PATRONES DESCRIPTIVOS DE LOS DISEÑOS DE CONTRASTACIÓN EN TAREAS CLÍNICAS[1]

Rafael Moreno
José Carmona
Rafael J. Martínez
Eva Trigo
Departamento de Psicología Experimental
Facultad de Psicología
Universidad de Sevilla

1. Este trabajo formó parte del Proyecto de Investigación PB89-0626 (1991-1994) financiado por la Dirección General de Investigación Científica y Técnica (DGICYT) del Ministerio de Educación y Ciencia de España. Correspondencia: Facultad de Psicología. Avda. San Francisco Javier s/n. 41005 Sevilla. E-mail:rmoreno@psicoexp.us.es

0. INTRODUCCIÓN

En el ámbito clínico, como en cualquier otro campo en que se desarrolle la práctica psicológica, el profesional necesita diseñar con la mayor validez posible la contrastación o puesta a prueba de sus hipótesis. Así ocurre por ejemplo cuando se requiere evaluar la plausibilidad de un diagnóstico o la eficacia de un tratamiento. En casos como éstos el profesional clínico y su trabajo se ven beneficiados al utilizar los conocimientos que la metodología aporta sobre los diseños de investigación.

En los diseños de investigación científica es usual distinguir dos componentes: la obtención planificada de datos sobre las variables en estudio y el análisis de esos datos para responder a las preguntas planteadas. El análisis de datos se ha desarrollado considerablemente, existiendo actualmente una aceptable estructuración de la numerosa casuística de procedimientos y técnicas existentes; el modelo estadístico lineal generalizado (McCullagh y Nelder, 1989). A diferencia, el conocimiento sobre la obtención planificada de datos es claramente mejorable; en lugar de un modelo que sintetice lo común a los distintos diseños, se ha llegado tan sólo a una serie de clasificaciones realizadas cada una con criterios diferentes. El número y tipo de variables independientes y dependientes, el papel activo o no del investigador, el número de observaciones por tratamiento, o el aspecto temporal implicado en todas esas operaciones son algunos de los criterios normalmente utilizados. Aunque todos éstos refieren aspectos relevantes, no existe un marco conceptual que organice y justifique a todos ellos y a las clasificaciones derivadas. Considerando que los avances en el conocimiento de las tareas implicadas en el diseño aumentan sin duda las posibilidades en los campos aplicados como el clínico, nuestro interés se centra en facilitar algunos avances en la temática señalada.

Una forma de abordar dicho objetivo consistiría en entender a los diseños exclusivamente en términos de los modelos formales utilizados en el análisis de datos. Pero esta alternativa podría llevar al olvido de los otros aspectos constitutivos de los diseños. Aunque el análisis es imprescindible para todo diseño, es sólo una parte de éste. Por ello se trataría de evitar que la atención al análisis de datos supusiese una consideración insuficiente del resto de operaciones. Esa relegación sería además poco aconsejable dado que la atención que se viene prestando a esas otras tareas es ya escasa; de hecho no es fácil hallar investigaciones por ejemplo sobre técnicas no estadísticas de control, a pesar de que los conocimientos al respecto no sean tan satisfactorios como cabría desear.

Explorando vías alternativas para concebir mejor a los diseños, tratamos de buscar aspectos que por generales estén presentes en todos ellos, cualquiera que sean sus especificidades y ámbitos de aplicación. Asumimos como punto de partida que en la mayoría de los casos un diseño se realiza para evaluar la razonabilidad o validez de una relación planteada, como las que en el campo clínico se plantean entre una intervención terapéutica y el comportamiento de los sujetos, o en las tareas de diagnóstico entre los diferentes síntomas que deben covariar entre sí para identificar un trastorno. De acuerdo con ello entendemos Martínez, Trigo y Moreno, 1993; Moreno, 1988; 1994; Moreno, Trigo y Martínez, 1989; que los diseños válidos implican la serie de requisitos expresados a continuación. Es preciso en ellos: *a*) considerar cuando menos una variable independiente (VI) y una variable depen-

diente (VD) de la relación en estudio; *b*) observar o producir como mínimo dos valores de la VI; *c*) registrar los valores de la VD que ocurran ante cada valor de la VI; *d*) y estudiar la posible covariación entre los valores de ambas variables, para así defender la relación entre ellas en la medida en que al variar la VI lo haga también la VD; por el contrario la ausencia de relación se ligará a la covariación nula, es decir a que los cambios en la VD no se produzcan o se den independientemente de los cambios en la VI. En cualquier caso, las garantías concedidas a esas conclusiones están ligadas a dos requisitos adicionales. *e*) El estudio de la covariación ha de hacerse mientras se controlen lo más posible otros factores, conocidos como variables extrañas o contaminadoras, que pueden covariar con la VD a la vez que la VI; ese control viene a suponer en todos los casos que los valores de tales factores no varíen a la vez que lo hagan los de la VI, impidiendo así que se confundan posibles covariaciones coincidentes. *f*) Aun con esas precauciones, las conclusiones sobre las covariaciones VI-VD no serán tenidas en cuenta como relevantes a menos que se obtengan a partir de un número razonable de datos.

Este conjunto de requisitos constituye una abstracción que pretende identificar lo común a todo diseño correcto; una regla general que pretende describir dichos diseños y por tanto hacer posible también las prescripciones sobre ellos. Entendemos pues que ese conjunto de requisitos es expresión de los criterios usados, implícitamente casi siempre, al procurar y evaluar la validez interna de todo diseño contrastador de relaciones entre variables.

De acuerdo con esta propuesta, los diferentes diseños y conceptos a ellos ligados podrían ser considerados formas específicas de cumplimiento o no de esos requisitos. Así por ejemplo, cuando en los diseños factoriales se estudian las relaciones de interacción o de segundo orden, el esquema propuesto se cumple al considerar como VD a una relación principal. Por su parte, los términos experimental y correlacional referencian dos formas de cumplir el requisito *b*) para obtener los valores de la VI, bien sea que los produzca el investigador o que éste aproveche la ocurrencia natural de aquéllos. El estudio de la existencia, forma y magnitud de las covariaciones puede ser considerado como el cumplimiento del requisito *d*) con diferentes criterios de evaluación: respectivamente identidad/diferencia, orden y magnitud entre los datos de la VD ante diferentes valores de la VI. Las estrategias de investigación pueden ser consideradas formas diferentes de conseguir los registros necesarios para la fiabilidad, requisito *f*), en función de que esos diferentes registros sean aportados por los mismos sujetos (estrategia longitudinal) o grupos diferentes de éstos (transversal). De manera semejante cabría presentar el resto de conceptos referidos a los diseños como especificaciones de uno o más de los requisitos señalados, al igual que la falta de alguno de éstos identificará algún tipo de error; así ocurre con los llamados diseños preexperimentales y cuasiexperimentales (Campbell y Stanley, 1966; Cook y Campbell, 1979), en los que se incumplen respectivamente los requisitos *b*) y/o *e*).

En definitiva, la descripción aportada por el conjunto de requisitos señalados supone una cierta integración conceptual de los diseños, aunque por supuesto admite desarrollos adicionales. En este sentido consideramos que la descripción mencionada está aún muy ligada a las operaciones específicas que se quieren describir. Por ello debería buscarse un mayor grado de abstracción. Sería deseable entender a los diseños, y por tanto también a los requisitos recién expuestos, en términos de algunas categorías más generales. Si implica-

ran un nivel más avanzado de abstracción aportarían mayor sencillez y potencia a la descripción deseada.

En otros trabajos (Moreno, Trigo y Martínez, 1989; 1991) ya adelantamos que podría haber tres relaciones u operaciones básicas conformadoras de los diseños. En los casos más simples serían: *a*) *Asignación* (A), entendida como correspondencia entre algún valor de cada una de las variables cuya covariación se pretende validar, es decir la relación VI-VD en el caso más simple; *b*) *Agrupación* (G) o correspondencia entre datos de la VD obtenidos ante un mismo valor del la VI, VD_{VIi}- VD_{VIi}; y *c*) *Comparación* (C) o relación entre los datos de la VD obtenidos ante valores distintos de la VI, es decir VD_{VIi}- VD_{VIj}, siendo $i \neq j$. Las tres clases de correspondencias contienen pues los mismos elementos, aunque con funciones diferentes en cada caso. La correspondencia de Asignación (A) se da entre valores de la VI y de la VD, la definitoria de Agrupación (G) se establece entre al menos dos datos de la VD ante un mismo valor de la VI, y la de Comparación (C) entre datos de la VD ante al menos dos valores distintos de la VI.

Estos conceptos deberían ser identificables a nivel empírico de forma fiable, para lo cual son necesarios criterios o índices adecuados. Con ellos sería conveniente poder especificar tanto los casos en los que se logren correctamente las correspondencias arriba definidas, como aquellos otros en los que ellas no ocurran. A tal fin se puede constatar que los requisitos contenidos en las tres correspondencias propuestas son los siguientes: *a*) existencia de al menos una VI y una VD, lo que es necesario para la asignación; *b*) existencia de al menos dos datos, diferentes o iguales, de la VD ante cada valor de la VI, lo que es imprescindible para la agrupación; *c*) existencia de al menos dos valores de la VI, lo cual es necesario para la categoría de comparación; y *d*) que los posibles cambios de la VI no se solapen con los cambios de cualquier otra variable distinta a la VD, para que así se cumplan sin confusión las condiciones definitorias de cada tipo de correspondencia.

