Revista de Investigación Educativa, 2009, Vol. 27, n.º 1, págs. 89-103

CÁLCULO DE LA FIABILIDAD Y CONCORDANCIA ENTRE CODIFICADORES DE UN SISTEMA DE CATEGORÍAS PARA EL ESTUDIO DEL FORO ONLINE EN E-LEARNING

*Juan Jesús Torres Gordillo**
juanj@us.es
*Víctor Hugo Perera Rodríguez**
vhperera@us.es

RESUMEN

Presentamos los resultados detallados del cálculo de la fiabilidad de un sistema de categorías para foros de debate online. Este trabajo se encuadra dentro una investigación sobre el estudio de la comunicación asincrónica en la formación a través de Internet. Hemos utilizado el coeficiente Kappa de Fleiss para tres codificadores. Nuestro coeficiente Kappa alcanza un valor k=0.77. Tomando varias tablas de interpretación del índice de diferentes autores, obtenemos un valor alto o bueno respecto a la fuerza de concordancia. La alta fiabilidad del sistema de categorías acredita que dicha herramienta pueda ser empleada por cualquier investigador en el ejercicio de la codificación, y en diferentes momentos, con garantías de que pueda aportar resultados que expliquen y faciliten la comprensión de los procesos de comunicación y enseñanza-aprendizaje en e-Learning.

Palabras clave: *fiabilidad entre codificadores, Kappa de Fleiss, sistema de categorías, foro online, e-Learning.*

* Dpto. Métodos de Investigación y Diagnóstico en Educación. Facultad de Ciencias de la Educación. Universidad de Sevilla. C/ Camilo José Cela, s/n. 41018 - Sevilla.

ABSTRACT

We offer detailed results measuring inter-rater reliability of a coding scheme of higher education online discussion boards. This is part of a piece of research on asynchronous communication in e-Learning. We have used Fleiss' Kappa coefficient (k) for three raters. Our Kappa coefficient reaches a value of k=0.77. If we consider various authors' interpretation tables of this index, this k value can be interpreted as a high or good value regarding the strength of agreement. The high reliability of this coding scheme allows it to be used by any researcher at any time, and guarantees results that explain the roles of communication and teaching-learning processes within e-Learning.

Key words: *inter-rater reliability, Fleiss' Kappa, coding scheme, online discussion board, e-Learning*

1. INTRODUCCIÓN

El presente trabajo se enmarca dentro de una investigación que tuvo como objetivo principal estudiar, indagar y analizar las posibilidades de la comunicación asincrónica como entorno de formación en cursos de postgrado desarrollados a través de Internet. Concretamente utilizamos la plataforma tecnológica WebCT. Para ello, se construyó y validó un sistema de categorías que permitiera analizar a posteriori los procesos comunicativos y de aprendizaje colaborativo a través del foro online. La técnica de investigación aplicada en este proceso fue el análisis de contenido.

En este artículo mostramos cómo se procedió al cálculo de la fiabilidad del sistema de categorías para el análisis del foro online en e-Learning, a través de la medición del acuerdo entre los codificadores. Una de las novedades que presentamos es que el cálculo se realiza para tres investigadores. Esto nos lleva a servirnos del coeficiente Kappa de Fleiss (Fleiss, 1981), permitiéndonos ofrecer al lector una perspectiva más avanzada de la técnica respecto al uso habitual que se da en muchas investigaciones que emplean Kappa de Cohen para dos codificadores (Cohen, 1960).

2. MARCO TEÓRICO

Durante algunos años, la falta de paradigmas o métodos de investigación motivó el escaso número de estudios rigurosos centrados en el aprendizaje en entornos de comunicación asincrónica (Marra, Moore & Klimczak, 2004). Ésta es una de las razones que nos conduce a examinar en detalle aspectos como la fiabilidad de los instrumentos utilizados.

Diversos trabajos preocupados por aclarar los conceptos y procedimientos relacionados con los criterios de rigor científico en la investigación cualitativa han supuesto un aliciente en el modo cómo abordar nuestra investigación (Sandín, 2000; Donoso, Figuera y Torrado, 2000; García, 2004).

Rourke et al. (2001a) llamaban la atención hace unos años sobre la falta de replicación de los modelos o sistemas de categorías presentados y publicados por otros autores en

torno al análisis de contenido en la comunicación mediada por ordenador. La siguiente cita deja clara evidencia de que la replicación debe ser el último eslabón en el proceso de construcción de un sistema de categorías fiable.

> "*La fiabilidad de un sistema de categorías puede ser vista como un continuum, comenzando con la estabilidad de un codificador (un codificador de acuerdo consigo mismo todo el tiempo), la fiabilidad entre codificadores (dos o más codificadores de acuerdo unos con otros), y, por último, la replicación (la capacidad de múltiples y distintos grupos de investigadores de aplicar un sistema de categorías de manera fiable). Además, el test definitivo de un sistema de categorías es la replicación*" (Rourke et al., 2001a).

