MÉTODOS DE ESCALAMIENTO EN LA OBTENCIÓN DE EVIDENCIAS DE VALIDACIÓN DE CONTENIDO: LA CONSTRUCCIÓN DE UNA ESCALA PARA EVALUAR RESPONSABILIDAD SOCIAL CORPORATIVA

SCALING METHODS IN GETTING CONTENT VALIDATION EVIDENCE: THE CONSTRUCTION OF A SCALE TO ASSESS CORPORATE SOCIAL RESPONSIBILITY

Marta de la Cuesta[1], Juan Diego Paredes Gázquez[1], Francisco Pablo Holgado-Tello[2] y María Isabel Barbero-García[2]

Universidad Nacional de Educación a Distancia (UNED)
[1]Facultad de Económicas y Empresariales UNED
[2]Facultad de Psicología, UNED

Cómo referenciar este artículo/How to reference this article:

De la Cuesta, M., Paredes-Gázquez, J. D., Holgado-Tello, F. P. y Barbero-García, M. I. (2013). Métodos de Escalamiento en la Obtención de Evidencias de Validación de Contenido: La Construcción de una Escala para Evaluar Responsabilidad Social Corporativa [Scaling Methods in Getting Content Validation Evidence: The Construction of a Scale to Assess Corporate Social Responsibility]. *Acción Psicológica, 10*(2), 27-40. http://dx.doi.org/10.5944/ap.10.2.11822

Resumen

El objetivo del estudio es probar la validez de una escala para medir la dimensión económica de la Responsabilidad Social Corporativa (RSC). La escala final, compuesta por 14 ítems, es construida a partir del juicio de expertos, empresas y grupos de interés. La importancia de los ítems de la escala es evaluada a través de la metodología de análisis de procesos jerárquicos (AHP), especialmente apropiada para tratar con constructos complejos como la RSC. Esta metodología también permite aplicar las comparaciones binarias de Thurtone para validad la escala, mostrándose que las comparaciones binarias permiten obtener múltiples evidencias de validez y facilitan el trabajo con constructos complejos.

Palabras Clave: Responsabilidad Social Corporativa (RSC), Análisis de procesos jerárquicos (AHP), Validez, Comparaciones binarias, Juicio.

Abstract

The aim of this study is test the validity of a scale for measuring the economic dimension of Corporate Social Responsibility (CSR). The final scale, composed by 14 items, is constructed attending to judgments of experts, companies

Correspondencia: Marta de la Cuesta. Facultad de Ciencias Económicas y Empresariales. UNED. Email: mcuesta@cee.uned.es.

Recibido: 16/04/2012
Aceptado: 01/06/2013

and stakeholders. Analytic hierarchy process (AHP), a methodology for dealing with complex constructs like CSR, is applied to the scale in order to assess the items. AHP methodology also allows applying Thurtone's pairwise comparisons methodology for the scale validation, showing that pairwise comparisons obtain multiple evidences of validity and eases working with complex constructs.

Keywords: Corporate Social Responsibility (CSR), Analytic hierarchy process (AHP), Validity, Pairwise comparisons, Judgements.

Introducción

En teoría, potencialmente cualquier test puede estar compuesto por un número infinito de ítems, pero desafortunadamente los sujetos evaluados no van a ser capaces de responder a un test de tales características. Por tanto, en el momento de construir el test se ha de determinar un número limitado de elementos que sirvan para evaluar el dominio de interés. Más específicamente la validez de contenido hace referencia al grado en que los elementos de un instrumento de medida constituyen una muestra relevante y representativa del posible universo de ítems que pudiéramos haber utilizado para medir el constructo de interés (APA, AERA, NCME, 1999; Haynes, Richard y Kubany, 1995; Netemeyer, Bearden y Sharma, 2005; Robinson y Stafford, 2006; Sireci, 2006). Es decir, se trata de una manera de operativizar un concepto abstracto (constructo) mediante elementos concretos y tangibles (ítems) (Trochim, 2002).

La representatividad se refiere al grado en que los elementos del test son proporcionales a las distintas dimensiones del constructo que se pretende medir, es decir, al grado en que las distintas facetas del dominio han sido adecuadamente muestreadas. Así mismo, la relevancia está en relación con el grado en que los ítems forman parte del dominio previamente definido. Por tanto, el objetivo básico en la validación de contenido consiste en tratar de garantizar que los ítems reflejen el contenido de las áreas implicadas en el constructo en una proporción adecuada (Netemeyer, Pulling y Bearden, 2002).

Habitualmente, un estudio de validez de contenido implica el desarrollo de un conjunto de ítems para su posterior análisis por parte de un grupo de jueces expertos. Este conjunto inicial de ítems ha de ser comprehensivo e incluir un elevado número de elementos a lo largo de cada una de las dimensiones que componen el constructo de interés. Con ello, se incrementa la probabilidad de que todas las dimensiones estén representadas adecuadamente. Un aspecto de gran importancia es tratar de definir de una manera clara y precisa el constructo a evaluar así como especificar la dimensionalidad del mismo.

Grosso modo, las distintas etapas a seguir en un estudio de validez de contenido son (Crocker y Algina, 1986): (a) definición del dominio del constructo; (b) elaboración de las especificaciones del test; (c) selección de un panel de expertos en el dominio; (d) establecimiento de un marco estructurado para el emparejamiento entre ítems y objetivos, y evaluación del grado de congruencia ítems-objetivos.