De forma paralela a estos requisitos surge la siguiente serie de errores posibles. *a*) El error que denominaremos de tipo 0 consiste en la falta de alguno de los elementos, VI y/o VD. Ello puede ocurrir en dos formas: por omisión de uno de los dos términos, o por aparecer como variable extraña alguna de las que el problema u objeto de estudio marque como VI o VD. *b*) El error que denominaremos 1 supone la presencia tan sólo de un dato de la VD ante cada valor de la VI. *c*) El error 2 consiste en la existencia de tan sólo un valor de la VI. Y *d*) el error 3 implica la coincidencia o confusión de cambios en los valores de la VI con los de al menos otra variable diferente a la VD.

Considerando la serie de criterios identificados y los correspondientes errores, hemos planteado el siguiente sistema de categorías referidas a correspondencias válidas:

a) Como se ha indicado, en general se puede entender por *asignación* la correspondencia entre algún valor de cada una de las variables, VI y VD. Sin embargo, en el contexto específico de los diseños contrastadores la asignación será considerada *válida* (Av) cuando incluya: las dos variables señaladas, al menos dos registros o datos de la VD ante cada valor de la VI, al menos dos valores diferentes de la VI, y que ninguna otra variable varíe a la vez que lo haga la VI.

b) La *agrupación* puede entenderse como correspondencia entre valores de la VD obtenidos ante un mismo valor de la VI. En el contexto de los diseños la agrupación será consi-

derada *válida* (Gv) cuando incluya: ambas variables, al menos dos registros de la VD ante cada valor de la VI, al menos dos valores diferentes de la VI, y que ninguna otra variable varíe a la vez que lo haga la VI.

c) Por su parte, siendo la *comparación* una correspondencia entre los valores de la VD obtenidos ante valores diferentes de la VI, en el contexto de los diseños será *válida* (Cv) cuando incluya ambas variables, al menos dos registros de la VD ante cada valor de la VI, al menos dos valores diferentes de la VI, y que ninguna otra variable varíe a la vez que lo haga la VI.

A las anteriores categorías hay que añadir las siguientes referidas a errores:

d) Asignación con error tipo 0 (A0) si en la correspondencia indicada falta la VI y/o la VD. Cuando ello ocurre no son posibles las categorías G ni C, y en términos estrictos ni tan siquiera la A. A pesar de ello identificamos este caso como error de A para señalar que el tipo 0 implica no lograr siquiera la categoría más básica del sistema.

e) Asignación y comparación con error 1 (Al y C1) si las relaciones señaladas se realizan con un solo dato de la VD ante cada valor de la VI. Las definiciones de agrupación y del error 1 hacen imposible la categoría G1.

f) Asignación y agrupación con error 2 (A2 y G2) si estas relaciones se realizan con un solo valor de la VI. Por definición la categoría C2 no existe.

g) Cualquiera de las tres relaciones con error 3 (A3, G3 y C3) cuando estas relaciones se establecen mientras que al menos otra variable distinta a la VI varía a la vez que ésta.

h) Asignación con errores 1 y 2 (A12) si aparecen los dos errores señalados; no se plantean G12 ni C12 al no ser posibles G1 ni C2.

i) Asignación y comparación con errores 1 y 3 (A13 y C13) si aparecen los dos errores señalados; no se plantea G13 por no ser posible G1.

Por otra parte, ni las categorías 23 ni por tanto tampoco las de tipo 123 son posibles con ninguna de las tres relaciones, porque el error 3 no puede darse junto con el 2: no es posible que las variables extrañas queden desigualmente repartidas en distintas condiciones de la VI cuando se utiliza un solo valor de ésta.

Como puede comprobarse el sistema propuesto presenta un número elevado de categorías. Aunque son tres las correspondencias básicas, la especificación de sus posibles errores eleva la cantidad de manera significativa. Como ello difiere de lo aconsejado normalmente conviene hacer el siguiente comentario. La usual sugerencia de limitar la cantidad de categorías es simplemente resultado de la constatación empírica de que ocurrirán menos problemas de registro mientras menos categorías haya que considerar. Sin embargo, entendemos que tal reducción debería ir siempre ligada a la relevancia del sistema de categorías que se quiera utilizar. Dicha reducción será conveniente sólo si el objeto de estudio lo permite. Entendemos que los problemas metodológicos no deben resolverse empobreciendo el objeto de estudio, sino adaptando la metodología a lo que se desee estudiar. Además, el número de categorías resulta relevante en función de la organización conceptual existente entre ellas. Será más complicado estudiar un sistema construido a partir de diversos criterios poco estructurados que otro, como el aquí planteado, en el que el elevado número de categorías es el resultado de especificar un corto número de criterios sencillos. En consecuencia, vamos a tratar de comprobar que es posible utilizar un sistema con categorías numerosas pero construidas de manera organizada.

Otra propiedad del sistema propuesto es su carácter progresivamente inclusivo o jerárquico, tanto en las categorías válidas por una parte como en las de errores por otra. En las válidas cada correspondencia es necesaria para la siguiente, consideradas en el orden A, G y C. De acuerdo con sus definiciones, la agrupación válida sólo es posible si previamente ha habido más de una asignación de valores de la VD ante cada valor de la VI. Igualmente, la comparación válida puede ser establecida sólo si previamente se han agrupado valores de la VD y al menos ante dos valores de la VI. En las categorías de errores se da una jerarquía similar. El carácter jerárquico del sistema implica que cuando se plantee una determinada categoría que necesita de una o más previas, habrá que considerar a éstas también como logradas aunque no hayan sido explicitadas independientemente. Así por ejemplo, si se plantea con claridad Cv sin hacer mención a pasos previos como Av y Gv, éstos han de anotarse también, ya que la definición de Cv así lo requiere. Para identificar estos casos se ha completado el sistema con la serie de categorías AG, GC, y AGC, las cuales pueden ser acompañadas de la especificación de validez o de error que en cada caso corresponda.

El carácter jerárquico señalado no es obstáculo sin embargo para defender la independencia de cada categoría, pues en otro caso sería artificial la distinción entre ellas. Por independencia entendemos que la identificación de cualquiera de ellas no conlleva inevitablemente la de alguna otra posterior. También implica entender que no existe una única secuencia posible en la aparición de las categorías. Por ejemplo, dos categorías de tipo válido conceptualmente inclusivas pueden verse intercaladas en la práctica por otras distintas, como por ejemplo las de error. En el mismo sentido, una categoría de error como Cl puede ser precedida por Al o por otras categorías como Av, Gv. Los datos de este estudio y otros posteriores revelarán si tales posibilidades teóricas ocurren en el terreno empírico.

En cualquier caso, y contando con que el sistema de categorías propuesto podría ser útil para un análisis detallado de la tarea de diseñar contrastaciones en ámbitos aplicados como el clínico, el objetivo prioritario de este trabajo es estudiar la validez de dicho sistema de categorías, lo cual implica al menos la siguiente serie de objetivos específicos.

En primer lugar deberá comprobarse si efectivamente es posible identificar las categorías propuestas en la práctica de los diseños de investigación clínica. Es claro que dicho sistema deberá permitir identificar cada una de las operaciones desarrolladas por quienes plantean los diseños, describiendo a cada una con categorías específicas. Ello supone evaluar las propiedades de claridad definitoria, exclusividad y exhaustividad del sistema de categorías (Anguera, 1981). En otras palabras, con tales categorías debería ser posible describir el conjunto de operaciones que los diseños implican. En este estudio sin embargo hemos tenido que excluir las constitutivas del análisis formal-estadístico, al ser de tipo cualitativo no probabilístico el realizado por nuestros sujetos.

Con el sistema de categorías propuesto pretendemos pues identificar las diferentes formas posibles, correctas o no, en que nuestros sujetos realizan las tareas de diseñar la contrastación de hipótesis. Aunque con carácter meramente exploratorio, podremos hacer un primer sondeo sobre frecuencias relativas de cada una de las categorías, de los errores derivados, y de otros aspectos tales como las estrategias transversales o longitudinales de investigación, las técnicas de control y los tipos generales de diseños planteados. Todo ello podría apuntar ciertas tendencias mayoritarias en la realización de los diseños. Asimismo ex-

ploraremos si estos aspectos cambian en sujetos de diferentes niveles de ingenuidad en tareas de contrastación. Ello puede sernos de ayuda a la hora de planificar en otros estudios un adecuado entrenamiento para el diseño de contrastaciones en ámbitos aplicados como el clínico, donde la práctica profesional puede condicionar fuertemente el modo de realizar dichas contrastaciones. Esta aproximación será completada con otra de carácter diacrónico, explorando posibles secuencias sistemáticas en que puedan presentarse las categorías.

