Apuntes de Psicología
1999, Vol. 17, número 3, pp. 175-192.
ISSN 0213-3334

Colegio Oficial de Psicólogos
(Andalucía Occidental) y
Universidad de Sevilla

Dimensiones estructurales de diseño para la evaluación de programas

Mª Teresa ANGUERA ARGILAGA
Universidad de Barcelona
Salvador CHACÓN MOSCOSO
Universidad de Sevilla

Resumen

Las circunstancias inestables que caracterizan el contexto en el que se realiza la evaluación de programas de intervención hacen que sea prácticamente imposible plantear diseños de evaluación estándares. El objetivo de este trabajo es describir una serie de dimensiones estructurales de diseños de evaluación que se puedan implementar de una manera flexible en los contextos de intervención. Desde estos referentes de diseño de evaluación se intentará potenciar la obtención de información con el mayor grado de validez científica, sin arraigar la idea de una lista de diseños que pueden aplicarse ante determinadas situaciones estándares. El esquema de análisis de partida lo estructuramos en tres grandes dimensiones estructurales de diseño: usuarios del programa, naturaleza de los datos, y momento temporal de registro. A partir de estos ejes se justificarán, por una parte, cómo la combinación de dichos criterios dará sentido al uso de un tipo de diseños u otros, y por otra se planteará cómo los distintos elementos de diseño pueden presentar distintas implicaciones respecto al estudio y neutralización de amenazas a la validez. Los elementos de diseño serán estructurados en contenidos referidos a la asignación a las condiciones del programa, las medidas previa, durante y posterior a la implementación del programa, la formación de grupos de comparación y la implementación del programa.

Palabras clave: evaluación, validez, diseño, programas.

Los autores agradecen los valiosos comentarios que el Prof. José López Ruiz ha realizado a un borrador inicial de este artículo

Dirección de los autores: Mª Teresa Anguera Argilaga. Facultad de Psicología. Departamento de Metodología de las Ciencias del Comportamiento. Paseo del Valle Hebrón, 171. 08035 Barcelona. *Correo electrónico:* tanguera@psi.ub.es

Salvador Chacón Moscoso. Facultad de Psicología. Departamento de Psicología Experimental. Avda. San Francisco Javier s/n. 41005 Sevilla. *Correo electrónico:* schacon@cica.es

Abstract

The unstable circumstances of programme evaluation contexts make standard programme evaluation designs practically impossible to provide. The main objective of this paper is to describe the structural dimensions of program evaluation designs in order to implement them in a flexible way in different evaluation contexts. These dimensions will try to enhance the scientific validity of the evaluative data obtained, without offering the idea of a series of design structures valid for certain standard situations. The structural dimensions are: clients of the programme, type of data, and recording moments. The combination of these dimensions will justify first, the use of different evaluative designs, and second the implications for studying and neutralizing threats to validity. Design elements will be structured around contents refering to programme conditions, pretest and posttest measures, comparison groups and type of implementation..

Key words: evaluation, validity, design, programme.

La evaluación de programas se ha convertido en una actividad obligada, tanto para los programas de intervención que ya están en funcionamiento, como para la implementación de nuevos programas. Como consecuencia de este hecho en los últimos veinte años, la investigación en evaluación de programas se ha convertido en una nueva disciplina metodológica, separada de las más generales de investigación social, educativa o de la salud. Esta situación se pone de manifiesto en la publicación de numerosos textos monográficos y revistas especializadas sobre el tema, la existencia de asociaciones tanto profesionales como académicas, la organización de cursos de doctorado y de postgrado sobre la temática, o la inclusión de estos contenidos en los planés de estudios universitarios.

En la época actual, y considerando la amplia diversidad de planteamientos evaluativos que se han desarrollado en la última década, nos encaminamos ya hacia un enfoque cada vez más integrador, signo inequívoco de mayor madurez y de que la evaluación de programas se está configurando como disciplina. En este sentido en la actualidad nos encontramos como factor común que cualquier proceso evaluativo persigue obtener una información de calidad sobre un determinado programa de acción. En nuestro caso consideramos información de calidad a la información válida obtenida desde los criterios de la metodología científica, y es por ello que consideramos el concepto global de validez científica como el pilar a partir del cual justificar el desarrollo de cualquier diseño de evaluación.

Por todo ello, en el presente trabajo defendemos que la evaluación de programas, en sus distintos aspectos, sigue las reglas del método científico, aunque la investigación evaluativa no se reduce a éstas, constituyendo hoy un área específica y separada de la anterior. Esta separación y la formación de un *corpus* propio, se debe a las peculiaridades de la investigación evaluativa, derivadas en primer lugar de su insistencia en los aspectos de *valor* de los programas (Scriven, 1980); en segundo lugar, la investigación evaluativa se diferencia por su *aplicabilidad*, tanto desde el punto de vista *formativo*, proporcionando *feedback* a los programadores y administradores de los programas (Vedung, 1993, 1996), como en la *toma de decisiones* sobre los resultados de éstos, su continuidad,

modificabilidad, etc. (Weiss, 1983; Wholey, 1987; Martínez-Arias, 1996).

Desde este planteamiento entendemos que la evaluación de programas se desarrolla a lo largo de un proceso lógico que sustancialmente no difiere del proceso de investigación en un ámbito aplicado. Se cuenta con una realidad compleja, pero tangible, con programas en que se implementan acciones que a veces no se ajustan al calendario, o que no se ejecutan por igual en todos los sujetos, pero en los cuales el qué, el cómo y el cuándo son registrables. El sector en que se ubica la necesidad, las características del entorno en que se enclava, y la propia naturaleza de la carencia condicionan la amplia casuística de programas, y sobre todo, les imponen fuertes limitaciones que chocan frontalmente con los requisitos que impone el rigor del método científico (Anguera y Chacón, en prensa).

De forma ilustrativa, podemos pensar en la *selección de sujetos*. En los manuales metodológicos, y simplificando mucho, se insiste en la práctica –casi se fuerza a su opción– del muestreo probabilístico si se opta por una vía deductiva, y por tanto, siempre que nos situemos bajo la cobertura de un marco teórico consolidado. Este muestreo probabilístico, por ser equiprobable, además de representativo, ¿garantizaría el componente de equidad en los potenciales usuarios? En el momento en que bajamos a la arena de lo cotidiano en evaluación de programas, si tenemos que evaluar un programa domiciliario de atención geriátrica se descarta de entrada una respuesta afirmativa, ya que serán los propios usuarios, o sus allegados, o los responsables de Servicios Sociales de la zona, los que tratarán de recabar la adscripción al programa. Aquí no cabe de plano el muestreo probabilístico (los distintos casos no son equiproblables, ni se eligen al azar, y ni siquiera se puede afirmar, al menos de forma general, que sean representativos del colectivo afectado por la necesidad), independientemente de que exista o no un determinado marco teórico relativo al problema y a su probable intervención, e independientemente también de que existan suficientes recursos (humanos, temporales, económicos, etc.) para que todos ellos sean atendidos. En consecuencia, no podemos hablar de equidad desde el mismo momento en que se plantease una selección mediante muestreo probabilístico de usuarios.