A partir de aquí encontramos diversos índices para cuantificar el grado de congruencia en las respuestas de los jueces sobre la asignación de los ítems a cada dimensión y su nivel de representatividad y utilidad. Entre ellos encontramos, por ejemplo, el CRV (*content validity ratio*) propuesto por Lawshe (1975) que se basa en un promedio que toma como referencia el número de jueces que consideran un ítem «no-necesario»; «útil» o «esencial» para medir el dominio al que ha sido asignado. Sin embargo, uno de los más utilizados (Osterlind, 1998), es el índice de congruencia propuesto por Rovinelli y Hambleton (1977) y Hambleton (1980) en el que para evaluar el grado de validez de contenido de cada ítem se le pide al juez que valore en una escala de tres puntos (-1; 0; 1) el grado en que el ítem está relacionado con la dimensión que trata de medir. Cuando un mismo ítem ha sido valorado con 1 por todos los jueces se obtiene un índice de congruencia de 1, que estaría indicando que el ítem ha sido emparejado a la misma dimen-

sión por todos los jueces. Recientemente se ha propuesto el Índice de Osterlind Restringido que amplía la escala de valoración a 5 puntos y cuyo comportamiento se ha mostrado más conservador en la selección de los ítems (Pérez-Gil et al., 2009).

Si nos detenemos más específicamente en la última etapa del proceso propuesto por Crocker y Algina (1986), y asumimos que el resultado final de un estudio de validación de contenido en el que se obtiene una medida de acuerdo o congruencia es equivalente a un intento para escalar los ítems en función del rasgo que se está midiendo (en este caso, congruencia, acuerdo, relevancia, o representatividad, por ejemplo), entonces podríamos entender que los métodos de escalamiento psicológico podrían concretar más específicamente el valor escalar de los ítems de acuerdo con su nivel de «validez de contenido», lo que permitiría obtener un ordenamiento escalar de los ítems. A través de las ordenaciones escalares de los ítems se obtendrían escalas de medida, con lo que la validación de contenido se beneficiaría de las ventjas que conllevaría enmarcar dicho procedimiento en la Teoría de la Medición.

En el párrafo anterior, se han introducido algunos conceptos fundamentales, como son Teoría de la Medición y Escalamiento. Respecto a la primera, para Suppes y Zinnes (1963) la medición consiste en representar un sistema relacional empírico a través de un sistema relacional numérico con una estructura similar. Esta teoría axiomática y formalizada aborda la medición en torno a tres grandes áreas: la representación, la unicidad y la significación. Mediante la representación se pretende encontrar una representación numérica del sistema empírico y guiar la construcción de la escala. De esta manera, el análisis formal del problema de la representación conduce a la formulación de las hipótesis necesarias para justificar una representación numérica dada y al desarrollo de procedimientos de construcción de escalas de medida. El problema de la unicidad hace referencia a la arbitrariedad de los números elegidos, es decir, pueden obtenerse distintas escalas de números para la misma variable o atributo del sistema sin romper la relación establecida en la representación. Por tanto, hay que determinar las transformaciones admisibles para la escala de tal forma que se guarde la correspondencia entre el sistema relacional empírico y numérico. Este es un punto central, ya que supone delimitar el tipo de escala y sus propiedades. Ejemplo de ello, es el sistema propuesto por Stevens (1951) dentro de la Teoría Clásica de la Medición en el que diferencia entre escalas de medida nominales, ordinales, de intervalo y razón y establece las transformaciones admisibles para cada una de ellas, recomendando los estadísticos a utilizar. Así, por ejemplo, las escalas resultantes tras aplicar el escalamiento de estímulos propuesto por Thurstone son de intervalo, lo que implica que los valores escalares obtenidos para los elementos además de igualdad-desigualdad (escalas nominales) y orden (escalas ordinales), también indican aditividad, es decir, las distancias entre los números tienen significado en relación con la propiedad medida. Por último, mediante la significación se plantea el problema de la interpretación de las escalas de medida y el significado de sus números a partir de las condiciones relacionales de las mismas.

Tras aplicar el escalamiento de estímulos propuesto por Thurstone, por ejemplo, en estudios de validación de contenido, obtendríamos ordenamientos escalares de los estímulos que se corresponderían con escalas de medida de intervalo.

Otro de los conceptos referidos anteriormente es el de escalamiento psicológico. En este sentido Torgerson (1958), refiriéndose a la medida de atributos psicológicos considera que el aspecto a ser medido es una propiedad que ocurre en una determinada cantidad que varía de un sujeto a otro, incluso dependiendo de la naturaleza del concepto dentro del mismo sujeto. A esta primera propiedad también se le llama «continuo psicológico». En segundo lugar, otra cuestión es qué propiedades de los números reales posee dicho continuo, con lo que se dirige directamente al problema de la medición referido anteriormente. Torgerson (1958) propone tres aproximaciones a la hora de abordar la problemática de construir un instrumento de medida: Métodos centrados en los sujetos, en los estímulos, y en las respuestas.

En los métodos centrados en los sujetos el constructor del instrumento de medida procura situar a los sujetos a lo largo del continuo. Es decir, la variabilidad encontrada entre los sujetos se debe a las diferencias individuales con respecto al atributo que se está midiendo. Este es el objetivo de la mayoría de las escalas de aptitudes, rendimiento y de algunos procedimientos desarrollados para la medida de actitudes como la técnica de Likert, o el diferencial semántico de Osgood. A pesar de la utilidad práctica de este procedimiento, para Torgerson (1958) no es posible evaluar las propiedades escalares de las puntuaciones derivadas de la asignación de valores numéricos a los sujetos sin recurrir a ningún modelo de escalamiento.

A diferencia de en el caso anterior, en los métodos centrados en los estímulos, la variación observada se debe a las diferencias entre los estímulos en relación con el rasgo latente. Por tanto, el objetivo de este tipo de métodos es localizar al conjunto de estímulos a lo largo del continuo. El origen de los métodos centrados en los estímulos se inició con los estudios de psicofísica con los trabajos de Weber y Fechner. El objetivo era analizar hasta qué punto la sensación humana era capaz de discriminar diferencias mínimas en estímulos físicos, como pesos o sonidos, por ejemplo. Toda esta corriente de investigación originó los modelos de escalamiento psicofísico. Por otro lado, Thurstone demostró que los estímulos psicológicos como las actitudes también podrían ser susceptibles de escalamiento dando origen a los métodos de escalamiento psicológico (ley del juicio categórico; o comparativo, por ejemplo). Es este tipo de escalamiento, el que daría soporte instrumental a la posibilidad de escalar ítems dentro del constructo que se quiere medir para la obtención de evidencias de validación de contenido.