1. METODO

1.1. Sujetos

En el estudio participaron 16 sujetos. Cinco de ellos eran estudiantes de 5º curso de Psicología de la Universidad de Sevilla y cuatro cursaban 5º de Filosofía, otros cuatro eran estudiantes de COU de Ciencias en el Instituto Nacional de Bachillerato Velázquez de Sevilla y tres cursaban COU de Letras en ese mismo Instituto. A diferencia de los tres últimos grupos, los sujetos de 5º de Psicología contaban en su curriculum con diversas asignaturas del Area de Metodología de las Ciencias del Comportamiento sobre diseño y contrastación de relaciones entre variables. Todos los sujetos se prestaron voluntariamente a participar en la investigación.

1.2. Materiales y aparatos

Se elaboraron 140 tarjetas de cartulina plastificada de 10x15 cm. Las tarjetas estaban divididas en 7 grupos de 20 tarjetas cada uno según el color de las mismas: naranja, blanco, gris, negro, salmón, verde y azul. Grupos de cinco tarjetas de los colores negro, gris y blanco tenían adheridos uno, dos, tres o ningún círculo de color naranja de 2 cm. de diámetro en una cara.

Las instrucciones sobre la tarea se entregaban por escrito en folios estándar de papel blanco. La ejecución de los sujetos fue recogida por una cámara de vídeo Sony F-550 de 8 mm. con trípode. La cámara estaba situada frente a la mesa en la que trabajaban los sujetos, donde se encontraban previamente dispuestas las tarjetas correspondientes. El investigador se situaba al lado de la cámara para controlar su correcto funcionamiento. Las sesiones de observación se realizaron en los centros de enseñanza de cada grupo de sujetos. Para la codificación se utilizó un magnetoscopio VHS Panasonic NV-F75 HQ y un monitor Philips 14GR1021.

1.3. Procedimiento

Se diseñó un estudio de observación interna (Anguera, 1989; Jorgensen, 1989) para registrar las actividades de los sujetos enfrentados a una tarea consistente en explicar qué

harían para evaluar la razonabilidad de una determinada relación entre variables. Pretendíamos así simular la tarea de diseñar la contrastación de una hipótesis, al igual que en otros estudios sobre estrategias de prueba de hipótesis, tareas de diagnóstico, juicios de covariación o descubrimiento de reglas (por ejemplo Allan, 1993; Klayman y Ha, 1987; Oaksford y Chater, 1994; Wasserman, Elek, Chatlosh y Baker, 1993; Young, 1995). En nuestro caso, para simplificar el contenido de la relación a contrastar optamos por plantear un tratamiento simple como la ingestión de café y un comportamiento psicológico convencional como los resultados en un test.

Cada sujeto fue enfrentado al mismo problema en dos ocasiones con una semana de intervalo. En la primera situación no se impusieron restricciones sobre el tipo de diseño a plantear, optando la mayoría de los sujetos por una estrategia transversal. Por ello en la segunda se les pidió que realizasen el estudio con un único sujeto, obligándolos así a adoptar una estrategia longitudinal de amplia utilización en la investigación clínica aplicada, obteniendo con ello una muestra más representativa de las situaciones de diseño.

Tras una breve charla informal de presentación del investigador y una vez familiarizado cada sujeto con la situación y el material, se proporcionaban las instrucciones en los siguientes términos: "Se trata de que lleves a cabo un estudio para poder llegar a confiar lo más posible en si el hecho de que un sujeto obtenga un resultado óptimo en el test depende de que tome tres tazas de café inmediatamente antes de la prueba".

En la segunda tarea las instrucciones eran las siguientes: "En esta ocasión se trata de que lleves a cabo el mismo estudio, es decir, un estudio para poder llegar a confiar lo más posible en si el hecho de que un sujeto obtenga un resultado óptimo en el test depende de que tome tres tazas de café inmediatamente antes de la prueba, pero utilizando para ello un único sujeto".

Excepto en estas instrucciones, en todo lo demás se procedió de igual forma en las dos situaciones. En ambas se indicaba a los sujetos que debían utilizar las tarjetas para representar los elementos del problema, y se proporcionaban las correspondencias de los colores de las tarjetas con dichos elementos (tabla 1).

Teniendo en cuenta dichas equivalencias, cada sujeto debía ir representando en la mesa las tareas necesarias para cumplir con el objetivo señalado. A la vez se le invitaba a que comentara verbalmente lo que iba realizando. Finalmente, se le pedía que una vez cumplido el objetivo propuesto avisara al investigador presente.

Ya que no en todos los casos el sujeto hacía uso de las tarjetas de acuerdo con las correspondencias indicadas, el investigador intervenía para aclarar el uso que les estaba dando: "¿Qué significa esta tarjeta?", "¿Por qué utilizas una tarjeta de este color?". Cuando el sujeto tenía dudas sobre cómo debía realizar la tarea e interrogaba al investigador, éste ofrecía respuestas del tipo "Tal como tú creas", "Hazlo a tu manera", "Lo que tú creas más conveniente", o bien se limitaba a aclarar las correspondencias entre las tarjetas y los elementos del problema. Por lo demás, el sujeto tenía libertad para organizar la tarea sin otras indicaciones adicionales.

Una vez que el sujeto avisaba de la finalización de la tarea o daba síntomas de ello, como cruzarse de brazos o parar de hablar y mirar al investigador, éste le interrogaba del siguiente modo: "¿Has terminado?". Y en caso de respuesta afirmativa añadía: "Con esto, ¿tú

Tabla 1. *Correspondencia entre las tarjetas y los elementos del problema a contrastar*

VARIABLE	*VALORES*	*REPRESENTACIÓN*
Sujetos	N	color naranja
V1: nº de cafés	0	0 círculos
	1	1 círculo
	2	2 círculos
	3	3 círculos
VD: resultados del test	buenos	color verde
	intermedios	color salmón
	malos	color azul
VE: hora del día	mañana	color blanco
	tarde	color gris
	noche	color negro

llegarías a confiar lo más posible en que los resultados óptimos dependen de tomar tres tazas de café inmediatamente antes de la prueba?". En caso de una respuesta negativa se volvía a enfrentar al sujeto con la tarea. Cuando respondía afirmativamente se le preguntaba además en qué se había fijado para concluir sobre la relación entre las variables. Con esta última pregunta complementaria a la observación realizada se trataba de obtener una mayor especificación de las operaciones de diseño realizadas. Tanto las instrucciones como la participación del observador pretendían que los sujetos informaran con claridad del diseño que utilizaban. No obstante, para preservar la validez interna del estudio, se evitaba interferir en las operaciones concretas que cada sujeto decidía llevar a cabo.

Respecto a las instrucciones cabe señalar otras cuestiones relevantes. En primer lugar sólo explicitaban un valor de la VI y de la VD, respectivamente tres tazas de café y un resultado óptimo en el test. Se intentaba así no sugerir el uso de más de un valor de la variable independiente, dando la posibilidad de encontrar sujetos que se limitasen a trabajar con el valor mencionado (lo que daría lugar a categorías de error tipo 2). En segundo lugar se utilizó intencionadamente el término "depende", tratando de delimitar la tarea como el estudio de una covariación causal entre variables en vez de una descripción no relacional. Por tanto no sería suficiente comprobar que con un valor de la VI (tres cafés) los valores de la VD no eran los esperados (resultados óptimos), ya que dichos resultados podrían obtenerse también con un valor distinto de la VI. En tercer lugar la tabla de equivalencias que se le mostraba al sujeto sólo explicitaba una variable extraña a controlar, la hora del día, con tres valores expresados mediante el color blanco, gris o negro de las tarjetas. En consecuencia sólo tendríamos en cuenta dicha variable para evaluar el cumplimiento del requisito de control. A pesar de que ello nos llevaría a una evaluación poco exigente de los diseños planteados por nuestros sujetos, optamos por tal restricción por varias razones. Por una parte y al ser la

mayoría de los sujetos ingenuos en la tarea, una situación más compleja en cuanto a las necesidades de control podría haber obstaculizado considerablemente sus ejecuciones. Por otra parte, con nuestra opción se contribuía a facilitar el proceso de codificación, lo cual es adecuado al objetivo exploratorio del presente trabajo.

La codificación de lo realizado por los sujetos fue llevada a cabo por dos parejas independientes de observadores. Con ello se intentó minimizar los errores, de forma que los cometidos por un investigador pudieran verse contrarrestados por su compañero. En consecuencia, el cálculo de la fiabilidad se realizó interparejas, y una vez que los miembros de cada una decían estar de acuerdo.

Las actuaciones de los sujetos se recogieron mediante un registro de eventos, aunque también se consideró importante conocer el momento de ocurrencia de los mismos (Altman, 1974; Mehm y Knutson, 1987). Para ello se dividió arbitrariamente cada registro de observación en intervalos regulares de 30 segundos. De esa manera, los eventos de interés activaban el registro, mientras que los intervalos permitían anotar el momento en el que era posible identificar el evento.