Estos autores continúan afirmando que el primer test de objetividad en los estudios de contenido pasa por ser la concordancia entre codificadores, entendida como el momento en el que diferentes codificadores, codificando cada uno el mismo contenido, llegan a las mismas decisiones de codificación.

Tradicionalmente, el método más empleado para medir la concordancia entre codificadores es el estadístico de acuerdo porcentual. Dicho estadístico refleja el número de acuerdo en función del número total de codificaciones realizadas. El coeficiente de fiabilidad de Holsti (1969, citado en Rourke et al., 2001a) proporciona una fórmula para calcular el acuerdo porcentual:

$$C.F. = 2m \; / \; n1 + n2$$

donde:

m = número de codificaciones donde los dos codificadores están de acuerdo
n1 = número de codificaciones realizadas por el codificador 1
n2 = número de codificaciones realizadas por el codificador 2

No obstante, como afirman algunos estadísticos, el acuerdo entre codificadores puede ser una medida inadecuada porque no tiene en cuenta el acuerdo al azar entre investigadores (Capozzoli, McSweeney & Sinha, 1999). Esto lo corrige el estadístico Kappa de Cohen (k), que se utiliza para dos codificadores, en *n* casos y para *m* categorías nominales exhaustivas y exclusivas mutuamente.

Archer et al. (2001) emplearon el coeficiente Kappa de Cohen para medir el acuerdo entre codificadores en un sistema de categorías sobre pensamiento crítico en foros online, obteniendo un k=0.74 en su última revisión.

Otros autores han descrito el foro online como un espacio de aprendizaje cuya comunicación puede ser estructurada para apoyar la creación de significados compartidos entre los miembros de grupos colaborativos. En este sentido, la dificultad para lograr niveles aceptables de concordancia entre codificadores ha llevado a que se desarrollen sistemas alternativos para la codificación de transcripciones. Así, Jonassen & Remidez (2005) describieron el modo de cómo codificar las conversaciones de foros online más estructurados y con opciones limitadas de interacción (aceptar, rechazar, ampliar, apoyar, hipótesis, punto importante, evidencia, aprendizaje, etc.). De acuerdo con Rourke et al.

(2001a), esto supone que indirectamente se esté facilitando la autocodificación general de la propia aportación.

3. SISTEMA DE CATEGORÍAS PARA ANALIZAR EL FORO ONLINE EN E-LEARNING

La primera fase del estudio consistió en construir un sistema de categorías para analizar la comunicación asíncrona en el foro en e-Learning. Dentro de ésta, llevamos a cabo la validación del propio sistema de categorías, realizando un estudio piloto. Partiendo del primer borrador (denominado *Sistema completo*), el proceso de análisis de los foros nos llevó a realizar continuos cambios de ajuste -fusión, integración, reestructuración y eliminación- en las categorías e indicadores del sistema creado. En cada subfase obtuvimos un nuevo sistema de categorías (llamados *Sistema corregido*, *Sistema corregido 1*, *Sistema corregido 2*, etc.), hasta llegar al último y definitivo (nombrado *Sistema corregido 5* o *Sistema definitivo*).

TABLA 1
POBLACIÓN DE LA INVESTIGACIÓN

NOMBRE del CURSO[1]	TIPO	MODALIDAD	ALUMNOS / TUTORES
Curso A	Experto (250 horas, 1 año)	Semipresencial	18 alumnos 4 tutores
Curso B	Formación complementaria (100 horas, 1 año)	A distancia (e-Learning)	86 alumnos 5 tutores
Curso C	Experto (250 horas, 1 año)	A distancia (e-Learning)	66 alumnos 8 tutores
Curso D	Doctorado (dos años)	Semipresencial	23 alumnos (1er año), 18 alumnos (2º año) 7 tutores
Curso E	Formación complementaria (100 horas, 6 meses)	A distancia (e-Learning)	24 alumnos 5 tutores
		TOTAL	217 alumnos / 29 tutores

1 Con "Nombre del Curso" nos referimos al nombre que asignamos a cada curso para reconocerlos en el estudio. No es el nombre oficial del curso.

El sistema de categorías definitivo, que puede ser consultado en otras publicaciones (Torres y Perera, 2005), se divide en tres dimensiones: cognitiva, social y didáctica. Cada una de éstas consta de categorías, subcategorías e indicadores.

4. POBLACIÓN Y MUESTRA DEL ESTUDIO

Nuestra selección se dirige en torno a la población de la investigación, es decir, a los foros de debate de los cursos de e-Learning en los cuales hemos participado durante dos cursos académicos. En consecuencia, no entendemos la *población* como los participantes potenciales de un estudio, sino en los términos que lo expresan Goetz y LeCompte (1988: 88): "*también los fenómenos no humanos y los objetos inanimados pueden constituir poblaciones. Los grupos humanos realizan sus actividades en escenarios y contextos, períodos de tiempo y circunstancias finitos y especificables. Cada uno de estos factores constituye una población limitada, a partir de la cual el investigador puede obtener muestras o seleccionar*".

