Para citar este artículo, le recomendamos el siguiente formato:

Vázquez Alonso, A., Manassero, M. A. y Acevedo, J. A. (2005). Análisis cuantitativo de ítems complejos de opción múltiple en ciencia, tecnología y sociedad: Escalamiento de ítems. *Revista Electrónica de Investigación Educativa,* 7 (1). Consultado el día de mes de año en: http://redie.uabc.mx/vol7no1/contenido-vazquez.html

Revista Electrónica de Investigación Educativa

Vol. 7, No. 1, 2005

Análisis cuantitativo de ítems complejos de opción múltiple en ciencia, tecnología y sociedad: Escalamiento de ítems

Quantitative Analysis of Complex Multiple-Choice Items in Science Technology and Society: Item Scaling

Ángel Vázquez Alonso (1)
angel.vazquez@uib.es
Departamento de Ciencias de la Educación
Universidad de las Islas Baleares

María Antonia Manassero Mas (1)
ma.manassero@uib.es
Departamento de Psicología
Universidad de las Islas Baleares

José Antonio Acevedo Díaz (2)
ja_acevedo@vodafone.es
Inspección educativa
Consejería de Educación de la Junta de Andalucía

(1) Edificio Guillem Cifre de Colonya,
Carretera de Valldemossa, Km. 7.5
07122, Palma de Mallorca, España

(2) Delegación Provincial de Huelva
Servicio de Inspección de Educación

C/ Los Mozárabes 8, 3ª planta
21071, Huelva, España

(Recibido: 14 de enero de 2005; aceptado para su publicación: 25 de febrero de 2005)

Resumen

La escasa atención prestada a la calificación y evaluación en la investigación de didáctica de las ciencias ha sido especialmente dañina para la educación en ciencia, tecnología y sociedad (CTS), debido a la naturaleza dialéctica, hipotética, cargada de valores y polémica de la mayoría de los temas CTS. Para superar los defectos metodológicos de los instrumentos de evaluación CTS usados en el pasado este artículo propone un instrumento empíricamente desarrollado: el Cuestionario de Opiniones sobre Ciencia, Tecnología y Sociedad (COCTS). Para mejorar la aplicación del cuestionario se han sugerido algunas propuestas metodológicas, como el modelo de respuesta múltiple y el cálculo de un índice actitudinal global. El paso final de estas propuestas metodológicas para la aplicación del COCTS requiere la clasificación previa en categorías de frases CTS. Este estudio describe el proceso de clasificación por medio de un procedimiento de escalamiento basado en un panel de jueces expertos, atendiendo al cuerpo de conocimientos de historia, epistemología y sociología de ciencia. La clasificación de las frases permite un fundamento sólido de los ítems CTS, haciéndolos útiles para la evaluación educativa, la investigación didáctica, así como para aumentar la confianza en sí mismos de los profesores en el desarrollo del currículo CTS en las aulas de ciencias.

Palabras clave: Ciencia, tecnología y sociedad (CTS); evaluación, cuestionario de opiniones, escalamiento de ítems.

Abstract

The scarce attention to assessment and evaluation in science education research has been especially harmful for Science-Technology-Society (STS) education, due to the dialectic, tentative, value-laden, and controversial nature of most STS topics. To overcome the methodological pitfalls of the STS assessment instruments used in the past, an empirically developed instrument (VOSTS, Views on Science-Technology-Society) have been suggested. Some methodological proposals, namely the multiple response models and the computing of a global attitudinal index, were suggested to improve the item implementation. The final step of these methodological proposals requires the categorization of STS statements. This paper describes the process of categorization through a scaling procedure ruled by a panel of experts, acting as judges, according to the body of knowledge from history, epistemology, and sociology of science. The statement categorization allows for the sound foundation of STS items, which is useful in educational assessment and science education research, and may also increase teachers' self-confidence in the development of the STS curriculum for science classrooms.

Key words: Science Technology Society (STS), evaluation, opinion survey, item scaling.

Introducción: la necesidad de evaluar los temas CTS

Muchos estudiosos en didáctica de las ciencias y diseñadores de planes de estudios coinciden en la necesidad de innovar la ciencia escolar mediante temas de ciencia, tecnología y sociedad (CTS), lo que se ha visto reflejado en las reformas de diversos países (Solomon y Aikenhead, 1994). Sin embargo, una seria dificultad para la innovación de la ciencia escolar a través de CTS es la dificultad de evaluar estos temas. La evaluación de contenidos CTS puede convertirse en un obstáculo curricular importante cuando los profesores deciden llevar a cabo la educación CTS, ya que pocos profesores desean incluir estos nuevos temas en sus lecciones de ciencias al no tener una idea clara de cómo evaluarlos (Bell, Lederman y Abd-El-Khalick, 2000).

La evaluación del aprendizaje general ha recibido escasa atención en la investigación de didáctica de las ciencias, con pocas excepciones (Kempa, 1986). Por ejemplo, la sección dedicada a la evaluación en el International Handbook of Science Education (Fraser y Tobin, 1998) es más corta que las otras secciones; además, la primera línea del primer artículo dedicado a la evaluación expresa: "esta sección (…) no se habría escrito hace diez años" (Tamir, 1998, p. 761). En un libro de McComas (2000) sobre la naturaleza de la ciencia sólo uno de los 21 capítulos se dedica a la evaluación. Estos hechos subrayan la escasa atención dada en el pasado y el creciente interés actual por la evaluación en la didáctica de las ciencias. Hace años, algunos especialistas en educación de ciencia tecnología y sociedad (CTS) expresaron su preocupación por la evaluación del aprendizaje de los estudiantes y una mayor coherencia entre la evaluación de los estudiantes y los objetivos educativos CTS (Hofstein, Aikenhead y Riquarts, 1988). Ellos demandaban alternativas a la evaluación tradicional, mediante instrumentos válidos y nuevos criterios de evaluación específicamente diseñados para la evaluación en el marco CTS. La justificación racional de la educación CTS es la multiplicidad de enfoques potenciales que permiten a los estudiantes asimilar el conocimiento científico del complejo y problemático mundo en el que viven. Así mismo, la educación CTS es compleja, diversa y está cargada de valores, los cuales son nuevos aspectos importantes en la educación CTS, en contraste con la ciencia escolar tradicional (Ziman, 1994). A pesar de la importancia creciente de la educación CTS en los currículos reformados de ciencia, la evaluación de los temas CTS aún está poco desarrollada y precisa de mejoras (Acevedo, 1997). La carga de valores CTS se agrega a las dificultades de enseñar ciencias y, sobre todo, de evaluar los temas CTS, donde el profesorado no sólo debe evaluar conocimientos científicos y procedimientos, como en las clases tradicionales, sino también valores de la ciencia. Los valores tienen un fuerte significado afectivo, arraigado en la capacidad humana para seleccionar entre diferentes alternativas y se integran de modo complejo en los contenidos y procedimientos CTS. Más allá de *conceptos*, *opiniones* o *creencias*, se necesitan estructuras más potentes para responder de modo fiable a los valores en la educación CTS. En los próximos

párrafos se sugiere el concepto de actitud como la estructura que permite esta evaluación integrada de los temas CTS.

Las actitudes en la educación científica

El concepto de actitud en psicología social surge de los problemas sociales (la actitud hacia la pena de muerte) y los políticos (votar candidatos), pero se ha extendido a otros muchos campos como el de la educación. Los profesores, en general, y los de ciencias, en particular, suelen tener un concepto de actitud de sentido común, relacionado con el interés (o desinterés) hacia el aprendizaje de la ciencia escolar; por ejemplo, cuando un estudiante participa poco en los debates en el aula de ciencia, los profesores estiman que tiene una actitud pobre hacia la ciencia. Sin embargo, el concepto de actitud es mucho más amplio, pues abarca cognición, conducta y sentimientos, siendo esta complejidad la principal razón por la que se usa en este trabajo, en vez de otros conceptos algo más simples, como *creencias* u *opiniones*, que también son bastante comunes en la bibliografía relativa a la educación científica. Por otro lado, la actitud es una estructura muy establecida en psicología social, que está mejor fundamentada que las creencias o las opiniones (Stahlberg y Frey, 1990).

La actitud es un constructo hipotético de los psicólogos sociales que puede definirse como:

> Una tendencia psicológica que se expresa evaluando una entidad particular con algún grado de agrado o desagrado (…) Una tendencia psicológica se refiere a un estado interior de la persona y *evaluar* se refiere a todos los tipos de respuestas de la evaluación, explícitas o implícitas, cognoscitivas, afectivas o conductuales [Traducción libre de los autores] (Eagly y Chaiken, 1993, p.1).

La expresión “entidad”, también conocida como objeto de la actitud, contiene los estímulos (cosas, ideas o personas) que suscitan las respuestas de la evaluación que expresan la actitud.

El uso del constructo *actitud* puede parecer un poco extraño aquí, pero se justifica porque muchos problemas CTS están cargados de valores. Requiere del estudiante (y del profesorado) no sólo conocimiento de hechos, sino también la adhesión a una posición o a una acción de acuerdo con esta posición. La actitud es la estructura que integra simultáneamente los componentes cognoscitivos, afectivos y conductuales, los cuales responden a los contenidos cargados de valores CTS en la educación científica. La cognición y comprensión siempre están presentes, pero el núcleo central de la actitud es la opción de la persona a lo largo del espectro de diversas posiciones actitudinales, entre el agrado y desagrado, que es característica de la actitud. El término *actitud* da mejor cuenta de los tipos de preguntas de la educación CTS, porque abarca al mismo tiempo cognición (conocimiento), sentimientos (acuerdo o desacuerdo) y conducta (por ejemplo, actuar para hacer el ambiente más saludable). Mientras el conocimiento de la

ciencia no siempre implica una actitud (por ejemplo, comprender la ley de gravedad y sus aplicaciones no genera diferentes actitudes, pues nadie discute su validez), en la base de cualquier actitud relacionada con la ciencia hay siempre algo de conocimiento (por ejemplo, adherirse a la posición de que el conocimiento de la ciencia es provisional), pues hay personas que aceptan esta posición (y otras que no), aunque su conocimiento del tema sea superficial.