Para la categorización se utilizaron como índices tanto las expresiones verbales de los sujetos como las actividades manipulativas que realizaban con las tarjetas. Los índices verbales eran imprescindibles para las categorizaciones, y podían utilizarse aisladamente si eran suficientemente significativos. No obstante algunos de tales índices quedaron excluidos por inespecíficos, como por ejemplo ocurría con la terminología técnica metodológica si no iba acompañada de otros elementos que demostrasen su uso adecuado. Así por ejemplo, nombrar simplemente una técnica de control no podía considerase un índice preciso. También se consideraba inespecífica la simple mención de los términos que aparecían en las instrucciones, como por ejemplo "resultados óptimos", "resultados intermedios" y "depende", por cuanto podían ser meras repeticiones y no índices significativos de lo realizado por el sujeto. En el anexo se muestran algunos índices verbales utilizados en la codificación. A diferencia de los índices verbales, los manipulativos no eran tenidos en cuenta si aparecían en solitario, ya que su significado podía ser ambiguo.

El proceso de categorización se realizó de acuerdo a una serie de pautas fijadas en las fases exploratorias del estudio. El proceso comenzaba con el visionado por parte de cada pareja de observadores de la grabación completa del sujeto y con la lectura de la transcripción íntegra de sus descripciones verbales. Con ello los observadores podían hacerse una idea de la ejecución global del sujeto, intentando evitar así que por conocer sólo logros parciales pudieran anticipar distintos resultados finales, sesgando con ello la codificación. En un segundo momento se identificaba cada uno de los elementos del problema utilizados por los sujetos en la contrastación: ocasiones de estudio, resultados en la prueba (VD), tazas de café (VI) y unión de éstas últimas a la variable hora del día. Este segundo estadio era fundamental para decidir si se estaba ante categorías de tipo 0, lo que de no suceder daría la posibilidad de los tipos 1, 2, 3 o válidas. La tercera etapa suponía determinar las diferentes asignaciones (A), agrupaciones (G) y/o comparaciones (C) planteadas por los sujetos con los elementos utilizados en cada caso. Por fin, el cuarto paso consistía en identificar qué requisitos se cumplían y cuáles no para cada una de las categorías de asignación, agrupación y comparación identificadas previamente, lo que suponía caracterizarlas como válidas (v) o con errores (1, 2 y/o 3).

El logro de los cuatro estadios de categorización hasta aquí descritos debía suponer la especificación de la categoría en cuestión. Sin embargo, previamente había que resolver un problema básico en cualquier tipo de registro: cómo decidir el corte del episodio global en diferentes unidades significativas de análisis. Esta partición no debía basarse ni en los cortes artificiales producidos por las intervenciones del investigador ni en las pausas que realizaba el propio sujeto, ya que éste podía continuar haciendo referencia a los mismos elementos y operaciones. Los cortes tampoco debían estar marcados por los intervalos previamente determinados de 30 segundos, ya que no se respetarían así las unidades naturales de conducta, principal característica de los registros de eventos (Irwin y Bushnell, 1980; Wright, 1960).

A diferencia, parecía necesario plantear una regla más significativa. Para recoger exhaustivamente lo planteado por los sujetos se dividiría el episodio global categorizando diferentes operaciones parciales. Por ejemplo, en caso de ocurrir por separado las diferentes categorías válidas que componen el sistema jerárquico (Av, Gv y Cv) no se consideraría suficiente categorizar sólo la última de ellas. Sería necesario cortar el episodio global para reflejar el logro de cada categoría sucesivamente, y diferenciarlo así del logro simultáneo de dos o más de ellas (AG, GC o AGC). De igual forma, si un mismo logro ocurriera más de una vez pero con diferentes elementos, sería categorizado en ambas ocasiones. Este sería el caso de dos asignaciones válidas cuando para una se eligen diferentes sujetos como ocasiones de estudio (estrategia transversal) y para la otra diferentes momentos temporales (estrategia longitudinal); o cuando en una se utiliza un sólo valor para la variable interviniente hora del día (mantenimiento constante) y varios de éstos en la otra (balanceo). Categorizando ambas operaciones por separado es posible recoger la diversidad de tareas realizadas, diferenciando este caso de aquellos en los que se llega a tal logro de una única forma. De esta forma quedan recogidos tanto los cambios considerados relevantes desde el propio sistema de categorías (de asignación a agrupación y a comparación), como los considerados relevantes desde un punto de vista metodológico más tradicional.

En otras ocasiones, por el contrario, no resultaría necesario dividir el episodio global. No se categorizarían los logros erróneos que en conjunción con otros terminasen por conformar categorías válidas, ya que lo planteado por los sujetos podría ser recogido exhaustivamente por el logro válido final. Por ejemplo, no se anotarían como dos errores de tipo 2 el asignar valores de la VD primero a un valor de la VI y después al otro, ya que se trataría en conjunto de la categoría Av (si se añadían además el resto de requisitos). No obstante, si una vez concluido el episodio completo no fuese posible identificar la categoría válida, se codificarían las incorrecciones cometidas.

Por último, conviene señalar que para especificar el proceso de categorización y codificación descrito fue necesario un trasvase constante entre los planos teórico y empírico. Las consideraciones conceptuales sobre lo que constituía un corte o un tipo de error significativo fueron evaluadas a la luz de las dificultades y soluciones que aportaban en el terreno de la fiabilidad; tuvimos en cuenta para ello las pruebas pilotos que realizamos con varios sujetos no incluidos en la muestra de este estudio. Al mismo tiempo la necesidad de dichos planteamientos conceptuales surgió precisamente de las dificultades en el trabajo de categorización. El entrenamiento de los observadores tuvo lugar, pues, mediante su participación en este proceso de toma de decisiones para definir con precisión el procedimiento de codificación (Hay, Nelson, y Hay, 1980, Martin y Bateson, 1991).

2. RESULTADOS

Conforme a la sugerencia de Harrop, Foulkes y Daniels (1989) comenzamos realizando un análisis visual de la matriz de acuerdos/desacuerdos (tabla 2). Definiendo acuerdo como aquella ocasión en la que coincidía la categoría identificada por las dos parejas de observadores en el mismo intervalo, la mayor parte de las categorizaciones, 76 de un total de 94, situadas en la diagonal principal, correspondían a acuerdos en la codificación.

Tabla 2. *Matriz de acuerdos/desacuerdos*

		PAREJA 1																
		Av	Gv	Cv	AGv	GCv	AGCv	A0	A1	A2	A13	G2	G3	AG2	C1	C13	AC1	∅	
PAREJA 2	Av	17																2	19
	Gv		6															1	7
	CV			6														1	7
	AGv				1		1												2
	GCv					7	1												8
	AGCv						12											1	13
	A0																	1	1
	A1								5										5
	A2									3								1	4
	A13										2								2
	G2											7						1	8
	G3												1						1
	AG2																	1	1
	C1														5				5
	C13															1			1
	AC1																2	1	3
	∅	1		1		1		1				1		1					6
		18	6	7	1	8	14		5	3	2	8	1	1	5	1	2	10	93

Los desacuerdos son mayoritariamente omisiones realizadas por alguna de las parejas: en diez ocasiones sólo codificó la pareja 2 y en seis ocasiones sólo lo hizo la 1 (ver categoría en la tabla 2). Tan sólo dos veces ambas parejas codificaron conjuntamente pero de forma diferente: la pareja 1 anotó dos AGCv mientras que la pareja 2 codificó un AGv y un GCv. Por último, con las categorías A0 y AG2 ocurre que en las dos ocasiones en que fueron codificadas produjeron desacuerdos por omisión de una de las parejas.

Para el análisis cuantitativo del grado de acuerdo entre observadores no se computaron los acuerdos producidos en intervalos de no ocurrencia, ya que habrían aumentado artificialmente el valor de los índices de acuerdo. En primer lugar se calculó el porcentaje de acuerdo según la expresión usual:

$$P = [Acuerdos/(Acuerdos + Desacuerdos)] \times 100 = 80.85\ \%$$

Sin embargo, en aras de una confirmación más exigente calculamos algunos de los índices para casos nominales que incluyen correcciones con base en el acuerdo esperado por azar: *Y* de Scott y *K* de Cohen. Estos índices comparten una misma expresión general:

$$I = (P_o - P_e)/(1 - P_e)$$

donde P_o es la proporción de acuerdos observados y P_e es la proporción de acuerdos esperados por azar. La diferencia entre los dos índices usados reside en la forma de calcular la proporción de acuerdos esperados por azar. Mientras ambos índices asumen que la distribución de las proporciones de las categorías para la población es conocida, el *Y* de Scott las supone idénticas para ambas parejas de observadores y el *K* acepta que dicha distribución de proporciones sea distinta (Cohen, 1960). Tanto el valor de *Y* (0.7869 para una P_e=0.1016) como el valor de *K* (0.7870 para una P_e=0.1010) superan el 0.75 considerado como valor mínimo de acuerdo más exigente (Harrop, Foulkes y Daniels, 1989; Suen y Lee, 1985). A pesar de ello sólo tomamos en consideración las codificaciones en las que hubo acuerdo.