Siguiendo con la selección de sujetos, si se plantea la cuestión desde la vía inductiva de forma obligada o deseada, sea porque no existe un determinado marco teórico de referencia, o porque no nos interesan los que hayan, deberá seleccionarse un caso único inicial al que le sigue una progresiva acumulación de casos afines, con el fin de llegar a encontrar regularidades en el comportamiento de todos estos casos y que se pueda ir trazando un esquema de funcionamiento del caso general. Yendo a la realidad del profesional, y desde la mayor flexibilidad que permite el planteamiento inductivo, si bien se establecerán criterios para la inclusión en el programa en función de las características técnicas de la necesidad y las condiciones físicas, personales, económicas, familiares, etc., del potencial usuario, es constatable que no siempre se tratará de casos *afines*, al menos en el sentido restrictivo del término. Consecuentemente, tampoco se llevaría a cabo la selección de sujetos de forma metodológicamente correcta, además de seguir cuestionándonos el concepto de equidad en los potenciales usuarios.

Pero podemos pensar en otros tipos de programas, como los institucionales relativos a preservar la capa de ozono o mantener limpia una ciudad o una superficie; o los sanitarios sobre hábitos higiénicos, o la deshabituación al tabaco o a drogas diversas; o los de educación vial. En todos ellos existen componentes que escapan -por la propia complejidad de la realidad- a la norma científica, lo cual da lugar a dos alternativas por las que entendemos que es fácil optar: se flexibiliza la norma, o se está al margen de toda norma.

La segunda posibilidad dejaría a la evaluación de programas fuera de un ámbito formal de estudio por disciplinas como la Psicología, Sociología, Medicina, Educación, etc. Se podrían describir *casos* aislados, como se ha hecho desde determinadas corrientes radicales de metodología cualitativa -ver, por ejemplo, Smith y Cantley (1985), o el número monográfico de *Qualitative Health Research* editado por Engel (1992), o el de *Qualitative Inquiry* editado por Reason y Lincoln (1996)-, pero es indudable que masivamente interesa ajustarnos a la lógica del procedimiento.

Esta lógica, sin embargo, que responde a los principios del positivismo científico, debe ser sensible a las específicas características individuales, situacionales, del programa, de los recursos disponibles, etc. La casuística es amplísima, y no es fácil lograr el equilibrio entre el rigor imprescindible y la flexibilidad adaptativa respecto el programa a evaluar.

Todas estas circunstancias inestables que caracterizan el contexto en el que se realizan las evaluaciones de programa de intervención hacen que sea prácticamente imposible plantear estructuras estándares de diseño apriorísticas. En este sentido el objetivo de este trabajo es describir una serie de dimensiones estructurales de diseño que puedan implementarse de una manera flexible en los contextos de intervención particulares. Desde estos referentes de diseño de evaluación se intentará potenciar la obtención de información con el mayor grado de validez científica, sin arraigar la idea de una serie de estructuras de diseño rígidas que sólo pueden aplicarse en situaciones estándares (Shadish, Cook y Campbell, en preparación).

Dimensiones estructurales de diseños de evaluación

En la literatura se proponen distintas formas de clasificar los diseños que podrían aplicarse a la evaluación de programas (podemos referirnos a algunos autores, como Arnau, Anguera y Gómez, 1990; Ato, 1991; Judd y Kenny, 1981; Kazdin, 1980; Kish, 1987; Owen y Rogers, 1999). A pesar de ello resulta difícil encajar estas clasificaciones en la práctica profesional de la evaluación de programas. Esta situación se da fundamentalmente porque en el contexto de la evaluación de las intervenciones es complejo encontrar modelos o referentes teóricos a partir de los cuales disponer de criterios con los que desarrollar un modelo de selección de los distintos componentes que tendría que analizar un programa de evaluación. Tengamos presente que no sólo nos referimos a la selección de usuarios del programa a evaluar, tal y como planteábamos en el ejemplo del apartado anterior; es decir, a qué población de sujetos nos referimos exactamente; también hemos de tener en cuenta cuáles son los distintos tipos de intervenciones teóricas que podrían utilizarse en cada casuística, cuáles son las variables que vamos a

medir, cuál es la delimitación realizada del contexto de intervención, en qué momento o momentos concretos y con qué recursos vamos a registrar los datos,... A estos interrogantes, que aunque aparentemente simples son de hondo calado, se les une el problema de conseguir una alta fiabilidad en las medidas (por ejemplo, para evaluar el impacto de un programa).

Todas estas circunstancias hacen que nos estemos refiriendo a un ámbito de trabajo en el que impera una filosofía falsacionista, ya que el grado de conocimiento previo y de control sobre el objeto de evaluación es reducido. En este sentido se habrán de prever cuáles pueden ser las principales amenazas a la validez, tanto de representatividad como de control, que puede presentar nuestra evaluación particular. El problema de esta situación es la dificultad de disponer de teorías sobre las posibles interacciones entre las variables en los contextos de intervención particular; más aún cuando unas mismas variables pueden interactuar de manera distinta en un mismo contexto en momentos temporales distintos.

En el marco de esta realidad descrita abogamos por justificar el desarrollo de un diseño de evaluación desde unos referentes dimensionales que se combinarán de distintas formas dependiendo del contexto de intervención particular donde se implementen. El esquema de partida lo estructuramos en tres grandes dimensiones: *Usuarios del programa, naturaleza de los datos,* y *momento temporal* (Anguera y Chacón, en prensa). Para establecerlas hemos partido de una idea original de Cattell (1952), que ya a mitad de siglo empezó a producir sus frutos. Utilizaba un paralelepípedo en que las aristas representaban personas, variables de medida y ocasiones de medida para poder ilustrar los datos en estudios de covariación. De esta forma, cada cara del paralelepípedo permite obtener una matriz bidimensional de puntuaciones, ya que la definen dos aristas, y la tercera cara no se muestrea, pero se fija. Muchos años después, Nessealroade y Hershberger (1993), en sus estudios sobre población, la adaptan para explicar la variabilidad intraindividual (Nesselroade, 1988, 1991) utilizando las mismas dimensiones de personas, variables de medida y ocasiones de medida.