Las actitudes en didáctica de las ciencias han tenido un camino largo y difícil que va desde su clarificación conceptual a su evaluación empírica. Gardner (1975) sugirió una distinción entre dos objetos actitudinales diferentes: las actitudes frente la ciencia y las actitudes científicas. En años recientes muchos autores han asumido esta distinción (Laforgia, 1988; Schibeci, 1983; Wareing, 1990). Por su parte, Hodson (1985) sugirió nuevos objetos actitudinales cuyos significados son especialmente importantes para la educación CTS: los aspectos sociales de la ciencia y de la ciencia escolar. La educación CTS es un enfoque innovador transversal en el currículo de ciencia escolar, que se centra en los valores de la ciencia (Aikenhead, 1994; Bybee, 1987; Vázquez y Manassero, 1997; Waks y Prakash, 1985), de modo que el concepto de actitud es la estructura que mejor abarca los objetivos de enseñanza y aprendizaje CTS, pues integra cognición, afecto y conducta. La gran cantidad de investigación actitudinal realizada en la didáctica de las ciencias ha ido añadiendo una variedad de objetos en este campo, cada uno de los cuales define una actitud diferente, lo que hace necesaria su sistematización. Para integrarlos ordenadamente, Vázquez y Manassero (1995) han sugerido una taxonomía de actitudes relacionadas con la ciencia que clasifica diferentes objetos actitudinales CTS (todos ellos relativos a la ciencia y tecnología) en tres dimensiones básicas y siete subdimensiones:

- Las actitudes hacia la ciencia y la tecnología escolar, su enseñanza y aprendizaje, los resultados de ciencia y tecnología escolar, etcétera.
- Las actitudes hacia las interacciones entre la ciencia, la tecnología y la sociedad: la imagen social de la ciencia y la tecnología; los aspectos sociales de la ciencia y la tecnología.
- Las actitudes hacia las características del conocimiento científico y tecnológico: las características de los científicos y tecnólogos; el constructivismo social en la ciencia y la tecnología; la naturaleza de la ciencia y la tecnología.

El concepto de actitud se ha usado en didáctica de las ciencias desde los inicios (Gardner, 1975; Gauld y Hukins, 1980; Haladyna y Shaughnessy, 1982; Munby 1983; Ormerod y Duckworth, 1975; Schibeci, 1983, 1984), pero la bibliografía actual emplea diversas denominaciones para referirse indistintamente al mismo constructo, tales como conceptos, creencias, opiniones e ideas, quizás debido a la influencia del enfoque constructivista del aprendizaje o simplemente por sentido común. Proponer el término “actitud” para representar los complejos problemas (cognoscitivos, conductuales y afectivos) integrados en la dialéctica de los temas CTS cargados de valores no es sólo una disputa nominal, sino un cambio teórico que puede ampliar el simple significado de ideas, creencias u opiniones en la

didáctica de las ciencias. Una discusión de las relaciones entre actitud y las demás denominaciones usadas en la bibliografía (creencias, opiniones, etc.), así como las razones por las que la actitud es un constructo que se ajusta mejor a las características de los problemas CTS –a la vez cognoscitivos, afectivos y conductuales– puede revisarse en otros lugares (Eagly y Chaiken, 1993, Manassero, Vázquez y Acevedo, 2004). Por último, las actitudes permiten resaltar los principales objetivos curriculares de los temas CTS que implican valores y el dominio afectivo de la educación, aunque hay también otros objetivos educativos CTS importantes, por ejemplo, en los dominios cognoscitivo o procedimental. Se subrayan así las actitudes como una importante estructura para lograr los objetivos del dominio afectivo (valores).

Medida de actitudes relacionadas con la ciencia

El diseño de los nuevos currículos de ciencias ha incluido los temas CTS como una consecuencia directa de la gran cantidad de investigación que ha mostrado que los estudiantes y los profesores no poseen concepciones adecuadas sobre estos temas CTS (véase, por ejemplo, Lederman [1992]).

La investigación de las actitudes relacionadas con la ciencia ha repetido algunos de los mismos fallos de la investigación de las actitudes generales realizada en psicología social, porque que se han subestimado e ignorado (Shrigley y Koballa, 1992). La conceptualización y medida de actitudes dentro de este campo han desarrollado dos tradiciones básicas: el *escalamiento psicofísico* y la *evaluación psicométrica*. El primero se basa en la graduación de los estímulos aplicados a las personas y en observar sus reacciones en una dimensión psicológica. El origen de la evaluación psicométrica tiene su base en los métodos de los tests mentales y psicológicos. Las técnicas de Likert y el diferencial semántico de Osgood corresponden a la tradición psicométrica, cuya validez está basada en la capacidad de cada ítem para representar adecuadamente los objetos actitudinales subyacentes a las escalas (casi siempre dados por supuesto).

Las controversias alrededor de la validez de los instrumentos y los procesos de evaluación de actitudes han sido y aún son frecuentes. Diversas revisiones (Gardner, 1975; Gauld y Hukins, 1980; Schibeci, 1984; Shrigley y Koballa, 1992) coinciden en las importantes limitaciones metodológicas de los instrumentos de evaluación de actitudes, y han criticado ampliamente los resultados obtenidos con ellos (Gardner, 1996). Las principales críticas pueden resumirse del siguiente modo:

- La inexacta definición del objeto de actitud en los instrumentos (validez de constructo) y la falta de correspondencia entre lo que se quiere medir y lo que realmente se mide (Gauld y Hukins, 1980).
- La falta de una legítima fundamentación epistemológica explícita del contenido del instrumento. Dada la naturaleza compleja y dialéctica de los contenidos de las actitudes relacionadas con la ciencia, la ausencia de especificaciones para

la base filosófica de las escalas debilita los resultados obtenidos con estos instrumentos y sus interpretaciones (Aikenhead, 1988; Gardner, 1975, 1996; Haladyna y Shaughnessy, 1982; Ormerod y Duckworth, 1975; Schibeci, 1984; Shrigley y Koballa, 1992).

- Los instrumentos violan implícitamente la hipótesis unidimensional, que es una condición para lograr medidas válidas. Los cuestionarios no tienen a menudo una única estructura común para todos los ítems e, incluso, a veces muestran una multidimensionalidad explícita (Bratt, 1984; Munby, 1983; Zeidler, 1984). En estos casos la actitud medida no es única, sino múltiple.

Otros problemas relacionados con la validez son los sesgos de los alumnos para satisfacer las expectativas de sus profesores y las dificultades de la formulación de las preguntas para evitar la denominada "percepción inmaculada" (la hipótesis implícita de que sujetos e investigador perciben el mismo significado de las frases en los cuestionarios). Por último, la confrontación entre el paradigma cuantitativo y el cualitativo, basada principalmente en las entrevistas clínicas y el análisis de casos, es otro problema polémico para la investigación de las actitudes. Las demandas de "dekuhnificar" esta discusión, romper los estereotipos y abrir el procedimiento, están dando lugar a enfoques para la integración de ambos métodos en lugar de excluir alguno de ellos (Shadish, 1995).

Los instrumentos de evaluación actitudinal CTS

Los contenidos CTS (la naturaleza de la ciencia y la tecnología, las relaciones entre la ciencia, la tecnología y la sociedad, etc.) se han ido elaborado por medio de los instrumentos actitudinales aplicados en la investigación. Una revisión de los instrumentos de evaluación de las actitudes CTS se ha mostrado en otros trabajos (Lederman, Wade y Bell, 1998; Vázquez y Manassero, 1995).

Como se ha indicado antes, la fiabilidad y la validez de los instrumentos de evaluación actitudinales, así como la interpretación parcial de los resultados, con frecuencia ha sido la principal fuente de problemas metodológicos en la evaluación CTS (Gardner, 1975, 1996). Aikenhead (1988) comparó la fiabilidad de diferentes técnicas e instrumentos –escalas Likert, preguntas abiertas, cuestionarios de opción múltiple empíricamente desarrollados y entrevistas semiestructuradas– y concluyó que los cuestionarios de opción múltiple desarrollados empíricamente están más equilibrados, pues reducen la ambigüedad, pueden aplicarse a muestras grandes que aumentan la representación de los resultados y pueden evitar las trampas metodológicas, tales como la doctrina de la percepción inmaculada de los cuestionarios desarrollados exclusivamente por los investigadores. Posteriormente, Lederman, Wade y Bell (1998) han revisado también los instrumentos que evalúan las concepciones de la naturaleza de ciencia y han cuestionado su validez debido a sus fallos de construcción y a las interpretaciones parciales de los resultados. Aunque estos autores concluyeron que la evaluación de los conceptos que tienen las personas sobre la naturaleza de la ciencia debe cambiarse por aproximaciones más cualitativas y abiertas, también

consideran que el cuestionario de papel y lápiz de opción múltiple Views on Science, Technology and Society (VOSTS) es un instrumento valioso para la evaluación de las opiniones de los estudiantes y un esfuerzo importante por ahondar en las razones que tienen los estudiantes para dar sus respuestas.

El VOSTS es un conjunto de 114 ítems empíricamente desarrolladas que abarcan la mayor parte de los contenidos CTS (Aikenhead, Fleming y Ryan, 1987; Aikenhead y Ryan, 1989, 1992). Asumiendo las mismas normas del VOSTS, Rubba y Harkness (1993) desarrollaron un conjunto de ítems de opción múltiple denominado Teacher's Belief About Science-Technology-Society (TBA-STS) para investigar las creencias de los profesores sobre los temas CTS. El COCTS tiene 100 ítems, seleccionados del VOSTS y del TBA-STS, que han sido adaptados a la cultura española. El COCTS se ha aplicado en una investigación sobre las actitudes CTS de una amplia muestra de estudiantes y profesores (Manassero y Vázquez, 1998).

Un ítem típico del COCTS muestra un texto inicial que plantea un tema CTS seguido por varias frases cortas, etiquetadas con una letra en orden alfabético; cada frase desarrolla diferentes razones para responder al ítem. Dentro de cada ítem, el conjunto de frases ofrece a la persona que responde una serie amplia de distintas posiciones sobre el tema; las actitudes se definen mediante la elección hecha o por medio de la valoración de cada frase. Esta manera de definir actitudes indica que el cuestionario no impone a la persona que responde ningún valor o modelo concreto sobre CTS; al contrario, las personas que responden pueden definir libremente las actitudes mediante sus elecciones de las frases proporcionadas. La Tabla I muestra un ejemplo de ítem, junto con las puntuaciones de los jueces a las frases correspondientes y la categoría asignada a cada una de ellas (según el criterio que se desarrollará en este artículo). La versión en castellano del COCTS, el manual de uso y las claves de puntuación se han incorporado a la extensa biblioteca del Test Collection of Educational Testing Service (accesible en línea en: www.ets.org).