Por último, los métodos basados en las respuestas, quizás, sean los más complejos (Crocker y Algina, 1986). Los datos son usados para escalar a los sujetos en el continuo psicológico a partir de las respuestas correctas; y al mismo tiempo los ítems o estímulos son escalados a partir de la cantidad del rasgo latente presente en el sujeto. Es decir, las variaciones observadas en las respuestas de los sujetos ante los estímulos se debe tanto a las diferencias entre los estímulos con respecto al atributo medido como a las diferencias entre los sujetos (Barbero, 1993). Ejemplos de este tipo de escalamiento lo encontramos en el escalograma de Guttman (1950), el escalamiento de Coombs (1953), o los modelos elaborados desde la Teoría de Respuesta al Ítem.

De las tres aproximaciones propuestas por Torgerson (1958) sólo los métodos centrados en los estímulos y en las respuestas permiten realizar pruebas de bondad de ajuste entre las observaciones empíricas y las pronosticadas por los distintos modelos propuestos. Por lo que también obtendríamos las ventajas que supone poder aplicar pruebas de bondad de ajuste en la obtención de evidencias de validación de contenido.

Como hemos comentado anteriormente, Thurstone (1925, 1927) retoma los trabajos de Fechner para aplicarlos a estímulos no necesariamente de naturaleza física, sino psicológica, iniciando de esta forma el escalamiento de estímulos psicológicos. Su interés se centraba en el escalamiento de estímulos, que no presentasen una dimensión física latente, tales como actitudes o intereses, y que por tanto no podían medirse con instrumentos de medida física (Barbero, 2010).

A partir de los trabajos de Thurstone (1925, 1927) en los que desarrolla la Ley del Juicio Comparativo y la Ley del Juicio Categórico ha sido prolija la investigación en relación con el escalamiento psicológico. De esta forma, podemos destacar a autores como Likert (1932), que desarrolla una técnica para el escalamiento de sujetos en continuos psicológicos unidimensionales, y Guttman (1944, 1947, 1950), o Coombs (1950, 1952) que proponen modelos para el escalamiento tanto de sujetos como de estímulos.

El objetivo de Thurstone era desarrollar un procedimiento a partir del cual se pudiera elaborar una escala en un continuo psicológico y situar en ella a los estímulos. Se trata por tanto de un modelo de escalamiento centrado en los estímulos, donde los sujetos hacen el papel de instrumentos de medida, y la variabilidad en-

contrada en sus respuestas es debida a diferencias en los estímulos respecto al atributo considerado. Esta es exactamente la lógica aplicable a los estudios de validación de contenido, es decir, habría que obtener un ordenamiento escalar de los ítems de acuerdo con su nivel de «validez de contenido» respecto al atributo medido.

Thurstone propuso dos procedimientos indirectos para la obtención de los valores escalares. Uno basado en comparaciones binarias entre los objetos, denominado Ley del Juicio Comparativo; y otro basado en la ordenación y clasificación de los estímulos en una serie de categorías preestablecidas o Ley del Juicio Categórico.

En resumen, en este trabajo se plantea la posibilidad de aplicar métodos de escalamiento como procedimiento para la obtención de evidencias de validación de contenido. Las principales ventajas de ello, vendrían dadas por la obtención de ordenamientos escalares de los ítems de acuerdo con su nivel de «validez de contenido». Atendiendo a la Teoría de la Medición propuesta por Stevens (1951), dichas escalas podrían ser de intervalo con lo que se conocería con exactitud la distancia entre los elementos de acuerdo con su «validez de contenido», por ejemplo.

La posibilidad apuntada en los párrafos anteriores, se aplica de manera empírica sobre la medida de la Responsabilidad Social Corporativa (RSC) de las empresas. En este contexto, uno de los principales problemas ha venido dado por el desarrollo de instrumentos y sistemas de ítems que definan y aborden en todas sus facetas dicho constructo tan complejo. Por ello, a continuación se presenta brevemente unos apuntes sobre ello.

De acuerdo a la Comisión Europea (2011), la RSC es la responsabilidad de las empresas por sus impactos en la sociedad. Esta definición no es la que adoptó la Comisión Europea (2001) por primera vez, que establecía que la RSC es un concepto a través del cual las empresas integran voluntariamente preocupaciones sociales y medioambientales tanto en sus operaciones comerciales como en sus interacciones con sus *grupos de interés* o grupos de interés, definidos éstos como cualquier agente que afecte a la actividad de la compañía o bien pueda verse afectado ella (Freeman, 1984). Este cambio en la definición refleja a la perfección el debate al que ha sido sometida la RSC desde que el término comenzó a aparecer en la literatura académica y empresarial allá por los años 50 del siglo XX (Carroll, 1999), debate que a día de hoy aún trata de especificar el contenido del concepto.