En evaluación de programas consideramos que existen tres referentes desde una perspectiva metodológica: a) a quiénes va dirigido el programa, ya que de lo contrario éste perdería su razón de ser, motivo por el que los *usuarios* ocupan el primer lugar; b) tipo de información que se obtiene, habitualmente de carácter cambiante a lo largo del proceso de implementación y en función de las diversas acciones que se llevan a cabo, por lo que la *naturaleza de los datos* es un referente obligado; y c) el carácter sincrónico o diacrónico del proceso de evaluación, dependiendo de si nos situamos en la evaluación sumativa o formativa, respectivamente (siguiendo con los presupuestos del modelo lineal *vs.* no lineal de Veney y Kaluzny -1984-). A continuación comentamos los aspectos más relevantes de cada una de ellas.

Usuarios del programa

Los usuarios del programa son los individuos en los que se detectó una necesidad y a quiénes van dirigidas las acciones del programa. Dicho en otros términos, y de forma genérica, son los individuos que contestan las preguntas de las entrevistas, rellenan los cuestionarios, y, en algunos casos, aceptan que se observe su actividad.

El conjunto de personas al que se destina el proyecto se le denomina usuarios, población-objetivo, población-meta, grupo-meta, o grupo focal.

Una vez establecida la población-objetivo y su localización espacial se pueden ya diferenciar los diferentes subcriterios desde los cuales se pueden fijar las características de los usuarios, así como delimitar el propio colectivo. En este sentido los usuarios pueden clasificarse dependiendo del *impacto del programa* (distinguiendo entre usuarios directos, indirectos y beneficiarios públicos en general), de su *nivel social* (distinguiéndose entre usuario activo, pasivo, potencial o allegado), de la *pluralidad* con la que sean delimitados (un individuo, un grupo de individuos, o la colectividad en general), de la *cobertura y extensión de uso*, en función del grado en que se accede a la población diana (Cohen y Franco, 1992), o del procedimiento de *selección* empleado para la realización de los distintos tipos de muestreo (Martínez Arias, 1995).

Naturaleza de los datos (grado de intervención)

En evaluación de programas no se dispone usualmente de instrumentos standard para la recogida de datos, de ahí que los instrumentos suelen ser en muchas ocasiones de elaboración propia. Esta situación provoca que los datos utilizados en evaluación son frecuentemente de carácter cualitativo y/o categórico; por ello puede decirse que la naturaleza del dato guarda una relación -no absoluta- con el grado de intervención o interacción con un programa. El uso de un instrumento standard, con sus normas de aplicación, registro y baremación, pueden suponer un mayor grado de intervención en el programa a evaluar que otro tipo de registros semi-estándares o de elaboración propia.

A su vez, el evaluador dispone de una gama de técnicas de recogida de datos que comprenden desde las que requieren una interacción mínima con un programa (como medidas discretas o revisión de los datos archivados) a las que implican una moderada interacción personal con la situación (como escalas, tests y encuestas) y las que requieren una interacción activa con los usuarios del programa (como observación y entrevistas).

Por todos estos motivos la naturaleza del dato puede ser un indicador (aunque no exclusivo) del grado de intervención que el proceso evaluativo está incorporando en el programa a evaluar.

En términos generales en esta segunda dimensión referida a la naturaleza del dato podemos contemplar diversos criterios:

a) Por una parte podemos referirnos a los datos cualitativos vs. cuantitativos. Se trata de una cuestión que ha resultado ser altamente polémica (Alvira, 1991; Anguera, 1989, 1995a; Cook y Reichardt, 1986; Fernández-Ballesteros, 1995; Filstead, 1986; Hernández, 1995; Ianni y Orr, 1986). Fue a partir de la década de los ochenta, cuando se empieza a preconizar un acortamiento de distancia entre ellos y a plantearse su uso complementario. Un evaluador no tiene por qué adherirse ciegamente a uno de ellos, sino que puede elegir datos de una u otra naturaleza, indistintamente, y combinarlos entre sí, si es que de esta forma logra una adaptación flexible a su problemática.

b) Por otra parte, según *características del instrumento* de recogida de datos nos podemos encontrar con datos de una u otra naturaleza, que guarda una relación -no absoluta- con el carácter del instrumento utilizado, cuestión que interactúa frecuentemente con el carácter cualitativo, cuantitativo o de complementariedad entre ellos.

Sin perder la vinculación con los dos subcriterios anteriores, hay también que considerar el *sistema de registro*, es decir la forma cómo se recoge la información (Anguera, 1995b; Hernández, 1995), optando por un sistema escrito, oral, mecánico, automático, icónico, etc., que facilite su almacenamiento.

El plano en que se sitúa el registro deberá permitir una necesaria elaboración posterior, y consecuentemente, la codificación hará posible la transformación de una información inicial, muchas veces narrativa, a un sistema de símbolos altamente estructurado y que permita un tratamiento cuantitativo.

Momento temporal de registro

La tercera dimensión hace referencia a cuándo se lleva a cabo la recogida de datos, y mantiene una indudable relación -aunque no coincidencia- con la evaluación sumativa y formativa.

Las posibilidades más diferenciadas, al margen de que quepan muchas posiciones intermedias relativas a puntos de corte en el proceso de implementación, son: un registro puntual, un seguimiento o un registro retrospectivo.

En el *registro puntual* la recogida de información tiene lugar sólo en un momento temporal, que suele ser una vez se terminó de implementar el programa. Habitualmente se utilizan instrumentos estándar para dar cuenta de los resultados (situación propia de la evaluación sumativa).

En cambio el *seguimiento, o continuidad prospectiva* sigue el curso de un proceso en la recogida de datos, por lo que se adapta particularmente bien a la evaluación formativa, ya que en cada fase del proceso cabe recoger y analizar los datos que se van obteniendo.

En el *registro retrospectivo* se plantea esencialmente el registro un tiempo después de haber terminado la implementación de un programa. Es frecuente la evaluación retrospectiva en estudios de impacto (social y ambiental). Los principales problemas que plantea el registro retrospectivo se refieren a la validez de la información recogida:

- Si es material de archivo, se pudo haber recogido mediante criterios distintos de los que ahora interesan para la evaluación.
- Si se requieren informantes, o se recoge la información de los usuarios, la información elicitada puede estar afectada de olvido, distorsión, o falta de contextualización adecuada por el tiempo transcurrido.