Tabla I. Resultados de la clasificación del ítem 10211 relativo a la definición de tecnología

Puntuaciones directas de los jueces											**Votos de jueces por categoría**			**Punt. media**	**Cate-goría**	**Texto de la pregunta**
J 1	**J 2**	**J 3**	**J 4**	**J 5**	**J 6**	**J 7**	**J 8**	**J 9**	**J 10**	**J 11**	**Ingen.**	**Plaus.**	**Adec.**			
																10211. Definir *qué es la tecnología* puede resultar difícil porque ésta sirve para muchas cosas. Pero la tecnología PRINCIPALMENTE es:
1	1	6	5	2	4	3	9	1	5	4	5	5	1	3.73	Plaus.	A. Muy parecida a la ciencia.
2	2	2	3	2	3	2	2	1	5	3	10	1	0	2.45	Ingen.	B. La aplicación de la ciencia.
4	6	6	2	7	5	5	4	6	5	5	1	9	1	5.00	Plaus.	C. Nuevos procesos, instrumentos, maquinaria, herramientas, aplicaciones, artilugios, ordenadores o aparatos prácticos para el uso de cada día.
4	5	5	2	7	5	4	5	1	5	2	3	7	1	4.09	Plaus.	D. Robots, electrónica, ordenadores, sistemas de comunicación, automatismos, máquinas.

8	6	5	6	6	5	5	1	5	6	9	1	8	2	5.64	Plaus.	E. Una técnica para construir cosas o una forma de resolver problemas prácticos.
6	7	5	6	6	5	5	6	5	8	6	0	9	2	5.91	Plaus.	F. Inventar, diseñar y probar cosas (por ejemplo, corazones artificiales, ordenadores y vehículos espaciales).
9	8	5	6	8	7	6	3	2	5	7	2	4	5	6.00	Adec.	G. Ideas y técnicas para diseñar y hacer cosas; para organizar a los trabajadores, la gente de negocios y los consumidores; y para el progreso de la sociedad.
6	5	4	4	4	6	5	1	1	6	8	2	8	1	4.55	Plaus.	H.* Saber cómo hacer cosas (por ejemplo, instrumentos, maquinaria, aparatos).

* Texto añadido de Rubba y Harkness (1993) al texto original del VOSTS.

El formato básico de respuesta propuesto para el VOSTS consiste en seleccionar una opción entre las diversas frases de cada ítem (Modelo de respuesta única [MRU]). El MRU es metodológicamente muy limitado, pues sólo permite comparaciones centradas en cada ítem particular, pero no pruebas de comparaciones test-retest o de comprobación de hipótesis. Rubba, Schoneweg-Bradford y Harkness (1996) propusieron puntuar la única respuesta de cada pregunta según un esquema de tres categorías (Realista / Meritoria / Ingenua) previamente asignado a las respuestas dadas por expertos, las cuales reflejan claramente el carácter actitudinal de la evaluación. Este esquema de puntuación en tres categorías fue sugerido por uno de los autores del VOSTS en una comunicación personal (Rubba y Harkness, 1993). En el caso de responder varias preguntas diferentes, las puntuaciones de las preguntas individuales se suman para conseguir una puntuación total que es internamente consistente y actitudinalmente significativa (actitud más o menos adecuada). Este método mejora el MRU simple, pero no diferencia de manera fiable ni adecuada las actitudes que mide. Además, el MRU tiene una segunda limitación importante: la actitud medida se basa en una sola elección no usa toda la información disponible en las restantes frases no seleccionadas.

Para superar estos serios inconvenientes, se ha sugerido un Modelo de Respuesta Múltiple (MRM), donde las personas que responden valoran todas las frases de cada ítem en una escala de 9 puntos para expresar su grado de acuerdo o desacuerdo. A continuación, las valoraciones de las frases se transforman en un índice actitudinal global (rango: -1, +1) mediante un método interpretativo que requiere una clasificación previa de cada frase en un escalamiento de tres categorías (Vázquez y Manassero, 1999):

- Adecuada (A): La frase expresa un punto de vista apropiado.
- Plausible (P): Aunque no es totalmente adecuada, la frase expresa algunos aspectos aceptables.
- Ingenua (I): La frase expresa un punto de vista que no es ni adecuado ni plausible.

El esquema de tres categorías para evaluar las respuestas de la pregunta no es un procedimiento del tipo *verdadero o falso* para buscar respuestas correctas; no

tiene nada que ver con un esquema absoluto porque es relativo, de acuerdo con el conocimiento dialéctico de la historia, epistemología y sociología de la ciencia. Se espera que la efectividad de la clasificación se mejore cuando el progreso de las controversias actuales pueda arrojar nueva luz sobre los temas CTS. Por otro lado, la clasificación tampoco tiene nada que ver con los tajantes procedimientos de calificación del tipo *correcto o incorrecto*, habituales en las pruebas y exámenes de ciencias, porque evalúa la actitud de los estudiantes a partir del conjunto de sus respuestas a todas las frases de cada ítem como *adecuadas*, *plausibles* e *ingenuas* y no solamente a partir de una de ellos.

El VOSTS se ha usado con el MRU para evaluar las actitudes de las personas encuestadas con fines descriptivos y formativos. Los propósitos sumativos están excluidos de las posibilidades del VOSTS, porque el MRU es incapaz de lograr cualquier puntuación sumativa, como se requiere tanto en investigación como en educación. El objetivo de este trabajo no es sustituir cualquier finalidad descriptiva o formativa de la evaluación con el COCTS, sino ampliar las posibilidades del instrumento, detallando más la descripción y mejorando ampliamente los objetivos formativos, aunque se mantengan todas las ventajas metodológicas anteriores.

Escalamiento de las frases mediante jueces

Un denominador común de la gran cantidad de investigación sobre las actitudes hacia la ciencia es la existencia de algunas creencias adecuadas, así como otras menos adecuadas. Sin embargo, el propósito de este estudio no es evaluar las puntuaciones de los estudiantes o los profesores, aunque una de las principales utilidades del método que se propone es la investigación que involucra valoraciones de creencias CTS. Para aplicar apropiadamente a la medida la escala anterior de tres categorías es necesaria la clasificación previa de todo el cuestionario; es decir, un escalamiento de los estímulos actitudinales representados por las diversas frases de cada ítem. La clasificación de las frases es un paso imprescindible para la evaluación válida y fiable de los temas CTS.

El objetivo de este artículo es presentar el método seguido y los resultados obtenidos para la clasificación de las frases del COCTS en una de las tres categorías a partir del escalamiento mediante jueces. Así, el centro del análisis está en las respuestas del grupo de jueces en lugar del análisis de cada frase, el cual necesitaría muchas páginas y tablas debido al elevado número de frases que se analizan.

Aunque el procedimiento general de escalamiento es similar al escalamiento de Thurstone o el de Guttman (Eagly y Chaiken, 1993), algunas áreas CTS aún son muy polémicas para que el acuerdo sobre la adecuación de una frase sea fácil de conseguir, incluso para los propios expertos (Alters, 1997a, 1997b; Smith, Lederman, Bell, McComas y Clough, 1997). Por ejemplo, McComas, Almazroa y Clough (1998) han sugerido algunos puntos de acuerdo general sobre la naturaleza de ciencia (empírica, racional, provisional, cargada de teoría), aunque

el núcleo real de las dificultades para definir concepciones adecuadas surge de los aspectos inciertos y dialécticos involucrados en muchos temas CTS. La provisionalidad se vuelve un rasgo central de los resultados alcanzados mediante la técnica de escalamiento, debido a la naturaleza dialéctica y polémica de los temas CTS. Rubba, Schoneweg-Bradford y Harkness (1996) clasificaron los 16 ítems del TBA-STS mediante cinco jueces e informaron que se encontraron dos de ellos *outliers*, lo que muestra la dificultad para evaluar los temas CTS. Recomendaron trabajar al menos con nueve jueces y usar como criterio para decidir la categoría de cada frase el acuerdo de al menos siete de los nueve jueces.

Por otro lado, algunos estudios han reportado que una persona puede manifestar creencias diferentes u opuestas (diversidad intrapersonal) sobre el mismo problema cuando se inspeccionan por medio de ítems que difieren en formato o contexto (Acevedo, 2000; Clough y Driver, 1986; Oliva, 1999; Taber, 2000). En la investigación general de actitudes, la presencia simultánea de actitudes opuestas en la misma persona es un hecho muy conocido que normalmente se atribuye al papel latente de éstas, sobre todo, cuando se refieren a temas no enseñados explícitamente, como es el caso de los temas CTS, al menos en el currículo de ciencias español actual. Estas actitudes opuestas se han considerado respuestas indiferentes, intermedias, incoherentes, ambivalentes o ambiguas. A pesar de su apariencia escasamente definida, con frecuencia se han atribuido a las debilidades metodológicas de los instrumentos de la medida; por ello, la atención se ha dirigido a mejorar la exactitud de éstos (Breckler, 1994). Sin embargo, algunos estudiosos reconocen esa inconsistencia en las actitudes como un problema complejo que surge de la falta de consistencia entre la conducta, la cognición o los afectos, aunque la pregunta relativa a las respuestas ambivalentes se refiere principalmente a la inconsistencia entre la actitud y la cognición (Eagly y Chaiken, 1993). Algunas investigaciones sostienen que las personas muy consistentes muestran actitudes más estables, más predictivas de la conducta y más resistente a la influencia por inducción, mientras que a las personas de baja consistencia les faltaría una actitud genuina (Chaiken, Pomerantz y Giner-Sorolla, 1995). Además, para completar el difícil cuadro de la investigación de las actitudes, las creencias complejas pueden asociarse con actitudes moderadas, pero también con actitudes extremas (Eagly y Chaiken, 1993). El MRM propuesto para el COCTS, junto con la clasificación obtenida en este estudio, permite la medida de estas actitudes contradictorias sobre un tema de modo natural, las cuales pueden coexistir en una misma persona, por medio del análisis individualizado de las respuestas a cada pregunta. Las propuestas para refinar la medida de la intensidad de la actitud pueden ayudar a tratar la ambivalencia o la inconsistencia en las actitudes relacionadas con la ciencia (véanse las contribuciones en Petty y Krosnick, 1995) y, desde esta perspectiva, el índice de actitud global puede representar un gran avance práctico para la medida de la actitud relativa a la ciencia.

En el marco de la evaluación de actitudes relacionadas con la ciencia, este estudio muestra el proceso de escalamiento de las diversas frases de los ítems del COCTS, mediante un panel de expertos que actúan como jueces. Este escalamiento es la

fase final de los procedimientos empíricos diseñados y expuestos en otro lugar (Vázquez y Manassero, 1999) para el desarrollo de un método cuantitativo que mida ítems complejos de opción múltiple, que permite el uso válido y fiable de las preguntas del COCTS en la evaluación de actitudes CTS, incluyendo también la verificación de hipótesis y las comparaciones entre grupos.

