propuestos inicialmente ha sido estudiar la evolución del índice de congruencia a lo largo de las distintas etapas del Delphi. Como hemos podido comprobar en los resultados encontrados respecto al índice de congruencia en las distintas etapas del Delphi, el índice de los ítems finalmente seleccionados mejoró a lo largo de las distintas etapas, mientras que el de los ítems excluidos empeoró a lo largo de las distintas etapas, reflejándose, como acabamos de comentar, la efectividad de los métodos de expertos en la elaboración de cuestionarios. Nos parece interesante señalar, como una futura vía de investigación la posibilidad de estudiar indicadores de validez con el cuestionario que se hubiera generado si hubiéramos tenido en cuenta sólo el criterio de los índices de congruencia tradicionales y compararlos con la validez encontrada en el cuestionario final obtenido a través de la metodología Delphi. Ya que no disponemos de datos de los cinco ítems que no formaron parte del cuestionario final y sí lo hubieran hecho en base al criterio de los índices de congruencia propuesto por Osterlind (1989), no podemos hacer esta comparación en el presente trabajo. Sin embargo, nos parece interesante plantearlo para futuros estudios: una comparación de criterios de calidad del instrumento (tanto validez como fiabilidad) entre los dos tipos de métodos de expertos, uno más cuantitativo y otro más flexible y cualitativo como es la metodología Delphi.

Por lo tanto, y a partir de los resultados encontrados, la aplicación de la metodología Delphi parece ser una buena herramienta en la elaboración de un cuestionario práctico y aplicable, para garantizar no sólo una alta validez de contenido, sino también posiblemente otros tipos de validez, así como una alta fiabilidad. El conocimiento científico se carac-

teriza por la utilización de un método reglado para la acumulación de evidencia empírica que nos permita estudiar determinados fenómenos. Sin embargo, el estudio de determinados fenómenos puede ser muy distinto si aplicamos una metodología basada en un acercamiento más cualitativo o más cuantitativo. La dicotomía cualitativo-cuantitativo es de sobra conocida y pone de manifiesto el amplio muro que separa las metodologías cualitativas de las cuantitativas. Desde el tipo de conocimiento (subjetivo versus objetivo) hasta los criterios de valoración del proceso de investigación (reflexivo-subjetivo versus objetivo-generalización de datos), pasando por el plan de investigación o las técnicas de recogida de información, selección de la muestra y análisis de datos…, se pone de manifiesto la enorme distancia existente entre las metodologías cualitativas y las cuantitativas (Leon y Montero, 2003). Sin embargo, la importancia de aunar ambas perspectivas ha sido puesta de manifiesto por algunos investigadores que han tratado de combinar ambos tipos de aproximación en el estudio de determinados fenómenos en ciencia a través de la idea de que las carencias que puede presentar, por ejemplo una aproximación más cuantitativa podrían ser, al menos contrarrestadas en parte, aplicando una metodología más cualitativa (Kaplan y Duchon, 1988). Los resultados del presente trabajo parecen ser un claro ejemplo: la metodología Delphi parece aportar flexibilidad en el trabajo con expertos durante la elaboración de un cuestionario para garantizar una alta validez de contenido, mejorando así las carencias de los procedimientos de expertos habituales.

Referencias

Abad, F.J., Olea, J., Ponsoda, V. y García, C (2011). *Medición en ciencias sociales y de la salud.* Madrid: Síntesis.

Anastasi, A. y Urbina, S. (1997). *Psychological Testing.* New York: Macmillan.

Baker, H. K. y Haslem, J. A. (1973). Information needs of individual investors. *Journal of Accountancy*, November, 64-69.

Calvo, F. y Díaz, M.A. (2004). Apoyo social percibido: características psicométricas del cuestionario Caspe en una población urbana geriátrica. *Psicothema*, *16*(4), 570-575.

Christie, C.A. y Barela, E. (2005). The Delphi Technique as a Method for Increasing Inclusion in the Evaluation Process. *Canadian Journal of Program Evaluation, 20*(1), 105-122.

Clark-Murphy, M. y Soutar, G. (2004). Individual Investor Preferences: A Segmentation Analysis. *Journal of Behavioural Finance, 6*(1), 6-14.

Cronbach, L. J. (1988). Five perspectives on the validity argument. En H. Wainer y H. I. Braun (Eds), *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Deville, C. W. (1996). An empirical link of content and construct validity evidence. *Applied Psychological Measurement, 20*, 127-139.

Ding, C. S. y Hershberger, S. C. (2002). Assessing content validity and content equivalence using structural equation models. *Structural Equation Modeling, 9,* 283-297.

Dorans, N. J. y Lawrence, I. M. (1987). *The internal construct validity of the SAT. Research Report.* Princeton, NJ: Education Testing Service.

Fundación de Estudios Financieros. (2003). Observatorio de Gobierno Corporativo de las Grandes Sociedades cotizadas en el Mercado de Valores Español (Ibex35). *Papeles de la Fundación,7.*

Furr, R.M. y Bacharach, V. R. (2008). *Psychometrics: an introduction*. Los Angeles: Sage Publications.