En la siguiente tabla se especifica la población final de la investigación con las características de cada curso:

Debemos aclarar que, en un primer momento, nuestra población la componían solamente los cursos "B", "C" y "D". Pero, a medida que avanzaban los análisis, decidimos ampliar la población con otros cursos en los cuales estábamos trabajando como tutores (curso "D") y con otros cursos con los que manteníamos alguna relación y se desarrollaban en la misma plataforma tecnológica WebCT (curso "A").

No obstante, nuestra intención no fue seleccionar todos los foros de estos cursos, puesto que los datos serían redundantes según el objetivo de estudio, además de ser innecesario y costoso en términos temporales. Nos atenemos, por tanto, a la *selección basada en criterios*, como lo han denominado Goetz y LeComte (1988)[2], tanto para identificar la población como para ir determinando la muestra (unidades de análisis o foros). Consiste en seleccionar casos con abundante información para estudios detallados (Patton, 1990) cuando alguien pretende entender algo sobre estos casos sin necesitar o desear generalizar sobre cada uno de los casos (McMillan y Schumacher, 2005). La finalidad de elegir el procedimiento de selección basada en criterios es buscar la *representatividad* de los datos. Según McMillan y Schumacher (ibídem, p. 407), se realiza para aumentar la utilidad de la información obtenida a partir de pequeños modelos, donde la información se obtiene sobre variaciones entre las subunidades. El poder y la lógica de este procedimiento consisten en que, con pocos casos estudiados en profundidad, se obtienen muchas aclaraciones sobre el tema (*abundante información*).

Con este objetivo, y dentro de las variantes de la selección basada en criterios (o muestreo intencionado), señaladas por Goetz y LeComte (1988) o McMillan y Schumacher (2005), llevamos a cabo la *selección por cuotas*, también llamada por Patton (1990) *muestreo de variación máxima*. Es una estrategia para aclarar diferentes aspectos sobre la cuestión del problema de investigación. Esta técnica, a diferencia de la *selección exhaustiva* (que cubre la totalidad de la población), se limita a un subconjunto de la población. Así, en el estudio comenzamos identificando los subgrupos relevantes, que en nuestro caso fueron dos grandes conjuntos: los *foros de carácter principal* (aquéllos de seguimiento

2 Otros autores lo llaman, de manera menos apropiada, muestreo intencional.

general del curso para tutorías, consultas y/o dudas) y los *foros específicos* (aquéllos para dudas de un tema específico, con una finalidad muy determinada o centrados en algún aspecto concreto). Nuestro interés se centraba en los foros generales, por atender al criterio de ser más ricos y variados en la información que proporcionan.

El siguiente paso fue seleccionar la muestra. La muestra definitiva fue de diez foros. La recogida de datos no fue puntual, sino un proceso que fue avanzando conforme obteníamos resultados de los análisis. Consistió en obtener un número arbitrario de unidades de análisis. En un primer momento, elegimos los foros generales más representativos, pero conforme se desarrollaban los análisis, fuimos rehaciendo la muestra (ampliando también la población, como decíamos más arriba), para garantizar el criterio de representatividad. Finalmente, llegamos a analizar todos los foros generales, apoyándonos en el criterio de *cubrir* todas las funciones que cumplían dentro de los cursos, a saber: unos sólo para la entrega de actividades y mantener la comunicación, a modo de dudas, o cuando se trata de cursos semipresenciales; otros sirven para tutorizar a nivel general el curso durante todo su desarrollo; otros más especializados que se centran en algún tema concreto; u otros sobre aspectos más técnicos.

5. LA FIABILIDAD DEL SISTEMA DE CATEGORÍAS: ACLARACIÓN DE CONCEPTOS Y ELECCIÓN DEL ESTADÍSTICO PARA EL CÁLCULO DE KAPPA

Cabe hacer una importante aclaración conceptual entre fiabilidad y concordancia entre codificadores. De una parte, la *concordancia* es un término más global que hace referencia a la medida en que dos o más codificadores están de acuerdo entre ellos. La entendemos como la proporción de acuerdos entre el número total de codificadores. Por ejemplo, hallamos la concordancia cuando dos codificadores se comparan entre sí. Por otra parte, la *fiabilidad* es un término más restrictivo que aprecia cuán precisa es una medida, esto es, cuánto se acerca a la *verdad*. Por ejemplo, hallamos la fiabilidad cuando un codificador se compara frente a un protocolo estándar asumido como *verdadero*.

En el contexto de nuestro estudio hacemos referencia a la *concordancia entre codificadores* durante el proceso de construcción del sistema de categorías; mientras que nos referimos a la *fiabilidad* en el momento en que los codificadores hacen uso del sistema de categorías definitivo sobre los datos textuales.