Hasta tal punto llega la cuestión que autores como Okoye (2009) reconocen que no es posible llegar a una definición de consenso sobre la RSC, argumentando que es un *concepto esencialmente debatido*, esto es, un concepto cuyo empleo implica inevitablemente disputas continuas sobre su uso apropiado. No obstante, que exista un incesante debate sobre la RSC no implica que no haya acuerdo sobre ciertas características que se consideran inalienables al concepto. De acuerdo a Henriques (2010), la práctica por parte de las empresas de la RSC suele incluir: (a) Atajar asuntos como son el cambio climático, las condiciones laborales en la cadena de suministro, los Derechos Humanos y, en líneas generales, cualquier aspecto económico, social o ambiental sobre el que la empresa tenga alguna repercusión, tal y como señalan De la Cuesta y Valor (2003). Esto es así porque la empresa es responsable, como agente moral, de los efectos que acarrean sus acciones (Argandoña, 2009); (b) Informar públicamente de tales asuntos para así rendir cuentas a sus grupos de interés. La rendición de cuentas es un concepto inherente al de responsabilidad (Painter-Morland, 2006) que, de acuerdo a la norma AA1000 (AccountAbility, 2008), consiste en el reconocimiento, asunción de responsabilidad y actitud transparente sobre los impactos de las políticas, decisiones, acciones, productos y desempeño asociado a una organización; (c) Dialogar activamente con los grupos de interés sobre tales asuntos en tanto en cuanto se ven afectados por la actividad de la compañía; d) Realizar acciones filantrópicas. Cohen (2010) indica que, en el contexto de RSC, la filantropía puede adoptar muchas formas: donaciones, colaboración con ONGs, patrocinios… y en definitiva, cualquier acción que contribuya al desarrollo social.

La RSC se establece como un concepto a través del cual una empresa es responsable, como agente moral, tanto de las acciones que realiza como de sus consecuencias, ante todos aquellos que se ven afectados por ellas, es decir, ante sus grupos de interés, por lo que estos últimos habrán de mantenerse convenientemente informados al respecto. De este modo, a través de la RSC la empresa contribuye al desarrollo social y ambiental del entorno en el que opera.

Se hace evidente entonces que la información es algo fundamental en la RSC. Existen muchos estándares, nacionales e internacionales, que intenta normalizar esa información con mayor o menor éxito, destacando sobre todo el estándar de la Global Reporting Initiative (GRI, 2011). Sin embargo, dichos estándares, así como la información sobre RSC en líneas generales, han sido fuertemente criticados alegándose sendas carencias (Hammond y Miles, 2004; Hess, 2008; Moneva, 2006). A este respecto, la Asociación Española de Contabilidad y Administración de Empresas (AECA, 2010) apunta a que las principales deficiencias de la información de sobre RSC son: (a) Cantidad de la información. La información disponible sobre las prácticas de RSC llevada a cabo por las empresas es abrumadora (informes de sostenibilidad de las empresas, opiniones de expertos, proveedores de información especializados en RSC, índices bursátiles), y este exceso de información contribuye a minimizar la relevancia de los aspectos esenciales para los grupos de interés (Moneva, 2007); (b) Calidad de la información. Hess (2008) observa que, las empresas ocultan información que les perjudica, publican sólo aquella que les beneficia o simplemente proporcionan información falsa (Hammond y Miles, 2004); (c) Comparabilidad. Si bien el estándar más popular es el de GRI, del que se han puesto de manifiesto severas debilidades (Hess, 2008; Moneva, 2006). El informe de CIMA y PWC (2011) revela problemas inherentes a la falta de comparabilidad de la información (Moneva, 2006); y (d) Fiabilidad de la información. A pesar de que la información de RSC suele someterse a un proceso de verificación externa para asegurar su fiabilidad (Accountability, 2006; Ioannou y Serafeim, 2011; Moneva, 2006), existen evidencias que apuntan a que la verificación de la información sobre RSC presenta severas carencias (Bouten et al., 2011; Hess, 2008; O'Dwyer y Owen, 2005).

Dada la dificultad a la hora de evaluar la información sobre RSC, el objetivo de este trabajo es obtener evidencias de validez de contenido de una serie de ítems de la dimensión económica de la RSC mediante el método de las comparaciones binarias de Thurstone. El estudio se centra en la dimensión económica de la RSC debido a que es la que más disenso suscita (Kakabadse et al., 2005), puesto que tradicionalmente ha sido la única valorada por las empresas.

Método

Para la identificación, evaluación y comparación de ítems se llevó a cabo el siguiente proceso, divido en 4 fases: (a) Revisión bibliográfica para la identificación de los principales ítems de la dimensión económica reflejados en normativas nacionales e internacionales de RSC; (b) Revisión por expertos de los ítems de la dimensión económica de la RSC identificados en la fase anterior; (c) Entrevistas semi-estructuradas a empresas y grupos de interés en las cuales se aplica la metodología de análisis de procesos jerárquico; (d) Aplicación del método de las comparaciones binarias de Thurstone.

Muestra

El muestreo realizado en cada una de las fases, es el siguiente:

a) Revisión bibliográfica de ítems de RSC: La revisión de diferentes fuentes de información sobre ítems de la dimensión económica de la RSC fue llevada a cabo por diversos académicos especializados en el tratamiento de la RSC.

b) Evaluación de ítems por parte de expertos: Los ítems seleccionados por los académicos fueron valorados

por 6 expertos en RSC de reconocida competencia en el mundo académico y empresarial.

c) Comparación de ítems por parte de empresas y grupos de interés en base a la metodología AHP: Con el objetivo de obtener un enfoque común entre los diferentes grupos de interés y solventar la confrontación de intereses planteada por Wood y Jones (1995), la muestra final estuvo compuesta por representantes de 5 instituciones: 3 empresas del IBEX-35 y 2 grupos de interés. Debido al carácter diferenciador de los accionistas como grupos de interés con características claramente diferenciadas y analizadas en el ámbito de la inversión socialmente responsable, los accionistas no forman parte de los grupos de interés analizados en el presente estudio. A los representantes de las 5 instituciones se les realizaron entrevistas semi-estructuradas de en torno a las dos horas de duración. La metodología AHP se aplicó durante la realización de las entrevistas a través del software Expert Choice.

d) Comparaciones binarias de Thurstone: La aplicación del método de las comparaciones binarias de Thurstone se realizó en base a las valoraciones de ítems de la dimensión económica de la RSC realizadas por las empresas y los grupos de interés.