Metodología de escalamiento

Una muestra válida de 16 jueces expertos españoles puntuó las frases de todas los ítems del COCTS. Todos ellos están acreditados como potenciales expertos competentes para emitir juicios válidos y fiables sobre las frases del COCTS, debido a su entrenamiento específico en temas CTS, independientemente de sus trabajos realizados o su formación inicial. Una mayoría de ellos (13) posee formación inicial científica (licenciados en física, química, biología o geología) y 3 de ellos son licenciados en filosofía. Los trabajos actuales son como profesores de educación secundaria (5), como asesores o formadores de profesores de ciencia (4) y como profesores universitarios e investigadores (7). La mayoría (12) participa activamente en investigación en didáctica de las ciencias y algunos de ellos (8) publican investigaciones concretas sobre temas CTS. Rubba, Schoneweg-Bradford y Harkness (1996) usaron sólo 5 jueces (profesores y científicos) y recomendaron un mínimo de 8 ó 9 jueces para lograr una clasificación más válida y fiable.

A los jueces se les proporcionaron los ítems del COCTS y se les pidió valorar cada frase en una escala de 9 puntos (1-9), donde debían expresar su desacuerdo o acuerdo respecto a los conocimientos actuales de historia, filosofía y sociología de la ciencia. Para asignar un significado más claro a los puntos de la escala, con el propósito de aumentar la exactitud y la fiabilidad de los juicios y obtener un voto simple para las valoraciones directas de los jueces, la escala de 9 puntos fue dividida en tres intervalos iguales:

- Puntuaciones de 1 a 3: categoría ingenua (I) (frases inadecuadas).
- Puntuaciones de 4 a 6: categoría plausible (P) (frases parcialmente adecuadas).
- Puntuaciones de 7 a 9: categoría adecuada (A) (frases apropiadas).

La escala de 9 puntos se eligió debido al uso de números enteros simples para evaluar las tres categorías, dándoles la oportunidad a los jueces de ampliar sus juicios, en lugar de restringirlos a una escala corta de 3 puntos que habría sido más simple para los propósitos cuantitativos, pero habría reducido mucho y limitado la expresión de opiniones ligeramente diferentes de los jueces. Estas asignaciones definen, al mismo tiempo una escala de tres categorías, que corresponden a rangos naturales con un significado fijo –adecuado, plausible e ingenuo–, y la escala entera original de 9 puntos, que permite a los jueces afinar

con más precisión sus valoraciones dentro de cada categoría y obtener una percepción clara del significado de cada puntuación.

El COCTS consta de 100 ítems con 637 frases para seleccionar como respuesta, cada una de ellas expresa actitudes, creencias y opiniones diferentes sobre el tema planteado en cada ítem. Cada uno de los 16 jueces se considera una variable pertinente del análisis, mientras las 637 frases que constituyen el COCTS serían los casos de la descripción. Este estudio pretende mostrar los procedimientos y análisis, empíricos y racionales, desarrollados para asignar cada frase a una de las tres categorías sugeridas (ingenua, plausible y adecuada), a partir de las puntuaciones directas de los jueces. El análisis se centra en los jueces, no en los ítems, de modo que el criterio para la asignación de categorías a las frases debe deducirse de las puntuaciones concedidas por los jueces. El primer criterio de clasificación será la puntuación media que éstos otorguen a cada frase, puesto que este parámetro sintetiza todos los matices de sus valoraciones. El segundo criterio será que la mayoría de jueces esté a favor de una categoría específica de acuerdo con las puntuaciones. La clasificación de las frases del COCTS en una de las tres categorías normalizadas se basará en estos dos criterios básicos, temperados por los modelos globales y los sesgos de las puntuaciones de los jueces.

Los jueces valoraron las 637 frases' pero cinco de ellos rechazaron puntuar cuatro ítems que plantean los siguientes temas: influencia de la cultura y la religión en el conocimiento científico, elegancia de las teorías científicas, naturaleza probabilística del conocimiento científico y el papel de un ser sobrenatural en la ciencia. Para estos ítems, las respuestas de los objetores se perdieron, lo que da lugar a un número de votos inferior.

Resultados del escalamiento

Las estadísticas descriptivas de las puntuaciones de los jueces (medias, variaciones, rangos, etc.) muestran la existencia de tendencias concretas en sus respuestas, como puntuaciones demasiado altas o demasiado bajas y el uso incompleto de las puntuaciones enteras de la escala de 9 puntos, etcétera (Tabla II). La diferencia entre las puntuaciones medias más alta (J4) y más baja (J12) de los jueces es de aproximadamente dos puntos (25% del rango de la escala). También las desviaciones típicas de las puntuaciones de los jueces varían mucho, con un mínimo de 1.38 puntos (J9) y un máximo de 2.79 (J13). Por otro lado, algunos jueces no usan los 9 puntos de la escala para evaluar las 637 frases de los ítems del COCTS. Dos de ellos (J4 y J9) no valoran ninguna frase con 1 punto y, además, J9 nunca usa la puntuación 2. Otro juez (J11) nunca usa la puntuación 9, siendo su puntuación máxima 8. En resumen, estos rasgos estadísticos de la distribución global de las puntuaciones de los jueces muestran algunos sesgos individuales en los juicios expresados mediante sus valoraciones.

Tabla II. Estadística descriptiva de las puntuaciones directas emitidas por los 16 jueces

Jueces	**J1**	**J2**	**J3**	**J4**	**J5**	**J6**	**J7**	**J8**	**J9**	**J10**	**J11**	**J12**	**J13**	**J14**	**J15**	**J16**
Media	4.23	4.79	4.26	5.54	4.11	4.84	4.23	4.89	5.34	4.67	4.19	3.81	3.91	4.87	4.64	5.09
Desv. estándar	2.74	2.46	2.22	1.83	2.17	2.15	1.65	2.32	1.38	2.23	1.58	2.61	2.79	2.35	2.23	2.51
Varianza	7.51	6.07	4.93	3.36	4.72	4.63	2.72	5.37	1.92	4.95	2.49	6.83	7.79	5.50	4.98	6.30
Mínimo	1	1	1	2	1	1	1	1	3	1	1	1	1	1	1	1
Máximo	9	9	9	9	9	9	9	9	9	9	8	9	9	9	9	9

Para mejorar la calidad de la toma de decisiones en la clasificación de las frases, estos sesgos plantean la posibilidad de eliminar algunos jueces. El rechazo de un juez debe fundamentarse bien y, por esa razón, se desarrollaron diversos análisis adicionales sobre las dimensiones subyacentes en las respuestas de los jueces (análisis factorial de componentes principales, análisis de *cluster* y análisis discriminante), para considerar con más detalle los estilos de respuesta de los jueces.

Los análisis de *cluster* y discriminante sirvieron para verificar la estabilidad de las clasificaciones de las frases del COCTS en cada una de las tres categorías. El primer tipo de análisis usa una medida de la distancia para asignar cada frase a un grupo, según la proximidad relativa entre los miembros de éste. Con las puntuaciones asignadas por los jueces el análisis de cluster agrupó las frases del COCTS en tres conglomerados (que se corresponden bien con las categorías *ingenua*, *plausible* y *adecuada*).

El análisis discriminante parte del conocimiento de la pertenencia de los individuos a algunos grupos, para deducir después la asignación de un nuevo individuo cuyo grupo no es conocido. Los resultados del análisis de cluster previo se usaron como asignación de pertenencia para el análisis discriminante y probar así la estabilidad de la asignación anterior. Los resultados del análisis discriminante muestran que los tres grupos empíricos de categorías exhiben una buena estabilidad en los dos análisis, de cluster y discriminante, de modo que valida el uso de las tres categorías para clasificar todas las frases del COCTS. Por tanto, los resultados de estos dos análisis confirman empíricamente la estabilidad y la potencia del sistema de clasificación en tres categorías de las frases del COCTS.

Análisis factorial de componentes principales

El análisis factorial de componentes principales (PCFA) tiene como objetivo reducir el número de variables originales y busca nuevas (llamadas factores), que resumen la varianza total de las variables originales en un número inferior de factores. Las variables dependientes del análisis factorial son los jueces, mientras que las variables independientes son las puntuaciones de éstos en las frases del COCTS. El análisis agrupa a los jueces según la intensidad de sus correlaciones mutuas mediante las 637 puntuaciones dadas a las frases, de modo que los jueces mejor interrelacionados tienden a agruparse en el mismo factor principal. El análisis de componentes principales permite la identificación de los jueces que sistemáticamente quedan fuera de los factores principales en las diferentes dimensiones.

Se ensayaron varios análisis factoriales exploratorios independientes para todas las frases y para cada una de las dimensiones del COCTS. Según los resultados de los análisis, se probaron las eliminaciones de algunos jueces para mejorar la máxima coherencia empírica entre jueces y la máxima parsimonia (simplicidad) en el número de factores obtenidos. Las sucesivas eliminaciones de algunos jueces en los diversos análisis factoriales mostraron que la varianza total explicada por los factores empíricos no disminuía significativamente, pero sí disminuyó el número de factores (mayor parsimonia) y mejoró la correlación y cohesión entre los jueces restantes.

La Tabla III muestra el análisis factorial de componentes principales para todos los jueces, el número de factores y la varianza que dichos factores explican para cada dimensión. El análisis factorial, utilizando todas las frases del cuestionario, agrupa a los 16 jueces en dos factores, que explican cada uno 33% y 20% de la varianza total. Para cada dimensión del COCTS se obtiene un número diferente de factores, que oscilan entre dos y cinco, aunque la mitad de las dimensiones (4 de 8) tiene una estructura de tres factores.

La información obtenida de estos análisis sugiere que al eliminar a algunos jueces se permitirá una reducción significativa del número de factores (aproximadamente la mitad de ellos) y también aumentará la coherencia, sin perder pluralidad de opiniones. Por otro lado, los diferentes análisis factoriales indican qué jueces quedan fuera de los factores principales y quiénes pueden ser los candidatos más probables a ser eliminados sin que disminuya el poder explicativo (aquéllos que reducen las variaciones explicadas significativamente). A partir de este análisis interpretativo, se eliminaron 5 de 16 jueces del grupo inicial y quedó un conjunto de 11 jueces para volver a comprobar en un nuevo análisis de factores.

El análisis de componentes principales de estos 11 jueces muestra el número de factores y la cantidad de varianza explicada en cada dimensión (véase la Tabla III). Para todas las frases del COCTS, los 11 jueces se agrupan en un único factor que explica 52% de la varianza total. Para cada dimensión, se obtiene un número variable de factores (uno a tres), aunque la mayoría de las dimensiones (cinco) tiene un solo factor. En suma, el grupo de 11 jueces tiene la mitad del número de

factores que 16 jueces, mientras aumenta significativamente la coherencia y la parsimonia en la descripción de los datos.