Hernández, J. M. (1996). Procedimientos de recogida de información en evaluación de programas. En R. Fernández-Ballesteros (Ed.). *Evaluación de programas. Una guía práctica en ámbitros sociales, educativos y de salud* (pp. 117-147). Madrid: Síntesis.

Hambleton, R.K. (1980). Test score validity and standard-setting methods. En R. A. Berk (Ed.). *Criterion-referenced measurement the state of the art.* (pp. 80-123). Baltimore: John Hopkins University Press.

Hambleton, R.K. (1984). Validating the test scores. En R. A. Berk (Ed.). *A guide to criterion referenced test construction.* (pp. 199-230). Baltimore: John Hopkins University Press.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31-36.

Kaplan, B. y Duchon, D. (1988). Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study. *MIS Quarterly, 12*(4), 571-586.

Kröger, E., Tourigny, A., Morin, D., Côté, L, Kergoat, M.J., Lebel, P., Robichaud, L., Imbeault, S., Proulx, S. & Benounissa, Z. (2007). Selecting process quality indicators for the integrated care of vulnerable older adults affected by cognitive impairment or dementia. *BMC Health Services Research, 7 (195).* doi:10.1186/1472-6963-7-195

Landeta, J. (1999). *El metodo Delphi.* Barcelona: Ariel.

Lease, R. C., Lewellen, W. G. y Schlarbaum, G. C. (1974). The individual investor: attributes and attitudes. *Journal of Finance, 11*, 413-438.

Leon, O. G. (2001). *Tomar decisiones difíciles.* (2ª Ed.) Madrid: McGraw-Hill.

León, O. G. y Montero, I. (2003). *Métodos de investigación en psicología y educación* (3ª ed.) Madrid: McGraw-Hill.

Linstone, A. y Turoff, M. (1975). *The Delphi Method: Technique and Aplications.* Massachusetts: Addison wesley publishing.

Muñiz, J. (2003). La validación de los test. *Metodología de las ciencias del comportamiento, 5*, 119-139.

Muñiz, J. (1998). *Teoría Clásica de los Tests* (3ª ed). Madrid: Pirámide.

Nagy, R. A. y Obenberger, R. W. (1994). Factors influencing individual investor behavior. *Financial Analysts Journal, 50*(4), 63-68.

Ochoa, B., Repáraz, C. y Polaino-Lorente, A. (1997). Validación de la escala cilc, de locus de control, en una muestra española de padres de niños hospitalizados. *Psicothema, 9*(1)*,* 89-103

Okoli, C. y Pawlowski, S.D. (2004) The Delphi method as a research tool: an example, design considerations and applications. *Information y Management, 42*, 15-29.

Oltman, P. K., Stricker, L. J. y Barrows, T.S. (1990). Analyzing test structure by multidimensional scaling. *Journal of Applied Psychology, 75*, 21-27.

Osterlind, S.J. (1989). *Constructing Test Items,* Boston. Kluwer Academic Publishers.

Pardo, A. y San Martin, R. (2004). *Análisis de datos en psicología.* Madrid: Pirámide.

Pascual-Ezama, D., San Martín, R., Gil-Gómez de Liaño, B. y Scandroglio, B. (2010). Elaboración y validación de una escala sobre las principales variables que afectan a la conducta de los inversores individuales en los mercados bursátiles. *Psicothema, 22*(4), 1010-1017.

Perera, C. y Toharia, J.J. (2006). Cómo son los pequeños accionistas en España. *Bolsa*, *151*, 14-17.

Potter, R. E. (1971). An empirical study of motivations of common stock investors. *Southern Journal of Business, 6*, 41-48.

Rodríguez-Carballeira, A., Escartín Solanelles, J., Visauta Vinacua, B., Porrúa García, C., & Martín-Peña, J. (2010). Categorization and Hierarchy of Workplace Bullying Strategies: A Delphi Survey. *The Spanish Journal of Psychology, 13 (1)*, 297-308.

Rogers, R. K. y Grant, J. (1998). Content analysis of information cited in reports of sell-side financial analysts. *C.F.A. Digest, 28*(2)*,* 9-30.

Rovinelli, R.J. y Hambleton, R. K. (1977). On the use of content specialist in the assessment of criterion-referenced test ítem validity. *Dutch Journal of Educational Research, 2*, 49-60.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research, 45,* 83-117.

Sireci, S. G. y Geisinger, K. F. (1995). Using subject matter experts to assess content representation: A MDS analysis. *Applied Psychological Measurement, 19*, 241-255.

Sireci, S. G. y Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16,* 17-31.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Warren, W. E., Stevens, R. E. y McConkey, C. W. (1990). Using Demographic and Lifestyle Analysis to Segment Individual Investors. *Financial Analysts Journal, Marzo/Abril*, 74-77.

Wenger, N. S., Shekelle, P. G., and the ACOVE investigators. Assessing care of vulnerable elders: ACOVE Project Overview. *Annals of Internal Medicine, 135 (8)*, 642-646.

(Artículo recibido: 15-02-2011, revisado: 15-1-2012, aceptado: 29-01-2012)