Las tres categorías generales (A, G y C) aparecieron de forma independiente: en ocho ocasiones la categoría A no fue seguida por la categoría G, y en otras cinco ocasiones G

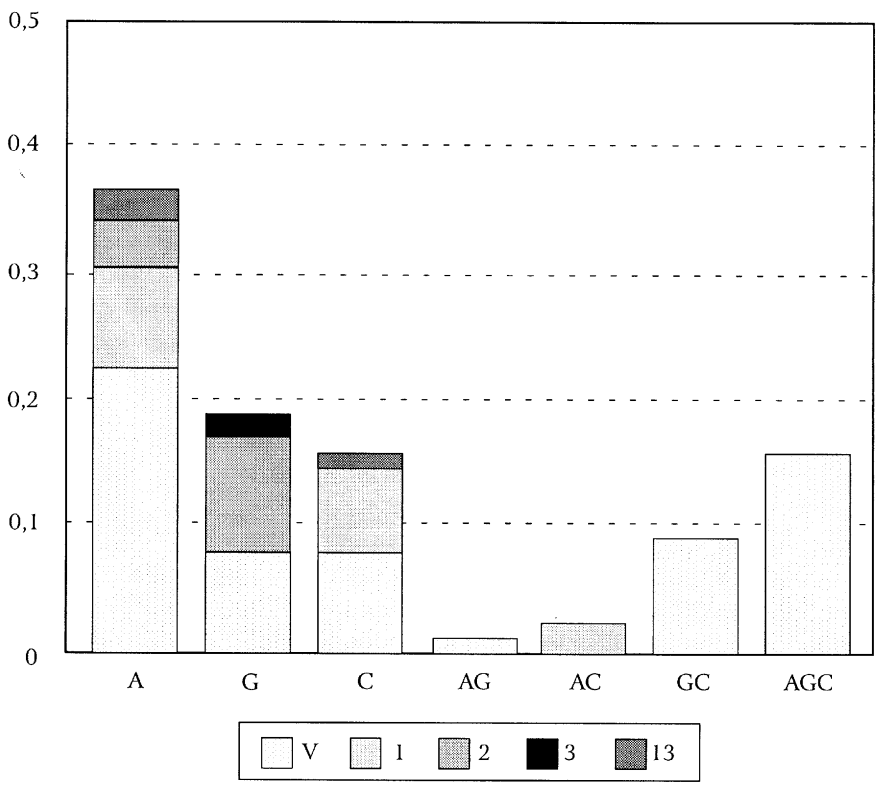

Figura 1. Proporción de categorías codificadas en función del tipo de logro (A, G y/o C) y el tipo de error (0, 1, 2 ó 3).

no fue seguida por C. La categoría más frecuente es Av con una proporción de 0.224, seguida de AGCv con 0.158 (fig. 1). A continuación se encuentran las categorías GCv y G2, con una proporción ambas de 0.092. Las siguientes son las categorías Gv, Cv y Al con una proporción de 0.079; C1 con una proporción de 0.066; A2 con 0.039; AC1 y A13 con 0.026; y por último aparecen una serie de categorías con la mínima frecuencia relativa posible (0.013), AGv, G3 y C13.

Hay que señalar además la ausencia de determinadas categorías o combinaciones de ellas que eran posibles en principio: con el error 2 falta la categoría AG; con el error 3 faltan las categorías A, C, AG, AC, GC y AGC; con el error 12 la categoría A; con el error 13 la categoría AC; y la categoría A0.

Agrupando las categorías según sean válidas o no y según su tipo de error, son más frecuentes las categorías válidas tanto como categorías aisladas (A, G y C) como en combinación (AG, GC y AGC), sumando una proporción de 0.645. Respecto a los errores, los más frecuentes son los de tipo 1 y los de tipo 2. El error tipo 1 se presenta en una proporción de 0.171, con las categorías A,C y AC. El error tipo 2 aparece como A o como G, reuniendo una proporción de 0.132 del total de categorías. A continuación aparece el error 13 (0.039) bien sea como A o como C; y por último el error tipo 3 que se registró como G con una proporción de 0.013.

Los resultados muestran que, de los 31 diseños planteados por los 16 sujetos, en las 19 ocasiones codificadas con las categorías válidas se realizó algún diseño correcto según la descripción dada en la literatura (Arnau, 1981-84; Barlow y Hersen, 1988). Diez casos, que fueron anotados con categorías incorrectas y ausencia de las válidas, correspondieron con diseños de tipo preexperimental (Campbell y Stanley, 1966), y otros dos con situaciones denominadas genéricamente de confusión de variables.

En la primera sesión los datos muestran un predominio de la estrategia transversal (0.687) frente a la longitudinal (0.313), tal y como puede observarse en la figura 2. En la segunda sesión todos los sujetos utilizaron la estrategia longitudinal, debido a nuestra petición de que la investigación se realizara con un solo sujeto. Teniendo en cuenta ambas sesiones, la técnica de control más usada de la variable extraña hora del día fue (fig. 3) el balanceo (0.709), seguida por la constancia (0.225), y sólo en dos ocasiones hubo confusión de VVEE (0.066). Por el número de condiciones, los diseños fueron mayoritariamente multicondicionales (0.483), seguidos por los bicondicionales (0.387) y por los que contenían un solo valor de la variable independiente (0.13), es decir por diseños de tipo preexperimental (fig. 4).

Se realizó también un análisis de los logros válidos para cada sesión en función de los grupos de referencia, 5º de Psicología, 5º de Filosofía, COU de Letras y COU de Ciencias (fig. 5 y 6). La consecución de tales logros no parece depender del conocimiento sobre las tareas de diseño, ya que al comparar la ejecución de los alumnos que tenían esa experiencia, Psicología, frente a los restantes que carecían de ella, no se encuentran diferencias significativas para $\alpha = .01$ ($X^2 = 2.92$; $p > .05$). De igual forma, los estudios de Ciencias, Psicología y COU-Ciencias, no parecen relacionarse con logros válidos en mayor medida que los estudios denominados de Letras, Filosofía y COU-Letras ($X^2 = 5.23$; $p = .02$). Los universitarios estudiados tampoco consiguen una mayor proporción de diseños válidos frente a los alum-

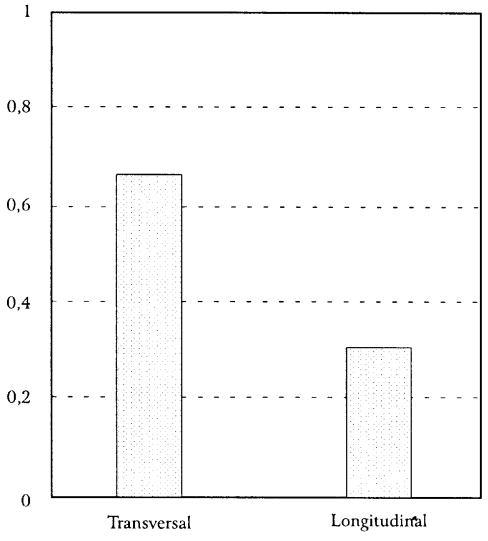

Figura 2. Proporción de estrategias de investigación usadas en la primera sesión.

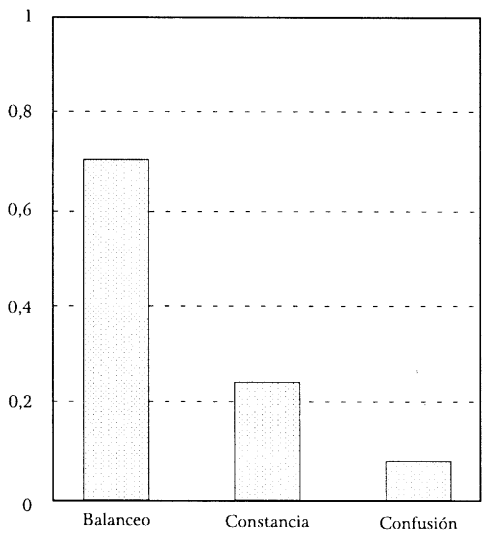

Figura 3. Proporción de técnicas de control utilizadas en ambas sesiones.

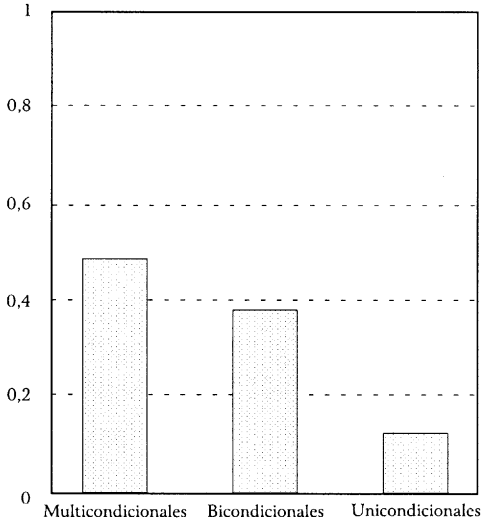

Figura 4. Proporción de diseños según el número de condiciones planteados en ambas sesiones.

nos de bachillerato (X^2=1.30; p=.25). Finalmente la familiaridad con la tarea que se les proponía no dio lugar a una ejecución más válida, al no encontrarse diferencias significativas entre la primera y la segunda sesión (McNemar, X^2=0 .20; p=.99), si bien este resultado podría deberse a una mayor dificultad de los diseños de caso único, que de hecho no fueron elegidos mayoritariamente en la primera sesión.