Tabla III. Comparación de los resultados de los análisis factoriales* para 8, 16 y 11 jueces.** Las varianzas múltiples se corresponden con el número de factores indicado

Temas del COCTS	Ítems/ Frases	Factores			Varianza explicada (%)		
		16 jueces	11 jueces	8 jueces	16 jueces	11 jueces	8 jueces
Todos los temas del COCTS.	100 / 637	2	1	1	33 / 20	52	54
Relaciones entre ciencia y tecnología.	10 / 70	2	1	1	32 / 27	53	58
Influencia de la sociedad en la ciencia y la tecnología.	15 / 101	3	1	1	30 / 16 / 14	52	55
Influencia de la ciencia y la tecnología en la sociedad.	22 / 140	3	2	1	24 / 20 / 18	32 / 29	54
Educación en ciencia y tecnología.	3 / 20	4	2	2	26 / 20 / 17 / 17	35 / 35	45 / 28
Características personales de los científicos y de género.	12 / 82	3	2	2	29 / 19 / 16	37 / 25	33 / 33
Construcción colectiva del conocimiento científico (Sociología interna de la ciencia).	14 / 84	3	1	1	32 / 15 / 13	52	53
Toma de decisiones tecnológicas.	4 / 27	5	3	2	25 / 17 / 15 / 11 / 10	37 / 18 / 17	40 / 31
Naturaleza del conocimiento científico (Epistemología de la ciencia).	20 / 113	2	1	1	38 / 19	56	57

* Método de componentes principales y rotación Varimax.
** Todos los jueces tienen experiencia en investigación sobre temas CTS.

Al probar la eliminación de tres jueces más (dejando sólo un grupo de 8 jueces que son quienes han investigado en el tema de CTS), la comparación de resultados con el conjunto de 11 jueces no muestra mejoras significativas. El número de factores es casi el mismo en ambos casos y las diferencias en la varianza explicada no son importantes (véase la Tabla III) excepto en dos dimensiones (una de ellas con sólo cuatro ítems). En otros términos, esta reducción adicional de jueces no supone una mejora sustancial de la coherencia global de los jueces. A partir de estos análisis, el grupo de 11 jueces parece el mejor conjunto posible para la clasificación de las frases del COCTS. La selección de 11 jueces válidos de los 16 disponibles permite inicialmente una notable reducción en el número de

factores, sin disminuir significativamente el porcentaje de varianza explicada. Un único factor explica más de la mitad de la varianza total para el conjunto completo de las frases del cuestionario, así como para la mitad de las dimensiones del COCTS, resultado que es particularmente pertinente para la parsimonia y simplicidad.

Las categorías de las frases del COCTS

Es importante notar que la reducción del número de jueces propuesta en el párrafo anterior se basa en criterios globales, no específicos o personales, y en el análisis general de las 637 puntuaciones emitidas por cada juez.

Reducir a 11 el número de jueces no tiene un impacto significativo en la clasificación de las frases del COCTS. La comparación de la distribución de la puntuación directa entre los grupos inicial y reducido de los jueces (16 y 11, receptivamente) muestra que las diferencias cuantitativas entre los dos grupos no son relevantes. No obstante, el grupo inicial de 16 jueces presenta frecuencias mayores en las puntuaciones centrales de la escala, mientras que el grupo reducido de 11 jueces tiene frecuencias mayores en las puntuaciones más bajas (véase la Figura 1).

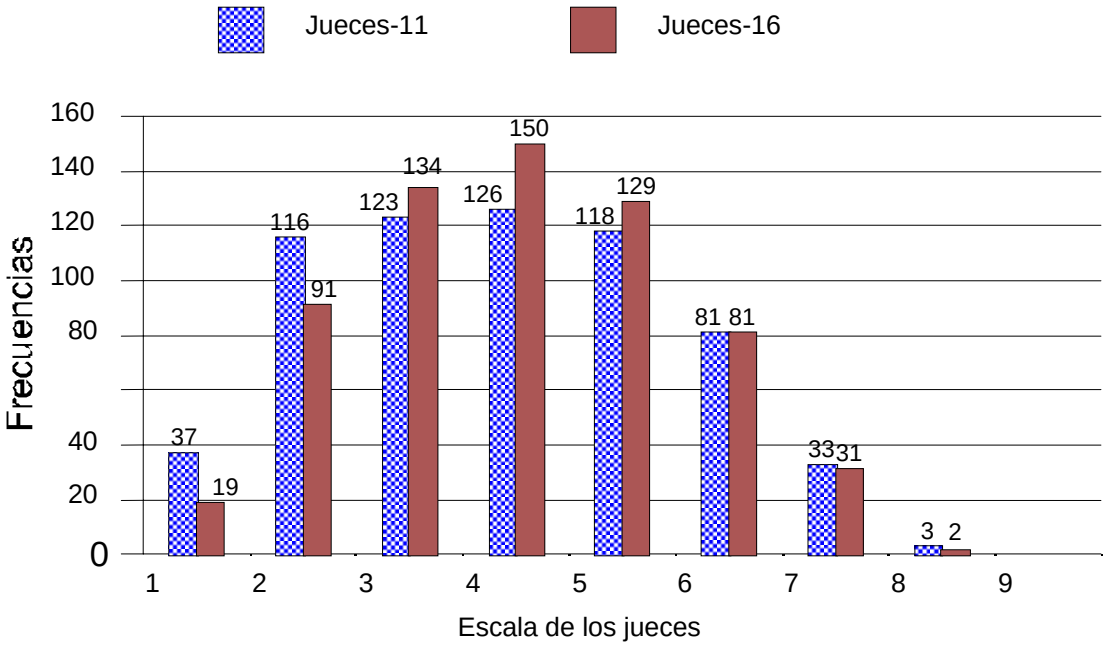

Figura 1. Distribución de las puntuaciones medias de las frases sobre los intervalos de la escala de puntuaciones de los once jueces

Por otro lado, la comparación de las clasificaciones de las frases basadas en el criterio del intervalo absoluto y las zonas de frontera difusas (las puntuaciones entre 3-4 y 6-7) muestra que el número total de frases que cambian su clasificación entre ambos modelos no es excesiva (aproximadamente 14%). La diferencia más importante entre ambos grupos de jueces aparece en la categoría *ingenua* (véase la Tabla IV).

Las puntuaciones de las respuestas directas del grupo de 11 jueces se usaron para clasificar todas las frases del COCTS (escalas) en cada una de las tres categorías (I/P/A), mediante la aplicación de dos criterios fundamentales: la puntuación media de la frase (la media aritmética de las puntuaciones individuales de los 11 jueces) y el voto mayoritario de los jueces a favor de una categoría (categoría asignada por la mayoría de jueces). En general, parece razonable que todas las frases cuyas puntuaciones medias caen dentro de los intervalos enteros naturales definidos para la escala presentada a los jueces deben asignarse a la categoría acorde correspondiente con su valor medio. Así, las puntuaciones medias comprendidas entre 1 y 3 puntos serían *ingenuas*, las puntuaciones medias entre 4 y 6 serían *plausibles* y entre 7 y 9 *adecuadas*. De acuerdo con este criterio, las frases cuyas puntuaciones medias pertenecen a los intervalos fronterizos (3, 4) y (6, 7), no se podrían clasificar (frases denominadas “difusas” en la Tabla IV).

Tabla IV. Distribución del número de frases en las tres categorías (*ingenuas*, *plausibles* y *adecuadas*), según sus puntuaciones medias (11 y 16 jueces) y tres criterios diferentes

Categorías	**Intervalo absoluto de puntos de corte en el medio I <=3.5 < P < 6.5<= A**	**Intervalo absoluto y fronteras indefinidas I <= 3; 4 >= P <=6; 7<= A**		**Voto de la mayoría de jueces**
	11 jueces	**11 jueces**	**16 jueces**	**11 jueces**
Ingenuas	202	153	101	220
Plausibles	364	258	261	252
Adecuadas	71	43	62	105
Difusos	0	183	209	58

En ausencia de sesgo o distorsión en las puntuaciones de los jueces, la clasificación de las frases cuyas puntuaciones medias se sitúan en los intervalos fronterizos –entre dos categorías vecinas– podrían clasificarse mediante el criterio simple de dividir el intervalo fronterizo en dos partes iguales, asignando cada puntuación media a la parte correspondiente del intervalo (véase la Tabla IV). Sin embargo, las puntuaciones de los jueces muestran dos sesgos importantes, uno que proviene directamente de la distribución de puntuaciones medias de todas las frases y otro debido al propio esquema de clasificación empleado.

La distribución de la puntuación media de las frases muestra un sesgo hacia las puntuaciones bajas (véanse la Tabla III y la Figura 1). Los indicadores principales

de este sesgo son las puntuaciones medias personales para la mayoría de los jueces, que son inferiores a 5 puntos (el punto central de la escala original), y la distribución general de la puntuación media de las frases (media = 4.43; DT = 1.60), que está sesgada hacia las puntuaciones más bajas. Asimismo, las puntuaciones medias de las frases tampoco están distribuidas homogéneamente a lo largo del rango de la escala, la cual tiene una longitud de ocho intervalos con nueve posiciones de valores enteros. Las puntuaciones medias de los jueces en las 637 frases varían entre 1.18 puntos (puntuación mínima) y 8.18 puntos (puntuación máxima), intervalo claramente asimétrico respecto al rango de la escala. La distancia entre este mínimo y máximo es de 7 puntos, lo que indica que la escala empírica proveniente de las puntuaciones de los jueces se acorta en una unidad (un intervalo menos) respecto a la escala original (ocho intervalos). Además, puede verse que la mayor reducción de la escala tiene lugar en el intervalo más alto del rango, entre 8 y 9 puntos, que está prácticamente vacío (véase la Figura 1). Esta distribución sesgada negativamente manifiesta el patrón que siguen los jueces de asignar menos puntuaciones altas que bajas.

El criterio de clasificación de las frases debería neutralizar este sesgo para representar con más fiabilidad las opiniones de los jueces. Una manera de compensar las consecuencias del sesgo podría consistir en mantener un escalamiento equilibrado de tres intervalos iguales, pero aplicando este escalamiento al rango de puntuaciones empírico [1.18 - 8.18]. Por ejemplo, si este intervalo real se distribuye homogéneamente entre las tres categorías, a cada una le correspondería una longitud de intervalo de 2.33 puntos, de modo que el posible escalamiento basado en las puntuaciones medias reales dibujaría este criterio del intervalo homogéneo con los siguientes puntos de corte: *ingenua*, menor que 3.52; *plausible*, entre 3.52 y 5.84; *adecuada*, superior a 5.84 puntos.

La aplicación de estos puntos de corte llevaría a una clasificación definida de las frases. Sin embargo, la clasificación resultante aún puede ser incompleta, pues sólo se basa en las puntuaciones medias de los jueces, mientras que otros aspectos significativos de sus puntuaciones quedan fuera. El primero de éstos surgiría de considerar las puntuaciones directas emitidas por los jueces en las frases como si fueran un voto registrado a favor de cada categoría. Parece de sentido común que las frases que logren una mayoría absoluta (seis o más votos favorables) o relativa (cinco votos) deberían asignarse a la categoría más votada. No obstante, el criterio de votación tiene también algún inconveniente, como pueden ser los frecuentes empates; cuando dos categorías alcanzan el mismo número de votos mayoritarios, el método de la votación no permite decidir la categoría, de modo que el criterio principal debería ser la puntuación media de los jueces.