Procedimiento

a) Revisión bibliográfica de ítems de RSC: Académicos especializados en la RSC realizaron una revisión de los ítems de RSC frecuentemente empleados para su evaluación, consultándose las siguientes fuentes: 1) Artículos académicos. Se consultaron artículos académicos de reconocida utilidad a la hora de evaluar la RSC como son los de AECA (2010), Ioannou y Serafeim (2011), Ruf et al. (1998), Waddock y Graves (1997) y Clarkson (1995); 2) Estándares nacionales e internacionales de RSC. Se recurrió al estándar internacional de RSC por excelencia, la Guía para la elaboración de memorias de sostenibilidad de GRI (2011), así como al Pacto Mundial de la Naciones Unidas (2003); 3) Organizaciones especializadas en el tratamiento de la RSC. El Observatorio de Responsabilidad Social Corporativa permitió la consulta y uso de los ítems que emplea para evaluar la RSC de acuerdo a la metodología que emplea para elaborar los análisis de las memorias del IBEX-35 (OBRSC, 2010); 4) Agencias de rating de sostenibilidad. Se dispuso de acceso a la base de datos de la consultora especializada en el tratamiento de la RSC Experts in Responsable Investment Solutions (EIRIS).

b) En total se seleccionaron 22 ítems de la dimensión económica de la RSC.

c) Evaluación de ítems por parte de expertos: Las valoraciones de los ítems se realizaron con puntuaciones que variaban en una escala de 0 a 3, donde 0 implicaba el ítem tenía una importancia nula para el experto, 1 implicaba que el ítem tenía una importancia baja, 2 implicaba que el ítem tenía una importancia media y 3 implicaba que el ítem tenía una importancia alta para el experto.

d) Comparación de ítems por parte de empresas y grupos de interés en base a la metodología AHP: En este estudio se emplea la metodología AHP como un medio a través del cual obtener la información necesaria para realizar las comparaciones binarias. De acuerdo a Saaty y Vargas (2000) AHP es una teoría de medición utilizada para inferir prioridades relativas en escalas absolutas a partir de comparaciones binarias discretas y continuas según una estructura jerárquica de diferentes niveles. Es

una herramienta flexible que permite medir los aspectos tangibles e intangibles de un problema uniendo valoraciones objetivas y subjetivas y valorándolas para obtener una clasificación en función de prioridades.

e) En la metodología AHP se realizan comparaciones binarias entre los diferentes criterios (y subcriterios si procede), cuya importancia ha de ser evaluada de acuerdo a diferentes alternativas que asignan a cada uno de ellos un valor numérico en función de la importancia que le dan los individuos entrevistados. En total, cada partícipe en el estudio realiza *n(n-1)/2* comparaciones binarias, donde *n* es el número de ítems. De este modo, la metodología AHP se basa en la realización de comparaciones binarias, las cuales posteriormente son analizadas conforme a lo establecido por Saaty (1980).

f) Comparaciones binarias de Thurstone: El procedimiento a través del cual las comparaciones binarias se emplean como instrumento de validación figura en Barbero (2007). Una vez los ítems son comparados por pares, se calcula la proporción de veces que un ítem ha sido preferido respecto a otro dividiendo el número de veces que ha sido elegido el ítem entre el tamaño de la muestra, obteniéndose una matriz de proporciones. A cada una de las proporciones que figuran en la matriz se le asigna una puntuación típica, obteniéndose entonces una matriz de puntuaciones típicas. A partir de esta última matriz podemos calcular un valor Z para cada estímulo que servirá para formar una escala de intervalos.

Resultados

a. Revisión bibliográfica para la identificación de los principales ítems reflejados en normativas nacionales e internacionales de RSC

En total, se identificaron un total de 22 ítems de la dimensión económica de la RSC. Estos ítems son frecuentemente empleados a la hora de informar sobre la dimensión económica de la RSC o evaluarla. Es pertinente realizar un filtro que, en base a al criterio de materialidad, permita identificar aquellos ítems que son más importantes.

b. Revisión por expertos de los ítems de RSC identificados en la fase anterior

La revisión de los expertos evidenció que determinados ítems, concretamente aquellos que fueron valorados con un 3, eran considerados más importantes que el resto. En total fueron 14 los ítems de la dimensión económica de la RSC los que obtuvieron una valoración de 3 por parte de los expertos.

c. Entrevistas semi-estructuradas a empresas y grupos de interés mediante la metodología AHP para la identificación de los ítems de RSC considerados materiales de forma conjunta por grupos de interés empresas

Los 14 ítems fueron valorados por los partícipes del estudio. La valoración consistía en comparaciones binarias en las que cada sujeto había de elegir una de las dos alternativas que se le daban. Cada sujeto realiza un total de 91 comparaciones entre ítems.

Al final de proceso se obtiene la valoración que las empresas y los grupos de interés dan a cada uno de los ítems. En la Tabla 1 se muestra la valoración que las empresas y los grupos de interés dieron a los ítems de la dimensión económica conforme a la metodología AHP. Cuanto mayor es la valoración, más consenso hay sobre la importancia del indicador.

Tabla 1.