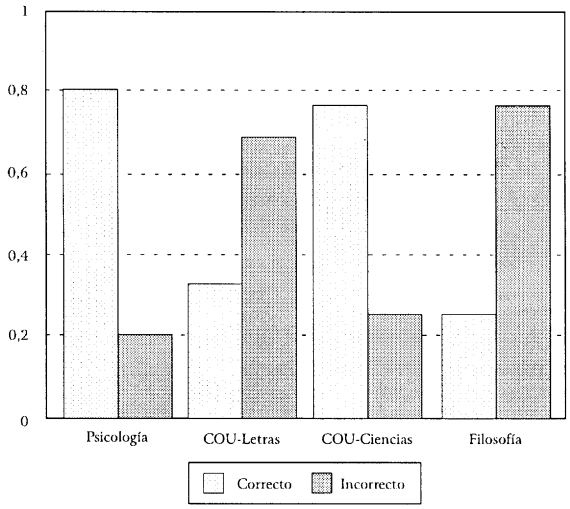

Figura 5. Proporciones de diseños correctos e incorrectos de cada grupo en la primera sesión

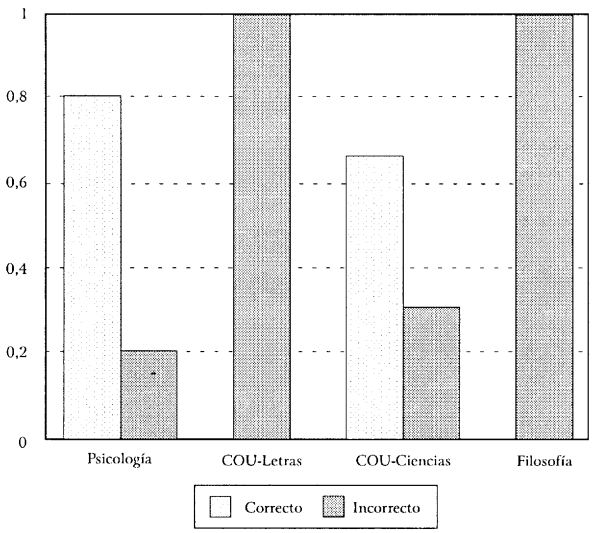

Figura 6. Proporciones de diseños correctos e incorrectos de cada grupo en la segunda sesión

Explorando además las posibles secuencias entre categorías, se ha realizado un análisis mediante el cálculo de la Z propuesto por Sackett (1979) con el programa ELAG 4.0 (Bakeman, 1983). Se han considerado todas las secuencias como un sólo grupo, con un nivel de confianza del 99%. Las categorías podían seguirse a sí mismas. Con estas restricciones el análisis secuencial mostró las siguientes cadenas de eventos significativas, tal como recoge la figura 7.

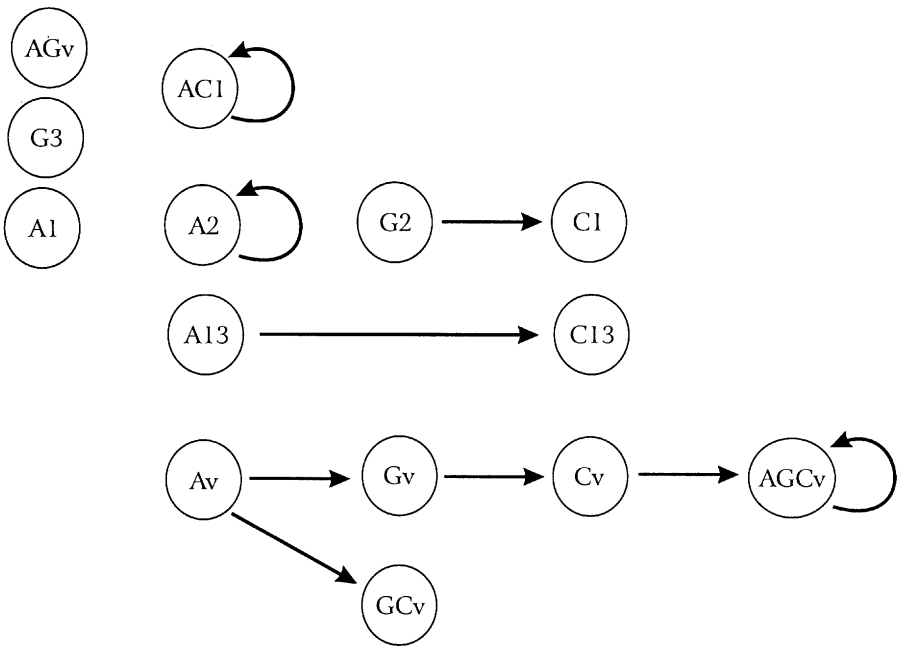

Figura 7. Secuencias de categorías con probabilidades condicionales significativas al 99%.

Las categorías AGv, Al, G3, GCv, Cl y C13, no eran seguidas de manera significativa por ninguna otra categoría. Tres categorías resultaron conectadas consigo mismas para el retardo +1: AC1 con Z=6.71; A2 con Z=3.70 y AGCv con Z=3.69. En cuanto a cadenas de eventos distintos consecutivos encontramos dos constituidas por categorías de error: G2 es seguida por Cl (Z=4.12), y A13 es seguida por C13 (Z=6.71). Finalmente, entre las categorías válidas encontramos las siguientes secuencias: Av-Gv (Z=2.62), Gv-Cv (Z=5.77) y Cv-AGCv (Z=3.08) formando una cadena, y por otro lado la secuencia Av-GCv (Z=3.18).

3. DISCUSIÓN

El grado de acuerdo obtenido entre las dos parejas de observadores hace pensar que las definiciones de las categorías son razonablemente adecuadas, lo que permite en muchos casos una identificación clara de los fenómenos. Cabe pensar también que los errores come-

tidos en la codificación pueden ser evitados en estudios posteriores. Los desacuerdos por omisión en la codificación de alguna de las parejas podrían ser solventados con un mayor entrenamiento de los observadores. Los desacuerdos por anotaciones diferentes se refieren a categorías muy concretas, por lo cual probablemente son ellas y no el sistema en su conjunto las que deben ser revisadas. Por una parte los índices de la categoría AGCv deben ser especificados en mayor grado, pues su similitud con los de AGv y GCv ha creado algunos problemas. Por otra, las categorías A0 y AG2 han resultado problemáticas, ocasionando desacuerdos cada vez que han aparecido. Ello tal vez esté señalando la conveniencia de considerar dichas categorías con mayor detalle en estudios posteriores.

El grado de acuerdo obtenido en un sistema formado por un número considerable de categorías supone una capacidad no desdeñable de diferenciación entre ellas. Los resultados confirman además la posibilidad de codificar cualquier categoría sin que necesariamente vaya seguida por la posterior en la jerarquía.

Con definiciones claras de cada categoría y con mutua exclusividad entre ellas es posible plantearse además que el conjunto recoge con exhaustividad el fenómeno en estudio. En nuestro trabajo se encuentran indicios de tal propiedad. En la realización de los sujetos se han identificado tanto diseños correctos como errores diversos. En los sujetos que hemos estudiado no hemos encontrado ni todos los diseños ni todos los errores posibles; sin embargo no encontramos razones para pensar que, estudiando otros sujetos y en circunstancias más complejas, no puedan aparecer los casos no encontrados en este estudio. Puesto que el sistema ha dado cuenta de lo realizado por los sujetos estudiados, cabe pensar que está preparado para identificar otras posibilidades que puedan aparecer en estudios posteriores.

A todo lo anterior no es ajeno el procedimiento utilizado. Probablemente resulta adecuado por ser el resultado de un largo proceso de ensayos y exploración, en el que se fueron solucionando problemas e inconvenientes de diversos tipos. Sin duda puede mejorarse aún, pero ciertas características del procedimiento no presentan ya problemas especiales. Por una parte el uso de índices múltiples, de tipo manipulativo y verbal, ha podido ser beneficioso al aportar información complementaria sobre lo que ocurría. Por otra parte el número considerable de categorías del sistema no fue un problema insalvable en la práctica. Como suponíamos, deducirlas con criterios claros a partir de un corto número de principios hizo manejable la casuística que los observadores encontraban y facilitó el trabajo con categorías de tipo molar. Suele considerarse que aunque la molaridad puede beneficiar la representatividad de los registros, la objetividad está en relación directa con el carácter molecular de la unidad de observación. Sin embargo no puede olvidarse que la correspondencia del aspecto molar-molecular con el de interpretación subjetiva-descripción objetiva es imperfecta y no necesaria (Anguera, 1988; Riba, 1991). Tener como referencia una adecuada organización de conceptos favorece el acuerdo en las codificaciones aunque las unidades de observación tengan carácter molar.