Por último, también debe tomarse en consideración otro sesgo estructural que proviene de la ordenación en tres categorías, adoptada para el escalamiento: *ingenua* (puntuaciones más bajas), *plausible* (puntuaciones intermedias) y *adecuada* (puntuaciones más altas). Sin embargo, es bien conocida la tendencia hacia el centro de las puntuaciones medias, que desplaza la clasificación hacia la

categoría intermedia (plausible). En efecto, las desviaciones potenciales de las puntuaciones más bajas, debidas a que algunos jueces evalúan las frases presumiblemente ingenuas con puntuaciones más altas de lo debido, sólo pueden tender hacia las puntuaciones mayores, desplazando siempre las puntuaciones medias de las frases ingenuas hacia arriba y contribuyendo a aumentar el número de frases en la categoría plausible. Análogamente, las puntuaciones más altas tienden a ser desplazadas hacia abajo, porque algunos jueces evalúan las frases adecuadas con puntuaciones más bajas, contribuyendo a aumentar también el número de frases plausibles. Por otro lado, para las frases potencialmente plausibles, las desviaciones parciales pueden tender hacia arriba, logrando puntuaciones más altas, así como hacia abajo, logrando puntuaciones más bajas. En general, el efecto neto medio de estas desviaciones al azar sería nulo para la categoría intermedia. Este sesgo estructural produce una inflación en la categoría central (frases plausibles) por ser topológicamente la categoría intermedia entre las otras dos. En resumen, el efecto final de estos sesgos sería siempre aumentar el número de frases en la categoría plausible.

Afrontar este sesgo estructural requeriría disminuir el número de frases plausibles, clasificando algunas de ellas como *ingenuas* o *adecuadas*. Las frases que probablemente son más susceptibles de esta compensación deben quedar situadas en las áreas fronterizas de la categoría central (plausible), por encima y por debajo. Una manera de tratar este sesgo sería reducir el rango de puntuaciones asignado a la categoría central (plausible), decidiendo los puntos de corte más apropiados para cada categoría. El modelo de intervalo homogéneo con su puntos de corte en 1.18 - 3.51 - 5.84 - 8.18 tiene su punto de corte intermedio superior a 3, la puntuación ingenua más alta en 3.51 puntos; las puntuaciones en la zona fronteriza entre 3 (extremo superior de las puntuaciones ingenuas) y 4 puntos (inicio de puntuaciones plausibles) podrían corresponder a cualquiera de las categorías *ingenua* o *plausible*. Parece bastante racional asignar las frases que quedan en la mitad inferior a la primera categoría (ingenua) y las frases que quedan en la mitad superior a la categoría central (plausible). Asimismo, el área fronteriza entre las categorías adecuadas y plausibles tiene su punto de corte en 5.85, por debajo de 6 puntos (la puntuación entera más alta asignada a la categoría plausible), dando lugar a un serio problema conceptual, ya que asignaría algunas frases con puntuaciones medias menores que 6 puntos (puntuación superior de la categoría plausible) a la categoría superior (adecuadas). Parece obvio que el mínimo punto de corte conceptualmente aceptable debe ser la puntuación de 6. Los resultados para las asignaciones de categoría, según el modelo de puntos de corte fijos a la puntuación 3.51 (para la frontera entre ingenua y plausible) y a la puntuación 6 (para la frontera entre plausible y adecuada) se muestran en la Tabla V. Con este modelo de puntos de corte, el número de frases en la categoría *plausible* es con mucho el más frecuente (casi la mitad de las frases).

Tabla V. Distribución de las frases en las tres categorías, según las puntuaciones medias por los 11 jueces, de acuerdo con el modelo del intervalo igual y el refinamiento que aplica al modelo de la mayoría de votos

Categorías	Modelo de intervalos iguales I<=3.51; 4>=P<=6; 6< A	Plausible con mayoría ingenua	Plausible con mayoría adecuada e ítem carente de categoría adecuada	Asignación de categorías definitiva
Ingenua	203	+17		220
Plausible	303	-17	-12	274
Adecuada	131		+12	143

En este punto, el análisis de los votos de la mayoría de los jueces podría mejorar la asignación de la categoría para ciertas frases. Algunas de ellas (17) localizadas en el área fronteriza inferior tienen puntuaciones medias superiores a 3.51 puntos (correspondiendo a la categoría plausible); pero la mayoría de los jueces les han asignado puntuaciones individuales correspondientes a la categoría *ingenua*. Parece lógico en estos pocos casos compensar el sesgo de la categoría intermedia (*plausible*); de esta manera, el criterio de la mayoría de los jueces prevalecería para asignar estas frases definitivamente a la categoría *ingenua*.

De modo semejante, algunas frases (19) localizadas en el área fronteriza superior, cuyas puntuaciones medias son menores que 6 puntos (puntuación entera más alta de la categoría plausible), reciben el apoyo de la mayoría de los jueces para asignarles la categoría *adecuada* en sus puntuaciones individuales. En este caso, no parece tan razonable aplicar el criterio de la mayoría de votos para clasificar todas estas frases como adecuadas sin ninguna otra condición adicional que lo haga estrictamente necesario, como ocurre cuando un ítem no tiene ninguna frase clasificada como adecuada. De este modo, sólo 12 de las 19 frases cumplieron ambas condiciones y se asignaron a la categoría *adecuada*, aunque sus puntuaciones medias fueran menores que 6 (véase un ejemplo en la Tabla VI).

Tabla VI. Texto y puntuaciones del ítem 20411, cuya clasificación muestra algunos rasgos especiales: ninguna de sus frases resulta clasificada como adecuada y un juez (J7) no valora este ítem

Puntuaciones directas de los jueces											Puntuaciones medias				Cate-goría	Texto del ítem
J 1	J 2	J 3	J 4	J 5	J 6	J 7	J 8	J 9	J 10	J 11	Ingen.	Plaus.	Adec.	Media		
																20411. Algunas culturas tienen un punto de vista particular sobre la naturaleza y los humanos. Los científicos y la investigación científica son afectadas por las creencias religiosas o éticas de la cultura donde se realiza el trabajo. Las creencias éticas y religiosas influyen en la investigación científica:
1	5	4	3	5	4	X	1	6	7	6	3	6	1	4.20	P	A. Porque algunas culturas quieren que se haga investigación específica cuyos resultados la beneficien.
9	7	4	5	5	4	X	4	6	7	8	0	6	4	5.90	P	B. Porque inconscientemente los científicos pueden elegir

																investigación que apoye las creencias de su cultura.
6	7	4	2	4	4	X	1	3	5	2	4	5	1	3.80	P	C. Porque la mayoría de los científicos no harían investigación que fuera contra su educación o sus creencias.
5	2	7	3	4	5	X	6	4	7	5	2	6	2	4.80	P	D. Porque todos reaccionamos de forma diferente ante nuestras culturas. Estas diferencias individuales de los científicos influyen en el tipo de investigación que hacen.
4	3	5	4	7	4	X	7	5	7	3	2	5	3	4.90	P	E. Porque grupos poderosos que representan a algunas creencias religiosas, políticas o culturales apoyarían determinados proyectos de investigación, o darían dinero para que no se hagan ciertas investigaciones.
																Las creencias éticas y religiosas NO influyen sobre la investigación científica:
4	1	3	4	4	4	X	1	4	4	1	4	6	0	3.00	I	F. Porque la investigación continúa a pesar de los enfrentamientos entre los científicos y ciertos grupos religiosos o culturales (por ejemplo, entre partidarios de la evolución y defensores de la creación).
5	2	4	4	3	5	X	1	4	3	4	4	6	0	3.50	I	G. Porque los científicos investigarán temas que son de importancia para la ciencia y ellos mismos, independientemente de las opiniones culturales o éticas.

Por último, estas asignaciones de clasificación para las frases del COCTS basadas en las puntuaciones medias y los votos de la mayoría de los jueces todavía dejan ocho ítems cuyas frases sólo reciben clasificaciones ingenuas o plausibles porque carecen de una frase adecuada. Como último recurso, la asignación de la categoría para estos ocho ítems difíciles requeriría un escrutinio cuidadoso de sus contenidos y, quizás, nuevas puntuaciones de jueces para mejorar su diseño y clasificación.

Conclusiones

En los últimos años, la medida de las actitudes relacionadas con la ciencia y la tecnología ha mejorado por la contribución tanto de la investigación cualitativa (entrevistas y portafolios) como de la investigación cuantitativa (véase la revisión de Lederman, Wade y Bell, 1998), donde sobresalen los cuestionarios desarrollados empíricamente como el VOSTS (Aikenhead y Ryan, 1989), el TBA-CTS (Rubba y Harkness, 1993) y la adaptación española de ambos denominada COCTS (Manassero y Vázquez, 1998). Estos instrumentos ofrecen algunas de las ventajas de los métodos cualitativos y cuantitativos y disminuyen algunos de sus inconvenientes. Con el fin de mejorar su aplicación se han sugerido y discutido

distintos modelos para responder los cuestionarios, diferente a sólo escoger la frase preferida en cada ítem.

El Modelo de Respuesta Múltiple (MRM) ofrece mejoras importantes a los otros modelos de respuesta, puesto que aborda toda la información disponible en cada ítem. Para llevar a cabo el MRM y su procedimiento de puntuación (calculando un índice actitudinal) es necesaria la clasificación previa de las diversas frases de cada ítem.

Este estudio se ha ocupado de informar la metodología de escalamiento aplicada mediante un panel de jueces expertos para clasificar las numerosas frases del COCTS en una de las tres categorías normalizadas, siguientes: *ingenua*, *plausible* y *adecuada*. El escalamiento de las frases del COCTS es una piedra angular para el cómputo de un índice actitudinal cuantitativo, global e independiente para cada ítem, como se ha explicado en otro lugar (Vázquez y Manassero, 1999). Esta propuesta muestra los progresos interpretativos y cuantitativos en la evaluación de las actitudes hacia la ciencia y en los estudios para la posterior investigación empírica, pues el índice actitudinal permite todo tipo de procedimientos de inferencia estadística (comparaciones entre grupos, verificación de hipótesis, etc.), habituales en este tipo de investigación.