El sentido que adopta todo trabajo de investigación, así como los resultados a los que se llega, dependerán esencialmente del sistema de categorías creado. Es por esto que debe evitarse caer en los *peligros* que suelen darse cuando se utilizan esquemas de codificación. Entre ellos se encuentra la posibilidad que tiene el investigador de intentar reflejar su deseo o perspectiva idiosincrásica. Para salvar esta situación, Bakeman y Gottman (1989) afirman que son necesarios: a) mantener a los investigadores ingenuos respecto a la(s) hipótesis de la investigación; b) trabajar con más de un investigador; y, c) evaluar en qué medida concuerdan. Para la comprensión de este último punto, hemos de clarificar los conceptos de *precisión*, *calibración* y *deterioro de la fiabilidad*:

- *Precisión*. Es la *razón conceptual* que consiste en la codificación similar que hacen de forma independiente dos o más investigadores sobre los mismos hechos y

eventos. En otras palabras, diferentes investigadores han codificado eventos semejantes de forma similar.

- *Calibración*. Es la *razón práctica* que consiste en asegurarse que los datos que tienen que registrar distintos investigadores no varían a lo largo del tiempo. Esto se consigue comparando cada codificación con las demás, o, mejor aún, evaluando a todos los codificadores respecto a algún protocolo estándar.
- *Deterioro de la fiabilidad*. Consiste en asegurarse que la codificación de un observador sea consistente a lo largo del tiempo.

Siguiendo las definiciones de la terna anterior, entendimos que debíamos interesarnos no sólo por la concordancia entre codificadores (esto es, precisión y calibración), sino también por la fiabilidad *intra-codificador* (o lo que es igual, el deterioro de la fiabilidad). Esto supuso que en el proceso de entrenamiento de los codificadores tuviéramos claro dos objetivos con relación a la evaluación de la concordancia de los codificadores. El primer objetivo se refería a la preocupación por entrenar a los codificadores de forma que fuesen altamente precisos y calibrados. Un segundo objetivo consistió en asegurar la consistencia en la codificación de cada investigador. En definitiva, buscábamos un estadístico que pudiera describir la concordancia respecto a cómo utilizan los investigadores el esquema de codificación.

Este interés nos llevó a realizar una revisión de la literatura que nos mostrara la variedad de estadísticos existentes en el campo de la investigación social para calcular la concordancia según variables diferentes. Después de precisar las condiciones de nuestro estudio, nos basamos inicialmente en los trabajos de Cohen (1960) para dar respuesta a los objetivos que nos propusimos. Este autor formuló el cálculo para la obtención de la probabilidad en la concordancia entre dos codificadores. Dicha probabilidad es conocida como *coeficiente Kappa de Cohen* (caracterizado con el símbolo *k*), que se define como un estadístico de concordancia entre dos investigadores que corrige el azar. Como es obvio, al ser una probabilidad, toma su valor en el intervalo [0, 1]. Ésta se representa según la fórmula:

$$K = \frac{P_o - P_c}{1 - P_c}$$

P_o se define como la proporción de concordancia observada realmente y se calcula sumando las marcas que representan la concordancia y dividiendo por el número total de ellas;
Pc es la proporción esperada por azar y se calcula sumando las probabilidades de acuerdo por azar para cada categoría.

Fleiss (1981) generalizó la aplicación del índice Kappa de Cohen para medir el acuerdo entre más de dos codificadores u observadores para datos de escala nominal y ordinal. Por tanto, dado que nuestro estudio considera tres investigadores en el proceso de codificación, empleamos el *Kappa de Fleiss*, ya que éste parte de la misma fórmula

que propone Cohen, pero generalizada para más de dos codificadores. El coeficiente Kappa de Fleiss añade el cálculo del sesgo del codificador (precisión-error) y el cálculo de la concordancia (calibración). La fórmula es la siguiente:

$$\overline{K} = 1 - \frac{n\,m^2 - \sum_{i=1}^{n}\sum_{j=1}^{r} x_{ij}^2}{n\,m\,(m-1)\sum_{j=1}^{r} \overline{p}_j\,\overline{q}_j}$$

Los símbolos de la fórmula vienen identificados por las siguientes correspondencias:

n: se corresponde con el número total de conductas o códigos a registrar;
m: identifica el número de codificaciones;
xij: define el número de registros de la conducta i en la categoría j;
r: indica el número de categorías de que se compone el sistema nominal;
p: es la proporción de acuerdos positivos entre codificadores;
q: es la proporción de acuerdos negativos (no acuerdos) en codificadores (1 - p)

No obstante, para facilitar el cálculo de nuestros coeficientes Kappa de Fleiss hemos empleado un software específico. Se trata de un programa informático que funciona bajo el sistema operativo MS-DOS. Fue creado por el francés Bonnardel[3]. Nosotros hemos utilizado la versión 1.0, y se conoce como *Fleiss v.1.0*. Nos permite obtener el cálculo del coeficiente hasta un máximo de veinticinco codificadores y de dos a veinticinco códigos. Concretamente, para esta investigación contábamos con tres investigadores y el número de códigos se introdujo por dimensiones, sin llegar a superar el límite de esos veinticinco códigos.

Para introducir los datos en el programa, se construyó una matriz de doble entrada, donde la fila representaba cada uno de los mensajes, y la columna cada código. De esta forma, cada celda de la matriz podría variar entre 0 y 3. En aquellos casos en los que se dio un 0, significó que dicho código no fue asignado al mensaje en cuestión por ninguno de los tres codificadores. El 1 nos diría que sólo un investigador consideró dicho código para el mensaje. Y así hasta llegar al número 3, que mostraría el acuerdo total entre los codificadores para un mensaje. De aquí se desprende que la suma en cada fila de la matriz es igual al número total de codificadores.