Valoración de los indicadores económicos aplicando AHP

Indicador	Distancia
Suma total de todo tipo de impuestos pagados, desglosados por países.	0.24
Subsidios recibidos, desglosados por países y regiones.	0.16
Gastos salariales totales desglosados por países o regiones.	0.12
Políticas, prácticas y proporción de gasto correspondiente a proveedores locales en lugares donde se desarrollen operaciones significativas.	0.07
Desglose Geográfico de Mercados.	0.07
Porcentaje de contratos pagados en conformidad con términos acordados.	0.07
Desglose de los proveedores por organizaciones y países.	0.04
Empresas con presencia en paraísos fiscales.	0.04
Procedimientos para la contratación local y proporción de altos directivos procedentes de la comunidad local en lugares donde se desarrollen operaciones significativas.	0.04
Impactos económicos indirectos. Principales externalidades asociadas a los productos y servicios de la organización.	0.04
Rango de las relaciones entre el salario inicial estándar y el salario mínimo local en lugares donde se desarrollen operaciones significativas.	0.04
Desarrollo e impacto de las inversiones en infraestructuras y los servicios prestados principalmente para el beneficio público mediante compromisos comerciales, pro bono, o en especie.	0.03
Consecuencias financieras y otros riesgos y oportunidades para las actividades de la organización debido al cambio climático.	0.03
Donaciones a la comunidad, sociedad civil u otros grupos en metálico y en especie, desglosadas por tipos y grupos.	0.01

La Figura 1 muestra gráficamente la información de la Tabla 1. Se observa que la práctica totalidad de los ítems se hayan muy próximos entre sí, lo que podrían ser indicios de validez convergente, mientras que hay 3 ítems que se encuentran muy aislados y distantes unos de otros, lo que podría indicar que son ítems que no se asocian al concepto de RSC.

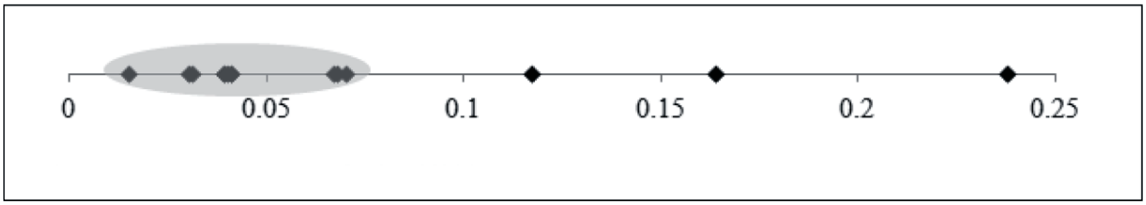

Figura 1. Dispersión de los ítems método AHP

d. Aplicación del método. Comparaciones binarias de Thurtone

Se agregaron las comparaciones binarias realizadas por los 5 partícipes en el estudio, recurriéndose al método de comparaciones binarias en matrices incompletas debido a que se acusó la existencia de datos perdidos. Tras realizar las pertinentes transformaciones y matriz, se obtiene la escala de ítems para la dimensión económica, escala que se muestra en la Tabla 2.

Tabla 2

Valoración de los indicadores económicos aplicando comparaciones binarias con matrices incompletas

Indicadores	Distancia
Suma total de todo tipo de impuestos pagados, desglosados por países.	0.63
Subsidios recibidos, desglosados por países y regiones.	0.63
Desglose Geográfico de Mercados.	0.41
Procedimientos para la contratación local y proporción de altos directivos procedentes de la comunidad local en lugares donde se desarrollen operaciones significativas.	0.40
Desglose de los proveedores por organizaciones y países.	0.36
Donaciones a la comunidad, sociedad civil u otros grupos en metálico y en especie, desglosadas por tipos y grupos.	0.30
Porcentaje de contratos pagados en conformidad con términos acordados.	0.24
Desarrollo e impacto de las inversiones en infraestructuras y los servicios prestados principalmente para el beneficio público mediante compromisos comerciales, pro bono, o en especie.	0.21
Política, prácticas y proporción de gasto correspondiente a proveedores locales en lugares donde se desarrollen operaciones significativas.	0.18
Consecuencias financieras y otros riesgos y oportunidades para las actividades de la organización debido al cambio climático.	0.18
Rango de las relaciones entre el salario inicial estándar y el salario mínimo local en lugares donde se desarrollen operaciones significativas.	0.17
Gastos salariales totales(sueldos, pensiones y otras prestaciones, e indemnizaciones por despido) desglosados por países o regiones.	0.07
Empresas con presencia en paraísos fiscales.	0.02
Impactos económicos indirectos. Principales externalidades asociadas a los productos y servicios de la organización.	0.00

Dado que el método empleado permite observar la distancia existente entre los ítems, en la escala se evidencian claramente 4 grupos de ellos. La Figura 2 evidencia que aplicando las comparaciones binarias de Thurtone existen diferencias entre los ítems, algo que no se observaba al aplicar la metodología AHP, en la cual todos los ítems se concentran en el mismo intervalo.

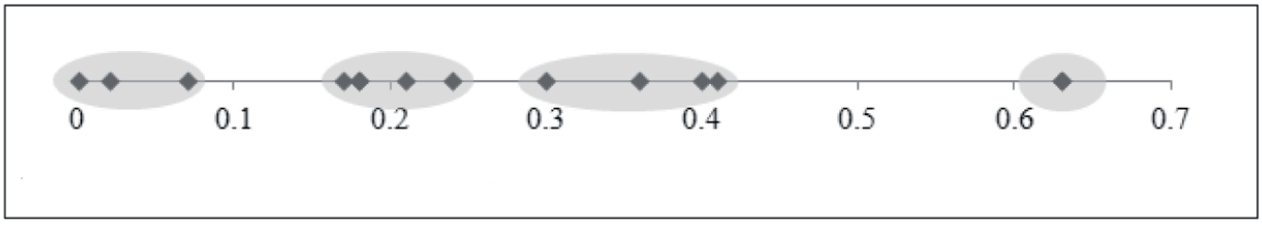

Figura 2. Dispersión de los ítems comparaciones binarias

Que los ítems se concentren en grupos concretos es un indicio de validez convergente que indica que los grupos de ítems están midiendo lo mismo. Así mismo, también hay indicios de validez discriminante, ya que la distancia de los grupos de ítems en la escala es considerable. Los ítems que obtienen una mayor valoración tratan sobre pagos e ingresos relacionados con la administración pública (impuestos y subvenciones), por lo que se pone de manifiesto la especial importancia que se da a la relación con las instituciones públicas, mientras que los ítems que obtienen una valoración menor tratan sobre aspectos sobre los que es difícil obtener información.