Configuración de diseños de evaluación

De acuerdo con el complejo marco en el que se desarrolla un programa de evaluación hemos justificado la dificultad de poder establecer una relación de diseños estándares de evaluación. Esta situación justifica la necesidad de configurar los diseños evaluativos de acuerdo con las necesidades y características del programa concreto, y por tanto desde una combinación

de las dimensiones de diseño anteriormente señaladas. El esquema básico resultante de la consideración conjunta de todos los referentes comentados lo podemos configurar a partir de cruce de las dimensiones de usuarios y temporalidad, en combinación con el grado de intervención o dominio que se tenga sobre el contexto de evaluación (ver figura 1).

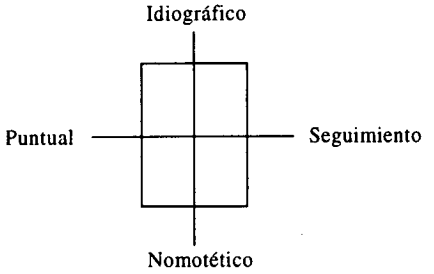

Figura 1. Esquema básico para la configuración de diseños de evaluación de programas (Anguera, 1995b).

Carácter Idiográfico vs. Nomotético (eje vertical)

El eje vertical se refiere al carácter idiográfico o nomotético en función de los usuarios del programa de intervención. No siempre adquiere la misma relevancia, pero plantea importantes cuestiones a nivel metodológico (Posavac y Carey, 1985) según se trate de sujetos individualmente considerados o de una colectividad (o muestra representativa de ella), si atendemos a la propuesta clásica de Allport (1942) en relación a los términos *idiográfico vs. nomotético.*

Ahora bien, a tal propuesta se han incorporado variantes adaptativas a las diversas situaciones evaluativas:

a) Se considerarán también como idiográficos estudios que amplían o restringen la propuesta clásica consistente en un individuo. Por una parte, entre los primeros se hallarán todos aquellos casos en que los usuarios son varios individuos entre los cuales existe un criterio de afinidad, agrupación, o reglas del juego a seguir; por ejemplo, un programa de intervención familiar, independientemente de cuántas personas componen aquella unidad familiar. Y, por otra parte, los que restringen el concepto clásico de idiográfico se centran en un solo nivel de respuesta, sea de un individuo único, o de varios; por ejemplo, si consideramos únicamente el nivel de conducta verbal, y evaluamos la resolución de un conflicto entre miembros de una familia a partir de la discusión de los respectivos puntos de vista y balance entre pros y contras de cada opción de solución.
b) Nomotéticas serán también aquellas variantes en que, independientemente de que tengamos un usuario o un grupo de usuarios, interesan varios niveles de respuesta. Así, en un programa de atención psicológica a enfermos infartados y sus familiares interesan los niveles de respuesta verbal y no verbal. Luego, nomotéticos serán todos aquellos diseños evaluativos en que se configura un elemento de pluralidad de unidades, sean individuos (propuesta clásica) o niveles de respuesta (variante posteriormente introducida).

Una vez delimitado el número de participantes en un estudio evaluativo (individuos sobre los que se interviene), el eva-

luador deberá decidir si todos deberán formar parte o no de la correspondiente evaluación, o incluso vincular esta decisión a distintas fases del proceso.

Los principales argumentos a favor de la inclusión del colectivo completo o de una muestra representativa del mismo son de índole metodológica (análisis de los efectos en toda la cobertura relativa a personal), pero también ética, y, en ocasiones, política. Así, si un centro hospitalario ofrece una unidad de atención psicológica a enfermos infartados por segunda vez, y es mayor la demanda que el número de plazas que puede cubrir el servicio, ¿se podría hablar de un criterio más "ético" que otros?, ¿cabe, desde los principios éticos, extraer una muestra representativa cuando todo el colectivo presenta un mismo tipo de necesidad demandada?

No faltan tampoco argumentos para el estudio evaluativo de los efectos de un programa en fases diferenciadas, de forma que a un primer análisis efectuado de forma nomotética le sigue un segundo basado en el estudio de sujetos individuales. La perspectiva idiográfica, desde la expansión e incidencia actual de la Psicología de las diferencias individuales, está alcanzando una gran relevancia en la implementación y evaluación de programas sociales y sanitarios. Cada vez los profesionales son más sensibles a la consideración diferencial de sujetos que, por su trayectoria vivida (circunstancias personales, *event-life*, rasgos de personalidad, etc.), requieren un análisis específico e individualizado de los efectos de un determinado programa de intervención. Pensemos en niños con transtornos comportamentales en el aula y con historias de vida absolutamente distintas (Herrero, 1989), o en enfermos con repetidos infartos que tenían muy diferente nivel de calidad de vida (Tuset, 1990), o en internos penitenciarios que cumplen una condena de igual duración a partir de un historial personal y delictivo completamente distinto (Redondo, 1992), o en deportistas que presentan determinadas peculiaridades en sus tácticas de juego (Hernández Mendo y Anguera, 1998).

Temporalidad del registro (eje horizontal)

La configuración básica de los diseños permite distinguir entre *registro puntual* y *seguimiento*. El registro puntual permitirá realizar un análisis de la situación en un momento dado en el tiempo, mientras que el seguimiento implica disponer de un determinado número de sesiones a lo largo del período de implementación del programa.

El criterio de temporalidad en el registro permite tener también en cuenta el punto de partida (previo, durante, o después de la aplicación de la intervención, o, expresado en otros términos, de la implementación del programa), y el periodo de cobertura en la recogida de datos (hasta el fin de la intervención, seguimientos puntuales periódicos hasta un determinado momento, igual con un seguimiento continuo, etc.).

Es muy fácil de argumentar cuál es el óptimo o ideal, partiendo del presupuesto de la existencia de recursos suficientes. Evidentemente, desde antes del inicio de la intervención, durante el tiempo que implique su puesta en práctica, y efectuando un seguimiento posterior a medio o largo plazo que posibilite un análisis riguroso de los efectos del programa.

Ahora bien, las distintas posibilidades que implica el barajar estos elementos, la necesidad de adecuarse a recursos generalmente limitados (Fienberg y Tanur, 1987),

y la propia naturaleza de la intervención, deben dar lugar a las decisiones relativas al registro (cómo, desde cuándo, hasta cuándo, con qué periodicidad, con qué garantías en la formación del personal que participa en la evaluación, etc.), en el más amplio sentido del término (Blanco y Anguera, 1991).