Sobre la base de esta inicial presunción de validez del sistema de categorías y del procedimiento seguido es posible explorar aspectos más particulares. La tarea propuesta no resultaba fácil de entrada, pero terminaba siendo asequible para el conjunto de los sujetos estudiados. Aunque los primeros intentos solían ser erróneos fueron mayoría los que termina-

ron exitosamente: los 19 problemas resueltos lo fueron por 11 de los 16 sujetos estudiados. También son mayoría las tareas categorizadas como válidas frente a los errores; y ello a pesar de que el número de categorías de error es superior y de que los errores aparecen también en algunos de los sujetos que acababan realizando correctamente la tarea. Se observa también que los sujetos no adoptan mayoritariamente los casos más sencillos para su tarea de diseñar contrastaciones: eligen más de dos valores de la VI de la investigación, y los combinan con varios valores de la variable extraña. Podría apuntarse entonces que la tarea de validación propuesta no es algo ajeno a las posibilidades de los sujetos, aunque no tengan especial experiencia en dicha tarea. Cometen errores, dudan y hacen cambios en los pasos a seguir, pero también en muchas ocasiones terminan por planificar la investigación conforme a los criterios convencionales de corrección. Finalmente destacar cómo en la primera sesión los sujetos eligieron mayoritariamente una estrategia transversal. Por tanto debería prestarse especial atención al entrenamiento de los profesionales clínicos en los diseños longitudinales, muy útiles en dicho ámbito.

Cada uno de los errores encontrados puede tener un significado diferente. El error 13 aparece ligado sólo a un sujeto que realiza con dicho error la sucesión de A y C. El tipo 2 muestra que los sujetos tratan a veces de validar una relación sin tener en cuenta al menos dos valores de la VI. No es de extrañar tal caso cuando la bibliografía ha tenido que recogerlo y darle el nombre específico de preexperimental para prevenir su uso. Por su parte el error tipo 1, el más frecuente, quizás refleje una falta de precaución por parte de los sujetos que lo cometen, al no plantearse la falta de representatividad que podría tener un sólo dato de la VD por condición. Puede que este error venga inducido por lo limitado del contexto en que se planteaba el problema; en esas condiciones ciertos sujetos pueden haber supuesto la existencia de un buen control a partir del cual inferir resultados válidos con un sólo dato por condición. No obstante, tal posibilidad también podría estar manifestando un exceso de confianza en el modo humano de investigar, olvidando las numerosas amenazas existentes para la validez incluso en investigaciones muy controladas. El error menos numeroso fue el tipo 3, es decir el que ilustra las situaciones de confundido. Es cierto que nuestros criterios evaluadores fueron poco exigentes puesto que los sujetos sólo tenían que controlar una variable extraña (la hora del día); por tanto podría ocurrir otra cosa en situaciones más abiertas o complejas.

En cuanto a las categorías, la asignación en cualquiera de sus posibilidades es la más frecuente y se da en todos los sujetos, lo que parece apoyar el papel básico que le concedíamos en nuestro sistema jerárquico de categorías. También se ve confirmada la posibilidad prevista de que aparecieran categorías G tras A, y C tras G; estas ocurrencias no eran necesarias y sin embargo han aparecido. De todos modos, estas secuencias se han dado de manera diferente en las categorías de error y en las válidas. Entre las de error se da una cierta variedad de secuencias, resultando significativas sólo dos categorías que se siguen a sí mismas (AC1 y A2) y dos cadenas (A13-C13 y G2-C1). Ello podría estar mostrando el modo tentativo en que los sujetos desarrollan la tarea de contrastación que se les propone. Van probando de maneras diferentes, sin un previo patrón claro y explícito. Las secuencias entre categorías válidas son en cambio más estables. La aparición de alguna de esas categorías era seguida significativamente por otra también válida, y sólo por ésta: Av era seguida de forma

significativa por Gv o GCv, mientras que Gv era siempre seguida por Cv y ésta siempre por AGCv. Además, una vez que aparecían las consideradas válidas no volvían a aparecer errores de manera significativa. En el mismo sentido, una vez que los sujetos resuelven la tarea correctamente insisten con otras realizaciones similares: es lo que parece indicar que la categoría que sigue a Cv sea siempre AGCv, y que ésta se vea seguida por realizaciones del mismo tipo en las que se introducen cambios en estrategias de investigación, técnicas de control o valores de la VI.

En conclusión, el estudio realizado parece permitir un acercamiento a la tarea de diseñar contrastaciones válidas, tanto en el ámbito clínico, de cara al cual se diseñó especialmente la tarea de caso único, como en cualquier otro ámbito de estudio de la Psicología. Tanto el sistema de categorías como el procedimiento parecen contener aspectos aprovechables. Las categorías han informado acerca de diversos aspectos de la planificación de las investigaciones. Se ha podido apreciar la correspondencia entre lo planteado por los sujetos y las categorías del sistema, tanto correctas como erróneas. Parece que el sistema señala las relaciones compartidas por los diferentes diseños de investigación correctos. Asimismo las informaciones obtenidas sobre secuencias encontradas, tipos de ellas, errores cometidos y proporción de soluciones válidas nos acercan a aspectos no tenidos normalmente en cuenta en los diseños.

De todos modos lo hallado debe ser tratado con cautela, pues su generalidad puede ser escasa. Se trataba de explorar posibilidades mínimas del sistema y del procedimiento de estudio. La generalidad de lo encontrado será algo a evaluar en estudios posteriores. A tal efecto son varias las direcciones posibles que pueden tomarse. Utilizar un mayor número de sujetos de diferentes edades y niveles de formación, incluyendo profesionales de distintos ámbitos, plantear problemas con mayor número de variables, de interacción y no sólo principales, plantearlos en contextos más abiertos, en los que se consideren más variables extrañas, orientando o no al sujeto sobre ellas y con instrucciones de diferentes estilos son algunos de los aspectos a través de los cuales evaluar la generalidad de lo encontrado. Teniendo en cuenta estos aspectos se podrá intentar generalizar los resultados obtenidos a campos aplicados como el clínico, caracterizados por contextos más abiertos y con un mayor número de variables extrañas.

Otra vía de generalización consistiría en relacionar más estrechamente este estudio con otros de objetivos similares. En los últimos años vienen apareciendo diversos trabajos (Campbell, 1993; Evans, 1990; Gholson, Shadish, Neimeyer, y Houts, 1989; Lovie, 1992; Moreno, Martínez, Trigo, Pérez Gil y Arambarri, 1994; Ribes, 1993) en los que se plantea y promueve un amplio programa de investigación sobre la ciencia y su método desde criterios y categorías psicológicas. Entre otros objetivos tal intento conlleva describir las pautas de acción de los investigadores para obtener información que complete los conocimientos derivados de otros puntos de vistas sobre el método. No cabe duda que la práctica de los investigadores es el referente más básico a partir del cual se han ido elaborando modelos y concepciones como las formales; por eso puede merecer la pena volver al referente base para recuperar aspectos no considerados. En este mismo sentido ya se ha comenzado a explorar (Moreno, 1994; Martínez Sánchez, 1993) la posibilidad de usar estas categorías para describir otros campos, tales como la generación de hipótesis y otros aspectos metodológicos como la medición, la validez y la fiabilidad.

BIBLIOGRAFIA

Allan, L.G. "Human contingency judgments: Rule based or associative?". A: *Psychological Bulletin,* 1993, núm. 114, pág. 435-448.

Altman, J. "Observational study of behavior: Sampling methods". A: *Behaviour,* 1974, núm. 49, pág. 227-267.

Anguera, M.T. "La observación (I). Problemas metodológicos". A: Fernandez-Ballesteros, R. A: Carrobles, J.A.I. (ed). *Evaluación conductual: Metodología y aplicaciones.* Madrid: Pirámide, 1981, pág. 292-333.

Anguera, M.T. *Observación en la escuela.* Barcelona: Graó, 1988.

Anguera, M.T. *Metodología de la observación en las ciencias humanas.* 4ªed. Madrid: Cátedra, 1989.

Arnau, J. *Diseños experimentales en psicología y educación.* Vols. I y II. México: Trillas, 1981-84.

Bakeman, R. "Computing Lag Sequential Statistics: The ELAG Program". A: *Behavior Research, Methods and Instruments,* 1983, núm. 15, pág. 530-535.

Barlow, D.; Hersen, M. *Diseños experimentales de caso único.* Barcelona: Martínez Roca, 1988.

Campbell, D.T.; Stanley, J.C. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally & Company, 1966 (trad. cast. en Buenos Aires: Amorrortu, 1973).

Campbell, D.T. "The Social Psychology of Scientific Validity". A: Shadish, W.R. (ed). *Social Psychology of Science.* Nueva York: Guilford, 1993.

Cohen, J.A. "A Coefficient of Agreement for Nominal Scales". A: *Educational and Psychological Measurement,* 1960, núm. 20, pág. 37-46.

Cook, T.D.; Campbell, D.T. *Quasi-Experimentation, Design and Analysis Issues for Field Settings.* Chicago: Rand McNally & Company, 1979.

Evans, T.D. "Understanding Understanding: On Psychology as a Science and Science as Psychology". A: Bjorgen, I.A. (ed). *Basic Issues in Psychology: An Scandinavian Contribution.* Soreidgrend (Norway): Sigma Forlag, 1990.