Por último, respecto a la salida de resultados que obtuvimos del programa, en primer lugar nos mostró la suma de los acuerdos totales por código. Después, nos ofreció todos los resultados, donde se incluye el valor global del índice Kappa de Fleiss, así como el p-valor asociado al contraste de hipótesis donde la hipótesis nula (Ho) es k=0.

Conocido el modo cómo funciona el software, pasaremos a continuación a presentar los distintos índices obtenidos.

3 El programa, conocido como Fleiss v.1.0, y creado por el francés Philippe Bonnadel, puede obtenerse en la dirección http://perso.worldonline.fr/kappa

6. EL PROCESO DE CONCORDANCIA ENTRE CODIFICADORES: CÁLCULO DEL *COEFICIENTE KAPPA DE FLEISS*

Nuestro estudio se basa en un sistema de categorías conformado por tres dimensiones. Cada dimensión comprende a su vez un número diferente de códigos. El cálculo de Fleiss tuvo en cuenta la proporción de posibles acuerdos que ocurrieron en cada dimensión. Así, por ejemplo, la dimensión cognitiva tuvo 250 codificaciones sobre un total de 250 mensajes (codificación excluyente). La pregunta que nos planteamos en ese momento consistió en conocer cuántos desacuerdos y acuerdos se dieron para ese número de mensajes. Recordamos que las dimensiones Social y Didáctica (excepto *Enseñanza Directa*) incluyen códigos que pueden formar parte de una codificación cruzada.

Como consideración importante para el cálculo de la concordancia, advertimos que el sistema de categorías empleado comprende varios sistemas nominales. Las dimensiones Social y Didáctica presentan sistemas nominales (como, por ejemplo, la categoría *Afectiva*) que lo diferencia del resto de los códigos que definen cada dimensión. Esto supuso que debiéramos considerar Kappas particulares para los subsistemas nominales de cada dimensión, ya que cada sistema nominal incluye una probabilidad de acuerdo distinta.

Alcanzar una alta fiabilidad en el sistema de categorías resultó ser una tarea ardua y repleta de continuas dificultades que debíamos salvar. La preocupación por lograr un elevado acuerdo en las tareas de codificación requirió de un gran esfuerzo de concentración y dedicación, de igual forma que la construcción de las tablas para el cálculo del *Coeficiente Kappa de Fleiss*, en las diferentes versiones que fuimos obteniendo. El conjunto de todo este proceso se realizó de un modo sistematizado.

En primer lugar, definimos lo que para el grupo de investigadores constituía un *acuerdo*. De este modo, identificamos el *acuerdo* entre codificadores como la coincidencia común en la identificación de los códigos sobre los mismos eventos o hechos. En caso contrario, entrábamos en situaciones de desacuerdo[4].

Llegados a un consenso sobre cómo debíamos los investigadores entender el acuerdo, en un siguiente paso se definió lo que para el grupo de codificadores iba a constituir una *unidad de codificación*. En nuestro caso, el límite de las unidades estaba perfectamente delimitado por cada mensaje, independientemente de la extensión del contenido textual. Por tanto, la concordancia no necesitó demostrarse para la determinación de límites en las unidades, esto es, *establecimiento de unidades*, sino para la asignación de los códigos, es decir, *codificación de eventos, conductas y pensamientos.*

Cabe mencionar que la estrategia de codificación seguida atendía a una codificación múltiple[5], donde los codificadores anotaban los diferentes eventos particulares que ocurrían en cada mensaje a partir de las tres dimensiones que conformaban el sistema de categorías. Para este caso concreto, diversos autores afirman que es más difícil la

4 También denominada por Bakeman y Gottman (1989) como 'error de omisión' o 'error de comisión'.

5 Bakeman y Gottman (1989) utilizan el término 'clasificación de eventos de forma cruzada'.

TABLA 2
CÁLCULO DE LOS COEFICIENTES KAPPA DE FLEISS

Índice de Kappa Fleiss[6] para el 'Sistema de Categorías' (corregido 2)

Foro 3 (30 mensajes y nº líneas entre 1681-2359) ~ 3 codificadores

Dimensión Cognitiva (19 acuerdos)	**Dimensión Social** (10 acuerdos)	**Dimensión Didáctica** (7 acuerdos y 8 acuerdos)	
		Resto de la Dimensión	*Enseñanza Directa*
k=0.64	k=0.33	k=0.23	k=0.27

k=0.37 (Kappa de Fleiss medio para la codificación del foro 3)

Índice de Kappa Fleiss para el 'Sistema de Categorías' (corregido 3)

Foro 5 (58 mensajes y nº líneas entre 1-1215) ~ 3 codificadores

Dimensión Cognitiva (35 acuerdos)	**Dimensión Social** (31 acuerdos)	**Dimensión Didáctica** (20 acuerdos y 22 acuerdos)	
		Resto de la Dimensión	*Enseñanza Directa*
k=0.60	k=0.53	k=0.34	k=0.38

k=0.46 (Kappa de Fleiss medio para la codificación del foro 5)