Finalmente, será conveniente distinguir, en este criterio relativo al carácter continuo o discreto del registro a lo largo del tiempo, entre la recogida de datos actuales (Plewis, 1985), y los retrospectivos (Holland y Rubin, 1988), como por ejemplo los referidos a material de archivo, tanto si se trata de datos censales o estadísticos, como de protocolos o informes personales (autobiográficos o realizados por terceras personas), siempre que se mantenga la homogeneidad de los criterios seguidos en su recogida y, en su caso, codificación (como categorías en un análisis de contenido de autoinformes), y no entorpezca su utilización el tan frecuente problema de los *missing data* (Little y Rubin, 1987).

Grado de intervención o dominio sobre el contexto de evaluación (interior vs. exterior del recuadro)

A estas dos dimensiones se ha de incorporar el referente del grado de intervención sobre la situación. Hemos intentado simbolizar esta dimensión presentando un recuadro interior en los ejes de la figura 1. Se pretende representar una superposición de dos planos distintos, de mayor a menor intervención respectivamente, dependiendo de si se está más hacia el centro o hacia el exterior de dicho recuadro interior. Con el término intervención en la situación hacemos referencia a la naturalidad de la situación, es decir el grado en que la relación de los sujetos, usuarios del programa, modifican sus interacciones naturales con el medio. Por ejemplo, en un programa de natación en la tercera edad en que los participantes realizan habitualmente ésta u otras actividades deportivas nos estaríamos refiriendo a un diseño de evaluación de baja intervención (este programa estaría representado en la parte exterior del recuadro). En cambio si nos interesase un diseño en el que un grupo de familias marginales de una gran ciudad han sido seleccionadas al azar y asignadas posteriormente a pueblos con bajo número de habitantes de otras provincias, para con ello disminuir la probabilidad de que los hijos de estas familias realicen conductas delictivas, se trataría de un tipo de programa con un alto grado de intervención (este programa estaría representado más hacia el interior del recuadro central de la figura 1).

El concepto de intervención no es dicotómico, no podemos establecer que haya o no intervención, se trata de un concepto de grado. De hecho no tendría sentido plantear la no existencia de intervención en tanto que el propio hecho de desarrollar una evaluación (con sistemas de registro, instrucciones a los usuarios en su caso,...) supone una intervención en sí misma. Los niveles de intervención se pueden desglosar de diversas formas. La opción que defendemos en este trabajo distingue entre diseños de intervención baja y diseños de intervención media-alta (parte exterior o interior, respectivamente, en el recuadro central de la figura 1). Se opta por esta dicotomización de la gradación en el nivel de intervención porque parece claro cuáles pueden ser los extremos de este eje bipolar, las denominadas metodologías naturalistas y experimentales, pero una vez que nos adentramos en ese teórico conti-

nuo es difícil establecer criterios que nos indiquen los límites a partir de los cuales referirnos a un tipo de metodología u otra. Más aún cuando en el ámbito de la intervención real suelen ser varios los procedimientos utilizados en un mismo programa de evaluación.

Desde esta argumentación, y obviando que en la práctica evaluativa se suelen utilizar conjuntamente distintos tipos de procedimientos, de una forma simplificada incluiríamos en los denominados diseños evaluativos de intervención baja a los diseños observacionales (diacrónicos, sincrónicos y mixtos), e incluiríamos en los diseños evaluativos de intervención media-alta a los diseños de grupo control no equivalente, de discontinuidad en la regresión y de series temporales interrumpidas.

A pesar de que hemos acabado desembocando en una interesante discusión, que sin duda requiere una mayor extensión en su tratamiento, consideramos que la cuestión no radica tanto en poder delimitar con precisión el grado en que se está interviniendo, sino más bien en tenerlo presente y estudiado en lo posible. Con ello se intentará conocer qué posibles fuentes de variación estamos incorporando en los datos registrados. A partir de este análisis detallado podremos disponer de más elementos de juicio a partir de los cuales valorar el grado de validez de la información registrada, y a partir de la cual se van a tomar decisiones que en la mayoría de los casos tienen una importante repercusión social.

Relación entre elementos de diseño evaluativo y neutralización de amenazas a la validez

Una vez descritas las distintas dimensiones estructurales a partir de las cuales diseñar la evaluación de un programa pasaremos a describir cómo los distintos elementos de diseño pueden ser tratados para reducir las distintas amenazas a la validez de la información recogida en el proceso evaluativo.

En todo diseño evaluativo hemos de tomar una serie de decisiones en las que combinamos de distintas formas las características estructurales de diseño previamente descritas. En este apartado hemos sistematizado estas decisiones sobre el diseño evaluativo, de tal forma que vamos a presentar una gradación de las variantes de diseño que se pueden presentar respecto al estudio y neutralización de amenazas a la validez. Estas variantes de diseño las hemos estructurado en contenidos referidos a la asignación a las condiciones del programa, la medida previa a la implementación del programa, la medida posterior a la implementación del programa, la formación de grupos de comparación y la implementación del programa.

En primer lugar mencionar que, cumpliéndose todas sus condiciones de aplicación, el experimento aleatorio, en el que el proceso de *asignación a las condiciones del programa es completamente conocido*, es el que presenta mayores garantías de validez. Normalmente en el experimento aleatorio se dispone de un modelo teórico contrastado en el que existen criterios de selección y asignación de los distintos componentes de programa. Todo ello posibilita un estudio sistemático de las distintas fuentes de variación en el diseño evaluativo, y por tanto un alto grado de conocimiento y control sobre la situación objeto de evaluación que permite el estudio y neutralización de un gran número de amenazas a la validez. A pesar de ello, el experimento aleatorio es difícil de ejecutar

(Boruch, 1997; Campbell y Russo, 1999) y es por ello que los diseños más potentes donde no es posible la asignación aleatoria son aquellos en los que es posible conocer, en la mayor medida posible, el procedimiento de asignación (por ejemplo en el diseño de discontinuidad en la regresión, Marcantonio y Cook, 1994). El problema del conocimiento de las reglas de asignación está directamente relacionado con la *conformación de grupos lo más similares posible*, para poder disponer de comparaciones válidas de los posibles efectos del programa. En los casos donde la asignación aleatoria no es posible se pueden utilizar técnicas como el *emparejamiento previo* de los sujetos antes de asignarlos a las condiciones del programa, aunque esto no siempre es factible y a su vez puede provocar más problemas que beneficios al añadir fuentes de error al proceso de asignación.