Gholson, B. et al. *Psychology of Science. Contributions to Metascience.* Nueva York: Cambridge University Press, 1989.

Harrop, A.; Foulkes, CH.; Daniels, M. "Observer Agreement Calculations: The Role of Primary Data in Reducing Obfuscation". A: *Bulletin of the British Psychological Society,* 1989, núm. 80, pág. 181-189.

Hay, L.R.; Nelson, R.O.; Hay, W. "Methodological problems in the use of participants observers". A: *Journal of Applied Behavior Analysis,* 1980, 13, 501-504.

Irwin, M.; Bushnell, M. *Observational Strategies for Child Study.* Nueva York: Holt, Rinehart & Winston, 1980 (trad. cast. en Madrid: Narcea, 1984).

Jorgensen, D.L. *Participant observation: A methodology for human studies.* Newbury Park, CA: Sage, 1989.

Klayman, J.; Ha, Y-W. "Confirmation, disconfirmation, and information in hypothesis testing". A: *Psychological Review,* 1987, núm. 94, pág. 211-228.

McCullagh, P. y Nelder, J. A. *Generalized linear models.* London: Chapman & Hall, 1989.

Lovie, A.D. *Context Science.* Nueva York: Harvester, 1992.

Martin, P.; Bateson, P. *La medición del comportamiento.* Madrid: Alianza Universidad, 1991.

Martínez, R.J.; Trigo, E.; Moreno, R. "Criterios para la clasificación de los diseños de investigación." A: *III Simposium de Metodología de las Ciencias Sociales y del Comportamiento.* Santiago de Compostela, 1993.

Martínez Sánchez, H. *El concepto de covariación como patrón descriptivo de aprendizaje*. Universidad de Sevilla, 1993. [Tesis doctoral no publicada]

Mehm, J.G.; Knutson, J.F. "A comparison of event and interval strategies for observational data analyssis assessments of observer agreement". A: *Behavioral Assessment*, 1987, núm. 9, pág. 151-167.

Moreno, R. "Prólogo". A: Barlow, D.H. A: Hersen, M. *Diseños experimentales de caso único*. Barcelona: Martínez-Roca, 1988.

Moreno, R. "Utilidad metodológica de una taxonomía de competencias relacionales". A: Hayes L. A: Ribes, E. A: López-Valadez, F. (coord) *Psicología Interconductual: Contribuciones en honor a J.R. Kantor*. Guadalajara: Universidad de Guadalajara, 1994, pág. 19-44.

Moreno, R. et al. "A Psychological Field Model of Scientific Method". A: *Revista Mexicana de Análisis de la Conducta / Mexican Journal of Behavior Analysis*, 1994, núm. 20 monografic issue, pág. 145-167.

Moreno, R.; Trigo, E.; Martínez, R.J. "Una aproximación a la dimensión psicológica del método de la ciencia". A: *I Simposium Nacional de Metodología de las Ciencias Humanas, Sociales y de la Salud*. Salamanca, 1989.

Moreno, R.; Trigo, E.; Martínez, R.J. "Competencias conductuales en los diseños de investigación". A: *II Simposium de Metodología de las Ciencias Humanas, Sociales y de la Salud*. Puerto de la Cruz, 1991.

Oaksford, M.; Chater, N. "A rational analysis of the slection task as optimal data selection". A: *Psychological Review*, 1994, núm. 101, pág. 608-631.

Riba, C. "El método observacional. Decisiones básicas y objetivos". A: Anguera, M.T. (dir). *Metodología observacional en la investigación psicológica (I): Fundamentación*. Barcelona: PPU, 1991.

Ribes, E. "La práctica de la investigación científica y la noción de juego de lenguaje". A: *Acta Comportamentalia*, 1993, núm. 1, pág. 63-82.

Sackett, G.P. "The Lag Sequential Analysis of Contingency and Cyclicity in Behavioral Interaction Research". A: Osofsky, J.D. (ed). *Handbook of Infant Development*. NuevaYork: Wiley, 1979.

Suen, H.K.; Lee, P.S. "Effects of the Use of Percentage Agreement on Behavioral Observation Reliabilities: A Reassessment". A: *Journal of Psychopathology and Behavioral Assessment*, 1985, núm. 7, pág. 221-234.

Wasserman, E.A et al. "Rating causal relations: Role of probability in judgments of response-outcome contingencies". A: *Journal of Experimental Psychology: Learning, memory and Cognition*, 1993, núm. 19, pág. 174-188.

Wright, H.F. "Observational Child Study". A: Mussen, P.H. (ed). *Handbook of Research Methods in Child Development*. Nueva York: Wiley, 1960, pág. 71-139.

Young, M.E. "On the origin of personal causal theories".A: *Psychonomic Bulletin & Review*, 1995, núm. 2, pág. 83-104.

ANEXO

Algunos de los índices verbales utilizados en la codificación fueron:

Categoría Av:

Av: Pongo 4 grupos de sujetos..., este grupo sin tomar café, 1 taza, 2 tazas y 3 tazas ... voy a poner los resultados. Ya está.

Av: ... un sujeto dándole las 3 tazas de café y al mismo sujeto no dándoselas y aplicándole después la prueba... haría esto mismo pero repitiéndolo un número de veces determinado, por ejemplo 8 veces y vería los resultados.

Categoría Gv:

Gv: ... con ningún café, fuera la hora que fuera, tenía estos resultados, con 1 y 2 cafés salía así y con 3 cafés salía bien por la mañana, por la tarde y por la noche.

Gv: ... y lo mismo hago por la tarde con 4 sujetos y 4 unidades de café distintas y dan los mismos resultados que por la mañana.

Categoría Cv:

Cv: ... en todos los casos obtiene los mejores resultados cuando toma 3 tazas de cafés.

Cv: ... y con 3 cafés ya sí obtendrían buenos resultados en el test.

Categoría AGv:

AGv: ... todos harían todas las pruebas. Entonces, por la mañana, sin tomar ningún café, daba cero resultados. Si tomaban 1 y 2 regular, y si toman 3 bien.

Categoría GCv:

GCv: ... el grupo que toma más cafés tiene mejores resultados, ... a medida que toman más cafés van teniendo mejores resultados.

GCv: ... si tomándose 3 cafés varía mucho que tomándose 1 o que tomándose 2 puedes llegar a la conclusión de si influye o no influye.

Categoría AGCv:

AGCv: ... un sujeto por la mañana toma 1, 2 y 3 cafés y da estos resultados, ... lo volvemos a hacer por la noche y en los 3 casos nos da los mismos resultados, es decir, que obtiene el mejor resultado con los 3 cafés.

AGCv: Yo tomaría 2 personas cada vez... le daría 1 café por la mañana, 2 cafés por la mañana, 3 cafés por la mañana... si les das 3 cafés el resultado no sería bueno, lo mejor que se tomaran 1 y lo peor que se tomaran 3.

Categoría A1:

A1: ... el mismo sujeto experimentando una vez con 1 café, 2 cafés y 3 cafés a la misma hora.... se verifican los resultados.

A1: ... a la misma hora con 0 cafés tiene resultados malos, con 1 café tendría resultados intermedios y con 2 resultados buenos. Si llegara a hacerlo con 3 yo puse 2 tarjetas verdes para decir que eran muy buenos.

Categoría A2:

A2: ... cojo al sujeto experimental por la mañana y le doy 3 cafés y me dan estos resultados.... luego lo cojo por la tarde, le doy 3 cafés y me dan estos resultados.

A2: ... el mismo sujeto.. por la noche y también le doy 3 cafés, y compruebo otra vez los resultados.

Categoría A13:

A13: ... por la mañana le daría 1 café que sería bueno hasta cierto punto, ... y 3 por la noche que sería negativo.

A13: Si bueno he puesto los 3 sujetos ... al de por la mañana le doy un café y le afectaría mucho, por la tarde intermedio y por la noche le doy 3 cafés y malos resultados.

Categoría G2:

G2: ... aplicándole al sujeto el tratamiento, o sea la VI, por la mañana, tarde y noche, para ver si afecta tomarse 3 cafés, a ver si afecta al resultado.

G2: ... siempre obtenemos los mismos resultados cada vez que el sujeto toma 3 tazas de café.

Categoría G3:

G3: ... con 1 café por la mañana obtienen estos resultados... y con 3 cafés fuera la hora que fuera lo harían así.

Categoría C1:

C1: ... un individuo toma 3 cafés por la mañana ... lo haría regular... Si tomara 2 pues tampoco lo haría mal.

C1: ... y ahora le doy 3 cafés, y en condiciones anteriores con 2 cafés los resultados han sido peores.

Categoría C13:

C13: ... por la mañana le he dado un café y tendría malos resultados... y por la noche le he dado 3 cafés y tendría los mejores resultados.

Categoría AC1:

AC1: ... en la mañana hacer tomar 3 tazas de café al sujeto y los resultados son óptimos... he hecho lo mismo con 2 tazas y el resultado era el mismo que con 3 tazas.

AC1: ... por la tarde con 3 tazas sólamente y el resultado es óptimo, por la tarde con 1 taza y sigue siendo el resultado óptimo.