La fortaleza y validez del índice actitudinal para la investigación cuantitativa es clara, pero al mismo tiempo también debe prestarse atención a las posibilidades interpretativas del mismo, respecto a los ítems del COCTS; por ejemplo, la capacidad de describir detalles cualitativos de las actitudes de los individuos respecto a una amplia gama de temas CTS. Por consiguiente, esta propuesta permite lograr los siguientes objetivos: Medir válidamente las actitudes, aprovechar toda la información disponible en cada ítem, perfilar interpretativamente las actitudes de los individuos y posibilitar la aplicación a las diferentes respuestas de los métodos cuantitativos avanzados, tales como análisis e inferencias estadísticos.

Rubba, Schoneweg-Bradford y Harkness (1996) al aplicar en su trabajo seminal un acuerdo general de mayoría entre cinco jueces encontraron diferencias muy significativas entre las opiniones de éstos sobre las interacciones CTS, hasta el extremo de que dos de ellos se situaban sistemáticamente fuera del rango. En consecuencia recomendaron aumentar el número de jueces. La clasificación por jueces de las frases del COCTS presentada aquí se basa en las puntuaciones de un panel de 11 jueces, extraído de un conjunto inicial de 16, por medio de criterios estadísticos globales y no individuales (las estadísticas descriptivas generales de las puntuaciones de los jueces y el análisis factorial de componente principales). Los criterios aplicados para obtener parsimonia y consistencia y evitar los sesgos han sido múltiples: las puntuaciones medias de los jueces, el acuerdo general de la mayoría de ellos y algunas correcciones de sesgos globales, como la definición de una escala del intervalo igual asimétrica y la compensación por la ausencia de frases de categoría adecuada en algunas preguntas. Después de aplicar racionalmente estos criterios por medio de la metodología descrita anteriormente,

las 637 frases de opción múltiple de todos los ítems del COCTS se han clasificado en una de las tres categorías, obteniéndose 143 frases en la categoría adecuada, 274 en la categoría plausible y 220 en la categoría ingenua.

La clasificación del COCTS puede tener un papel en el desarrollo del currículo CTS para las aulas de ciencia, como una guía curricular para la amplia gama de temas de esta área. De hecho, cada pregunta puede verse como una aproximación introductoria a las diferentes posiciones sobre el tema CTS planteado, las cuales se formulan en las frases correspondientes. Asimismo, la gran variedad de temas de los ítems del COCTS puede permitir al profesorado de ciencias seleccionar, ordenar y desarrollar los contenidos CTS para las distintas etapas y niveles, seleccionar el conjunto apropiado de ítems para desarrollar los temas y emplear los ítems del COCTS como si fueran un guión. Éstos pueden utilizarse en las aulas de ciencias como contenidos CTS explícitos, de varias maneras, como fomentar las discusiones entre los estudiantes, buscar casos históricos, leer textos ilustrativos o buscar autores que defiendan o critiquen las diferentes opciones. Teniendo en cuenta que los profesores de ciencias reciben poca formación en los temas CTS, las recomendaciones curriculares anteriores también son válidas para la formación inicial o continua del profesorado que se dirige a poner al día y mejorar las actitudes de los profesores hacia los temas CTS (Abd-El-Khalick, Bell y Lederman, 1998; Akerson, Abd-El-Khalick y Lederman, 2000).

Por otra parte, los procesos de escalamiento por medio de los jueces siempre tienen un desajuste estadístico que desafía la calidad global del COCTS como instrumento de evaluación CTS. Este desajuste proviene de la variabilidad de juicio causada por el carácter dialéctico de los temas CTS y por el propio proceso de clasificación, ya que involucra el uso de parámetros estadísticos afectados por errores humanos aleatorios. Estas fuentes de error estadístico inherentes al escalamiento de los jueces constituyen un paso valioso, pero provisional, que necesita mejorar mediante procesos iterativos de realimentación. Por ejemplo, Rubba, Schoneweg-Bradford y Harkness (1996) lograron el acuerdo por medio de la discusión abierta y directa con sus cinco jueces, una tarea que está más allá del alcance de este estudio porque los jueces integrantes del panel provienen de lugares distantes y, por consiguiente, la tarea de revisión es realmente difícil. Un juez (científico) comentó que la tarea de clasificar los ítems CTS fue “una de las tareas más duras que he realizado jamás” (Rubba, Schoneweg-Bradford y Harkness, 1996, p. 396). En su lugar, se propuso a los investigadores de didáctica de las ciencias discutir los resultados que se han publicado con detalle (Manassero, Vázquez y Acevedo, 2001). Aunque pueda parecerlo, no se pretende sostener que exista un amplio acuerdo dentro de las comunidades de ciencia, historia, filosofía y sociología de la ciencia; si éste fuera el caso, obviamente no hubiéramos apelado a los jueces para que clasificaran las posiciones. Aunque estos resultados puedan interpretarse como una prueba definitiva de la falta de acuerdo general en CTS, hay que destacar que, como conclusión del trabajo de los jueces son posibles algunos acuerdos parciales sobre los temas CTS (McComas, Clough y Almazroa, 2000). Así pues, la clasificación de las frases no constituye ningún tipo de sistema absoluto para clasificar rígidamente las respuestas como

correctas o falsas, sino sólo para perfilar aproximadamente las actitudes de las personas.

Las los ítems del COCTS y del VOSTS se desarrollaron empíricamente, como un esfuerzo por evitar los fallos comunes de los instrumentos de evaluación actitudinales aplicados en la bibliografía mencionada en la introducción. La naturaleza empírica del COCTS es, ciertamente, un aspecto metodológico fuerte, pero el proceso de escalamiento por medio de los jueces muestra algunas debilidades específicas. Algunos de ellos expresaron descontento y críticas sobre preguntas concretas; por ejemplo, los jueces fueron incapaces de encontrar alguna frase adecuada en varios ítems. Por otro lado, un juez reclamó que algunas preguntas necesitaban una distribución más equilibrada y simétrica de opciones para representar mejor el rango completo de las respuestas potenciales (entre los dos extremos actitudinales). Además, el proceso de escalamiento por medio de los jueces desplegó otras dificultades que necesitan ser consideradas, tales como: diseñar ítems más equilibradas que puedan presentar las tres categorías en cada pregunta (por ejemplo, aumentando el número de opciones en algunos ítems). Estas sugerencias son especialmente críticas porque afectan el carácter empíricamente desarrollado de los ítems, pues algunos de ellos tendrían que ser modificados o reconstruidas y quizás perderían entonces sus COCTS propiedades empíricas iniciales (validez de constructo).

La atención a todas estas sugerencias representaría una mejora neta para la calidad del COCTS, porque tales demandas surgen de la especialización de los jueces y no es probable que pudieran surgir del conocimiento simple de personas poco expertas, como los jóvenes encuestados como base de su desarrollo empírico. Sin embargo, para mantener el carácter empíricamente desarrollado del COCTS, su mejora debe hacerse con cautela; por ejemplo, por medio de la comprobación empírica previa de las nuevas frases que pudieran incluirse.

Por último, se debe destacar que el enfoque actitudinal en los temas CTS puede parecer extraño a muchos que preferirían los términos “concepciones”, “opiniones” o “creencias”. Todos estos enfoques pueden ser complementarios, pero estimamos que el enfoque actitudinal parece más global, mientras el enfoque conceptual se limita al conocimiento de hechos o procedimientos. Esta diferencia puede verse en un ejemplo clásico, como sería la distinción entre *poseer* el conocimiento sobre los efectos adversos del tabaco en la salud y la *actitud personal* (acuerdo o desacuerdo) hacia el acto de fumar; es decir, la comprensión de los efectos de tabaco en la salud es un factor importante que puede afectar la actitud personal, pero no es el único factor determinante y, de hecho, hay personas legas en este asunto que no fuman en absoluto y médicos que son grandes fumadores. El mismo esquema se aplica a la naturaleza cargada de valores de los temas CTS; una persona puede entender los procesos de calentamiento global del planeta, pero al mismo tiempo puede tener un comportamiento contradictorio con esa comprensión. Nuestra posición es que la responsabilidad de la ciencia escolar debe ser contribuir significativamente a la comprensión pública de la ciencia en nuestra sociedad, lo que sin duda puede

interpretarse en términos del conocimiento, pero también puede referirse a las convicciones cívicas, compromisos y conductas de los estudiantes.

El enfoque actitudinal integra de modo bastante natural conocimientos, afectos y conductas y subraya la conciencia de que la ciencia escolar tiene que alcanzar finalidades educativas de valores, los cuales implican comprensión pero también una opción personal; metas que deben ser importantes para todos los estudiantes. En general, educar en valores en este campo no significa inculcar un cierto conjunto de ellos sobre la ciencia y la tecnología; más bien significa exponer el conjunto completo de valores en torno a un tema (cada ítem del COCTS despliega una amplia gama de posibles posiciones sobre un tema), explorarlos, discutirlos y facilitar la adhesión personal cuando sea posible. Sin embargo, en algunos casos concretos los expertos están de acuerdo (por ejemplo, la naturaleza provisional del conocimiento científico, la creatividad en la metodología científica, la influencia de la sociedad en la ciencia, etc.) en que educar para lograr valores también significa señalar la respuesta adecuada.

Metodológicamente, el procedimiento de escalamiento presentado aquí aporta un sistema más completo para evaluar el perfil en un tema CTS, mostrando explícitamente el marco de referencia para calcular las puntuaciones (Manassero y Vázquez, 2002; Manassero, Vázquez y Acevedo, 2003). Por otro lado, el método cuantitativo no excluye la interpretación de los datos a través de un análisis cualitativo; de hecho, el análisis más simple basado en cada frase permite conseguir rápidamente conclusiones cualitativas (Manassero, Vázquez y Acevedo, 2004). Sin embargo, el objetivo de la aplicación del COCTS no es tanto clasificar personas o grupos según el grado de adecuación de sus actitudes, sino ayudar a describir y comparar con detalle sus complejas posiciones.

Para concluir, conviene señalar que este estudio no asume que exista un acuerdo general sobre los temas CTS, aunque el acuerdo parcial también está surgiendo en algunos ítems específicos (McComas, Clough y Almazroa, 2000; Vázquez, Acevedo y Manassero, 2004); no obstante, hay un reconocimiento explícito de que la principal dificultad para evaluar los temas CTS es la falta de acuerdo general. En este caso, la clasificación presentada aquí es una aproximación simple, pero explícita, de un marco de referencia para la evaluación, que se somete a las críticas de la comunidad de investigadores y expertos en educación científica. En suma, la aplicación de esta clasificación en los futuros proyectos de investigación relativos a la evaluación de actitudes relacionadas con la ciencia y los temas de CTS por medio del MRM y la nueva métrica, deberían ofrecer la prueba empírica de estos rasgos del COCTS.

Referencias

Abd-El-Khalick, F., Bell, R. L. y Lederman, N. G. (1998). The nature of science and instructional practice: making the unnatural natural. *Science Education, 82* (4), 417-436.