Respecto a las medidas previas a la implementación del programa, se ha comprobado que las *múltiples medias previas* cuanto más numerosas sean mejor, en tanto servirán para analizar los efectos de maduración, los artefactos de regresión y el estudio de los posibles efectos de instrumentación y medida. En ocasiones no cabe realizar medidas múltiples previas, por lo que al menos deberíamos registrar *una medida previa*. Si tampoco es posible hay algunas alternativas como la *medida previa de muestras independientes* (con el serio problema de la representatividad de la muestra elegida) o la posibilidad de tomar *medidas retrospectivas*, preguntando a los sujetos, o desarrollar *medias aproximadas a la variables medidas en el programa*. En estos dos últimos casos se corre un alto riesgo de falta de fiabilidad en las medidas e incluso el uso de variables inadecuadas en el caso de la configuración de un modelo de medida aproximado, de ahí que en la práctica se usen poco.

En cuanto a las medidas posteriores a la implementación del programa, siempre vamos a partir de al menos *una medida posterior* a la que, si es posible, le deberíamos de añadir *múltiples medidas posteriores* a la intervención que pudiésemos comparar con un patrón de medidas previamente elaborado desde una teoría sustantiva. En esta misma lógica es en la que se basa el uso de *variables dependientes no equivalentes*, es decir, disponer de dos constructos plausibles, en el que uno de ellos se espera que sea afectado por el programa y el otro, aunque no lo sea, pueda servir para analizar y descartar las mismas amenazas a la validez que afectan a la variable de interés.

Respecto al grupo de comparación, nos referimos a la necesidad de disponer de un referente válido a través del cual podamos valorar que hubiese ocurrido si el programa no hubiese intervenido. En definitiva lo que interesa es conformar grupos semejantes que difieran potencial y exclusivamente en haber recibido o no el programa; por este motivo el uso de las *cohortes* es mejor que el uso de *grupos no equivalentes* (obtenidos por procedimientos no aleatorios), en tanto que la semejanza entre las cohortes será normalmente mayor que entre grupos de sujetos, usuarios de los programas, que no comparten un mismo contexto. En último término se pueden utilizar *múltiples grupos de comparación no equivalentes* con objeto de explorar más amenazas a la validez y poder triangular los datos de tal forma que nos permita una estimación más precisa del rango de variación de los efectos del programa. De todos modos se ha de tener en cuenta que si la configuración de un grupo de control no

equivalente implicaba riesgos precisamente en cuanto a su no comparabilidad, estos riesgos se multiplicarán al aumentar los grupos. En ocasiones los grupos de comparación son construidos, como por ejemplo mediante la obtención de puntuaciones desde la *extrapolación de la regresión*, el uso de *grupos normativos* o el uso de *datos secundarios*; no obstante todas estas últimas variantes suelen presentar problemas de representatividad y fiabilidad.

Por otra parte, la implementación del programa es un elemento fundamental para llevar a cabo la interpretación de sus efectos. Es este caso, en la posible gradación de las variantes se puede priorizar el uso de *replicaciones de la intervención intercambiada*, siempre que el efecto del programa no sea persistente, en tanto supone beneficios para la validez interna y externa. Si esto no fuese posible la siguiente tentativa podrían ser los *diseños de reversión* «ABAB», (donde A indica presencia de intervención y B ausencia de intervención), para seguidamente situar al mismo nivel de relevancia, en cuanto al logro de la validez, *la intervención invertida* y la *supresión de intervenciones*, éste último siempre que el efecto no sea persistente. En último lugar mencionaríamos el diseño con *dosificación en la exposición de la intervención*, ya que puede presentar serios problemas de selección, que a su vez pueden interactuar con las distintas variantes de intervención implementadas.

Después de hacer una jerarquización de las principales variantes existentes en los elementos de diseño, hemos de plantear que no existe un único diseño ideal. El mejor diseño para una evaluación depende de las hipótesis concretas que se quieran evaluar, de los distintos tipos posibles de amenazas a la validez que se puedan encontrar, lo cual a su vez depende del conocimiento previo que se tenga de estudios anteriores sobre la temática y de las características de contexto donde se va a evaluar el programa. El conjunto de estas circunstancias pueden condicionar el uso de unos determinados elementos de diseños u otros. Por ejemplo, en las evaluaciones dirigidas al estudio de los efectos de un programa se suele utilizar el diseño de grupo control no equivalente con sólo una medida previa y posterior a la implementación del programa. De todos modos, tal y como se ha expuesto a lo largo de este apartado los diseños se pueden mejorar incluyendo múltiples medidas previas, variables dependientes no equivalentes, múltiples grupos de comparación, o la manipulación deliberada de las intervenciones implementadas. A su vez, es importante resaltar que estos elementos de diseño deberían ser elegidos una vez estudiadas las potenciales amenazas a la validez, y no antes. En este sentido se ha querido enfatizar que sin el uso de la asignación aleatoria, la realización de evaluaciones válidas se ha de basar en una buena selección de elementos de diseño que posibilite la neutralización del mayor número de posible de amenazas a la validez, lo que a su vez llevará asociado la obtención de un mayor cantidad de datos, de más calidad, a través de los cuales será posible realizar un mayor número de análisis, y más complejos, que sirvan para apoyar o no las inferencias obtenidas.

A modo de conclusión

La evaluación de programas ha estado fuertemente condicionada por la tradición experimentalista, que ha dominado la investigación desde la metodología científi-

ca. Lógicamente la posibilidad de delimitar con precisión nuestro objeto de estudio (por otra parte, casi siempre planteado en términos relacionales-causales), junto con la capacidad absoluta para introducir modificaciones, es decir manipular tanto el mencionado objeto de estudio como su contexto delimitador, hace posible una clasificación estándar de las casuísticas (diseños experimentales), así como una priorización de las posibles alternativas a seguir.

Por el contrario, en evaluación de programas el objeto de estudio (programa de intervención) es complejo (tanto en cantidad como en variedad de variables y sus posibles relaciones), y la responsabilidad de su delimitación no recae exclusivamente en el evaluador, sino más bien en un conjunto de implicados cuyas prioridades pueden estar en ocasiones contrapuestas. Al mismo tiempo dicho objeto de estudio se inserta en un contexto sociopolítico variable que lo modula de forma continua. Por otra parte, la posibilidad de modificar, o manipular, el objeto de estudio viene condicionada por multiplicidad de casuísticas difícilmente controlables por el encargado de realizar la evaluación.

Desde esta situación en este trabajo se ha querido enfatizar que para el diseño de la evaluación de un programa nos basamos más en un criterio de *plausibilidad* del conocimiento de la mayor parte posible de variables que puedan implicar amenazas a la validez de la información. De ahí que los elementos de diseños planteados se basen por una parte en estudiar que posibles amenazas de validez existentes en el contexto de evaluación, y por otra en intentar neutralizar o al menos minimizar sus efectos en la mayor medida posible. Esta circunstancia hace que la calidad de los diseños evaluativos dependa de los estudios previos existentes y de la sistematización de las posibles amenazas a la validez con las que nos podemos encontrar en los distintos ámbitos de intervención. En la mayor parte de los diseños de evaluación no existe un mecanismo ómnibus como la asignación aleatoria que pueda neutralizar las amenazas a la validez de una manera directa. Por tanto los controles se han de basar en los recursos de los diseños planificados directamente ligado a la robustez de las teorías y medidas utilizadas.