Discusión y conclusiones

La aplicación del método de las comparaciones binarias en matrices incompletas ha traído consigo muchas ventajas en cualquier proceso de medición como son:

1. Escalar ítems. Se ha analizado si efectivamente los ítems se escalan tal y como establece la metodología AHP, todos muy próximos entre sí, o lo que es lo mismo, todos igual de importantes. Se ha evidenciado que en realidad existen varias dimensiones distintas, a las cuáles los sujetos entrevistados otorgan diferente importancia.
2. Instrumento de validación de contenidos. La proximidad de ítems en la escala evidencia validez convergente, la valoración que obtienen por parte de los partícipes es la misma, mientras que la distancia entre los ítems es un indicio de validez discriminante.
3. Atajar las opiniones neutras. Tradicionalmente al hablar de RSC, sobre todo las empresas subrayan que todos los aspectos son muy importantes; sin embargo, las comparaciones binarias hacen que el partícipe muestre realmente lo que piensa sobre lo que se le está preguntando.
4. Facilitar las comparaciones. Al ser la RSC un constructo multidimensional y evaluarse a partir de muchos ítems se puede perder la noción de lo que se está considerando. Las comparaciones binarias, al ser un método basado en pares de ítems, hacen que el sujeto siempre tenga claro lo que está comparando
5. Visión global concepto validez. El método de las comparaciones binarias nos brinda la oportunidad de evaluar la concurrencia de múltiples y distintas evidencias de validez, la cual ha de evaluarse recurriendo a un proceso basado en múltiples fuentes que la evidencian y que requiere la combinación tanto de argumentos teóricos como empíricos, algo que contribuye a entender el concepto de validez como un todo y no como la adición de distintos tipos de validez, ya sea ésta de contenido, criterios, constructo, convergente, divergente o de contenido.
6. Aportar información complementaria al Análisis Factorial. El método de las comparaciones binarias se muestra como un complemento interesante al Análisis Factorial a la hora de identificar y validar constructos, siendo a la vez fácil de calcular e interpretar.

Entendemos que las comparaciones binarias se ha mostrado un método un método muy adecuado para investigar la RSC, ya que permite evaluar actitudes subjetivas que implica la RSC (ley del juicio comparativo), facilita la comparación entre ítems y pone en relieve el pensamiento subconsciente del sujeto (es decir, aunque el sujeto diga que todos los ítems o dimensiones tienen la misma importancia, cuando se compara se ve que no).

No obstante no hemos de dejar de lado algunas limitaciones como que el método de las comparaciones binarias exige a los entrevistados realizar un gran número de comparaciones binarias, con el consecuentemente agotamiento mental que eso conlleva y la disponibilidad de tiempo que exige. Por otra parte, un número reducido de comparaciones binarias implica que la escala se compone de un número reducido de ítems, por lo que su capacidad para representar el constructo, en este caso la dimensión económica de la RSC, será menor. Por último, dada las limitaciones temporales y presupuestarias, el número de partícipes en el estudio ha sido reducido, siendo necesaria la realización de un estudio que considere a un mayor número de partícipes para corroborar los resultados expuestos en estas líneas

Referencias

AccoutAbility (2008). *AA1000 Accountability Principles Standard 2008*. AccoutAbility.

AccoutAbility (2006). *The Materiality Report: Aligning Strategy, Performance and Reporting*. London: AccoutAbility.

AECA (2010). *Normalización de la Información sobre Responsabilidad Social Corporativa [Standardization of Information on Corporate Social Responsibility].* Asociación Española de Contabilidad y Administración de Empresas (AECA), Serie Responsabilidad Social Corporativa: Documento n.º 7.

American Psychological Association, American Educational Research Association and National Council on Measurement in Education (1966, 1985, 1999). *Standards for educational and psychological test*. Washington: Author.

Argandoña, A. (2009). ¿Puede la responsabilidad social corporativa ayudar a entender la crisis financiera? *[Can corporate social responsibility help understand the financial crisis].* Documento de investigación DI-790, IESE Business School - Universidad de Navarra.

Barbero, M. I. (1993). *Psicometría II: Métodos de elaboración de escalas [Psychometry II: Scaling methods].* Madrid, España: UNED.

Barbero, M. I. (2007). *Métodos de elaboración de escalas [Scaling methods]*. Madrid, España: UNED.

Barbero, M. I. (2010). Introducción a la Psicometría. En I. Barbero (Coord.), E. Vila y F. P. Holgado, *Psicometría* (pp. 1-44) [Psychometry]. Madrid, España: Sanz y Torres.

Bouten, L., Everaert, P., Van Liedekerke, L., De Moor, L. y Christiaens, J. (2011). Corporate social responsibility reporting: A comprehensive picture? *Accounting Forum*, *35*(3), 187-204.

Carroll, A. B. (1999). Corporate Social Responsibility. Evolution of a Definitional Construct. *Business & Society*, *38*(3), 268-295.

CIMA y PWC (2011). *Tomorrow's corporate reporting - A critical system at risk*. Chartered Institute of Management Accountants and PricewaterhouseCoopers.

Cohen, J. (2010). *Philanthropy*. En W. Visser, D. Matten, M. Pohl, M. y N. Tolhurst (Eds.), *The A to Z of corporate social responsibility* (pp. 315-316). West Sussex, England: Wiley.

Comisión Europea (2011). *Estrategia renovada de la UE para 2011-2014 sobre la responsabilidad social de las empresas [Renewed EU Strategy for 2011-2014 on the social responsibility of business].* Comisión de las Comunidades Europeas, COM (2011) 681 final, Bruselas.