Índice de Kappa Fleiss para el 'Sistema de Categorías' (corregido 4)

Foro 8 (98 mensajes y nº líneas entre 1-1279) ~ 3 codificadores

Dimensión Cognitiva (87 acuerdos)	**Dimensión Social** (91 acuerdos)	**Dimensión Didáctica** (91 acuerdos y 93 acuerdos)	
		Resto de la Dimensión	*Enseñanza Directa*
k=0.67	k=0.62	k=0.58	k=0.65

k=0.63 (Kappa de Fleiss medio para la codificación del foro 8)

Índice de Kappa Fleiss para el 'Sistema de Categorías' (corregido 5, definitivo)

Todos los foros (10 foros: 2039 mensajes y nº líneas entre 1-41348 líneas) ~ 3 cod.

Dimensión Cognitiva (1936 acuerdos)	**Dimensión Social** (1950 acuerdos)	**Dimensión Didáctica** (1923 acuerdos y 1944 acuerdos)	
		Resto de la Dimensión	*Enseñanza Directa*
k=0.88	k=0.69	k=0.64	k=0.87

k=0.77 (Kappa de Fleiss medio para la codificación de todos los foros)

6 Las probabilidades que presentamos tras el cálculo de Kappa de Fleiss aparecen redondeadas a dos decimales.

determinación de la concordancia; circunstancia por la que decidimos centrarnos en el cálculo estadístico de la fiabilidad para cada una de las tres dimensiones de forma separada. De este modo obtuvimos una tabla *Kappa* para cada esquema de clasificación o dimensión: *Cognitiva*, *Social*, y dentro de *Didáctica* consideramos dos opciones, una primera para *Enseñanza Directa*; y una segunda, llamada *Resto*, que se refería al conjunto de categorías restantes que no incluía la *Enseñanza Directa*.

En segundo lugar, identificamos y anotamos en diferentes tablas los acuerdos y desacuerdos. Este procedimiento se llevó a cabo mediante una actividad manual donde para cada unidad codificada los tres codificadores fuimos señalando una marca sobre el papel. Una vez finalizada cada una de las sesiones, se contabilizó las marcas que indicaban acuerdos y desacuerdos en los códigos dentro de cada unidad de registro (mensaje) para facilitar la obtención del valor de *Kappa*. A partir de este momento, dichos datos constituyeron las cifras que fueron sustituidas en la fórmula que finalmente adoptamos.

A continuación presentamos los cálculos y resultados de todos los coeficientes Kappa de Fleiss realizados durante las distintas subfases (estudio piloto y validación completa). Más adelante, en el siguiente punto, nos detendremos en la interpretación de estos resultados.

7. VALORACIÓN DE LA FIABILIDAD DEL SISTEMA DE CATEGORÍAS

La disposición para valorar algo implica necesariamente contar con criterios previos que nos permitan enjuiciar aquello que es objeto de evaluación. Así, para interpretar el valor del coeficiente Kappa, es útil disponer de alguna escala de valoración. En nuestra revisión de la literatura hemos encontrado algunas aproximaciones que los autores siempre proponen reconociendo cierta arbitrariedad.

Fleiss (1981) ofrece una clasificación de los Kappas que nos puede ayudar a interpretar los coeficientes obtenidos. Este autor caracteriza como *Regulares* los *Kappas* que se hayan entre 0.40 y 0.60, *Buenos* de 0.61 a 0.75, y *Excelentes* por encima de 0.75.

TABLA 3
INTERPRETACIÓN DEL ÍNDICE KAPPA DE FLEISS (FLEISS, 1981)

Interpretación del índice Kappa de Fleiss (Fleiss, 1981)	
Valor de K	***Fuerza de concordancia***
0.40 – 0.60	Regular
0.61 – 0.75	Buena
> 0.75	Excelente

Por su parte, Altman (1991) propone una clasificación algo más amplia. Los coeficientes registran valores que van desde 0 a 1, siendo 0 el valor donde hay mayor desacuerdo entre investigadores y 1 el punto donde encontramos mayor acuerdo. Su clasificación indica que los *Kappas* pueden ser *Pobres* (0 a 0.20), *Débiles* (0.21 a 0.40), *Moderados* (0.41 a 0.60), *Buenos* (0.61 a 0.80) y *Muy buenos* (0.81 a 1.00). Nosotros basaremos nuestras interpretaciones en esta clasificación, por ser más completa. La siguiente tabla resume su propuesta:

TABLA 4
INTERPRETACIÓN DEL ÍNDICE KAPPA DE FLEISS (ALTMAN, 1991)

Interpretación del Índice Kappa (Altman, 1991)	
Valor de K	***Fuerza de concordancia***
< 0,20	Pobre
0,21 – 0,40	Débil
0,41 – 0,60	Moderada
0,61 – 0,80	Buena
0,81 – 1,00	Muy buena