Lógicamente además del control conseguido mediante estrategias de diseño podrían utilizarse controles analíticos, no obstante las decisiones de diseño son previas a los análisis de datos y en gran medida condicionarán la calidad de tales análisis. Es por ello que hemos priorizado el estudio de los elementos de diseño; esto no implica una infravaloración de los criterios de análisis, en tanto que en las decisiones de diseño el tipo de análisis posterior y la potencia de éstos también conforman criterios de decisión a la hora de optar por distintas variantes de diseño. Esta argumentación no implica que no se reconozcan técnicas de análisis ya existentes, que en sí mismas pueden depurar los diseños evaluativos, como por ejemplo, el análisis de supervivencia (Yamaguchi, 1991), la regresión múltiple y logística (Reichard y Bormann, 1994), o los modelos lineales jerárquicos (Kreft y de Leeuw, 1998).

Todas las circunstancias descritas nos han conducido a enfatizar la validez del dato desde el que se realiza la evaluación, en vez de abogar por un listado de posibles diseños estándares de difícil aplicabilidad práctica. Lo que sucede es que dicho criterio de validez no es absoluto, sino más bien relativo en tanto supone un juicio valorativo global en el que se han de integrar evi-

dencias empíricas, planteamientos teóricos y criterios pragmáticos utilitaristas. Ante esta situación enfatizamos la necesidad de participación de los implicados en el proceso de intervención-evaluación con objeto de lograr que dicha valoración global sea consensuada (Carey y Smith, 1992; Fetterman, 1997; Lobosco y Newman, 1992; Brandon, Newton y Harman, 1993; Camasso y Dick, 1993), y en la medida de lo posible acorde con las teorías de programación social existentes (Chen, 1990; Chen y Rossi, 1992; Gottfredson, 1984). En resumen, el criterio de validez defendido en este trabajo está directamente relacionado con un criterio de pragmatismo conceptual (Fishman, 1991), en el sentido de lograr que el juicio valorativo realizado sea útil en tanto plantee la recogida de una información válida, necesaria para la resolución de problemas en el contexto en que se desarrolla.

Referencias

Allport, G.W. (1942). *The use of personal documents in psychological science.* Nueva York: Social Science Research.

Alvira, F.(1991). *Metodología de la evaluación de programas.* Madrid: Centro de Investigaciones Sociológicas.

Anguera, M.T. (1989). Innovaciones en la metodología de evaluación de programas. *Anales de Psicología, 5,* 13-42.

Anguera, M.T. (1995a). Metodología de la evaluación: Evaluación cualitativa frente a evaluación cuantitativa. En Equipo del Gabinete Psicotécnico Municipal de Torrent (Recop.) *La evaluación...proceso final?* (pp.27-36). Torrent: Ajuntament de Torrent.

Anguera, M.T. (1995b). Diseños. En R. Fernández-Ballesteros (Ed.), *Evaluación de programas sociales: Una guía práctica en ámbitos sociales, educativos y de salud* (pp. 149-172). Madrid: Síntesis.

Anguera, M.T. y Blanco, A. (1991). *Evaluación de programas en Servicios Sociales: alternativas metodológicas.* Informe de la investigación subvencionada por la Comisión Interministerial de Ciencia y Tecnología (CICYT).

Anguera, M.T. y Chacón, S. (en prensa). Bases metodológicas. En M.T. Anguera Argilaga (Dir.). *Evaluación de programas sociales y sanitarios: un abordaje metodológico.* Madrid: Síntesis.

Arnau, J., Anguera, M.T. y Gómez, J. (1990). *Metodología de la investigación en ciencias del comportamiento.* Murcia: Servicio de publicaciones de la Universidad de Murcia.

Ato, M. (1991). *Investigación en ciencias del comportamiento. I: Fundamentos.* Barcelona: P.P.U.

Blanco, A. y Anguera, M.T. (1991). Sistemas de codificación. En M.T. Anguera (Ed.), *Metodología observacional en la investigación psicológica, Vol. I* (pp. 193-239). Barcelona: P.P.U.

Boruch, R.F. (1997). *Randomized field experiments for planning and evaluation: A practical guide.* California: Sage.

Brandon, P.R., Newton, B.J. y Harman, J.W. (1993). Enhancing validity through beneficiaries' equitable involvement in identifying and prioritizing homeless children's educational problems. *Evaluation and Program Planning, 6,* 287-293.

Camasso, M.J. y Dick, J. (1993). Using multiattribute utility theory as a priority-setting tool in human services planning. *Evaluation and Program Planning, 16,* 295-304.

Campbell, D.T. y Russo, M.J. (1999). *Social experimentation*. Londres: Sage.

Carey, M.A. y Smith, M.W. (1992). Enhacement of validity through qualitative approaches: incorporating the patient's perspectives. *Evaluation and the Health Professions, 15* (4), 107-114.

Cattell, R.B. (1952). The three basic factor analytic research designs - their interrelationships and derivatives. *Psychological Bulletin, 49,* 499-520.

Chen, H. (1990). *Theory-driven evaluations.* Londres: Sage.

Chen, H-T. y Rossi, P.H. (1992). *Using theory to improve program and policy evaluations.* Nueva York: Greenwood Press

Cohen, E. y Franco, R. (1992). *Evaluación de proyectos sociales.* Buenos Aires: Grupo Editor Latinoamericano.

Cook, T.D. y Reichardt, Ch.S. (eds.) (1986). *Métodos cualitativos y cuantitativos en investigación evaluativa* (pp. 25-58). Madrid: Morata.

Engel, J.D. (Ed.) (1992). Issues of methodology in qualitative inquiry. *Qualitative Health Research, 2* (4), número monográfico.

Fernández Ballesteros, R. (1995). *Evaluación de programas. Una guía practica en ámbitos sociales, educativos y de salud.* Madrid: Síntesis.

Fetterman, D.M. (1997). Empowerment evaluation and acreditation in higher education. En E. Chelimnsky y W.R. Shadish. *Evaluation for the 21st Century. A Handbook* (pp. 381-395). Londres: Sage.