Acevedo, J. A. (1997). Ciencia, tecnología y sociedad (CTS). Un enfoque innovador para la enseñanza de las ciencias. *Revista de Educación de la Universidad de Granada, 10*, 269-275.

Acevedo, J. A. (2000). Algunas creencias sobre el conocimiento científico de los profesores de Educación Secundaria en formación inicial. *Bordón, 52* (1), 5-16. Consultado en la sección Sala de Lecturas CTS+I de la OEI, el 10 de enero de 2005 en: http://www.campus-oei.org/salactsi/acevedo18.htm.

Aikenhead, G. S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of Research in Science Teaching, 25* (8), 607-629.

Aikenhead, G. S. (1994). Consequences to learning science trough STS: a research perspective. En J. Solomon y G. S. Aikenhead (Eds.), *STS education: International perspectives on reform* (pp. 169-186). Nueva York: Teachers College Press.

Aikenhead, G. S., Fleming, R. W. y Ryan, A .G. (1987). High school graduates' beliefs about science-technology-society: I. Methods and issues in monitoring students views. *Science Education, 71* (2), 145-161.

Aikenhead, G. S. y Ryan, A. G. (1989). *The development of a multiple choice instrument for monitoring views on science-technology-society topics* (Final report of SSHRCC Grant). Saskatoon, Canadá: University of Saskatchewan, Department of Curriculum Studies.

Aikenhead, G. S. y Ryan, A. G. (1992). The development of a new instrument: "Views on science-technology-society" (VOSTS). *Science Education, 76* (5), 477-491. También disponible en: http://www.usask.ca/education/people/aikenhead/vosts_2.pdf

Akerson, V. L., Abd-El-Khalick, F. y Lederman, N. G. (2000). Influence of a reflective explicit activity-based approach on elementary teachers' conceptions of nature of science. *Journal of Research in Science Teaching, 37* (4), 295-317.

Alters, B. J. (1997a). Whose nature of science? *Journal of Research in Science Teaching, 34* (1), 39-55.

Alters, B. J. (1997b). Nature of science: a diversity or uniformity of ideas? *Journal of Research in Science Teaching*, *34* (10), 1105-1108.

Bell, R.L., Lederman, N. G. y Abd-El-Khalick, F. (2000). Developing and acting upon one's conception of the nature of science: A follow-up study. *Journal of Research in Science Teaching, 37* (6), 563-581.

Bratt, M. (1984). Further comments on the validity studies of attitude measures in science education. *Journal of Research in Science Teaching, 21* (9), 951.

Breckler, S. J. (1994). A comparison of numerical indexes for measuring attitude ambivalence. *Educational and Psychological Measurement, 54* (2), 350-365.

Bybee, R. W. (1987). Science education and the science-technology-society (S-T-S) theme. *Science Education, 71* (5), 667-683.

Clough, E. E. y Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education, 70* (4), 473-496.

Chaiken, S., Pomerantz, E. M. y Giner-Sorolla, R. (1995). Structural consistency and attitude strength. En R. E. Petty y J. A. Krosnick, *Attitude strength. Antecedents and consequences* (pp. 387-412). Mahwah, NJ: LEA.

Eagly, A. H. y Chaiken, S. (1993). *The psychology of attitudes*. Forth Worth TX: Harcourt Brace College Publishers.

Fraser, B. J. y Tobin, K. G. (Eds.). (1998). *International handbook of science education*. Dordrecht, Países Bajos: Kluwer Academic Publishers.

Gardner, P. L. (1975). Attitude measurement: A critique of some recent research. *Education Research, 17* (2)101-105.

Gardner, P. L. (1996). The dimensionality of attitude scales: a widely misunderstood idea. *International Journal of Science Education, 18* (8), 913-919.

Gauld, C. F. y Hukins, A. A. (1980). Scientific attitudes: A review. *Studies in Science Education, 7*, 129-161.

Haladyna, T. y Shaughnessy, J. (1982). Attitudes towards science: A quantitative synthesis. *Science Education, 66* (4), 547-563.

Hodson, D. (1985). Philosophy of science, science, and science education. *Studies in Science Education, 12*, 25-57.

Hofstein, A., Aikenhead, G. y Riquarts, K. (1988). Discussions over STS at the Fourth IOSTE Symposium. *International Journal of Science Education, 10* (4), 357-366.

Kempa, R. (1986). *Assessment in science*. Cambridge: Cambridge University Press.

Laforgia, J. (1988). The affective domain related to science education and its evaluation. *Science Education, 72* (4), 407-421.

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching, 29* (4), 331-359.

Lederman, N. G., Wade, P. D. y Bell, R. L. (1998). Assessing the nature of science: What is the nature of our assessments? *Science and Education, 7* (6), 595-615.

Manassero, M. A. y Vázquez, A. (1998). *Opinions sobre ciència, tecnologia i societat*. Palma de Mallorca: Conselleria d'Educació, Cultura i Esports.

Manassero, M. A. y Vázquez, A. (2002). Instrumentos y métodos para la evaluación de las actitudes relacionadas con la ciencia, la tecnología y la sociedad. *Enseñanza de las Ciencias, 20* (1), 15-27.

Manassero, M. A., Vázquez, A. y Acevedo, J. A. (2001). *Avaluació del temes de ciència, tecnologia i societat*. Palma de Mallorca: Conselleria d'Educació i Cultura.

Manassero, M. A., Vázquez, A. y Acevedo, J. A. (2003). *Cuestionario de Opiniones sobre Ciencia, Tecnología y Sociedad (COCTS)*. Princeton, NJ: Educational Testing Service. También disponible en: http://www.ets.org/testcoll/

Manassero, M. A., Vázquez, A. y Acevedo, J. A. (2004). Evaluación de las actitudes del profesorado respecto a los temas CTS: nuevos avances metodológicos. *Enseñanza de las Ciencias, 22* (2), 299-312.

McComas, W. F. (Ed.). (2000). *The nature of science in science education: Rationales and strategies*. Dordrecht, Países Bajos: Kluwer Academic Publishers.

McComas, W. F., Almazroa, H. y Clough, M. P. (1998). The Nature of Science in Science Education: An Introduction. *Science and Education, 7* (6), 595-615.

McComas, W. F., Clough, M. P. y Almazroa, H. (2000). The role and character of the nature of science in science education. En W. F. McComas (Ed.), *The nature of science in science education: Rationales and strategies* (pp. 3-39). Dordrecht, Países Bajos: Kluwer Academic Publishers.

Munby, H. (1983). Thirty studies involving the “Scientific Attitude Inventory”: What confidence can we have in this instrument? *Journal of Research in Science Teaching, 20* (2), 141-162.

Oliva, J. M. (1999). Algunas reflexiones sobre las concepciones alternativas y el cambio conceptual. *Enseñanza de las Ciencias, 17* (1), 93-107.

Ormerod, M. B. y Duckworth, D. (1975). *Pupils attitudes' to science: a review of research*. Windsor, Reino Unido: NFER Publishing.

Petty, R. E. y Krosnick, J. A. (1995). *Attitude strength. Antecedents and consequences*. Mahwah, NJ: LEA.

Rubba, P. A. y Harkness, W. L. (1993). Examination of preservice and in-service secondary science teachers' beliefs about Science-Technology-Society interactions. *Science Education, 77* (4), 407-431.

Rubba, P. A., Schoneweg-Bradford, C. S. y Harkness, W. L. (1996). A new scoring procedure for the Views on Science-Technology-Society instrument. Internacional. *Journal of Science Education, 18* (4), 387-400.

Schibeci, R. A. (1983). Selecting appropriate attitudinal objectives for school science. *Science Education, 67* (5), 595-603.

Schibeci, R. A. (1984). Attitudes to science: Un update. *Studies in Science Education, 11*, 26-59.

Shadish, W. R. (1995). The quantitative-qualitative debates: 'Dekuhnifying' the conceptual context. *Evaluation and Program Planning, 18*, 47-49.

Shrigley, R. L. y Koballa Jr., T. R. (1992). A decade of attitude research based on Hovland's learning model. *Science Education, 76* (1), 17-42.

Smith, M. U., Lederman, N. G., Bell, R. L., McComas, W. F. y Clough, M. P. (1997). How great is disagreement about the nature of science: A response to Alters. *Journal of Research in Science Teaching, 34* (10), 1101-1103.

Solomon, J. y Aikenhead, G. (Eds.). (1994). *STS education: International perspectives on reform*. Nueva York: Teachers College Press.

Stahlberg, D. y Frey, D. (1990). Actitudes I: estructura, medida y funciones. En M. Hewstone, W. Stroebe, J. P. Codol y G. M. Stephenson (Dirs.), *Introducción a la Psicología Social* (pp. 149-170). Barcelona: Ariel.

Taber, K. S. (2000). Multiple frameworks? Evidence of manifold conceptions in individual cognitive structure. *International Journal of Science Education, 22* (4), 399-418.

Tamir, P. (1998). Assessment and evaluation in science education. Opportunities to learn and outcomes. En B. J. Fraser y K. G. Tobin (Eds.), *International handbook of science education* (pp. 761-790). Dordrecht, Países Bajos: Kluwer Academic Publishers.

Vázquez, A., Acevedo, J. A. y Manassero, M. A. (2004). Consensos sobre la naturaleza de la ciencia: evidencias e implicaciones para su enseñanza. *Revista*

Iberoamericana de Educación, edición digital. Consultado el 3 de enero de 2005, en: http://www.campus-oei.org/revista/deloslectores/702Vazquez.PDF.

Vázquez, A. y Manassero, M. A. (1995). Actitudes relacionadas con la ciencia: una revisión conceptual. *Enseñanza de las Ciencias, 13* (3), 337-346.

Vázquez, A. y Manassero, M. A. (1997). Una evaluación de las actitudes relacionadas con la ciencia. *Enseñanza de las ciencias, 15* (2), 199-213.

Vázquez, A. y Manassero, M. A. (1999). Response and scoring models for the "Views on Science-Technology-Society" instrument. *International Journal of Science Education, 21* (3), 231-247.

Waks, L. J. y Prakash, M. S. (1985). STS education and its three step-sisters. *Bulletin of Science, Technology and Society, 52* (2), 105-116.

Wareing, C. (1990). A survey of antecedents of attitudes toward science. *Journal of Research in Science Teaching, 27* (4), 371-386.

Zeidler, D. L. (1984). Thirty studies involving the "scientific attitude inventory": what confidence can we have in this instrument. *Journal of Research in Science Teaching, 21* (3), 341- 342.

Ziman, J. (1994). The rationale of STS. Education is in the approach. En J. Solomon y G. Aikenhead (Eds.), *STS education: International perspectives on reform* (pp. 21-31). Nueva York: Teachers College Press.