Una de las ventajas que nos proporcionan las tablas *Kappa* es la representación gráfica del desacuerdo. Una simple inspección ocular nos revela de inmediato cuáles fueron los códigos que presentaron una mayor confusión y cuáles casi nunca. Para optimizar el cálculo de *Kappa de Fleiss*, y con ello obtener versiones de los sistemas de categorías más fiables, pusimos especial atención sobre los desacuerdos más graves. De hecho, en

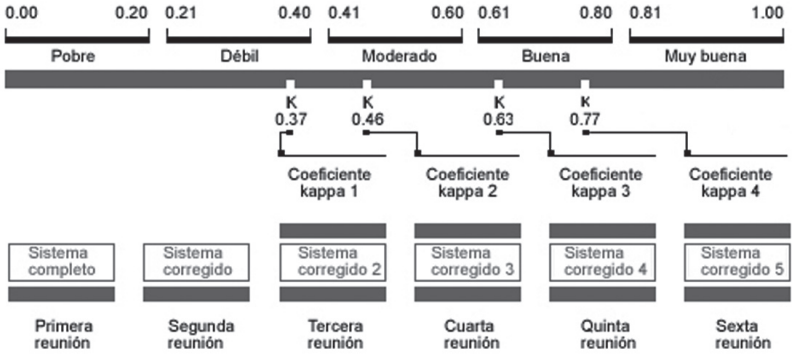

Figura 1
Valoración de los coeficientes Kappa de Fleiss.

nuestra investigación aparecieron continuos desacuerdos que abrieron diversos espacios de discusión entre los investigadores para alcanzar un consenso común.

Cuando hablamos de *fiabilidad* entendemos que cualquier investigador que utilice el sistema de categorías tendría que obtener resultados fidedignos y rigurosos. También se refiere a que el sistema pueda ser aplicado a cualquier foro de discusión. Incluso que sea aplicado por un mismo investigador en distintos momentos.

Teniendo claro lo anterior, nos disponemos a representar gráficamente los distintos Kappas hallados, en función de las diferentes reuniones y sistemas de categorías utilizados en cada subfase del proceso de validación. En la figura 1 describimos todos los Kappas, teniendo en cuenta la clasificación de la interpretación de Altman (parte superior de la gráfica), y los valores *k* alcanzados en cada reunión, dependiendo del sistema de categorías empleado en cada caso.

Observamos que en la primera validación del estudio piloto (tercera reunión), donde utilizábamos el *Sistema Corregido 2* para codificar una parte seleccionada al azar del foro 3, obtuvimos el Coeficiente Kappa 1 con un valor k=0.37. Según la clasificación de Altman, estamos ante un Kappa *débil*, que implica poco acuerdo entre codificadores, llevándonos a un concepto bajo de fiabilidad.

Esto nos condujo a continuar con un nuevo proceso de validación en el estudio piloto (segunda validación). Tomando el *Sistema corregido 3* codificamos el foro 5 (cuarta reunión), alcanzando un valor k=0.46 (Coeficiente Kappa 2). Esto se traduce en un Kappa *moderado*, que aunque supera al valor anterior, no logra un acuerdo satisfactorio para asegurar la fiabilidad.

Pasamos a una tercera validación en el estudio piloto. Fue codificado el foro 8 empleando el *Sistema corregido 4*. El valor del Coeficiente Kappa 3 fue de k=0.63, que nos llevaba a un nivel *bueno* en la clasificación de Altman. Esto proporcionó una confianza mayor en el acuerdo entre codificadores al haber logrado un Kappa aceptable. Por ende, se tomó la decisión de finalizar el estudio piloto y pasar a la codificación completa de todos los foros con el sistema de categorías resultante (hechas las oportunas modificaciones).

Por último, en la sexta y última reunión de codificación, pusimos en común todas las codificaciones de la muestra completa de foros. Se realizaron con el *Sistema corregido 5* (definitivo). Alcanzamos el nivel *bueno* en la clasificación de Altman, con un valor k=0.77 (Coeficiente Kappa 4). Podemos concluir que logramos un acuerdo alto y fiable entre los tres codificadores.

8. CONCLUSIONES

En este artículo hemos contribuido a clarificar el modo de cómo realizar el cálculo de la fiabilidad y la concordancia entre codificadores en estudios donde se toma el análisis de contenido y/o análisis del discurso como técnicas principales de investigación. Los estudios revisados nos indican que el cálculo de la fiabilidad se ha venido obteniendo a partir del acuerdo entre dos codificadores, utilizando para tal fin el índice Kappa de Cohen. Es por ello que en este trabajo hayamos querido mostrar en detalle el procedimiento para el cálculo de la fiabilidad en aquellos casos en los que el número de codificadores que intervienen es mayor que dos.