Fienberg, S.B. y Tanur, J. (1987). The design and analysis of longitudinal surveys: Controversies and issues of costs and continuity. En R.F. Boruch y R.W. Pearson (Eds.), *Designing research with scarce resources.* Nueva York: Springer-Verlag.

Filstead, W.J. (1986). Una experiencia necesaria en la investifación evaluativa. En T.D. Cook y Ch.S. Reichardt (Eds.), *Métodos cualitativos y cuantitativos en investigación evaluativa* (pp. 59-79). Madrid: Morata.

Fishman, D.B. (1991). An introduction to the experimental versus the pragmatic paradigm in evaluation. *Evaluation and Program Planning, 14,* 353-363.

Gil, J. (1994). *Análisis de datos cualitativos. Aplicaciones a la investigación educativa.* Barcelona: P.P.U.

Gottfredson, G.D. (1984). A theory-ridden approach to program evaluation. A method of stimulating researcher-implementer collaboration. *American Psychologist, 39 (10),* 1101-1112.

Hernández López, J.M. (1995). Procedimiento de recogida de información en evaluación de programas. En R. Fernández-Ballesteros (Ed.), *Evaluación de programas sociales: Una guía práctica en ámbitos sociales, educativos y de salud* (pp. 117-147). Madrid: Síntesis.

Hernández Mendo, A. y Anguera, M.T. (1998). Análisis de coordenadas polares en el estudio de las diferencias individuales de la acción de juego. En M.P. Sánchez y M.A. Quiroga (Coords.), *Perspectivas actuales en la investigación psicológica de las diferencias individuales* (pp. 84-88). Madrid: Centro de Estudios Ramón Areces.

Herrero, M.L. (1989). *Encidencia de la historia personal en el comportamiento en el aula: Estudio observacional analítico.* Tesis Doctoral no publicada. Barcelona: Universidad de Barcelona.

Holland, P.W. y Rubin, D.B. (1988). Causal inference in retrospective studies. *Evaluation Review, 12* (3), 203-231.

Ianni, F.A. y Orr, M.T. (1986). Hacia un acercamiento entre las metodologías cuantitativas y cualitativas. En T.D. Cook y Ch.S. Reichardt (Eds.), *Métodos cualitativos y cuantitativos en investigación evaluativa* (pp. 131-146). Madrid: Morata.

Judd, C. y Kenny, D. (1981). *Estimating the impact of social interventions.* Cambridge M.A.: Cambridge University Press.

Kazdin, A. E. (1980). *Research design in clinical psychology.* Nueva York: Harper y Row.

Kish, L. (1987). *Survey sampling.* Nueva York: John Wiley

Kreft, I. y de Leeuw, J. (1998). *Introducing multilevel modeling.* Londres: Sage.

Little, R.J. y Rubin, D.B. (1987). *Statistical analysis with missing data.* Nueva York: Wiley

Lobosco, A.F. y Newman, D.L. (1992) Stakeholder information needs. Implications for evaluation practice and policy development in early childhood special education. *Evaluation Review, 16 (5),* 443-463.

Marcantonio, R.J. y Cook, T.D. (1994). Convincing Quasi-experiments: The Interrupted Time Series and Regression-Discontinuity Designs. En J.S. Wholey, H.P. Hatry y K.E. NewComer (Eds.) *Handbook of Practical Program Evaluation* (pp.133-154). San Francisco: Jossey-Bass.

Martínez Arias, R.M. (1995). *Psicometría: teoría de los tests psicológicos y educativos.* Madrid: Síntesis

Martínez-Arias, R.M. (1996). Metodología de la investigación en drogodependencias (Programa docente). Universidad Complutense de Madrid.

Nesselroade, J.R. (1988). Sampling and generalizability: Adult development and aging research issues examined within the general methodological framework of selection. En K.W. Schaie, R.T. Campbell, W. Meredith y S.C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 13-42). Nueva York: Springer.

Nesselroade, J.R. (1991). Interindividual differences in intraindividual changes. En J.L. Horn y L. Collins (Eds.), *Best methods for measuring change.* Washington: American Psychological Association.

Nesselroade, J.R. y Hershberger, S.L. (1993). Intraindividual variability: Methodological issues for population health research. En K. Dean (Ed.), *Population health research. Linking theory and methods* (pp. 74-94). Londres: Sage.

Owen, J.M. y Rogers, P.J. (1999). *Program evaluation. Forms and approaches.* Londres: Sage.

Plewis, I. (1985). *Analysing change. Measurement and explanation using longitudinal data.* Nueva York: Wiley.

Posavac, E.J. y Carey, R.G. (1985). *Program evaluation. Methods and case studies.* Nueva York: Prentice-Hall.

Reason, P. y Lincoln, Y.S. (Eds.) (1996). Quality in Human Inquiry (Special issue). *Qualitative Inquiry, 2* (1), número completo.

Redondo, S. (1992). *Evaluar e intervenir en prisiones.* Barcelona: P.P.U.

Reichard, C. y Bormann, C. (1994). Using regression models to estimate program effects. En J.S.Wholey, H.P. Hatry, K.E. NewComer (Eds.)

Handbook of Practical Program Evaluation (pp. 417-455). San Francisco: Jossey-Bass.

Shadish, W., Cook, T. y Campbell, D (en preparación). *Experimental and quasiexperimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Scriven, M. (1980). *The logic of evaluation*. California: Edgepress.

Smith, G. y Cantley, C. (1985). Policy evaluation: The use of varied data in a study of a psychogeriatric service. En R. Walker (Ed.), *Applied qualitative research* (pp. 156-174). Aldershot: Gower.

Tuset, A. (1990). *Análisis de las respuestas al test de Rorschach de un grupo de sujetos afectados de un primer infarto de miocardio*. Tesis Doctoral no publicada. Barcelona: Universidad de Barcelona.

Vedung, E. (1993). Utilización de la evaluación. *Revista de Servicios Sociales y Política Social, 2,* 69-80.

Vedung, E. (1996). *Public policy and program evaluation*. Londres: Transaction Publishers

Veney, J.E. y Kaluzny, A.D. (1984). *Evaluation and decision making for health services program*. Nueva York: Prentice Hall.

Weiss, C.H. (1983). The stakeholder approach to evaluation: origins and promise. En A.S. Bryk (Ed.) *Stakeholder-based evaluation* (pp. 3-14). San Francisco: Jossey-Bass.

Wholey, J.S. (1987). Evaluability assessment: developing program theory. En L. Bickman (Ed.) *Using program theory* (pp. 77-93). San Francisco: Jossey-Bass.

Yamaguchi, K. (1991). *Event History Analysis*. Londres: Sage.