Comisión Europea (2001). *Fomentar un marco europeo para la responsabilidad social de las empresas [Promoting a European framework for corporate social responsibility].* Comisión de las Comunidades Europeas, COM (2001) 366 final, Bruselas.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57, 145-158.*

Coombs, C. H. (1952). *A theory of psychological scaling*. Engineering Research Institute Bulletin, Ann Arbor: University of Michigan Press.

Coombs, C. H. (1953). Theory and methods of social measurement. En L. Festinger y D. Katz (Eds.), *Research methods in the behavioral sciences*. Nueva York: Dryden Press.

Crocker, L. y Algina, J. (1986). *Introduction to classical and modern test theory.* Nueva York: Holt, Rinehart and Winston.

De la Cuesta, M. y Valor, C. (2003). Responsabilidad social de la empresa. Concepto, medición y desarrollo en España [Social responsibility of business. Concept, measurement and development in Spain]. *Boletín ICE Económico*, (2755), 7-19.

Freeman, R. (1984). *Strategic Management: A Stakeholder Perspective*. Boston: Pitman.

Hammond, K. y Milles, S. (2004). Assessing quality assessment of corporate social reporting: UK perspectives. *Accounting Forum*, *28*(1), 61–79.

GRI (2011). *Sustainability Reporting Guidelines* (Version 3.1.). Amsterdam: Global Reporting Initiative (GRI).

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150.

Guttman, L. (1947). On Festinger´s evaluation of scale analysis. *Psychological Bulletin, 44*, 451-465.

Guttman, L. (1950). The basics of scalogram analysis. En S. A Stouffer, L. Guttman y E. Suchman (Eds.), *Measurement in social science.* Princenton: Princenton University Press.

Haynes, S., Richard, D. y Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.

Henriques, A. (2010*). Corporate Impact. Measuring and Managing Your Social Footprint*. London, Whasingston: Earthscan.

Hess, D. (2008). The three pillars of corporate social reporting as new governance regulation: disclosure, dialogue, and development. *Business Ethics Quarterly*, *18*(4), 447-482.

Ioannou, I. y Serafeim, G. (2011). *What Drives Corporate Social Performance? International Evidence from Social, Environmental and Governance Scores*. Working paper 11-016, Harvard Business School, Cambridge.

Kakabadse, N. K., Rozuel, C. y Lee-Davies, L. (2005). Corporate social responsibility and stakeholder approach: a conceptual review. *International Journal of Business Governance and Ethics*, *1*(4), 277-302.

Likert, R. S. (1932). Technique for the measurement of attitudes. *Archives of psychology, 140*, 44-53.

Moneva (2007). El marco sobre información de la responsabilidad social de las organizaciones. *Ekonomiaz*, *65*(2), 284-317.

Moneva, J. M., Archel, P. y Correa, C. (2006). GRI and the camouflaging of corporate unsustainability. *Accounting Forum*, *30*(2), 121–137.

Netemeyer, R., Bearden, W. y Sharma, S. (2005). *Scaling procedures. Issues and applications.* Londres, UK: Sage.

Netemeyer, R. G., Pulling, C. y Bearden, W. (2002). Observations on some key psychometric properties of paper-and-pencil measures. En A. G. Woodside y E. M. Moore (Eds.), *Essays by distinguished marketing scholars of the Society for Marketing Advances* (pp. 115-138). Amsterdam, Países Bajos: JAI.

O'Dwyer, B. y Owen, D. L. (2005). Assurance Statement Practice in Environmental, Social and Sustainability Reporting. *British Accounting Review*, *37*(2), 205–229.

Okoye, A. (2009). Theorising Corporate Social Responsibility as an Essentially Contested Concept: Is a Definition Necessary? *Journal of Business Ethics*, *89*(4), 613-627.

Painter-Morland, M. (2006). Redefining Accountability as Relational Responsiveness. *Journal of Business Ethics*, *66*(1), 89-98.

Pérez-Gil, J. A., Chacón, S., Holgado, F. P., Sanduvete, S., Lozano, J. A., Sánchez, M. y Muñoz, N. (2009, septiembre). Validez de Contenido: Índice de Osterlind restringido. Una propuesta de modificación para el cálculo de adecuación de los ítems de un test. Comunicación presentada

en el *XI Congreso de Metodología de las Ciencias Sociales y de la Salud*. Málaga.

Robinson, S. y Stafford, M. (2006). *Testing and measurement. A user–friendly guide*. Londres, UK: Sage.

Saaty, T. L. y Vargas, L. G. (2001): *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*. Norwell: Kluwer Academic Publishers.

Sireci, S. G. (2006). Content validity. En N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Londres, UK: Sage.

Stevens, S. S. (1951). *Handbook of experimental psychology.* Nueva York: Wiley.

Suppes, P. y Zinnes, J. L. (1963). Basic measurement theory. En R. D. Luce, R. R. Brush y E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 1-76). Nueva York: Wiley

Thurstone, R. (1925). A method of scaling psychological and educational test. *Journal of educational psychology, 16*, 433-451.

Thurstone, R. (1927). A law of comparative judgment. *Psychological Review, 34*, 272-286.

Torgerson, W. S. (1958). *Theory and methods of scaling.* Nueva York: Wiley.

Trochim, A. (2002). *Construct validity*. Recuperado de http://trochim.human.cornell.edu/kb/const-val.htm.

Wood, D. J. (2010). Measuring Corporate Social Performance: A Review. International *Journal of Management Reviews*, *12*(1), 50-84.

Wood, D. J. y Jones, R. E. (1995). Stakeholder Mismatching: A Theorical Problem in Empirical Research on Corporate Social Performance. *The International Journal of Organizational Analysis*, *3*(3), 229-267.