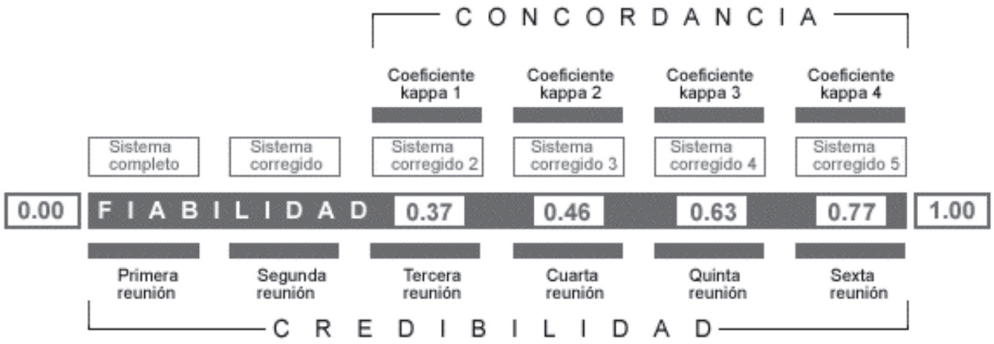

Figura 2
Fiabilidad, concordancia y credibilidad de la investigación.

Tomando la valoración de nuestros coeficientes kappa, mostramos la figura 2 que simplifica la relación entre los conceptos de fiabilidad, concordancia y credibilidad de nuestra investigación.

La *concordancia*, que mide el acuerdo entre los investigadores, llega por los distintos coeficientes Kappa que hemos hallado (cuatro en total). El valor ascendente que hemos ido consiguiendo, pasando de un nivel *débil* (k=0.37) a uno *bueno* (k=0.77), nos ofrece además una *fiabilidad* alta del sistema de categorías. Esto nos asegura que cualquier investigador puede alcanzar resultados semejantes al aplicarlo a otros foros online en los que tengan lugar procesos de enseñanza-aprendizaje mediante e-Learning, así como en diferentes momentos (evitando el deterioro de la fiabilidad). Por su parte, la *credibilidad* nos viene proporcionada por las continuas revisiones y reuniones mantenidas durante el proceso de construcción del sistema de categorías. De este modo, se garantiza que las interpretaciones puedan ajustarse a la realidad estudiada.

9. REFERENCIAS BIBLIOGRÁFICAS

ALTMAN, D.G. (1991). Practical statistics for medical research. New York: Chapman and Hall.

ARCHER, W. et al. (2001). A framework for analysing critical thinking in computer conferences. Paper presented at EURO-CSCL Conference 2001 (21-24 marzo). Maastricht (Holanda). http://www.ll.unimaas.nl/euro-cscl/ programme.htm (25/01/2008).

BAKEMAN, R. y GOTTMAN, J.M. (1989). Observación de la interacción: introducción al análisis secuencial. Madrid: Morata.

CAPOZZOLI, M., McSWEENEY, L. & SINHA, D. (1999). Beyond kappa: A review of interrater agreement measures. The Canadian Journal of Statistics, 27(1), 3-23.

COHEN, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

DONOSO, T., FIGUERA, P. y TORRADO, M. (2000). Análisis y validación de una escala para medir la conducta exploratoria. Revista de Investigación Educativa, 18 (1), 201-220.

FLEISS, J.L. (1981). Statistical methods for rates and proportions. New York: John Wiley and Sons.

GARCIA, R. (2004). Diagnóstico de la Teleformación: construcción y validación de un escalograma Guttman. Revista de Investigación Educativa, 22 (1), 277-302.

GOETZ, J.P. y LeCOMPTE, M.D. (1988). Etnografía y diseño cualitativo en investigación cualitativa. Madrid: Morata.

JONASSEN, D. & REMIDEZ, Jr., H. (2005). Mapping alternative discourse structures onto computer conferences. International Journal Knowledge and Learning, 1 (1/2), 113-129.

MARRA, R.M.; MOORE, J.L. & KLIMCZAK, A.K. (2004). Content analysis of online discussion forums: a comparative analysis of protocols. Educational Technology Research and Development (ETR&D), 52(2), 23-40.

McMILLAN, J.H. y SCHUMACHER, S. (2005). Investigación educativa. 5ª ed. Madrid: Pearson Educación.

PATTON, M.Q. (1990). Qualitative evaluation and research methods. 2nd ed. Beverly Hills: Sage Publications.

ROURKE, L. et al. (2001a). Methodological issues in the content analysis of computer conference transcripts. International Journal of Artificial Intelligence in Education, 12, 8-22.

ROURKE, L. et al. (2001b). Assessing social presence in asynchronous text-based computer conferencing. Journal of Distance Education / Revue de l'enseignement à distance, 14 (2). http://cade.athabascau.ca/vol14.2/rourke_et_al.html (25/01/2008)

SANDIN, M.P. (2000). Criterios de validez en la investigación cualitativa: de la objetividad a la solidaridad. Revista de Investigación Educativa, 18 (1), 223-242.

TORRES, J.J. & PERERA, V.H. (2005). Studying Collaborative Learning in Online Discussion Forums. In *ICTE in Regional Development*, 118-121. Valmiera (Latvia): Vidzeme University College.

Fecha de recepción: 13 de mayo de 2008.
Fecha de aceptación: 16 de diciembre de 2008.